TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Análisis sobre la evolución del precio de la gasolina y el gasoil

Autor: Peilei Xu

Director: Francisco Javier Sierra Martínez

Departamento: Econometría, Estadística y Economía aplicada

Convocatoria: Septiembre del 2022

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resumen

En este trabajo, pretende analizar el impacto de los factores que provocan el aumento del precio de los carburantes. Concretamente, el estudio se centrará en la gasolina, el gasóleo y gasóleo para calefacción.

Para poder llevar a cabo este estudio, se ha recopilado los datos mediante la técnica de Web scraping y se ha modelizado cada serie con la metodología Box-Jenkins. Finalmente, se ha llevado a cabo una comparación de los valores predichos con los valores observados para medir el impacto del cambio estructural sufrido a causa de la pandemia y otros factores socio-económicos.

Palabras claves: Series temporales, modelització, ARIMA, carburantes.

The aim of this work is to analyze the impact of the factors that cause the increase in the price of fuels, specifically the study will focus on gasoline, diesel and heating oil.

The data that has been used to implement the study has been collected using the Web scraping technique and each series has been modelled with the Box-Jenkins methodology. Finally, a comparison of the predicted values with the observed values was carried out to measure the impact of structural change due to the pandemic and other socio-economic factors.

Key words: Time series, modelling, ARIMA, fuels.

Classificación AMS:

- 68N19 Other programming techniques (object-oriented, sequential, concurrent, automatic, etc.)
- 91B70 Stochastic models
- 37M10 Time series analysis
- 62M20 Prediction; filtering

Agradecimientos

Primero de todo, me gustaría agradecer a mi tutor Francisco Javier Sierra Martinez por la dedicación a este proyecto durante estos meses. Siempre me ha atendido muy rápidamente cuando lo he necesitado y me ha ayudado a resolver mis dudas a medida que he ido profundizando y avanzando en el proyecto.

En segundo lugar, me gustaría agradecer a toda mi familia y amigos que me han brindado apoyo durante este periodo, especialmente, quiero mencionar a Nora, gracias por toda la paciencia, la ayuda y por acompañarme en este proceso.

Finalmente, agradezco todo lo aprendido en estos cuatro años. Estoy orgullosa de ver hasta donde he podido llegar y tengo muchas ganas de seguir descubriendo y aprendiendo de la vida.

Índice general

Classificación AMS II

Agradecimientos III

Índice de cuadros VI

Listings VII

Índice de figuras IX

1. Introducción 1

2. Base de datos 6

2.1. Web Scrapping 6

2.2. Métodos de imputación 10

3. Análisis descriptivo 12

3.1. Análisis de los impuestos 12

3.2. Cambio estructural 14

3.3. Estadística descriptiva de las series 17

4. Metodología 21

4.1. Box-Jenkins 21

4.1.1. Modelo autorregresivo 24

4.1.2. Modelo media móvil 26

4.1.3. Etapas 28

5. Análisis ARIMA: Gasolina 34

5.1. Transformaciones previas 34

5.2. Identificación 38

5.3. Estimación 39

5.4. Validación 42

5.5. Predicción 45

6. Resumen de los análisis e interpretación **48**
6.1. Resultados de los análisis 48
6.2. Interpretación de las predicciones 49
6.2.1. Gasolina 49
6.2.2. Gasoil 50
6.2.3. Diésel para la calefacción 51

7. Conclusiones **53**

Bibliografia **57**

A. Análisis ARIMA: Gasoil **59**
A.1. Transformaciones previas 59
A.2. Identificación 61
A.3. Estimación 63
A.4. Validación 64
A.5. Predicción 66

B. Análisis ARIMA: Diésel para calefacción **67**
B.1. Transformaciones previas 67
B.1.1. Identidicación 69
B.1.2. Estimación 71
B.2. Validación 73

C. Código **75**

Índice de cuadros

3.1. Porcentaje impuestos 13
3.2. Resultados del contraste de Chow 16
3.3. Análisis descriptivo 17
3.4. Análisis descriptivo para el precio de la gasolina por año estudiado . . 18
3.5. Análisis descriptivo para el precio del diésel por año estudiado 18
3.6. Análisis descriptivo para el precio del diésel para la calefacción por año estudiado 19

4.1. Ejemplos para el proceso de estimación 29

5.1. Resultados de las pruebas de normalidad 42
5.2. Comparación de valores reales y predichos, la diferencia y la tasa de variación 47

6.1. Identificación de los modelos óptimos para los diferentes carburantes . 49
6.2. Comparación de valores reales y predichos, la diferencia y la tasa de variación 50
6.3. Comparación de valores reales y predichos, la diferencia y la tasa de variación 51
6.4. Comparación de valores reales y predichos, la diferencia y la tasa de variación 52

A.1. Resultados de las pruebas numéricas para diésel 65

B.1. Resultados de las pruebas numéricas para diésel de calefacción 73

Listings

2.1. Código aplicado en R para la obtención automática de los enlaces . . 8
2.2. Código aplicado en R para la lectura de los datos 9
2.3. Código aplicado en R para la imputación de datos faltantes 11
5.1. Código aplicado en R para la estimación 39
C.1. Código aplicado en R para el proyecto 75

Índice de figuras

1.1. Divisa euro vs dólar 2
1.2. Esquema del proceso estadístico 3

2.1. Datos originales 8
2.2. Datos semanales 9
2.3. Datos mensuales 10
2.4. Dato faltante 10
2.5. Dato faltante imputado 11

3.1. Precios de los carburantes con impuesto 12
3.2. Precios de los carburantes sin impuesto 13
3.3. Evolución de los impuestos 14
3.4. Tipos de cambios estructurales 15
3.5. Puntos máximos y mínimos de los precios 19

4.1. Ejemplos gráficos de series temporales no estacionarias y estacionarias 22
4.2. Gráficas del ACF de una serie no estacionaria y estacionaria 24
4.3. Etapas del proceso de Box-Jenkins 28

5.1. Gráfica ACF de la serie gasolina 34
5.2. Box-plot y gráfica de media-varianza 35
5.3. Monthplot de la serie gasolina 36
5.4. Media de la serie gasolina con una diferenciación 37
5.5. Gráfica ACF de la serie gasolina transformada 37
5.6. Grafica de ACF y PACF de la serie gasolina transformada 38
5.7. Salida de estimación del modelo $MA(4)$ para X_t 40
5.8. Tabla resumen con los principales resultados para la estimación de los modelos 40
5.9. Tabla resumen de los modelos de la serie gasolina 41
5.10. QQplot e Histograma de residuos 42
5.11. Gráfica de residuos y raíz cuadrada de residuos absoluto 43
5.12. Resultado de la prueba Breusch-Pagan 43

5.13. Gráficos para el ACF y el PACF de los residuos 44
5.14. Resultado de la prueba de Durbin-Watson 44
5.15. Gráfica de p-valores para la prueba de Ljung-Box 45
5.16. Comparación de valores reales y predichos 46

A.1. Gráfica ACF de la serie diésel 59
A.2. Box-plot y gráfica de Monthplot 60
A.3. Gráfica ACF de la serie gasoil transformada 61
A.4. Gráfica ACF y PACF de la serie gasoil transformada 61
A.5. Salida de estimación del modelo $MA(3)$ y $AR(2)$ para X_t 63
A.6. Salida de estimación del modelo $MA(3)$ y $AR(2)$ para W_t 64
A.7. Gráficos de validación de la serie de diésel 65
A.8. Comparación de valores reales y predichos diésel 66

B.1. Gráfica ACF de la serie diésel para la calefacción 67
B.2. Box-plot y Monthplot de la serie calefacción 68
B.3. Gráfica ACF de la serie calefacción transformada 69
B.4. Gráfica ACF y PACF de la serie calefacción transformada 69
B.5. Tabla resultante de las estimaciones con diésel.cal transformada . . . 71
B.6. Tabla resultante de la estimación con la serie diésel.cal 72
B.7. Gráficas de validación de la serie de calefacción 73

Capítulo 1

Introducción

Durante los últimos meses, el precio de los diferentes carburantes han incrementado de manera brusca y, cada mes, se observa cómo estos van superando el récord histórico de España.

El aumento de los precios es debido a diferentes factores socio-económicos que está viviendo el país y el mundo actualmente. En concreto, se puede definir tres causas principales que han provocado el aumento desproporcionado de los precios (*¿Por qué sube el precio de la gasolina y el diésel? — Noticias Coches.net*, [s.f.]).

En primer lugar, los grandes cambios en el ámbito político. Las diferentes decisiones políticas tomadas a causa de los conflictos internacionales que se han ido produciendo han provocado cambios en la evolución del precio en el sector energético. Por ejemplo, en los diferentes medios de comunicación han ido relatando el conflicto que se está viviendo entre Ucrania y Rusia y cómo este ha afectado en los diferentes ámbitos mundialmente.

A causa de este gran problema, en febrero del año 2022, la Unión Europea impuso una serie de sanciones contra Rusia, entre otras, una sanción económica y restricciones con la compraventa de productos. En otras palabras, la Unión Europea impuso una serie de restricciones a la importación y la exportación de diversas materias primas con Rusia. Teniendo en cuenta que Rusia es el tercer país productor de petróleo, restringir la importación y exportación ha supuesto la disminución de oferta en el mercado, incrementando así, el precio (*Medidas restrictivas de la UE contra Rusia por sus actos en Ucrania (desde 2014) - Consilium*, [s.f.]).

En segundo lugar, estudiando la evolución de los precios de los carburantes y haciendo una investigación de los factores con alta probabilidad en desencadenar un cambio en la evolución de estos, se puede determinar que la subida no se inició en

el 2022, sino mucho antes. Esto es debido el aumento del precio del barril *Brent*.

El precio del petróleo aumentó nada más empezar la COVID-19. En concreto, inició en abril del 2020, ya que se paró a los trabajadores debido a las cuarentenas y otras medidas tomadas para evitar la propagación del coronavirus. Por ese motivo, había menos oferta de carburante en el mercado. Esto no se vio reflejado en los precios de manera inmediata, por el hecho de que la población no usaba en gran cantidad los combustibles. Pero, a partir del junio del 2021 aproximadamente, se empezaron a aflojar las restricciones y con ello, creció la demanda para usar el vehículo. Ante esta alta necesidad, las organizaciones de los diferentes países exportadores de petróleo no quisieron aumentar la producción para mantener el precio alto o incluso, incrementarlo.

En tercer lugar, el último factor principal de los grandes cambios sufridos en los últimos meses en cuanto al precio de los combustibles es el precio de las divisas. Los cambios que sufren las diferentes monedas afectan directamente al precio del producto. Teniendo en cuenta que la compra del petróleo se realiza mediante la moneda del dólar, a continuación, se muestra una gráfica con la evolución de la divisa de euro contra el dólar durante el último año.

Figura 1.1: Divisa euro vs dólar

Fuente: Euros a Dólares estadounidenses — Convierta 1·EUR a USD — Xe, *[s.f.]*

En la figura 1.1 se puede observar que, durante el último año, el euro se devalúa continuamente frente al dólar, y, los últimos datos indican que, un euro tiene menor valor que un dólar, cosa que nunca se había observado desde que se empezó a cotizar la moneda del euro. La devaluación del euro ha provocado que se pague más en caso de que el precio se mantenga y mucho más en este contexto.

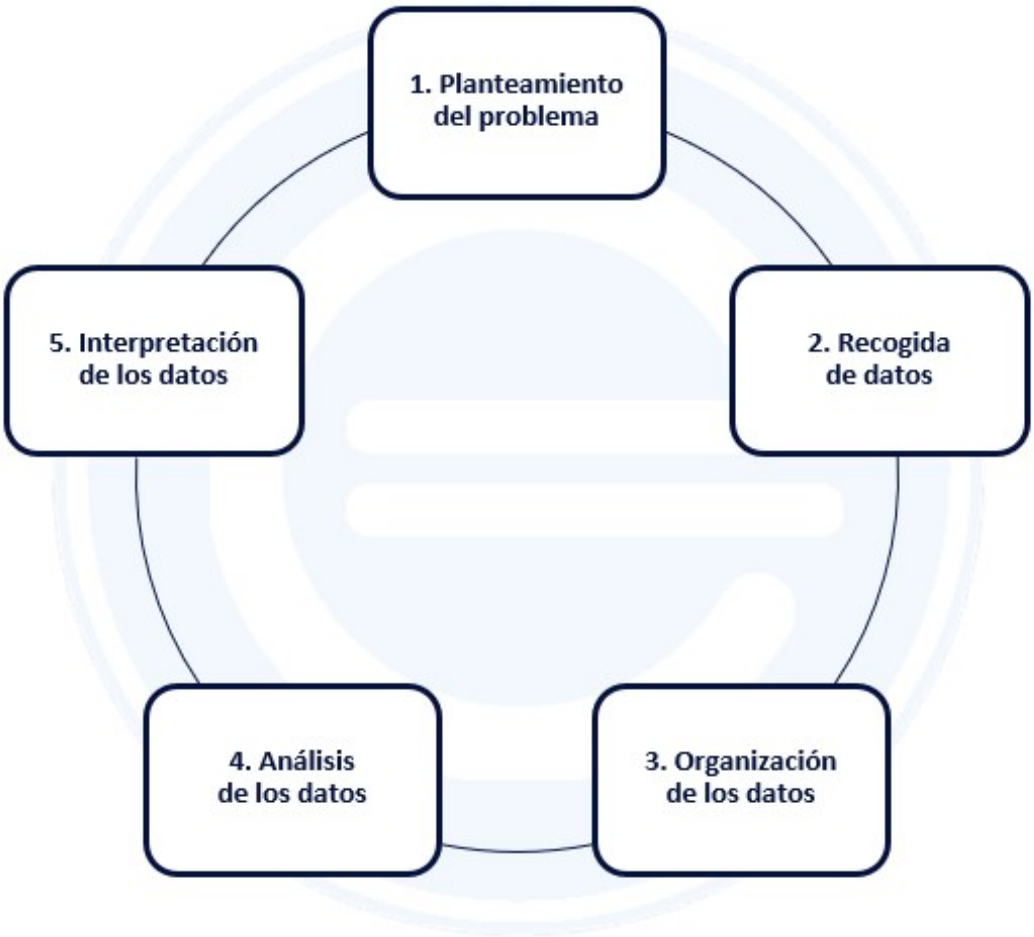
Teniendo en cuenta todos estos factores que se han comentado anteriormente, al ser uno de los temas actuales y con alta atención tanto de los medios como de la población en general, el objetivo principal de este estudio es analizar el impacto económico que ha supuesto a la ciudadanía los cambios tan extremos y a corto plazo que se han observado en el precio de los carburantes. Este proyecto quiere analizar e informar de cuáles han sido realmente los cambios vividos en cuanto al precio

del carburante y cómo ha afectado, por ejemplo, al llenar el depósito de un coche. De este modo, se podrá contrastar analíticamente si la información que se ha ido proporcionando en los últimos meses en los telediarios es real o, por lo contrario, el cambio es menor a lo notificado.

Para llevar a cabo el estudio, se realizará un análisis de series temporales sobre los distintos tipos de carburantes, principalmente, la gasolina súper 95 y el gasóleo porque son los carburantes más usados al día a día para propulsar el vehículo.

En esta memoria se describe el proceso completo del estudio y las herramientas usadas para la investigación divididas según el proceso estadístico como se representa en la figura 1.2.

Figura 1.2: Esquema del proceso estadístico



Fuente: Proceso estadístico - Qué es, definición y concepto — 2022 — Economipedia, *[s.f.]*

En primer lugar, se debe plantear el problema, sus causas y motivaciones para llevar a cabo el análisis y cuáles son las hipótesis iniciales. Esta etapa del proceso estadístico correspondería a la introducción de este proyecto. Para las siguientes etapas del proceso se podrían describir como:

- **Base de datos:** en este apartado se realizará una recopilación del precio histórico de carburante gasolina 95, gasóleo y gasóleo para la calefacción mediante la mecánica *Web Scraping*. Una vez obtenidos los precios, se realizará una detección e imputación de los datos faltantes. En esta sección se haría referencia a la segunda etapa del proceso estadístico, la recogida de los datos.

- **Análisis descriptivo:** para entrar en la tercera etapa, se debe organizar los datos. Por ello, una vez conseguidas las series completas, se presentará un análisis descriptivo global y anual sobre cada una de las series.

- **Metodología:** antes de iniciar con la siguiente etapa del proceso estadístico, se debe presentar la parte teórica que respalda los cálculos e interpretaciones de los análisis de datos. Por ello, en este apartado se presentará la metodología usada para llevar a cabo el análisis de series temporales de manera teórica, es decir, se define el esquema de la metodología Box-Jenkins.

- **Análisis ARIMA (Gasolina):** en esta etapa se llevará a cabo la explicación de los análisis llevados a cabo para determinar el modelo que mejor se adapta al carburante que se está estudiando y de esta manera realizar las predicciones de los precios. Finalmente, en cada uno de los análisis se puede consultar cada una de las etapas de la metodología Box-Jenkins aplicada al caso sujeto a estudio y la comparativa de los valores predichos (como hubiera evolucionado el precio del carburante sin los cambios estructurales sufridos) y los valores reales observados. El análisis de este carburante se mostrará paso a paso el proceso a seguir para los diferentes combustibles.

- **Resumen de los análisis e interpretación:** teniendo en cuenta el proceso llevado a cabo para la gasolina, en este capítulo se resumirán las conclusiones a las cuales se ha llegado con el análisis de cada uno de los carburantes y se especificaran diferentes interpretaciones a los cambios bruscos de la evolución del precio.

- **Conclusiones:** en este apartado se recoge, para acabar la investigación, las conclusiones del estudio explicando si realmente el incremento de precio vivido en este último periodo es más grande de lo que ya se esperaba estadísticamente. Además, se hará una valoración de la importancia del proyecto y el trabajo llevado a cabo.

Para poder realizar una correcta interpretación de los diferentes resultados que se irán planteando en las diferentes secciones del proyecto, es necesario definir qué se está considerando como precio de los carburantes de comercio.

Para poder empezar, hay que tener en cuenta que el petróleo se cotiza internacionalmente. Por este motivo, la compraventa se realiza mediante dólares.

Por otra parte, el carburante se comercializa como cualquier otro producto. A la hora de realizar una compra de carburantes, se está pagando un precio base y también unos impuestos. El precio base está formalizado por la cotización internacional de carburantes, y este factor presenta alrededor de un 42 y un 50 % del precio de comercio final. Además de la cotización, también se incluye en el precio base el coste de distribución, el beneficio mayorista y el beneficio minorista, que ocupa entre un 11 y un 14 % del precio de venda. Sobre los impuestos, aparte del IVA, los carburantes tienen otro tipo de impuesto, el impuesto especial de hidrocarburos (IEH) y este tipo de impuesto cambia según los distintos tipos de carburantes. El IVA y la IEH suman entre un 39 y un 44 % del precio en el surtidor (*Todo lo que debes saber sobre el precio de los carburantes · AOP*, [s.f.]).

Capítulo 2

Base de datos

Los datos que se han utilizado para poder llevar a cabo este estudio se han conseguido a través de la técnica *web scraping*.

2.1. Web Scrapping

Uno de los problemas que se presentan al inicio de un proyecto analítico es la obtención de los datos. Esto es debido a que los datos publicados en las diferentes páginas web solo proporcionan su lectura mediante los diferentes navegadores web. En otras palabras, no hay opción de guardar o descargar estos datos. Por este motivo, y para dar una solución eficaz en la primera fase del proyecto, se ha usado la técnica de *web scraping* y así, dar inicio al análisis.

Históricamente, para llevar a cabo las diversas búsquedas, se desarrollaron los *web crawlers*, es decir, robots que rastrean todas las páginas web para proporcionar a los usuarios la información que necesitan y dirigirlos a la web adecuada. A continuación, se llevó a cabo la investigación y el desarrollo del *web scraping* con el objetivo de extraer los datos, tanto estructurados como no estructurados, que los diferentes usuarios han querido consultar. Esta necesidad de creación de una nueva técnica para llevar a cabo la extracción o lectura automática de la información consultada en internet, nace a raíz de la optimización del proceso de tratamiento de grandes cantidades de información en un corto tiempo y sin usar una gran cantidad de memoria (*Servicios De Web Scraping: Cómo Comenzó y Qué Sucederá en El Futuro — Octoparse*, [s.f.]).

Existen diferentes modos de *scraping* y en general, se diferencian entre dos tipos, *scraping manual* y *scraping automático*

El primer tipo de *scraping* es el más usado entre los usuarios de las páginas web. El uso de las funciones implementadas en el teclado para llevar a cabo las operaciones de copiar y pegar información de una página web se le llama *scraping manual*. Esta técnica se suele aplicar solo cuando se requiere informaciones muy concretas debido a su ineficiencia capturando y separando los datos.

Por otro lado, el segundo tipo es el *scraping automático* que se usa al trabajar con grandes cantidades de información. En este proyecto, al estar tratando con diferentes datos para cada uno de los combustibles a estudiar, se requiere trabajar con una gran cantidad de información numérica, y las funciones para copiar y pegar no son las opciones óptimas. En cambio, el *scraping automático* ayuda a capturar la información de forma rápida y ordenada. Gran parte de la información con la que se trabajará en el análisis son datos no estructurados en formato *HTML* que con ayuda del *web scraping* se ha podido transformar en datos estructurados y almacenarlos para utilizarlos en el proceso de análisis.

Para aplicar la técnica automática de captación de información, es necesario el uso de un algoritmo o software. Existen distintos métodos para hacerlo, el primero y el más común de ellos, es mediante el uso de *bots*. Estos *bots* están programados para aplicar diferentes tareas definidas previamente por un programador o informático de manera automática.

En el desarrollo de este proyecto, para extraer información de una página web se ha aplicado la alternativa automática mediante un *parser* o analizador sintáctico. De este modo, se ha cambiado la estructura de una pieza de texto y guardado la información. Al tener esta nueva estructura, ha sido necesario la aplicación de un análisis de textos, aunque este método presente una alta complejidad y suela usarse solo por programadores avanzados usando la función *grep* de *Unix* (Bambenek y col., 2009).

En concreto, para llevar a cabo este estudio se ha creado mediante el *software R*, un código para que se lea los datos automáticamente de la página web.

En la figura 2.1, se presenta de la página original de donde se han extraído los datos para el análisis. En esta, se presentan los precios de la gasolina, el diésel y el diésel para la calefacción, con y sin el impuesto. Además, la tabla con los diferentes valores se va actualizando cada semana con el precio medio de todas las gasolineras de España.

Figura 2.1: Datos originales

https://datosmacro.expansion.com/energia/precios-gasolina-diesel-calefaccion/espana?anio

datosmacro.com Datos Países Análisis Diccionario € $ Search

Precios de los derivados del petróleo: España

¿Quieres que lo enviémos a tu correo?

Precios al consumidor de productos petrolíferos España

Fecha	Super 95	Super 95 (Sin imp.)	Diesel	Diesel (Sin imp.)	Diesel Cal.	Diesel Cal. (Sin imp.)
30/05/2022	1,968 €	1,154 €	1,852 €	1,152 €	1,369 €	1,035 €
23/05/2022	1,940 €	1,131 €	1,867 €	1,164 €	1,336 €	1,007 €
16/05/2022	1,898 €	1,096 €	1,887 €	1,181 €	1,355 €	1,023 €
09/05/2022	1,878 €	1,080 €	1,911 €	1,201 €	1,393 €	1,054 €
02/05/2022	1,837 €	1,046 €	1,873 €	1,169 €	1,352 €	1,021 €
25/04/2022	1,817 €	1,029 €	1,847 €	1,148 €	1,375 €	1,039 €
11/04/2022	1,591 €	0,842 €	1,614 €	0,955 €	1,314 €	0,990 €
04/04/2022	1,613 €	0,860 €	1,647 €	0,982 €	1,315 €	0,990 €
28/03/2022	1,818 €	1,030 €	1,837 €	1,139 €	1,391 €	1,053 €
21/03/2022	1,813 €	1,026 €	1,798 €	1,107 €	1,339 €	1,010 €
14/03/2022	1,845 €	1,052 €	1,817 €	1,123 €	1,393 €	1,055 €
07/03/2022	1,680 €	0,916 €	1,581 €	0,928 €	1,175 €	0,874 €

Fuente: Precios de los derivados del petróleo: España 2022 — Datosmacro.com, *[s.f.]*

La web *datasmacro* almacena los precios de cada año en una ventana por separado. Para poder llevar a cabo este estudio, es necesario obtener todos los datos desde el año 2010 hasta el actual, por lo tanto, será necesario leer 13 ventanas diferentes. Este proceso se realiza fácilmente creando un contador en R empezando por 2010 a 2022, y mediante la función *paste0* (Becker, 2018), se genera los trece enlaces que corresponden a cada uno de los años sujetos a estudio.

```
1 library(rvest)
2 contador <- c(2010:2022)
3 paginas <- paste0("https://datosmacro.expansion.com/energia/precios
    -gasolina-diesel-calefaccion/espana?anio=",contador)
```

Listing 2.1: Código aplicado en R para la obtención automática de los enlaces

Como se puede ver en el código presentado, para llevar a cabo el *scraping* de los datos, se ha utilizado el paquete *rvest* creado por Hadley Wickham y sirve para ayudar a raspar los datos de una página web, es decir, documentos con formato *HTML*.

```
code <- read_html(url)
tablas <- html_table(code)
```

Listing 2.2: Código aplicado en R para la lectura de los datos

Principalmente, se han usado las funciones presentadas en el código anterior 2.2 dónde la primera se usa para leer las páginas y la segunda función es muy útil para raspar la información de la tabla en formato *HTML*, ya que se los guarda directamente como un *data.frame* dentro del programa.

Aunque ha leído con éxito los datos, no siempre tiene un formato correcto para poder manipularlos en el análisis. Por ejemplo, en esta fase del proyecto, se pudo detectar en el programa como no se reconocían los datos leídos en formato numérico debido a que todos llevan el símbolo del euro. Para convertirlos en datos numéricos, se ha aplicado la función *substr* para guardar los cinco primeros caracteres de los datos. En otras palabras, se ha excluido el símbolo de euro, y después, se ha convertido en datos numéricos mediante la función *as.numeric*. En definitiva, se trata de realizar una depuración de los datos.

Aplicadas ya las diferentes funciones y cambios para trabajar con una base de datos limpia y avanzar con el análisis, se puede observar los datos como en la imagen 2.2.

Figura 2.2: Datos semanales

Description: df [604 x 9]

Fecha <chr>	Super 95 <dbl>	Super 95 (Sin imp.) <dbl>	Diesel <dbl>	Diesel (Sin imp.) <dbl>	Diesel Cal. <dbl>
25/01/2010	1.100	0.512	0.996	0.518	0.629
18/01/2010	1.104	0.515	1.001	0.523	0.633
11/01/2010	1.109	0.519	1.009	0.529	0.637
04/01/2010	1.090	0.503	0.988	0.511	0.608
22/02/2010	1.108	0.519	1.005	0.526	0.621
15/02/2010	1.099	0.511	0.988	0.511	0.614
08/02/2010	1.105	0.516	0.994	0.517	0.619
01/02/2010	1.095	0.508	0.987	0.511	0.620
29/03/2010	1.169	0.571	1.053	0.567	0.668
22/03/2010	1.159	0.562	1.045	0.560	0.664

1-10 of 604 rows | 1-6 of 9 columns Previous 1 2 3 4 5 6 ... 61 Next

Fuente: Elaboración propia

Analizando la figura anterior, se puede observar que los datos recopilados son datos semanales, por este motivo, se ha considerado necesario añadir una variable llamada *mes* para mostrar el número del mes que corresponde a cada dato. Una vez generada esta nueva variable, se calcula la media de todos los datos que pertenecen al mismo mes mediante la función *aggregate*.

Después de aplicar todas las funciones mencionadas, se ha conseguido los datos mensuales con los que se trabajará.

Figura 2.3: Datos mensuales

Group.1 <chr>	Super 95 <dbl>	Super 95 (Sin imp.) <dbl>	Diesel <dbl>	Diesel (Sin imp.) <dbl>	Diesel Cal. <dbl>
01/2010	1.100750	0.5122500	0.998500	0.5202500	0.6267500
02/2010	1.101750	0.5135000	0.993500	0.5162500	0.6185000
03/2010	1.147800	0.5528000	1.033800	0.5504000	0.6564000
04/2010	1.179667	0.5803333	1.082000	0.5926667	0.7000000
05/2010	1.174600	0.5760000	1.087000	0.5968000	0.7050000
06/2010	1.168500	0.5707500	1.090250	0.5997500	0.7050000
07/2010	1.169000	0.5512500	1.081000	0.5725000	0.7070000
08/2010	1.164000	0.5456000	1.086800	0.5764000	0.6992000
09/2010	1.166500	0.5480000	1.092500	0.5812500	0.6990000
10/2010	1.168750	0.5497500	1.096500	0.5845000	0.7060000

Fuente: Elaboración propia

La base definitiva presenta seis series mensuales, que son, el precio de gasolina, el precio del diésel y el precio del diésel para la calefacción, con y sin impuesto.

Observando la figura 2.3, al agrupar los precios semanales, se han podido ver 149 observaciones mensuales, que corresponden al intervalo de enero del 2010 a junio del 2022, donde las observaciones de los primeros 11 años se utilizan para crear el modelo y los datos recogidos en 2021 y 2022, se utilizaran para testear con los valores pronosticados con el modelo.

Después de obtener las series, se ha detectado que no se presenta ningún valor correspondiente al mes de julio del 2020 (figura 2.4).

Figura 2.4: Dato faltante

	Fecha <chr>	Super 95 <chr>	Super 95 (Sin imp.) <chr>	Diesel <chr>	Diesel (Sin imp.) <chr>	Diesel Cal. <chr>	Diesel Cal. (Sin imp.) <chr>	mes <dbl>	any <chr>
150	07/2020	NA	NA	NA	NA	NA	NA	7	2020

1 row

Fuente: Elaboración propia

Los valores faltantes pueden provocar problemas a la hora de hacer análisis de series temporales. Hay ocasiones que se puede destacar y únicamente hacer el análisis con los datos disponibles, pero en este proyecto se ha decidido realizar la imputación de datos, que se trata de asignar a los valores faltantes otros valores estimados.

2.2. Métodos de imputación

Para afrontar el problema detectado dentro de la depuración de los datos hay que llevar a cabo la estimación de los valores faltantes mediante métodos de imputación. Existen distintos métodos de imputación, entre otros, hay métodos de imputación simple, métodos de imputación basados en la máxima verosimilitud y basados en el *machine learning* (*RPubs - Imputación de datos*, [s.f.]).

Los métodos de imputación simple se basan en la estimación de un solo valor mediante la asignación del valor estimado a los valores faltantes. Por ejemplo, usando la media, estimando una regresión, usando la interpolación, etc. Pero, estos modelos pueden provocar una subestimación de los errores estándar.

El método basado en máxima verosimilitud se define en la implementación de un modelo identificando los coeficientes que maximicen el ajuste a los datos observados y se hace una estimación a los valores faltantes. Este método es muy eficiente y no provoca subestimación de los errores estándar pero, el proceso de implementación supone mayor coste computacional y complicación al anterior.

Por último, otra de las alternativas para la imputación de los datos es el método de *machine learning*. El método conocido como KNN(*K-Nearest-Neighbor*) trata de buscar las observaciones que tienen características muy similares y asignar a los valores faltantes la media de los vecinos seleccionados (Peterson, 2009).

Teniendo en cuenta que en este caso se trata de un análisis de series univariantes, es decir, solo tenemos una variable a estudiar, no se podrá realizar los métodos de máxima verosimilitud ni de *machine learning*. Por lo tanto, se ha decidido llevar a cabo la imputación con el método de interpolación lineal basado en la estimación del valor de una función entre dos valores conocidos.

```
library(imputeTS)
for(j in 2:ncol(x)) {
  x[[j]] <- as.numeric(x[[j]])
  x[[j]] <- ts(x[[j]],start = 2010,frequency = 12)
  x[[j]] <- na_interpolation(x[[j]], option='linear')
}
```

Listing 2.3: Código aplicado en R para la imputación de datos faltantes

En el código 2.3 se presenta el bucle creado para la implementación de la estimación de valores faltantes detectados de manera automática en los datos usados.

Figura 2.5: Dato faltante imputado

	Fecha	Super 95	Super 95 (Sin imp.)	Diesel	Diesel (Sin imp.)	Diesel Cal.	Diesel Cal. (Sin imp.)	mes	any
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
150	07/2020	1.13	0.46125	1.03275	0.4745	0.5215	0.3345	7	2020

1 row

Fuente: Elaboración propia

Como se aprecia en la figura 2.5, la serie ya esta completa y sin ningún valor faltante.

Capítulo 3

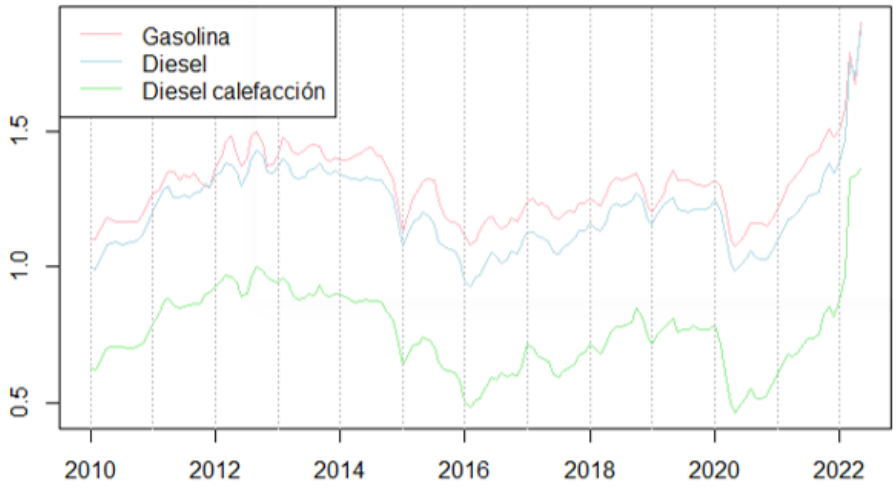
Análisis descriptivo

Ya presentada la base de datos con la que se va a trabajar y aplicados todos los tratamientos previos al inicio del análisis, en este capítulo se presenta un análisis descriptivo para entender mejor el comportamiento de los datos.

3.1. Análisis de los impuestos

Como se ha explicado en la introducción del proyecto, para llevar a cabo el análisis explicativo de la evolución de los precios de los carburantes, se ha tenido en cuenta la presencia o no de los impuestos aplicados. Por ello, se ha iniciado el estudio con un breve análisis descriptivo de la evolución de los precios de los principales carburantes aplicando el impuesto.

Figura 3.1: Precios de los carburantes con impuesto



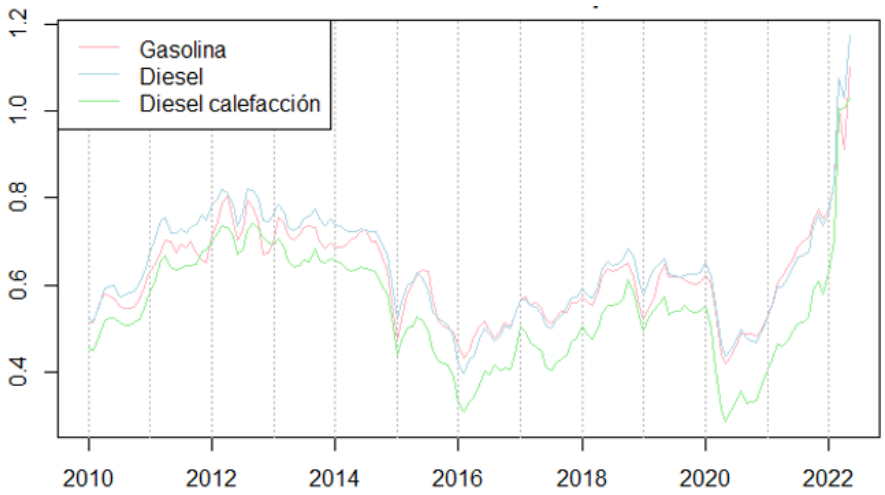
Fuente: Elaboración propia

Observando la figura 3.1 se puede ver que generalmente, durante el tiempo de estudio, los tres tipos de carburantes evolucionan con un patrón muy similar. Además, se destaca como el precio de la gasolina siempre ha sido el más elevado, a excepción de los últimos datos dónde el precio del diésel crece hasta coincidir con los valores presentados por la gasolina.

El precio del diésel para la calefacción se mantiene a niveles más bajos durante el periodo analizado, siendo así, el más económico de los tres tipos analizados. Cabe destacar que no solo presenta un menor precio, sino que también se le aplica un menor porcentaje de impuestos.

A continuación, para llevar a cabo el proyecto y detectar si hay diferencias en el patrón evolutivo del precio teniendo en cuneta o no los impuestos, se representará de igual forma la evolución de los precios de los tres tipos de carburantes sin impuestos.

Figura 3.2: Precios de los carburantes sin impuesto



Fuente: Elaboración propia

En la imagen 3.2, se observa de manera clara que el patrón evolutivo de los precios sin impuestos no se diferencian entre los diferentes tipos de combustibles a diferencia de lo observado con los precios con impuesto. Otro punto a destacar es que, sin aplicar los impuestos, el precio de diésel es el más elevado entre los tres.

En la tabla 3.1, se analizará los impuestos numéricamente.

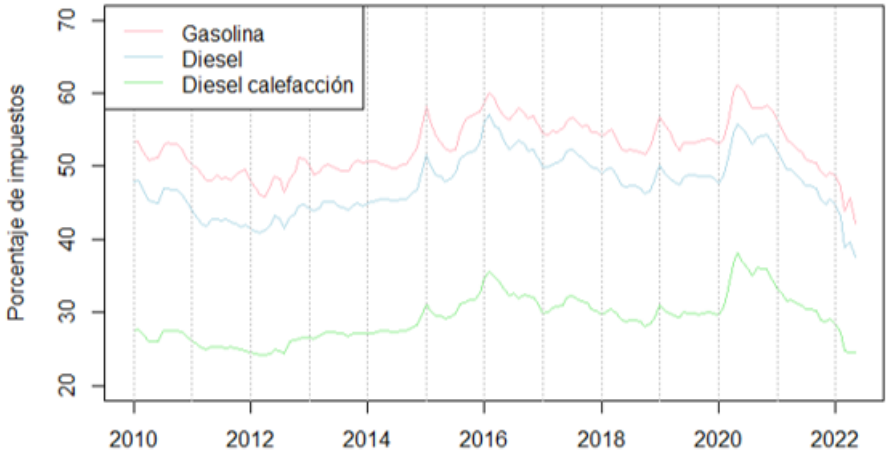
Cuadro 3.1: Porcentaje impuestos

Impuestos	**Precio medio con impuesto**	**Precio medio sin impuesto**	**Porcentaje de impuesto medio**
Gasolina	1.3020	0.6208	52.3 %
Diésel	1.2097	0.6390	47.2 %
Diésel para la calefacción	0.7625	0.5442	28.6 %

Fuente: Elaboración propia

El precio medio de gasolina durante los últimos 12 años es 1.302 euros por litro, y de cada litro, el 52,3 % de su precio son impuestos. El diésel tiene un impuesto un poco más bajo que la gasolina, en el periodo estudiado presenta un 47,2 %. Para el diésel para la calefacción se puede ver un 28.6 % de impuesto aplicado en el precio total del combustible.

Figura 3.3: Evolución de los impuestos



Fuente: Elaboración propia

En la figura 3.3 se observa que el porcentaje de los impuestos se comporta de manera inversa al precio de comercio. En otras palabras, durante los años 2010 al 2016, el precio de comercio se presenta como una parábola cóncava, con el precio máximo entre el 2012 y el 2013, y el mínimo en el 2016. Pero si se observa la gráfica con el porcentaje de impuestos que se presentan como parte del precio del combustible, el año 2012 es cuando se registra el menor valor de impuestos y a partir de allí, va incrementando hasta el año 2016. Cabe destacar también que los datos de los últimos años, en concreto a partir del final del año 2020, el precio de comercio creció de manera brusca, pero el porcentaje de los impuestos disminuyó.

Teniendo en cuenta que los precios de los carburantes a estudiar en este proyecto con y sin aplicar impuestos evolucionan de la misma manera, pero a diferente escala, se ha decidido seguir con el proyecto y cumplir así los objetivos solo con los precios con impuestos. Esto es debido a que es la cantidad total que paga cada uno de los consumidores de los carburantes (por ejemplo, a la hora de repostar).

3.2. Cambio estructural

Llevado a cabo un primer análisis descriptivo de los precios en el periodo de estudio, en esta sección se tendrá en cuenta la posible presencia de cambios estructurales en los datos registrados.

Observando la evolución de las tres series, se puede plantear que se han presentado varios cambios estructurales, en concreto, en el año 2020.

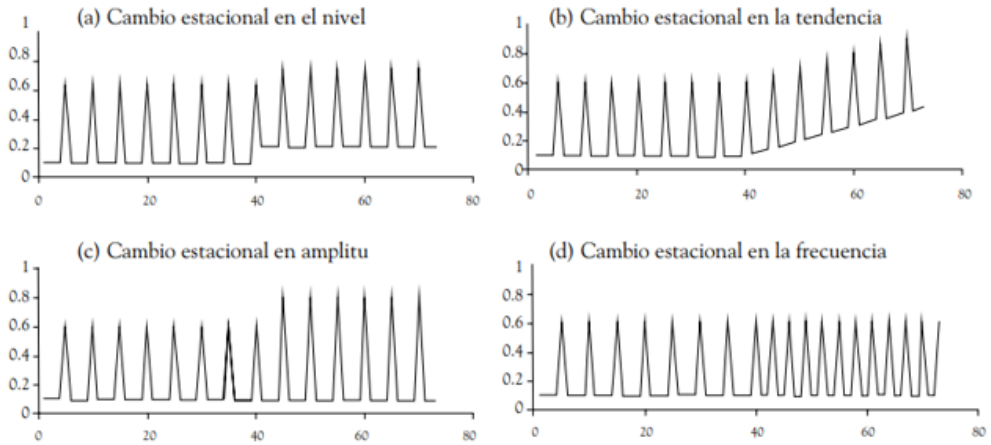
Es importante estudiar y tener en cuenta que los datos de una serie temporal pueden verse afectados por determinados acontecimientos. En este proyecto, se analizará los acontecimientos que hayan podido afectar al precio de los combustibles y de este modo producir cambios estructurales en los datos que deseamos analizar.

Cambio estructural: podemos definir como cambio estructural cuando hay un cambio instantáneo o permanente e inesperado en una variable recogida a lo largo del tiempo.

Teniendo en cuenta la definición de cambio estructural, si se estima un modelo con una serie sin tener en cuenta los cambios estructurales puede llevar a varias consecuencias. Por ejemplo, las predicciones pueden ser poco fiables, ya que las predicciones efectuadas con la metodología *Box-Jenkins* asume que todos los factores se mantienen estables a lo largo del tiempo.

Existen distintos tipos de cambios estructurales que se pueden ver visualmente en la figura 3.4. En primer lugar, está el cambio en el nivel que ocurre cuando sucede un evento que afecta el nivel de la serie permanentemente. También puede aparecer un cambio en la tendencia a causa de un evento que cambió la tendencia de la serie a creciente o decreciente. Por último, se pueden presentar cambios en la estacionalidad de la serie por causa de eventos que afectan la estructura estacional, modificando la amplitud o la frecuencia (Sánchez, 2008).

Figura 3.4: Tipos de cambios estructurales



Fuente: Sánchez, 2008

Una manera basada en la estadística para detectar la existencia de un cambio estructural es hacer el contraste de *Chow* (Zeileis y col., 2002), donde las hipótesis son:

$$H_0 : Permanencia\ o\ estabilidad\ estructural$$

$$H_1 : Cambio\ estructural$$

Para poder aplicar esta prueba, primeramente se divide los datos en dos series, a la primera se le llama *T1* y a la segunda *T2* y la T representa la serie completa. Cabe mencionar que para realizar esta prueba es necesario el mismo número de observaciones en cada grupo.

Como había mencionado anteriormente, se puede pensar que hay varios cambios estructurales a lo largo de los 12 años estudiados, pero al planear aplicar el método Box-Jenkins para este proyecto, teniendo más en cuenta los datos recientes que los pasados, se ha considerado analizar solo el cambio que se produjo en el año 2020. Por lo tanto, se define *T1* los datos de junio del 2018 hasta el mayo del 2020 y el *T2* serían las observaciones de junio de 2020 hasta el mayo de 2022.

Para construir el estadístico F del contraste de *Chow*, se usa la siguiente fórmula:

$$F_0 = \frac{SQE_T - (SQE_{T_1} + SQE_{T_2})/k}{(SQE_{T_1} + SQE_{T_2})/T - 2k} \sim F_{k,T-2k;\alpha}$$

Donde SQE es la suma del error al cuadrado de cada periodo o subperiodo y k representa el número de parámetros.

Teniendo en cuenta que si $F_0 \leq F_{(k,T-2k;\alpha)}$ se rechaza la hipótesis nula y se concluye que existe un cambio estructural. En caso contrario, no hay suficiente evidencia estadística para rechazar la hipótesis nula.

A continuación, se mostrará una tabla indicando el estadístico F, el *p-valor* y el resultado para cada uno de los carburantes.

Cuadro 3.2: Resultados del contraste de Chow

Contraste de Chow	**Estadístico F**	**P-valor**	**Resultado**
Gasolina	10.384	0.0008074	Rechazar H_0
Diésel	6.2113	0.007977	Rechazar H_0
Diésel para la calefacción	8.4439	0.002195	Rechazar H_0

Fuente: Elaboración propia

En los tres casos mostrados en la tabla 3.2 se ve un *p-valor* menor a un nivel de significación del 5 %, por lo tanto, se rechaza la hipótesis nula del contraste de

hipótesis, concluyendo así, que existe un cambio estructural para cada una de las tres series.

En este análisis, al tener un pequeño número de datos con el que trabajar para llevar a cabo el proyecto y cumplir los objetivos propuestos al inicio, aun haber determinado significativo el cambio estructural, no hay los datos mínimos necesarios para solo trabajar con una submuestra. En otras palabras, en este caso se debería construir un modelo por cada subperiodo por separado, ya que la metodología *Box-Jenkins* requiere al menos de 50 datos para la muestra para elaborar un modelo adecuado, pero, no hay suficientes datos en el último subperiodo. Por lo tanto, se ha decidido construir el modelo sin tenerse en cuenta este cambio estructural y hacer una comparación entre los valores predichos y valores observados para saber cómo evolucionarían si todos los factores se mantienen estables y que se está pagando de más por estos cambios.

3.3. Estadística descriptiva de las series

Para finalizar con el análisis descriptivo de los datos del proyecto, se llevará a cabo una estadística descriptiva con los principales estimadores para conocer más como se distribuyen los precios a lo largo del periodo estudiado.

Cuadro 3.3: Análisis descriptivo

Estimadores	**Mínimo**	**Q1**	**Mediana**	**Media**	**Q3**	**Máximo**
Gasolina	1.07875	1.1805167	1.2885000	1.2801365	1.367250	1.5010
Diésel	0.92540	1.0844375	1.2029250	1.1906533	1.302563	1.4325
Diésel para la calefacción	0.46350	0.6401833	0.7545417	0.7502096	0.875750	1.0030

Fuente: Elaboración propia

En los 11 años estudiados se puede ver en la tabla 3.3 que se llega a la misma conclusión que se ha descrito en los apartados anteriores. El precio de la gasolina es el precio más elevado entre los tres tipos de carburantes, con una media de 1.28 euros por litro seguido del gasoil con un 1.19 euros por litro. Por otra parte, el precio del gasoil para la calefacción es el más bajo entre los tres, con una media de 0.75 euros por litro.

Teniendo en cuenta que entre los objetivos planteados para este proyecto es explicar como han evolucionado los precios de los diferentes carburantes, es conveniente profundizar más en el análisis descriptivo y llevar a cabo un desglose por año de cada uno de los carburantes. En el cuadro 3.4 presentado, se puede ver como los

estadísticos media y mediana proporcionan resultados iguales hasta los céntimos, concluyendo que hay una distribución bastante simétrica en los datos. Cabe destacar que el precio medio parece haber crecido en el periodo estudiado y en los últimos años está volviendo a los niveles del inicio.

Cuadro 3.4: Análisis descriptivo para el precio de la gasolina por año estudiado

Año	**Mínimo**	**Q1**	**Mediana**	**Media**	**Q3**	**Máximo**
2010	1.10075	1.159950	1.168625	1.168625	1.175867	1.235667
2011	1.26760	1.297271	1.319375	1.318096	1.341312	1.351667
2012	1.36650	1.374875	1.410875	1.424460	1.466312	1.501000
2013	1.38825	1.411125	1.427000	1.430561	1.449888	1.477000
2014	1.24040	1.387313	1.403700	1.387526	1.414750	1.439250
2015	1.13025	1.165687	1.216225	1.228160	1.292875	1.327000
2016	1.08140	1.133375	1.159125	1.150264	1.172550	1.201000
2017	1.17460	1.202063	1.223950	1.218235	1.236167	1.250500
2018	1.22600	1.248363	1.308125	1.290454	1.327812	1.345400
2019	1.20300	1.290438	1.307825	1.296739	1.320038	1.356250
2020	1.07875	1.124167	1.160750	1.172925	1.188950	1.321750

Fuente: Elaboración propia

En el cuadro 3.5 se presenta de igual forma los principales estadísticos descriptivos por cada año del precio del diésel. La tabla verifica y reafirma las conclusiones a las que se ha llegado anteriormente.

Cuadro 3.5: Análisis descriptivo para el precio del diésel por año estudiado

Año	**Mínimo**	**Q1**	**Mediana**	**Media**	**Q3**	**Máximo**
2010	0.99350	1.0692000	1.086900	1.076807	1.093500	1.159333
2011	1.20260	1.2573625	1.270250	1.269422	1.290937	1.308250
2012	1.29750	1.3455000	1.354000	1.365690	1.389500	1.432500
2013	1.32700	1.3377500	1.357125	1.357975	1.372125	1.398000
2014	1.18200	1.3126875	1.322583	1.306731	1.327800	1.343750
2015	1.02250	1.0721875	1.106675	1.116893	1.170767	1.203250
2016	0.92540	0.9700625	1.024800	1.014068	1.051313	1.088667
2017	1.05020	1.0816250	1.108833	1.101011	1.128525	1.135667
2018	1.13450	1.1663625	1.220675	1.204904	1.236625	1.271800
2019	1.15550	1.2078875	1.214625	1.214315	1.221500	1.259000
2020	0.98625	1.0248000	1.033708	1.069369	1.076925	1.246000

Fuente: Elaboración propia

Para acabar con el desglose por año del análisis descriptivo de los diferentes carburantes, se presenta con el mismo formato los resultados para el diésel para la calefacción en el cuadro 3.6.

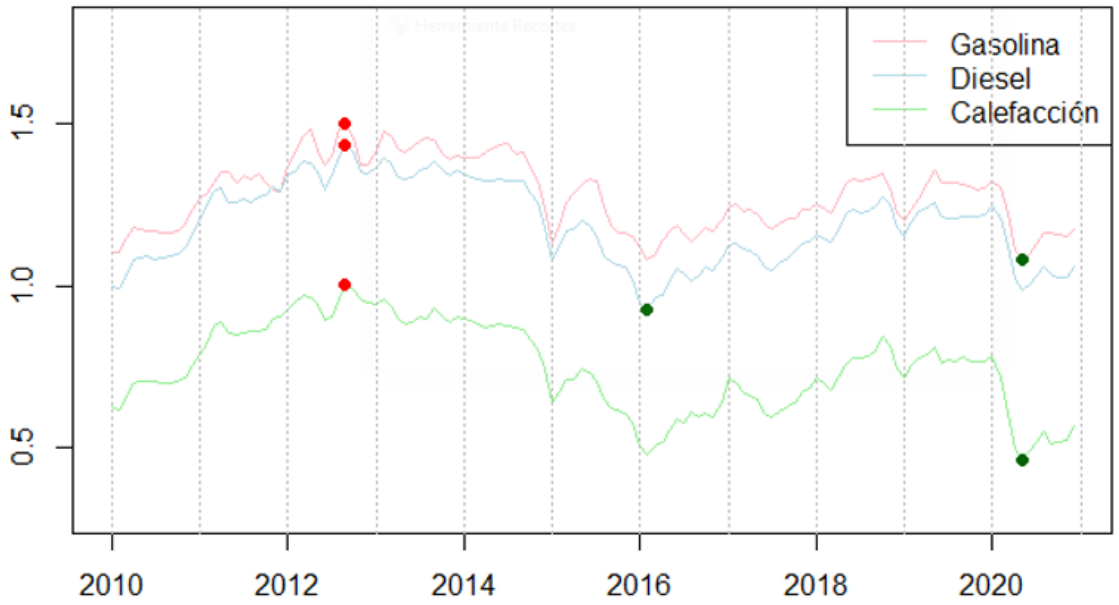
Cuadro 3.6: Análisis descriptivo para el precio del diésel para la calefacción por año estudiado

Año	Mínimo	Q1	Mediana	Media	Q3	Máximo
2010	0.61850	0.6883500	0.702500	0.691286	0.706250	0.755333
2011	0.79020	0.8519375	0.861750	0.860976	0.879937	0.904666
2012	0.89225	0.9387500	0.954875	0.950834	0.965250	1.003000
2013	0.88200	0.8953125	0.902625	0.911395	0.933562	0.961250
2014	0.74480	0.8585375	0.874375	0.858208	0.880800	0.899250
2015	0.57900	0.6201875	0.662450	0.666093	0.713166	0.742750
2016	0.48060	0.5147083	0.587375	0.565809	0.600075	0.638333
2017	0.59480	0.6250000	0.657866	0.657866	0.677437	0.716400
2018	0.68100	0.7172250	0.767750	0.760176	0.786437	0.846600
2019	0.71625	0.7658750	0.768750	0.769722	0.777775	0.811500
2020	0.46350	0.5091667	0.522550	0.563366	0.577950	0.786000

Fuente: Elaboración propia

Para finalizar con esta sección, se analizará el gráfico 3.5 donde se presenta la evolución del precio de los tres carburantes y de manera visual, se pueden identificar con puntos rojos el momento en que cada una de las series alcanzó su valor máximo y de color verde el momento en que presento el precio mínimo en el periodo.

Figura 3.5: Puntos máximos y mínimos de los precios



Fuente: Elaboración propia

Se puede ver de manera clara que el precio máximo de las tres series se encuentra en la misma fecha, en concreto, en septiembre del 2012, donde la gasolina presenta un precio de 1.50 euros por litro, el gasoil con 1.43 y el diésel para calefacción estaba a 1.003 euros por litro.

Por otra parte, el precio mínimo de la gasolina y el diésel para calefacción

coinciden en mayo del año 2020, con un precio de 1.08 y 0.46 euros por litro respectivamente. Por lo que hace al gasoil, este presenta el menor valor con 0.93 euros por litro en enero del 2016.

Con los diferentes análisis iniciales para estudiar la evolución de los precios de los carburantes se ha podido concluir con diversos métodos que las tres series evolucionan de manera similar. Durante el periodo que constituye del 2010 al 2016, los precios se presentan en forma de parábola cóncava. Este patrón que se ha detectado en este periodo se repite para los años del 2016 a 2020, con lo cual, se concluye que la tendencia de las series no es constante.

Capítulo 4

Metodología

En este capítulo se recogerán las diferentes teorías y definiciones para explicar los procedimientos que se llevaran a cabo a lo largo del proyecto y análisis explicativo de series temporales.

4.1. Box-Jenkins

La metodología *Box-Jenkins* proviene de Gorge E.P. Box y Gwilym M.Jenkins, autores que formalizaron el análisis de series temporales estocásticas univariante en los años 70. Los autores desarrollaron esta tipología de modelos de series temporales teniendo en cuenta que la variable dependiente del modelo es una combinación de los valores pasados de esta, es decir, los valores futuros (predicción) de la variable se explican solamente mediante el comportamiento pasado de los valores de la propia variable. Este tipo de modelo es apropiado para hacer predicciones a corto plazo, ya que se da más importancia a las observaciones más cercanas en el tiempo que las alejadas en un momento determinado. Genéricamente, el modelo Box-Jenkins también se denomina modelo *ARIMA*, que es un resumen de sus tres términos *AR* (Autorregresivo), *I* (Integrado) y *MA* (Medias Móviles).

Otro concepto muy importante en los análisis de series temporales es el ruido blanco. Un ruido blanco es una serie que presenta la media 0, una varianza constante y los retardos no están correlacionados.

Para poder aplicar la metodología *ARIMA*, la serie que se está analizando debe cumplir dos importantes características (Pankratz, 2009):

- La serie tiene que ser estacionaria en media (tendencia constante).

$$Media : E(X_t) = E(X_{t+h}) = \mu$$

- La serie tiene que ser estacionaria en varianza (varianza constante).

$$Varianza : V(X_t) = V(X_{t+h}) = \sigma^2$$

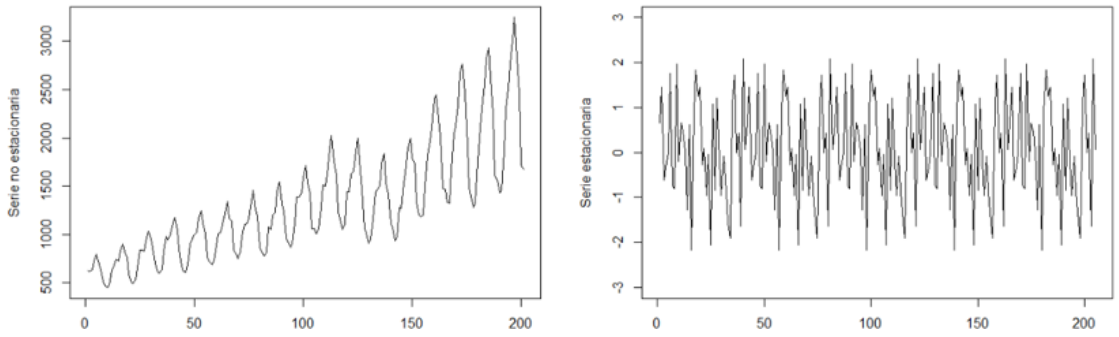
Si la serie cumple estas condiciones, decimos que la serie es estacionaria. La serie puede ser estacionaria en media y/o varianza. Para comprobar la estacionariedad de la serie, se observa la gráfica de la función de autorregresión simple (ACF), que tiene la siguiente expresión:

$$\rho(h) = \rho(X_t, X_{t+h}) = \frac{E((X_t - \mu)(X_{t+h} - \mu))}{\sqrt{E(X_t - \mu)^2 E(X_{t+h} - \mu)^2}} = \frac{\gamma(h)}{\gamma(0)}$$

Si una serie es estacionaria, su función de autorregresión decae rápidamente hacia cero, en caso contrario, la serie no es estacionaria. Sin embargo, no todas las series cumplen estas características, pero mediante unas sencillas transformaciones podemos conseguir que la serie sea estacionaria.

A continuación, se muestra en la figura 4.1 una gráfica en la izquierda como ejemplo de una serie no estacionaria y en la derecha de una serie temporal estacionaria.

Figura 4.1: Ejemplos gráficos de series temporales no estacionarias y estacionarias



Fuente: Elaboración propia

Observando ambas gráficas se puede ver que la figura de la izquierda no cumple la estacionariedad ni en media ni en varianza, ya que van aumentando con el tiempo. En cambio, la gráfica de la derecha presenta la media y varianza constantes.

Si la serie no cumple la estacionariedad en media, se debería de aplicar la transformación usando la diferenciación regular.

Para diferenciar de manera regular una serie de datos, se define una nueva variable W_t, que es una transformación de la serie original de la siguiente forma:

$$W_t = Z_t - Z_{t-1}, \quad t = 2, 3, ..., n$$

Ahora, la serie W_t es la serie Z_t aplicando una diferenciación. En caso de que aplicando la primera diferencia regular tampoco se consiga una media constante, se pueden aplicar tantas diferenciaciones como sean necesarias. Se define como una segunda diferenciación:

$$W_t = (Z_t - Z_{t-1}) - (Z_t - Z_{t-1}), \quad t = 3, 4, ..., n$$

Aparte de las diferenciaciones regulares, también están las diferenciaciones estacionales. Este tipo de diferenciación se aplica cuando una serie tiene una media no constante estacionalmente. En este caso, se aplicaría una diferenciación estacional de orden s, donde s representa el ciclo del periodo. Por ejemplo, si los datos son mensuales $s = 12$ observando una estacionalidad mensual, si los datos son diarios $s = 7$ representa una estacionalidad semanal.

A continuación, se muestra la fórmula de la primera diferenciación estacional:

$$W_t = Z_t - Z_{t-s}, \quad t = 2, 3, ..., n$$

Y si aún no se consigue una media estacional constante, se aplica una segunda diferenciación:

$$W_t = (Z_t - Z_{t-s}) - (Z_t - Z_{t-s}), \quad t = 3, 4, ..., n$$

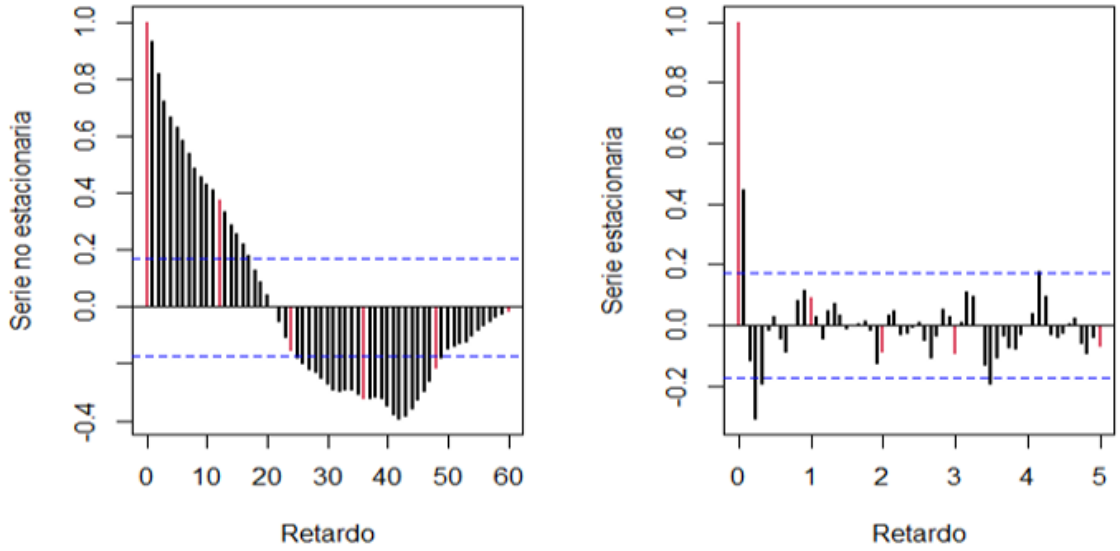
A efectos prácticos, la primera diferenciación estacional se necesita con bastante frecuencia, la segunda solo ocasionalmente y una tercera o más diferenciaciones casi nunca es necesario.

Una serie temporal, aparte de ser no estacionaria en media también, puede presentar no estacionariedad en la varianza. Para resolver este problema y poder trabajar con esta metodología, se aplica una transformación logarítmica a la serie original.

En la imagen 4.2, se muestra una gráfica comparando la gráfica de la función de autorregresión de una serie estacionaria en la derecha de la imagen, y otra no

estacionaria en la izquierda.

Figura 4.2: Gráficas del ACF de una serie no estacionaria y estacionaria



Fuente: Elaboración propia

Tal y como se había mencionado anteriormente, el modelo *ARIMA* se define por tres términos, *AR* (Autorregresivo), *I* (Integrado) y *MA* (Medias Móviles).

Aparte de analizar la estacionariedad de la serie, el *ACF* también se utiliza para identificar el orden del modelo medias móviles, observando la significancia de los primeros retardos. Para identificar el orden del modelo autorregresivo, se usa la gráfica de la función de autocorrelación parcial (PACF), que mide la relación dentro de una serie de datos y es útil para dar una idea de los patrones de los datos disponibles.

$$corr(X_{t+h}, X_t | X_{t+h-1}, ..., X_{t+1})$$

Estas funciones se usan como una guía para escoger uno o más modelos como candidatos del modelo final que mejor se ajusta a los datos y así, explicar mejor la evolución de las series temporales.

4.1.1. Modelo autorregresivo

El modelo $AR(p)$, modelo Autorregresivo de orden p retrocede p periodos para realizar la predicción, es decir, el valor actual de la serie es una combinación lineal de la serie estudiando los p periodos anteriores. El modelo autorregresivo es un modelo de regresión en que tanto la variable dependiente como la variable explicativa, son la misma variable. La variable dependiente está explicada por las observaciones procedentes de la misma variable (Villavicencio, 2010). La fórmula matemática que define el modelo es:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + Z_t$$

Expresada en otros términos, donde B representa el operador retroceso ($BX_t = X_{t-1}$), la fórmula que define el modelo sería:

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) X_t = Z_t, \quad Z_t \sim N(0, \sigma_z^2)$$

Donde X_t es el valor resultante para el mes t, y los ϕ son los parámetros que multiplican a las observaciones del día anterior hasta el mes $t - p$ y Z_t representa el ruido blanco.

4.1.1.1. Modelo $AR(1)$

Un proceso autorregresivo de orden 1, es decir, un $AR(1)$, considerando $\mu = 0$ y la varianza constante, tendrá la siguiente expresión:

$$X_t = \phi X_{t-1} + Z_t, \quad Z_t \sim N(0, \sigma_z^2)$$

Expresada de otra manera, la anterior expresión equivale a:

$$(1 - \phi B) X_t = Z_t, \quad Z_t \sim N(0, \sigma_z^2)$$

Los coeficientes de autocorrelación para el proceso $AR(1)$ es:

$$\rho_h = \frac{\gamma(h)}{\gamma(0)} = \phi^h$$

4.1.1.2. Modelo $AR(2)$

Un proceso autorregresivo de orden 2, es decir, un $AR(2)$, considerando $\mu = 0$ y la varianza constante, tendrá la siguiente expresión:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t, \quad Z_t \sim N(0, \sigma_z^2)$$

De la misma forma que se ha definido anteriormente, la expresión para este modelo mediante el uso del operador retroceso (B) se define como:

$$(1 - \phi_1 B - \phi_2 B^2) X_t = Z_t, \quad Z_t \sim N(0, \sigma_z^2)$$

Los coeficientes de autocorrelación para el proceso $AR(2)$ se definen como:

$$\rho_h = \frac{\gamma(h)}{\gamma(0)} = \phi_1\rho_{h-1} + \phi_2\rho_{h-2}, \quad h = 1, 2, 3, ...$$

Donde $\gamma(0)$ es la función de autocorrelación, $\gamma(h)$, la función de autocovarianza en orden h. Para una expresión detallada de autocorrelación y autocorrelación parcial, se puede consultar el artículo Pankratz, 2009.

4.1.2. Modelo media móvil

El modelo $MA(q)$ o el modelo *Moving Average* (Media móvil) de orden q es un modelo de regresión que depende de los errores pasados y tiene la siguiente expresión:

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + ... + \theta_q Z_{t-q}, \quad Z_t \sim N(0, \sigma_z^2)$$

Expresada de otra manera, usando el operador retroceso, se puede definir como:

$$X_t = (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q) Z_t, \quad Z_t \sim N(0, \sigma_z^2)$$

Donde X_t es el valor resultante para el mes t, y los θ son los parámetros que multiplican a los errores del mes anterior hasta el mes $t - p$ y Z_t representa el ruido blanco.

4.1.2.1. Modelo $MA(1)$

Un modelo media móvil del orden 1, es decir, $MA(1)$, considerando $\mu = 0$ y la varianza constante, tendrá la siguiente expresión:

$$X_t = Z_t + \theta Z_{t-1}, \quad Z_t \sim N(0, \sigma_z^2)$$

También, se puede definir el modelo con la siguiente expresión matemática:

$$X_t = (1 + \theta_1 B) Z_t, \quad Z_t \sim N(0, \sigma_z^2)$$

La función de autocorrelación se define de la forma que se muestra a continuación:

$$\rho_1 = \frac{\gamma(1)}{\gamma(0)} = \frac{\theta}{1+\theta^2}$$

$$\rho_h = \frac{\gamma(h)}{\gamma(0)} = 0,\ h > 1$$

4.1.2.2. Modelo MA(2)

Un modelo media móvil del orden 2, es decir, $MA(2)$, considerando $\mu = 0$ y la varianza constante, tendrá la siguiente expresión:

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}, \quad Z_t \sim N(0, \sigma_z^2)$$

Que es equivalente a la siguiente formulación:

$$X_t = (1 + \theta_1 B + \theta_2 B^2) Z_t, \quad Z_t \sim N(0, \sigma_z^2)$$

Con funciones de autocorrelación:

$$\rho_1 = \frac{\gamma(1)}{\gamma(0)} = \frac{-\theta}{1+\theta^2}$$

$$\rho_h = \frac{\gamma(h)}{\gamma(0)} = 0,\ h > 1$$

Donde $\gamma(0)$ es la función de autocorrelación, $\gamma(h)$ es la función de autocovarianza en orden h. Para la demostración de la función autocorrelación y autocorrelación parcial, se puede consultar el artículo publicado en Pankratz, 2009.

Un modelo $ARIMA(p, d, q)$ tiene como expresión general:

$$\phi_p(B)(1-B)^d X_t = \theta_q(B) Z_t$$

Que se define por tres componentes principalmente. En primer lugar, encontramos el parámetro p que representa el orden del modelo autorregresivo, el parámetro d identifica el número de las veces que se han aplicado diferenciaciones regulares a la

serie cuando su media no es constante en el tiempo, y el parámetro q representa el orden de un modelo de media móvil.

Si la serie tiene componente estacional, la serie es nombrada como *Seasonal ARIMA* o $SARIMA(p,d,q)(P,D,Q)_s$ y presenta como expresión general la siguiente fórmula:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D X_t = \theta_q(B)\Theta_Q(B^s)Z_t$$

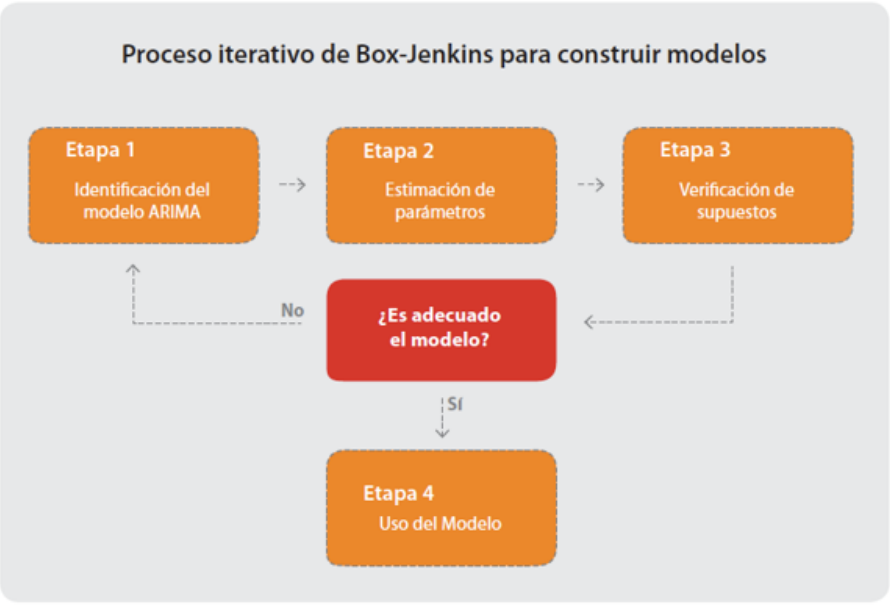
Donde S indica el ciclo del periodo, D es el número de diferenciaciones estacionales que se han aplicado sobre la serie temporal, y P y Q representan el orden del modelo AR y MA teniendo en cuenta solo los retardos múltiples de S.

4.1.3. Etapas

En esta sección se procederá a la presentación y explicación de cada una de las etapas que componen la metodología *Box-Jenkins*.

Box y Jenkins propusieron un procedimiento basado en tres etapas para encontrar un modelo óptimo mediante la identificación, la estimación y finalmente, la validación de este. Hay que tener en cuenta que definieron como requisito antes de empezar el proceso la verificación de que la serie temporal que se usará para el análisis sea estacionaria para poder aplicar el modelo $ARIMA$.

Figura 4.3: Etapas del proceso de Box-Jenkins



Fuente:
http://ingnsist.blogspot.com/2017/05/metodo-de-box-jenkins-la-metodologia.html

En la imagen 4.3, se puede ver el esquema que definieron para seguir paso a paso y así, conseguir un buen modelo para cumplir los objetivos o llevar a cabo el testeo de las hipotesis plateadas al inicio del proyecto (Pankratz, 2009):

4.1.3.1. Primera etapa: Identificación

La etapa de identificación se trata de escoger uno o más modelos como candidatos. En esta etapa, se identifica la parte del modelo MA observando la significancia de los primeros retardos con la función de autocorrelación (ACF) y la parte del modelo AR mediante la función de autocorrelación parcial ($PACF$). Aparte de los primeros retardos, también es necesario estudiar los primeros retardos múltiples a s. En caso de que estos resulten significativos, se determinará que existe un componente estacional en los datos que deberá recogerse y tenserse en cuenta dentro del modelo mediante el planteamiento de un SARIMA (*seasonal ARIMA*). Como se ha mencinado anteriormente, este tipo de modelo sigue una estructura $SARIMA(p,d,q)(P,D,Q)_s$ donde (p,d,q) son los parámetros para la parte regular del modelo y $(P,D,Q)_s$ identifican la parte estacional. Si no se presenta una componente estacional en los datos, el modelo se convierte automáticamente en un modelo $ARIMA(p,d,q)$.

4.1.3.2. Segunda etapa: Estimación

Una vez llevada a cabo la identificación de los posibles modelos, siguiendo el esquema de la metodología *Box-Jenkins*, hay que estimar los parámetros que definen los modelos planteados. En esta etapa hay que tomar en consideración que un modelo AR siempre será invertible y un modelo MA siempre será estacionario, por ello, habrá que cerciorarse de la invertibilidad del modelo MA y estacionariedad del modelo AR. Finalmente, para acabar con esta etapa, es necesario confirmar que cada uno de los parámetros que componen el modelo son significativos.

Cuadro 4.1: Ejemplos para el proceso de estimación

Modelo	**Expresión**	**Invertibilidad**	**Estacionariedad**
AR(1)	$Z_t = (1-\phi B)X_t$	-	$\lvert\phi\rvert < 1$
AR(2)	$Z_t = (1-\phi_1 B-\phi_2 B^2)X_t$	-	$\lvert\phi_1 \pm \phi_2\rvert < 1$
MA(1)	$X_t = (1-\theta B)Z_t$	$\lvert\theta\rvert < 1$	-
MA(2)	$X_t = (1-\theta B-\theta_2 B^2)Z_t$	$\lvert\theta_1 \pm \theta_2\rvert < 1$	-

Fuente: Elaboración propia

Como se puede ver en el cuadro 4.1 anterior se muestran diversas expresiones de diferentes modelos concretos para ejemplificar el procedimiento de estimación de la metodología *Box-Jenkins*. Teniendo en cuenta que B representa el operador retroceso, también se presentan las condiciones que se deben cumplir para determinar

que el modelo es invertible y estacionario en cada uno de los casos.

Una vez asegurado que los modelos son invertibles y estacionarios, se planteará un análisis de la significancia de cada parámetro definiendo como hipótesis a contrastar:

$$H_0 : \phi_i = 0$$

$$H_1 : \phi_i \neq 0$$

Donde $\hat{\phi}_i \approx N(\phi_i, \sigma_{\phi_i})$, $\hat{t} = \frac{\phi}{se(\phi_i)} \approx t - Student_{T-k}$.

En términos prácticos, dentro del *software* que se utilizará para llevar a cabo el análisis, se hará una división entre el valor estimado con su error al cuadrado. Con el resultado de esta operación, si el valor resultante es mayor que dos en valor absoluto, se podrá concluir que el parámetro es estadísticamente significativo para pertenecer al modelo plateado y de este modo, explicar la serie temporal que se está analizando.

En caso de que se haya identificado más de un modelo, será necesario decidir cuál de ellos es el óptimo según los objetivos planteados al inicio del análisis. Para seleccionar el mejor modelo, se hará una comparación entre ellos para ver cuál es el modelo que explica mejor los datos mediante el valor que se obtiene con el AIC (criterio de información de Akaike) que se calcula de la siguiente manera:

$$AIC = -2logL(\phi, \theta, \sigma^2|X) + 2p$$

Donde p es el número de parámetros y $L(\phi, \theta, \sigma^2|X)$ hace referencia a la expresión del valor máximo para la función de verosimilitud.

Se puede determinar que el modelo que presenta menor valor de AIC será el modelo que mejor se ajusta a los datos. Una vez seleccionado el modelo, se pasará a la tercera etapa.

4.1.3.3. Tercera etapa: Validación

La tercera etapa de esta metodología se recoge la validación de los modelos escogidos. Tal y como se ha mencionado al inicio del capítulo, los modelos $SARIMA(p, d, q)(P, D, Q)_s$ se definen por la siguiente expresión matemática:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D X_t = \theta_q(B)\Theta_Q(B^s)Z_t$$

Y la parte estocástica del modelo es representada por $Z_t \sim N(0, \sigma_z^2)$.

Por lo tanto, al presentar dentro del modelo una parte aleatoria habrá que verificar los siguientes aspectos y de este modo dar por válido el modelo.

- Normalidad
- Homogeneidad de la varianza
- Independencia

Para estudiar la normalidad del término estocástico del modelo se usará tanto un análisis numérico como uno gráfico. Debido a que las pruebas numéricas no son muy concretas y precisas en caso de presencia de valores atípicos, la ayuda y uso de gráficas para estudiar la distribución normal para los errores del modelo será necesario.

Entre los diferentes contrastes para la determinación de distribución normal, se ha seleccionado en este proyecto el contraste de *Shapiro-wilk* (*normality test*), la prueba de *Anderson-Darling test* y por último el *Jarque-Bera test*, ya que cada uno define el estadístico teniendo en cuenta la media, los cuartiles y otros criterios básicos y muy característicos para el estudio de la presencia de normalidad. De todos modos, aún presentar el estadístico de contraste diferente para cada una de las pruebas, el contraste de hipótesis para todas las pruebas se define de la siguiente forma:

H_0 :los residuos provienen de una población normalmente distribuida

H_1 :los residuos no provienen de una población normalmente distribuida

Para acompañar y verificar que las pruebas numéricas son correctas, se plantea de manera gráfica el *qqplot* y el histograma para visualizar de manera sencilla la distribución del término estocástico del modelo.

Siguiendo con el proceso de validación, la segunda característica que se debe de cumplir y validar es la homogeneidad de la varianza. Para llevar a cabo su estudio, se representarán en un gráfico los residuos y la raíz cuadrada de residuos en valor absoluto. Del mismo modo que se ha hecho anteriormente, además de la validación visual mediante el análisis de los diferentes gráficos, también se puede realizar la prueba de *Breusch-Pagan* definiendo el contraste de hipótesis como:

H_0 :Los residuos tienen varianza constante

H_1 :Los residuos no tienen varianza constante

Para acabar las diferentes comprobaciones del cumplimiento de las características que debe presentar el modelo, se debe analizar la independencia. Por eso, se presentará el gráfico tanto para el ACF como para el $PACF$ de los residuos para analizarla de manera visual. Para verificar las conclusiones a las cuales se ha llegado mediante el estudio de los diferentes gráficos, también se realizarán el *Ljung-box test* y el *Durbin-Watson test*.

Para cada una de las pruebas numéricas se define ligeramente diferente el contraste de hipótesis, siendo estos del siguiente modo. Para la prueba de *Durbin-Watson test* el contraste define la hipótesi nula y alternativa como:

H_0 :Los datos se distribuyen de forma independiente

H_1 :Los datos no se distribuyen de forma independiente

En cuanto a la prueba de *Ljung-box test* las hipótesis se definen como:

H_0 :Las autocorrelaciones hasta retardo k son iguales a cero

H_1 :Las autocorrelaciones hasta retardo k no son iguales a cero

Para finalizar con el esquema presentado por la metodología, cumplida la tercera etapa validando las diferentes características que deben presentar los residuos del modelo escogido, se procederá con la predicción de valores futuros usando el modelo que se ha considerado óptimo y ha pasado las diferentes etapas satisfactoriamente.

Como último tratamiento para iniciar con la predicción de valores futuros es conveniente analizar la estabilidad del modelo ajustando el modelo sin los últimos datos. En otras palabras, se debe testear que el modelo escogido se ha adaptado correctamente a los datos y sigue explicando con los mismos parámetros los datos que ya conocemos. Para considerar que el modelo es estable, se deberá observar que no hay una gran variación de coeficientes entre los datos que se han usado para crear el modelo y los datos que se han usado para validar la estabilidad del modelo. Además, se deben mantener los mismos signos.

4.1.3.4. Predicción

Escogido ya el mejor modelo para la serie temporal que se está estudiando, finalmente, se procederá a la predicción de los últimos valores y se comparará con los datos reales para averiguar si el modelo hace buena predicción o no.

Para cuantificar la capacidad predicativa del modelo escogido, se estudiarán los valores que se obtienen con los indicadores $RMSPE$ (*Root Mean Square Percentage Error*) y $MAPE$ (*Mean Absolute Percentage Error*).

- **RMSPE:** es un indicador que muestra el porcentaje del error cuadrático medio.

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{y}_i}{y_i})^2}$$

- **MAPE:** es un indicador que muestra el error porcentual absoluto.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i - \hat{y}_i}{y_i}|$$

Capítulo 5

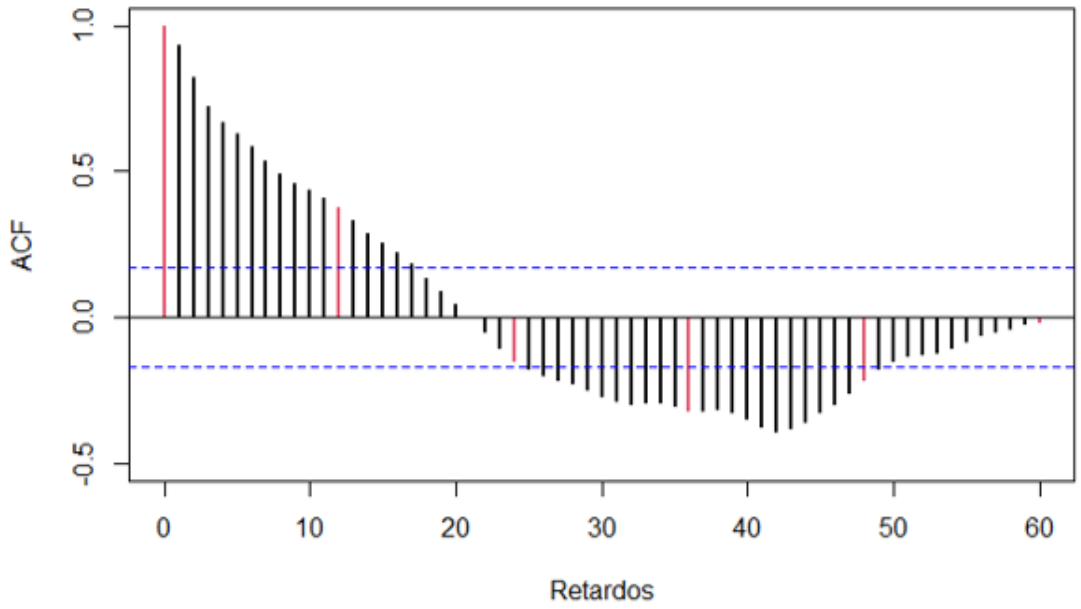
Análisis ARIMA: Gasolina

En este capítulo, se aplicará la metodología *ARIMA* definida en el apartado anterior y así, encontrar un modelo adecuado para la serie de la gasolina súper 95.

5.1. Transformaciones previas

Antes de empezar a identificar los posibles modelos, es necesario estudiar la estacionariedad de la serie.

Figura 5.1: Gráfica ACF de la serie gasolina

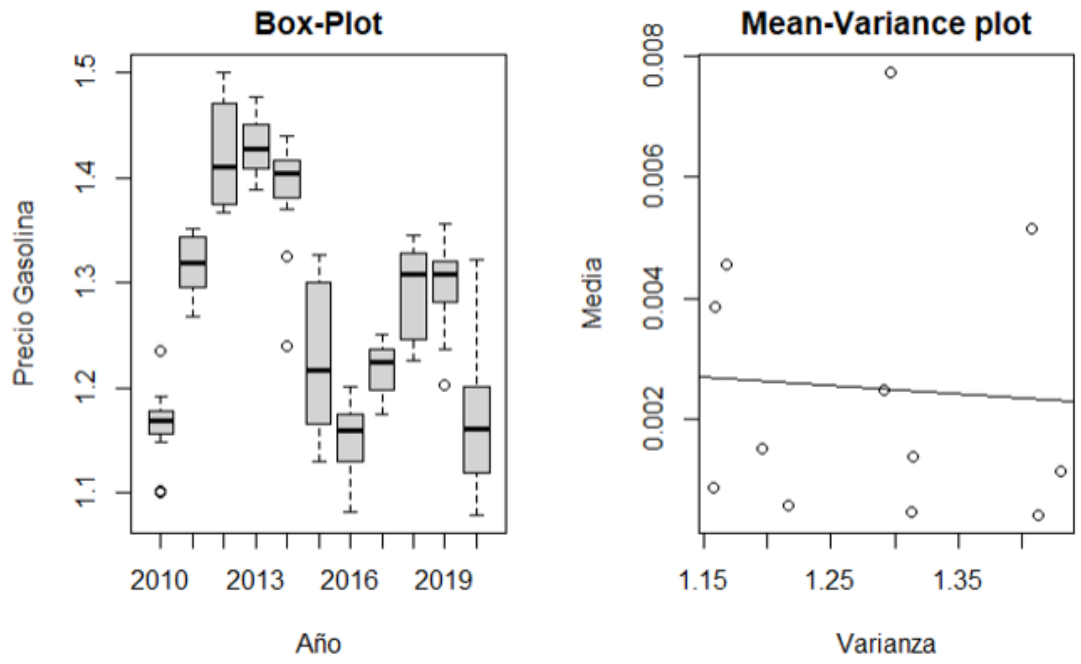


Fuente: Elaboración propia

En la gráfica que se presenta en la figura 5.1 se observa el *ACF* para el precio de la gasolina. Con este recurso visual, se puede determinar si la serie realmente es estacionaria o no. En este caso, se observa que los retardos decrecen muy lentamente

hacia cero, por lo tanto, se puede determinar que la serie no es estacionaria. Para poder llevar a cabo un análisis con esta metodología, es necesario identificar qué componente de la serie se le tiene que aplicar una transformación. Es decir, se llevará a cabo un estudio de la media y varianza.

Figura 5.2: Box-plot y gráfica de media-varianza



Fuente: Elaboración propia

En la figura 5.2 se representan en gráficos *box-plot* y un gráfico de media-varianza los datos del carburante y así, identificar si la media y varianza son constantes a lo largo del tiempo.

Examinando las líneas negras del *boxplot* y la posición de los puntos en el eje y del gráfico media-varianza, se ve que la media de cada año no se mantiene en el mismo nivel, por lo tanto, se puede decir que la media de la serie del precio de gasolina no es constante en el tiempo.

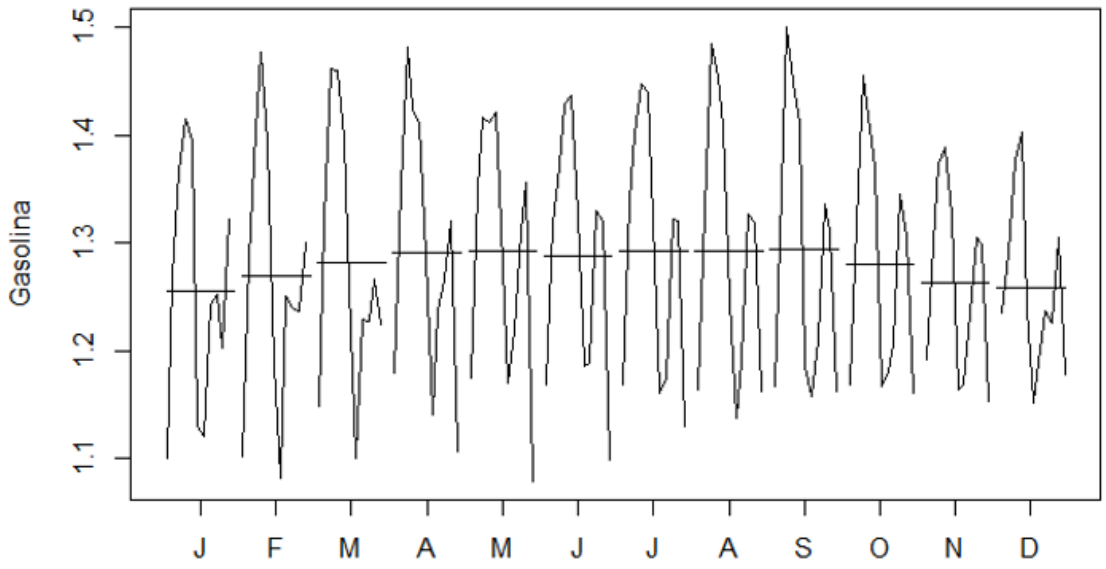
Por lo que hace a la varianza, si se estudian los rangos intercuartílicos del *boxplot*, se ve claramente que la varianza no es constante, ya que se presentan muchas diferencias entre las diferentes amplitudes de las cajas. Además, con la representación gráfica de la media-varianza se llega a la misma conclusión.

Como se ha definido en el capítulo de la metodología de *Box-Jenkins*, para solucionar el problema de la varianza se ha aplicado una transformación logarítmica sobre la serie original.

Aplicada la primera transformación a la serie temporal, se realizará el análisis de la estacionalidad de la serie aplicando el *monthplot*, donde se agrupan los datos

de cada año que corresponden el mismo mes.

Figura 5.3: Monthplot de la serie gasolina



Fuente: Elaboración propia

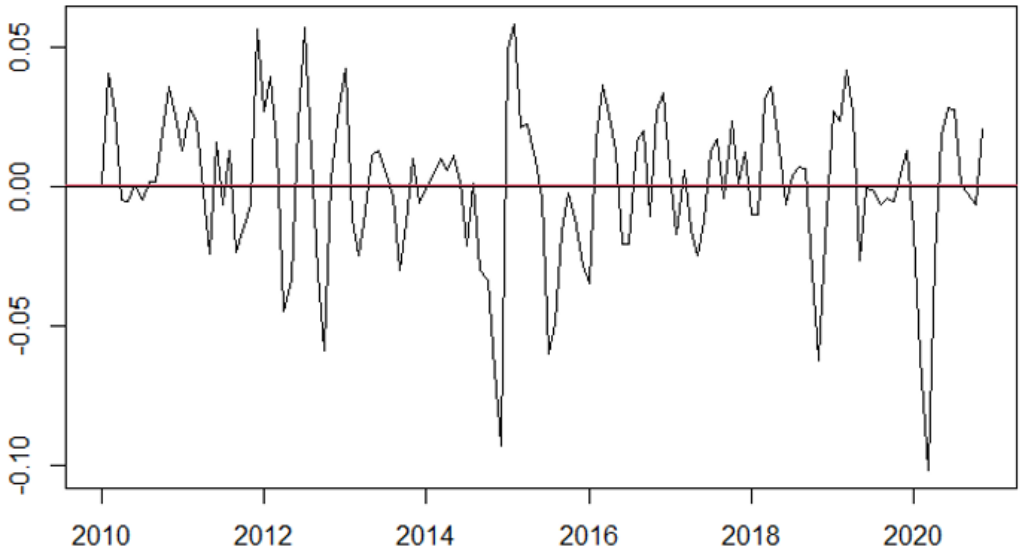
Como se muestra en la figura 5.3, el gráfico puede presentar indicios de que el precio de abril a septiembre es más elevado que el resto del año. Pero teniendo en cuenta la escala del *monthplot*, la diferencia que se presentan es menos de 5 céntimos, considerando así, que la serie del precio de gasolina no muestra una variación en la estacionalidad estadísticamente significativa. De este modo se podría concluir que no es necesario aplicar ninguna diferenciación estacional.

Como se había advertido con la ayuda de la figura 5.2, analizando la última etapa de posibles transformaciones previas que puedan aplicarse a la serie temporal de este carburante, se había concluido que la media no era constante a lo largo del tiempo. Es por ello, de la misma forma que se ha definido dentro del capítulo anterior, será necesaria la aplicación de otra transformación a los datos y así, obtener una serie temporal estacionaria.

Para solucionar la no constancia de la media, se debe aplicar las diferenciaciones regulares necesarias para remover las raíces unitarias.

En la imagen 5.4, se muestra el gráfico de la serie temporal constituida por los precios de la gasolina después de aplicar una diferenciación regular. Cabe destacar que la línea roja representa la media de la serie y la línea negra representa el cero. De esta forma, se puede ver que la media de la serie corresponde a la línea negra, es decir, la media es constante y centrada en cero.

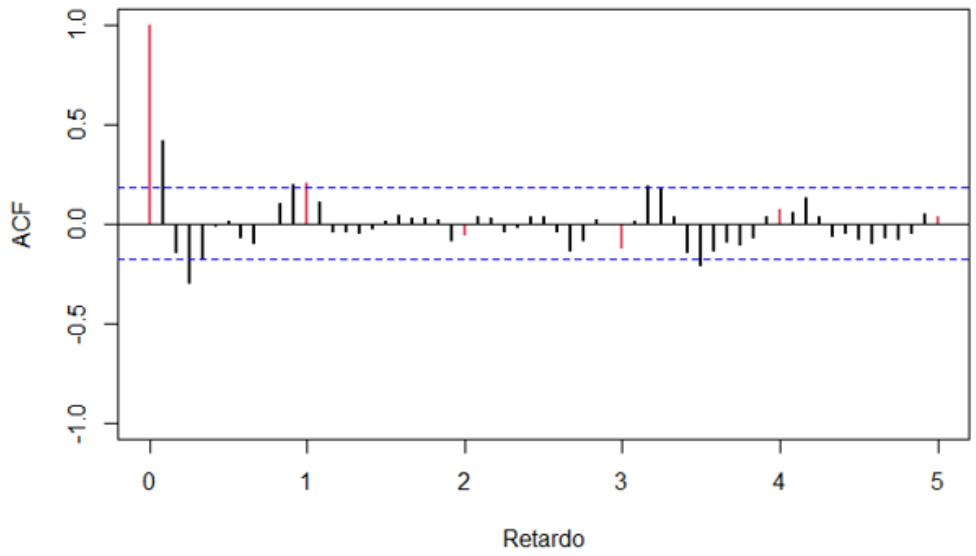
Figura 5.4: Media de la serie gasolina con una diferenciación



Fuente: Elaboración propia

Para poder determinar que las transformaciones aplicadas sobre los datos eran las necesarias y con ellas se ha obtenido una serie estacionaria preparada para iniciar con la metodología *Box-Jenkins*, se comprobará mediante la gráfica del ACF de la serie transformada.

Figura 5.5: Gráfica ACF de la serie gasolina transformada



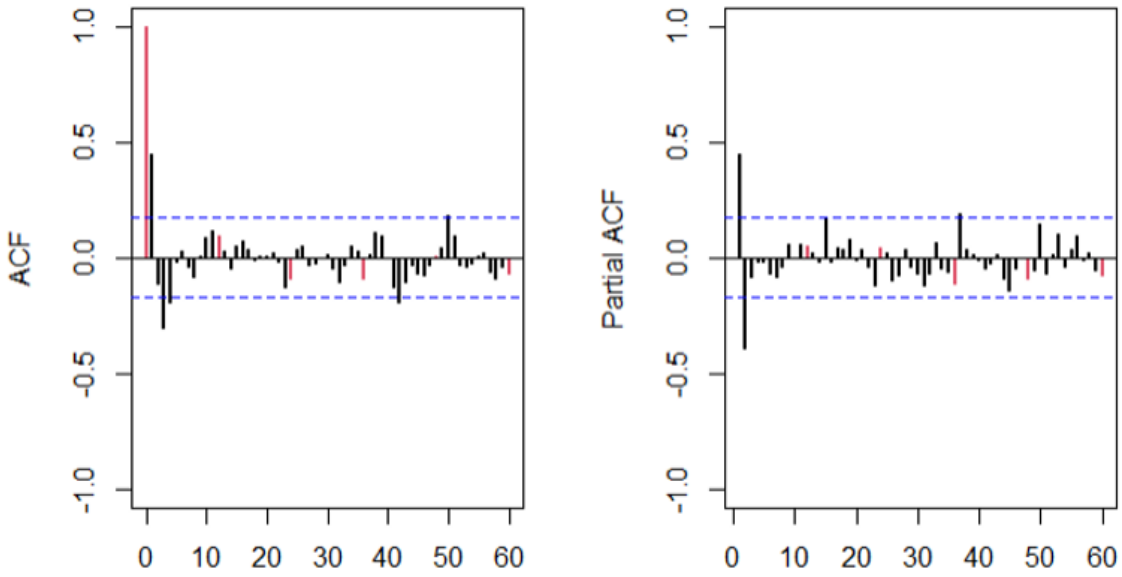
Fuente: Elaboración propia

En la figura 5.5, se pueden ver que los retardos decaen de manera rápida hacia 0 indicando que la serie de los precios de la gasolina con las transformaciones aplicadas es estacionaria.

5.2. Identificación

Una vez asegurada la estacionariedad de la serie, se llevará a cabo la identificación de los posibles modelos que se puede ajustar al precio de la gasolina a través de las gráficas ACF y $PACF$. Para una mejor identificación se muestra en color rojo los retardos múltiples de 12. Además, para determinar los retardos superiores a 0, se mostrará los gráficos con bandas de confianza al 95 %.

Figura 5.6: Grafica de ACF y PACF de la serie gasolina transformada



Fuente: Elaboración propia

Con el gráfico de la función ACF, se pueden identificar y proponer para la parte regular un modelo $MA(4)$. Aun así, teniendo en cuenta el principio de parsimonia, sería mejor un modelo con menos parámetros. Por otra parte, observando los retardos múltiples a s, en este caso s es igual a 12, se aprecia que ninguno de los retardos estacionales es significativo, es decir, la parte estacional presenta ruido blanco.

En la misma figura 5.6, con la gráfica de la función de autocorrelación parcial ($PACF$) se puede plantear o sugerir un modelo $AR(2)$ para la parte regular. Del mismo modo que se ha visto con el ACF, la parte estacional también presenta ruido blanco, confirmado de esta manera que la serie no tiene componente estacional.

Teniendo en cuenta las transformaciones a la serie y el análisis llevado a cabo para la identificación de posibles modelos explicativos para el carburante, se han propuesto los siguientes modelos:

- **Modelo 1:** $MA(4)$ para W_t;

$$W_t = (1-B)^d(1-B^s)^D ln(X_t), \quad d = 1, \ D = 0, \ s = 12;$$

La formula compacta del modelo es:

$$(1-B)X_t = \theta_4(B)Z_t$$

Sustituyendo cada polinomio característico, se obtiene:

$$(1-B)ln(X_t) = (1+\theta_1 B+\theta_2 B^2+\theta_3 B^3+\theta_4 B^4)Z_t$$

- **Modelo 2:** $AR(2)$ para W_t;

$$W_t = (1-B)^d(1-B^s)^D ln(X_t), \quad d=1,\ D=0,\ s=12;$$

La formula compacta del modelo es:

$$\phi_2(1-B)X_t = Z_t$$

Sustituyendo cada polinomio característico, se obtiene:

$$(1-\phi_1 B-\phi_2 B^2)(1-B)ln(X_t) = Z_t$$

5.3. Estimación

En esta sección se presentará los procedimientos de la segunda parte de la metodología *Box-Jenkins* para el tratamiento de la serie temporal.

En concreto, se ajustará los modelos propuestos mediante la función $ARIMA$ del programa informático *R-studio* usando la serie estacionaria W_t (el código 5.1 es el usado para la estimación con el lenguaje R). Se debe recordar que la serie W_t es estacionaria, en este caso corresponde a la serie original aplicando la transformación logarítmica y una diferenciación regular.

```
1 library(stats)
2 modelo = arima(serie_estacionaria, order=c(p,d,q))
```

Listing 5.1: Código aplicado en R para la estimación

Aplicada la función correspondiente para llevar a cabo la estimación, para el primer modelo definido en la sección anterior se ha obtenido la salida que se puede ver en la figura 5.7. En la imagen se puede ver los resultados de las estimaciones llevados a cabo en R. En primer lugar, se vuelve a escribir la función usada con el nombre de la serie que se ha guardado y los parámetros que se han tomado en consideración. En segundo lugar, se presentan los coeficientes estimados para el

modelo y los errores de estimación de cada uno de los parámetros. Finalmente, se pueden visualizar diferentes estadísticos para decidir si la estimación es buena o no.

Figura 5.7: Salida de estimación del modelo $MA(4)$ para X_t

```
Call:
arima(x = d1lngas, order = c(0, 0, 4))

Coefficients:
         ma1      ma2      ma3      ma4  intercept
      0.5761  -0.0361  -0.2840  -0.1975     0.0003
s.e.  0.0860   0.0972   0.0978   0.0845     0.0021

sigma^2 estimated as 0.0005058:  log likelihood = 310.93,  aic = -609.86
```

Fuente: Elaboración propia

De este mismo modo, en la figura 5.8 se ha presentado una tabla con los resultados de todos los parámetros para ambos modelos que se han identificado con anterioridad, los estadísticos *t-ratios*, el valor de log-verosimilitud maximizado, la varianza estimada y el valor del AIC correspondiente a cada modelo. Con estos resultados, se estudiará la significancia de cada uno de los parámetros, teniendo en cuenta que la significancia se determina mediante la división de los coeficientes estimados y el error estándar, resultando con el valor de los *t-ratios*. Si el *t-ratio* resulta mayor que dos en valor absoluto, se considera que el parámetro es significativo y necesario que permanezca en el modelo.

Figura 5.8: Tabla resumen con los principales resultados para la estimación de los modelos

Resultados

	Dependent variable:	
	lnDiesel	
	MA(4) (1)	AR(2) (2)
ma1	0.510	
	t = 5.672	
ma2	-0.0004	
	t = -0.004	
ma3	-0.144	
	t = -1.484	
ar1		0.550
		t = 6.589
ar2		-0.307
		t = -3.670
intercept	0.0004	0.0004
	t = 0.129	t = 0.129
Observations	131	131
Log Likelihood	283.944	285.045
sigma2	0.001	0.001
Akaike Inf. Crit.	-557.889	-562.090

Note: t = T-statistic value = coeff/SE(coeff)

Fuente: Elaboración propia

En primer lugar, se observa que el intercepto de ambos modelos son no significativos, por lo cual, se habría de modelizar con la serie original aplicando la transformación e introducir la diferenciación regular junto con otros componentes del modelo usando la función $ARIMA$. Además del intercepto, se ve que, para el modelo $MA(4)$, solo el parámetro θ_1 es significativo. Por este motivo, según el principio de parsimonia, se debe reducir el modelo a un $MA(1)$ porque no se han considerado necesarios los otros parámetros para explicar la evolución de los datos a lo largo del periodo estudiado.

Con la figura 5.9 se mostrará con el mismo formato de la tabla anterior pero modelizando con la serie X_t.

Figura 5.9: Tabla resumen de los modelos de la serie gasolina

Resultados

	Dependent variable:	
	lngas	
	MA(4) (1)	AR(2) (2)
ma1	0.576	
	t = 6.703	
ma2	-0.036	
	t = -0.370	
ma3	-0.284	
	t = -2.902	
ma4	-0.198	
	t = -2.338	
ar1		0.631
		t = 7.869
ar2		-0.402
		t = -5.000
Observations	131	131
Log Likelihood	310.915	310.688
sigma2	0.001	0.001
Akaike Inf. Crit.	-611.830	-615.377

Note: t = T-statistic value = coeff/SE(coeff)

Fuente: Elaboración propia

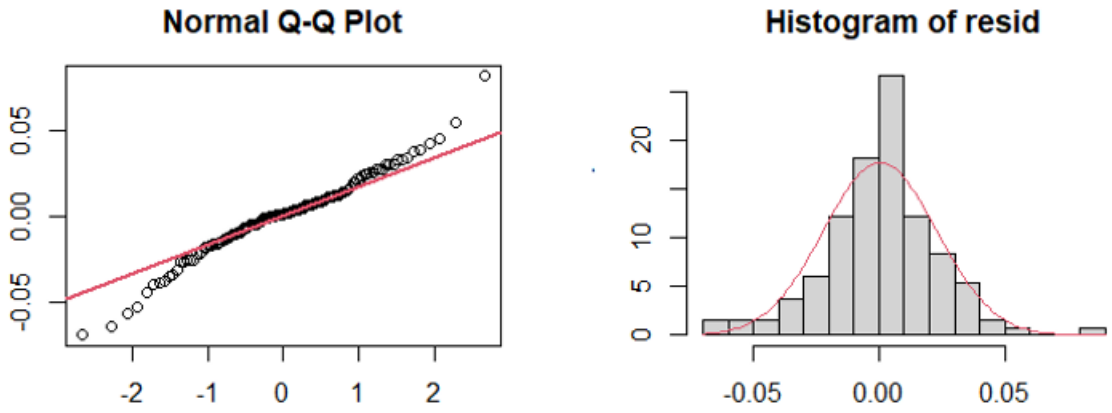
Con estos nuevos resultados se puede comprobar que todos los parámetros son significativos, excepto por θ_2 para el modelo $MA(4)$. De todas formas, debido a que los parámetros θ_3 y θ_4 son significativos, no se aplicará el principio de parsimonia reduciendo el modelo.

Comparando los resultados de ambos modelos, concretamente en la columna 4, el modelo $AR(2)$ presenta menor valor del AIC con un valor de -615.377 indicando que este modelo explica los datos mejor que el otro modelo propuesto para esta serie temporal. Por lo tanto, se decide continuar con el proceso definido en la metodología y realizar la siguiente etapa con el modelo $AR(2)$.

5.4. Validación

En esta tercera etapa, se tendrá que validar el modelo analizando la normalidad, la variabilidad y la independencia de los residuos del modelo.

Figura 5.10: QQplot e Histograma de residuos



Fuente: Elaboración propia

Para iniciar con la validación de la normalidad de los residuos, se muestra el *qqplot* y el histograma en la figura 5.10. En ambos gráficos de la imagen se observa la presencia de valores atípicos. La presencia de estos valores inusuales pueden suponer en un error en los contrastes numéricos para la verificación de la normalidad, por ello, se debe tener en mayor cuenta que los datos siguen bastante bien la distribución normal de manera visual.

Una de las opciones para determinar la normalidad de los datos es llevar a cabo las tres pruebas para la normalidad *Shapiro-Wilk Normality test*, *Anderson-Darling test* y *Jarque-Bera test*, con las hipótesis:

H_0 :los residuos provienen de una población normalmente distribuida

H_1 :los residuos no provienen de una población normalmente distribuida

Cuadro 5.1: Resultados de las pruebas de normalidad

Prueba	**Estadístico**	**P-valor**
Shapiro-Wilk Normality	$W = 0{,}97229$	0.0084
Anderson-Darling	$A = 1{,}2012$	0.0038
Jarque-Bera	$\chi^2 = 14{,}527$	0.0007

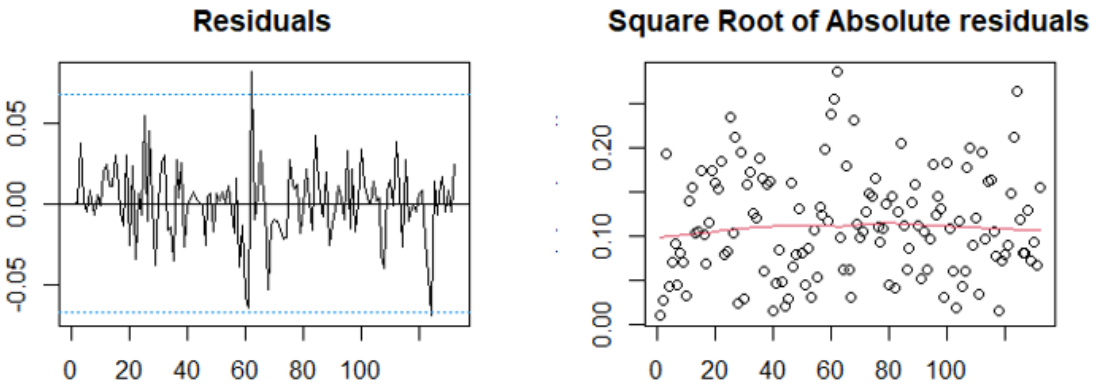
Fuente: Elaboración propia

Se recogen en el cuadro 5.1 los resultados de cada una de las pruebas numéricas. En todas ellas se ha obtenido unos *p-valores* menores que $\alpha = 0{,}05$ y de este modo, rechazando la hipótesis nula del contraste.

Cabe destacar que aún teniendo el rechazo de las pruebas numéricas, tal y como se ha mencionado en el apartado correspondiente en la metodología, los rechazos de las pruebas numéricas son usuales con la presencia de valores atípicos. Por este motivo, según el criterio analítico y los gráficos observados, se concluirá que el precio de la gasolina presenta normalidad en su término estocástico del modelo explicativo.

Siguiendo con la validación de las principales características que debe presentar el modelo, se estudiará la homogeneidad de los residuos.

Figura 5.11: Gráfica de residuos y raíz cuadrada de residuos absoluto



Fuente: Elaboración propia

En la figura 5.11, se muestra primero la gráfica de la distribución de los residuos y al lado la de la raíz cuadrada de los residuos al cuadrado. Con el primer gráfico, el *residual plot*, se observa que los residuos varían levemente y de manera bastante similar. Se puede destacar unos puntos concretos que sobresalen de las bandas de significancia y probablemente corresponden al valor atípico visto anteriormente. Finalmente, con la gráfica de la raíz cuadrada de los residuos absolutos, se observa que los datos presentan una varianza bastante constante.

Además de las gráficas, también se ha aplicado la prueba de *Breusch-Bagan* para estudiar la homoscedasticidad definiendo como hipótesis:

H_0 :Los residuos tienen varianza constante

H_1 :Los residuos no tienen varianza constante

Figura 5.12: Resultado de la prueba Breusch-Pagan

```
        studentized Breusch-Pagan test

data:  resid(model) ~ I(obs - resid(model))
BP = 0.70547, df = 1, p-value = 0.401
```

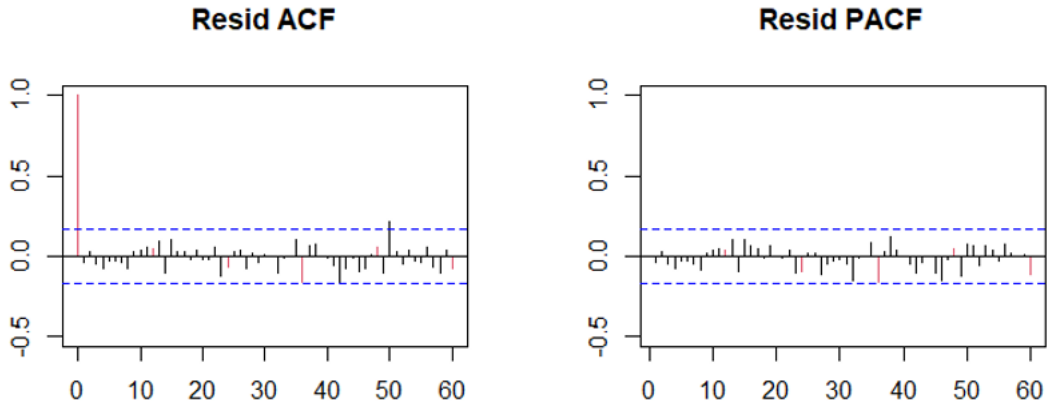
Fuente: Elaboración propia

La figura 5.12 presenta los resultados de la prueba de *Breusch-Pagan*. En ella se

puede ver como el estadístico BP asume un valor de 0.70547 con su correspondiente *p-valor* de 0.401 que es mayor que el nivel de significación $\alpha = 0{,}05$. De esta forma se concluye que los residuos tienen varianza constante, aceptando así, la hipótesis nula.

Para finalizar con esta última etapa del proceso, se estudiará la independencia de los residuos mediante el análisis de los gráficos ACF y $PACF$.

Figura 5.13: Gráficos para el ACF y el PACF de los residuos



Fuente: Elaboración propia

En la imagen 5.13 se percibe que todos los retardos se encuentran dentro del intervalo de confianza al 95 % de significación, excepto el retardo 50 de la gráfica de ACF. Este puede ser debido tanto por el valor atípico que se ha contemplado al inicio del estudio como por el 5 % de probabilidad a que algún retardo se encuentre por fuera de los límites de confianza.

Si se quiere contrastar o aplicar algún otro método para verificar la conclusión que se ha llegado mediante los gráficos, se puede llevar a cabo la prueba de *Durbin-Watson* la cual es un contraste de hipótesis numérico definido como:

H_0 :Los datos se distribuyen de forma independiente

H_1 :Los datos no se distribuyen de forma independiente

Figura 5.14: Resultado de la prueba de Durbin-Watson

```
        Durbin-Watson test

data:  resid(model) ~ I(1:length(resid(model)))
DW = 2.0743, p-value = 0.6335
alternative hypothesis: true autocorrelation is greater than 0
```

Fuente: Elaboración propia

En la imagen 5.14 se puede ver el resultado del contraste y como este ha devuelto un estadístico DW de 2.0743 con su *p-valor* correspondiente de 0.6335. Al ser el *p-*

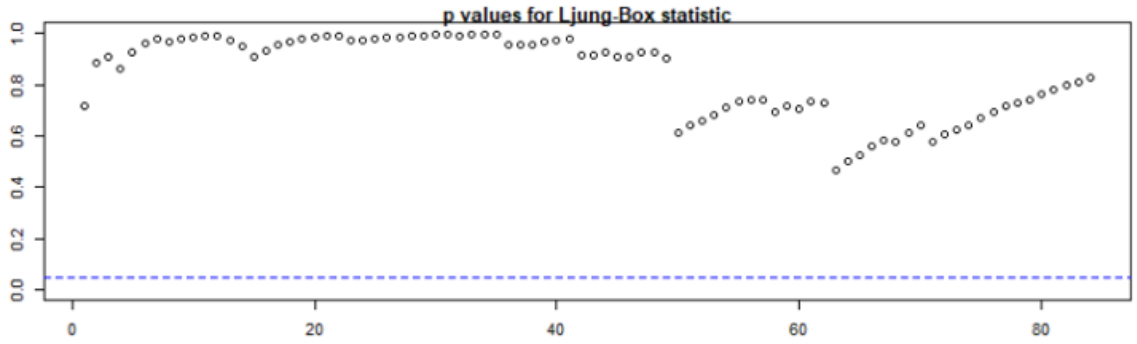
valor mayor que el nivel de significación al 95 %, no hay evidencias estadísticas significativas para rechazar la hipótesis nula y se concluye que no existe correlación entre los residuos.

Además de la prueba de *Durbin-Watson*, para verificar que el resultado obtenido es correcto, se ha decidido aplicar la prueba de *Ljung-Box* definiendo las hipótesis como:

H_0 :Las autocorrelaciones hasta retardo k son iguales a cero

H_1 :Las autocorrelaciones hasta retardo k no son iguales a cero

Figura 5.15: Gráfica de p-valores para la prueba de Ljung-Box



Fuente: Elaboración propia

Graficando hasta $k = 84$ retardos los resultados de la prueba, como se puede ver en la figura 5.15, se ve que todos los p-valores se sitúan por encima de la línea que representa el nivel de significación de $\alpha = 0{,}05$. Por lo tanto, se puede concluir que las autocorrelaciones hasta el retardo 84 son iguales a cero, en otras palabras, los retardos son independientes.

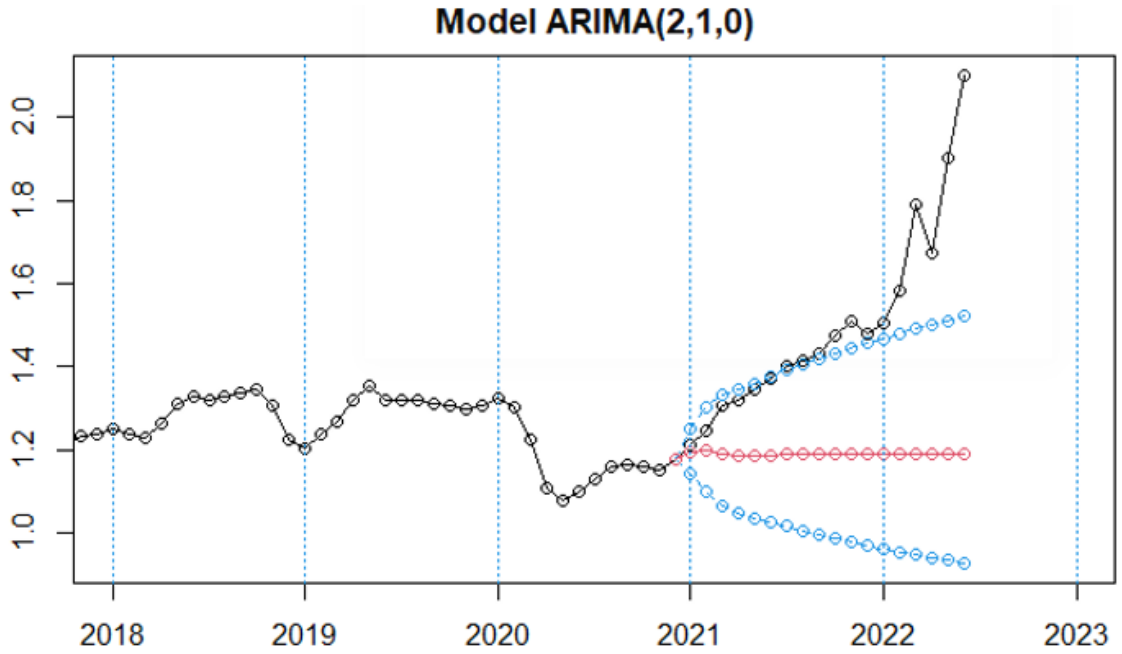
Concluyendo todas las pruebas y gráficas, se puede decir que el modelo $ARIMA(2, 1, 0)$ está validado para explicar los datos.

5.5. Predicción

Teniendo en cuenta que el objetivo de este proyecto es comparar los precios de los diferentes carburantes si no se hubieran presenciado los diferentes factores que han hecho cambiar de manera brusca el precio como se ha mencionado en la introducción, se realizará en este apartado la predicción *in-sample*. Esto trata de coger un subconjunto de datos, en este caso, los datos del año 2010 hasta 2020 incluido para pronosticar los siguientes valores y así, compararlos con los resultados reales.

A continuación, en el gráfico de la figura 5.16, se puede ver la comparación de los valores reales con los valores predichos. En este caso, los valores predichos representan el precio de la gasolina si todos los factores externos a la evolución natural del precio de la gasolina no hubieran ocurrido.

Figura 5.16: Comparación de valores reales y predichos



Fuente: Elaboración propia

Los puntos negros del gráfico son los datos reales del precio de gasolina súper 95. Los puntos rojos hacen referencia a las predicciones puntuales aplicando el modelo $ARIMA(2,1,0)$ y los puntos azules representan las bandas de confianza con un nivel de significación del 95 %.

De esta forma, mediante los indicadores presentados en la metodología, este modelo presenta un $RMSPE$ de 0.22 y un $MAPE$ de 0.19. Es decir, teniendo en cuenta que estos indicadores son mejores cuando más se acerquen a cero, podemos decir que el modelo no hace una buena predicción a corto plazo. Con la figura 5.16 se verifica como las estimaciones del modelo se encuentran dentro de las bandas de confianza en los primeros meses, pero a causa del cambio estructural, el modelo planteado no tiene la capacidad de predecir correctamente a largo plazo.

Finalmente, se muestra una tabla numérica con los valores reales, predichos, la diferencia entre los dos precios y la tasa de variación calculada de la siguiente forma:

$$Diferencia = x_{obs} - x_{pred}$$

$$TV = (\frac{x_{obs} - x_{pred}}{x_{pred}}) * 100$$

Cuadro 5.2: Comparación de valores reales y predichos, la diferencia y la tasa de variación

Año	**Mes**	**Valores reales**	**Valores predichos**	**Diferencia**	**Tasa de variación**
2021	Enero	1.213667	1.195506	0.018161	1.5 %
2021	Febrero	1.246250	1.197293	0.048957	4.1 %
2021	Marzo	1.305200	1.190937	0.114263	9.6 %
2021	Abril	1.320333	1.186234	0.134099	11.3 %
2021	Mayo	1.345800	1.185810	0.159990	13.5 %
2021	Junio	1.370750	1.187429	0.183321	15.4 %
2021	Julio	1.402250	1.188621	0.213629	18 %
2021	Agosto	1.416600	1.188722	0.227878	19.2 %
2021	Setiembre	1.431000	1.188307	0.242693	20.4 %
2021	Octubre	1.474000	1.188004	0.285996	24.1 %
2021	Noviembre	1.510200	1.187980	0.322220	27.1 %
2021	Diciembre	1.481333	1.188087	0.293247	24.7 %
2022	Enero	1.506000	1.188163	0.317837	26.8 %
2022	Febrero	1.583750	1.188169	0.395581	33.3 %
2022	Marzo	1.789000	1.188142	0.600858	50.6 %
2022	Abril	1.673667	1.188122	0.485544	40.9 %
2022	Mayo	1.904200	1.188121	0.716079	60.3 %
2022	Junio	2.102667	1.188128	0.914539	77 %

Fuente: Elaboración propia

En el cuadro 5.2 se pueden ver las diferencias entre los cambios pronunciados del precio de la gasolina y los valores que según el modelo creado deberían haber sucedido. Por ejemplo, con los valores obtenidos en junio del año 2022, se considera que, se ha incrementado casi un 0.91 euros por litro de gasolina. Es decir, si todos los factores externos se hubieran mantenido estables sin afectar a la evolución del precio y viendo que ha habido una tasa de variación del 77 % esto significa que, repostando un depósito de 50 litros se está pagando unos 45 euros de más.

Capítulo 6

Resumen de los análisis e interpretación

En este capítulo se recogerán los principales resultados obtenidos por los análisis de los carburantes estudiados y se llevará a cabo su interpretación según los objetivos planteados al inicio del proyecto.

6.1. Resultados de los análisis

De la misma forma que se ha visto el proceso de análisis *ARIMA* para la gasolina, en los anexos A y B se encuentra el proceso analítico según la metodología *Box-Jenkins* para los otros carburantes. En cada uno de los apartados se identifican diferentes modelos y se selecciona el óptimo teniendo en cuenta el indicador AIC, si las estimaciones son significativas y si finalmente se valida.

Teniendo en cuenta estos procesos, se han decidido como modelos óptimos para explicar la evolución del precio en el periodo estudiado de los carburantes los presentados en el cuadro 6.1.

Llevada a cabo la identificación de los mejores modelos para cada uno de los combustibles, se ha hecho la estimación de los parámetros que constituyen el modelo tanto en el análisis dentro del informe como en los anexos correspondientes a cada uno de los combustibles estudiados. De esta forma, se ha podido hacer la última fase de la metodología correspondiente a la validación de los modelos según las premisas definidas en la metodología. Los términos de perturbación de cada uno de los modelos deben cumplir un seguido de requisitos para contemplar el modelo como válido.

Cuadro 6.1: Identificación de los modelos óptimos para los diferentes carburantes

Gasolina	$ARIMA(2,1,0)$ para X_t
Forma compacta del modelo *ARIMA* para $ln(X_t)$ con $d=1$	$\phi_2(B)(1-B)^d ln(X_t) = Z_t$
Sustituyendo cada uno de los polinomios característicos se obtiene:	$(1-\phi_1 B-\phi_2 B^2)(1-B)ln(X_t) = Z_t$
Gasoil	$ARIMA(0,1,2)$ para X_t
Forma compacta del modelo *ARIMA* para $ln(X_t)$ con $d=1$	$\phi_2(B)(1-B)^d ln(X_t) = Z_t$
Sustituyendo cada uno de los polinomios característicos se obtiene:	$(1-\phi_1 B-\phi_2 B^2)(1-B)ln(X_t) = Z_t$
Diésel para la calefacción	$ARIMA(0,1,1)$ para X_t
Forma compacta del modelo *ARIMA* para $ln(X_t)$ con $d=1$	$\phi_2(B)(1-B)^d ln(X_t) = Z_t$
Sustituyendo cada uno de los polinomios característicos se obtiene:	$(1-\phi_1 B-\phi_2 B^2)(1-B)ln(X_t) = Z_t$

Fuente: Elaboración propia

Para la gasolina, como se ha visto en el capítulo anterior, se han podido dar por validadas las diferentes características del término de perturbación. Es decir, el residuo del modelo estaba distribuido según la normalidad de manera homogénea e independiente.

Del mismo modo, se puede consultar el anexo A donde se lleva a cabo paso por paso el análisis del diésel, concluyendo también, que las diferentes premisas para el término de perturbación se validan satisfactoriamente.

En cambio, para el diésel para la calefacción, no se ha encontrado un modelo que valide todas las características necesarias en el residuo del modelo. Por ello, se ha decidido seguir trabajando con el modelo propuesto debido a su optimalidad explicando la evolución del precio aunque no cumpla la normalidad y homogeneidad.

6.2. Interpretación de las predicciones

Finalmente, en esta sección se dará respuesta al principal objetivo de este proyecto, se analizará e interpretará cómo debería haber evolucionado el precio en estos últimos meses donde se han registrado precios altos de manera inesperada y brusca por causa de factores externos.

6.2.1. Gasolina

Para la gasolina, tal y como se ha visto dentro del apartado de análisis, los valores predichos por el modelo se encuentran entre el 1.18 euros y 1.19 euros. Es

decir, el precio de la gasolina no hubiera incrementado de manera drástica si no se hubieran presenciado factores socio-económicos significativos que rompieron la serie temporal haciéndola cambiar de nivel rápidamente. Los valores reales han pasado de 1.21 euros en enero del 2021 a crecer hasta 2.10 euros en junio del 2022.

Cuadro 6.2: Comparación de valores reales y predichos, la diferencia y la tasa de variación

Año	**Mes**	**Valores reales**	**Valores predichos**	**Diferencia**	**Tasa de variación**
2021	Enero	1.213667	1.195506	0.018161	1.5 %
2021	Febrero	1.246250	1.197293	0.048957	4.1 %
2021	Marzo	1.305200	1.190937	0.114263	9.6 %
2021	Abril	1.320333	1.186234	0.134099	11.3 %
2021	Mayo	1.345800	1.185810	0.159990	13.5 %
2021	Junio	1.370750	1.187429	0.183321	15.4 %
2021	Julio	1.402250	1.188621	0.213629	18 %
2021	Agosto	1.416600	1.188722	0.227878	19.2 %
2021	Setiembre	1.431000	1.188307	0.242693	20.4 %
2021	Octubre	1.474000	1.188004	0.285996	24.1 %
2021	Noviembre	1.510200	1.187980	0.322220	27.1 %
2021	Diciembre	1.481333	1.188087	0.293247	24.7 %
2022	Enero	1.506000	1.188163	0.317837	26.8 %
2022	Febrero	1.583750	1.188169	0.395581	33.3 %
2022	Marzo	1.789000	1.188142	0.600858	50.6 %
2022	Abril	1.673667	1.188122	0.485544	40.9 %
2022	Mayo	1.904200	1.188121	0.716079	60.3 %
2022	Junio	2.102667	1.188128	0.914539	77 %

Fuente: Elaboración propia

Con el cuadro 6.2 se puede ver como la diferencia entre los valores observados por la situación actual y los valores predichos (si no se hubieran presentado estos factores externos) va aumentando considerablemente durante el periodo de estudio. Acaba llegando a una diferencia de un punto aproximadamente, es decir, casi un euro de diferencia por litro.

Como se ha podido ver en los diferentes medios de comunicación, se ha comprobado de manera estadística mediante este análisis como la tasa de variación ha aumentado hasta llegar a pagar un 77 % más por la gasolina. Por ejemplo, en junio del 2022 sabiendo que la diferencia es de 0.91 euros y la tasa de variación es de un 77 %, repostando un coche que tenga un depósito de 50 litros, se ha estado pagando 45 euros más si no se hubiera presenciado la inflación de los precios.

6.2.2. Gasoil

Para el gasoil, observando el análisis presentado en el anexo A, los valores predichos por el modelo se encuentran entre el 1.07 euros y 1.08 euros. Es decir, el precio del diésel no habría fluctuado creciendo de manera pronunciada sin presentarse factores socio-económicos significativos externos que provocaron un cambio estructural

a la serie temporal del carburante durante el periodo de estudio. Los valores reales han pasado de 1.09 euros en enero del 2021 a crecer hasta 2 euros en junio del 2022.

Cuadro 6.3: Comparación de valores reales y predichos, la diferencia y la tasa de variación

Año	**Mes**	**Valores reales**	**Valores predichos**	**Diferencia**	**Tasa de variación**
2021	Enero	1.097000	1.078091	0.018909	1.8 %
2021	Febrero	1.128000	1.081131	0.046869	4.3 %
2021	Marzo	1.177600	1.081131	0.096469	8.9 %
2021	Abril	1.180667	1.081131	0.099536	9.2 %
2021	Mayo	1.203800	1.081131	0.122669	11.3 %
2021	Junio	1.234500	1.081131	0.153369	14.2 %
2021	Julio	1.262250	1.081131	0.181119	16.8 %
2021	Agosto	1.265000	1.081131	0.183869	17 %
2021	Setiembre	1.277000	1.081131	0.195869	18.1 %
2021	Octubre	1.344000	1.081131	0.262869	24.3 %
2021	Noviembre	1.382000	1.081131	0.300869	27.8 %
2021	Diciembre	1.348667	1.081131	0.267536	24.7 %
2022	Enero	1.382800	1.081131	0.301669	27.9 %
2022	Febrero	1.470750	1.081131	0.389619	36 %
2022	Marzo	1.758250	1.081131	0.677119	62.6 %
2022	Abril	1.702667	1.081131	0.621536	57.5 %
2022	Mayo	1.878000	1.081131	0.796869	73.7 %
2022	Junio	1.999333	1.081131	0.918203	84.9 %

Fuente: Elaboración propia

Con el cuadro 6.3 se puede apreciar como la diferencia entre los valores observados por la situación actual y los valores predichos si no se hubieran presentado estos factores externos han crecido bastante durante los meses que se han considerado para el análisis. La diferencia empezó siendo de un céntimo hasta llegar en el último mes a los 0.92 euros de diferencia por cada litro comprado.

Como se ha podido ver en los diferentes medios de comunicación, se ha comprobado de manera estadística mediante este análisis como la tasa de variación del precio ha crecido superando más del 50 % a partir de marzo del 2022 y llegando a registrar un porcentaje de 85 % en el último mes por el gasoil. Esto indica que, por ejemplo, en junio del 2022 sabiendo que la diferencia es de 0.92 euros y la tasa de variación es de un 85 %, repostando un coche que tenga un depósito de 50 litros, se ha estado pagando 45 euros más si no se hubiera presenciado la inflación de los precios en estos últimos meses.

6.2.3. Diésel para la calefacción

Finalmente, para acabar con las interpretaciones de los análisis de los diferentes carburantes se ha llevado a cabo la lectura de los resultados para el diésel para la calefacción. Si se consulta el análisis presentado en el anexo B, los valores predichos por el modelo se encuentran alrededor del 0.59 euros. Es decir, el precio del diésel

para la calefacción no habría cambiado si no se hubieran presentado factores socioeconómicos externos a la evolución del precio que provocaron un cambio estructural y cambiando así, el nivel presente en el combustible. Los valores reales han pasado de 0.61 euros en enero del 2021 a crecer hasta 1.52 euros en junio del 2022.

Cuadro 6.4: Comparación de valores reales y predichos, la diferencia y la tasa de variación

Año	**Mes**	**Valores reales**	**Valores predichos**	**Diferencia**	**Tasa de variación**
2021	Enero	0.6073333	0.5904470	0.016886	2.9 %
2021	Febrero	0.6345000	0.5904470	0.044053	7.5 %
2021	Marzo	0.6776000	0.5904470	0.087153	14.8 %
2021	Abril	0.6706667	0.5904470	0.080220	13.6 %
2021	Mayo	0.6866000	0.5904470	0.096153	16.3 %
2021	Junio	0.7147500	0.5904470	0.124303	21.1 %
2021	Julio	0.7357500	0.5904470	0.145303	24.6 %
2021	Agosto	0.7378000	0.5904470	0.147353	25 %
2021	Setiembre	0.7535000	0.5904470	0.163053	27.6 %
2021	Octubre	0.8287500	0.5904470	0.238303	40.4 %
2021	Noviembre	0.8550000	0.5904470	0.264553	44.8 %
2021	Diciembre	0.8183333	0.5904470	0.227886	38.6 %
2022	Enero	0.8762000	0.5904470	0.285753	48.4 %
2022	Febrero	0.9682500	0.5904470	0.377803	64 %
2022	Marzo	1.3245000	0.5904470	0.734053	124.3 %
2022	Abril	1.3346667	0.5904470	0.744220	126 %
2022	Mayo	1.3610000	0.5904470	0.770553	130 %
2022	Junio	1.5163333	0.5904470	0.925886	156.8 %

Fuente: Elaboración propia

Con el cuadro 6.4 se puede apreciar como la diferencia entre los valores observados por la situación actual y los valores predichos si no se hubieran presentado estos factores externos han aumentado la diferencia significativamente durante los últimos meses considerados en el análisis. La diferencia empezó siendo de 0.61 euros hasta llegar en el último mes a 1.52 euros de diferencia por cada litro comprado.

En este caso, se ha comprobado de manera estadística con el análisis temporal del carburante como la tasa de variación del precio ha crecido superando más del 50 % en febrero del 2022 y llegando a registrar un porcentaje de 157 % en el último mes por el diésel para la calefacción. Con todos estos análisis e indicadores se puede determinar, por ejemplo, en junio del 2022 la diferencia ha sido de 1.52 euros y la tasa de variación con un porcentaje del 157 %. Es decir, teniendo en cuenta que el gasto del diésel para calefacción depende estrictamente del uso, si se enciende la calefacción en el hogar durante 10 horas al día, se consumiría aproximadamente 370 litros de gasoil mensualmente. En términos de dinero equivaldría a que se ha estado pagando en un caso parecido al ejemplo 560 euros al mes, pero si no se hubiera presentado esta inflación de los precios solo se pagaría 218 euros aproximadamente con el precio predicho.

Capítulo 7

Conclusiones

Para finalizar con el informe para el análisis de los principales carburantes de uso diario en España, hay que destacar que se han cumplido los objetivos presentados al inicio.

En primer lugar, haciendo uso del proceso estadístico, se ha llevado a cabo un análisis temporal para cada una de las series constituidas por los datos registrados del precio de la gasolina, el diésel y el diésel para la calefacción. Estos tres combustibles son los más habituales en el uso diario entre la población española.

Teniendo en cuenta que el objetivo principal de este estudio es conocer las diferencias de los precios en los últimos meses registrados, afectados por diversos factores socio-económicos y cómo hubieran evolucionado los precios si no se hubiera roto la evolución mediante un cambio estructural, se ha llevado a cabo con la metodología *Box-jenkins* un análisis para identificar la gravedad de los cambios ocurridos.

Iniciando con la etapa de recogida de datos según el proceso estadístico, se ha decidido ir más allá de lo aprendido durante los estudios e investigar la manera de llevar a cabo una lectura de datos oficiales de manera automática y óptima según los métodos informáticos presentes en el *software R*. En este caso, al seleccionar unos datos que se encontraban tabulados según el lenguaje *HTML*, se ha decidido implementar un código basado en la teoría del *web scraping*. En otras palabras, se ha identificado el código con menor coste computacional para leer los datos de manera rápida y automáticamente de la página donde están publicados.

En segundo lugar, entrando en la etapa de organización de datos, se ha llevado a cabo un análisis descriptivo de los diferentes carburantes resultando con la identificación de media y varianza no constantes y sin presencia de estacionalidad en las series temporales. Además, cabe destacar un estudio adicional mediante la prueba

de *Chow* del cambio estructural sufrido en el periodo de estudio. Con este test se ha podido concluir que efectivamente, la evolución del precio ha sufrido un cambio estructural en los últimos meses haciendo que el nivel de la serie incrementara de manera rápida.

Para obtener los valores predichos del precio sin alterarse por factores externos de cada uno de los combustibles estudiados, se ha hecho un análisis para determinar cuál es el mejor modelo para explicar la evolución de los precios. De este modo, se ha obtenido que los modelos óptimos son el $ARIMA(2,1,0)$ para la gasolina, el $ARIMA(0,1,2)$ para el diésel y el $ARIMA(0,1,1)$ para el diésel para la calefacción.

Con estos modelos se ha podido calcular la tasa de variación para cada mes de los diferentes carburantes resultando con un 77 % de variación en junio del 2022 para la gasolina, un 85 % para el gasoil y un 157 % para el diésel para la calefacción. Esto a efectos prácticos ha supuesto a la población española pagar 45 euros más por repostar un coche con el depósito de 50 litros en junio, tanto para gasolina como para el gasoil. Además, para la calefacción se ha pagado aproximadamente 300 euros más por consumirla durante 10 horas en un hogar gastando 370 litros en este tiempo.

Esto muestra como para un ciudadano con base salarial mínima de 950 euros al mes, en caso de consumir alguna de estas materias primas para el coche y así ir a trabajar o para otras acciones, deben abonar un gran porcentaje para cubrir estos gastos, dejando una sensación de pobreza en la población. En caso de no estabilizarse este crecimiento, en una época más fría y con necesidad de calefacción en los pueblos que acostumbran registrar bajas temperaturas, la población española con un base salarial mínima de 950 euros al mes podrían llegar a gastar 300 euros de más aproximadamente solo para mantener la calefacción encendida durante 10 horas. Con este proyecto, ya detectado el problema y la gran inflación de los precios, deben plantearse medidas preventivas y resolutivas de temática social y económica para hacer frente a la nueva situación que se está viviendo.

Para acabar con este proyecto, no solo se ha conseguido mostrar estos resultados tan impactantes por lo que hace a la evolución de los precios de los carburantes, sino que también este proyecto ha sido una ocasión para llevar a cabo una investigación con una temática distinta a las que he llegado a ver durante mis estudios, he necesitado informarme más en indicadores económicos, estar actualizada en el mundo económico y además, iniciar un nuevo hábito para informarme y coger facilidad en la búsqueda de información relevante para este caso.

Durante mis estudios he tenido más facilidad o he encontrado más de mi agrado investigar y programar análisis que tuvieran un coste computacional bajo. Con este

proyecto, he intentado no solo dejarme guiar por mis instintos de programar, sino dedicarle más tiempo a la investigación de artículos científicos y estadísticos, noticias económicas y comprender los diferentes procesos que me ayudan a dar solución a mis hipótesis y objetivos. Aun así, también he querido dedicar un tiempo para aprender el lenguaje *LaTeX*, ya que es muy útil para la redacción de artículos o proyectos científicos como este.

Finalmente, este ha sido un proyecto con interés más social y económico que me a brindado una gran satisfacción relacionarlo con mis estudios y demostrarme que con lo aprendido a lo largo del grado puedo dar respuesta a preguntas muy diversas y de ámbitos muy diferentes.

Bibliografía

¿Por qué sube el precio de la gasolina y el diésel? — Noticias Coches.net, [s.f.] **online**[visitado 2022-07-22]. Disp. desde: https://www.coches.net/noticias/porque-sube-el-precio-de-la-gasolina-y-el-diesel.

BAMBENEK, John y KLUS, Agnieszka, 2009. *grep Pocket Reference: A Quick Pocket Reference for a Utility Every Unix User Needs*. O'Reilly Media.

BECKER, Richard, 2018. *The new S language*. CRC Press.

Euros a Dólares estadounidenses — Convierta 1·EUR a USD — Xe, [s.f.] **online**[visitado 2022-07-28]. Disp. desde: https://www.xe.com/es/currencyconverter/convert/?Amount=1&From=EUR&To=USD.

Medidas restrictivas de la UE contra Rusia por sus actos en Ucrania (desde 2014) - Consilium, [s.f.] **online**[visitado 2022-07-14]. Disp. desde: https://www.consilium.europa.eu/es/policies/sanctions/restrictive-measures-against-russia-over-ukraine/.

PANKRATZ, Alan, 2009. *Forecasting with univariate Box-Jenkins models: Concepts and cases*. John Wiley & Sons.

PETERSON, Leif E, 2009. K-nearest neighbor. *Scholarpedia*. Vol. 4, n.º 2, pág. 1883.

Precios de los derivados del petróleo: España 2022 — Datosmacro.com, [s.f.] **online**[visitado 2022-07-25]. Disp. desde: https://datosmacro.expansion.com/energia/precios-gasolina-diesel-calefaccion/espana?anio.

Proceso estadístico - Qué es, definición y concepto — 2022 — Economipedia, [s.f.] **online**[visitado 2022-06-23]. Disp. desde: https://economipedia.com/definiciones/proceso-estadistico.html.

RPubs - Imputación de datos, [s.f.] **online**[visitado 2022-07-25]. Disp. desde: https://rpubs.com/ydmarinb/429757.

SÁNCHEZ, Paola Andrea, 2008. Cambios estructurales en series de tiempo: una revisión del estado del arte. *Revista Ingenierías Universidad de Medellín*. Vol. 7, n.º 12, págs. 115-140.

Servicios De Web Scraping: Cómo Comenzó y Qué Sucederá en El Futuro — Octoparse, [s.f.] **online**[visitado 2022-06-23]. Disp. desde: https://www.octoparse.es/blog/como-comenzo-y-sucedera-en-futuro#div3.

Todo lo que debes saber sobre el precio de los carburantes · AOP, [s.f.] **online**[visitado 2022-08-01]. Disp. desde: `https://www.aop.es/blog/2022/04/13/preguntas-respuestas-precio-carburantes/`.

VILLAVICENCIO, Jhon, 2010. Introducción a series de tiempo. *Puerto Rico*.

ZEILEIS, Achim; LEISCH, Friedrich; HORNIK, Kurt y KLEIBER, Christian, 2002. strucchange: An R package for testing for structural change in linear regression models. *Journal of statistical software*. Vol. 7, págs. 1-38.

Apéndice A

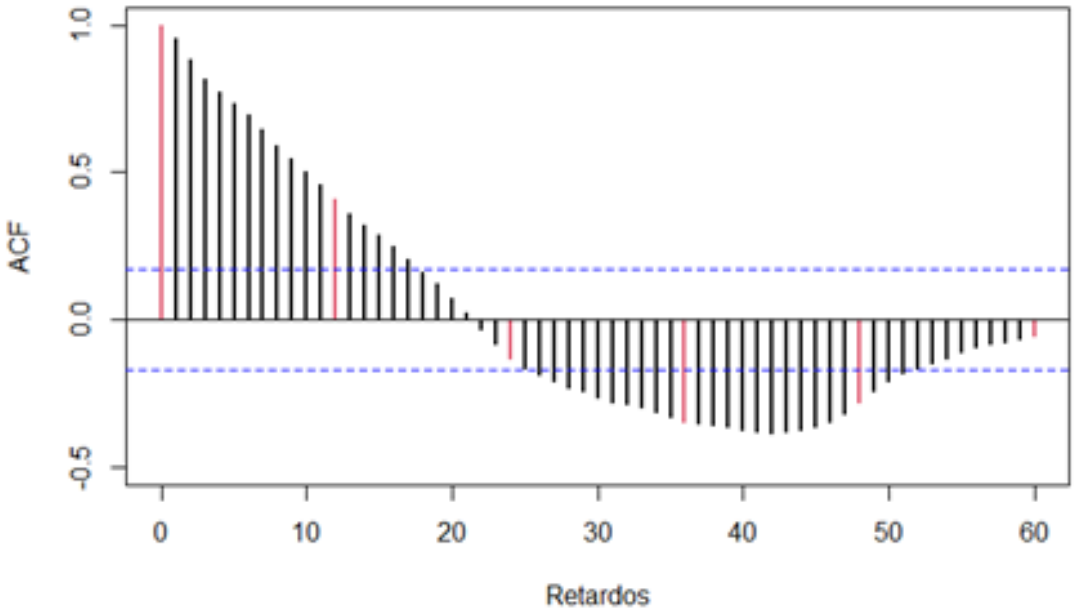
Análisis ARIMA: Gasoil

En este capítulo, se aplicará la metodología *ARIMA* para encontrar un modelo adecuado para la serie del gasoil.

A.1. Transformaciones previas

Antes de empezar a identificar los posibles modelos, es necesario estudiar la estacionariedad de la serie.

Figura A.1: Gráfica ACF de la serie diésel



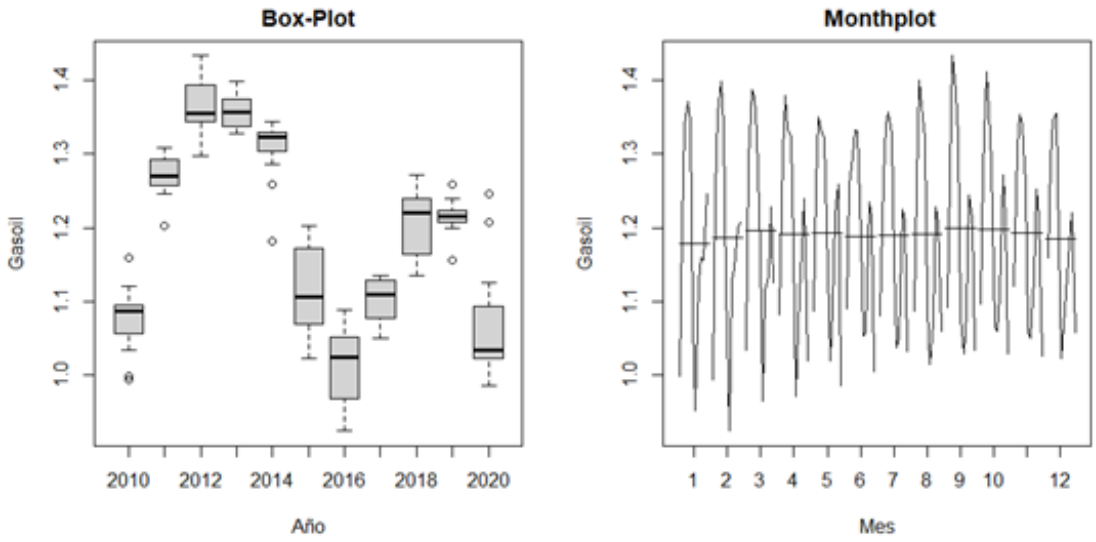
Fuente: Elaboración propia

En la gráfica que se presenta en la figura A.1 se observa que los retardos de *ACF* se decrecen muy lentamente hacia cero, con lo cual se representa que la serie

no es estacionaria.

A continuación, se llevará a cabo los análisis sobre la media y la varianza se la serie, para estudiar las transformaciones necesarias para conseguir una serie estacionaria.

Figura A.2: Box-plot y gráfica de Monthplot



Fuente: Elaboración propia

En la figura A.2 se representan los gráficos *box-plot* para analizar la estabilidad de la media y de la varianza de, y también el *Monthplot* para estudiar la estacionaliedad de la serie diésel.

Analizando las líneas negras dentro de cada caja, que representa la media de cada uno de los años estudiados, y se concluye que, las líneas no se mantienen en el mismo nivel y por lo tanto, la media no es constante. Por otra parte, se observa que la amplitud de caja de cada año son diferentes, con lo cual, se puede decir que la varianza tampoco es constante.

Observando el *monthplot*, las diferencias de medias entre meses no son considerables, y se concluye que la serie no hay componente estacional.

Resumiendo los análisis previos, se ha de aplicar una transformación logarítmica y una diferenciación regular sobre la serie original para conseguir una serie estacionaria.

A continuación, se muestra la en la imagen A.3la gráfica ACF de la serie transformada.

Figura A.3: Gráfica ACF de la serie gasoil transformada

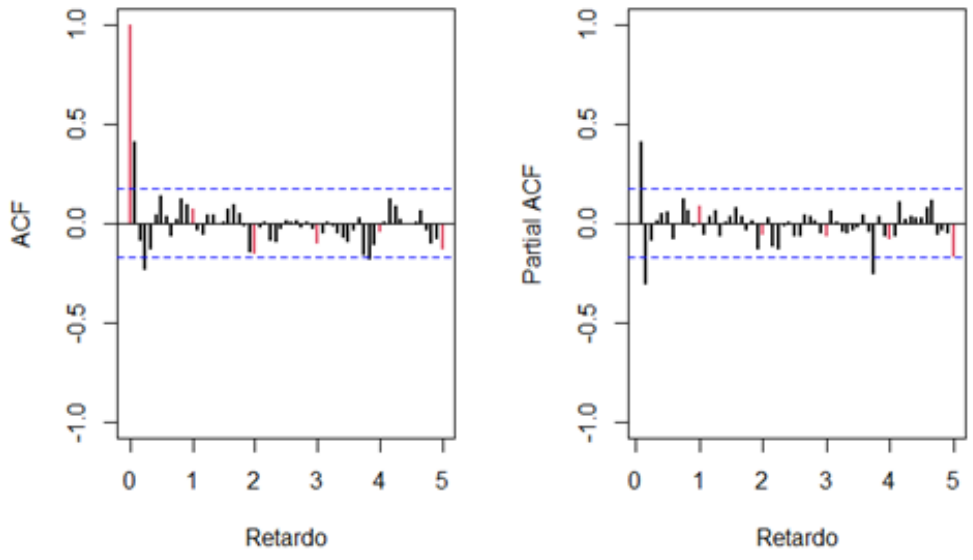
Fuente: Elaboración propia

Se observa que los retardos decaen rápidamente hacia 0, por lo tanto, se puede decir que la serie ya es estacionaria.

A.2. Identificación

Una vez asegurado la estacionariedad de la serie, se llevará a cabo la identificación de los posibles modelos que se puede ajustar en la serie a través de la gráfica ACF y $PACF$. Para una mejor identificación, se muestra en color rojo los retardos múltiples de 12.

Figura A.4: Gráfica ACF y PACF de la serie gasoil transformada



Fuente: Elaboración propia

Con el gráfico de la función ACF de la imagen A.4, se puede proponer para parte regular un modelo $MA(3)$. Fijando en los retardos múltiples de s, en este caso, s es igual a 12, se ve que ninguno de los retardos estacionales es significativo, es decir, la parte estacional presenta el ruido blanco.

Con la gráfica de la función de autocorrelación parcial ($PACF$), se podría sugerir un $AR(2)$ para la parte regular, y el parte estacional, también se presenta el ruido blanco. Del mismo modo que se ha visto con el ACF, la parte estacional también presenta ruido blanco confirmado de esta manera que la serie no tiene componente estacional.

Teniendo en cuenta las transformaciones a la serie y el análisis llevado a cabo para la identificación de posibles modelos explicativos para el carburante, se han propuesto los siguientes modelos:

- **Modelo 1:** $MA(3)$ para W_t;

$$W_t = (1-B)^d(1-B^s)^D ln(X_t), \quad d=1,\ D=0,\ s=12;$$

La formula compacta del modelo es:

$$(1-B)X_t = \theta_3(B)Z_t$$

Sustituyendo cada polinomio característico, se obtiene:

$$(1-B)ln(X_t) = (1+\theta_1 B+\theta_2 B^2+\theta_3 B^3)Z_t$$

- **Modelo 2:** $AR(2)$ para W_t;

$$W_t = (1-B)^d(1-B^s)^D ln(X_t), \quad d=1,\ D=0,\ s=12;$$

La formula compacta del modelo es:

$$\phi_2(1-B)X_t = Z_t$$

Sustituyendo cada polinomio característico, se obtiene:

$$(1-\phi_1 B-\phi_2 B^2)(1-B)ln(X_t) = Z_t$$

A.3. Estimación

A continuación, se ajusta los modelos propuestos mediante la función $ARIMA$ de la programa $R-studio$ utilizando la serie estacionaria W_t. Recuerda que la serie W_t es la serie estacionaria, en este caso, es la serie original aplicando la transformación logarítmica y una diferenciación regular.

Figura A.5: Salida de estimación del modelo $MA(3)$ y $AR(2)$ para X_t

Results

	Dependent variable:	
	lnDiesel	
	MA(3) (1)	AR(2) (2)
ma1	0.505	
	t = 5.703	
ma2	-0.010	
	t = -0.090	
ma3	-0.165	
	t = -1.780	
ar1		0.538
		t = 6.457
ar2		-0.305
		t = -3.642
intercept	0.001	0.001
	t = 0.207	t = 0.192
Observations	131	131
Log Likelihood	282.345	283.048
sigma2	0.001	0.001
Akaike Inf. Crit.	-554.690	-558.095
Note:	t = T-statistic value = coeff/SE(coeff)	

Fuente: Elaboración propia

Con la imagen A.5, se puede ver que, los parámetros ϕ_2 y ϕ_3 del modelo $MA(3)$ no son significativos, por lo tanto, según el principio de parsimonia, se podría reducir el modelo a un $MA(1)$. Por otro lado, se ve que el intercepto de ambos modelos resultan un $t-ratio$ menor de dos en valor absoluto, en consecuencia, se habrá de modelizar con la serie original aplicando la transformación logarítmica.

Seguidamente, se mostrará en la imagen A.6 los resultados de la modelización con la serie sin aplicar la diferenciación previamente. Se presentará los resultados del modelo MA(3), AR(2) y otro modelo que se mencionará más delante.

Figura A.6: Salida de estimación del modelo $MA(3)$ y $AR(2)$ para W_t

Results

	Dependent variable:		
	lnDiesel		
	MA(3) (1)	MA(2) (2)	AR(2) (3)
ma1	-0.437	-0.517	
	t = -4.356	t = -4.590	
ma2	-0.436	-0.483	
	t = -5.239	t = -6.039	
ma3	-0.128		
	t = -1.305		
ar1			-0.105
			t = -1.246
ar2			-0.295
			t = -3.520
Observations	130	130	130
Log Likelihood	276.296	275.503	258.335
sigma2	0.001	0.001	0.001
Akaike Inf. Crit.	-544.592	-545.007	-510.669
Note:	t = T-statistic value = coeff/SE(coeff)		

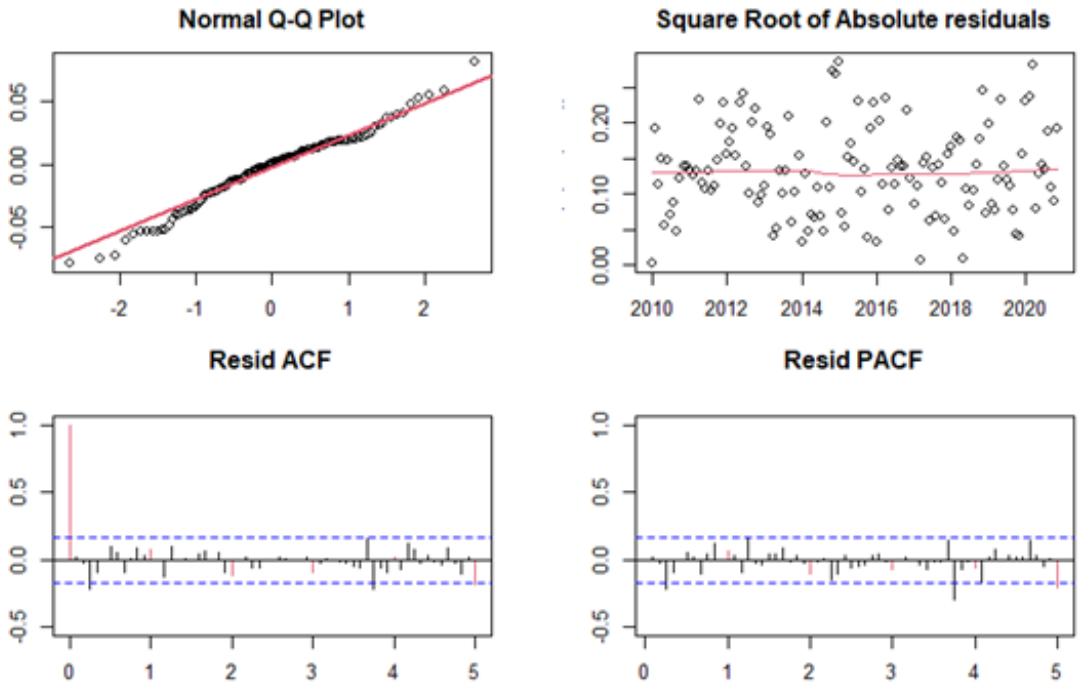
Fuente: Elaboración propia

Fijando en las salidas del modelo $MA(3)$, el parámetro ϕ_3 no es significativo, por lo tanto, se reduce el modelo a un $MA(2)$, y se ve que ahora, todos sus parámetros son significativos. Comparando los valores de AIC, resulta que el modelo $MA(2)$ es el modelo que explica mejor los datos, por lo tanto, se llevara a cabo la validación de dicho modelo.

A.4. Validación

En esta tercera etapa, se tendrá que validar el modelo analizando la normalidad, la variabilidad y la independencia de los residuos del modelo.

Figura A.7: Gráficos de validación de la serie de diésel



Fuente: Elaboración propia

Con el $qq-plot$, se ve que la mayoría de los residuos se ajustan bastante bien a la normalidad, menos unos pocos puntos muy concretos. Para la homogeneidad, se ve que los residuos presentan una varianza bastante estable. Analizando la independencia con la gráfica ACF y $PACF$ de los residuos, se ve que, la mayoría de los datos se sitúan dentro de las bandas de confianzas, excepto el tercer y el cuadragésimo quinto retardo.

Seguidamente, se mostrará en la cuadro A.1 los resultados de las pruebas numéricas para estudiar la normalidad, homogeneidad e independencia.

Cuadro A.1: Resultados de las pruebas numéricas para diésel

Prueba	**tipo**	**Estadístico**	**P-valor**
Shapiro-Wilk Normality	Normalidad	$W = 0{,}98114$	0.0064
Anderson-Darling	Normalidad	$A = 1{,}0233$	0.0038
Jarque-Bera	Normalidad	$\chi^2 = 14{,}527$	0.0007
Breusch-Pagan	Homogeneidad	$BP = 16{,}233$	0.0007
Durbin-Watson	Independencia	$DW = 1{,}9355$	0.0007

Fuente: Elaboración propia

El resultado para la prueba del $Anderson-Darling$ ha salido no significativo, con un $p-valor$ de 0.01043, menor que el nivel de la significación de 0.05. Por lo tanto, según esta prueba, se rechaza la normalidad de los residuos. En cambio, la prueba de $Shapiro-Wilk$ y $Jarque-Bera$, resultan un $p-valor$ de 0.06563 y

0.3327 respectivamente, ambos superan el nivel de significación, y no hay evidencia para rechazar la hipótesis nula.

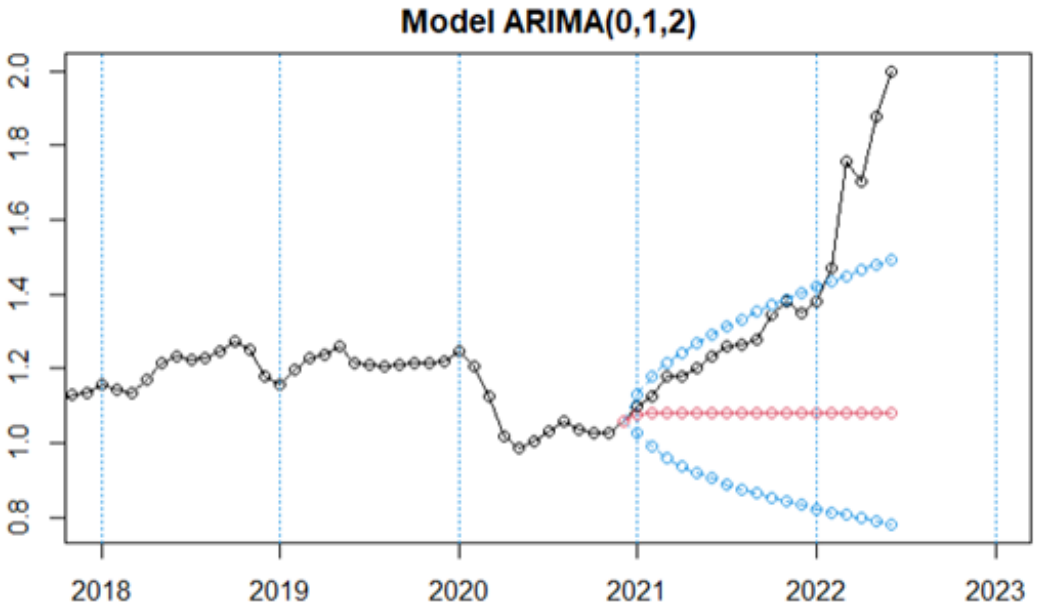
La prueba de $Durbin - Watson$ resulta un estadístico DW de 1.936 y un $p - valor$ de 0.3232, mayor que 0.05, por lo tanto, no hay evidencia para rechazar la hipótesis nula y se concluye que no existe correlación entre los residuos.

Y, para la homogeneidad, aplicando la prueba de $Breusch - Pagan$, resulta un estadístico BP de 16.233 y un $p - valor$ de 0.0001, con lo cual, se rechazaría la hipótesis nula. Pero, observando la distribución de los residuos en la gráfica de raíz cuadrada de residuos absolutos, se ve que, son bastante homogéneos. Concluyendo todas las pruebas y gráficas, se puede decir que el modelo $ARIMA(0, 1, 2)$ está validado para explicar los datos.

A.5. Predicción

A continuación, en el gráfico de la figura 5.16, se puede ver la comparación de los valores reales con los valores predichos.

Figura A.8: Comparación de valores reales y predichos diésel



Fuente: Elaboración propia

Los puntos negros de la gráfica son los datos reales del precio del gasoil. Los puntos rojos son los valores predichos aplicando el modelo $ARIMA(0, 1, 2)$ a la serie hasta el año 2020 y los puntos azules, las bandas de confianzas con un nivel de significación 95 %.

Apéndice B

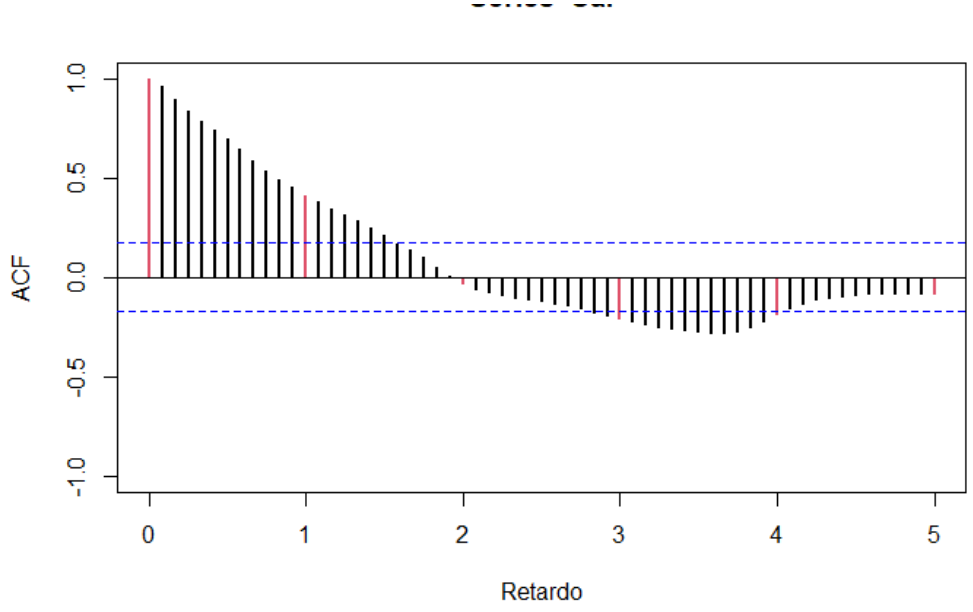
Análisis ARIMA: Diésel para calefacción

En este capítulo, se aplicará la metodología *ARIMA* para encontrar un modelo adecuado para la serie del diésel para la calefacción.

B.1. Transformaciones previas

Antes de empezar a identificar los posibles modelos, es necesario estudiar la estacionariedad de la serie.

Figura B.1: Gráfica ACF de la serie diésel para la calefacción

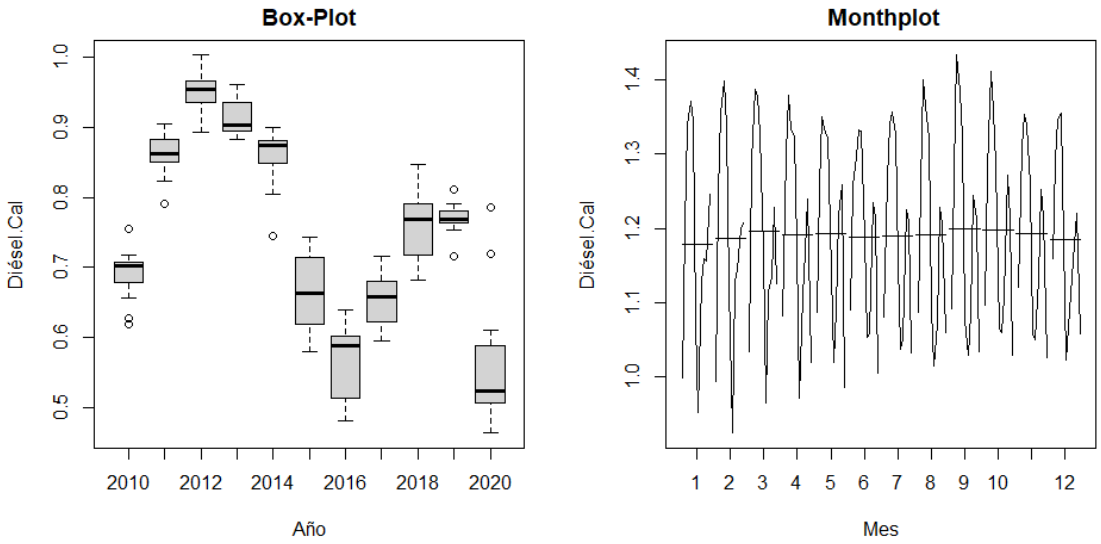


Fuente: Elaboración propia

Se ve en la iB.1,los retardos se decrecen muy lentamente hacia cero, con lo cual se refiere que la serie no es estacionaria. A continuación, se llevará a cabo los análisis sobre la media y varianza.

Seguidamente, se mostrará un gráfico de $box-plot$ y $monthplot$ para analizar las transformaciones necesarias para conseguir una serie estacionaria.

Figura B.2: Box-plot y Monthplot de la serie calefacción

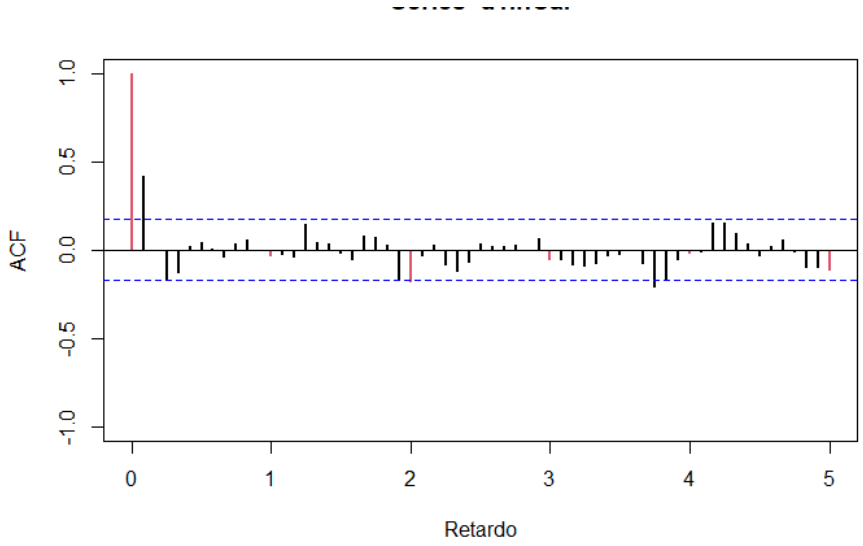


Fuente: Elaboración propia

La gráfica de $box-plot$ indica que la serie no presenta una varianza constante, donde el año 2015 y 2016 son los años que presenta mayor variabilidad y el 2019, menor. Fijando las líneas negras dentro de cada *boxplot*, se ve que el precio medio de cada año no es constante. Analizando la estacionalidad con el *monthplot*, se concluye que las diferencias entre la media de los meses no son considerables. Concluyendo los análisis previos, se habrá de aplicar una transformación logarítmica y una diferenciación regular a la serie para conseguir una serie estacionaria.

Una vez aplicado las transformaciones, se revisa si la serie transformada ya es estacionaria o no. A continuación, se muestra la gráfica de ACF de la serie transformada.

Figura B.3: Gráfica ACF de la serie calefacción transformada



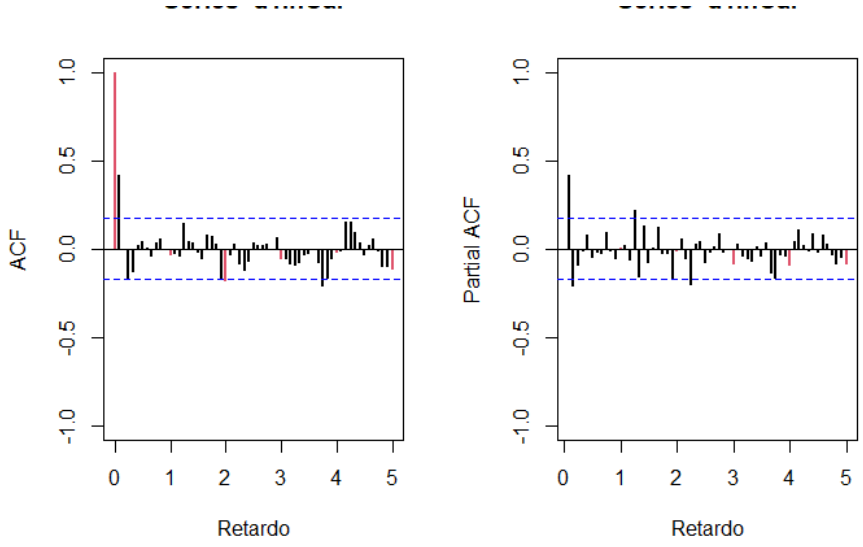
Fuente: Elaboración propia

Se ve en la imagen B.3,los retardos decaen rápidamente hacia cero, por lo tanto, se puede decir que la serie ya es estacionaria.

B.1.1. Identidicación

Una vez asegurado la estacionariedad de la serie, se llevará a cabo la identificación de los posibles modelos que se puede ajustar en la serie a través de la gráfica ACF y $PACF$. Para una mejor identificación, se mostrará en color rojo los retardos múltiples de 12.

Figura B.4: Gráfica ACF y PACF de la serie calefacción transformada



Fuente: Elaboración propia

Con la gráfica ACF, se puede proponer para parte regular un $MA(2)$. Fijando

en los retardos múltiples de s, en este caso, se identifica un $SMA(2)$. Esta situación es extraña porque con el resultado del apartado anterior, se considera que no hay componente estacional, pero de toda manera, se seguirá analizando según los modelos identificados con la gráfica de ACF.

Con la gráfica de la función de autocorrelación parcial ($PACF$), se podría sugerir un $AR(2)$ para la parte regular, y el parte estacional, se presenta el ruido blanco.

Teniendo en cuenta las transformaciones a la serie y el análisis llevado a cabo para la identificación de posibles modelos explicativos para el carburante, se han propuesto los siguientes modelos:

- **Modelo 1:** $MA(2)SMA(2)$ para W_t;

$$W_t = (1-B)^d(1-B^s)^D ln(X_t), \;\; d=1, \; D=0, \; s=12;$$

La formula compacta del modelo es:

$$\theta_2(B)(1-B)(1-B^{12})ln(X_t) = \Theta_2(B^{12})Z_t$$

Sustituyendo cada polinomio característico, se obtiene:

$$(1-\theta_1 B-\theta_2 B^2)(1-B)(1-B^{12})ln(X_t) = (1+\Theta_1^{12}+\Theta_2^{12})$$

- **Modelo 2:** $AR(2)SMA(2)$ para W_t;

$$W_t = (1-B)^d(1-B^s)^D ln(X_t), \;\; d=1, \; D=0, \; s=12;$$

La formula compacta del modelo es:

$$\phi_2(B)(1-B)(1-B^{12})ln(X_t) = \Theta_2(B^{12})Z_t$$

Sustituyendo cada polinomio característico, se obtiene:

$$(1-\phi_1 B-\phi_2 B^2)(1-B)(1-B^{12})ln(X_t) = (1+\Theta_1^{12}+\Theta_2^{12})$$

B.1.2. Estimación

A continuación, se ajusta los modelos sugeridos mediante la función $ARIMA$ utilizando la serie estacionaria W_t. Recuerda que la serie W_t es la serie original aplicando la transformación logarítmica y una diferenciación regular.

Seguidamente, se mostrará una tabla de resultados con todos los parámetros, t-ratios, valor de log-verosimilitud maximizada, la varianza estimada y el valor de AIC de cada uno de los dos modelos propuestos

Figura B.5: Tabla resultante de las estimaciones con diésel.cal transformada

Results

	Dependent variable:	
	d1lnCal	
	MA(2)SMA(2) (1)	AR(2)SMA(2) (2)
ma1	0.476	
	t = 5.309	
ma2	0.110	
	t = 1.014	
ar1		0.487
		t = 5.469
ar2		-0.193
		t = -2.181
sma1	-0.091	-0.086
	t = -0.837	t = -0.825
sma2	-0.201	-0.189
	t = -1.893	t = -1.812
intercept	-0.001	-0.001
	t = -0.208	t = -0.253
Observations	131	131
Log Likelihood	229.103	229.779
sigma2	0.002	0.002
Akaike Inf. Crit.	-446.205	-447.558

Note: t = T-statistic value = coeff/SE(coeff)

Fuente: Elaboración propia

Primero de todo, se puede ver en la imagen B.5 que tanto el intercepto como los dos parámetros que representa el componente estacional, no son significativos, por lo tanto, se llevara a cabo la modelización con la serie original aplicando la transformación logarítmica.

Seguidamente, se mostrará una tabla de resultados modelizando con la serie sin aplicar la diferenciación previamente, y se la introduce con la función $ARIMA$ del programa $R{-}Studio$. Se presentará los resultados del modelo $MA(2)SMA(2)$,$AR(2)SMA(2)$. Aparte de estos dos, también mostrará los resultados del modelo $MA(1)$ y $AR(2)$, quitando los parámetros no significativos identificados anteriormente.

Figura B.6: Tabla resultante de la estimación con la serie diésel.cal

Results

	Dependent variable:			
	lnCal			
	MA(2)SMA(2) (1)	MA(1) (2)	AR(2)SMA(2) (3)	AR(2) (4)
ma1	0.477	0.474		
	t = 5.327	t = 6.255		
ma2	0.110			
	t = 1.021			
ar1			0.488	0.519
			t = 5.487	t = 6.031
ar2			-0.193	-0.222
			t = -2.175	t = -2.572
sma1	-0.090		-0.084	
	t = -0.828		t = -0.813	
sma2	-0.199		-0.187	
	t = -1.886		t = -1.801	
Observations	131	131	131	131
Log Likelihood	229.081	226.666	229.747	228.065
sigma2	0.002	0.002	0.002	0.002
Akaike Inf. Crit.	-448.162	-449.333	-449.494	-450.129

Note: t = T-statistic value = coeff/SE(coeff)

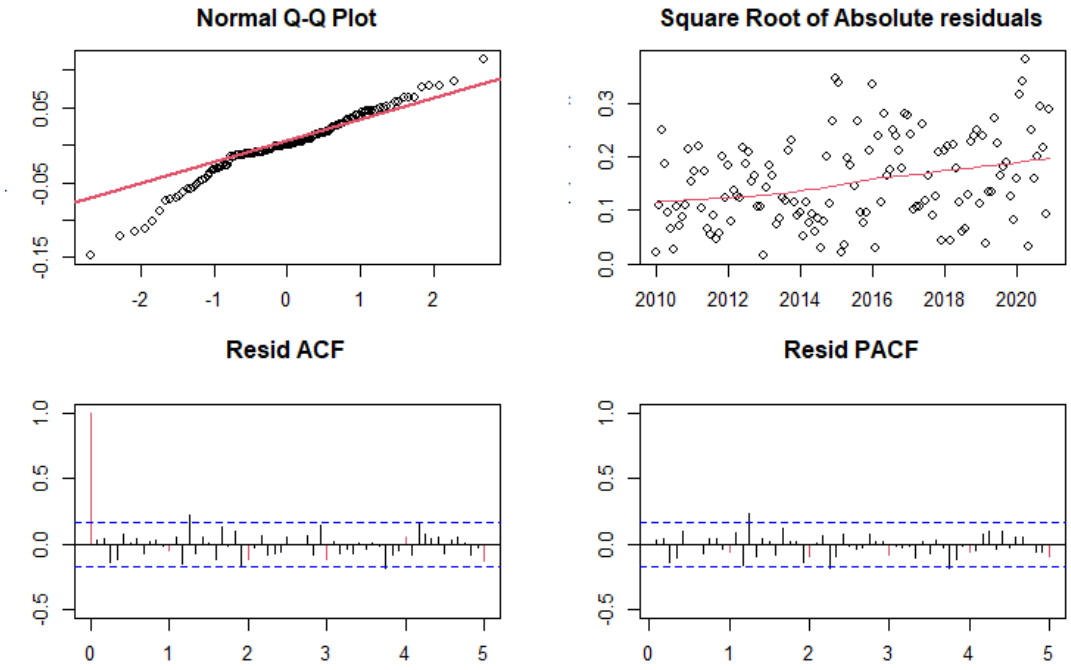
Fuente: Elaboración propia

Se observa en la imagen B.6, modelizando en la serie X_t, los parámetros que explica el componente estacional siguen siendo no significativos. Por lo tanto, se ha de escoger un modelo entre $MA(1)$ y $AR(2)$. Observando los valores de AIC, se ve que el modelo que presenta menor AIC es el modelo $AR(2)$, pero, la diferencia que lleva respecto el modelo $MA(1)$ no es considerable. Debido al principio de parsimonia, se ha decidido llevar a cabo la validación del modelo $MA(1)$.

B.2. Validación

Para esta etapa, se llevará a cabo el análisis sobre la normalidad, variabilidad e independencia de los residuos del modelo.

Figura B.7: Gráficas de validación de la serie de calefacción



Fuente: Elaboración propia

Con las gráficas se ve que hay bastantes residuos que no se ajusta a la normalidad. Para la homogeneidad, se puede pensar que el primer parte de residuos son menos variables que los restos y para la independencia, se ve que, la mayoría de los datos se sitúan dentro de las bandas de confianzas. A continuación, se llevará a cabo las pruebas numéricas.

Cuadro B.1: Resultados de las pruebas numéricas para diésel de calefacción

Prueba	**tipo**	**Estadístico**	**P-valor**
Shapiro-Wilk Normality	Normalidad	$W = 0{,}9627$	0.001107
Anderson-Darling	Normalidad	$A = 1{,}6769$	0.000253
Jarque-Bera	Normalidad	$\chi^2 = 17{,}333$	0.000172
Breusch-Pagan	Homogeneidad	$BP = 8{,}6717$	0.003233
Durbin-Watson	Independencia	$DW = 1{,}9117$	0.2742

Fuente: Elaboración propia

El resultado de todas las pruebas de normalidad no ha salido significativo, los $p - valores$ son mucho menor que el nivel de significación de 0.05, por lo tanto, hay evidencia para rechazar la hipótesis nula, se concluye que los residuos no siguen una distribución normal.

La prueba de homogeneidad resulta un estadístico BP de 8.6713 y un $p-valor$ de 0.003233, menor que el nivel de significación, por lo tanto, hay evidencia para rechazar la hipótesis nula y que los residuos no se distribuyan de una manera constante.

Y por último, aplicando la prueba de $Durbin-Watson$, resulta un estadístico DW de 1.9117 y un $p-valor$ de 0.2742, con lo cual, no hay evidencia para rechazar la hipótesis nula y que los residuos son independientes.

Al no poder validar la normalidad y homogeneidad de los residuos tanto numéricamente como gráficamente, no ha podido dar por válido el modelo $MA(1)$. Una vez conseguido este resultado, se ha llevado a cabo la validación del modelo $AR(2)$, pero tampoco ha podido conseguir la validación de la normalidad ni homogeneidad. Por lo tanto, se ha decidido pronosticar con el modelo $MA(1)$, aunque, los valores que serán predichos con este modelo no podrán representa la evolución del precio de diésel para calefacción, los resultados se pueden utilizar como referencia.

Apéndice C

Código

```
## --------
library(rvest)

contador <- c(2010:2022)
paginas <- paste0("https://datosmacro.expansion.com/energia/precios
    -gasolina-diesel-calefaccion/espana?anio=",contador)
x <- data.frame()

for (pagina in paginas){
  # Defino la pagina
  url <- pagina # Una pagina del contador
  # Leo el codigo
  code <- read_html(url)
  # Obtengo el nodo de las tablas enteras:
  tablas <- html_table(code)
  length(tablas) # Aqui veo que son "n" tablas
  # Genero una tabla vacia para irla llenando para esta
    p      g i n a  particular
  tabla_parcial <- data.frame()
  tabla_parcial <- tabla_parcial[-length(tablas),]

  for (j in length(tablas)){
  tabla_parcial <- rbind.data.frame(tabla_parcial, tablas[[j]])#
    Hacemos un loop sobre las tablas
  tabla_parcial$mes <- tabla_parcial$Fecha
  tabla_parcial$mes <- substr(tabla_parcial$mes,4,5) #Crear un
    variable que solo coge el numero del mes
  tabla_parcial <- tabla_parcial[-nrow(tabla_parcial),] #Quitar
    ultima observacion
  tabla_parcial <- tabla_parcial[order(tabla_parcial$mes),] #
    Ordenar dataframe segun mes
  tabla_parcial <- as.data.frame(tabla_parcial)
```

```
27
28      for(j in 2:ncol(tabla_parcial)) {
29        tabla_parcial[[j]] <- substr(tabla_parcial[[j]], 1, 5) #
      Quitar el simbolo de euro
30        tabla_parcial[[j]] <- sub(",",".", tabla_parcial[,j],fixed=
      TRUE) #Sustituir coma por punto
31        tabla_parcial[[j]] <- as.numeric(substr(tabla_parcial[[j]],
      1, 5)) #convertir en numerica
32       }
33       tabla_parcial$Fecha2 <- substr(tabla_parcial$Fecha,4,10) #Solo
      mostrar mes y  a       o
34       tabla_parcial<-aggregate(tabla_parcial[, 2:8], list(tabla_
      parcial$Fecha2), mean) #sacar media segun mes
35
36    }
37    # a la tabla que va a estar capturando todo:
38    x <- rbind.data.frame(x,tabla_parcial)
39  }
40  x$mes <- as.numeric(x$mes)
41  x$any <- as.numeric(substr(x$Group.1,4,8))
42
43
44
45  ## ------
46  missing <- c("07/2020","NA","NA","NA",07,2020)
47  x <- rbind(x,missing)
48  x$mes <- as.numeric(x$mes)
49  x <- x[order(x$any,x$mes),]
50  x[127,]
51
52
53  ## ------
54  library(imputeTS)
55  for(j in 2:ncol(x)) {
56    x[[j]] <- as.numeric(x[[j]])
57    x[[j]] <- ts(x[[j]],start = 2010,frequency = 12)
58    x[[j]] <- na_interpolation(x[[j]], option='linear')
59  }
60  x[127,]
61
62
63  ## ------
64  x1 <- x[1:132,]
65
66
67  ## -----
68  gas1 <- ts(x1$`Super 95`,start = 2010,frequency = 12)
```

```
dies1 <- ts(x1$Diesel,start = 2010,frequency = 12)
cal1 <- ts(x1$`Diesel Cal.`,start = 2010,frequency = 12)
ts.plot(gas1,dies1,cal1,gpars = list(col=c("lightpink","lightblue",
   "lightgreen")))
legend("topright",legend=c("Gasolina","Diesel","Calefacci n"),col
   = c("lightpink","lightblue","lightgreen"), lty = 1)
abline(v=2010:2022,col="grey",lty=3)


gas2 <- ts((x1$`Super 95`-x1$`Super 95 (Sin imp.)`),start = 2010,
   frequency = 12)
dies2 <- ts((x1$Diesel - x1$`Diesel (Sin imp.)`),start = 2010,
   frequency = 12)
cal2 <- ts((x1$`Diesel Cal.`-x1$`Diesel Cal. (Sin imp.)`),start =
   2010,frequency = 12)
ts.plot(gas2,dies2,cal2,gpars = list(col=c("lightpink","lightblue",
   "lightgreen")),ylim=c(0,1))
legend("topright",legend=c("Gasolina","Diesel","Calefacci n"),col
   = c("lightpink","lightblue","lightgreen"), lty = 1)
abline(v=2010:2022,col="grey",lty=3)



## -----
###Contraste Chow Gasolina###
T1 <- x$`Super 95`[102:125]
T2 <- x$`Super 95`[126:149]

#load strucchange package
library(strucchange)

#perform Chow test
sctest(T1 ~ T2, type = "Chow", point = 10)

###Contraste Chow Diesel###

T1 <- x$Diesel[102:125]
T2 <- x$Diesel[126:149]
#perform Chow test
sctest(T1 ~ T2, type = "Chow", point = 10)


###Contraste Chow Diesel Calefaccion###

T1 <- x$`Diesel Cal.` [102:125]
T2 <- x$`Diesel Cal.` [126:149]
#perform Chow test
```

```
sctest(T1 ~ T2, type = "Chow", point = 10)




## ------
summary(x1$`Super 95`)
summary(x1$Diesel)
summary(x1$`Diesel Cal.`)



#por a o
gasolina <- ts(x1$`Super 95`,start = 2010,frequency = 12)

media_gas <- as.numeric(aggregate(gasolina, nfrequency = 1, FUN =
    mean))
max_gas <- as.numeric(aggregate(gasolina, nfrequency = 1, FUN = max
    ))
min_gas <- as.numeric(aggregate(gasolina, nfrequency = 1, FUN = min
    ))
mediana_gas <- as.numeric(aggregate(gasolina, nfrequency = 1, FUN =
     median))
q1_gas <- as.numeric(aggregate(gasolina, nfrequency = 1, FUN =
    quantile)[c(2,7,12,17,22,27,32,37,42,47,52)])
q3_gas <-  as.numeric(aggregate(gasolina, nfrequency = 1, FUN =
    quantile)[c(4,9,14,19,24,29,34,39,44,49,54)])

descrip_gas <- cbind(min_gas,q1_gas,mediana_gas,media_gas,q3_gas,
    max_gas)
descrip_gas

diesel <- ts(x1$Diesel,start = 2010,frequency = 12)

media_diesel <- as.numeric(aggregate(diesel, nfrequency = 1, FUN =
    mean))
max_diesel <- as.numeric(aggregate(diesel, nfrequency = 1, FUN =
    max))
min_diesel <- as.numeric(aggregate(diesel, nfrequency = 1, FUN =
    min))
mediana_diesel <- as.numeric(aggregate(diesel, nfrequency = 1, FUN
    = median))
q1_diesel <- as.numeric(aggregate(diesel, nfrequency = 1, FUN =
    quantile)[c(2,7,12,17,22,27,32,37,42,47,52)])
q3_diesel <-  as.numeric(aggregate(diesel, nfrequency = 1, FUN =
    quantile)[c(4,9,14,19,24,29,34,39,44,49,54)])

```

```
143 descrip_diesel <- cbind(min_diesel,q1_diesel,mediana_diesel,media_
       diesel,q3_diesel,max_diesel)
144 descrip_diesel
145
146
147
148 cal <- ts(x1$`Diesel Cal.`,start = 2010,frequency = 12)
149
150 media_cal <- as.numeric(aggregate(cal, nfrequency = 1, FUN = mean))
151 max_cal <- as.numeric(aggregate(cal, nfrequency = 1, FUN = max))
152 min_cal <- as.numeric(aggregate(cal, nfrequency = 1, FUN = min))
153 mediana_cal <- as.numeric(aggregate(cal, nfrequency = 1, FUN =
       median))
154 q1_cal <- as.numeric(aggregate(cal, nfrequency = 1, FUN = quantile)
       [c(2,7,12,17,22,27,32,37,42,47,52)])
155 q3_cal <-  as.numeric(aggregate(cal, nfrequency = 1, FUN = quantile
       )[c(4,9,14,19,24,29,34,39,44,49,54)])
156
157 descrip_cal <- cbind(min_cal,q1_cal,mediana_cal,media_cal,q3_cal,
       max_cal)
158 descrip_cal
159
160
161 gas2 <- ts(x1$`Super 95`,start = 2010,frequency = 12)
162 dies2 <- ts(x1$Diesel,start = 2010,frequency = 12)
163 cal2 <- ts(x1$`Diesel Cal.`,start = 2010,frequency = 12)
164 ts.plot(gas2,dies2,cal2,gpars = list(col=c("lightpink","lightblue",
       "lightgreen")),ylim = c(0.3,1.8))
165 legend("topright",legend=c("Gasolina","Diesel","Calefacci n"),col
       = c("lightpink","lightblue","lightgreen"), lty = 1)
166 abline(v=2010:2021,col="grey",lty=3)
167 points(2012.65,1.501,pch = 19,col = "red")
168 points(2020.33,1.079,pch = 19,col = "darkGreen")
169 points(2012.65,1.4325,pch = 19,col = "red")
170 points(2016.083,0.9254,pch = 19,col = "darkGreen")
171 points(2012.65,1.003,pch = 19,col = "red")
172 points(2020.33,0.462,pch = 19,col = "darkGreen")
173
174
175
176 ## -----
177 acf(gasolina,ylim=c(-1,1),lag.max=60,col=c(2,rep(1,11)),lwd=2)
178
179
180 par(mfrow=c(1,2))
181 gasolina <- as.numeric(gasolina)
182 boxplot(gasolina~x1$any)
```

```
m=apply(matrix(gasolina,ncol=12),2,mean)
v=apply(matrix(gasolina,ncol=12),2,var)
plot(v~m,main="Mean-Variance plot") # no constante
abline(lm(v~m))


## ----setup, include=FALSE-----
lngas=log(gasolina)
plot(lngas,type = "o")
par(mfrow=c(1,2))
boxplot(lngas~x1$any)
m=apply(matrix(lngas,ncol=12),2,mean)
v=apply(matrix(lngas,ncol=12),2,var)
plot(v~m,main="Mean-Variance plot")
abline(lm(v~m))


## -------
gass <- ts(x1$`Super 95`, frequency = 12)

monthplot(gass, ylab  = "Gasolina", main = "Monthplot")


## -------
d1lngas<-ts(diff(lngas),start = 2010,frequency = 12)
plot(d1lngas,main="d1lngas")
abline(h=0)
abline(h=mean(d1lngas), col=2)


## ------
par(mfrow=c(1,2))
acf(d1lngas,ylim=c(-1,1),lag.max=60,col=c(2,rep(1,11)),lwd=2, xlab
    = "Retardo" )
pacf(d1lngas,ylim=c(-1,1),lag.max=60,col=c(rep(1,11),2),lwd=2,xlab
    = "Retardo" )


## ----
mod1A=arima(d1lngas, order=c(0,0,4))
mod2A=arima(d1lngas, order=c(2,0,0))
#cat("Modelo 2A \nT-ratios:",round(mod2B$coef/sqrt(diag(mod2B$var.
    coef)),2))
mod1B=arima(lngas, order=c(0,1,4))
mod2B=arima(lngas, order=c(2,1,0))

```

```
## ----------
library(stargazer)

stargazer(mod1A, mod2A,title="Resultados", type="text",notes.append
    = FALSE,  report = "vct",notes = c("t = T-statistic value =
   coeff/SE(coeff)"),
  digits = 3,
  column.labels = c("MA(4)","AR(2)"))

stargazer(mod1B,mod2B,title="Resultados", type="text",notes.append
   = FALSE,  report = "vct",notes = c("t = T-statistic value =
   coeff/SE(coeff)"),
  digits = 3,
  column.labels = c("MA(4)","AR(2)"))


## ------
#Normalidad
model = mod2B
resid=model$residuals
par(mfrow=c(1,2),mar=c(3,3,3,3))
qqnorm(resid)
qqline(resid,col=2,lwd=2)
hist(resid,breaks=20,freq=FALSE)
curve(dnorm(x,mean=mean(resid),sd=sd(resid)),col=2,add=T)



##################Validation###################################
validation=function(model){
 s=frequency(get(model$series))
  resid=model$residuals
  par(mfrow=c(2,2),mar=c(3,3,3,3))
  #Residuals plot
  plot(resid,main="Residuals")
  abline(h=0)
  abline(h=c(-3*sd(resid),3*sd(resid)),lty=3,col=4)
  #Square Root of absolute values of residuals (Homocedasticity)
  scatter.smooth(sqrt(abs(resid)),main="Square Root of Absolute
   residuals",
                 lpars=list(col=2))

  #Normal plot of residuals
  qqnorm(resid)
  qqline(resid,col=2,lwd=2)

  ##Histogram of residuals with normal curve
```

```
hist(resid,breaks=20,freq=FALSE)
curve(dnorm(x,mean=mean(resid),sd=sd(resid)),col=2,add=T)


#ACF & PACF of residuals
par(mfrow=c(1,2))
acf(resid,ylim=c(-1,1),lag.max=60,col=c(2,rep(1,11)),lwd=1)
pacf(resid,ylim=c(-1,1),lag.max=60,col=c(rep(1,11),2),lwd=1)
par(mfrow=c(1,1))


#ACF & PACF of square residuals
par(mfrow=c(1,2))
acf(resid^2,ylim=c(-1,1),lag.max=60,col=c(2,rep(1,11)),lwd=1)
pacf(resid^2,ylim=c(-1,1),lag.max=60,col=c(rep(1,11),2),lwd=1)
par(mfrow=c(1,1))

#Ljung-Box p-values
par(mar=c(2,2,1,1))
tsdiag(model,gof.lag=7*12)
cat("\n
 ----------------------------------------------------------------------\
 n")
print(model)

#Stationary and Invertible
cat("\nModul of AR Characteristic polynomial Roots: ",
    Mod(polyroot(c(1,-model$model$phi))),"\n")
cat("\nModul of MA Characteristic polynomial Roots: ",
    Mod(polyroot(c(1,model$model$theta))),"\n")

#Model expressed as an MA infinity (psi-weights)
psis=ARMAtoMA(ar=model$model$phi,ma=model$model$theta,lag.max=36)
names(psis)=paste("psi",1:36)
cat("\nPsi-weights (MA(inf))\n")
cat("\n--------------------\n")
print(psis[1:20])

#Model expressed as an AR infinity (pi-weights)
pis=-ARMAtoMA(ar=-model$model$theta,ma=-model$model$phi,lag.max
 =36)
names(pis)=paste("pi",1:36)
cat("\nPi-weights (AR(inf))\n")
cat("\n--------------------\n")
print(pis[1:20])

## Add here complementary tests (use with caution!)
```

```
313 ##-------------------------------------------------------
314 cat("\nNormality Tests\n")
315 cat("\n--------------------\n")
316
317 ##Shapiro-Wilks Normality test
318 print(shapiro.test(resid(model)))
319 sw<-round(shapiro.test(resid(model))$p.value,4)
320 suppressMessages(require(nortest,quietly=TRUE,warn.conflicts=
     FALSE))
321 ##Anderson-Darling test
322 print(ad.test(resid(model)))
323 ad<-round(ad.test(resid(model))$p.value,4)
324
325 suppressMessages(require(tseries,quietly=TRUE,warn.conflicts=
     FALSE))
326 ##Jarque-Bera test
327 print(jarque.bera.test(resid(model)))
328 jb<-round(jarque.bera.test(resid(model))$p.value,4)
329
330 cat("\nHomoscedasticity Test\n")
331 cat("\n--------------------\n")
332 suppressMessages(require(lmtest,quietly=TRUE,warn.conflicts=FALSE
     ))
333 ##Breusch-Pagan test
334 obs=get(model$series)
335 print(bptest(resid(model)~I(obs-resid(model))))
336 bp<-round(bptest(resid(model)~I(obs-resid(model)))$p.value,4)
337
338 cat("\nIndependence Tests\n")
339 cat("\n--------------------\n")
340
341
342
343 ##Durbin-Watson test
344 print(dwtest(resid(model)~I(1:length(resid(model)))))
345
346 dw<-dwtest(resid(model)~I(1:length(resid(model))))$p.value
347
348 ##Ljung-Box test
349 cat("\nLjung-Box test\n")
350 print(t(apply(matrix(c(1:4,(1:4)*s)),1,function(el) {
351   te=Box.test(resid(model),type="Ljung-Box",lag=el)
352   c(lag=(te$parameter),statistic=te$statistic[[1]],p.value=te$p.
     value)})))
353 ##************End of complementary tests**********************
     ****************
354 lj<-round(t(apply(matrix(c(1:4,(1:4)*s)),1,function(el) {
```

```
355     te=Box.test(resid(model),type="Ljung-Box",lag=el)
356     c(lag=(te$parameter),statistic=te$statistic[[1]],p.value=te$p.
        value)}))[,3],4)
357
358 ################## Fi Validaci n ('Validation')
        ##################################
359
360
361
362 resumen<-data.frame(Pruebas=1:18)
363 colnames(resumen)<-paste0("mod1B")
364 rownames(resumen)<-c("Shapiro-Wilks Normality p-value","Anderson-
        Darling p-value","Jarque-Bera p-value","Breusch-Pagan p-value","
        Durbin-Watson p-value",
365           "Ljung-Box (lag 1) p-value","Ljung-Box (lag 2) p-value","
        Ljung-Box (lag 3) p-value","Ljung-Box (lag4) p-value",
366           "Ljung-Box (lag 12) p-value","Ljung-Box (lag 24) p-value",
        "Ljung-Box (lag 36) p-value","Ljung-Box (lag 48) p-value",
367           "Log Likelihood","AIC","RMSPE", "MAPE","Mean Length")
368
369   resumen[1,1]=sw[1]
370   resumen[2,1]=ad[1]
371   resumen[3,1]=jb[1]
372   resumen[4,1]=bp[1]
373   resumen[5,1]=dw[1]
374   resumen[6,1]=lj[1]
375   resumen[7,1]=lj[2]
376   resumen[8,1]=lj[3]
377   resumen[9,1]=lj[4]
378   resumen[10,1]=lj[5]
379   resumen[11,1]=lj[6]
380   resumen[12,1]=lj[7]
381   resumen[13,1]=lj[8]
382   resumen[14,1]=model$loglik
383   resumen[15,1]=model$aic
384   resumen[16,1]=NA
385   resumen[17,1]=NA
386   resumen[18,1]=NA
387   return(resumen)
388
389 }
390
391 ## -----
392 model=mod2B
393
394 validation(model)
395
```

```
396
397
398 ## ----fig.height=3,fig.width=8-------
399 #Normalidad
400 model = mod2B
401 resid=model$residuals
402 par(mfrow=c(1,2),mar=c(3,3,3,3))
403 qqnorm(resid)
404 qqline(resid,col=2,lwd=2)
405 hist(resid,breaks=20,freq=FALSE)
406 curve(dnorm(x,mean=mean(resid),sd=sd(resid)),col=2,add=T)
407
408 #Homogeneidad
409 par(mfrow=c(1,2),mar=c(3,3,3,3))
410 plot(resid,main="Residuals")
411 abline(h=0)
412 abline(h=c(-3*sd(resid),3*sd(resid)),lty=3,col=4)
413 #Square Root of absolute values of residuals (Homocedasticity)
414 scatter.smooth(sqrt(abs(resid)),main="Square Root of Absolute
    residuals",
415                lpars=list(col=2))
416
417 #Independencia
418 par(mfrow=c(1,2),mar=c(3,3,3,3))
419 acf(resid,ylim=c(-0.5,1),lag.max=60,col=c(2,rep(1,11)),lwd=1, main
    = "Resid ACF")
420 pacf(resid,ylim=c(-0.5,1),lag.max=60,col=c(rep(1,11),2),lwd=1, main
     = "Resid PACF")
421
422
423 ## ----fig.height=7,fig.width=8----------
424 #Ljung-Box p-values
425
426 par(mar=c(2,2,1,1))
427 tsdiag(model,gof.lag=7*12)
428 ?tsdiag
429
430
431 ## ---------
432 ultim=c(2020,12)                        #Dic 2020
433 serie1 <- x[1:150,] #Serie competo: 2010-2022 abril
434 lnserie1 <- log(serie1$`Super 95`)
435 serie2 <- x1[1:132,] #Serie sin observacion del ultimo a o:
    2010-20121
436 lnserie2 <- log(serie2$`Super 95`)
437
438 #Fit the model to the complete series: lnserie1
```

```
(mo1B=arima(lnserie1,order=c(2,1,0)))
#Fit the model to the subset series (without 2020 data): lnserie2
(mod1B2=arima(lnserie2,order=c(2,1,0)))


## -------------
obs <- x1$`Super 95`[120:132]
cat("mod1B2")
#Fit the model to the subset series (without 2019 data): lnserie2
(mod1B2=arima(lnserie2, order=c(2,1,0)))
model=mod1B2
##### Use subset series lnserie2 to predict 2019 data
pred=predict(mod1B2,n.ahead=12)                                    #
   outputs point predictions and corresponding standard errors:for
   year 2019
se<-ts(c(0,pred$se),start=ultim,freq=12)                          #
   Standard errors for point predictions
pr<-ts(c(tail(lnserie2,1),pred$pred),start=ultim,freq=12)  #point
   predictions
tl<-ts(exp(pr-1.96*se),start=ultim,freq=12)
tu<-ts(exp(pr+1.96*se),start=ultim,freq=12)
pr<-ts(exp(pr),start=ultim,freq=12)              #pr
(mod.EQM1=sqrt(sum(((obs-pr)/obs)^2)/12))   # Error = obs - pred
(mod.EAM1=sum(abs(obs-pr)/obs)/12)
(mod.ML1=sum(tu-tl)/12)

resumen1<-validation(model)
resumen1[16,1]=mod.EQM1
resumen1[17,1]=mod.EAM1
resumen1[18,1]=mod.ML1

cat("mod2B2")
#Fit the model to the subset series (without 2019 data): lnserie2
(mod2B=arima(lnserie2, order=c(2,1,0)))
model=mod2B
##### Use subset series lnserie2 to predict 2019 data
pred=predict(mod2B,n.ahead=12)                                     #
   outputs point predictions and corresponding standard errors:for
   year 2019
se<-ts(c(0,pred$se),start=ultim,freq=12)                          #
   Standard errors for point predictions
pr<-ts(c(tail(lnserie2,1),pred$pred),start=ultim,freq=12)  #point
   predictions
tl<-ts(exp(pr-1.96*se),start=ultim,freq=12)
tu<-ts(exp(pr+1.96*se),start=ultim,freq=12)
pr<-ts(exp(pr),start=ultim,freq=12)              #pr
(mod.EQM3=sqrt(sum(((obs-pr)/obs)^2)/12))   # Error = obs - pred
```

```
(mod.EAM3=sum(abs(obs-pr)/obs)/12)
(mod.ML3=sum(tu-tl)/12)
resumen3<-validation(model)
resumen3[16,1]=mod.EQM3
resumen3[17,1]=mod.EAM3
resumen3[18,1]=mod.ML3


tablef<-cbind.data.frame(resumen1,resumen3)
stargazer(tablef, summary=FALSE, type="text")


## ---
pred=predict(mod1B2,n.ahead=18)                                    #
   outputs point predictions and corresponding standard errors:for
   year 2019
pr<-ts(c(tail(lnserie2,1),pred$pred),start=ultim,freq=12)  #point
   predictions

se<-ts(c(0,pred$se),start=ultim,freq=12)                           #
   Standard errors for point predictions

#Prediction Intervals (back transformed to original scale using exp
   -function)
tl<-ts(exp(pr-1.96*se),start=ultim,freq=12)
tu<-ts(exp(pr+1.96*se),start=ultim,freq=12)
pr<-ts(exp(pr),start=ultim,freq=12)              #predictions in
   original scale

#Plot of the original series and out-of-sample predictions: only
   time window 2015-2019 shown
gas <- ts(x$`Super 95`,start = 2010,frequency = 12)
gas <- ts(gas[1:150],start = 2010,frequency = 12)
ts.plot(gas,tl,tu,pr,lty=c(1,2,2,1),col=c(1,4,4,2),xlim=ultim[1]+c
   (-2,+3),type="o",main="Model ARIMA(2,1,0)")
abline(v=(ultim[1]-3):(ultim[1]+3),lty=3,col=4)

TV = round(((gas-pr)/pr)*100,1)
dd <- as.data.frame((previs=window(cbind(gas,pr,diferencia=round(
   gas-pr,6),TV),start=ultim)))









## -------
acf(x1$Diesel,ylim=c(-0.5,1),lag.max=60,col=c(2,rep(1,11)),lwd=2,
```

```
    xlab = "Retardos")
516
517
518 ## ----fig.width=10,fig.height=5------
519 Diesel <- ts(x1$Diesel,start = 2010,frequency = 12)
520 Media y varianza no constante, no estacionalida, aplicar log y
    diferenciacion regular
521
522
523 ## -------
524 lndies <- log(Diesel)
525 d1lndies <- ts(diff(lndies),start = 2010,frequency = 12)
526
527 par(mfrow=c(1,2))
528 acf(d1lndies,ylim=c(-0.5,1),lag.max=60,col=c(2,rep(1,11)),lwd=2,
    xlab = "Retardos")
529 pacf(d1lndies,ylim=c(-0.5,1),lag.max=60,col=c(2,rep(1,11)),lwd=2,
    xlab = "Retardos")
530
531
532
533
534 ## ------
535 mod1A <- arima(lnDiesel, order=c(0,0,3))
536 cat("Modelo A \nT-ratios:",round(mod1A$coef/sqrt(diag(mod1A$var.
    coef)),2))
537 mod1AA <- arima(lnDiesel, order=c(0,0,1))
538 cat("Modelo A \nT-ratios:",round(mod1AA$coef/sqrt(diag(mod1AA$var.
    coef)),2))
539
540 mod2A <- arima(lnDiesel, order=c(2,0,0))
541 cat("Modelo A \nT-ratios:",round(mod2A$coef/sqrt(diag(mod2A$var.
    coef)),2))
542
543 library("stargazer")
544 stargazer(mod1A, mod2A,title="Results", type="text",notes.append =
    FALSE,  report = "vct",
545 notes = c("t = T-statistic value = coeff/SE(coeff)"),
546   digits = 3,
547   column.labels = c("MA(3)","AR(2)"))
548
549
550
551 ## -------
552 mod1B=arima(lnDiesel, order=c(0,1,3))
553 cat("Modelo B \nT-ratios:",round(mod1B$coef/sqrt(diag(mod1B$var.
    coef)),2))
```

```
554 mod1BB=arima(lnDiesel, order=c(0,1,2))
555 cat("Modelo B \nT-ratios:",round(mod1BB$coef/sqrt(diag(mod1BB$var.
    coef)),2))
556 mod2B=arima(lnDiesel, order=c(2,1,0))
557 cat("Modelo B \nT-ratios:",round(mod2B$coef/sqrt(diag(mod2B$var.
    coef)),2))
558
559
560 ## -------
561 library("stargazer")
562 stargazer(mod1B, mod1BB,mod2B,title="Results", type="text",notes.
    append = FALSE,  report = "vct",
563 notes = c("t = T-statistic value = coeff/SE(coeff)"),
564   digits = 3,
565   column.labels = c("MA(3)","MA(2)","AR(2)"))
566
567
568 ## --------
569 model=mod1BB
570
571 validation(model)
572
573
574 ## ----fig.height=3,fig.width=8-----
575 #Normalidad
576
577 hist(resid,breaks=20,freq=FALSE)
578 curve(dnorm(x,mean=mean(resid),sd=sd(resid)),col=2,add=T)
579
580
581 ##Shapiro-Wilks Normality test
582 print(shapiro.test(resid(model)))
583 sw<-round(shapiro.test(resid(model))$p.value,4)
584 suppressMessages(require(nortest,quietly=TRUE,warn.conflicts=FALSE)
    )
585 ##Anderson-Darling test
586 print(ad.test(resid(model)))
587 ad<-round(ad.test(resid(model))$p.value,4)
588 suppressMessages(require(tseries,quietly=TRUE,warn.conflicts=FALSE)
    )
589 ##Jarque-Bera test
590 print(jarque.bera.test(resid(model)))
591 jb<-round(jarque.bera.test(resid(model))$p.value,4)
592
593 #Homogeneidad
594 par(mfrow=c(1,2),mar=c(3,3,3,3))
595 plot(resid,main="Residuals")
```

```
596 abline(h=0)
597 abline(h=c(-3*sd(resid),3*sd(resid)),lty=3,col=4)
598 #Square Root of absolute values of residuals (Homocedasticity)
599 scatter.smooth(sqrt(abs(resid)),main="Square Root of Absolute
        residuals",
600                    lpars=list(col=2))
601
602 #Independencia
603 par(mfrow=c(1,2),mar=c(3,3,3,3))
604 acf(resid,ylim=c(-0.5,1),lag.max=60,col=c(2,rep(1,11)),lwd=1, main
        = "Resid ACF")
605 pacf(resid,ylim=c(-0.5,1),lag.max=60,col=c(rep(1,11),2),lwd=1, main
         = "Resid PACF")
606
607 ## ----fig.height=7,fig.width=8---
608 par(mar=c(2,2,1,1))
609 tsdiag(model,gof.lag=7*12)
610
611
612 ## ------------------
613
614 ultim=c(2020,12)                          #Dic 2020
615
616 serie1 <- x1[1:150,] #Serie competo: 2010-2022 abril
617 lnserie1 <- log(serie1$Diesel)
618 serie2 <- x1[1:132,] #Serie sin observacion del ultimo a o:
        2010-20121
619 lnserie2 <- log(serie2$Diesel)
620
621 #Fit the model to the complete series: lnserie1
622 mod1B=arima(lnserie1,order=c(0,1,2))
623 #Fit the model to the subset series (without 2020 data): lnserie2
624 mod1B2=arima(lnserie2,order=c(0,1,2))
625
626
627 ## ----------
628 pred=predict(mod1B2,n.ahead=18)                                    #
        outputs point predictions and corresponding standard errors:for
        year 2019
629 pr<-ts(c(tail(lnserie2,1),pred$pred),start=ultim,freq=12)  #point
        predictions
630
631 se<-ts(c(0,pred$se),start=ultim,freq=12)                    #
        Standard errors for point predictions
632
633 #Prediction Intervals (back transformed to original scale using exp
        -function)
```

```
634 tl<-ts(exp(pr-1.96*se),start=ultim,freq=12)
635 tu<-ts(exp(pr+1.96*se),start=ultim,freq=12)
636 pr<-ts(exp(pr),start=ultim,freq=12)              #predictions in
    original scale
637
638 #Plot of the original series and out-of-sample predictions: only
    time window 2015-2019 shown
639 diesel <- ts(x$Diesel[1:150],start = 2010,frequency = 12)
640 ts.plot(diesel,tl,tu,pr,lty=c(1,2,2,1),col=c(1,4,4,2),xlim=ultim
    [1]+c(-2,+3),type="o",main="Model ARIMA(0,1,2)")
641 abline(v=(ultim[1]-3):(ultim[1]+3),lty=3,col=4)
642 diferencia=round(diesel-pr,6)
643 obs=window(diesel,start=ultim)
644 TV = round(((diesel-pr)/pr)*100,1)
645
646
647 ## ----------
648 Cal <- ts(x1$`Diesel Cal.`,start = 2010,frequency = 12)
649 acf(Cal,ylim=c(-1,1),lag.max=60,col=c(2,rep(1,11)),lwd=2,xlab = "
    Retardo")
650
651 ## ----fig.width=10,fig.height=5------
652 par(mfrow=c(1,2))
653 boxplot(Cal~x1$any, ylab = "Diesel para calefacci n", xlab="A o",
    main="Box-Plot")
654 monthplot(Cal,ylab = "Diesel para calefacci n",main = " Monthplot"
    ,xlab="Mes")
655
656
657
658 ## ----------
659 lnCal=log(Cal)
660 d1lnCal<-diff(lnCal)
661
662
663 ## ----------
664 par(mfrow=c(1,2))
665 acf(d1lnCal,ylim=c(-1,1),lag.max=60,col=c(2,rep(1,11)),lwd=2,xlab="
    Retardo")
666 pacf(d1lnCal,ylim=c(-1,1),lag.max=60,col=c(rep(1,11),2),lwd=2,xlab=
    "Retardo")
667
668
669
670 ## ----------
671 mod1A=arima(d1lnCal, order=c(0,0,2),seasonal=list(order=c(0,0,2),
    period=12))
```

```
672 mod2A=arima(d1lnCal, order=c(2,0,0),seasonal=list(order=c(0,0,2),
        period=12))
673
674
675
676 library("stargazer")
677 stargazer(mod1A, mod2A,title="Results", type="text",notes.append =
        FALSE,  report = "vct",
678 notes = c("t = T-statistic value = coeff/SE(coeff)"),
679   digits = 3,
680   column.labels = c("MA(2)SMA(2)","AR(2)SMA(2)"))
681
682
683 ## ----------
684 mod1B=arima(lnCal, order=c(0,1,2),seasonal=list(order=c(0,0,2),
        period=12))
685 mod1BBB=arima(lnCal, order=c(0,1,1))
686 mod2BB=arima(lnCal, order=c(2,1,0))
687 mod2B=arima(lnCal, order=c(2,1,0),seasonal=list(order=c(0,0,2),
        period=12))
688
689
690 stargazer(mod1B,mod1BBB,mod2B,mod2BB,title="Results", type="text",
        notes.append = FALSE,  report = "vct",
691 notes = c("t = T-statistic value = coeff/SE(coeff)"),
692   digits = 3,
693   column.labels = c("MA(2)SMA(2)","MA(1)","AR(2)SMA(2)","AR(2)"))
694
695
696 ## ---------
697 model = mod1BBB
698 validation(model)
699
700
701 ## ----fig.height=3,fig.width=8--------
702 #Normalidad
703 model = mod1BBB
704 resid=model$residuals
705 par(mfrow=c(1,2),mar=c(3,3,3,3))
706 qqnorm(resid)
707 qqline(resid,col=2,lwd=2)
708 hist(resid,breaks=20,freq=FALSE)
709 curve(dnorm(x,mean=mean(resid),sd=sd(resid)),col=2,add=T)
710
711
712 ##Shapiro-Wilks Normality test
713 print(shapiro.test(resid(model)))
```

```
sw<-round(shapiro.test(resid(model))$p.value,4)
suppressMessages(require(nortest,quietly=TRUE,warn.conflicts=FALSE))
##Anderson-Darling test
print(ad.test(resid(model)))
ad<-round(ad.test(resid(model))$p.value,4)
suppressMessages(require(tseries,quietly=TRUE,warn.conflicts=FALSE))
##Jarque-Bera test
print(jarque.bera.test(resid(model)))
jb<-round(jarque.bera.test(resid(model))$p.value,4)

#Homogeneidad
par(mfrow=c(1,2),mar=c(3,3,3,3))
plot(resid,main="Residuals")
abline(h=0)
abline(h=c(-3*sd(resid),3*sd(resid)),lty=3,col=4)
#Square Root of absolute values of residuals (Homocedasticity)
scatter.smooth(sqrt(abs(resid)),main="Square Root of Absolute residuals",
                lpars=list(col=2))

#Independencia
par(mfrow=c(1,2),mar=c(3,3,3,3))
acf(resid,ylim=c(-0.5,1),lag.max=60,col=c(2,rep(1,11)),lwd=1, main = "Resid ACF")
pacf(resid,ylim=c(-0.5,1),lag.max=60,col=c(rep(1,11),2),lwd=1, main = "Resid PACF")

## ----fig.height=7,fig.width=8-----
par(mar=c(2,2,1,1))
tsdiag(model,gof.lag=7*12)


## ----fig.height=5,fig.width=8-------
?par
par(mfrow=c(2,2),mar=c(3,3,3,3))
model = mod1BBB
resid=model$residuals
qqnorm(resid)
qqline(resid,col=2,lwd=2)

scatter.smooth(sqrt(abs(resid)),main="Square Root of Absolute residuals",
                lpars=list(col=2))
acf(resid,ylim=c(-0.5,1),lag.max=60,col=c(2,rep(1,11)),lwd=1, main = "Resid ACF")
```

```
754 pacf(resid,ylim=c(-0.5,1),lag.max=60,col=c(rep(1,11),2),lwd=1, main
        = "Resid PACF")
755
756
757 ## ----------------
758
759
760 ultim=c(2020,12)                          #Dic 2020
761
762 serie1 <- x[1:150,] #Serie competo: 2010-2022 abril
763 lnserie1 <- log(serie1$`Diesel Cal.`)
764 serie2 <- x[1:132,] #Serie sin observacion del ultimo a o:
       2010-20121
765 lnserie2 <- log(serie2$`Diesel Cal.`)
766
767
768 #Fit the model to the complete series: lnserie1
769 (mod=arima(lnserie2,order=c(0,1,1)))
770
771
772
773 ## ----------------------
774 pred=predict(mod,n.ahead=18)                                  #outputs
       point predictions and corresponding standard errors:for year
       2019
775 pr<-ts(c(tail(lnserie2,1),pred$pred),start=ultim,freq=12)  #point
       predictions
776
777 se<-ts(c(0,pred$se),start=ultim,freq=12)                     #
       Standard errors for point predictions
778
779 #Prediction Intervals (back transformed to original scale using exp
       -function)
780 tl<-ts(exp(pr-1.96*se),start=ultim,freq=12)
781 tu<-ts(exp(pr+1.96*se),start=ultim,freq=12)
782 pr<-ts(exp(pr),start=ultim,freq=12)              #predictions in
       original scale
783
784 #Plot of the original series and out-of-sample predictions: only
       time window 2015-2019 shown
785 cal <- ts(x$`Diesel Cal.`[1:150],start = 2010,frequency = 12)
786 ts.plot(cal,tl,tu,pr,lty=c(1,2,2,1),col=c(1,4,4,2),xlim=ultim[1]+c
       (-2,+3),type="o",main="Model ARIMA(0,1,1)")
787 abline(v=(ultim[1]-3):(ultim[1]+3),lty=3,col=4)
788 diferencia=round(cal-pr,6)
789 obs=window(cal,start=ultim)
```

```
790 TV = round(((cal-pr)/pr)*100,1)
```

Listing C.1: Cdigo aplicado en R para el proyecto