UNIVERSIDAD DE COSTA RICA

SISTEMA DE ESTUDIOS DE POSGRADO

IMPLEMENTACIÓN DE MODELOS ESTADÍSTICOS PARA LA ESTIMACIÓN DE LA DEMANDA DE COMBUSTIBLES EN COSTA RICA

Trabajo final de investigación aplicada sometido a la consideración de la Comisión del Programa de Estudios de Posgrado en Estadística para optar al grado y título de Maestría Profesional en Estadística

ALLAN QUESADA ROJAS

Ciudad Universitaria Rodrigo Facio, Costa Rica

2022

DEDICATORIA

A Dios, a quién debo todo, pues con su mano me ha sostenido, me ha cuidado, y con su apoyo me ha permitido salir adelante.

A mi esposa, a quién amo de un modo incondicional, pues con su amor, nobleza y apoyo, ha luchado por cada proyecto que hemos emprendido y llenado de alegría cada paso.

A mis padres y familia, quienes me formaron con los valores más grandes, que guardo siempre en mi corazón, además han sido un soporte y un ejemplo de inspiración.

A nuestros futuros hijos, que desde ya los esperamos con amor.

AGRADECIMIENTOS

Agradezco enormemente al profesor Guaner, por todo su apoyo, anuencia y orientación. Desde el inicio me apoyó e instruyó para que este proyecto fuera una realidad.

A Marcela y a Karla por toda la ayuda brindada, por esos aportes que enriquecieron el proceso, siempre con gran admiración he agradecido cada contribución.

A mi familia porque siempre me han apoyado y han sido mi soporte.

A mi esposa, por ser mi mayor bendición y mi motor en la búsqueda de mis sueños.

"Este trabajo final de investigación aplicada fue aceptado por la Comisión del Programa de Estudios de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Estadística"

GUANER DAVID ROJAS ROJAS (FIRMA)
PERSONA FISICA, CPF-05-0328-0120.
Fecha declarada: 16/09/2022 01:27:59 PM
Razón: Visto Bueno
Lugar: Montes de Oca Contacto: guaner.rojas@ucr.ac.cr

Dr. Guaner Rojas Rojas

Profesor guía

MARCELA ALFARO CORDOBA (FIRMA)
Digitally signed by MARCELA ALFARO CORDOBA (FIRMA)
Date: 2022.09.23 17:49:06 -07'00'

Dra. Marcela Alfaro Córdoba

Lectora

KARLA PATRICIA MONTERO VIQUEZ (FIRMA)
Firmado digitalmente por KARLA PATRICIA MONTERO VIQUEZ (FIRMA)
Fecha: 2022.09.22 06:59:11 -06'00'

MSc. Karla Montero Víquez

Lectora

ALLAN GERARDO QUESADA ROJAS (FIRMA)
Firmado digitalmente por ALLAN GERARDO QUESADA ROJAS (FIRMA)
Fecha: 2022.08.31 14:23:08 -06'00'

Allan Quesada Rojas

Sustentante

TABLA DE CONTENIDO

RESUMEN

La pandemia del Covid-19, ha constituido un reto enorme para la humanidad, generando importantes desafíos en las diversas áreas del conocimiento, y la Estadística no es la excepción, pues muchos modelos, requieren el desarrollo de diversas modificaciones en su tratamiento, para la adecuada internalización de este cambio tan importante en el contexto.

De ahí nace la motivación del presente trabajo, el cual busca generar un aporte, que permita dar diversas alternativas para la estimación de series de tiempo o cronológicas, las cuales han sido expuestas a una perturbación fuerte como es el caso de la pandemia, que ocasiona importantes errores al momento de pronosticar el comportamiento a futuro de la serie.

En este estudio, se empleó la información de la demanda de combustible de Costa Rica de 2010 a 2020 en litros para los productos Súper (RON 95), Regular (RON 91) y Diésel (Diésel 50 ppm), que son los combustibles de mayor consumo a nivel nacional, son de gran importancia en la economía del país, y han sido una de las demandas más afectadas por la pandemia. Para lograr lo anterior, se desarrollaron diversos modelos estadísticos de series de tiempo, como por ejemplo los modelos de suavizamiento exponencial, modelos ARIMA, modelos de Series de Tiempo Estructurales Bayesianas y modelos de minería de datos como XGBoost y el uso de Vectores de Soporte de Regresión, empleando datos estadísticos de las series de ventas de combustible, datos de movilidad de Google y datos asociados a la evolución del Covid-19.

El fin de utilizar este conjunto de modelos, fue aprovechar las diversas características y bondades que cada uno de ellos ofrece, y de este modo poder enfrentar un escenario de estimación complejo, comprendiendo que, en diversas ocasiones los mejores resultados se obtienen de emplear diversos modelos, por ejemplo para la gasolina regular, el mejor resultado se obtuvo por medio de un modelo XGBoost (7,55% de Error Absoluto Medio), mientras que para el Diésel el mejor resultado fue del modelo Bayesiano (8,84% de Error Absoluto Medio). Además, se observó que en ocasiones el trabajo conjunto de todas estas técnicas puede ser una gran alternativa, como se observó en el caso de la gasolina Súper, en la cual se empleó el modelo conjunto estimado por medio de Vectores de Soporte de Regresión, utilizando como variables explicativas, las estimaciones de los otros modelos y permitió obtener los mejores resultados con 9,91% de Error Absoluto Medio.

ABSTRACT

The Covid-19 pandemic has constituted a huge challenge for humanity, generating important challenges in the various areas of knowledge, Statistics is no exception, since many models require the development of various modifications in their treatment, for the adequate internalization of this important change in context.

Hence the motivation for this work was born, which seeks to generate a contribution, that will allow giving various alternatives for the estimation of time series, which have been exposed to a strong disturbance such as the pandemic, which causes important errors when forecasting the future behavior of the series.

In this study, the information on the demand for fuel in Costa Rica from 2010 to 2020 in liters was used for the Super (RON 95), Regular (RON 91) and Diesel (Diesel 50 ppm) products, which are the most popular and important in the country's economy and have been one of the demands most affected by the pandemic. To achieve this, various statistical models for time series were used, such as exponential smoothing models, ARIMA models, Bayesian Structural Time Series models and data mining models such as XGBoost and the use of Support Vectors of Regression, using statistical data from the fuel sales series, Google mobility data and data associated with the evolution of Covid-19.

The purpose of using this set of models was to take advantage of the various characteristics and benefits that each of them offers, and thus be able to face a complex estimation scenario, understanding that, on various occasions, the best results are obtained from using different models, for example, for regular gasoline, the best result was obtained through an XGBoost model (7.55% Mean Absolute Error), while for Diesel the best result was from the Bayesian model (8.84% Absolute Error Medium). In addition, it was observed that sometimes the joint work of all these techniques can be a great alternative, as was observed in the case of Super gasoline, in which the joint model estimated by means of Regression Support Vectors, using as explanatory variables, the estimates of the other models and allowed to obtain the best results with 9.91% of Mean Absolute Error.

LISTA DE CUADROS

LISTA DE GRÁFICOS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

AIC	Criterio de información de Akaike
Aresep	Autoridad Reguladora de los Servicios Públicos
BCCR	Banco Central de Costa Rica
BIC	Criterio de información Bayesiano
BSTS	Modelos de Series de Tiempo Bayesianas Estructurales
Covid-19	Enfermedad causada por el nuevo coronavirus conocido como SARS-CoV-2
MAPE	Error Porcentual Absoluto Medio
RECOPE	Refinadora Costarricense de Petróleo
RMSE	Raíz del Error Cuadrático Medio
ARIMA	Modelo estacional autorregresivo integrado de media móvil
SVR	Vectores de Soporte de Regresión

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Allan Quesada Rojas, con cédula de identidad 207020415, en mi condición de autor del TFG titulado "Implementación de modelos estadísticos para la estimación de la demanda de combustibles en Costa Rica".

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** [X] **NO** *[ ]

***En caso de la negativa favor indicar el tiempo de restricción: _____ año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

ALLAN GERARDO QUESADA ROJAS (FIRMA)

Firmado digitalmente por ALLAN GERARDO QUESADA ROJAS (FIRMA)
Fecha: 2022.08.31 14:24:27 -06'00'

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

1. INTRODUCCIÓN

Los hidrocarburos constituyen la principal fuente de energía de la humanidad (Frey et al, 2009, pág. 30), por esta razón, los escenarios y proyecciones energéticas son de gran importancia para la toma de decisiones políticas y estratégicas (Culka, 2016, pág. 2), pues logran influir de modo considerable en el crecimiento económico y social de los diversos países, por ello académicos y expertos en energía han trabajado en diversos estudios analizando los determinantes de los niveles anteriores, actuales y futuros de la demanda y precio del petróleo (Frey et al, 2009, pág. 29).

Los precios de los hidrocarburos influyen de modo importante en la determinación de costos operativos de diversas empresas a nivel mundial, como por ejemplo las aerolíneas (Alquist et al., 2011, pág. 3), y la industria del transporte en general, en la cual, el gasto en combustibles llega a alcanzar hasta un 32.5% del costo variable operativo por kilómetro (Ruiz et al, 2014, pág. 239).

En la actualidad, la pandemia del Covid-19, ha ocasionado una disminución sustancial en la demanda de combustibles, nunca antes vista en la historia, es por ello, que tal y como indica Koyama & Suehiro (2020), las estimaciones que se desarrollen sobre el comportamiento de dicha variable resultan de gran relevancia y se encuentran propensas a altos niveles de error y complicación al momento de la estimación.

Además, esta fuerte reducción en el consumo de los hidrocarburos está ocasionando una disminución en los ingresos de la Refinadora Costarricense de Petróleo (RECOPE), así como de las empresas encargadas del flete de combustible, Peddler y las estaciones de servicio ubicadas en todo el país. Debido a lo anterior, es fundamental realizar una adecuada estimación de la demanda de gasolina súper, regular y diésel, en especial en periodos tan complejos como los acontecidos, los cuales tal y como mencionan Koyama & Suehiro (2020, pág. 8) podrían ocasionar una reducción importante en el consumo global de hidrocarburos que rondaría según sus estimaciones entre -9.3% y -12.8%.

A su vez, tal y como indica Rystad Energy (2021), el sector de los hidrocarburos es de los más afectados ante la pandemia del Covid-19 y muestra una alta correlación con la evolución

de los casos de la enfermedad, lo anterior, se explica en gran medida, pues las acciones y restricciones a la movilización, influyen de un modo directo y en una magnitud relevante sobre el consumo de combustibles, por la reducción en el transporte de personas.

De la revisión de referencias bibliográficas, se logró observar que, en condiciones regulares, la aplicación de modelos de series de tiempo ARIMA han ofrecido una adecuada estimación de la demanda energética, lo cual se evidencia en los trabajos desarrollados por Khedmatia & Ghalebsaz-Jeddib (2018), Bhattacharyya y Timilsina (2009) y Alquist et al., (2011). Sin embargo, utilizar un modelo de Series de Tiempo Estructurales Bayesianas tal y como se desarrolló en los estudios Li & Ngan (2019), Suzuki et al., (2020), Kitamura (2018) y Scott & Varian (2014), permitiría modelar y capturar los cambios estructurales como el ocasionado por el Covid-19 y lidiarían con el problema de tener relativamente pocos datos que representen este cambio de comportamiento en el consumo.

De igual manera, es importante mencionar que el uso del modelo XGBoost aplicado en series de tiempo, contribuye en la modelación de relaciones no lineales, con lo cual se podrían mejorar los pronósticos de series de tiempo, principalmente en estimaciones de demanda del sector energético que son propensas a tener este tipo de comportamiento no lineal, tal y como se muestra en los trabajos de Garnier & Belletoile (2019) y Zhou, Li, Shi, & Qian (2019).

Por último, con base en la experiencia de modelos híbridos como los desarrollados por Pai & Lin (2005) y Ogcu, Demirel, & Zaim (2012), se ha observado que su uso podría mejorar la capacidad predictiva del modelo, debido en gran medida a su la versatilidad y robustez. Al respecto, una referencia importante es el abordaje desarrollado por Barbosa de Alencar et al., (2017), en el cual, los resultados de otros modelos estimados por redes neuronales y ARIMA, se utilizaron como insumo de un modelo de minería de datos, lo cual permitió aprovechar los beneficios de cada tipo de modelo para realizar las estimaciones y según los resultados obtenidos en su trabajo, dicho modelo obtuvo los menores errores de predicción en todas las estimaciones.

Por todo lo anterior, se considera relevante desarrollar el trabajo en este periodo relativamente complejo para las estimaciones de demanda de hidrocarburos, y dado que no

hay evidencia de trabajos en los cuales se haya desarrollado un análisis similar para Costa Rica, resulta novedoso el abordaje y se podrían aprovechar las bondades de la metodología desarrollada en el presente trabajo para lograr mejorar las estimaciones empleadas en el cálculo de precios internos y la elaboración de políticas públicas energéticas que se realicen en el país.

A continuación, se referencia y conceptualiza el tema a tratar, se describe los datos y metodología utilizados, se presentan los principales resultados y se finaliza el documento con un apartado de conclusiones.

2. OBJETIVOS

Este proyecto se realiza con los objetivos que se exponen a continuación.

2.1. Objetivo general

Implementar un conjunto de modelos de estimación para la demanda de combustibles en Costa Rica con datos de 2010 a 2020.

2.2. Objetivos específicos

1. Describir el comportamiento de la demanda de combustible gasolina súper, regular y diésel en Costa Rica, para los años 2010 a 2020.

2. Aplicar modelos de series de tiempo, bayesianos y de minería de datos para la estimación de la demanda de la gasolina súper, regular y diésel de Costa Rica.

3. Generar un modelo de minería de datos que utilice como variables explicativas las estimaciones de demanda de gasolina súper, regular y diésel de otros modelos.

3. MARCO TEÓRICO

El sector energético de una economía es de vital importancia pues constituye un insumo fundamental en los procesos productivos, dentro de este sector se encuentra el subconjunto asociado a los combustibles derivados de hidrocarburos, que está delimitado específicamente por aquellas fuentes energéticas basadas en petróleo y sus derivados, las cuales se utilizan en su mayoría para el transporte. Por lo anterior, en esta sección se desarrolla un marco de referencia sobre los estudios relacionados con el efecto ocasionado por el Covid-19 sobre la demanda de combustibles derivados de hidrocarburos, así como los diversos trabajos de investigación desarrollados para la estimación de la demanda energética, es importante aclarar que, pese a que muchas de las investigaciones están basadas en el sector energético como un todo, son aplicables al subconjunto de hidrocarburos.

3.1.Efectos del Covid-19 sobre la demanda de hidrocarburos

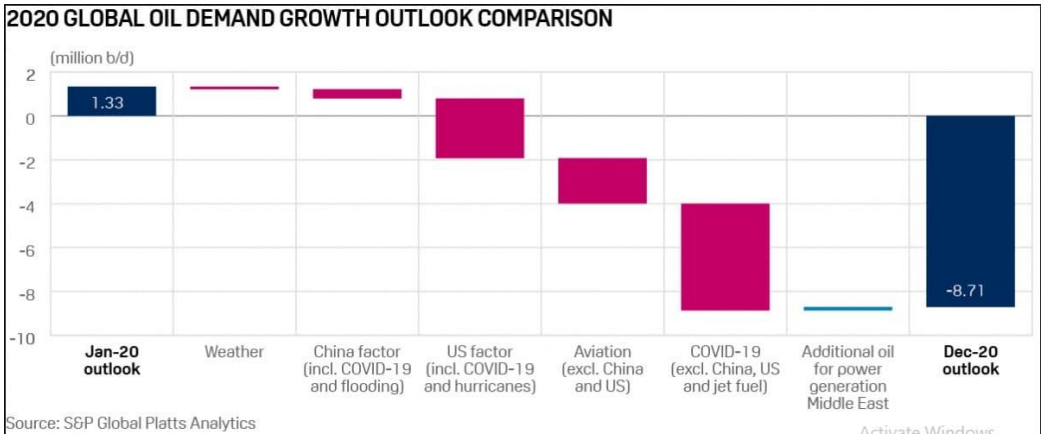
La Pandemia fue una disrupción importante en la demanda de una amplia variedad de productos y servicios. Para algunos sectores, esto fue una interrupción a corto plazo, posteriormente la demanda retornó a niveles anteriores en 6-12 meses, mientras que para otros sectores esto fue un cambio estructural con un impacto a largo plazo en forma de aumento/caída sostenida de la demanda (Sagar, 2021).

Se han realizado diversos estudios tratando de determinar los posibles efectos del Covid-19 en la demanda de los diversos productos y servicios, para Japón por ejemplo se estimaron los impactos en algunas industrias, y se observó que, para el caso de combustibles se esperaba una disminución cercana al -6%, pues el factor principal para la determinación del posible impacto, está relacionado especialmente con la restricción del movimiento y el cambio en el hábitos y actividades de las personas (Suzuki et al., 2020, pág. 7).

De este modo, los brotes epidémicos conducen al endurecimiento de restricciones de movimiento o de una mayor duración, que a su vez repercuten en una contracción de la demanda y por consiguiente de la actividad económica (Arce, 2020, pág. 83). De este modo, es claro que el nivel de afectación no es homogéneo entre industrias, pues aquellas que dependan en mayor medida de la movilidad de las personas, serán las más afectadas, por ello,

el consumo de combustibles es una de las demandas con mayor afectación, lo cual se puede evidenciar en la siguiente gráfica que muestra la disminución global en la demanda de combustibles de enero a diciembre de 2020:

Gráfico 1. Variación en millones de barriles por día, de la demanda mundial de petróleo, según diversos tipos de factores explicativos, enero a diciembre 2020.



Fuente: extraído de (Rizvi, 2021, pág. 1).

Esta disminución en la demanda de hidrocarburos a nivel mundial repercutió en los precios internacionales, los cuales experimentaron los mínimos históricos, dado que, en el segundo semestre de 2020, se presentó un exceso de oferta de petróleo sin precedentes, el cual se fue reduciendo en el tiempo (Rizvi, 2021, pág. 1).

De este modo, el aumento de casos por Covid-19 y las consecuencias asociadas a los bloqueos y las medidas de distanciamiento social llevaron a una desaceleración de la demanda en el cuarto trimestre de 2020 y pronósticos reservados para 2021 (Smith, 2020).

Pese a ello, Messler (2021) indica que, en inicios de 2021, los precios del petróleo se han recuperado hasta el punto en que lograron igualar los niveles anteriores al Covid-19. Hay dos factores clave para entender esta recuperación: 1) El primero son las restricciones a la producción de los productores estadounidenses y del cartel OPEP +. 2) El segundo es la recuperación de la demanda, la cual superaría a la oferta en junio de 2021, aunque no lograría alcanzar en plenitud los niveles pre pandémicos.

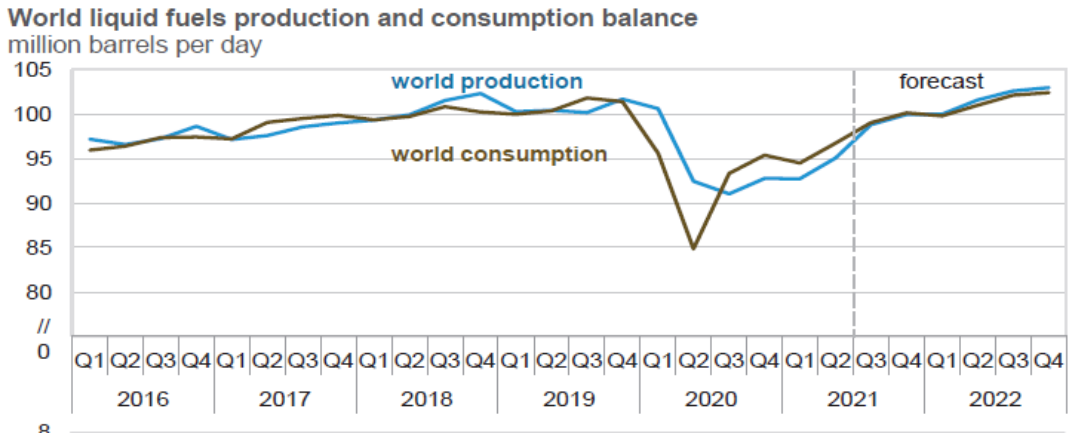
Para comprender esta dinámica tan compleja y estrechamente vinculada, es necesario analizar los factores que están influyendo en la oferta y la demanda del petróleo, y que por consiguiente están afectando al mercado internacional. En primera instancia, se procederá a analizar los factores que están influyendo sobre la oferta.

El informe de Primary Vision Network (PVN) dice que son "cautelosamente optimistas" con respecto a la recuperación de la actividad de exploración y producción en 2021, lo anterior, pues existen altos niveles de incertidumbre en el mercado y se puede esperar que los productores sigan buscando ahorros de costos siempre que sea posible, ya que durante 2020 la mayoría de estas empresas dedicadas a la extracción de petróleo tuvieron pérdidas importantes, y en 2021, están tratando de sanar sus balances (Rizvi, 2021).

Es por esta razón que los productores estadounidenses han prometido repetidamente que los días de crecimiento a cualquier precio quedaron en el pasado y sus objetivos son mantener los niveles de producción actuales o mantener la tasa de explotación a un nivel rentable. En lugar de crecer, los productores se han centrado en reparar los balances dañados a causa de las amortizaciones masivas de activos durante los últimos dos años y recompensar a los inversores pacientes con dividendos más altos a medida que se expanden los márgenes (Messler, 2021). En resumen, por parte de la oferta, se espera un comportamiento cauteloso, buscando en la medida de lo posible no generar un exceso de oferta que pueda presionar el precio a la baja, pues el objetivo de estas empresas es tener un nivel de rentabilidad cercano a las metas pactadas con sus accionistas.

En relación con la demanda, la Agencia Internacional de Energía (EIA), espera que en 2021 el consumo de combustible destilado sea aproximadamente igual a los niveles de 2019 mientras que en 2022, espera que el consumo de destilados supere los niveles de 2019. (EIA, 2021, pág. 9), lo anterior, se puede evidenciar en la siguiente gráfica:

Gráfico 2. Diferencial esperado entre la oferta y demanda de petróleo, enero 2016 a diciembre 2022.

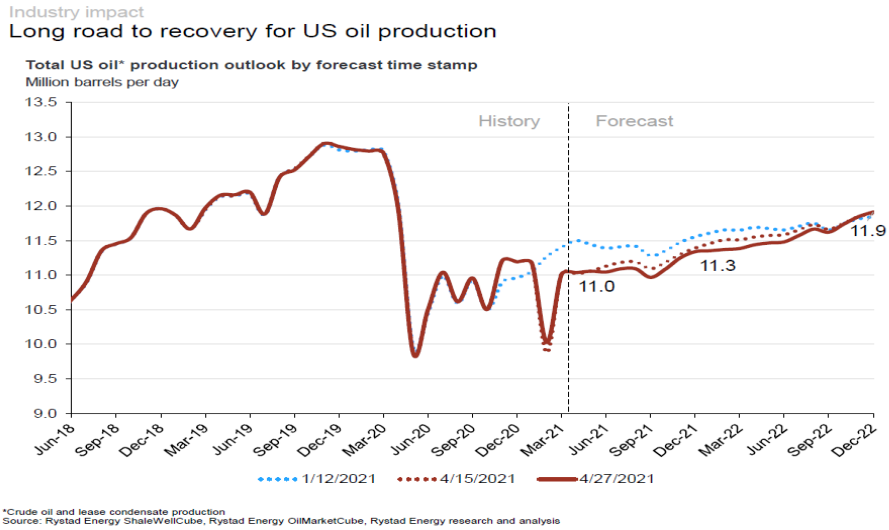


Fuente: Extraído de (EIA, 2021, pág. 24)

Sobre este punto Rizvi (2021) indica que la Agencia Internacional de Energía (EIA) espera que la recuperación de la demanda de petróleo sea más lenta en 2021 de lo que se pensaba anteriormente. Con los informes de nuevos bloqueos en Europa, la demanda en el sector del transporte probablemente se verá afectada. A su vez, dicho autor indica que el análisis de S&P Global Platts sugiere que la demanda será 2,4 mbpd más baja que los niveles de 2019.

Lo anterior, también es compartido por Rystad Energy (2021), quienes consideran que la evolución de la pandemia aún mantiene un nivel de incertidumbre importante, pues pese a que se ha observado que la demanda ha logrado una importante recuperación en 2021, esta ha tenido ciertos deslices, pues se ha observado que la vacunación no ha avanzado tasas muy elevadas, y se observa poco porcentaje de la población con 2 dosis; incluso en algunos países se observa un nivel de estancamiento en la vacunación, que ocasiona que la reducción de los efectos del Covid-19 no se logren evidenciar en el corto plazo. Además, como se muestra en la siguiente gráfica, las proyecciones sobre la recuperación de la oferta de petróleo estadounidense han sufrido dos correcciones a la baja, explicados por la inestabilidad observada en el crecimiento de la demanda y porque este nivel de incertidumbre no propicia el aumento de la producción de EEUU, que fue de las más golpeadas a nivel financiero durante 2020.

Gráfico 3. EEUU, producción de petróleo en millones de barriles por día, junio 2018 a diciembre 2022.



Fuete: Extraído de (Rystad Energy, 2021, pág. 32)

Pese a los pronósticos de Platts, de la EIA y de Rystand Energy, existe otra tendencia, encabezada por la Organización de Países Exportadores de Petróleo y sus aliados (OPEP+) quienes estiman que la demanda aumentará en 5,95 millones de barriles por día (bpd) en 2021 lo que representa en términos porcentuales un aumento de 6,6%, inclusive dicha organización indica que la recuperación económica mundial continúa, respaldada significativamente por un estímulo fiscal y monetario muy fuerte y que la recuperación se inclina mucho hacia la segunda mitad de 2021 (CNBC, 2021).

Por lo anterior, este cartel de producción de petróleo, aún continua con su plan de recortes, a fin de no generar un exceso de oferta, pero propicia la reducción gradual de estas cuotas, inyectando cada mes una mayor cantidad de petróleo, con la confianza de que la demanda se mantendrá creciendo pese a los declives que pueda ocasionar el Covid-19.

Esta dinámica del mercado internacional es de vital importancia para Costa Rica al ser un país 100% importador de combustibles derivados de hidrocarburos, además este proceso de importación, distribución y comercialización mayorista está bajo un monopolio estatal, que ha otorgado a Recope la exclusividad para el desarrollo de estas actividades. Por ello, el

análisis de las ventas de este operador aproxima de una manera razonable la demanda nacional de combustibles derivados de hidrocarburos.

Hecho este resumen inicial sobre los efectos que el Covid-19 está ocasionando en el mercado internacional, es claro que la pandemia influye de un modo muy importante en el mercado de los hidrocarburos, especialmente en la demanda, la cual depende en gran medida del transporte y movilidad de personas.

Para evidenciar lo anterior, se pueden analizar los datos del año 2019, en donde se observó que de las 4 350 millones de toneladas de petróleo que fueron procesadas en refinerías a nivel mundial, cerca del 61% fue empleada para el transporte (IEA, 2019), sin embargo, para el caso de Costa Rica, este porcentaje es significativamente superior, al rondar cerca del 78% (SEPSE, 2019), lo cual demuestra como la movilidad es un determinante fundamental de la demanda de combustibles derivados de hidrocarburos, y de un modo especial en Costa Rica, donde su vinculación es aún más fuerte.

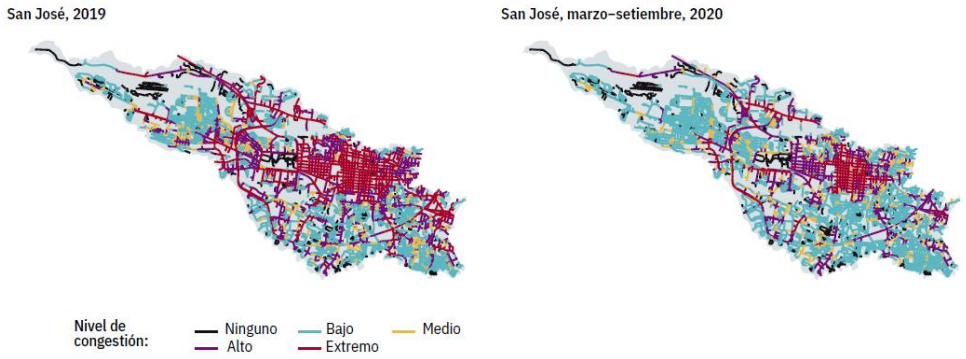
Por lo anterior, se procederá a analizar el comportamiento de la movilidad durante el periodo pandémico y se empleará como una variable explicativa en el análisis de la demanda de las gasolinas y el diésel.

En esta línea de análisis, se estudiaron informes relacionados con la movilidad en Costa Rica, y se concluyó que, durante el año 2020, se presentó una importante reducción en la congestión vehicular tal como se indica a continuación:

> Las restricciones a la movilidad impuestas por el Poder Ejecutivo para el control de la pandemia redujeron la intensidad de la congestión al grado que rompieron la relación territorial de esa movilidad a partir de abril, de acuerdo con los análisis estadísticos ejecutados (Programa Estado de la Nación, 2020, pág. 231).

Lo anterior, se evidencia en los siguientes mapas comparativos de la congestión vehicular (jams):

Figura 1. Cantidad de jams promedio por segmento[a/] en el cantón central de San José, según nivel de intensidad, 2019-2020.

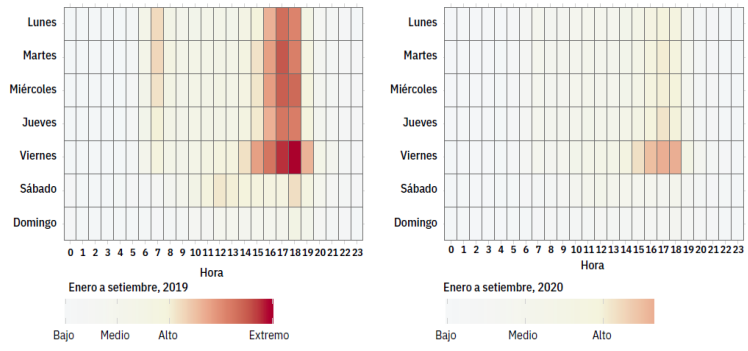


a/ Se entiende por jams el conteo de congestión por segmento de carretera (cien metros aproximadamente) reportado por la aplicación Waze para un lapso de tiempo definido.
Fuente: Gómez Campos et al., 2020, con datos de Waze-MOPT.

Fuente: Extraído de (Programa Estado de la Nación, 2020, pág. 240)

Complementando lo anterior, se presentan los siguientes gráficos de calor que muestran la comparación de los datos de 2019 y 2020, evidenciando una clara disminución de la congestión vehicular por día y por semana, explicados por la reducción drástica de la movilidad posterior al inicio de la pandemia en Costa Rica.

Gráfico 4. Costa Rica, nivel de congestión vial por día de la semana, según hora, enero a diciembre 2020.



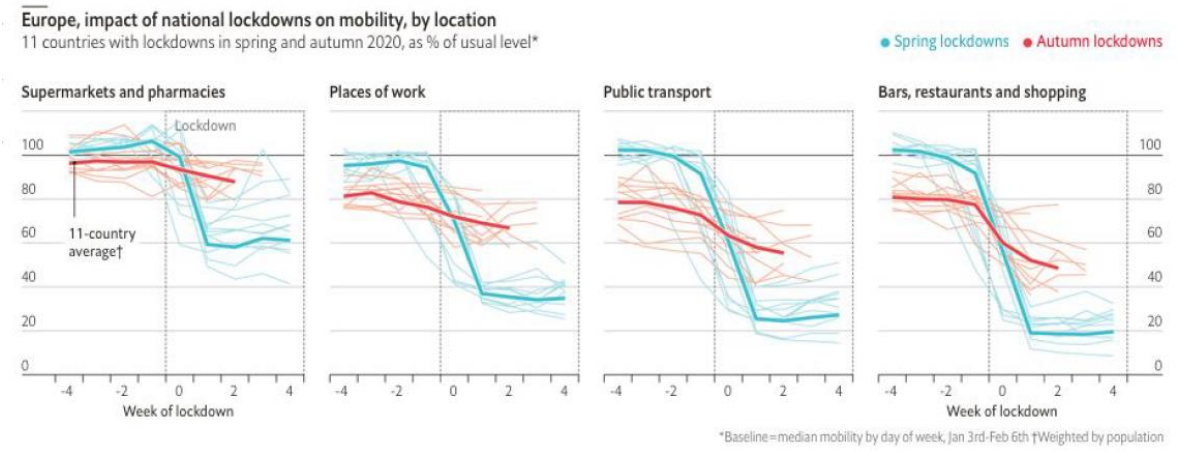
Fuente: Gómez Campos et al., 2020, con datos de Waze-MOPT.

Fuente: Extraído de (Programa Estado de la Nación, 2020, pág. 241)

El impacto del Covid-19 para nuestro país, también se evidencia en resultados como los externados por Valverde (2020), quien indica que la recaudación fiscal en Costa Rica durante 2020 fue menor a la de periodos anteriores, indicando que, a setiembre de 2020 los recursos por impuestos cayeron un -11,46% respecto al mismo mes de año pasado. Por tipo de impuestos esta fue la situación de los principales: IVA: decreció ¢67 196 millones (-5,76%), Impuesto de Renta: cayó ¢118 695 millones (-8,93%,) y el Impuesto de Combustibles: cayó ¢70 635 millones (-17,52%). Nótese cómo la pandemia ocasionó una disminución de la actividad económica que afectó la recaudación de impuestos indirectos asociados al consumo, afectó también la recaudación por rentas o utilidades reportadas, pero sobre todo afectó la recaudación del Impuesto de Combustibles, explicado por una disminución importante en el consumo y por ende la importación de estos bienes, en el periodo pandémico, evidenciando el importante efecto de las restricciones de movilidad (Arce, 2020, pág. 8).

En conclusión, el nivel de incertidumbre en la demanda es muy alto, y estará condicionado por la evolución del Covid-19 y las restricciones a la movilidad que se desarrollen, en este aspecto, es importante mencionar que la respuesta de la movilidad, ante las restricciones sanitarias está cambiando en el tiempo; como se observa en la siguiente gráfica para Europa, durante primavera (línea azul), se observó una clara reducción en los patrones de movilidad observados, sin embargo durante Otoño (línea roja) la movilidad se redujo ante las restricciones sanitarias, pero en una magnitud menos pronunciada. Esto nos indica que las personas modifican en menor medida sus patrones de movilidad, ante las restricciones sanitarias, comenzando a acostumbrarse a una “nueva normalidad”, en la cual se respetan las condiciones sanitarias, pero tratan de continuar en la medida de lo posible con su movilidad habitual.

Gráfico 5. Europa, comparación del índice de movilidad durante las restricciones sanitarias entre primavera y otoño de 2020.



Fuente: Extraído de (Arce, 2020, pág. 4)

De este modo, se observa que la movilidad está logrando converger y tener un patrón menos volátil, al tiempo que las actividades productivas tienden a acomodarse a estas nuevas condiciones y la sociedad trata de buscar un nuevo equilibrio en la forma de organización y producción, a fin de no estancarse en una pandemia, cuyo final aún no es previsible. Este punto es de gran importancia, y fue analizado por Koyama & Suehiro (2020), quienes consideran que existen dos escenarios que se deben analizar al momento de proyectar la demanda de combustibles, el primero en el que se asume una pandemia de mediana duración, y otra que asume una pandemia de más larga duración, llama la atención como se observan diferencias de cerca de 25% en los efectos sobre la demanda entre un escenario y otro (Koyama & Suehiro, 2020, pág. 4).

A nivel mundial se observa que, pese a que las vacunaciones han aumentado, el porcentaje de la población vacunada, aún sigue siendo bajo, y las proyecciones manejadas por Rystad Energy (2021), es que se llegará al 50% en la mayoría de países hasta el último trimestre de 2021, y al 75% de vacunados hasta el 2022, y según los resultados observados en los últimos días se observa que sobrepasar el umbral del 60% ha sido complicado en la mayoría de países, con lo cual los porcentajes de vacunación tienden a estancarse cerca de dicho monto y con

las nuevas variantes gestadas, se observa una nueva ola pandémica en la mayoría de países, lo cual reafirma los altos niveles de incertidumbre sobre la evolución de la pandemia y de la demanda de hidrocarburos.

Cuadro 1. Mes esperado para alcanzar determinados porcentajes de población vacunada según la región, 2021.

Expected month of vaccination thresholds
Dark text: Current report; Light text: Previous report

Region	25% vaccinated		50% vaccinated		75% vaccinated	
East Asia	Aug-21	Jul-21	Dec-21	-	Jun-22	May-22
Middle East	Sep-21	-	Jan-22	-	May-22	Apr-22
North America	Apr-21	-	Jul-21	-	Sep-21	-
South America	Jul-21	Jun-21	Oct-21	-	Mar-22	Jan-22
South Asia	Aug-21	Jul-21	Nov-21	-	Mar-22	Feb-22
Southeast Asia	Aug-21	-	Dec-21	Nov-21	Apr -22	Mar-22
Western Europe	May-21	-	Jul-21	-	Oct-21	Sep-22

Fuente: Extraído de (Rystad Energy, 2021, pág. 20)

Un último elemento para considerar en el análisis de la demanda de combustible es el cambio en los gustos y preferencias de los agentes económicos, por ejemplo la continuación de funciones teletrabajables y el avance de la movilidad eléctrica que comienza a sustituir los combustibles fósiles, cuyo nivel de penetración aún es bajo, sin embargo, se han dado importantes avances, por ejemplo: Volkswagen y CATL están produciendo baterías que han alcanzado el umbral de 100 dólares, mientras que Elon Musk anunció que fabricarán sus propias celdas de batería, con una reducción de costos de más del 50%, al tiempo que EEUU, está valorando empezar a extraer su propio litio (Fuentes, 2020).

Todo lo anterior, en resumen, nos muestra que la proyección de la demanda de combustible tuvo un cambio drástico a partir de 2020, que ocasionó que lo modelos enfrentarán un escenario muy complejo a nivel de predicción. Esto se evidenció en el pico de la pandemia, pues la mayoría de los modelos de series de tiempo en producción no vieron la repentina

caída de la demanda. En pocas palabras, la mayoría de los modelos de pronóstico, que fueron entrenados antes de la pandemia, fueron de poca utilidad y señalaron la necesidad de reentrenarlos con una nueva perspectiva que atienda los cambios de política de los gobiernos y la reacción del público ante los mismos (Sagar, 2021).

Sagar (2021) indica que los modelos de pronóstico ahora deben utilizar un enfoque de múltiples frentes para tomar en cuenta una amplia variedad de factores externos, como los cambios en las políticas gubernamentales y su impacto posterior, además, debe haber un sistema para generar alertas cuando la salud del modelo se deteriore más allá de las normas aceptables. En ese momento, se necesita una intervención oportuna para verificar si el deterioro en el estado del modelo se debe al proceso de generación de datos subyacente, que generalmente se observa a través de cambios en la distribución, o debido a choques temporales en el sistema. De este modo, es necesario la modificación de los modelos de estimación para que: 1) se ajusten a un nivel de incertidumbre y cambios fuertes como los ocasionados por la pandemia, 2) puedan aprovechar al máximo los activos de información existentes y generados por diversas fuentes, 3) tengan una versatilidad que permita incorporar diversos efectos y 4) permitan monitorear los errores de estimación y mejorar la interpretabilidad de los resultados, para responder rápidamente en caso de determinar que el modelo no logra adaptarse a los cambios que se puedan estar observando en la realidad.

3.2. Métodos de estimación de la demanda de hidrocarburos

Antes de iniciar con la descripción específica de los resultados de estudios relacionados con la estimación de la demanda de hidrocarburos y los aprendizajes alcanzados para la presente investigación, es importante realizar una introducción general sobre los diversos métodos empleados en este tipo de estimaciones.

Tal y como indican Gairifo y Dias (2009), Frey et. al. (2009), Gotham (2009), Wadud et al (2011) y Zhang y Yang (2015), los enfoques de uso final, híbridos y basados en escenarios, se emplean principalmente para pronósticos de mediano y largo plazo, razón por la cual, las tendencias de cambios tecnológicos, sustitución de fuentes de energía, hábitos de consumo,

expectativas de crecimiento demográfico y económico, así como las tendencias de configuración productiva son de gran relevancia para dichos plazos

En la presente investigación, el periodo de análisis es de corto plazo, y por consiguiente se puede asumir que las características más estructurales tienden a mantenerse relativamente estables (O`Ryan, 2008), además se incluyen covariables relevantes que pueden estar influyendo de modo importante en el consumo de hidrocarburos y que podrían captar cambios de patrones como por ejemplo el efecto ocasionado en el corto plazo por el Covid-19.

A continuación, se procede a mostrar un cuadro en el cual se detallan las principales metodologías para la estimación de la demanda de energía, así como las principales ventajas, desventajas y requerimientos de información típicos.

Cuadro 2. Ventajas, Desventajas y Requerimientos de Información de Metodologías de Proyección de Demanda de Energía.

Metodología	Ventajas	Desventajas	Requerimientos de Información Típicos
Tendenciales (Series de Tiempo)	Útil para predicciones temporales.	No considera "*driving forces*" No incluyen causalidad y no pueden identificar cuando surgen contradicciones.	Series históricas sociales, demográficas, económicas, etc. Por ejemplo: PIB, Población, consumos, etc.
Econométricas	Especialmente útiles en el corto y mediano plazo.	No captura cambios estructurales. Según expertos este método no necesariamente resulta en mejores predicciones que las tendenciales (Huss, 1985)	Series históricas sociales, demográficas, económicas, etc. Por ejemplo: PIB, Población, consumo, etc.
Análisis de Uso Final	Fácil de incorporar cambios tecnológicos anticipados. Permite capturar efectos de saturación. Permite distintos niveles de agregación.	Puede llevar a pronósticos de demanda mecánicos sin referencia alguna al comportamiento óptimo de los agentes ni variaciones en patrones de consumo debido a cambios demográficos, económicos o culturales.	Intensivo en datos. Requiere consumos energéticos sectoriales, desagregados tanto como sea posible, en general, los sectores desagregados en subsectores representativos con datos de diferentes tipos de consumos.
Enfoques Combinados/ Híbridos	Permite incluir en las estimaciones las inquietudes de ingenieros y economistas		Intensivo en datos. Consumos sectoriales desagregados y series de datos que sustenten el análisis econométrico.
Análisis de Escenarios	Los supuestos quedan explícitos (transparencia)	Escenarios son débiles cuando se asume que los "drivers" claves del análisis permanecen inalterados en forma indefinida	Intensivo en datos, pues requiere consumos energéticos sectoriales, desagregados tanto como sea posible, en general, los sectores desagregados en subsectores representativos con datos de diferentes tipos de consumos.

Fuente: extraído de (O`Ryan, 2008, págs. 41-42)

Como se observa en el cuadro anterior, existen diversas alternativas, muchas de las cuales requieren información específica del sector y la participación de profesionales de diversas áreas, sin embargo, en esta investigación se emplearán únicamente los enfoques estadísticos, específicamente los basados en series de tiempo frecuentistas y bayesianos, y se incluye una metodología adicional que son los modelos estadísticos basados en minería de datos.

En función de lo anterior, se procede a detallar la formulación teórica de los diversos modelos, y los resultados observados en diversas investigaciones en las que se han empelado estos métodos.

3.3.Modelo Holt Winters y ARIMA

Tal y como indica (Hernández, 2011, pág. 44), el modelo de suavizamiento exponencial Holt Winters multiplicativo se basa en las siguientes ecuaciones:

$$a_t = \propto \frac{Z_t}{S_{t-s}} + (1-\propto)(a_{t-1} + b_{t-1})$$

$$b_t = \beta\,(a_t - a_{t-1}) + (1-\beta)b_{t-1}$$

$$S_t = \gamma \frac{Z_t}{a_t} + (1-\gamma)S_{t-s}$$

$$P_{t+m} = (a_t - b_t m)S_{t-s+m}$$

Donde P_{t+m}, representa la estimación desarrollada para el periodo $t+m$ luego a_t y b_t son constantes para cada periodo, $\propto$, β y γ son los parámetros por estimar, Z_t representa la variable de serie de tiempo que se está analizando y S_t es un índice de estacionalidad.

Tal y como se logra apreciar, esta técnica parte de un conjunto de valores iniciales, y por medio de un proceso iterativo, logra realizar una proyección basada en una función lineal, en la cual el valor de a_t y b_t se obtienen de promedios ponderados, basados en los valores de periodos precedentes a_{t-1} y b_{t-1}, es decir, de los valores históricos de la serie de tiempo y

de los índices de estacionalidad que se obtengan de la información, al tiempo que el sistema de ecuaciones permite estimar con base en la información histórica los parámetros $\propto$, β y γ.

Otro de los modelos ampliamente difundidos para el desarrollo de estimaciones de series de tiempo, son los modelos ARIMA (Autorregresivo Integrado de Medias Móviles), los cuales de conformidad con Pankratz, (1983, pág. 281) se pueden representar por medio de la siguiente ecuación:

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s{}^D\widetilde{Z}_t = \Theta_Q(B^s)\theta_q(B)a_t$$

Esta ecuación presenta una notación muy comprimida de un conjunto de operaciones que normalmente se representan por medio de la notación ARIMA(p,d,q)(P,DQ)[s], de este modo $\boldsymbol{\phi_p(B)}$ representa un polinomio de rezagos de grado p para la parte regular o tendencial, es lo que normalmente se asocia al componente conocido como Autorregresivo de la serie $\boldsymbol{\widetilde{Z_t}}$, de ese modo si el modelo es de tipo AR1, en esta notación se representaría por ARIMA(1,0,0)(0,0,0)[12] dado que, en este caso p =1, y nos indica que el valor observado en t tiene un nivel de asociación con el dato observado en el mes previo $t-1$, también se presenta el operador de rezagos $\boldsymbol{\nabla^d = (1-B)^d}$ el cual indica cuantas veces se debe diferenciar la serie $\boldsymbol{\widetilde{Z_t}}$ para convertirla en una serie estacionaria en su parte regular, de ese modo si $d = 1$, implica que se debe hacer una diferencia.

También se tiene el componente $\boldsymbol{\Phi_P(B^s)}$ que representa un polinomio de rezagos de grado P para la parte estacional, es lo que normalmente se asocia al componente conocido como Autorregresivo estacional de la serie $\boldsymbol{\widetilde{Z_t}}$, es decir con respecto al mismo mes pero de los años anteriores, de ese modo si el modelo es de tipo ARIMA(0,0,0)(1,0,0)[12] es porque en este caso $P = 1$, e indica que el valor observado en t tiene un nivel de asociación con el dato observado en el mismo mes pero del año previo, es decir en $t-12$, también tenemos el operador de rezagos $\boldsymbol{\nabla_s{}^D = (1-B^S)^D}$ el cual indica cuantas veces se debe diferenciar estacionalmente la serie $\boldsymbol{\widetilde{Z_t}}$, para hacerla estacionaria, recordando que la diferencia estacional estaría dada por $\boldsymbol{Z_t - Z_{t-12}}$, de ese modo si $D = 1$, es porque se debe hacer una diferencia estacional para buscar la estacionariedad.

Además, se tiene del otro lado de la ecuación las medias móviles de los errores, en este caso se modela la variable $\widetilde{\boldsymbol{Z}_{\boldsymbol{t}}}$, haciendo uso también de rezagos observados de los errores en la estimación de meses previos de un modo similar a como se realiza con la parte autorregresiva, pero en función de los errores previos, estos componentes ayudan a corregir la serie tomando en consideración los desvíos que está tendiendo el modelo en el valor esperado con respecto al valor observado.

El componente $\boldsymbol{\theta_q(B)}$ que representa un polinomio de rezagos de grado q para la parte regular, es lo que normalmente se asocia al componente conocido como media móvil de los errores $\boldsymbol{a_t}$ de ese modo si el modelo es de tipo MA1 o lo que en esta notación sería ARIMA(0,0,1)(0,0,0)[12] es porque en este caso $q = 1$, y nos indica que el valor observado en t tiene un nivel de asociación con el error observado en el mes previo $t - 1$. También se tiene el componente $\boldsymbol{\Theta_P}(B^s)$ que representa un polinomio de rezagos de grado Q para la parte estacional, es lo que normalmente se asocia al componente conocido como media móvil de los errores serie $\boldsymbol{a_t}$ con respecto al mismo mes pero de los años anteriores, de ese modo si el modelo es de tipo ARIMA(0,0,0)(0,0,1)[12] es porque en este caso $Q = 1$, y nos indica que el valor observado en t tiene un nivel de asociación con el dato de error observado en el mismo mes pero del año previo es decir en $t - 12$.

En resumen, el modelo ARIMA, es otro modelo, que logra explotar al máximo los datos observados, para lo cual trata de modelar la tendencia y la estacionariedad a partir de componentes autorregresivos de la misma serie en estudio ya sea con respecto a meses inmediatos previos o mismos meses de años previos, además busca modelar adecuadamente los ciclos por medio de un componente de media móvil, que estará en función de las desviaciones del valor observado con respecto al esperado, ya sea de periodos inmediatos o periodos estacionales, con lo cual busca de modo implícito captar el proceso generador de datos y por consiguiente, reflejar las condiciones que se están presentando en la serie para desarrollar los pronósticos.

Sobre este modelo se debe recalcar que se deben desarrollar las pruebas respectivas para asegurar el cumplimiento de los supuestos de este tipo de modelación, elementos que se probarán en el desarrollo del trabajo.

En relación con estos modelos de series de tiempo, tal y como muestra (Lee & Huh, 2017, pág. 2) constituyen el principal método de estimación para la demanda de combustible. Estos modelos tienen una amplia difusión, desde la década de 1960, dado que, a partir de esta fecha los modelos estadísticos y econométricos se han utilizado comúnmente para pronosticar el consumo de hidrocarburos, por ejemplo, del gas natural (Zhang & Yang, 2015, pág. 216).

Lo anterior, se explica en parte por su simplicidad y bondad de ajuste, lo que los ha convertido en un punto de referencia prioritaria a nivel de pronóstico. Se han desarrollado una amplia gama de modelos, los cuales se pueden dividir en tres grupos, dependiendo de sus supuestos sobre el proceso de generación de datos: secuencias de martingala, modelos autorregresivos y especificaciones de reversión a la media (Frey et al, 2009, pág. 31).

Un ejemplo de esta variedad de modelos se puede notar en el trabajo desarrollado por Khedmatia & Ghalebsaz-Jeddib (2018), en el cual se propusieron cinco modelos de series de tiempos de la demanda de petróleo para países miembros de la OCDE, en este trabajo se incluyeron modelos ARIMA estacionales, modelo de pronóstico Holt-Winters y de modelos de función de transferencia.

Según Consultores en Economía Dinámica SPA (2011, pág. 3) los modelos ARIMA “son construcciones estadísticas los cuales han resultado ser muy precisos en hacer proyecciones de corto plazo. Pero por su naturaleza 100% estadísticos estos modelos son incapaces de generar explicaciones o historias económicas”, de este modo, bajo condiciones regulares permiten hacer proyecciones con bajos niveles de error, sin embargo, no permiten contemplar cambios en el contexto que puedan modificar sustancialmente el comportamiento esperado.

Por esta razón, se han desarrollado un conjunto de modelos de más largo plazo, que buscan generar una relación entre variables, de modo que, si alguna variable se aleja de una senda esperada, se manifieste una especie de corrección, que busque en próximos periodos, volver al valor propuesto en la senda esperada.

De manera más general, los modelos de corrección de errores (ECM) son diseñados para capturar movimientos hacia un nivel de equilibrio, de modo que la variable en estudio tiende a ajustarse a las desviaciones de este equilibrio (Frey et al, 2009, pág. 32).

Por lo anterior, una de las prácticas comunes para el desarrollo de pronósticos de series de tiempo es la siguiente:

> La estrategia general es proyectar los primeros seis meses con modelos ARIMA y luego el resto del horizonte de proyección se completa con modelos de cointegración (o modelos auto regresivos). La estrategia fue priorizar en el corto plazo la precisión de las proyecciones, en cambio en el mediano y largo plazo se le dio importancia a las "historias" de las proyecciones en función de variables exógenas (Consultores en Economía Dinámica SPA, 2011, pág. 4)

En relación con estos proyecciones de largo plazo, se han desarrollado modelos de Vectores Autorregresivos Estructurales, los cuales, por medio de un modelo de ecuaciones dinámicas simultáneas con un conjunto de restricciones de signos en la identificación, logran modelar diversas relaciones con base en la teoría económica, el conocimiento institucional y otra información ajena, a fin de proveer de una mayor interpretación económica y el significado causal de las estimaciones del modelo, este tipo de modelaciones han sido empleadas en el sector de hidrocarburos a través modelo de mercado de petróleo estructural introducido por Kilian y Murphy en el año 2014 (Economou et al, 2017, págs. 14-15).

Por todo lo anterior, en la presente investigación se utilizarán modelos Holt Winters y ARIMA para el pronóstico de corto plazo de las ventas de estos principales combustibles derivados de hidrocarburos en Costa Rica.

3.4.Modelo de Series de Tiempo Bayesianas Estructurales

Una de las principales críticas a los modelos de series de tiempo convencionales, es que estos suponen implícitamente que los factores que influyen en la variable objetivo tendrán los mismos efectos en el futuro, sin embargo se pueden presentar cambios en el entorno de modo

intempestivo como por ejemplo la reducción abrupta en la movilidad ocasionada por las restricciones del Covid-19, y los modelos de series de tiempo, pese a contar con mecanismos de análisis de intervención, no logran contar con una incorporación oportuna de estos efectos en las estimaciones, pues requieren al menos de un tiempo prudencial para estimar los efectos e incorporarlos en el modelo.

Tradicionalmente para lidiar contra esta incertidumbre, se procedía a realizar encuestas a expertos en la materia, propiciando un enfoque subjetivo que busca orientar los pronósticos, sin embargo, en ocasiones la opinión de los expertos resulta insuficiente ya que puede provocar sesgos influenciados por la falta de información completa, es por ello que se ha optado por complementar la confiabilidad de los pronósticos mediante la utilización de modelos de pronóstico bayesianos. (Lee & Huh, 2017, pág. 3)

Una evaluación de la incertidumbre debe proporcionar una comprensión de la confiabilidad del resultado de un modelo. El criterio experto es importante y debe ser incluido, pero debe ser relativizado por hechos estadísticos y un reproducible método de cálculo (Culka, 2016, pág. 3).

Ante esta situación, lo modelos bayesianos han logrado aumentar su popularidad, y son empleados cada vez más en la estimación de la demanda de hidrocarburos, estos modelos cobran relevancia pues logran incluir dentro de los diversos mecanismo de muestreo y simulación, un conjunto de escenarios respaldados por distribuciones de probabilidad esperados, que permiten incorporar elementos que mejoran los resultados ante escenarios de alta incertidumbre (Lee & Huh, 2017, pág. 3).

El modelo de Series de Tiempo Bayesianas Estructurales (BSTS), de conformidad con lo expuesto por Scott y Varian (2014, pág. 8) se puede resumir de la siguiente forma:

$$y_t = \mu_t + \tau_t + \beta^T x_t + \varepsilon_t$$

$$\mu_t = \mu_{t-1} + \delta_{t-1} + u_t$$

$$\delta_t = \delta_{t-1} + v_t$$

$$\tau_t = -\sum_{s=1}^{S-1} \tau_{t-s} + w_t$$

Como se logra apreciar en este sistema de ecuaciones, la estimación de la serie de interés y_t, posee un conjunto de componentes, en primera instancia un valor tendencial μ_t que logra modelar el componente de trayectoria en el tiempo, el cual será igual al valor previo observado, más una "pendiente" que estará dada por δ, la cual por temas de exogeneidad en el cálculo será la del periodo previo más una perturbación v_t, de modo que la pendiente está determinada estocásticamente por un proceso autorregresivo.

Además, se adiciona un componente estacional, determinado por τ_t, el cual estará en función de la suma de los efectos estacionales de los $S-1$ periodos previos, es decir del efecto que las variables estacionales de los meses previos dentro del ciclo generan sobre la variable dependiente, y que como indican los autores, a modo de comparación pueden ser entendidas como variables dicotómicas por mes con coeficientes dinámicos restringidos.

Otro componente que se agrega al modelo es el relacionado con las covariables $\beta^T x_t$, estas variables exógenas buscan agregar información relevante, que permita incluir información útil al modelo sobre fenómenos relacionados, que no logran ser captados de modo directo en el componente estacional o tendencial, y que por tanto podrían contribuir en la explicación de la variabilidad de la variable dependiente, en nuestro caso por ejemplo las variaciones de movilidad ocasionadas por el Covid-19. Además, se incluyen las variables ε_t, u_t y w_t representan los respectivos errores.

Los modelos bayesianos hacen Cadenas de Markov Monte Carlo (MCMC), las cuales facilitan la inclusión de muchas influencias potenciales por medio de las diversas iteraciones, logrando aprovechar los medios computacionales (Culka, 2016, pág. 8).

Este tipo de modelos, han permitido por medio del uso de la probabilidad posterior, contribuir a abordar el problema de incertidumbre de los modelos, elemento que no se aborde de modo directo en otros métodos existentes. En términos de precisión, incluso se han observado

mejores resultados de predicción que otros métodos, como el modelo estructural de Gray y redes neuronales artificiales. (Zhang & Yang, 2015, pág. 219).

Un modelo de pronóstico inmediato efectivo debe considerar tanto el comportamiento pasado de la serie que se modela, así como los valores de las señales contemporáneas más fácilmente observables. Choi y Varian en 2009 y 2012 demostraron que los datos de búsqueda de Google podrían usarse como un método efectivo para incorporar señales externas para un modelo de predicción inmediata, pero sus métodos requerían seleccionar cuidadosamente el conjunto de predictores que se utilizarán (Scott & Varian, 2014, pág. 4).

Este resultado es de gran relevancia en esta investigación, pues la reducción de la movilidad constituye el principal factor explicativo de las modificaciones observadas en la demanda de hidrocarburos y por consiguiente el uso de información en tiempo real, sobre la movilidad aproximada por medio de dispositivos GPS, puede constituir una variable de gran relevancia para las estimaciones que se deban desarrollar.

Las covariables explicativas pueden ser incorporadas en los modelos de series de tiempo convencionales, sin embargo el modelo BSTS propuesto por Scott & Varian (2014) permite la inclusión de un gran número de covariables explicativas y utiliza un mecanismo de ponderación para que se mantengan con coeficientes diferentes a cero, aquellas variables más relevantes. Además, usa un algoritmo de muestreo MCMC para la distribución posterior, lo que permite la utilización de un gran número de modelos potenciales y por medio de promedios de los resultados determina los pronósticos finales (Scott & Varian, 2014, pág. 5).

Como subproducto del análisis bayesiano, el modelo BSTS también proporciona informes convincentes que indican la probabilidad marginal de inclusión posterior para cada predictor, y un desglose gráfico de cómo el modelo ha distribuido la variabilidad de las series de tiempo entre diferentes componentes del estado (Scott & Varian, 2014, pág. 5).

Todo lo anterior ha hecho que se utilicen de modo complementario modelos ARIMA con modelos bayesianos debido a la capacidad de capturar patrones en los datos y rendimiento de pronóstico superiores (Kitamura, 2018, pág. 72), de modo similar a como se realiza en el presente trabajo.

3.5.Modelo XGBoost

Tal y como se mencionó anteriormente, uno de los principales supuestos de los métodos de regresión es que los patrones en los datos históricos se mantendrán en el futuro (Pavlyshenko, 2019, pág. 10), sin embargo se ha observado que el uso de modelos de redes neuronales, logran manejar eficazmente incertidumbre, y se ha convertido en una herramienta común para pronosticar el consumo de hidrocarburos (Zhang & Yang, 2015, pág. 216).

El modelo XGBoost, es una técnica de aprendizaje automático recientemente dominante para la predicción de series de tiempo y para la selección de características relevantes a partir de patrones observados (Abbasi et al., 2019)

El modelo XGBoost, es un método de conjunto en el cual se combinan Kestimaciones de varios modelos, cada uno de los cuales utiliza como insumos covariables explicativas, y por medio de algún mecanismo de consenso determinan el valor final estimado para $\hat{y}_i$, lo anterior, se expresa matemáticamente de la siguiente manera (Zhou, Li, Shi, & Qian, 2019, pág. 4):

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i)$$

Sobre este método, se debe indicar que se utiliza tanto para problemas de regresión como para problemas de clasificación, en este trabajo se empleará para estimar la regresión de la demanda de hidrocarburos, y cada uno de los k-ésimos modelos estimados por este método serán árboles de regresión. Este método se diferencia de otros métodos similares por la utilización del método de optimización de Newton en vez del método de gradiente (Zhou, Li, Shi, & Qian, 2019, pág. 4). De este modo, encuentra los parámetros óptimos minimizando la función de pérdida o de error denotada como función $l(\hat{y}_i, y_i)$ (la cual para regresión podría ser la Raíz del Error Cuadrático Medio), tal y como se muestra a continuación:

$$L(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k)$$

En este caso n representa el tamaño de la muestra y como se indicó anteriormente, la función $l(\hat{y}_i, y_i)$ es la función de pérdida o error y por consiguiente está en función de los valores observados (y_i) y estimados ($\hat{y}_i$) de la variable dependiente, mientras que la función $\Omega(f_k)$ es la función de complejidad del árbol de decisión para cada k-ésimo modelo, la cual viene definida de la siguiente forma:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\eta\|\omega\|^2$$

En esta función de complejidad, el factor T representa el número de nodos de hojas de los árboles de decisión, ω representa el peso de los nodos de hojas, γ controla el alcance de la penalización por complejidad para la estructura del árbol y η controla el grado de regularización de f_k. En resumen, este componente realiza una regularización para los modelos, a fin de que no se presente un sobreajuste por memorizar patrones de datos, y de este modo propiciar la adecuada generalización de los resultados.

Dado que es difícil para el modelo de conjunto de árboles minimizar la función de pérdida con los métodos tradicionales en el espacio euclidiano, el modelo utiliza la forma aditiva, por medio de la cual se agrega una modificación a la función de pérdida a partir de los resultados del modelo "más débil" o con peores resultados de estimación, es decir el mecanismo XGBoost, busca mejorar los resultados de estimación, para ello, trata de ponderar cada vez más los árboles con peores resultados, a fin de que el modelo logre ir mejorando, es decir, que se enfoque en las deficiencias que actualmente posee a fin de lograr solventarlas y de esta manera lograr una mayor reducción del error o pérdida global. En términos matemáticos, esto se expresa de la siguiente forma (Zhou, Li, Shi, & Qian, 2019, pág. 4):

$$L^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i{}^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

Donde $\hat{y}_i{}^{(t)}$ es la predicción del *i*-esimo dato en la *t*-ésima iteración y f_t es el modelo de árbol de decisión "más débil" o con peores resultados en el proceso de aprendizaje en la *t*-ésima iteración.

La optimización de Newton realiza una expansión de segundo orden de Taylor en la función de pérdida indicada en la ecuación anterior, lo anterior pues la aproximación de segundo orden ayuda a minimizar la función de pérdida de manera eficiente (Zhou, Li, Shi, & Qian, 2019, págs. 4-5). A continuación, se realiza la expansión de Taylor antes mencionada:

$$L^t = \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i) \right] + \Omega(f_t)$$

Note que, el componente $\Omega(f_t)$ estará determinado por la forma funcional indicada anteriormente y el componente $l\left(y_i, \hat{y}_i^{(t-1)}\right)$ es constante, por consiguiente, podría ser suprimido para efectos de la minimización, dado que su derivada es cero, por lo anterior, la función a optimizar estaría dada por (Zhou, Li, Shi, & Qian, 2019, pág. 5):

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i) \right] + \gamma T + \frac{1}{2} \eta \sum_{j=1}^{T} {\omega_j}^2$$

Donde:

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$$

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$$

Es decir, la función g representa el gradiente o derivada de primer orden de la expansión de Taylor, y la función h representa el hessiano o derivada de segundo orden de la expresión de Taylor.

En resumen, el modelo XGBoost, permite emplear un conjunto de árboles de regresión que por consenso determinan el valor estimado de una variable, al ser un proceso iterativo, busca explotar al máximo la capacidad computacional para realizar la estimación respectiva, para lograrlo usa una minimización de la función de pérdida o error, la cual es modificada por medio de métodos eficientes de optimización, y que incorpora un refuerzo en función de los resultados más débiles, buscando priorizar esas deficiencias y tratando de corregirlas en el proceso iterativo.

Los resultados alcanzados por Barton, van Kasteren, & Birk (2015), muestran como al utilizar un conjunto de datos conocido por ser exitoso con el análisis de series de tiempo, el cual recopilaba información sobre ventas de varios productos en varias semanas, se lograron mejores resultados de estimación utilizando el modelo XGBoost, en comparación con otros modelos de series de tiempo tradicionales.

En el futuro, se espera utilizar el enfoque propuesto para pronosticar otras series de tiempo de energía, como la velocidad del viento, la carga de electricidad y los precios de las emisiones de carbono, o la demanda energética (Zhou, Li, Shi, & Qian, 2019, pág. 12).

3.6. Modelo de minería de datos que utiliza como variables explicativas las estimaciones de otros modelos.

Antes de iniciar con el desarrollo de este tipo de modelaciones es importante definir el concepto de minería de datos, el cual se detalla a continuación:

> La minería de datos puede definirse como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos.
>
> […]
>
> Las herramientas de minería de datos permiten extraer patrones, tendencias y regularidades para describir y comprender mejor los datos y para predecir comportamientos futuros. (Pérez & Santin, 2007, págs. 1-2)

De igual manera, tal y como indica Sivaramakrishnan & Suchithra (2017), la minería de datos se compone de la clasificación, regresión y agrupación de los datos, y se complementa con el aprendizaje de máquinas, el cual se enfoca en la predicción y clasificación, por medio de la apropiación de patrones aprendidos a través del uso de algoritmos entrenados. De este modo, la minería de datos se concentra en encontrar nuevo conocimiento a partir de la información, de modo que aprovecha el aprendizaje de máquinas para desarrollar modelos que permitan mejorar las estimaciones a partir del uso intensivo de los datos existentes.

Por lo anterior, para efectos de la presente investigación se entenderá que los modelos de minería de datos estarán basados en el uso intensivo de información para obtención de tendencias, patrones y comportamientos esperados, por medio de la aplicación de algoritmos de aprendizaje de máquinas.

De esta manera, para la predicción de series de tiempo que poseen patrones complejos y dinámicos con incertidumbre, se han empleado modelos híbridos o modelos compuestos, los cuales buscan beneficiarse de las ventajas de cada modelo y obtener un rendimiento óptimo de la previsión global a partir del máximo aprovechamiento de la información (Barbosa de Alencar et al., 2017, pág. 4). Otro elemento importante por recalcar es que la combinación de modelos casi siempre mejora la generalización (Brownlee, 2021).

Este enfoque de utilizar los resultados de múltiples modelos, en las predicciones del conjunto de validación como regresores de entrada para los modelos en un siguiente nivel fue denominado enfoque de apilamiento por Pavlyshenko (2019).

De modo general y utilizando como base lo planteado por Pai & Lin (2005, págs. 498-499), se puede plantear la siguiente forma funcional:

$$\hat{Y}_{t+1} = f\left(\widehat{M1}_{t+1}, \widehat{M2}_{t+1}, \dots, \widehat{Mk}_{t+1}\right)$$

Donde $\hat{Y}_{t+1}$ representa el valor estimado para la variable dependiente en el periodo $t+1$, el cual será calculado a partir de una función f la cual tendrá como variables explicativas los valores estimados para la variable dependiente en el periodo $t+1$, del modelo 1 ($\widehat{M1}_{t+1}$), del modelo 2 ($\widehat{M2}_{t+1}$), y así sucesivamente hasta el k-ésimo modelo ($\widehat{Mk}_{t+1}$),

A partir del uso de modelos de Máquinas de Soporte Vectorial (SVM), se ha logrado estimar relaciones no lineales con resultados muy prometedores (Pai & Lin, 2005, pág. 497). Este modelo fue desarrollado por Vapnik en 1995 y su objetivo final es establecer un margen máximo en el proceso de clasificación a fin de tener un buen desempeño y una alta precisión predictiva (Cristianini y Shawe, 2000). Los modelos SVM se han ampliado para problemas con variables dependientes continuas por medio de Regresiones de Vectores de Soporte

(SVR) de un modo eficiente (Ogcu, Demirel, & Zaim, 2012, pág. 1579), es por ello que, se emplearán en la presente investigación para estimar la función f.

Para profundizar en el desarrollo del modelo de SVR, se procede a detallar la especificación matemática tomando como base (Pai & Lin, 2005, págs. 498-499). Se parte de la siguiente ecuación:

$$\hat{Y}_t = w * \emptyset(x) + b$$

Donde $\emptyset(x)$ es una función característica, que determina el espacio vectorial o transformación no lineal desde el espacio de entrada x (Pai & Lin, 2005, pág. 498) , donde x en este caso son las estimaciones obtenidas de los modelos previos en el periodo de entrenamiento. Los coeficientes w, indican el nivel de uniformidad de la estimación, en este caso, el inverso estaría asociado a la distancia entre categorías o niveles, de este modo, se esperan valores adecuados en estos coeficientes que permitan un buen nivel de discriminación y por consiguiente de precisión para el valor de $\hat{Y}_t$ esperado, por su parte la b es una constante del modelo. Ambos parámetros se estiman minimizando la siguiente función:

$$R = C * \frac{1}{n} * \sum_{i=1}^{n} L\left(Y_i - \hat{Y}_i\right) + \frac{1}{2}\|w\|^2$$

En este caso tendríamos que C sería un parámetro de regularización, similar a cómo se indicó para el modelo XGBoost, por su parte Y_i es el valor observado y $\hat{Y}_i$ es el valor estimado por medio de SVR, además $L\left(Y_i - \hat{Y}_i\right)$ es una función de pérdida o error y w es el parámetro de uniformidad (inverso de la distancia) antes indicado, y por tanto su minimización permite discriminar cuál es el valor de Y_i, dado que un valor muy alto de w nos daría prácticamente una estimación constante, mientras que un valor bajo, nos permitiría tener diversas estimaciones de Y_i, las cuales esperamos sean similares al valor observado.

Tal y como indica Pai & Lin (2005, pág. 499), es posible realizar una transformación de la ecuación 20, para expresarla en una optimización lagrangiana tal y como sigue:

$$R(w,\zeta,\zeta^*) = \frac{1}{2}ww^T + C^* \sum_{i=1}^{n}(\zeta_i + \zeta^*{}_i)$$

Sujeto a:

$$w * \emptyset(x) + b - Y_i \leq \varepsilon + \zeta^*{}_i$$

$$Y_i - w * \emptyset(x) + b \leq \varepsilon + \zeta_i$$

De esta manera ε representa el error de convergencia máximo permitido y las variables $\zeta_i \, y \, \zeta^*{}_i$, representan implícitamente las diferencias entre valores reales y estimados en términos positivos y por ello lo que se busca es que sean mínimas. Al resolver el problema lagrangiano se obtendría el valor de w y b que minimizan los errores y maximizan la distancia o capacidad de discriminación.

En relación con la experiencia en el uso de modelos combinados, se observó que es posible mejorar la previsión de la demanda en el comercio electrónico ya que comparten información y pronósticos de diversos modelos, permitiendo explotar características cruzadas y efectos no lineales que existen en este tipo de transacciones, al tiempo que permiten realizar predicción, con muy poca historia de los productos. Al evaluar estos modelos globales en un contexto real, con un número realista de productos, se superaron los resultados observados hasta ese momento en la previsión de la demanda de estos productos (Garnier & Belletoile, 2019, pág. 6).

De este modo, los diferentes modelos de pronóstico pueden complementarse entre sí por medio de un modelo híbrido o conjunto, por ejemplo, se han observado modelos conjuntos a partir de estimaciones de ARIMA y SVR, presentando una mejora en el rendimiento de la predicción. Teóricamente, así como empíricamente, la hibridación de dos modelos diferentes logra reducir errores de previsión. Sin embargo, la selección estructurada de parámetros óptimos del modelo híbrido o conjunto es de gran interés (Pai & Lin, 2005, págs. 503-504).

En la literatura reciente, se observan varios ejemplos de este enfoque, por ejemplo Li & Ngan (2019) desarrolló un modelo de pronóstico heterogéneo híbrido que combina el modelo ARIMA y dos clasificadores de aprendizaje automático (una SVR y una red neuronal) para

el pronóstico de series de tiempo, también para la estimación de la demanda energética se han realizado combinaciones de modelos estadísticos alternativos y de inteligencia artificial como se observa en los trabajos desarrollados por Barbosa de Alencar et al (2017), Pai & Lin (2005), Ming, Bao, Hu, & Xiong (2014) y Pavlyshenko (2019).

Es importante indicar que, existe consenso relativo respecto de las bondades de los enfoques híbridos ya que éstos permiten reducir los errores sistemáticos del uso de un enfoque único (O`Ryan, 2008, pág. 34). Lo anterior, también ha sido analizado, en el pronóstico de la demanda energética, ya que, en comparación con el método único, el método conjunto puede proporcionar resultados más precisos, permitiendo que este enfoque se vuelva cada vez más popular (Zhang & Yang, 2015, pág. 217).

A su vez se ha observado que, debido a las limitaciones inherentes de los enfoques econométricos y de uso final en la estimación de la demanda energética, ahora hay un mayor interés por el enfoque híbrido o conjunto para este tipo de estimaciones (Bhattacharyya & Timilsina, 2009, págs. 96-97).

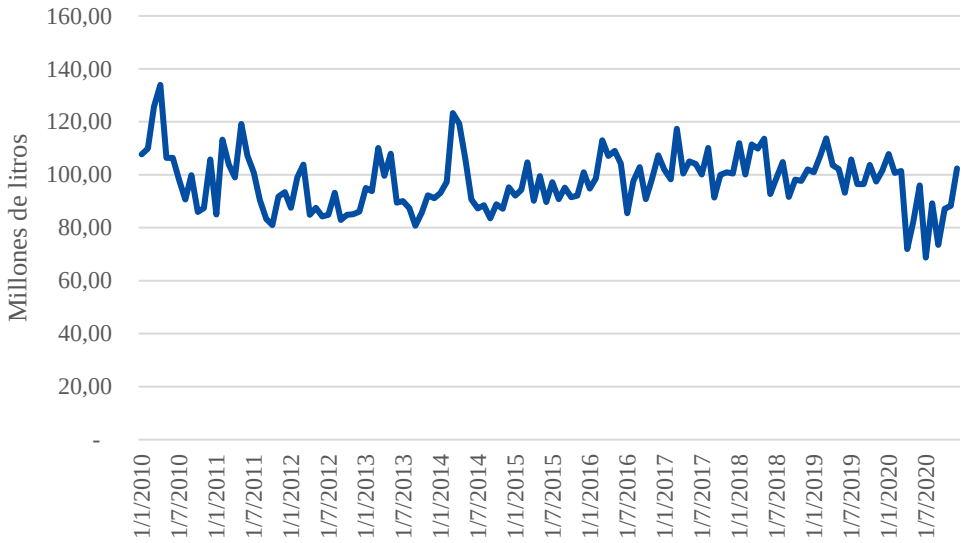
4. METODOLOGÍA

A continuación, se procede a detallar la metodología desarrollada en la presente investigación:

4.1.Datos utilizados

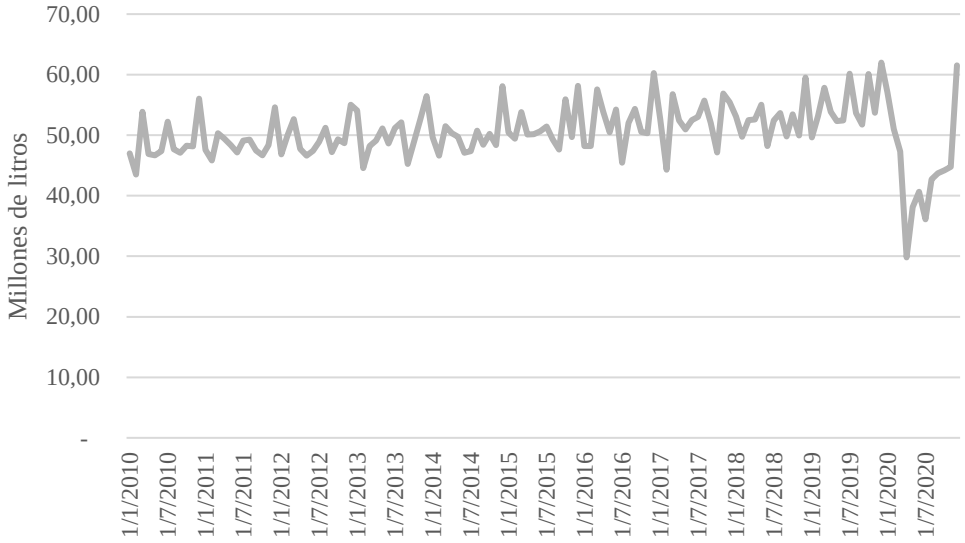
Los datos por utilizar serán las ventas totales mensuales de RECOPE para la gasolina súper, regular y diésel de Costa Rica por tipo de combustible de 2010 a 2020, esta será la variable dependiente sobre la cual se realizan las estimaciones de su comportamiento a futuro, a continuación, procedemos a mostrar el comportamiento histórico para cada combustible, en el periodo indicado.

Gráfico 6. Costa Rica, ventas de Diésel en millones de litros por mes, enero 2010 a diciembre 2020.



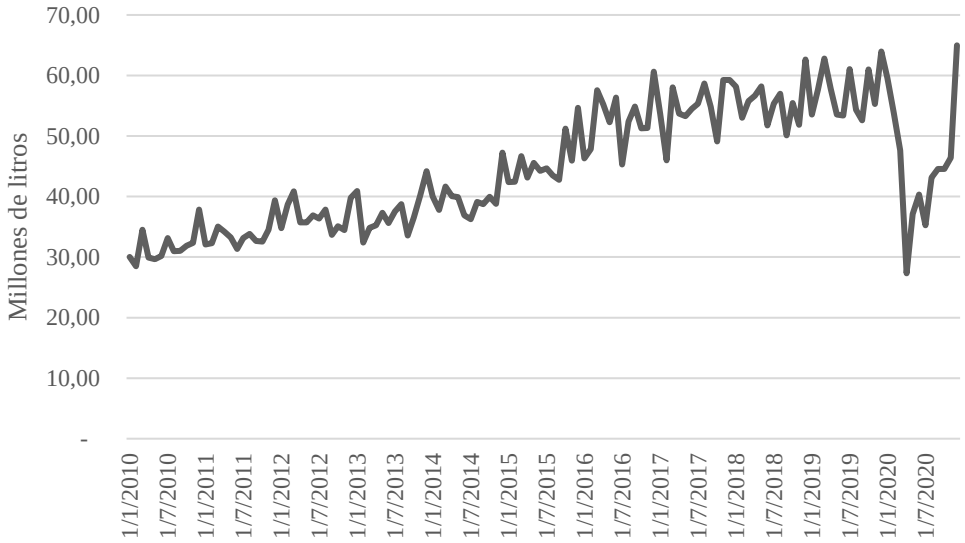
Fuente: Elaboración propia con datos de Aresep.

Gráfico 7. Costa Rica, ventas de gasolina RON 91 (Regular) en millones de litros por mes, enero 2010 a diciembre 2020.



Fuente: Elaboración propia con datos de Aresep.

Gráfico 8. Costa Rica, ventas de gasolina RON 95 (Súper) en millones de litros por mes, enero 2010 a diciembre 2020.



Fuente: Elaboración propia con datos de Aresep.

Tal y como se logra apreciar en las gráficas anteriores, existe una clara disminución en las ventas de combustibles durante el periodo inicial de la pandemia sin embargo para el mes de diciembre del 2020 se observa una rápida recuperación, que permitió situarse en valores similares al mismo mes del año pasado.

Se observa en la gráfica que el comportamiento de las series asociadas a las ventas de la gasolina regular y súper es relativamente similar durante el periodo pandémico, con efecto muy similares, aunque se logra observar a nivel histórico una mayor volatilidad de las ventas de gasolina súper, por su parte, el diésel tiende a mantenerse más estable en el tiempo, sin embargo el efecto de la pandemia, tuvo implicaciones diferentes en este producto, el cual no tuvo un golpe tan violento como los otros productos analizados, pero tuvo un comportamiento oscilante y con una recuperación menos evidente, con excepción del mes de diciembre en donde se logra alcanzar valores similares a los del mismo mes del año pasado.

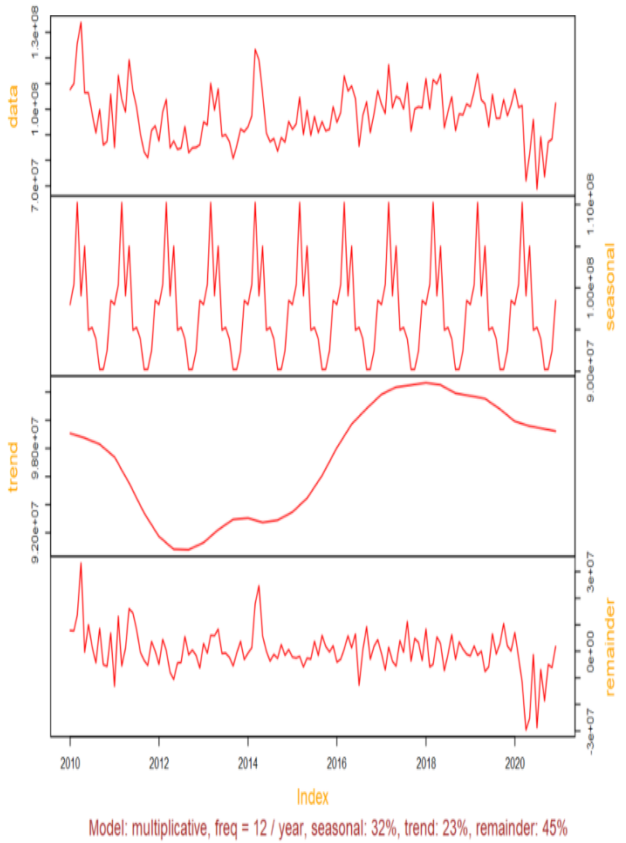
Para profundizar en los análisis del comportamiento histórico, se procedió a realizar una descomposición de estas series de tiempo, a fin de analizar con más detalle, la tendencia, la estacionalidad y el componente irregular. Como se observa en los gráficos 9, 10 y 11, es claro que cada una de las series tiene un componente estacional marcado, el cual es relativamente similar entre la gasolina Regular y Súper, pero difiere del Diésel, además de un componente tendencial relativamente creciente, aunque se observa en el Diésel, una desaceleración en los últimos años. Además, el efecto de la pandemia se evidencia con gran claridad dentro del componente irregular, lo cual demuestra que es un fenómeno de fuerte impacto, y que no puede ser explicado por los componentes regulares de la serie de tiempo, de modo que requiere un manejo especial al momento de incluirlo en los modelos.

Tal y como se observa en dichas gráficas, se evidencia que las series de ventas de Diésel, gasolina RON 91 (Regular) y RON 95 (Súper), presentan un comportamiento que permitiría la utilización de modelaciones de series de tiempo, y se esperaría que modelos de suavizamiento exponencial y modelos ARIMA logren buenos resultados en sus estimaciones, muestra de ello, es que al observar los pronósticos realizados por RECOPE y compararlos

con los datos reales observados para el periodo 2016-2019, el Error Porcentual Absoluto Medio (MAPE) era de 4,76% en el Diésel, 4,51% en la gasolina regular y de 5,05% en la súper.

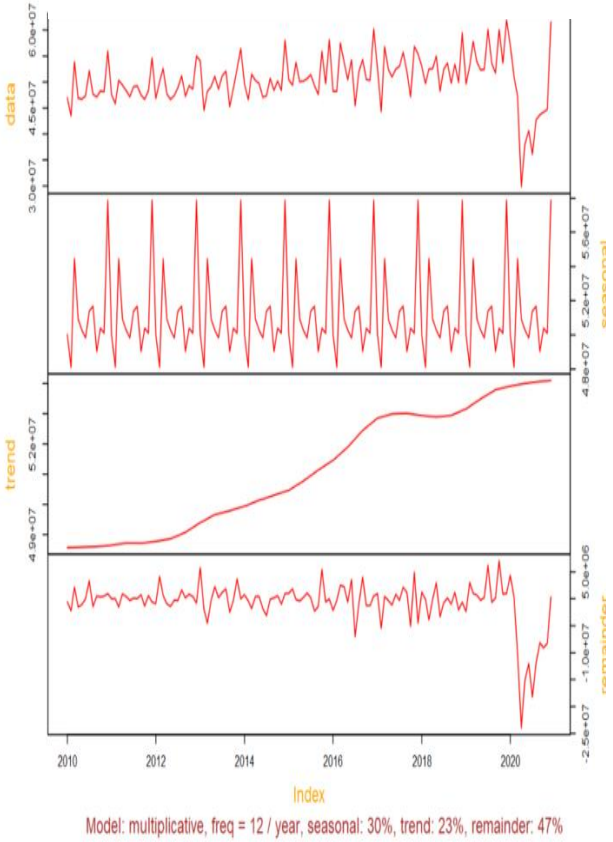
Pese a lo anterior, al analizar el año 2020, el choque ocasionado por el Covid-19 sobre el consumo de combustible no logró ser explicado por este tipo de modelos, ya que como se indicó anteriormente, los efectos de la pandemia no logran ser recogidos en los componentes regulares de una serie de tiempo, y por tanto no estaban incorporados en el proceso generador de datos, lo que en consecuencia, ocasionó que el modelo no tuviera conocimiento, sobre el posible efecto que podría ocasionar en la demanda de combustible, esto ocasionó que el MAPE de los pronósticos desarrollados por RECOPE para 2020 fueran de 14,94% para el Diésel, 18,74% para la gasolina regular y 22,39% para la gasolina súper.

Gráfico 9. Costa Rica, componentes de la serie de tiempo del consumo Diésel en litros por mes, enero 2010 a diciembre 2020.



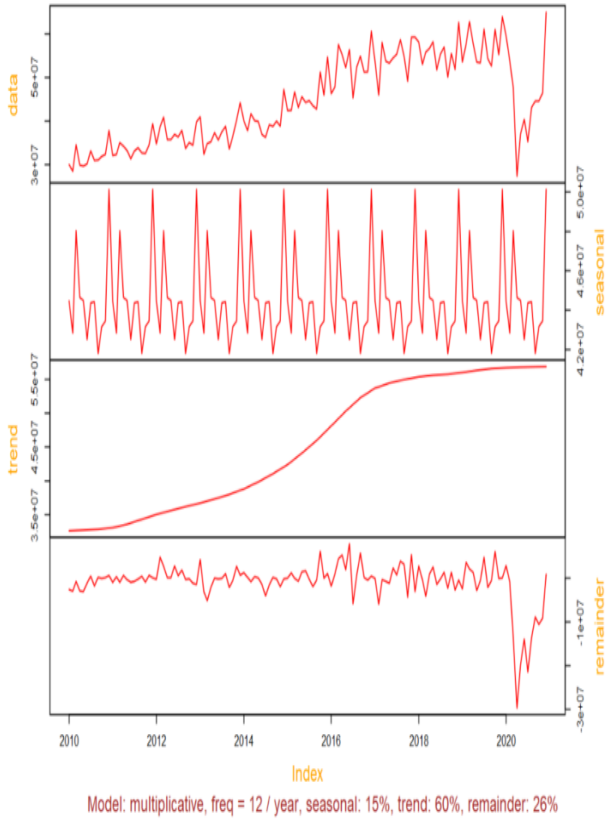
Fuente: Elaboración propia con datos de Aresep, empleando el objeto visual time series descomposition de Power BI.

Gráfico 10. Costa Rica, componentes de la serie de tiempo del consumo de gasolina RON 91 (Regular) en litros por mes, enero 2010 a diciembre 2020.



Fuente: Elaboración propia con datos de Aresep, empleando el objeto visual time series descomposition de Power BI.

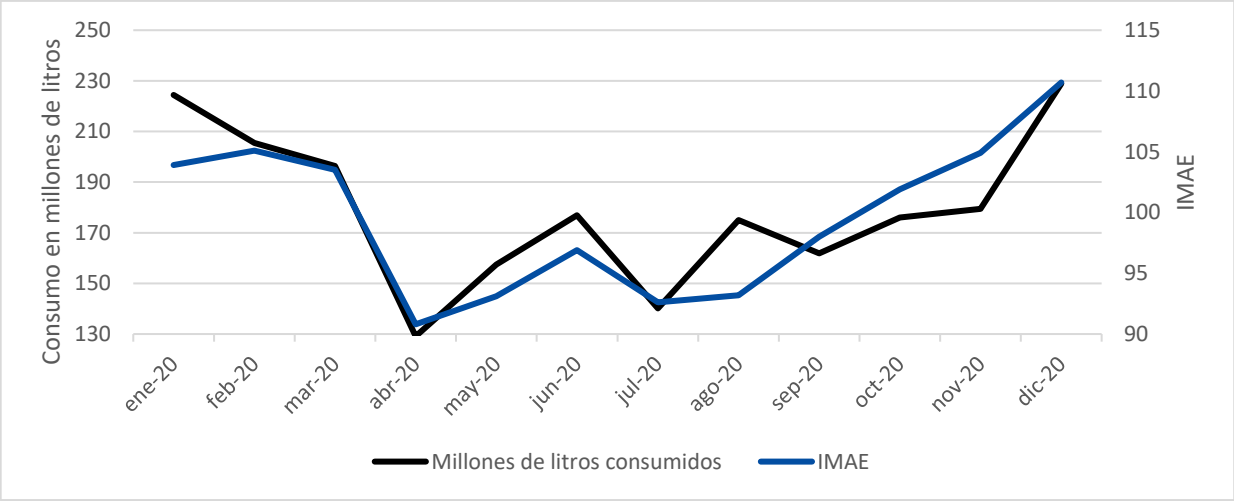
Gráfico 11. Costa Rica, componentes de la serie de tiempo del consumo de gasolina RON 95 (Súper) en litros por mes, enero 2010 a diciembre 2020.



Fuente: Elaboración propia con datos de Aresep, empleando el objeto visual time series descomposition de Power BI.

Otra variable relevante en el análisis es el Índice Mensual de Actividad Económica, el cual como se observa tienen una relación importante con el consumo de combustibles, e incluso se observa cierto comportamiento adelantado del consumo de combustibles, de modo que se podría pensar que dicho consumo, al ser un insumo importante en la producción, podría ayudar a predecir el comportamiento de la actividad económica, por lo anterior, la adecuada estimación de esta variable, ayudará a predecir hasta cierto punto la evolución del IMAE, en especial al contemplar que este indicador se publica con rezago temporal de cerca de mes y medio.

Gráfico 12. Costa Rica, Índice Mensual de Actividad Económica (IMAE) y consumo mensual de combustibles derivados de hidrocarburos seleccionados en millones de litros, enero 2020 a diciembre 2020.



Fuente: Elaboración propia con datos del Consejo Monetario Centroamericano y de Aresep.

En resumen, estas serían las variables obtenidas de información pública para el desarrollo de los modelos:

Cuadro 3. Variables obtenidas de información pública para el desarrollo de los modelos, 2021.

Variable	Descripción	Periodicidad	Unidad
FECHA	Mes al cual hace referencia la información.	Mensual	dd/mm/aa
PRODUCTO	Nombre del producto o combustible que se está analizando en este caso sería Súper, Regular y Diésel.		
LITROS	Cantidad de litros consumidos en el mes de referencia.	Mensual	Litros
MES T-2	Cantidad de litros consumidos en el periodo t-2, es decir dos meses antes, al mes de referencia.	Mensual	Litros
MES A-1	Cantidad de litros consumidos doce meses antes, al mes de referencia. Es decir, el mismo mes del año anterior.	Mensual	Litros

Fuente: Elaboración propia.

4.2.Creación de variables explicativas adicionales

Debido a las condiciones de alta incertidumbre, se requiere realizar una modelación especial para 2020 que permita realizar las estimaciones con un nivel de error menor, y que pueda ser empleada para contemplar efectos importantes, por medio de variables explicativas que

contribuyan en el poder explicativo y que se puedan estimar con base en información real existente al momento del pronóstico.

Estas estimaciones son complejas (en especial en 2020) pues se enfrentan a escenarios cambiantes en función de ciertas condiciones existentes, al tiempo que son modelaciones que se desarrollan para los estudios tarifarios que realiza Aresep, de modo que se desarrollan los segundos viernes de cada mes, y aplicarán el mes siguiente, razón por la cual tendrán un rezago temporal de 2 meses, lo que complica aún más la estimación. Para comprender de mejor manera lo anterior, tomemos por ejemplo el mes de junio de 2020, la fecha de inicio del estudio fue el 12 de junio de 2020, y se contó con información real de consumo, hasta mayo, y como el estudio entraría a regir en julio, se debió pronosticar el consumo para ese mes, de ese modo se debía pronosticar julio, con información histórica hasta mayo.

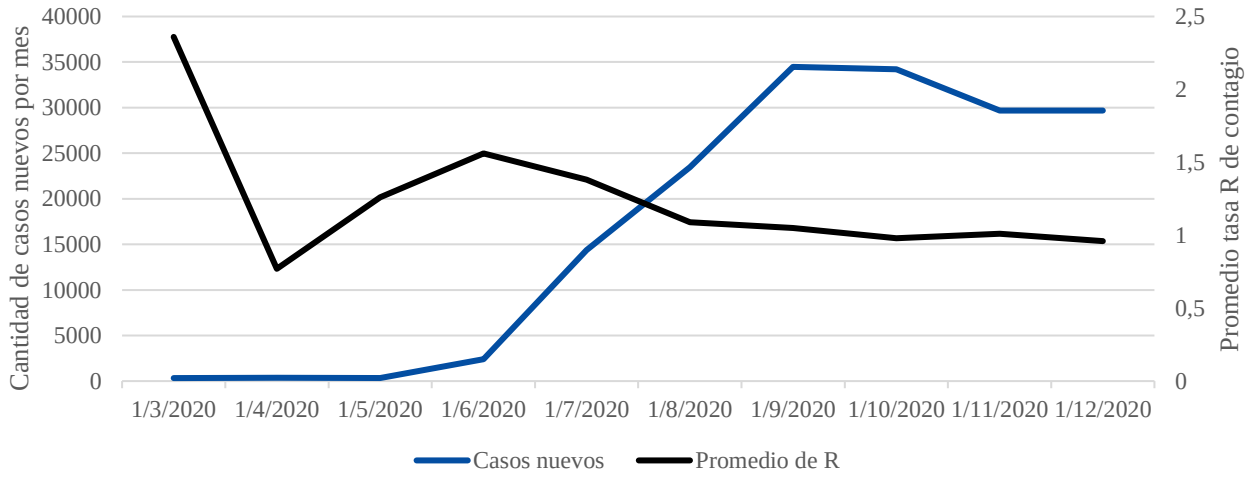
Por lo anterior, se construyeron un conjunto de variables explicativas adicionales por medio de las fórmulas y criterios, a partir de información real existente, tal y como se muestra a continuación:

En primera instancia, se trató de aproximar una variable que permitiera incluir el efecto de la pandemia de modo directo, es por esta razón que se utilizaron datos relacionado con la evolución de la enfermedad, de modo específico se seleccionaron dos variables relevantes: 1) los datos de la cantidad de casos nuevos por Covid-19 y 2) la tasa R de reproducción del virus, esta información histórica, se descargó de la plataforma suministrada por el Observatorio del Desarrollo de la Universidad de Costa Rica, específicamente de la dirección: https://app.powerbi.com/view?r=eyJrIjoiMjU3M2NkNjQtMGIyOS00ZjRmLWE3NjYtNDE2OWNkZjIxZTdjIiwidCI6ImFkNjNmZDZmLWE4OTctNDljZS1hZWU5LTRmYzYxNzY1NjY4YSJ9&pageName=ReportSection.

Esta información, se maneja con un detalle diario, sin embargo para efectos del presente trabajo se requería de modo mensual, por esta razón se obtuvo un promedio diario de casos para cada mes, mientras que para la tasa R, se tomó un valor promedio de los datos diarios,

en relación con la tasa de contagio R, esta es estimada por la Universidad de Costa Rica, su mecanismo de estimación está explicado en Rosero (2021) y básicamente se estima por medio de "un cociente entre el número de casos el día t y el promedio ponderado de los casos ocurridos en días anteriores" (Rosero, 2021, pág. 1), esta variable indica el ritmo de variación de los contagios, y por consiguiente una tasa mayor a 1, indica que hay proliferación y por consiguiente cada vez hay una mayor cantidad de población infectada. A continuación, se muestran las gráficas que detallan el comportamiento de la tasa R y la cantidad de contagios por mes para el año 2020:

Gráfico 13. Costa Rica, cantidad de casos nuevos de Covid-19 por mes y tasa R promedio por mes, enero 2020 a diciembre 2020.



Fuente: Elaboración propia con datos del Observatorio del Desarrollo UCR

Tal y como se muestra en la gráfica, la tasa promedio R de contagios, tiende a ser un indicador relativamente adelantado, del comportamiento esperado de la pandemia, razón por la cual, entre más alto sea su valor, se espera una mayor cantidad de casos a futuro, y constituye una variable importante, pues es utilizada por las autoridades para tomar decisiones sobre las restricciones sanitarias a emplear.

Como se mencionó en la sección relacionada sobre el efecto que el Covid-19 tiene sobre la demanda de hidrocarburos, es posible afirmar que la pandemia, genere un efecto indirecto, el cual dependerá en gran medida de la movilidad de los individuos, de este modo, se espera que si aumentan las restricciones a la movilidad por parte del gobierno, o se desarrollan medidas para evitar el desplazamiento de las personas, y al ser el transporte el principal uso de los combustibles, es de esperar que haya una reducción en el consumo de las gasolinas y el diésel.

Por lo anterior, se considera de gran importancia contar con una variable que permita aproximar la movilidad de los individuos, para ello se aprovecharon los datos de Google del 15 de febrero hasta la actualidad, específicamente de la información de movilidad de las comunidades ante el Covid-19 para Costa Rica, extraídos de la página https://www.google.com/covid19/mobility/, en dónde se calcula la variación porcentual con respecto a un valor de referencia[1] de la cantidad de dispositivos electrónicos con conectividad y que cumplen con el umbral de privacidad y calidad de Google, estos datos son suministrados de modo diario y se agregarán de modo mensual, a través de un promedio, para su utilización en los análisis respectivos.

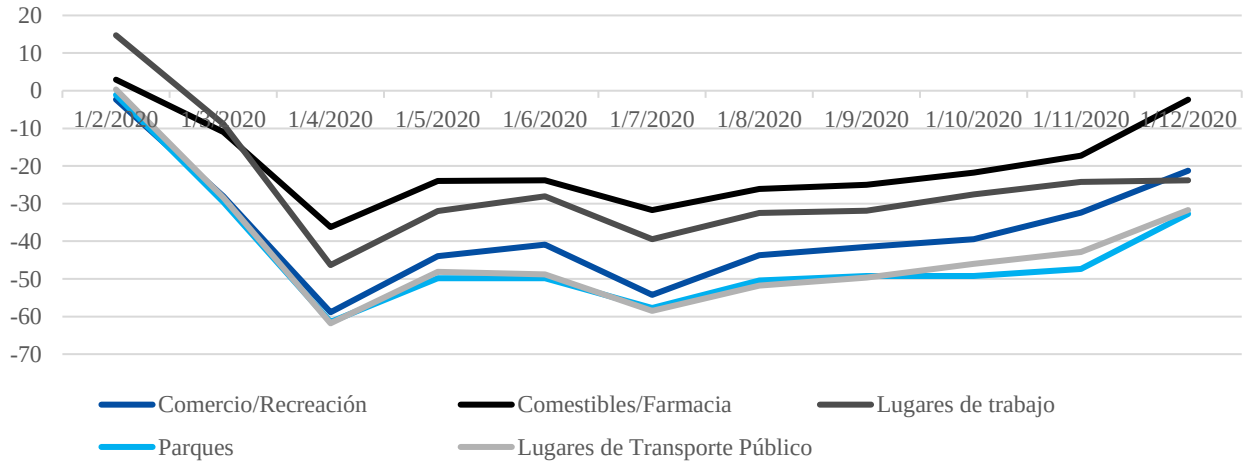
Tal y como se indicó anteriormente, el Programa Estado de la Nación (2020), desarrolló una aproximación, empleando datos de Waze, sin embargo, tal y como indica dicho ente, el nivel de correlación entre los datos de Waze y Google es muy elevada, lo cual permite pensar que logran aproximar de un modo similar el nivel de movilidad de los vehículos (Programa Estado de la Nación, 2020, pág. 237).

Es importante indicar que “los registros de Google han servido de insumo en la literatura comparada para evaluar la movilidad de personas durante la pandemia” (Programa Estado de la Nación, 2020, pág. 231), razón por la cual resultan de gran relevancia en el análisis que se va a desarrollar. Además, Google suministra información para diferentes tipos de lugares: Comercio/ Recreación, Comestibles/Farmacia, Lugar de Trabajo, Parques y lugares de

[1] El valor de referencia, que es el valor medio de cada día de la semana, se calcula durante un periodo de 5 semanas, desde el 3 de enero al 6 de febrero del 2020.

Transporte Público, y se logra observar una muy alta correlación entre todos los tipos de lugares lo cual se puede evidenciar en el siguiente gráfico que muestra los datos promedio para cada mes:

Gráfico 14. Promedio mensual de variación diaria de movilidad con respecto a un valor de referencia prepandemia, suministrado por Google para Costa Rica, febrero 2020 a diciembre 2020.



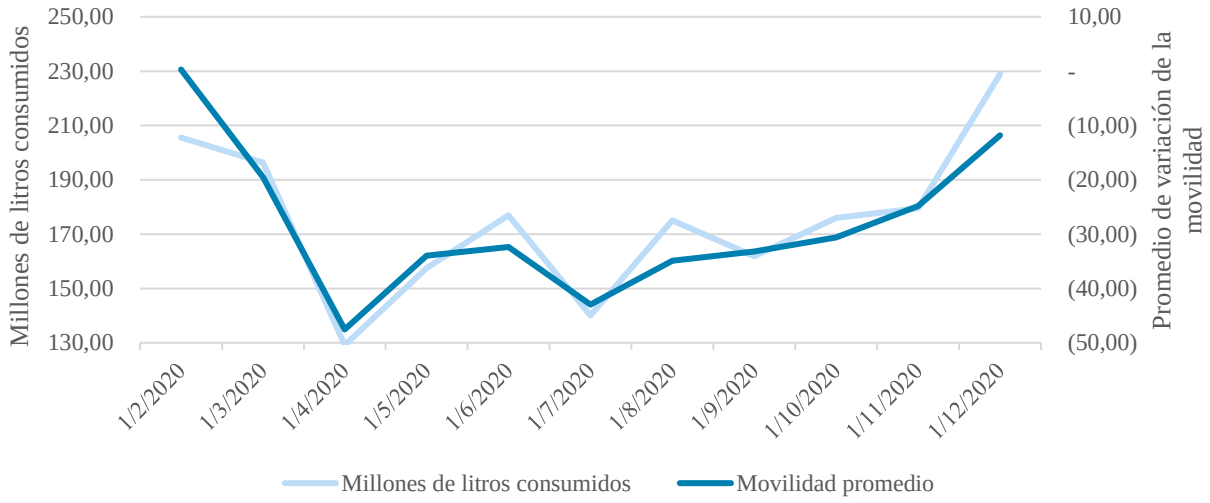
Fuente: Elaboración propia con datos del reporte de movilidad de Google.

Debido a la alta correlación de los datos, se seleccionaron para el análisis de los modelos, dos series de movilidad que presentaron la mayor relación con el consumo de combustible, para realizar los anterior, se determinó cuáles tenían una mayor correlación con el consumo de hidrocarburos, obteniendo como resultado que las series mensuales de Comestible/Farmacia y Comercio/Recreación, lograron correlaciones de Pearson de 92% y 87%, y superiores a las restantes series que presentaron valores inferiores al 80%.

Durante el desarrollo de diversas modelaciones se observó que la utilización de un promedio de estas dos series lograba mejores resultados en las estimaciones, además dado que se tiene que trabajar con un rezago temporal cercano a los 2 meses con respecto al mes a estimar, se

deben hacer estimaciones de la movilidad con pocos datos, por lo cual, resulta más conveniente la utilización de una sola medida de movilidad, que estará dada por el promedio simple en cada mes, del dato de movilidad medio de Comestible/Farmacia y Comercio/Recreación, a continuación se muestra un detalle del promedio desarrollado por mes y se denota como el comportamiento en el tiempo es muy similar a la demanda de combustibles derivados de hidrocarburos seleccionados:

Gráfico 15. Costa Rica, promedio mensual de variación diaria de movilidad media con respecto a un valor de referencia prepandemia y consumo de combustibles derivados de hidrocarburos seleccionados, febrero 2020 a diciembre 2020.



Fuente: Elaboración propia con datos del reporte de movilidad de Google y de Aresep.

Al momento de realizar los pronósticos, no se cuenta directamente con el dato de movilidad para el mes a pronosticar, tal y como se mencionó anteriormente, por ejemplo, para la estimación de julio 2020, se tiene el dato de movilidad promedio hasta mayo, por ello, para emplear estos datos de la mejor manera, dada la importancia y volatilidad de esta variable, se requiere realizar algún tipo de estimación que permita su utilización en los modelos respectivos. Dado que se tienen pocos datos, no es posible hacer una estimación de serie de tiempo, entonces se propone lo siguiente:

Para el mes previo a la estimación, se tomará el dato de movilidad promedio de los primeros 8 días (los cuales siempre estarán disponibles a la fecha de corte) y con este dato y el de los dos meses previos se desarrollará un promedio móvil. Por ejemplo, para la estimación de julio, se tomará el dato promedio de los primeros 8 días de junio, y se calculará un promedio móvil, de ese dato, junto con los datos promedio de mayo y abril, el dato obtenido será el estimado para junio. Sin embargo, note que el dato que se requiere es el de movilidad de julio, para este no se puede seguir el mismo procedimiento, entonces se tomará el dato estimado en junio, y se le aplicará una tasa de crecimiento, dada por la variación del promedio estimado en junio, contra el homólogo de mayo, de modo que, si el dato medio móvil decreció de mayo a junio, se espere un decrecimiento similar de junio a julio.

A continuación, se realiza una formulación matemática del procedimiento antes descrito:

$$PROM\_MOV\_E_{t+2} = PROM\_MOV\_E_{t+1} * \left(\frac{PROM\_MOV\_E_{t+1}}{PROM\_MOV\_E_{t}}\right) \quad Ecuación\ 1$$

Donde:

$PROM\_MOV\_E_{t+2}$ = Promedio esperado para la variación de la movilidad en el periodo t+2 (en el ejemplo sería julio).

$PROM\_MOV\_E_{t+1}$ = Promedio esperado para la variación de la movilidad en el periodo t+1 (en el ejemplo sería junio).

$PROM\_MOV\_E_{t}$ = Promedio esperado para la variación de la movilidad en el periodo t (en el ejemplo sería mayo).

Mientras que el promedio esperado para cada periodo se determina por medio de un promedio móvil de la siguiente manera:

$$PROM\_MOV\_E_{t+1} = \frac{MM_{t-1} + MM_t + MI_{t+1}}{3} \quad Ecuación\ 2$$

Donde:

$PROM\_MOV\_E_{t+1}$ = Promedio esperado para la variación de la movilidad en el periodo t+1 (en el ejemplo sería junio).

MM_{t-1} = Promedio observado de la variación de la movilidad de todo el mes t-1 (en el ejemplo sería abril).

MM_t = Promedio observado de la variación de la movilidad de todo el mes t (en el ejemplo sería mayo).

MI_{t+1} = Promedio observado de la variación de la movilidad en los primeros 8 días del mes t+1 (en el ejemplo sería junio).

De esta manera a partir de las observaciones previas se hace una estimación de la movilidad esperada en dos meses y este dato constituiría el insumo base a utilizar en la estimación consumo de combustible del mes $t + 2$, aprovechando al máximo todos los datos observados de movilidad hasta el corte del estudio.

De igual manera, de modo auxiliar se utilizó una variable dicotómica cuyo valor es igual a 1 en abril y julio, y representa las dos restricciones sanitarias más fuertes que se dieron para evitar las aglomeraciones con motivo de las vacaciones y que buscaban limitar los desplazamientos al mínimo.

Para la utilización de la información relacionada con casos de Covid-19 se procedió de un modo similar al tratamiento empleado para la estimación de la movilidad, dado que el desfase de dos meses en la información hace imposible la inclusión contemporánea y se presentan pocos datos para realizar una estimación por medio de un modelo. Por lo anterior , se procedió a estimar una cantidad de casos diarios promedio para el mes, el cual se estima de un modo análogo a lo indicado en las ecuaciones 1 y 2, es decir usando una tasa de crecimiento y los datos de promedios móviles, tal y como se explicó anteriormente.

En resumen, estas serían las variables creadas adicionalmente por medio de las fórmulas y criterios antes descritos, a partir de información real existente, y que se emplearon en el desarrollo de los modelos:

Cuadro 4. Variables explicativas creadas adicionalmente para el desarrollo de los modelos, 2021.

Variable	Descripción	Periodicidad	Unidad
RESTRICCION	Variable dicotómica, que tomar valor de 1 en los meses de abril y julio de 2020, al ser los meses en los cuales se presentaron las mayores restricciones.	Mensual	Valores de 0 ó 1.
PROM_MOV_E	Movilidad diaria promedio esperada de los individuos en Costa Rica, para el mes de referencia calculado por medio de la ecuación 1.	Diario, y se hace un promedio por mes	Puntos porcentuales
PROM_COVID_E	Cantidad diaria promedio esperada de casos nuevos por Covid-19 en Costa Rica para el mes de referencia.	Diario, y se hace un promedio por mes	Cantidad de casos
TASA R-2	Promedio mensual de la tasa R diaria de contagio, para el periodo t-2, es decir dos meses antes, al mes de referencia.	Diario, y se hace un promedio por mes	Tasa de contagio

Fuente: Elaboración propia.

A continuación, se procede a desarrollar la metodología específica que se empleó en cada uno de los modelos estimados:

4.3.Modelo Holt Winters y ARIMA

Para la estimación del modelo Holt Winters multiplicativo, tal y como se mostró en el marco teórico, se emplean únicamente los datos de la serie de tiempo a estimar, es decir, se utiliza la información histórica de la variable LITROS, con ella se determinan los índices de estacionalidad, se definen valores iniciales para a_{t-1} y b_{t-1} y por medio del siguiente sistema de ecuaciones se estima los coeficientes $\propto$, β y γ y con ellos, se estiman los litros para los periodos siguientes.

$$a_t = \propto \frac{LITROS_t}{S_{t-s}} + (1-\propto)(a_{t-1} + b_{t-1})$$

$$b_t = \beta\,(a_t - a_{t-1}) + (1-\beta)b_{t-1}$$

$$S_t = \gamma \frac{LITROS_t}{a_t} + (1-\gamma)S_{t-s}$$

$$\widehat{LITROS}_{t+m} = (a_t - b_t m)S_{t-s+m}$$

En función de lo anterior, el proceso de estimación es algorítmicamente sencillo y por medio del comando programado en el software R se estiman los valores de los coeficientes y las estimaciones de litros para los periodos siguientes.

El segundo modelo desarrollado es el modelo ARIMA, el cual como se explicó en el marco teórico contempla una ecuación basada en rezagos de componentes autorregresivos y estacionales de la serie histórica a estimar y de las desviaciones observadas en periodos previos entre el valor estimado y observado. Además, incluye componentes de diferenciación regular o estacional, a fin de buscar la estacionariedad de la serie.

El orden de los rezagos y diferencias a utilizar en el componente regular o estacional, son los que terminan determinando la estructura del modelo a estimar, y para ello, se emplea la notación ARIMA(p,d,q)(P,DQ)[s]. En resumen, para seleccionar el modelo lo que se debe determinar son los valores de p,d,q,P,D y Q, y para establecer dichos valores se emplearon dos métodos:

- El uso de la función autoarima de R, la cual se basa en el algoritmo de Hyndman-Khandakar (2008), el cual se basa en un proceso iterativo de los parámetros y comparación a partir de criterios de información como el AIC y el BIC.
- Se utilizó un modelo seleccionado a partir del análisis visual de los correlogramas, determinando a partir de los patrones teóricos, cuál podría ser la estructura del modelo respectivo.

Esto se realizó para cada combustible, de modo que se obtuvo dos posibles especificaciones para cada producto. Para los 6 modelos seleccionados, se procedió a realizar la estimación correspondiente, se analizaron los estadísticos AIC, BIC, y MAPE de validación cruzada y se seleccionó la especificación para cada producto que presentó los mejores resultados.

Una vez determinada la especificación funcional se estiman los coeficientes, por medio de la siguiente ecuación:

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s{}^D LITROS_t = \Theta_Q(B^s)\theta_q(B)a_t$$

Una vez realizada la estimación se deben revisar los diferentes supuestos de este modelo:

Sobre los coeficientes estimados se realiza la prueba de hipótesis de igualdad a cero para valorar su significancia, de igual manera se analizan las raíces de cada uno, a fin de determinar si se encuentran dentro del círculo unitario, pues de esa manera se podrá validar la invertibilidad y por consecuencia la estabilidad del modelo.

Después de validados las hipótesis sobre los coeficientes se analizan los supuestos sobre los residuos, es decir se debe verificar la ausencia de autocorrelación, para ello se utilizará el valor p del estadístico Q, de la prueba de autocorrelación de Ljung-Box. Además, se analizó la normalidad, por medio del análisis de la distribución de los residuos y la prueba de Shapiro-Wilk de normalidad, y para el análisis de homocedasticidad se utilizó la prueba ARCH-LM. En relación con la estacionariedad de la serie, se realizó un análisis visual y se realizó la prueba de Dickey-Fuller, cuya hipótesis nula es que la serie tiene raíz unitaria

4.4.Modelo de Series de Tiempo Bayesianas Estructurales

Para el caso del modelo de Series de Tiempo Bayesianas Estructurales, de conformidad con lo indicado en el marco teórico, se estimó el siguiente sistema de ecuaciones:

$$LITROS_t = \mu_t + \tau_t + \beta_1 * (MES\ T - 2)_t + \beta_2 * (MES\ A - 1)_t + \beta_3 * RESTRICCION_t + \beta_4 * PROM\_MOV\_E_t + \beta_5 * PROM\_COVID\_E_t + \beta_6 * (TASA\ R - 2)_t + \varepsilon_t$$

$$\mu_t = \mu_{t-1} + \delta_{t-1} + u_t$$

$$\delta_t = \delta_{t-1} + v_t$$

$$\tau_t = -\sum_{s=1}^{S-1} \tau_{t-s} + w_t$$

Note que, bajo este modelo se deben estimar los coeficientes β_i para cada variable explicativa, mientras que los demás componentes asociados a parámetros que buscan aproximar la tendencia y la estacionalidad se determinan de modo endógeno por medio del sistema en cuestión a partir de valores iniciales en un proceso iterativo.

Para el desarrollo de los cálculos respectivos, se asume que las variables ε_t, u_t y w_t se distribuyen de modo normal con media cero, y los demás parámetros a estimar se estiman de modo bayesiano, empleando una distribución a priori normal y una varianza estimada por medio de una distribución a priori Gamma (Scott & Varian, 2014, pág. 11).

Para obtener las distribuciones a posterior, se utilizó un muestreo de Gibbs, a partir de las distribuciones condicionales de los parámetros, por medio de simulaciones de Cadenas de Markov Monte Carlo (MCMC), a continuación, se detallan los pasos seguidos según (Scott & Varian, 2014, pág. 12):

- Se determinaron las variables latentes de los componentes por medio de la simulación suavizada de Durbin y Koopman.
- Posteriormente se realizan las distribuciones condicionales de cada uno de los parámetros, en función de las variables latentes estimadas en el punto anterior y de

las distribuciones condicionales de los otros parámetros por medio de un muestreo de Gibbs.

- Posteriormente con las variables latentes y los parámetros estimados, se generan las estimaciones de los valores de $LITROS_t$.

En relación con las estimaciones a desarrollar, estos se deben estimar a partir de la distribución posterior estimada, y con base en las funciones de densidad, se estima la integral, que nos suministra la probabilidad esperada y con esta se determina el valor esperado para cada predicción (Scott & Varian, 2014, pág. 13).

$$p\left(\widetilde{LITROS}_{t+m} \middle| LITROS_t, X_t\right) = \int p\left(\widetilde{LITROS}_{t+m} \middle| \phi\right) p(\phi | LITROS_t, X_t) d\phi$$

Donde ϕ, representa los coeficientes a estimar: $\delta_t, \mu_t, \tau_t, \beta$ y la X_t representa el conjunto de covariables: $(MES\ T-2)_t$, $(MES\ A-1)_t$, $RESTRICCION_t$, $PROM\_MOV\_E_t$, $PROM\_COVID\_E_t$, $(TASA\ R-2)_t$

En resumen, la estimación estaría dada por la esperanza matemática de los litros para los periodos $t+m$, con base en una distribución posterior, la cual utiliza los parámetros estimados del modelo iterativo, y la información disponible de las variables explicativas, lo que en el fondo provee una especie de promedio bayesiano de los distintos valores esperados con base en la distribución posterior.

Una vez, realizado los cálculos anteriormente descritos, por medio del paquete BSTS de R, se procederá a revisar los coeficientes estimados para las covariables, los signos y magnitudes resultantes y la probabilidad de inclusión calculada por el mismo algoritmo, con base en estos resultados se analizará si guarda consistencia con el comportamiento teórico esperado para cada producto. También se procederá a verificar la bondad de ajuste con base en la estimación desarrollada en el proceso MCMC y si el promedio de dicho proceso iterativo se aproxima al valor real observado, también se realizará esta comparación para el modelo, con y sin covariables a fin de observar si se reduce el error absoluto acumulado, lo anterior también se logra por medio de las visualizaciones que provee el paquete de R.

Por último, con base en la teoría bayesiana no se debe realizar una revisión de supuestos en los residuos, sin embargo, es necesario, validar la convergencia del modelo, lo anterior se analiza con base en la desviación estándar de los resultados observados entre las distintas iteraciones, esto sería relativamente similar al estadístico Gelman and Rubin's que busca comparar las varianzas entre cadenas y dentro de las cadenas, sólo que en el presente trabajo se realiza de un modo gráfico, para ello, se revisa para cada iteración si los valores estimados tienden a ser relativamente similares dentro de la cadena, lo que implica que la desviación estándar tiende a decrecer, y también se espera que tienda a estabilizarse entre cada iteración, es decir, se espera que estas desviaciones tengan una tendencia decreciente al inicio entre cada iteración y que posteriormente se aplane la curva como señal de estabilización, lo que implica la convergencia, pues nos indica que las cadenas están arrojando resultados parecidos entre iteraciones, para cada valor estimado.

4.5.Modelo XGBoost

En relación con el modelo XGBoost, tal y como me mencionó en el marco teórico, busca generar un conjunto de K estimaciones a partir de un proceso de remuestreo con reemplazo, y a partir de la combinación de dichos resultados, permite obtener el resultado final. En nuestro caso, lo anterior se logra reflejar en la siguiente ecuación:

$$\widehat{LITROS_t} = \sum_{k=1}^{K} f_k((MES\,T-2)_{\,t}, (MES\,A-1)_t, \quad RESTRICCION_t, PROM\_MOV\_E_t, \\ PROM\_COVID\_E_t \;\; (TASA\,R-2)_t) \quad Ecuación\ 3$$

Cada una de las funciones f_k a utilizar son árboles de regresión, y para la determinación de los parámetros óptimos, se realiza un proceso iterativo, en el cual con base en los errores de estimación o diferencias entre el valor real y el estimado, se busca ir mejorando el modelo entre cada iteración, para ello se busca potenciar los resultados del modelo "más débil", es decir busca enfocarse en las deficiencias y de esa manera ir logrando corregirlas a fin de mejorar las estimaciones, para lograrlo se realiza una optimización de la siguiente función:

$$L^t = \sum_{t=1}^{n} l(LITROS_t, \widehat{LITROS}_t + f_p(x_t)) + \Omega(f_p) \quad Ecuación\ 4$$

Donde t representa cada periodo de la serie de tiempo, desde el primer periodo hasta el último (n) y donde x_t representa el conjunto de variables explicativas, es decir: $x_t = (MES\ T-2)_t, (MES\ A-1)_t, RESTRICCION_t, PROM\_MOV\_E_t, PROM\_COVID\_E_t, (TASA\ R-2)_t$.

En la ecuación 4 básicamente lo que se tiene es una función de error convencional, por ejemplo, la raíz del error cuadrático medio (RMSE), con el componente adicional de que los litros estimados se obtienen por medio de $\widehat{LITROS}_t + f_p(x_t)$, es decir se obtienen de la combinación de los modelos, tal y como se mencionó en la ecuación 3, pero se le adiciona el resultado del modelo "más débil" $(f_p(x_t))$, es por medio de esta adición que se da una mayor ponderación a este modelo, buscado mejorarlo para la próxima iteración por medio de parámetros óptimos que reduzcan el error previo, de esa manera se espera que después de varias iteraciones los modelos "más débiles" se hayan fortalecido en este proceso, y permitan de esta manera lograr contribuir en una mejor estimación de la variable dependiente. Aunado a lo anterior, se reitera que la expresión $\Omega(f_p)$ es una función que se adiciona para evitar el sobre ajuste del modelo, tal y como se explicó en el marco teórico.

Para determinar los litros estimados a futuro, lo que se debe hacer es utilizar la ecuación 35, con las variables explicativas para el periodo que se desee estimar, en dicha ecuación, se deberán emplear los parámetros óptimos que se hayan determinado a partir del proceso iterativo.

Para la obtención de los resultados a asociados a estos modelos se deben definir un conjunto de parámetros iniciales, a continuación, se detallan cuáles fueron los parámetros empleados:

Cuadro 5. Parámetros utilizados en los modelos XGBoost estimados para las series de demanda de combustibles derivados de **hidrocarburos, 2021.**

Rubro	Parámetro
Máximo de iteraciones	Entre 100 y 200
Máximo de árboles	Entre 6 y 20
Penalización por complejidad	0
Parámetro de regularización	0,2
Tipo de modelo	Árboles de regresión

Fuente: Elaboración propia

Es importante indicar que se desarrollaron diversos modelos para cada producto variando los parámetros indicados en el cuadro anterior, sin embargo, para todos los productos estos fueron los parámetros que ofrecieron los mejores resultados de estimación.

Una vez estimados los modelos, la librería de R utilizada para este modelo, permite determinar el porcentaje de importancia de cada variable explicativa, el cual se determina en el proceso iterativo, a partir de la capacidad de discriminación que logra aportar cada variable en los árboles de regresión, específicamente en la función de ganancia de discriminación por Entropía o Gini, es decir, determina para cada variable, como su utilización permite mejorar el índice de Entropía o Gini entre cada nodo, y de este modo, como permite mejorar la determinación de cuál podría ser el valor estimado, dadas las variables explicativas.

Los modelos de minería de datos tratan el mecanismo de datos como desconocido (Breiman, 2001) y por tanto existe una diferencia en el número de suposiciones dado que los modelos de minería de datos no requieren que se especifique la distribución de la variable dependiente o independiente (Deoras, 2017), lo que hace que estas técnicas sean menos restrictivas ya que permite ser utilizada con los mínimos supuestos posibles (Beltrán, 2005, págs. 20-21).

Por lo anterior para los modelos de minería de datos, lo que se verificará es que el modelo converja, lo cual se verificará en los resultados obtenidos en el software y la capacidad predictiva por medio de funciones de pérdida o de error y por consiguiente no es necesario realizar ningún análisis de cumplimiento de supuestos.

4.6. Modelo de minería de datos que utiliza como variables explicativas las estimaciones de otros modelos.

Uno de los elementos metodológicos más importantes en el presente trabajo, fue la construcción de un modelo de minería de datos que utilice como variables explicativas los resultados obtenidos por otros modelos.

Lo anterior, se representa en términos matemáticos de la siguiente manera:

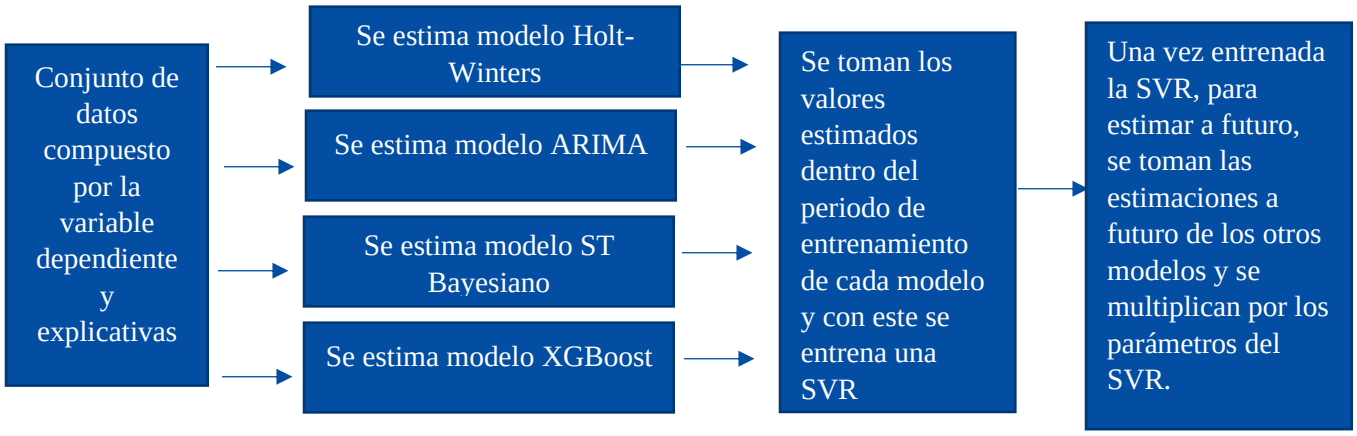
$$\widehat{LITROS}_{t+1} = f\left(\widehat{MHW}_{t+1}, \widehat{MAR}_{t+1}, \widehat{MBST}_{t+1}, \widehat{MXG}_{t+1}\right) \; Ecuación\ 5$$

De este modo, note como los litros estimados por medio de este modelo, están en función de los litros estimados por el modelo Holt Winters ($\widehat{MHW}_{t+1}$), por los litros estimados del modelo ARIMA ($\widehat{MAR}_{t+1}$), los litros estimados por el modelo de series de tiempo bayesianas ($\widehat{MBST}_{t+1}$) y las estimaciones del modelo XGBoost ($\widehat{MXG}_{t+1}$).

Tal y como se mencionó en el marco teórico, para la determinación de la función f que permitirá la conjugación de las estimaciones de los diversos modelos, se utilizará una regresión de Soporte Vectorial (SVR), la cual ha mostrado adecuados en trabajos similares.

Para comprender de mejor manera el proceso desarrollado, se presenta un breve esquema del proceso algorítmico:

Figura 2. Proceso seguido para la determinación del modelo conjunto.



Fuente: Elaboración propia.

De esta manera, la ecuación 5 puede ser reexpresada para el periodo de entrenamiento, empleando la forma funcional de la SVR tal y como sigue:

$$\widehat{LITROS_t} = w * \emptyset\left(\widehat{MHW_t}, \widehat{MAR_t}, \widehat{MBST_t}, \widehat{MXG_t}\right) + b \quad Ecuación\ 6$$

Recordando que $\emptyset$ es una función característica que determina el espacio vectorial o transformación no lineal, de las covariables explicativas, es decir permite realizar un reescalamiento de los datos de entrada para lograr la adecuada discriminación de los datos y de esa manera lograr una estimación más precisa de la variable dependiente.

Por su parte w representa los coeficientes estimados para las variables explicativas, que en este caso dichas covariables son las estimaciones de los otros modelos y b representa los coeficientes de constantes estimadas en el modelo. Todos estos coeficientes se determinan por medio de la minimización de una función de pérdida o error, tal y como se explicó en el marco teórico, la cual para el caso específico en cuestión estaría determinada de la siguiente manera:

$$R = C * \frac{1}{n} * \sum_{i=1}^{n} L\left(LITROS_t - \widehat{LITROS_t}\right) + \frac{1}{2}\|w\|^2$$

Donde C sería un parámetro de regularización, similar a cómo se indicó para el modelo XGBoost, por su parte $LITROS_t$ es el valor observado y $\widehat{LITROS_t}$ es el valor estimado por medio de SVR, además L es una función de pérdida o error, además se incluye el término $\frac{1}{2}\|w\|^2$, el cual se incluye en la optimización a fin de minimizar este componente y de esta manera maximizar la heterogeneidad o discriminación, es decir, por medio de este método se busca minimizar los errores de estimación, obteniendo los coeficientes w y b, los cuales a su vez, tienen la condición implícita de buscar la mejor discriminación o determinación de los valores estimados.

De modo similar al modelo XGBoost, el modelo SVR no realizan ningún supuesto sobre la distribución de los datos, ni realiza algún supuesto teórico elemental que sea necesario de verificar, si no que se debe valorar la convergencia por medio de los resultados arrojados en el software, y la capacidad predictiva del modelo para emular los datos reales, de igual manera, es necesario determinar un conjunto de parámetros iniciales, los cuales se describen a continuación:

Cuadro 6. Parámetros utilizados en los modelos SVR estimados para las series de demanda de combustibles derivados de **hidrocarburos, 2021.**

Variable explicativa	Diésel	Regular	Súper
Tipo de modelo	eps-regression	eps-regression	eps-regression
Parámetro de regularización	1	1	1
Parámetro de curvatura	0,25	0,25	0,25
Margen máximo permitido	0,1	0,1	0,1

Fuente: Elaboración propia

En el desarrollo del modelo se hicieron diversas estimaciones cambiando los parámetros indicados en el cuadro 6, sin embargo, se determinó de un análisis exploratorio de diversos valores, que estos parámetros iniciales ofrecían los mejores resultados observados, de igual manera para la determinación del kernel o función característica $\emptyset$ se procedió a realizar las estimaciones con valores lineales, radiales, sigmoideos y polinomiales, dejando para cada producto la forma funcional que ofrecía los mejores resultados dentro del periodo de entrenamiento.

Una vez obtenidos los resultados de los coeficientes, se procedió a realizar una interpretación del signo y magnitud asociado a cada variable explicativa, a fin de valorar cuáles son los modelos que están incluyendo en mayor medida en las estimaciones obtenidas.

Para el desarrollo de los modelos se utilizó el software estadístico R y su IDE R Studio, empleando las librerías: readxl, ggplot2, forecast, astsa, itsmr, tseries, zoo, tidyr, aTSA, FinTS, dplyr, xgboost, lattice, caret, Ckmeans.1d.dp, bsts y e1071.

4.7.Métodos de comparación de modelos.

Como se ha indicado anteriormente, en el presente trabajo se realizarán comparaciones de diversos modelos que buscan estimar adecuadamente la demanda de los combustibles derivados de hidrocarburos seleccionados, por esta razón es necesario determinar un método que permita una adecuada comparación.

Uno de los puntos a analizar en la comparación de los modelos, es que cumplan con los supuestos asociados, tal y como se mencionó anteriormente, para el caso de los modelos ARIMA esto implica el análisis de estacionariedad, invertibilidad, estabilidad de los coeficientes y el análisis de los supuestos de los residuos, mientras que por su parte para el modelo de series de tiempo bayesianas, esto se asocia al análisis de los valores de los coeficientes estimados y la convergencia del modelo, mientras que para los modelos de minería de datos, no se realiza ningún supuesto teórico que deba ser constatado, sino que sus resultados deberán valorarse en función de su capacidad predictiva y de generalización fuera del conjunto de entrenamiento.

En relación con la comparación entre modelos, se podría emplear el principio de la navaja de Occam, el cual indica que, en igualdad de condiciones, las teorías simples son preferibles a las complejas. El aspecto clave para su uso, es el predicado de que "todo lo demás es igual", pues antes de escoger la opción más sencilla, se debe verificar que las condiciones sean similares, y en esta verificación, es necesario analizar la habilidad predictiva, en especial en aquellos casos en que las estimaciones a futuro sean la preocupación más importante

(Brownlee, Ensemble Learning Algorithm Complexity and Occam's Razor, 2020), es decir se utilizará dicho principio cuando se obtengan resultados relativamente similares.

El uso de predictores y modelos complejos puede resultar desagradable, sin embargo puede ser necesario en aquellos casos en que la precisión predictiva es una prioridad (Breiman, 2001, pág. 208), por esta razón pese a que el uso de modelos bayesianos o de minería de datos pueden parecer modelos de estimación más complejos, deben valorarse en función de la capacidad predictiva que ofrecen, en especial en contextos de difícil estimación, como por ejemplo el periodo pandémico.

Para validar los resultados obtenidos de los diferentes modelos, se determina un período de entrenamiento y un período de prueba. Tal y como se indicó anteriormente al momento de hacer estimaciones de la demanda de combustibles derivados de hidrocarburos en estudios tarifarios tramitados en Aresep, se tiene un rezago temporal de 2 meses en la información histórica, por lo anterior, y con el fin de hacer comparables las estimaciones realizadas por Recope, contra las estimaciones que se desarrollarán en el presente trabajo, se procedió a utilizar un periodo de comparación de 6 meses, específicamente de julio a diciembre de 2020, y se utilizó información con un rezago de 2 meses en cada estimación, de este modo para estimar julio, se utilizarán datos históricos hasta mayo.

Una vez estimados los modelos, se realiza el pronóstico y se comparan los resultados contra el valor real de la demanda en los meses de prueba, por medio de una función de pérdida o de error, la cual nos permitirá indicar que tan cercanas están las estimaciones de los valores reales y por consiguiente para este periodo cuál modelo ofrece los mejores resultados. Para tales comparaciones se emplean la raíz del error cuadrático medio (RSME) y el valor del error porcentual absoluto medio (MAPE) los cuales se especifican en las siguientes ecuaciones (Alquist et al., 2011, pág. 10):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_t - \widehat{Y}_t)^2} \qquad MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_t - \widehat{Y}_t}{Y_t}\right|$$

5. RESULTADOS

A continuación, se procede a exponer los resultados obtenidos para cada uno de los modelos, y se realiza un análisis comparativo de cada uno.

5.1.Estacionariedad de la serie

En primera instancia se observan los resultados del análisis de la estacionariedad de la serie, para ello se realiza un análisis visual, el cual tal y como se puede verificar en los anexos 2, 3 y 4 muestra que la serie no presenta indicios de estacionariedad, por lo cual se procede a aplicar logaritmo, una diferenciación de nivel y una diferenciación estacional, posteriormente se aplica la prueba de Dickey-Fuller, la cual tal y como se indicó en el marco metodológico posee la hipótesis nula de que la serie tiene raíz unitaria. Como se observa en el siguiente cuadro para los tres productos se rechaza la hipótesis, lo cual reafirma la idea de que aplicando esas operaciones las series son estacionarias.

Cuadro 7. Estadístico y valor p de la prueba Dickey-Fuller de raíz unitaria para las series de demanda tipo de producto, 2021.

Producto	Estadístico DF	Valor p
Diésel	-17,71	0,01
Regular	-15,84	0,01
Súper	-15,36	0,01

Fuente: Elaboración propia.

Posteriormente se procedió a analizar para cada serie los modelos a desarrollar para estimar el comportamiento futuro de las series, a continuación, se procede con el detalle respectivo:

5.2.Modelo Holt Winters y ARIMA

El primer modelo que se desarrolló fue un modelo de suavizamiento exponencial Holt-Winters, el cual de conformidad con lo indicado en el apartado metodológico tendrá un

componente tendencial y estacional multiplicativo, en el siguiente cuadro se presentan los parámetros obtenidos:

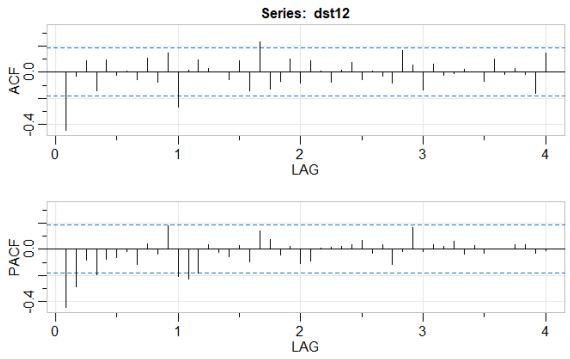
Cuadro 8.Parámetros obtenidos para el modelo Holt-Winters para las series de demanda tipo de producto, 2021.

Producto	Diésel	Regular	Súper
Alpha	0,220	0,377	0,496
Beta	0,014	0,000	0,000
Gamma	0,549	0,174	0,168

Fuente: Elaboración propia.

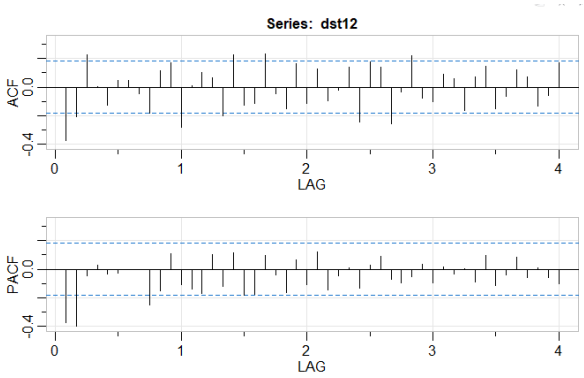
Por su parte, para el desarrollo de los modelos ARIMA, se analizaron los correlogramas de las series, posterior a la aplicación de las transformaciones respectivas para su estacionariedad, los resultados de estas series modificadas se muestran a continuación:

Gráfico 16. Diésel, correlograma de la serie del logaritmo de demanda en litros, con diferencia estacional y regular.



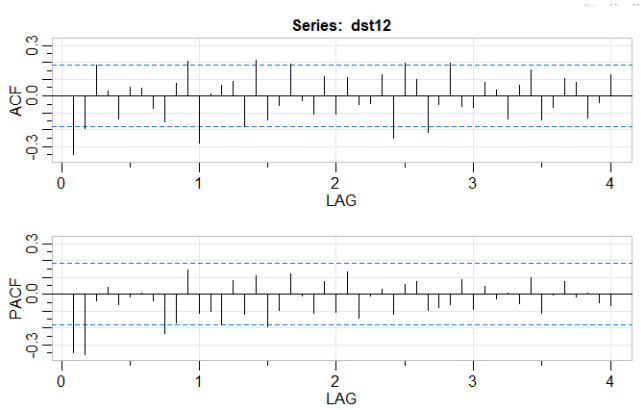
Fuente: Elaboración propia

Gráfico 17. Regular, correlograma de la serie del logaritmo de demanda en litros, con diferencia estacional y regular.



Fuente: Elaboración propia

Gráfico 18. Súper, correlograma de la serie del logaritmo de demanda en litros, con diferencia estacional y regular.



Fuente: Elaboración propia

Como se indicó en la metodología, para la determinación de los modelos ARIMA, se utilizaron dos mecanismos: el primero el uso de la función autoarima de R, la cual se basa en el algoritmo de Hyndman-Khandakar (2008) y también se utilizó un modelo seleccionado a partir del análisis visual de los correlogramas, determinando a partir de los patrones teóricos, cuál podría ser la estructura del modelo respectivo, empleando estos dos métodos se escogieron las siguientes especificaciones funcionales:

Cuadro 9. Especificación del modelo ARIMA seleccionado para las series de demanda por tipo de producto, según mecanismo de selección, 2021.

Producto	Especificación correlograma	Autoarima
Diésel	ARIMA(1,1,2)(1,1,1)[12]	ARIMA(1,0,0)(1,0,1)[12]
Regular	ARIMA(2,1,2)(1,1,0)[12]	ARIMA(1,1,1)(0,0,2)[12]
Súper	ARIMA(2,1,2)(1,1,0)[12]	ARIMA(0,1,1)(1,0,0)[12]

Fuente: Elaboración propia

Para los 6 modelos seleccionados, por medio de los estadísticos AIC, BIC, y MAPE de validación cruzada, y se determinó que, en todos los casos, la especificación por

correlograma, ofrecía mejores resultados que la especificación autoarima, además ofrecieron resultados similares en el cumplimiento de supuestos, por lo anterior, se procedió a utilizar estos modelos como los representantes de la modelación ARIMA. Por lo anterior, y con el fin de sintetizar la redacción, se mostrarán únicamente los resultados de estos modelos seleccionados, sin embargo, en el código desarrollado se pueden determinar los resultados de todos los modelos.

A continuación, se procede a mostrar los coeficientes estimados:

Cuadro 10. Coeficientes estimados para los modelos ARIMA de las series de demanda tipo de producto, 2021.

Componente del modelo	Diésel		Regular		Súper	
	Coeficiente	Desviación	Coeficiente	Desviación	Coeficiente	Desviación
Autorregresivo regular 1 (AR1)	-0,707*	0,162	-0,596	0,392	-0,484	0,672
Autorregresivo regular 1 (AR2)			-0,150	0,204	-0,037	0,262
Media móvil regular 1 (MA1)	0,091	0,150	-0,049	0,386	-0,048	0,667
Media móvil regular 1 (MA2)	-0,658*	0,104	-0,309	0,337	-0,328	0,516
Autorregresivo estacional 1 (SAR1)	-0,015	0,148	-0,395*	0,105	-0,369*	0,105
Media móvil estacional 1 (SMA1)	-0,723*	0,128				
Media móvil estacional 2 (SMA2)						

* Indica que son significativamente distintos de cero al 95% de confianza

Fuente: Elaboración propia

Sobre estos coeficientes se realizaron las pruebas de significancia, y como se observa muchos no son significativos al 95% de confianza, pero son los que guardan mayor consistencia con los patrones teóricos en el correlograma y ofrecen mejores resultados que la especificación del autoarima, por lo anterior, se dio mayor ponderación al ofrecer una mayor capacidad

predictiva. Además, como se observa en los anexos 5, 6 y 7 las raíces se encuentran dentro del círculo unitario, y por consiguiente serían invertibles.

También se realizó un análisis del cumplimiento de los supuestos asociados a cada modelo, a continuación, se muestran los resultados obtenidos.

Cuadro 11.Valor p de las pruebas de hipótesis para verificación de los supuestos de los modelos ARIMA para las series de demanda por tipo de producto, 2021.

Supuesto a analizar	Diésel	Regular	Súper
Ausencia de autocorrelación	0,806	0,477	0,781
Normalidad	0,023	0,000	0,000
Homocedasticidad	0,010	0,000	0,000

Fuente: Elaboración propia

Como se evidencia del cuadro anterior, los residuos presentan problemas, de heterocedasticidad y no normalidad, sin embargo del análisis visual, que se puede realizar en los anexos 8 y 9 y 10, se observa que estos se podrían explicar en gran medida por algunos valores extremos ocasionados por la pandemia, por lo anterior, se considera que estos modelos no cumplen a cabalidad con los supuestos de los modelos, sin embargo no se deben descartar pues poseen un nivel predictivo aceptable, y podría competir adecuadamente con los demás modelos.

5.3.Modelo de Series de Tiempo Bayesianas Estructurales

El tercer modelo estimado es el modelo de Series de Tiempo Bayesianas, en este método, se procedió a realizar un modelo, basado únicamente en tendencia y estacionalidad, y otro que incluye las covariables seleccionadas en el presente trabajo, al realizar los análisis de las capacidad predictiva del modelo por medio del RMSE y el MAPE, se observó que para todos los productos, posee mejores resultados el modelo con covariables, por ello para sintetizar la redacción sólo se mostrarán los resultados para este tipo de modelos con variables explicativas, a continuación se muestran el promedio de coeficiente estimado:

Cuadro 12. Coeficientes promedio estimados y probabilidad de inclusión por variable para los modelos BSTS de las series de demanda por tipo de producto, 2021.

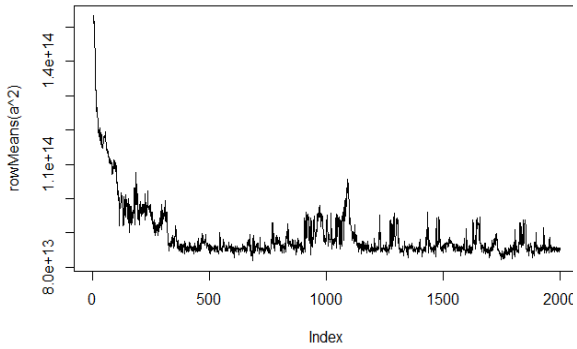
Variable	Diésel		Regular		Súper	
	Promedio	Probabilidad de inclusión	Promedio	Probabilidad de inclusión	Promedio	Probabilidad de inclusión
PROM_MOV_E	498398,80	0,77	599180,00	1,00	481489,70	1,00
RESTRICCION	-3930700,00	0,14	-79757,01	0,01	-45816,25	0,01
TASA R-2	0,00	0,04	-1441,61	0,00	171,62	0,00
MES A-1	23793,62	0,03	0,00	0,01	0,00	0,00
MES T-2	-235692,10	0,02	0,00	0,01	0,00	0,01
COVID CR-2	0,00	0,01	3414,73	0,02	2170,58	0,01
Intercepto	0,00	0,00	0,00	0,00	0,00	0,00

Fuente: Elaboración propia

Como se logra observar, de los resultados obtenidos se nota que el promedio diario esperado de movilidad es el coeficiente más importante, pues agrega un mayor valor explicativo, además la restricción y la tasa R, lograron tener una participación importante en el caso del Diésel, además como se observa en el análisis visual de los anexos 11, 12 y 13, los resultados del proceso iterativo MCMC arroja resultados similares a los de la serie original y por consiguiente nos muestra que efectivamente hay una buena bondad de ajuste, al tiempo que se observa que la distribución acumulada del error absoluto de predicción con respecto al valor promedio de la distribución de iteraciones es menor en el segundo modelo que incluye covariables, es decir que la inclusión de covariables en efecto contribuye a mejorar los resultados.

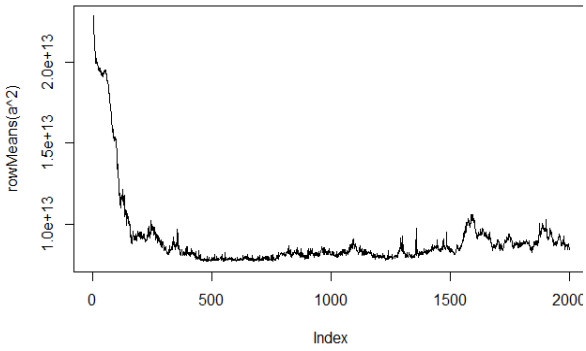
El otro elemento que se analiza para estos modelos es la convergencia del proceso iterativo MCMC, a continuación, se muestran los gráficos de la varianza de los resultados observados entre las distintas iteraciones, las cuales muestran una clara reducción conforme se fueron desarrollando las iteraciones y que los últimos valores fueron relativamente similares, lo cual nos hace pensar que efectivamente se presentó la convergencia de cada uno de los modelos.

Gráfico 19. Diésel, desviación estándar de los resultados observados entre las distintas iteraciones del modelo BSTS.



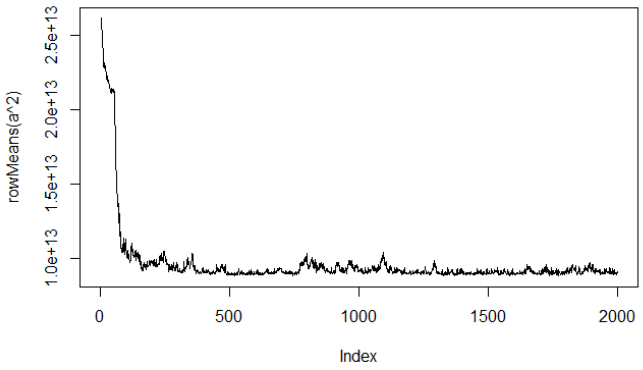
Fuente: Elaboración propia

Gráfico 20. Regular, desviación estándar de los resultados observados entre las distintas iteraciones del modelo BSTS.



Fuente: Elaboración propia

Gráfico 21. Súper, varianza de los resultados observados entre las distintas iteraciones del modelo BSTS.



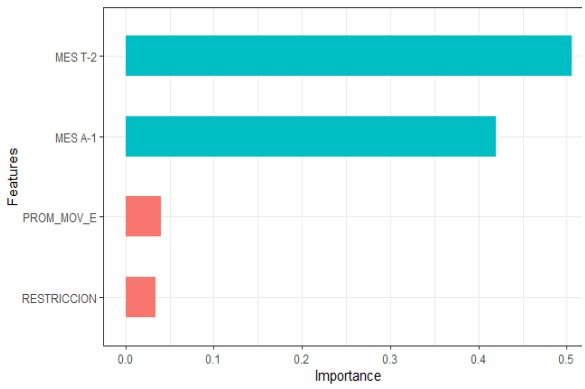
Fuente: Elaboración propia

5.4.Modelo XGBoost.

En relación con el modelo XGBoost, se procedió a realizar el proceso iterativo, de optimización asociado al algoritmo en cuestión, para lo cual se definieron los siguientes parámetros:

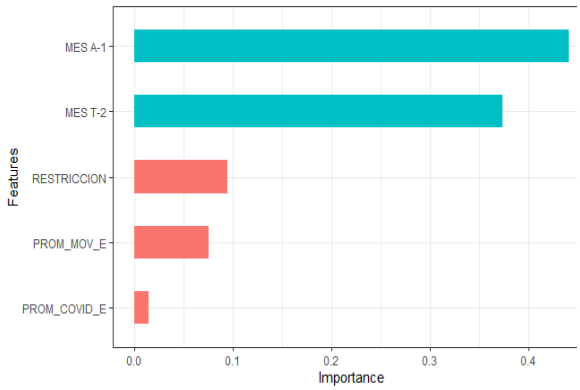
A continuación se muestra el detalle de un porcentaje de importancia que determina el modelo XGBoost para identificar las variables de influyen en mayor medida en los resultados de estimación:

Gráfico 22. Diésel, porcentaje de importancia obtenido para cada variable en el modelo XGBoost.



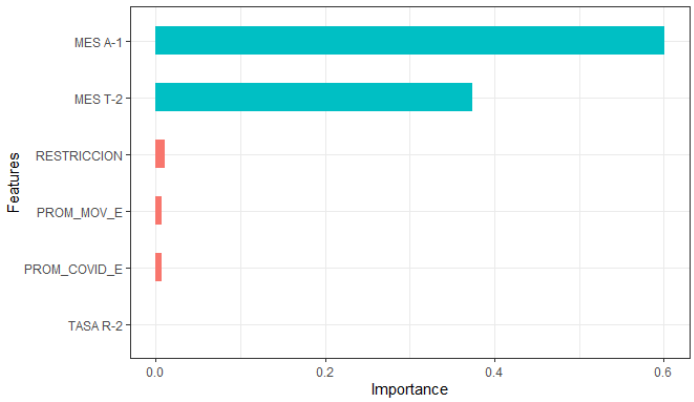
Fuente: Elaboración propia

Gráfico 23. Regular, porcentaje de importancia obtenido para cada variable en el modelo XGBoost.



Fuente: Elaboración propia

Gráfico 24. Súper, porcentaje de importancia obtenido para cada variable en el modelo XGBoost.



Fuente: Elaboración propia

Se observa de los resultados, que el modelo logra emplear la información del rezago de los litros de dos meses previos y el rezago de un año previo de un modo importante, emulando hasta cierto punto la lógica de los modelos ARIMA, además logra aprovechar la información de la movilidad esperada, la tasa R de los 2 meses previos y la cantidad esperada de contagios por Covid-19, se nota como para la gasolina regular estas variables cobran más importancia que para los otros productos, lo cual como veremos más adelante contribuye para minimizar los errores.

5.5. Modelo de minería de datos que utiliza como variables explicativas las estimaciones de otros modelos.

En relación con el modelo SVR que utiliza como variables explicativas las estimaciones de otros modelos, se realizaron las respectivas estimaciones y se observó que el algoritmo de un modo robusto logró incorporar adecuadamente las estimaciones de los otros modelos y en todos los casos alcanzó la convergencia.

En relación con el kernel o función característica $\emptyset$ se realizaron las pruebas de diversas formas funcionales y se determinó que para los productos Diésel y Regular los mejores resultados para el periodo de entrenamiento se lograron por medio formas funcionales lineales, mientras que para el caso de la gasolina Súper se obtuvo una especificación polinomial.

A su vez se obtuvieron los coeficientes en cada modelo, los cuales se muestran a continuación, con excepción de los de la gasolina súper, los cuales al ser un kernel polinomial, se estiman en un espacio ∅ diferente y se estimaron cerca de 102 coeficientes, por ello, sólo se muestran los resultados para regular y diésel, los coeficientes estimados fueron los siguientes:

Cuadro 13. Coeficientes estimados para los modelos SVR que usan como variable explicativa los resultados de modelos previos de las series de demanda tipo de producto, 2021.

Variable explicativa	Diésel	Regular
Constante	-0,186	-0,017
XGBoost	0,061	-0,080
ARIMA	-0,029	0,303
BSTS	0,365	-0,208
Holt-Winters	-0,211	0,002

Fuente: Elaboración propia

Como se aprecia para el modelo estimado, los valores proporcionados por el modelo BSTS y Holt Winters son los más incluyentes para el producto Diésel, mientras que para la gasolina Regular son los modelos BSTS y ARIMA los que poseen coeficientes más importantes, en algunos casos con signo positivo, pues se considera que subestiman el valor esperado, o en signo negativo pues se espera que sobrestimen los posibles valores a alcanzar.

En el caso del modelo para la gasolina súper, como se indicó anteriormente se seleccionó la forma funcional polinomial, la cual logró por medio de patrones no lineales emplear los diferentes resultados para alcanzar una mejor aproximación entre el valor real y el estimado.

5.6.Estimaciones a futuro y comparación de resultados

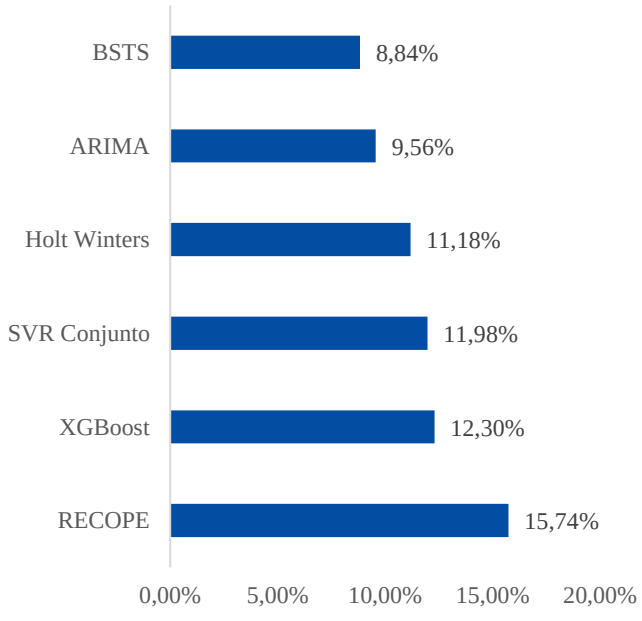
Por último, se presentan los resultados obtenidos para el periodo de julio a diciembre de 2020, en este proceso de validación fuera de muestra, se recuerda que existe un rezago temporal de dos meses previos, de este modo para julio se utilizó información real hasta mayo, y así

sucesivamente y cada modelo empleó toda la información disponible según su forma funcional y variables incluidas, además para su comparación, se utilizó el RMSE y el MAPE, tal y como se indicó en la sección metodológica. Además, se incluyen las estimaciones realizadas por Recope, en los estudios tarifarios, para establecer un parámetro de comparación.

En los próximos 6 gráficos, se muestran los resultados de estos indicadores por producto, en primera instancia se revisará el detalle del diésel, posteriormente de la gasolina regular y por último de la gasolina súper, cada uno por separado con el fin de poder profundizar en su análisis.

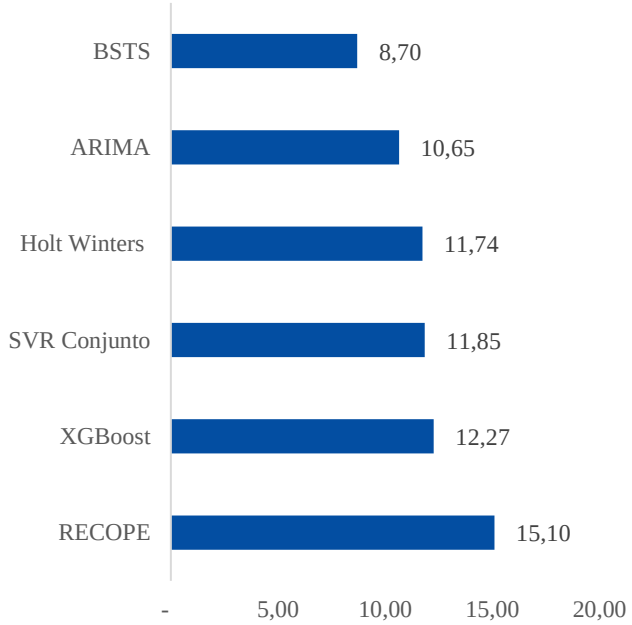
Los resultados en cuestión para el diésel son los siguientes:

Gráfico 25. Diésel, MAPE calculado para las estimaciones de julio a diciembre 2020.



Fuente: Elaboración propia

Gráfico 26. Diésel, RMSE calculado para las estimaciones de julio a diciembre 2020.



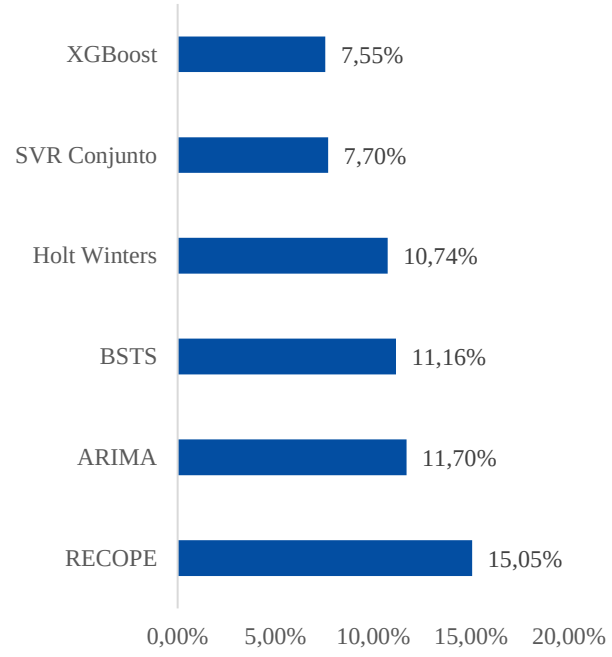
Fuente: Elaboración propia

Como se logra observar, para el caso del diésel, los resultados muestran que el uso del modelo BSTS, muestra resultados relativamente similares entre el RMSE y el MAPE, lo cual permite

observar que este tipo de modelación en series de tiempo que han tenido una perturbación fuerte como la experimentada por el Covid-19, pero con un efecto menos severo que el observado en las gasolinas (como se verá más adelante), permiten resultados similares a los modelos de series de tiempo más usuales pero con mejores resultados al contemplar por medio de estas covariables la incertidumbre existente en el mercado, como se refleja en el anexo 14. Además, este resultado es consistente con el hecho de que, para el diésel, las covariables tuvieron probabilidades inclusión más altas que en otros productos, contribuyendo con el poder explicativo del modelo.

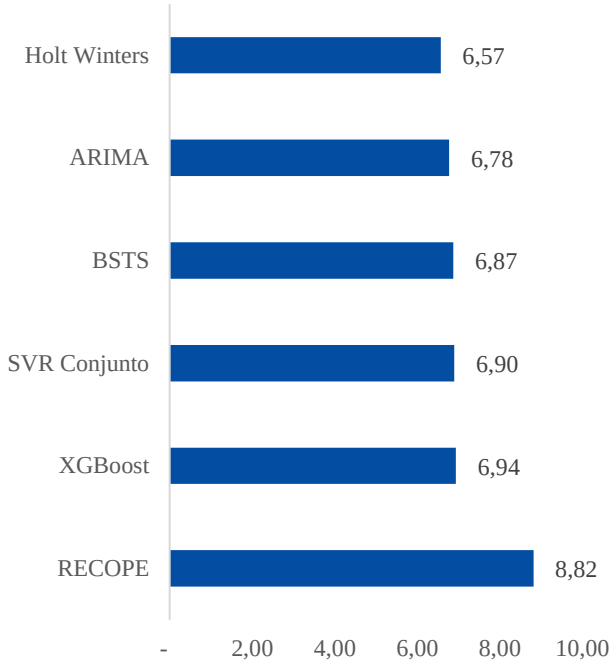
Por otra parte, para la gasolina regular, se observa que los resultados entre MAPE y RMSE difieren, pues el primer indicador nos arroja que el modelo XGBoost posee un error porcentual medio de 7,55%, logrando el mejor resultado de todos los otros modelos, aunque relativamente similar al del modelo SVR Conjunto, sin embargo, esto no se cumple con el RMSE, el cual muestra resultados muy similares entre todos los modelos con excepción de la estimación realizada por Recope.

Gráfico 27. Regular, MAPE calculado para las estimaciones de julio a diciembre 2020.



Fuente: Elaboración propia

Gráfico 28. Regular, RMSE calculado para las estimaciones de julio a diciembre 2020.

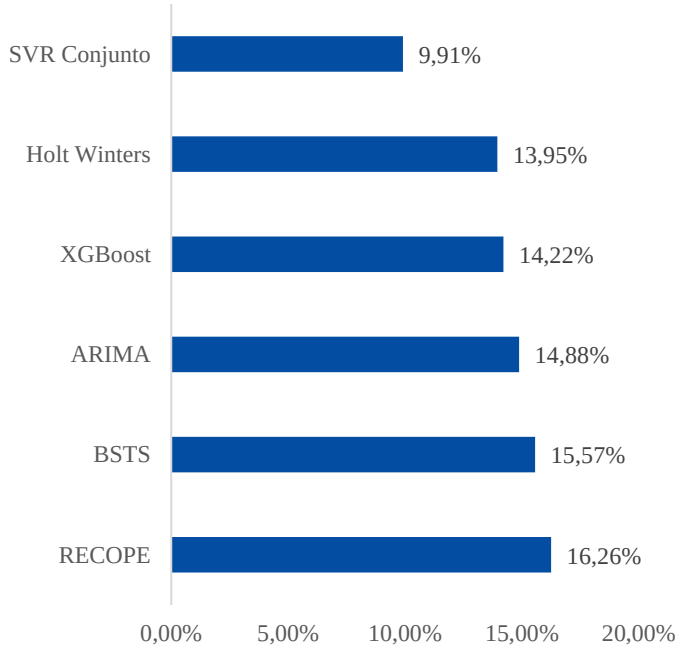


Fuente: Elaboración propia

Lo anterior, se explica en gran medida por el mes de diciembre, el cual se aleja sustancialmente de las estimaciones desarrolladas por los modelos, sin embargo, si se excluyera este resultado ambos indicadores arrojarían que el modelo XGBoost presenta los mejores resultados, lo cual también se puede validar en el anexo 15.

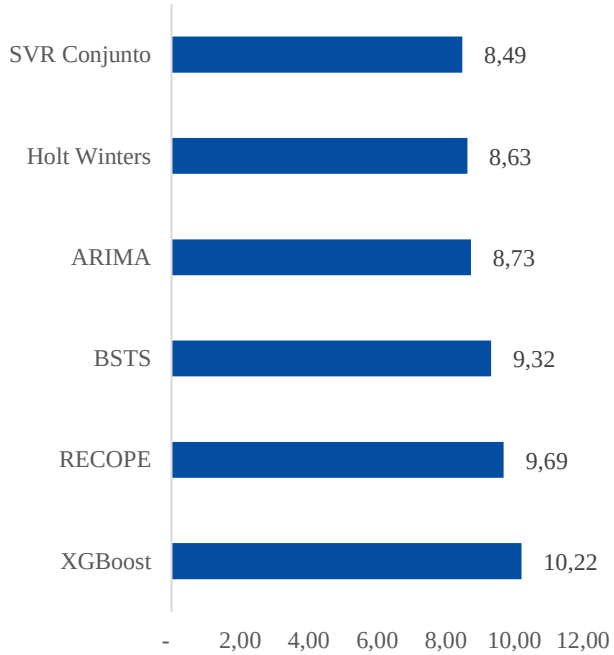
Por último, para la gasolina súper se muestran los resultados más interesantes, pues en este producto es donde se observa que el Covid-19, tuvo una mayor afectación, y los errores de estimación fueron más altos, sin embargo, el modelo SVR Conjunto logró mejorar sustancialmente el valor indicado en el valor porcentual absoluto medio (MAPE).

Gráfico 29. Súper, MAPE calculado para las estimaciones de julio a diciembre 2020.



Fuente: Elaboración propia

Gráfico 30. Súper, RMSE calculado para las estimaciones de julio a diciembre 2020.



Fuente: Elaboración propia

De un modo similar a la gasolina regular, el valor del RMSE se ve influenciado por el valor de diciembre el cual se aleja del valor real, pues no se esperaba un repunte tan importante como el observado, sin embargo ambos indicadores nos muestran que el modelo conjunto es el que ofrece mejores resultados, además llama la atención como efectivamente este modelo logra aprovechar las bondades de cada uno de los modelos individuales, pues pese a que los demás modelos ofrecen un MAPE cercano al 15%, este modelo logra resultados cercanos al 10%, lo anterior se explica por el hecho de que algunos modelos ofrecieron mejores resultados en ciertas condiciones y otros modelos ofrecieron mejores resultados en otras, entonces el modelo conjunto permitió aprovechar las bondades de cada modelo al momento de la estimación, logrando priorizar los resultados de los modelos dependiendo de las condiciones, y permitiendo por tanto que la conjugación ofreciera mejores resultados que el

uso de un solo modelo de modo específico, lo anterior se logra observar de un modo más claro en los resultados observados en el anexo 16.

6. CONCLUSIONES

Del trabajo desarrollado en la presente investigación se pueden extraer las siguientes conclusiones:

1. El Covid-19 generó un impacto sobre la demanda de combustibles nunca antes visto y de dimensiones difíciles de predecir, de hecho, es una de las demandas con mayor afectación a nivel nacional e internacional, pues las restricciones sanitarias, afectan la movilidad de los agentes y esto a su vez impacta el consumo de combustibles.
2. Por lo anterior, las estimaciones del comportamiento de la demanda de combustibles derivados de hidrocarburos son muy complejas pues dependen de las restricciones a la movilidad, las cuales a su vez dependen de la evolución del virus, y como se mostró en la sección relacionada con los efectos del Covid-19. Diferentes entes con un alto respaldo técnico están constantemente cambiando sus pronósticos, por el alto nivel de incertidumbre, lo que demuestra que esta realidad implica una gran dificultad para los modelos que se desarrollen.
3. Ante un escenario como este, la Estadística permite utilizar información constante sobre la evolución del virus y factores que puedan influir en la movilidad, por ello el uso de información como la cantidad de casos nuevos, el seguimiento satelital de la movilidad, la tasa de contagio y el nivel de vacunación, son de gran importancia, pues como se demostró en los modelos empleados en el presente trabajo, la inclusión de esas variables logra mejorar de un modo considerable los resultados de estimación, al permitir ofrecer información oportuna, confiable y adecuada para predecir el comportamiento del consumo de combustibles.
4. La computación ha logrado por medio del uso de infraestructura y técnicas para grandes datos proveer casi en tiempo real información relevante, por lo cual se puede aprovechar estas nuevas técnicas computacionales como oportunidades que brinda la era tecnológica para mejorar el desarrollo de modelos y toma de decisiones. Lo anterior es de gran importancia, pues como se mencionó en el trabajo, los patrones de movilidad están cambiando en el tiempo, pues estamos ingresando en una “nueva

normalidad", por ello no se pueden mantener modelos estáticos, hay que estar mejorándolos constantemente y desarrollando un continuo monitoreo.

5. Además del uso de información oportuna, actualmente hay importantes avances en el desarrollo de modelos de series de tiempo, los modelos ARIMA y de suavizamiento exponencial, han sido de gran ayuda y ofrecen buenos resultados, sin embargo no son la única opción, y por tanto es adecuado emplear nuevas metodologías como las Series de Tiempo Bayesianas Estructurales (BSTS) y modelos de minería de datos como el XGBoost y la Regresión de Soporte Vectorial (SVR), los cuales como se evidenció en el trabajo pueden ofrecer muy buenos resultados e inclusive superiores a los métodos más tradicionales.
6. Como se observó de los análisis realizados, el modelo BSTS ofreció el mejor resultado para el consumo de Diésel con un MAPE de 8,84%, el modelo XGBoost ofreció el mejor resultado para el consumo de gasolina Regular con un MAPE de 7,55%, y el resultado que más llama la atención es que el modelo SVR ofreció los mejores resultados para la gasolina Súper con una MAPE de 9,91%. Este resultado es de gran importancia pues demostró que en periodos de alta incertidumbre y con cambios importantes en el patrón de consumo, el uso de diversos métodos de modelación es importante (pues para cada combustible el mejor resultado se obtuvo por métodos diferentes), además evidenció que los modelos se pueden integrar en un modelo conjunto como se desarrolló en el SVR, en donde se utilizaron como variables explicativas los resultados de los otros modelos, y esto permitió aprovechar las bondades que cada modelo ofrecía para obtener mejores resultados en vez de utilizar cada modelo por separado, ofreciendo una nueva alternativa para escenarios de predicción complejos en donde los modelos pueden complementarse.
7. Estos resultados son consistentes con los observados en investigaciones previas, pues tal y como se mencionó en la sección del marco teórico, se ha coincidido en que los modelos de minería de datos han ofrecido muy buenos resultados en la estimación de la demanda energética. Otra coincidencia importante con otras investigaciones es que el modelo XGBoost y la combinación de diversos modelos de minería de datos,

bayesianos y series de tiempo ha logrado mejores resultados, en comparación con los métodos tradicionales, para la estimación de series de tiempo con patrones complejos no lineales.

7. REFERENCIAS

Abbasi et al. (Marzo de 2019). *Short Term Load Forecasting Using XGBoost*. Obtenido de Researchgate: https://www.researchgate.net/publication/331746834

Alquist et al. (2011). *Forecasting the Price of Oil.* Washington, United States: International Finance Discussion Papers, Board of Governors of the Federal Reserve System.

Arce, J. L. (2020). *Costa Rica: Perspectivas económicas y políticas.* San José, Costa Rica: FCS Análisis y Estrategia.

Barbosa de Alencar et al. (2017). Different Models for ForecastingWind Power Generation: Case Study. *Energies*, 1-27.

Barton, J., van Kasteren, A., & Birk, J. (2015). *Comparing ARIMA and XGBoost algorithms on multiple time series golf data.* Stockholm, Sweden: KTH Royal Institute of Technology.

Beltrán, B. (2005). *Minería de datos.* Puebla, México: Benemérita Universidad Autónoma de Puebla.

Bhattacharyya, S., & Timilsina, G. (2009). *A Comparative Study of Energy Demand Models.* Washington, United States: The World Bank, Policy Research Working Paper 4866.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, 199-231.

Brownlee, J. (21 de Diciembre de 2020). *Ensemble Learning Algorithm Complexity and Occam's Razor.* Obtenido de Machine Learning Mastery: https://machinelearningmastery.com/ensemble-learning-and-occams-razor/

Brownlee, J. (14 de May de 2021). *A Gentle Introduction to Ensemble Diversity for Machine Learning*. Obtenido de Machine Learning Mastery: https://machinelearningmastery.com/ensemble-diversity-for-machine-learning/

CNBC. (13 de April de 2021). *Consumer News and Business ChanneL*. Obtenido de OPEC raises 2021 oil demand growth forecast on hope pandemic wanes: https://www.cnbc.com/2021/04/13/oil-opec-raises-2021-demand-growth-forecast-on-hope-pandemic-wanes.html

Consultores en Economía Dinámica SPA. (2011). *Modelos de proyección de demanda de combustibles.* Santiago, Chile: Comisión Nacional de Energía.

Culka, M. (2016). Uncertainty analysis using Bayesian Model Averaging: a case study of input variables to energy models and inference to associated uncertainties of energy scenarios. *Energy, Sustainability and Society*, 2-24.

Deoras, S. (3 de Noviembre de 2017). *Analytics indiamag*. Obtenido de How Is Machine Learning Different From Statistics: https://analyticsindiamag.com/machine-learning-different-statistics/

Economou et al. (2017). *A Structural Model of the World Oil Market: The Role of Investment Dynamics and Capacity Constraints in Explaining the Evolution of the Real Price of Oil.* Oxford, England: Oxford Institute for Energy Studies.

EIA. (2021). *Short-Term Energy Outlook July.* Washington, United States: U.S. Energy Information Administration.

Frey et al. (2009). Econometric Models for oil price forecasting a critical survey. *CESifo Forum*, 29-44.

Fuentes, V. (3 de December de 2020). *Ya hemos alcanzado el peak oil: el petróleo llega a su punto de no retorno y empieza la era dorada de los coches eléctricos*. Obtenido de Motor Pasion: https://www.motorpasion.com/industria/hemos-alcanzado-peak-oil-petroleo-llega-a-su-punto-no-retorno-empieza-era-dorada-coches-electricos

Gairifo, R. M., & Dias, M. (2009). *Implementation of demand forecasting models for fuel oil.* Lisboa, Portugal: Universidade Técnica de Lisboa.

Garnier, R., & Belletoile, A. (2019). *A multi-series framework for demand forecasts in E-commerce.* Bordeaux, France: Université Paris Seine.

Gotham, D. (2009). *Methods for Forecasting Energy Supply and Demand.* Purdue University, Indiana, United States.: Institute of Public Utilities.

Hernández, Ó. (2011). *Introducción a las Series Cronológicas.* San José, Costa Rica: Editorial Universidad de Costa Rica.

Hyndman, RJ y Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27*, 1-22.

IEA. (2019). *Sankey Diagram.* Obtenido de International Energy Agency: https://www.iea.org/sankey/

Khedmatia, M., & Ghalebsaz-Jeddib, B. (2018). Three Approaches to Time Series Forecasting of Petroleum Demand in OECD Countries. *Journal of Optimization in Industrial Engineering, Vol.11, Issue 2, Summer and Autumn*, 17-24.

Kitamura, A. (2018). *Forecasting Japan's spot LNG prices using.* Tokio, Japan: Erasmus University Rotterdam.

Koyama, K., & Suehiro, S. (2020). *Demand for Oil, Natural Gas, and LNG Facing the Worst Global Economic Conditions since the Great Depression.* Tokio, Japan: The Institute of Energy Economics, Japan.

Lee, C.-Y., & Huh, S.-Y. (2017). Forecasting Long-Term Crude Oil Prices Using a Bayesian Model with Informative Priors. *Sustainability*, 1-15.

Li, L., & Ngan, C.-K. (2019). A Weight-adjusting Approach on an Ensemble of Classifiers for Time Series Forecasting. *Proceedings of the 2019 3rd Internacional Conference on Information System and Data Mining*, 65-69.

Messler, D. (10 de February de 2021). *How Much Higher Can Oil Prices Go?* Obtenido de Oil Price: https://oilprice.com/Energy/Energy-General/How-Much-Higher-Can-Oil-Prices-Go.html

Ming, W., Bao, Y., Hu, Z., & Xiong, T. (2014). Multistep-Ahead Air Passengers Traffic Prediction with Hybrid ARIMA-SVMs Models. *The Scientific World Journal, Volume 2014*, 1-14.

O`Ryan, R. (2008). *Diseño de un Modelo de Proyección de Demanda.* Santiago, Chile: Universidad de Chile .

Ogcu, G., Demirel, O., & Zaim, S. (2012). Forecasting Electricity Consumption with Neural Networks and Support Vector Regression. *Procedia - Social and Behavioral Sciences 58*, 1576-1585.

Pai, P.-F., & Lin, C.-S. (2005). AhybridARIMAand support vector machines model in stock price forecasting. *Sciense Direct Omega 33*, 497-505.

Pankratz, A. (1983). *Forecasting with univariate Box-Jenkins models.* New York. United States: John Wily & Sons. Inc.

Pavlyshenko, B. M. (2019). Machine-Learning Models for Sales Time. *Data 4,15*, 1-11.

Pérez, C., & Santin, D. (2007). *Minería de datos. Técnicas y herramientas.* Madrid, España: Paraninfo.

Programa Estado de la Nación. (2020). *Informe del Estado de la Nación 2020.* San José, Costa Rica: Consejo Nacional de Rectores.

Quanying et al. (2020). Crude oil price analysis and forecasting: A perspective of “new triangle". *Energy Economics, Elsevier, vol. 87*.

Rizvi, O. (9 de January de 2021). *Will Oil Demand Recover In 2021?* Obtenido de Oil Price: https://oilprice.com/Energy/Crude-Oil/Will-Oil-Demand-Recover-In-2021.html

Rosero, L. (2021). *Método para la estimación de la tasa R de Covid-19.* San José, Costa Rica: Universidad de Costa Rica.

Ruiz et al. (2014). Modelo estadístico que permite observar el impacto de los factores que inciden en el rendimiento de combustible. *Revista Electrónica Nova Scientia Nº 14 Vol. 7*, 236-253.

Rystad Energy. (2021). *Covid-19 Report May.* Houston, United States: Rystad Energy.

Sagar, R. (9 de January de 2021). *How Pandemic Affected The Time Series Models In Production: An Insider's Perspective*. Obtenido de Analytics India Mag: https://analyticsindiamag.com/time-series-model-production-pandemic-effects/

Scott, S., & Varian, H. (2014). Predicting the Present with Bayesian Structural Time Series. *Inderscience Publishers*, 4-23.

SEPSE. (2019). *Costa Rica Balance Energético Nacional 2019.* Obtenido de Secretaría de Planificación del Subsector de Energía: https://sepse.go.cr/wp-content/uploads/2020/12/diagrama-sankey-2019-g-191120.pdf

Sivaramakrishnan, G., & Suchithra, S. (2017). *A DETAILED STUDY ON MACHINE LEARNING TECHNIQUES FOR DATA MINING.* International Conference on Trends in Electronics and Informatics: IEEE.

Smith, G. (14 de December de 2020). *Crude Oil Prices Fall as OPEC Cuts Demand Forecasts and Lockdowns Tighten*. Obtenido de Investing: https://www.investing.com/news/commodities-news/crude-oil-prices-fall-as-opec-cuts-demand-forecasts-and-lockdowns-tighten-2368312

Soldo, B. (2012). *Forecasting natural gas consumption.* Zagreb, Croacia: Croatian electrical company group.

Suzuki et al. (2020). *Predicting the Industry Impacts of Covid-19 Using a Bayesian Structural Time Series Model.* Nomura, Japan: Nomura Research Institute.

Valverde, L. (20 de Octubre de 2020). *Déficit llegó en setiembre a un 6,75% del PIB*. Obtenido de CR hoy: https://www.crhoy.com/economia/deficit-llego-en-setiembre-a-un-675-del-pib/

Wadud et al. (2011). Modeling andforecastingnaturalgasdemandin Bangladesh. *Energy Policy 39*, 7372-7380.

Zhang, W., & Yang, J. (2015). Forecasting natural gas consumption in China by Bayesian Model Averaging. *Energy Reports*, 216-220.

Zhou, Y., Li, T., Shi, J., & Qian, Z. (2019). A CEEMDAN and XGBOOST-Based Approach to Forecast Crude Oil Prices. *Hindawi, Complexity*, 1-15.

8. ANEXOS

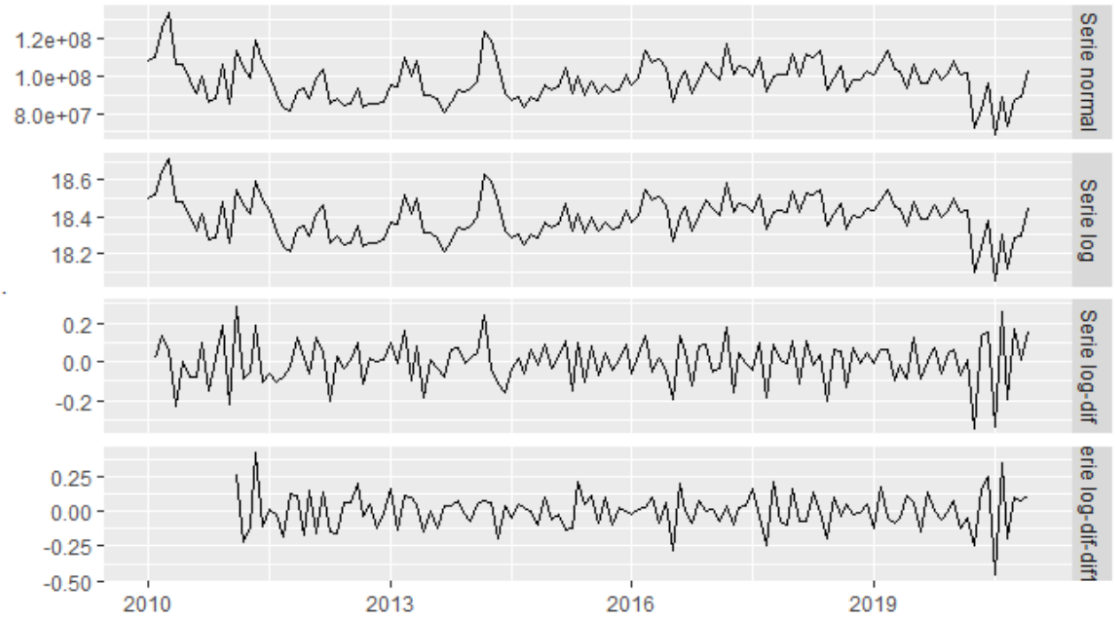
Anexo 1. Tabla con referencias de distintos tipos de modelos empleados en la estimación de la demanda energética.

Table 6
Overview of used forecasting tools.

Forecasting tools	*References*
Hubbert curve model	Hubbert [3], Hubbert [4], Al-Jarri and Startzman [5], Al-Fattah and Startzman [27], Siemek et al. [30], Cavallo [8], Imam et al. [37], Reynolds and Kolodziej [62], Maggio and Cacciola [6] and Valero and Valero [7]
Statistical models	Balestra and Nerlove [9], Beierlein et al. [76], Piggott [12], Herbert et al. [13], Herbert [14], Liu and Lin [16], Erdogdu [71], Brabec et al. [50], Lee and Singh [17], Sailor and Munoz [24], Aras and Aras [35], Gorucu and Gumrah [34], Huntington [46], Timmer and Lamb [48], Sanchez-Ubeda and Berzosa [47], Vondracek et al. [44], Brabec et al. [58], Brabec et al. [57], Yoo et al. [60], Behrouznia et al. [75], Azadeh et al. [66]
Artificial neural networks	Werbos [15], Brown et al. [18], Brown and Iftekhar [20], Suykens et al. [22], Khotanzad and Elragal [25], Khotanzad et al. [26], Gorucu et al. [32], Elragal [38], Khotanzad and Elragal [25], Khotanzad et al. [26], Viet and Mandziuk [40], Musilek et al. [43], Kizilaslan and Karlik [52], Kizilaslan and Karlik [59], Tonkovic et al. [63], Dombayci [73]
Grey prediction model	Xie and Li [64], Ma and Wu [61], Chen et al. [65], Ma and Li [67], Xu and Wang [69]
Conditional demand analysis	Bartels et al. [23], Aydinalp-Koksal and Ugursal [53]
Econometric model	Berndt and Watkins [11], Nagy [19], Gelo [78]
Mathematical model	Gil and Deferrari [36], Simunek and Pelikan [55]
Expert system	Smith et al. [21]
Stochastic Gompertz innovation diffusion model	Gutierrez et al. [39]
Dynamical system model	Li et al. [68]
Simulated annealing	Toksari [72]

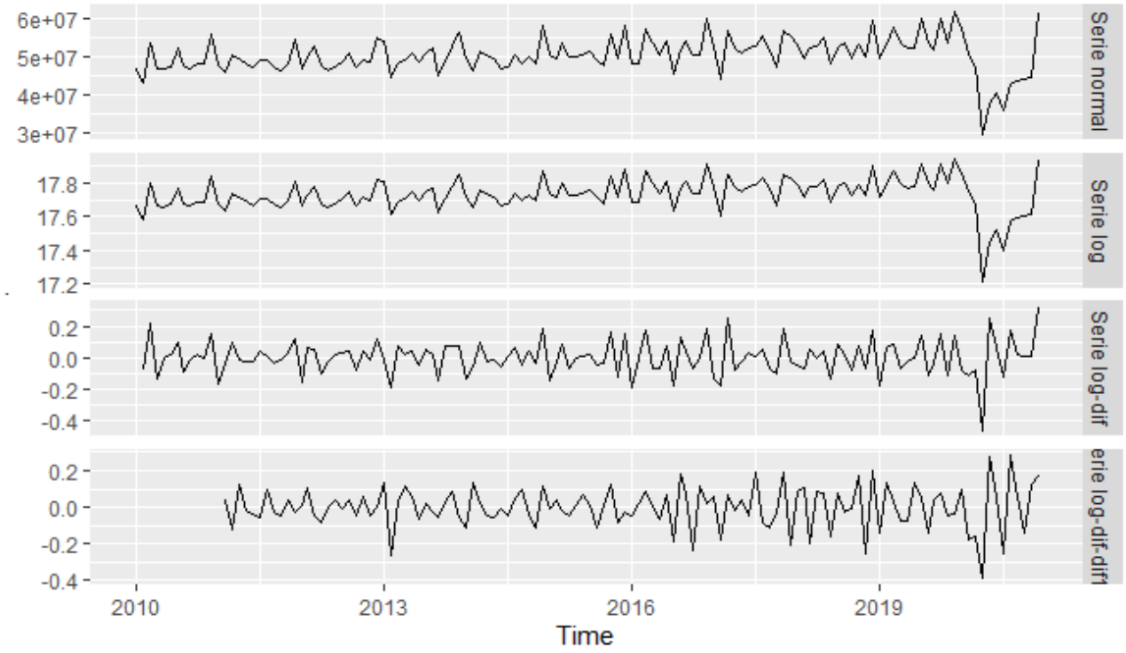
Fuente: (Soldo, 2012, pág. 11)

Anexo 2. Diésel, comportamiento de la serie de consumo en litros, en logaritmo, con una diferencia regular y con una diferencia estacional.



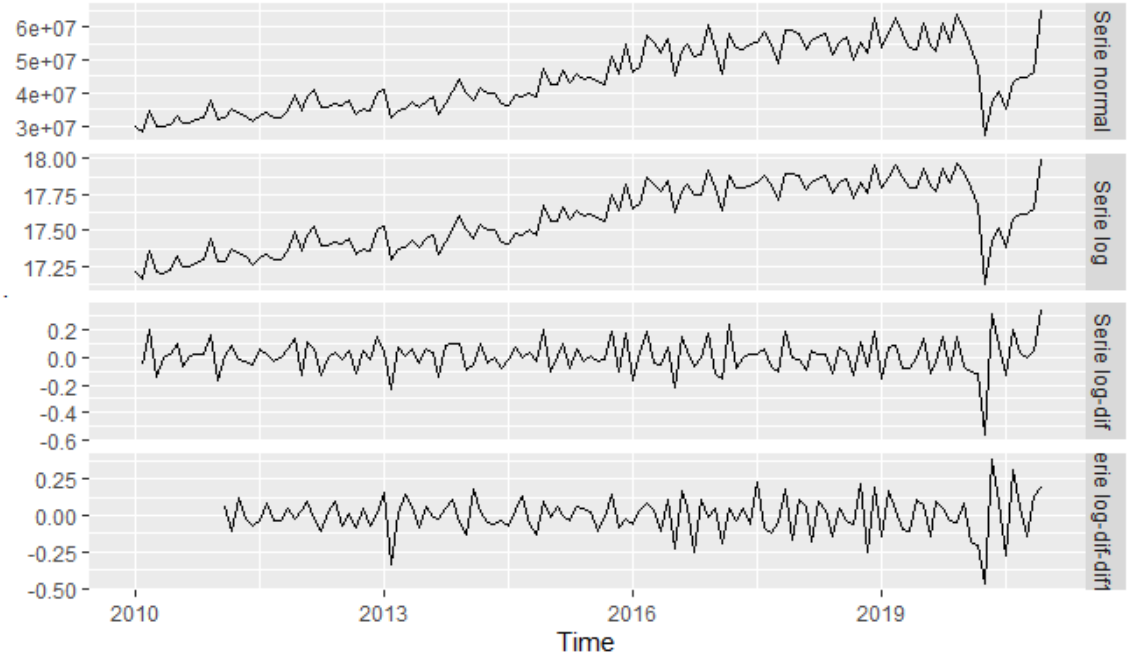
Fuente: Elaboración propia

Anexo 3. Regular, comportamiento de la serie de consumo en litros, en logaritmo, con una diferencia regular y con una diferencia estacional.



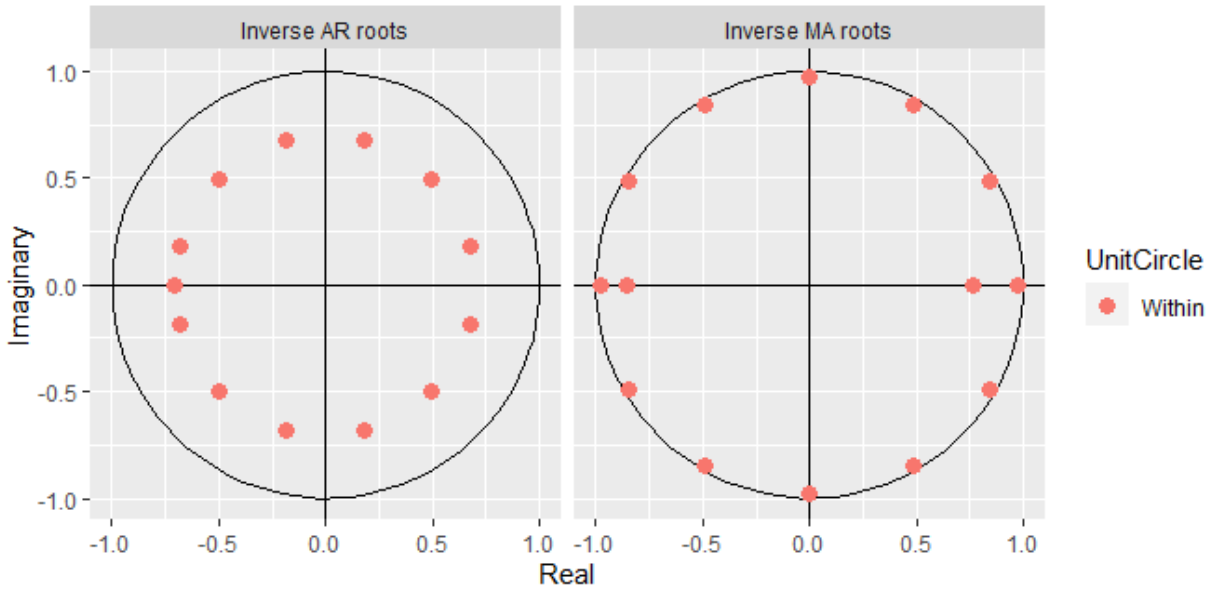
Fuente: Elaboración propia

Anexo 4. Súper, comportamiento de la serie de consumo en litros, en logaritmo, con una diferencia regular y con una diferencia estacional.



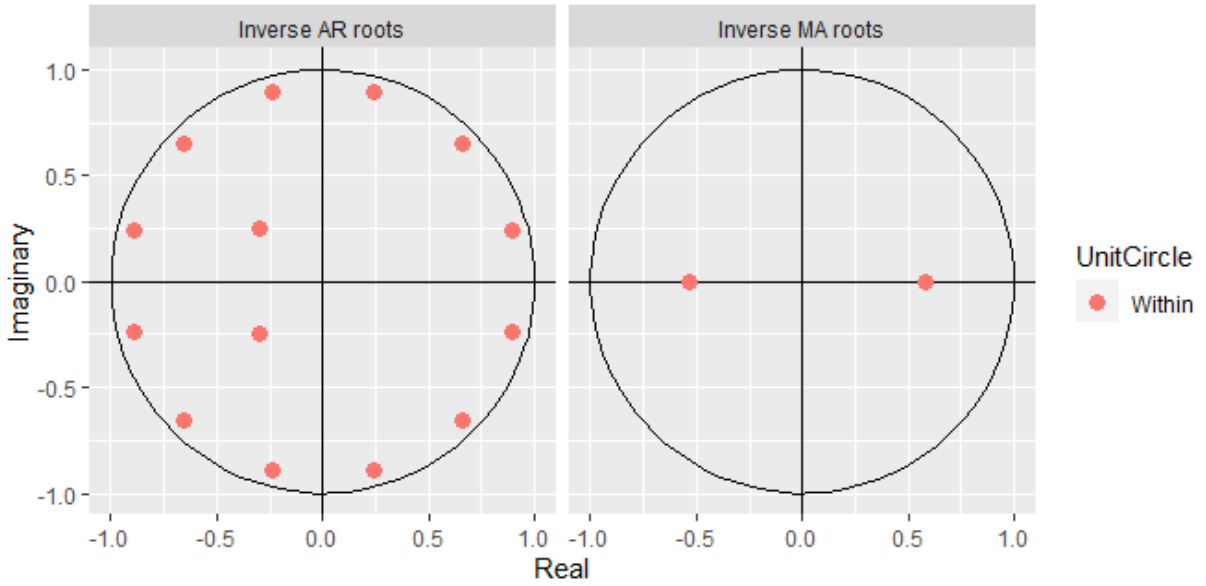
Fuente: Elaboración propia

Anexo 5. Diésel, circulo unitario de las raíces asociados a los modelos ARIMA, para el análisis de invertibilidad del modelo.



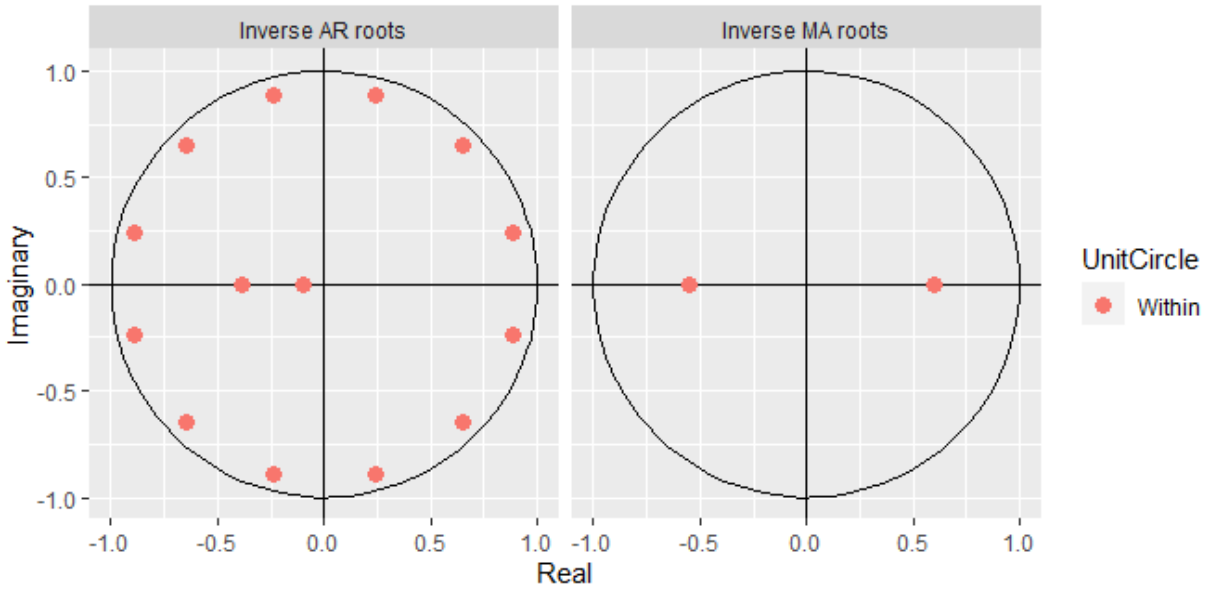
Fuente: Elaboración propia

Anexo 6. Regular, circulo unitario de las raíces asociados a los modelos ARIMA, para el análisis de invertibilidad del modelo.



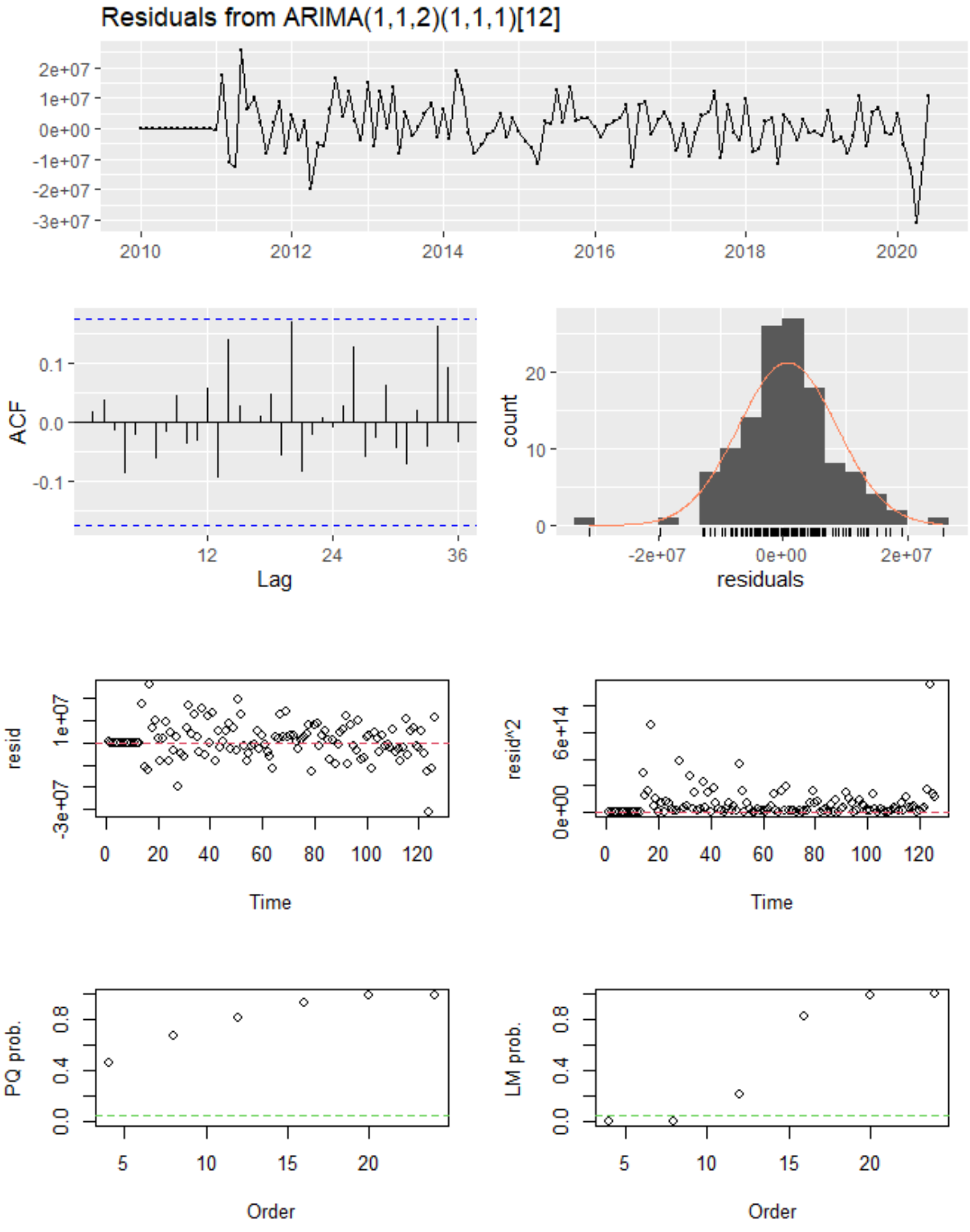
Fuente: Elaboración propia

Anexo 7. Súper, circulo unitario de las raíces asociados a los modelos ARIMA, para el análisis de invertibilidad del modelo.



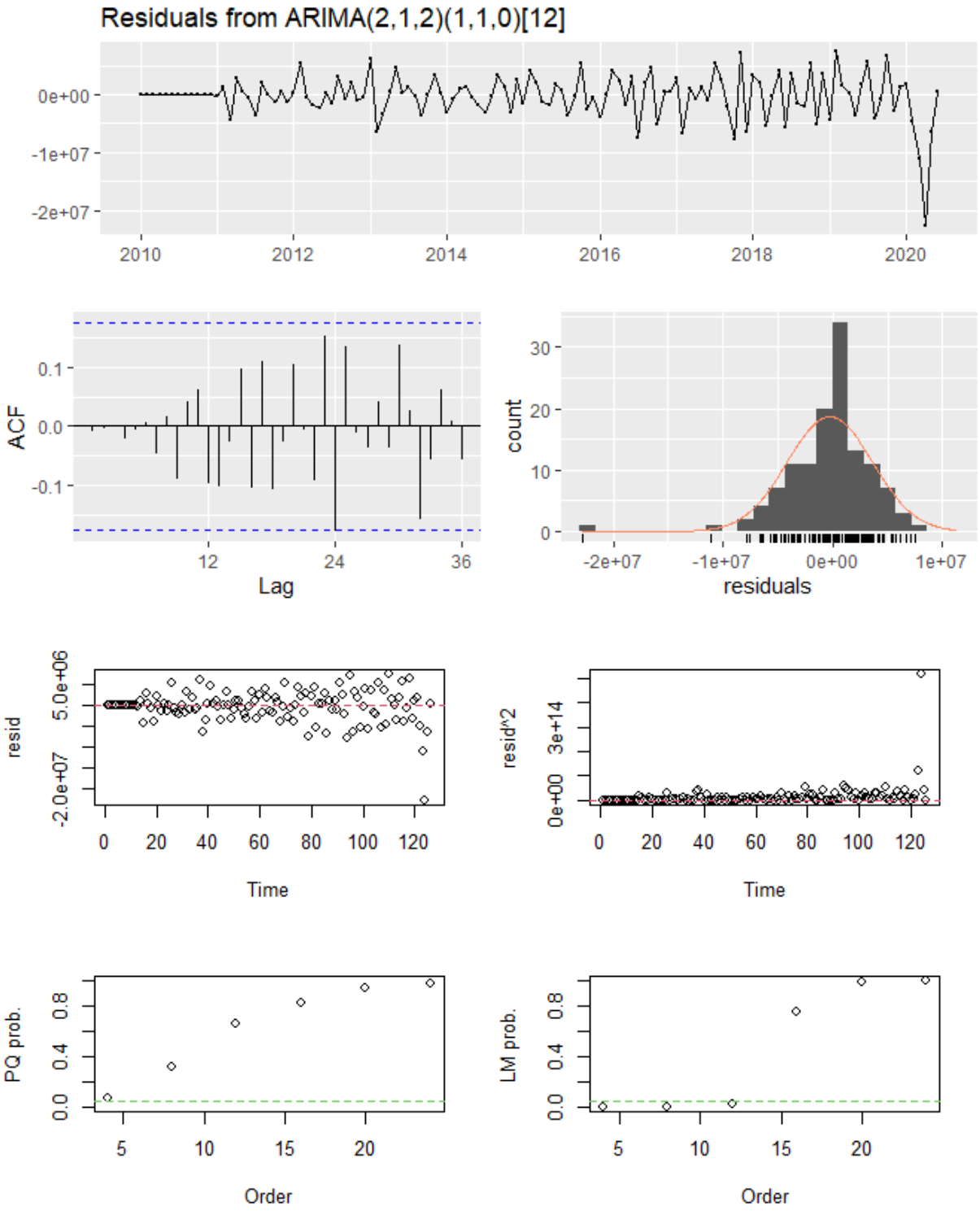
Fuente: Elaboración propia

Anexo 8. Diésel, comportamiento de los residuos para el análisis del cumplimiento de los supuestos de los modelos ARIMA.



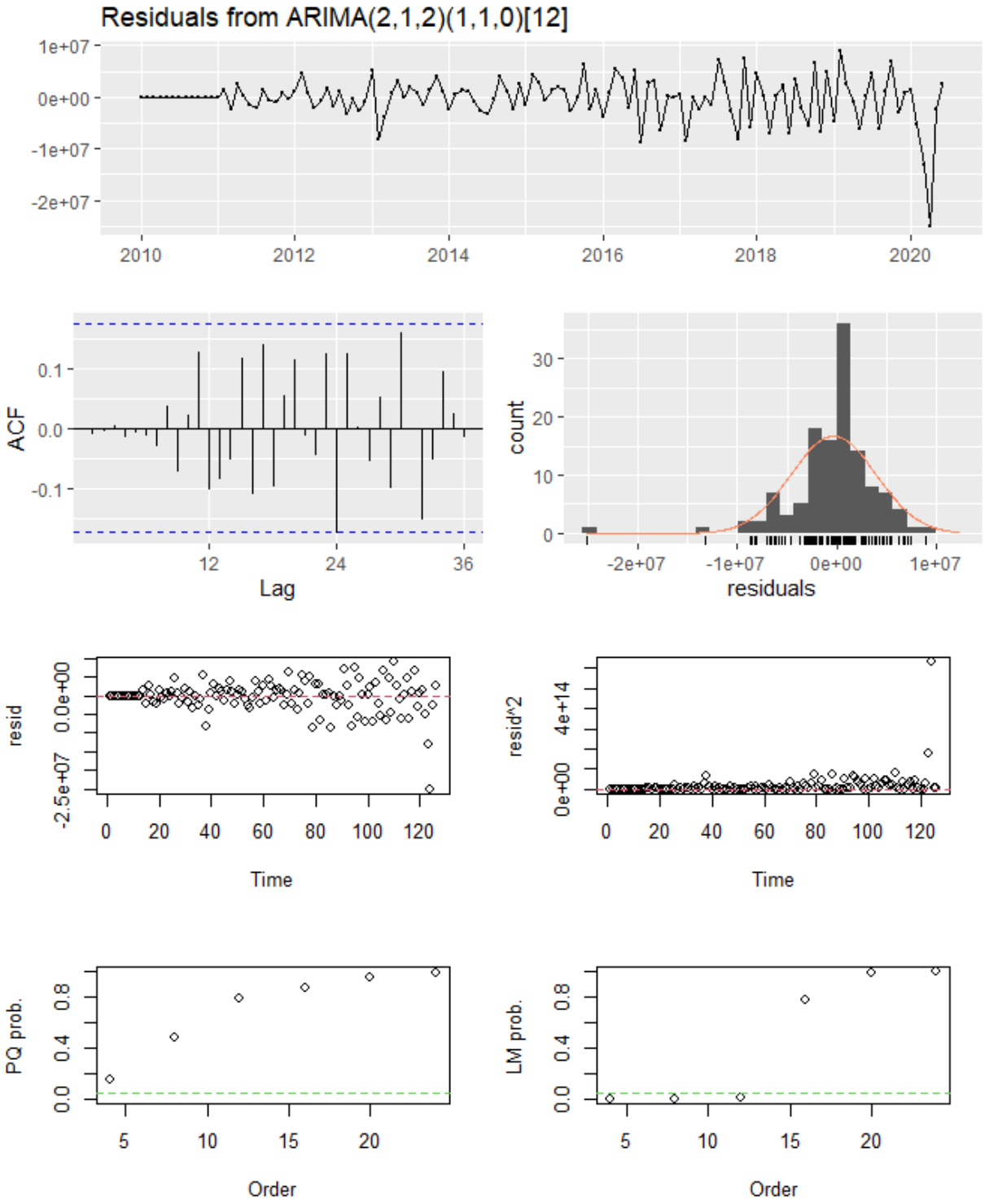
Fuente: Elaboración propia

Anexo 9. Regular, comportamiento de los residuos para el análisis del cumplimiento de los supuestos de los modelos ARIMA.



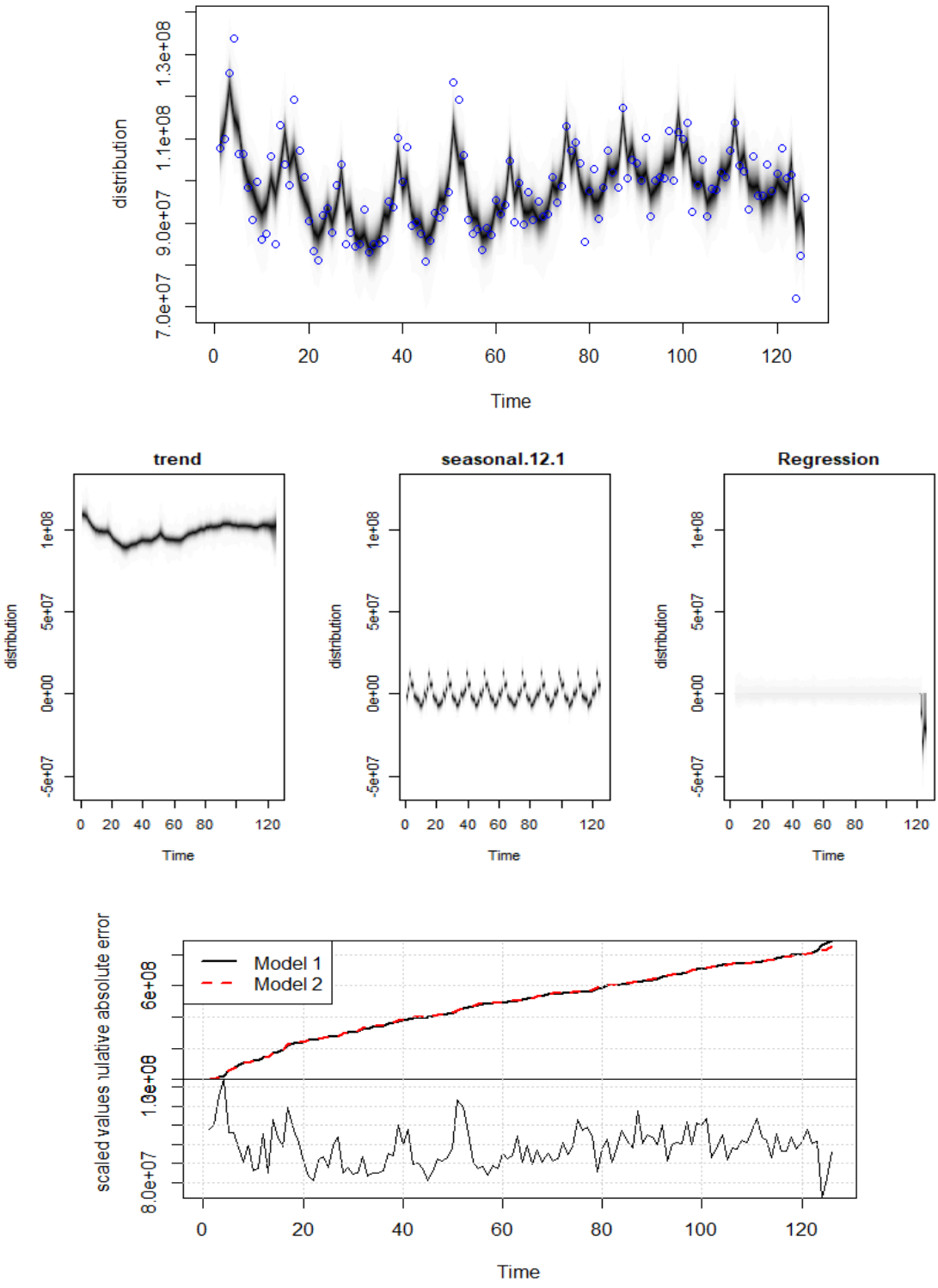
Fuente: Elaboración propia

Anexo 10. Súper, comportamiento de los residuos para el análisis del cumplimiento de los supuestos de los modelos ARIMA.



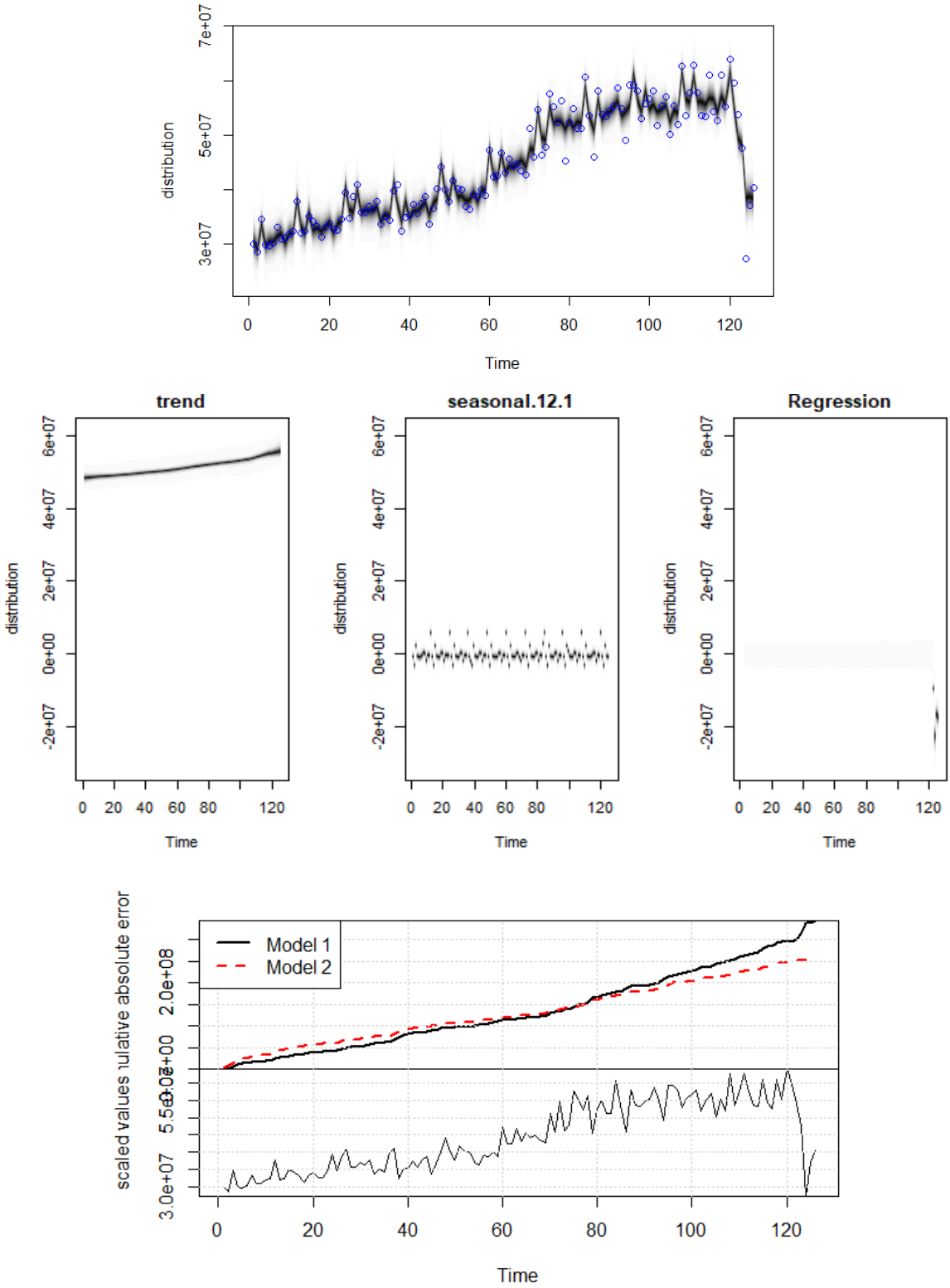
Fuente: Elaboración propia

Anexo 11. Diésel, resultados del proceso iterativo MCMC para la estimación del modelo BSTS con covariables.



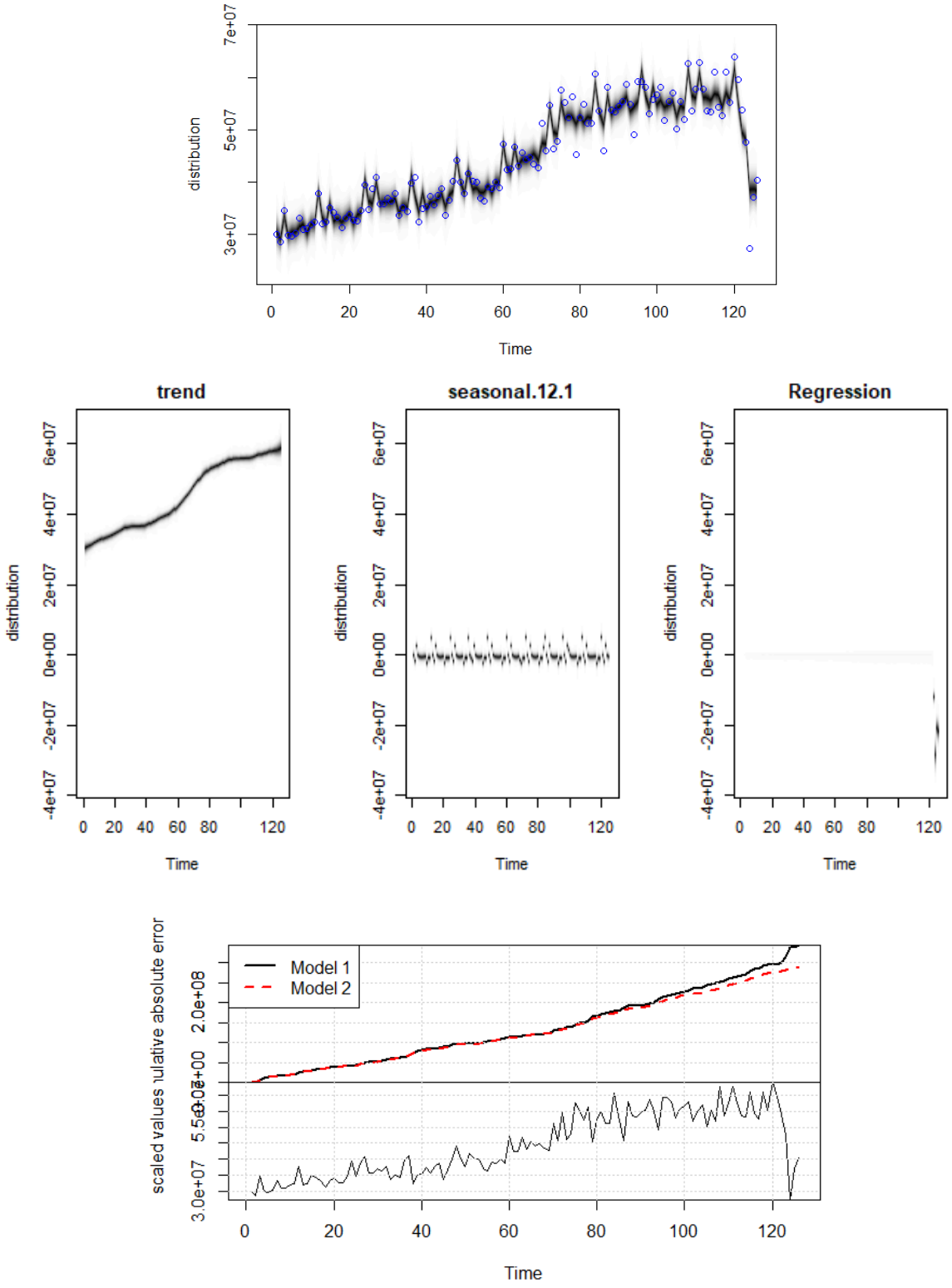
Fuente: Elaboración propia

Anexo 12. Regular, resultados del proceso iterativo MCMC para la estimación del modelo BSTS con covariables.



Fuente: Elaboración propia

Anexo 13. Súper, resultados del proceso iterativo MCMC para la estimación del modelo BSTS con covariables.



Fuente: Elaboración propia

Anexo 14. Diésel, comparación de las estimaciones desarrolladas por cada modelo para el periodo de julio a diciembre 2020.

Fuente: Elaboración propia

Anexo 15. Regular, comparación de las estimaciones desarrolladas por cada modelo para el periodo de julio a diciembre 2020.

Fuente: Elaboración propia

Anexo 16. Súper, comparación de las estimaciones desarrolladas por cada modelo para el periodo de julio a diciembre 2020.

Fuente: Elaboración propia