Lab 2 - Series de tiempo

Oscar Godoy - Rafael Dubois

2022-08-05

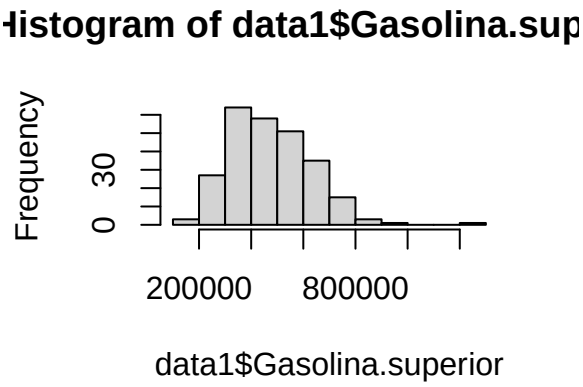
Datos y limpieza

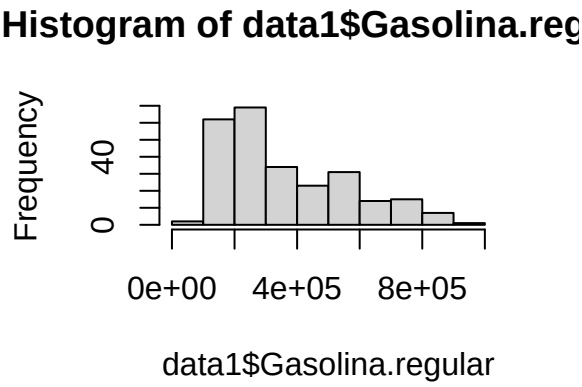
Los sets de datos incluyen información sobre importación y consumo de gasolina en Guatemala por los últimos 22 años. Iniciando en enero del 2001 hasta junio del 2022, se registraron la cantidad de galones de gasolina superior, regular y diesel que se importó y consumió en el país de forma mensual. El set de datos 1 contiene información sobre importación, mientras que el segundo de consumo. Los datos se encuentran bajo la carpeta con nombre "Estadísticas de Comercialización de Hidrocarburos", disponible en el sitio web del Ministerio de Energía y Minas.

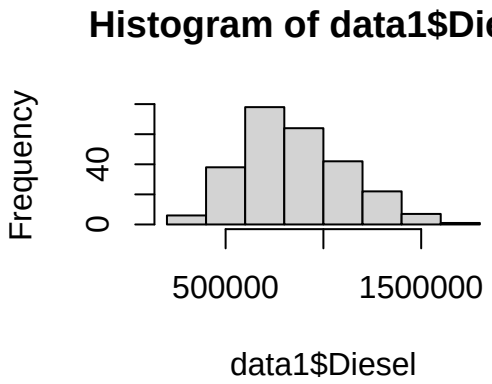
Inicialmente, los conjuntos de datos se presentan en archivo excel. Estos se llevaron a un formato csv y se filtraron sus estadísticas relevantes. A cada set de datos se le agregó una columna correspondiente al año y al mes. Al finalizar, obtenemos dos dataframes (uno de importación y otro de consumo) cada uno con 258 filas (una por cada mes desde enero 2001 a junio 2022) y cada uno con 5 variables: Year, Month, Gasolina.superior, Gasolina.regular y Diesel.

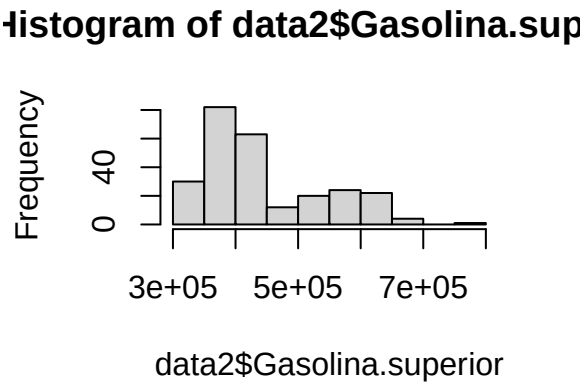
Exploración de datos

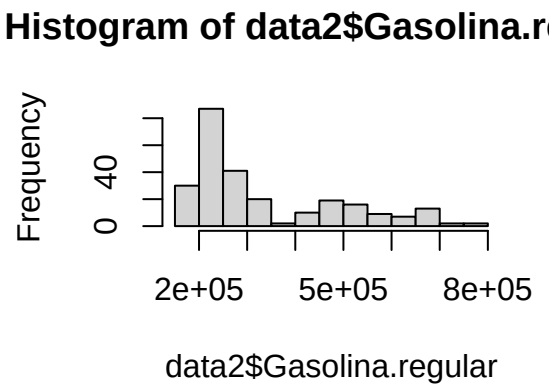
Contamos con 6 variables de interés, todas continuas. Entonces, investigamos la distribución de dichas variables con histogramas. Podemos ver que todas las distribuciones se encuentran sesgadas a la derecha. No se observa que los datos se encuentren distribuidos normalmente, aunque en su mayoría puede decirse que las curvas son mesocúrticas.

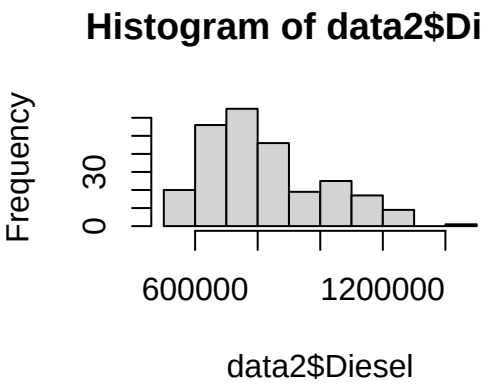




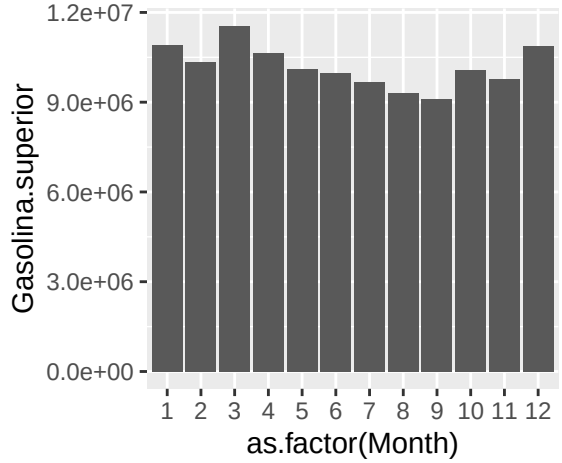


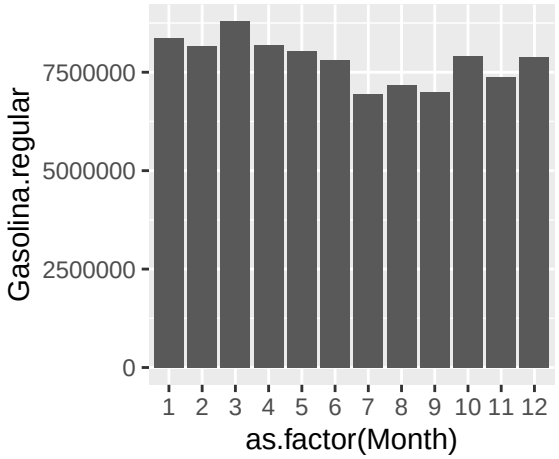


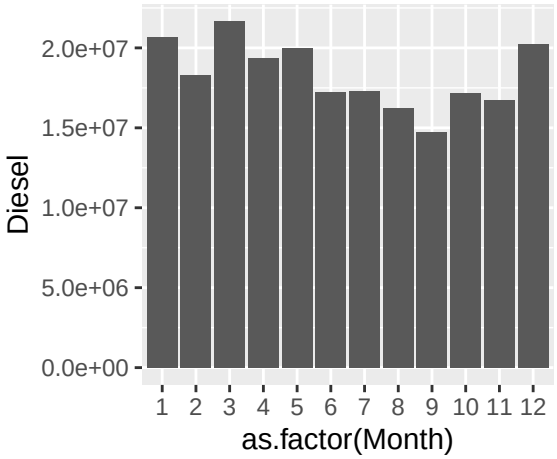


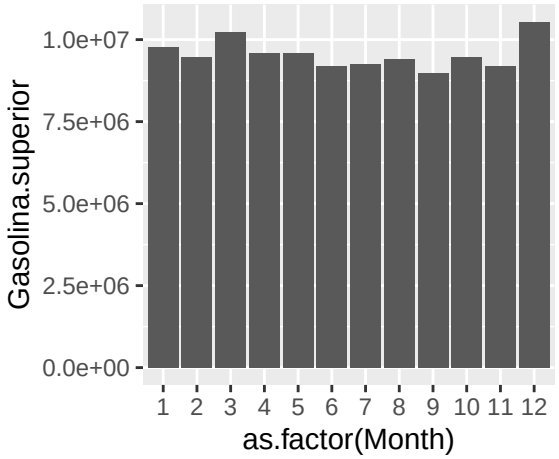


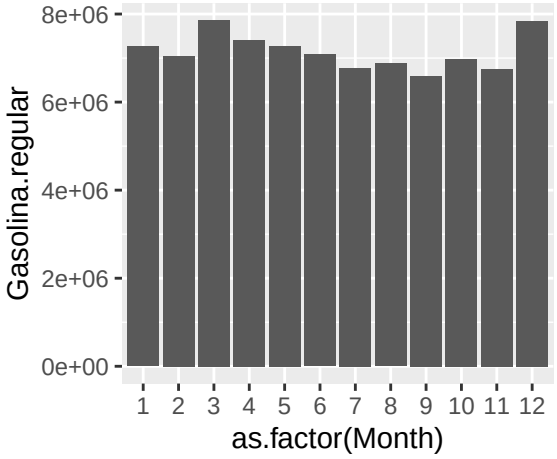
Luego, procedemos a analizar la data por mes. Curiosamente notamos que, tanto el volumen de importación como de consumo, ven sus máximos valores al inicio y al final del año. Alrededor del mes de septiembre el país parece importar y consumir menos hidrocarburos. Esta tendencia se ve más marcada en la importación de gasolina superior, la importación de diesel y el consumo de diesel.

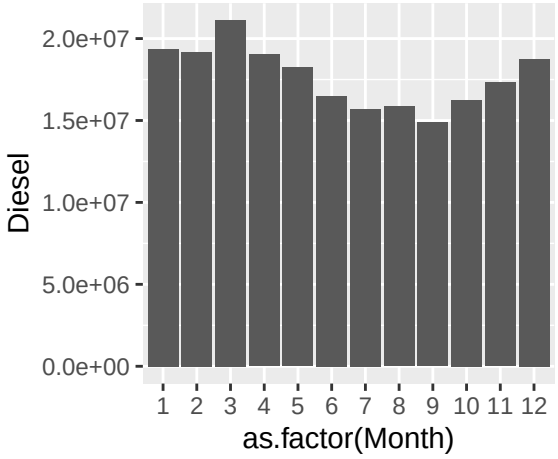




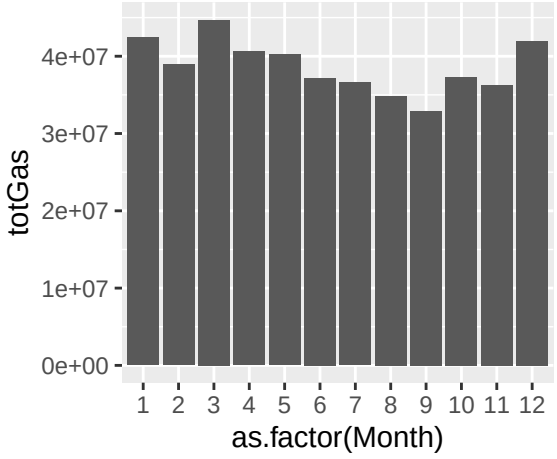


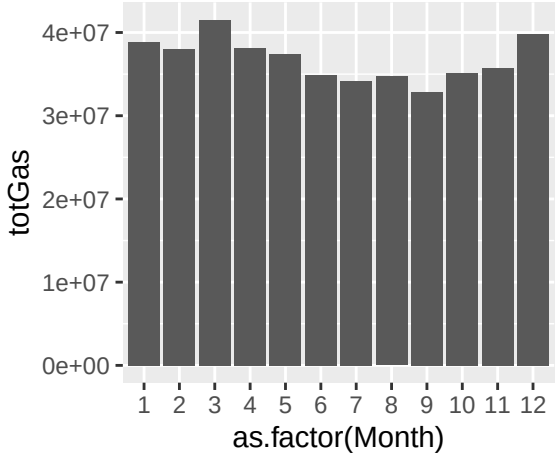






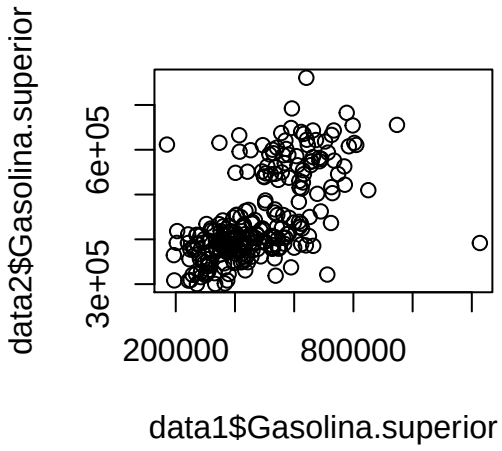
Totalizando la cantidad de gasolina importada o consumida por mes, tomando en cuenta cada uno de los tres tipos de gasolina, obtenemos dos gráficos que resumen la información de los anteriores diagramas de barras. Como era de esperarse, el volumen de importación y consumo de hidrocarburos sigue siendo predominante en los meses de enero, marzo y diciembre. Esto sugiere una cierta estacionalidad del set de datos, y queda pendiente verificarlo al estudiar las series de tiempo de los datos.

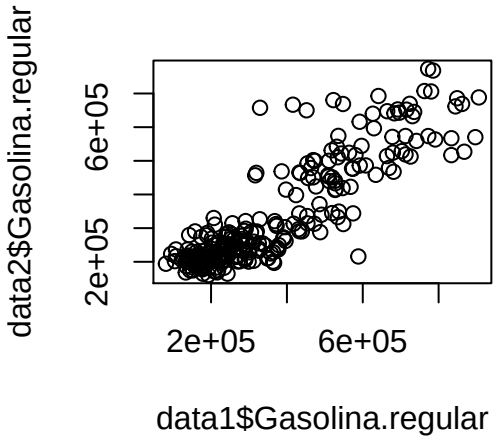


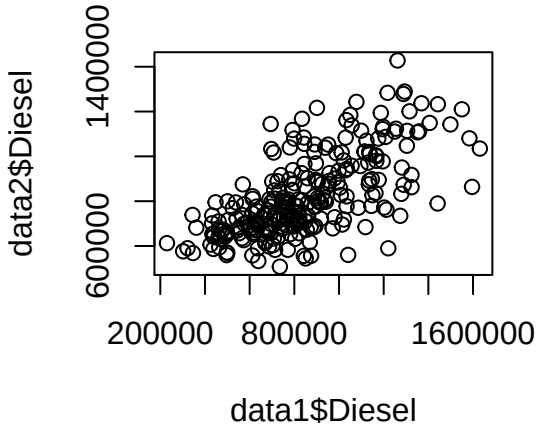


Finalmente, notamos la relación entre la importación y el consumo de gasolina para cada uno de los tipos de gasolina. Graficando en un diagrama de dispersión la entrada de importación y consumo para cada tipo de

gasolina, por cada mes, obtenemos la siguiente nube de puntos. Vemos que hay una correlación positiva entre cada par de variables. Esto es de esperarse, pues la decisión del país en importar más hidrocarburos se ve reflejada en una mayor demanda de este producto por el público.







Series de tiempo

A partir de las 6 variables de interés investigadas anteriormente creamos 6 series de tiempo, con los datos desde enero del 2001 hasta junio 2022. Cada una de estas series contiene una frecuencia de 12 meses. Además, a partir de este momento también creamos series de tiempo de entrenamiento, con las cuales trabajaremos principalmente durante el resto del laboratorio. Estas series de tiempo inician en enero del 2001 pero finalizan en diciembre del 2019. De nuevo, cada una de estas series de entrenamiento también posee una frecuencia de 12 meses.

```
start(trainSup1)
```

```
## [1] 2001    1
```

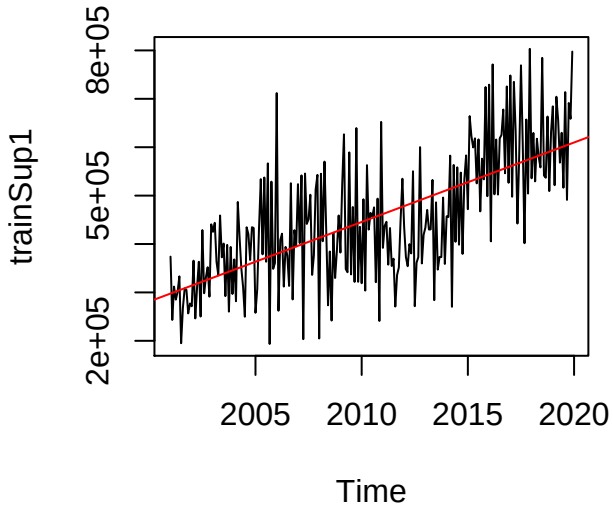
```
end(trainSup1)
```

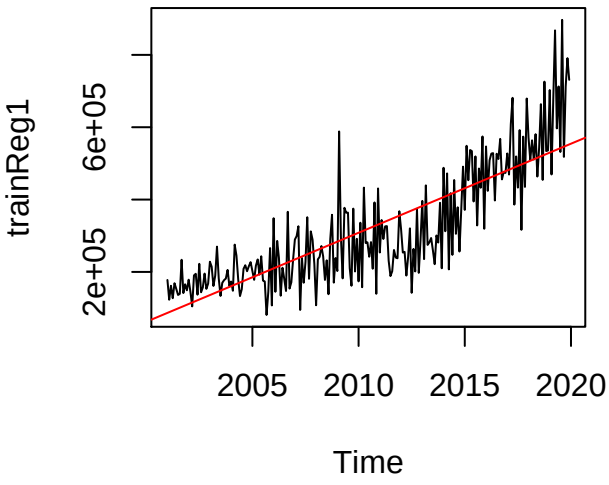
```
## [1] 2019   12
```

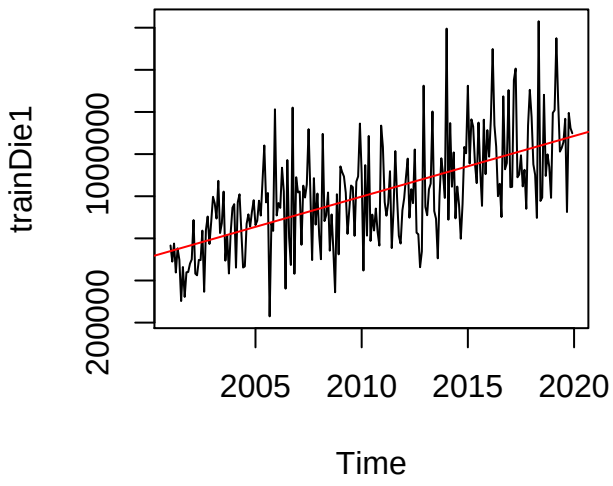
```
frequency(trainSup1)
```

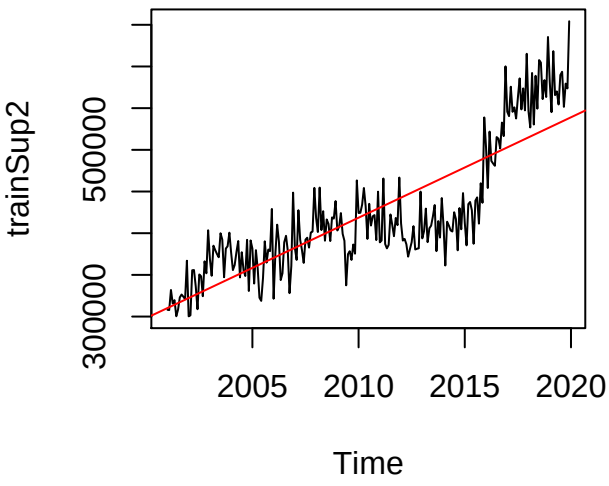
```
## [1] 12
```

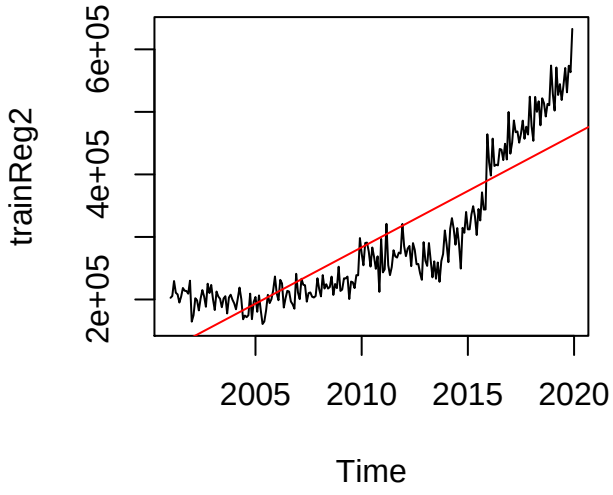
Ahora, damos un primer vistazo a nuestras series de tiempo de entrenamiento. Cada una de las 6 series posee una tendencia, es decir, ninguna es constante en media. El consumo e importación de hidrocarburos parecen ir en aumento desde el 2001 hasta el 2019 en el país. Por otra parte, vemos que los volumenes de gasolina son muy variantes por año. Cada set de datos parece poseer estacionalidad, pero la varianza de los set de datos no parece ser constante. Esto es especialmente cierto para las series de tiempo de importación, las cuales oscilan de forma más variada.

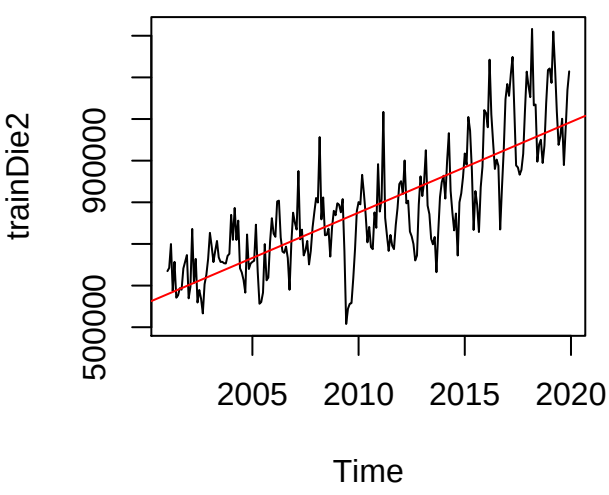




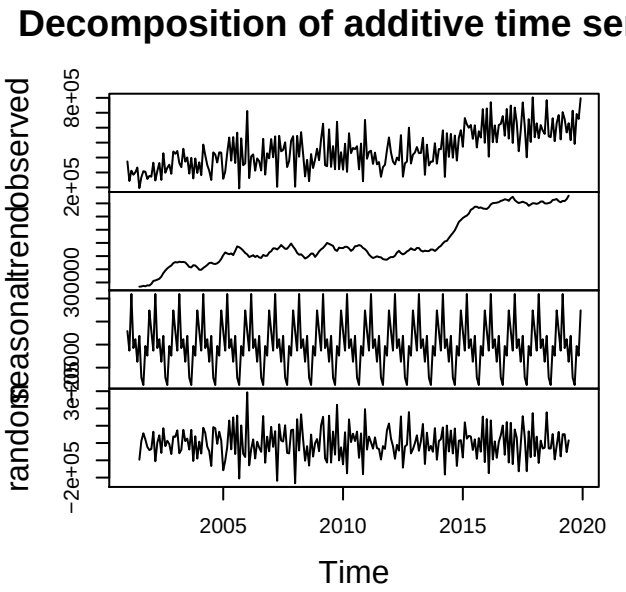


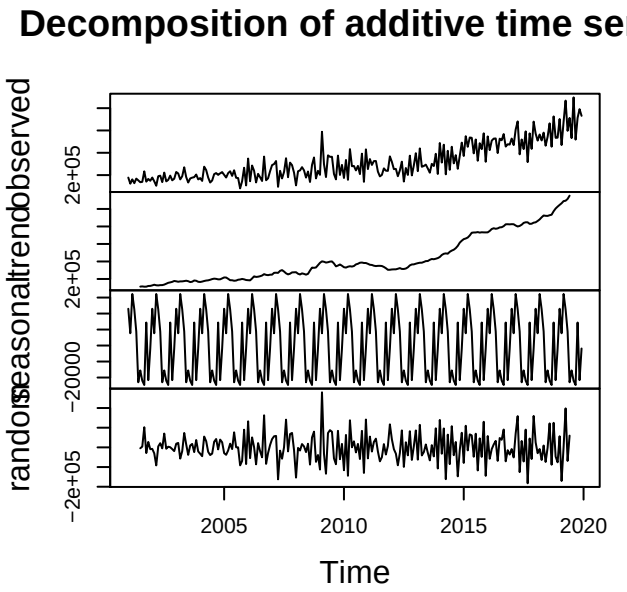


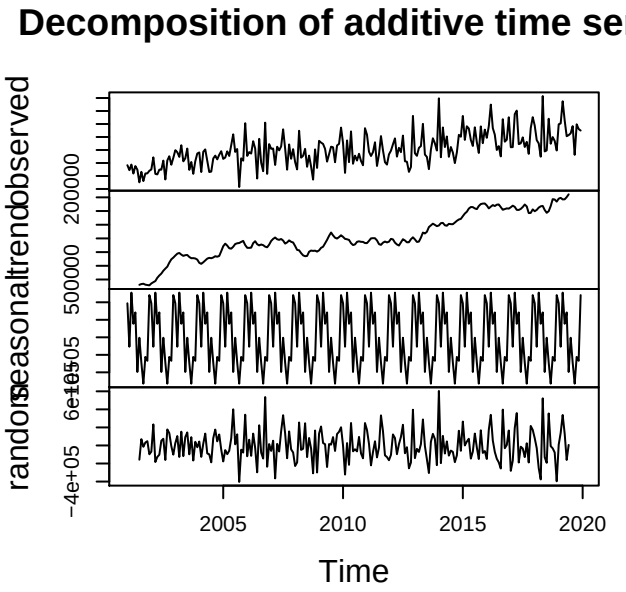


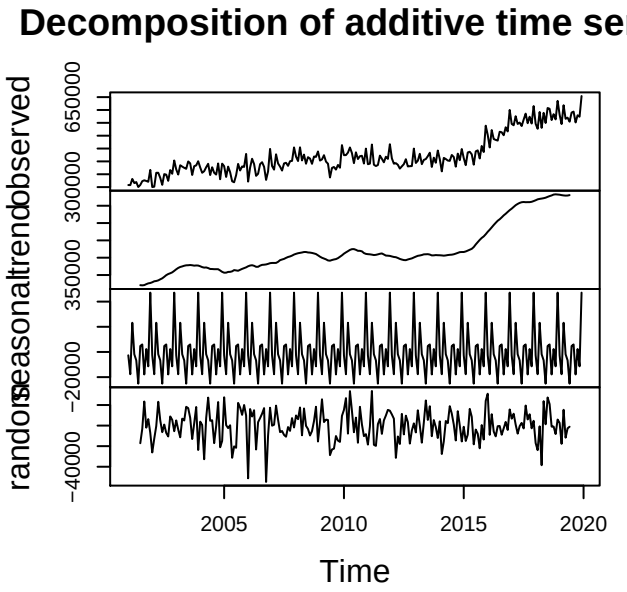


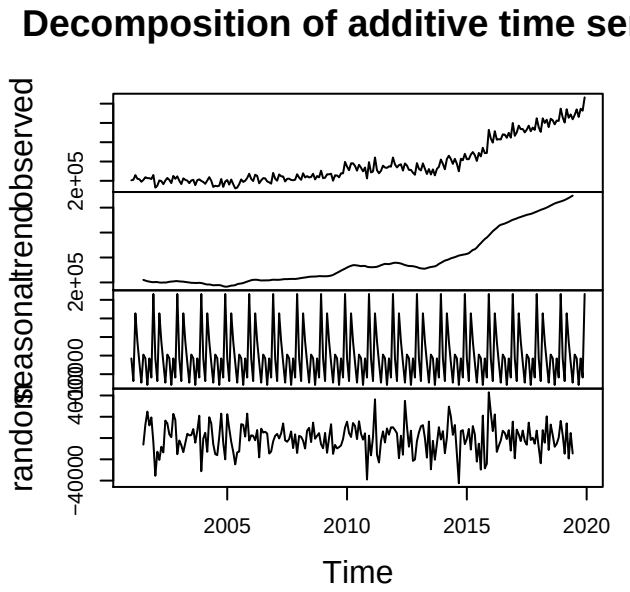
Procedemos a descomponer las series vistas en sus componentes de tendencia, estacionalidad y aleatoriedad. Como esperabamos, cada serie posee una media en aumento durante los 19 años analizados. Por su parte, puede verse que la estacionalidad es tentativamente anual. Esto se observa al contar la cantidad de picos o ciclos que hay entre intervalos de tiempo definidos en el gráfico de estacionalidad. Para intervalos de tiempo de 5 años, pueden contarse 5 de estos picos.

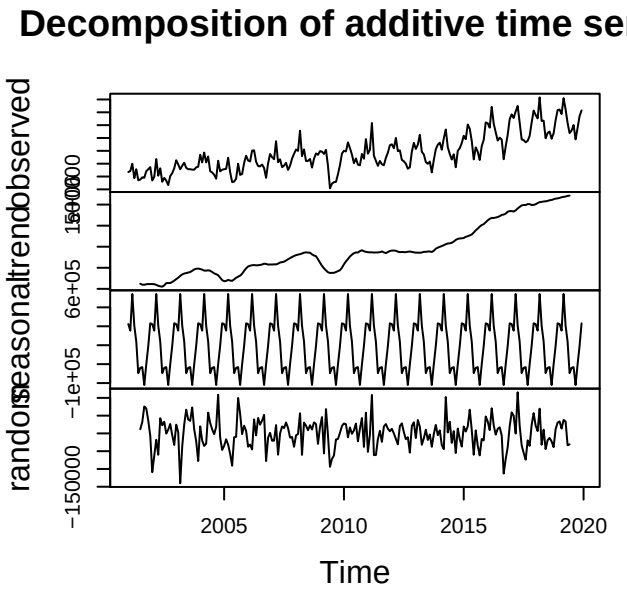

Decomposition of additive time seri
observed
trend
seasonal
random
2e+05
8e+05
300000
3e-08
−2e+05
2005
2010
2015
2020
Time


Decomposition of additive time seri
observed
trend
seasonal
random
2e+05
2e+05
−20000
−2e+05
2005
2010
2015
2020
Time


Decomposition of additive time seri
observed
trend
seasonal
random
200000
500000
6e-10
−4e+05
2005
2010
2015
2020
Time

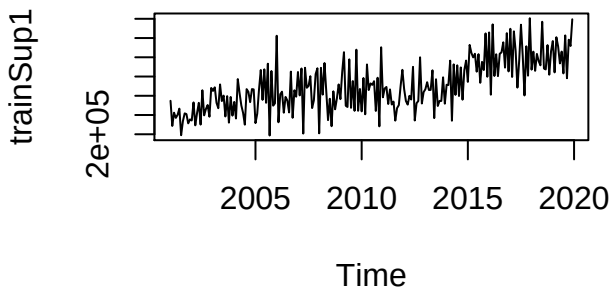

Decomposition of additive time seri
observed
trend
seasonal
random
650000
300000
350000
−20000
−40000
2005
2010
2015
2020
Time

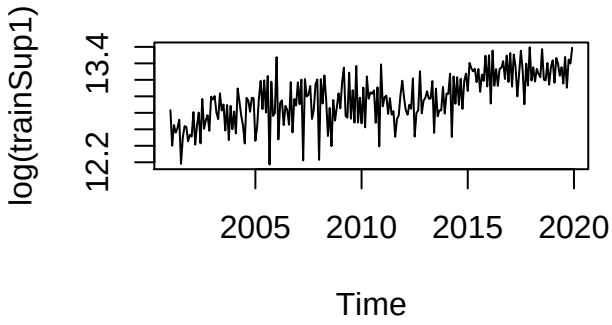


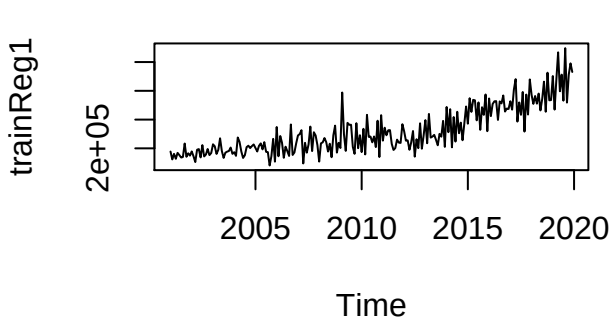


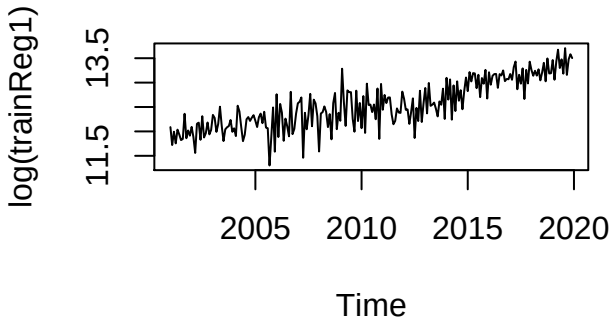
Búsqueda de parámetros ARIMA

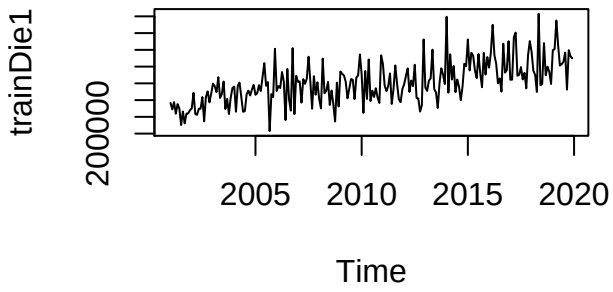
Dado que ninguna de nuestras variables posee varianza constante, estas aún no se encuentran listas para ser introducidas a un modelo ARIMA. En especial, vemos que las series de tiempo de importacion poseen mucha variabilidad. La transformación logarítmica mejora un poco esta condición, pero esto no es suficiente. Idealmente, necesitamos otra transformación que anule por completo esta variabilidad en varianza, pero por el momento analizamos las series de consumo. Estas sí parecen esatbilizarse en media al aplicarles dicha transformación.

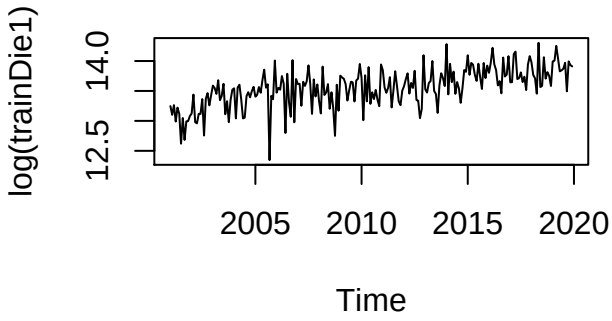


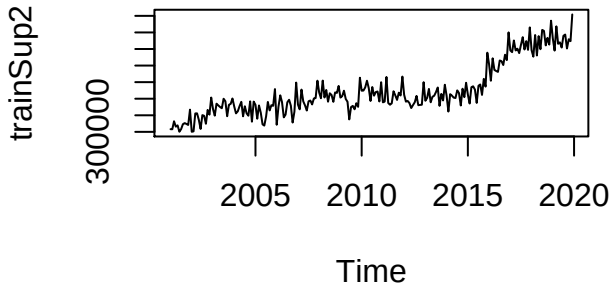


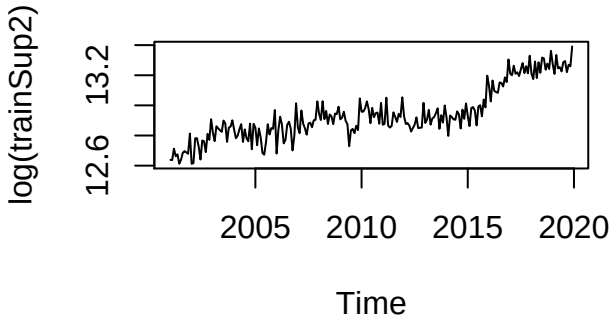

trainReg1
2e+05
2005
2010
2015
2020
Time

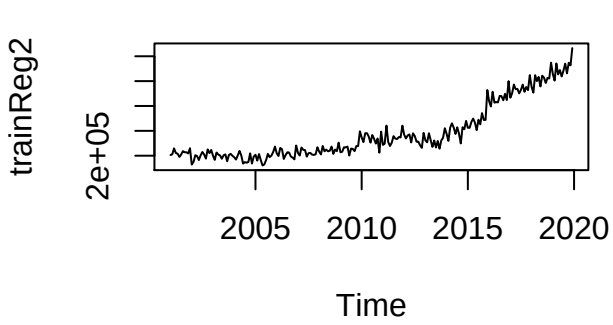

log(trainReg1)
11.5
13.5
2005
2010
2015
2020
Time

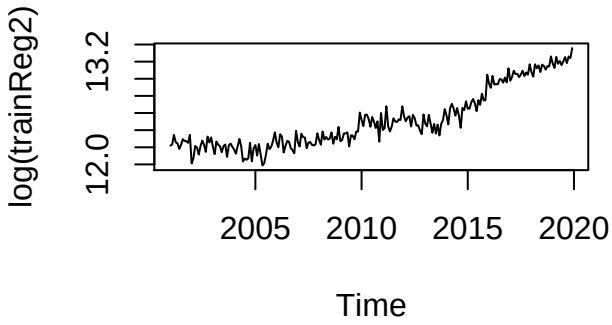

trainDie1
200000
2005
2010
2015
2020
Time

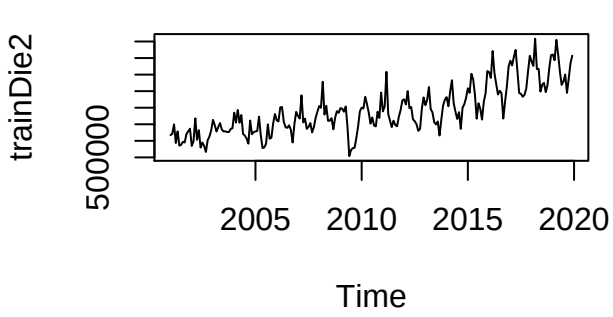

log(trainDie1)
12.5
14.0
2005
2010
2015
2020
Time

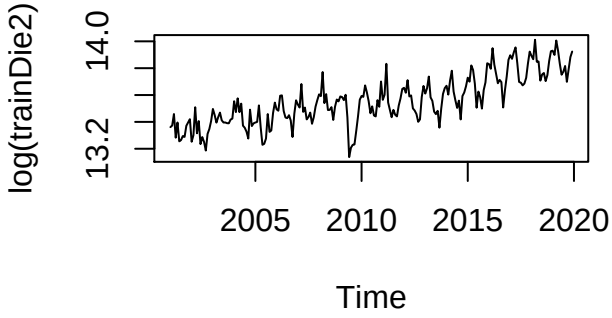

trainSup2
300000
2005
2010
2015
2020
Time


log(trainSup2)
12.6
13.2
2005
2010
2015
2020
Time









Por otra parte, corroboramos si las series de tiempo son estacionarias en media. Realizamos la prueba aumentada de Dickey-Fuller y la prueba para la existencia de raíces unitarias. Para cada una de las series, los p-valores obtenidos son mucho mayores a 0.01, indicando la no estacionariedad en media de las series. Por lo tanto, diferenciamos una vez cada una de las series y corremos las mismas pruebas de nuevo. Esto resuelve el problema en cada una de las series, pues el p-valor se ve reducido a menos de 0.01 en todos los casos. Con esto, recuperamos que $d = 1$ para cada serie. Por brevedad, mostramos únicamente las pruebas realizadas para la serie de tiempo de importación de gasolina superior, para su versión original y su versión diferenciada.

```
# sup1: d=1
adfTest(trainSup1)
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 1
##   STATISTIC:
##     Dickey-Fuller: -0.485
##   P VALUE:
##     0.4613
##
## Description:
```

```
##  Fri Aug 05 14:42:00 2022 by user: oscar
```

```
unitrootTest(trainSup1)
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 1
##   STATISTIC:
##     DF: -0.485
##   P VALUE:
##     t: 0.505
##     n: 0.5717
##
## Description:
##  Fri Aug 05 14:42:00 2022 by user: oscar
```

```
adfTest(diff(trainSup1))
```

```
## Warning in adfTest(diff(trainSup1)): p-value smaller than printed p-value
```

```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 1
##   STATISTIC:
##     Dickey-Fuller: -19.5518
##   P VALUE:
##     0.01
##
## Description:
##  Fri Aug 05 14:42:00 2022 by user: oscar
```

```
unitrootTest(diff(trainSup1))
```

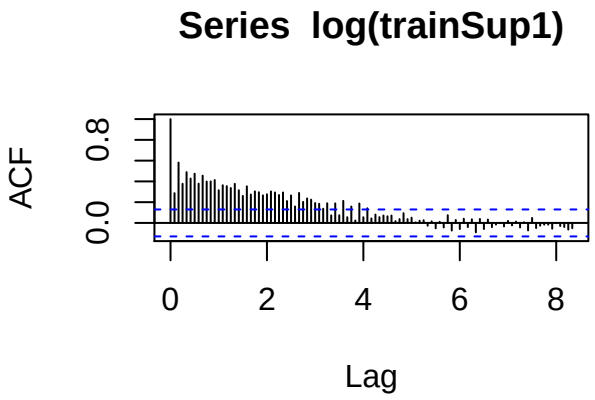
```
##
## Title:
##  Augmented Dickey-Fuller Test
##
## Test Results:
##   PARAMETER:
##     Lag Order: 1
##   STATISTIC:
##     DF: -19.5518
##   P VALUE:
##     t: < 2.2e-16
##     n: 0.001727
##
## Description:
##  Fri Aug 05 14:42:00 2022 by user: oscar
```

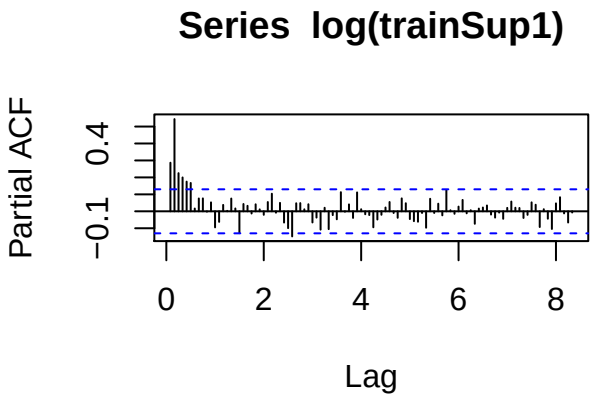
Luego, buscamos los parámetros p y q para los modelos ARIMA. Para ello nos apoyamos con los gráficos de correlación y de correlación parcial para cada serie. Buscamos cuantos retardos le toma a cada gráfico ser anulado, y a partir de estas cifras anotamos los correspondientes parámetros p y q. Para cada serie, encontramos:

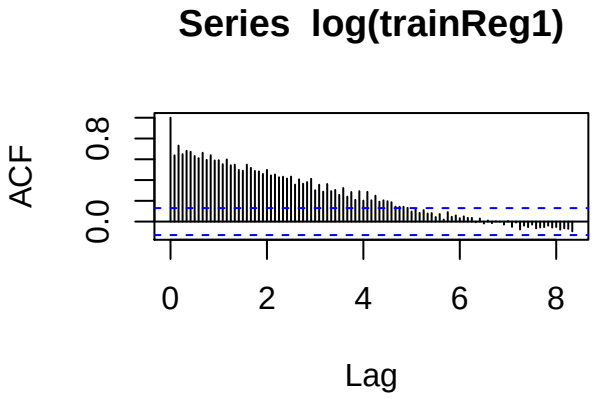
- Importación - Superior: $p = 1$ y $q = 4$.
- Importación - Regular: $p = 1$ y $q = 5$.

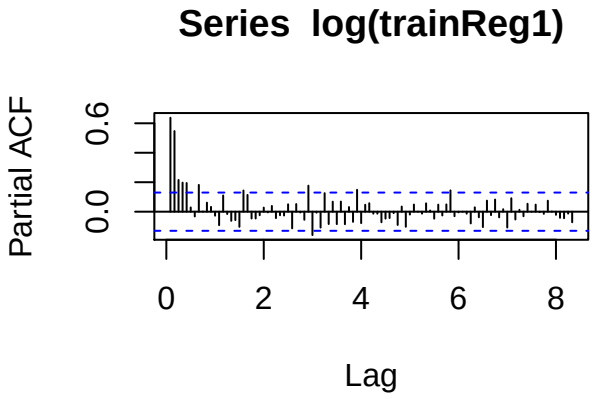
- Importación - Diesel: $p = 2$ y $q = 4$.
- Consumo - Superior: $p = 1$ y $q = 4$.
- Consumo - Regular: $p = 1$ y $q = 5$.

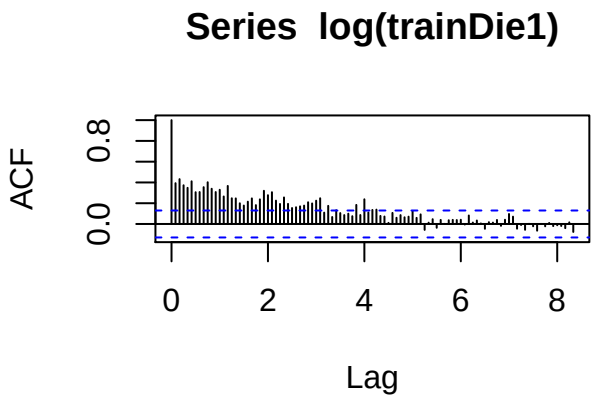
- Consumo - Diesel: $p = 1$ y $q = 6$.

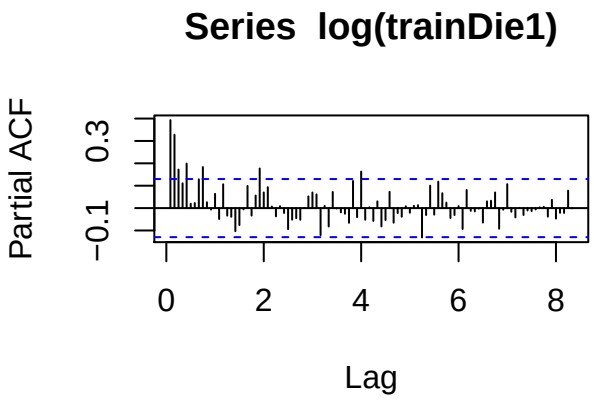


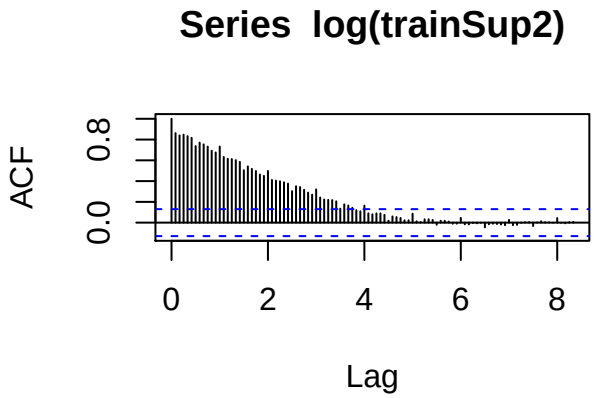


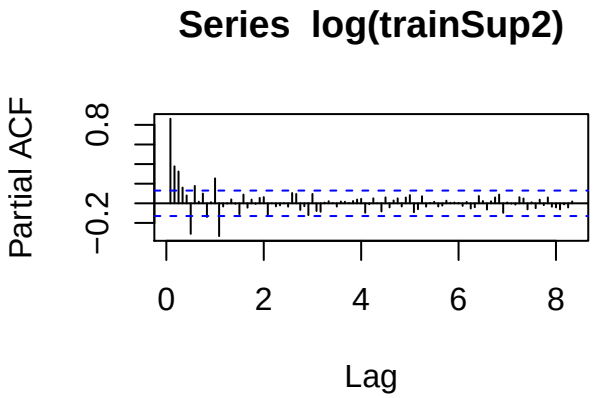


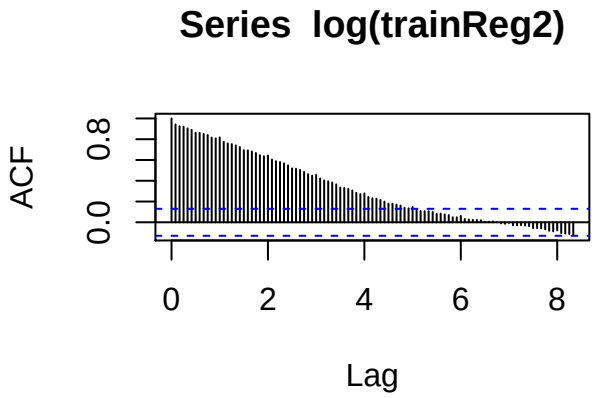


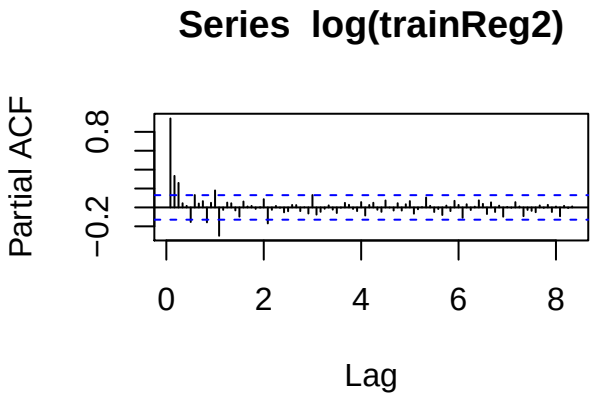

Series log(trainDie1)
ACF
0.0
0.8
0
2
4
6
8
Lag

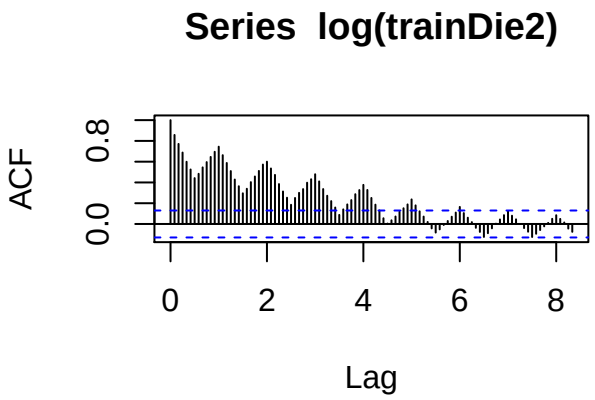

Series log(trainDie1)
Partial ACF
−0.1
0.3
0
2
4
6
8
Lag

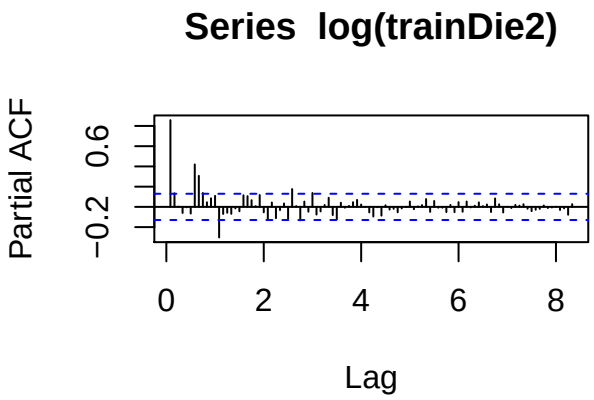

Series log(trainSup2)
ACF
0.0
0.8
0
2
4
6
8
Lag


Series log(trainSup2)
Partial ACF
−0.2
0.8
0
2
4
6
8
Lag


Series log(trainReg2)
ACF
0.0
0.8
0
2
4
6
8
Lag

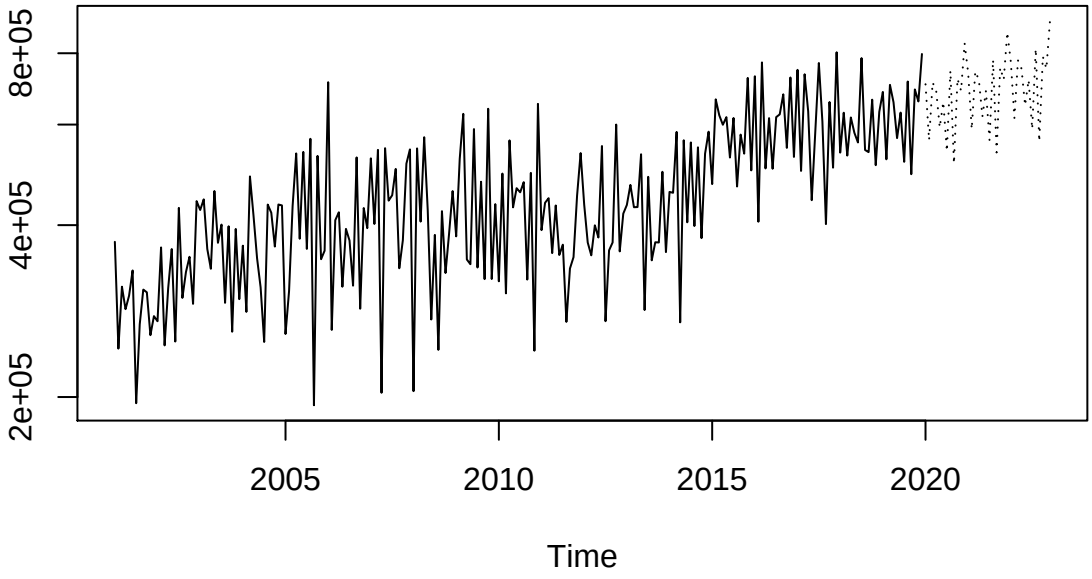

Series log(trainReg2)
Partial ACF
−0.2
0.8
0
2
4
6
8
Lag

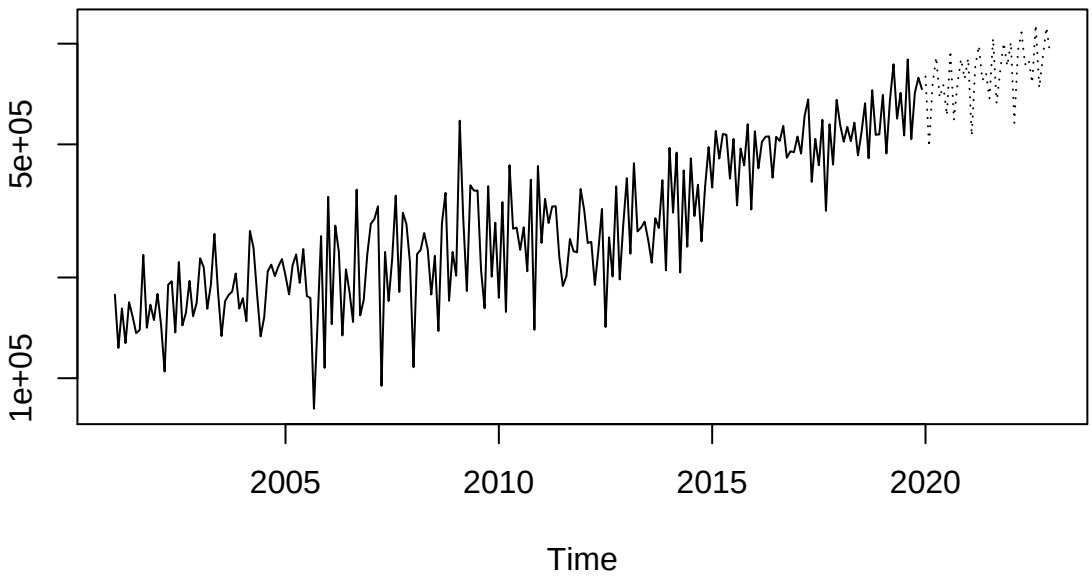


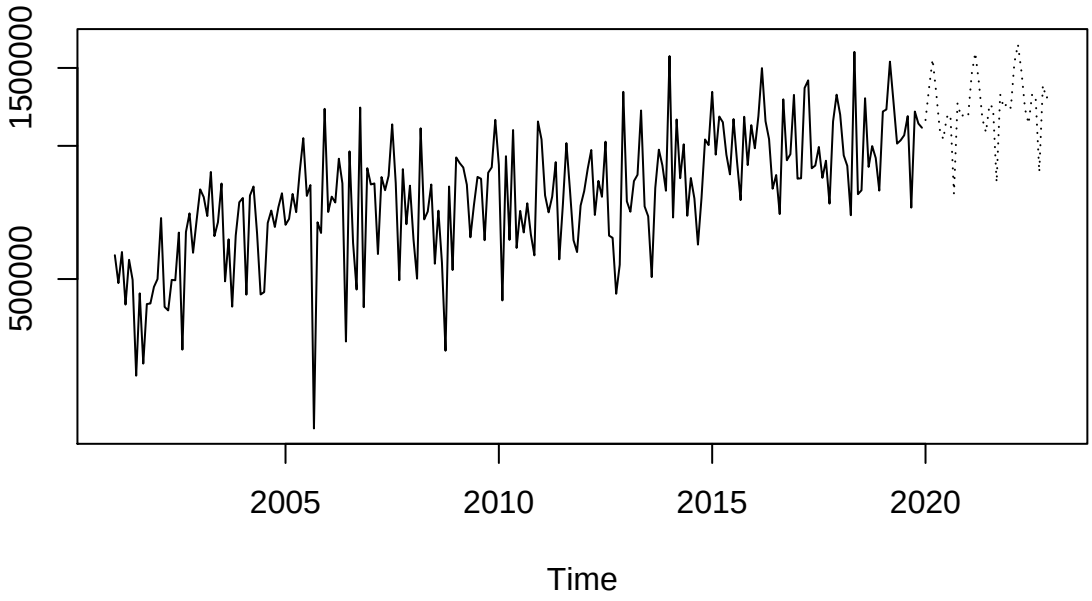


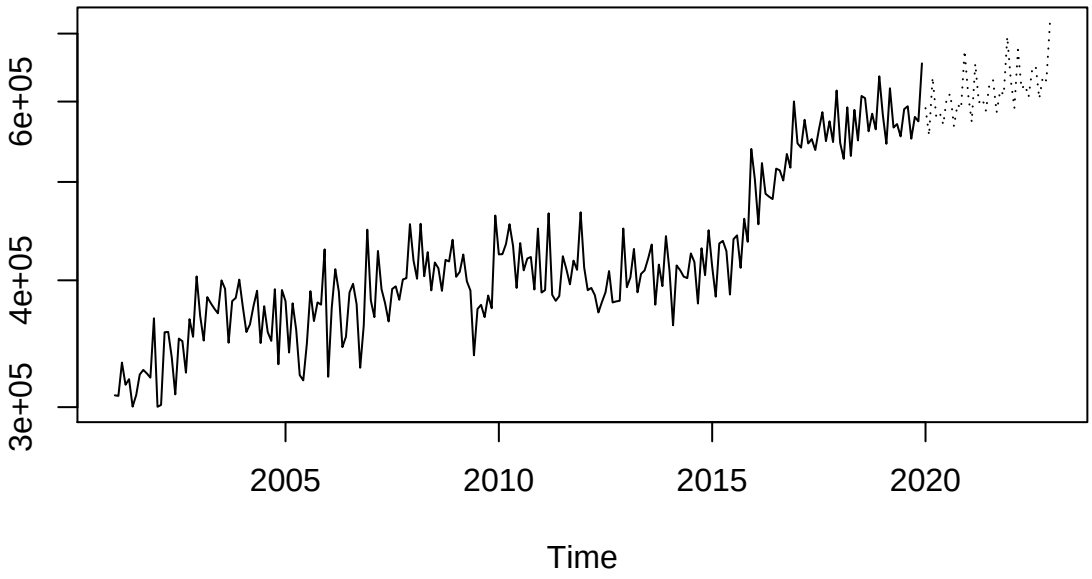
Construcción de los modelos y predicciones

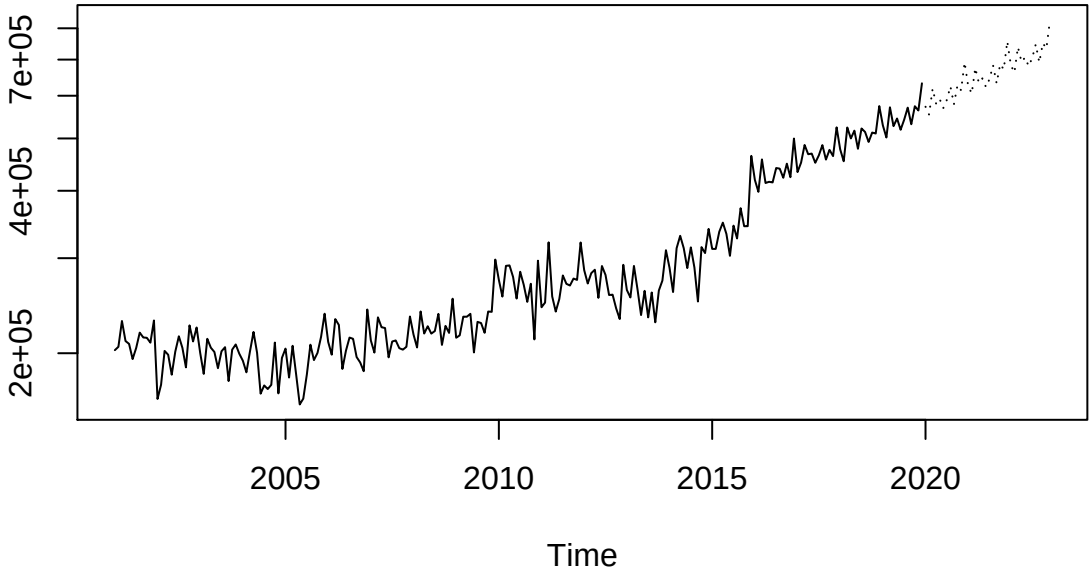
Con los parametros encontrados de los modelos ARIMA, introducimos manualmente estos para la construcción de nuestros 6 modelos. Empleamos estos modelos para predecir los volumenes de importación y consumo de hidrocarburos por los próximos 3 años, es decir, para los años 2020, 2021 y 2022. Los gráficos de predicción pueden verse a continuación.

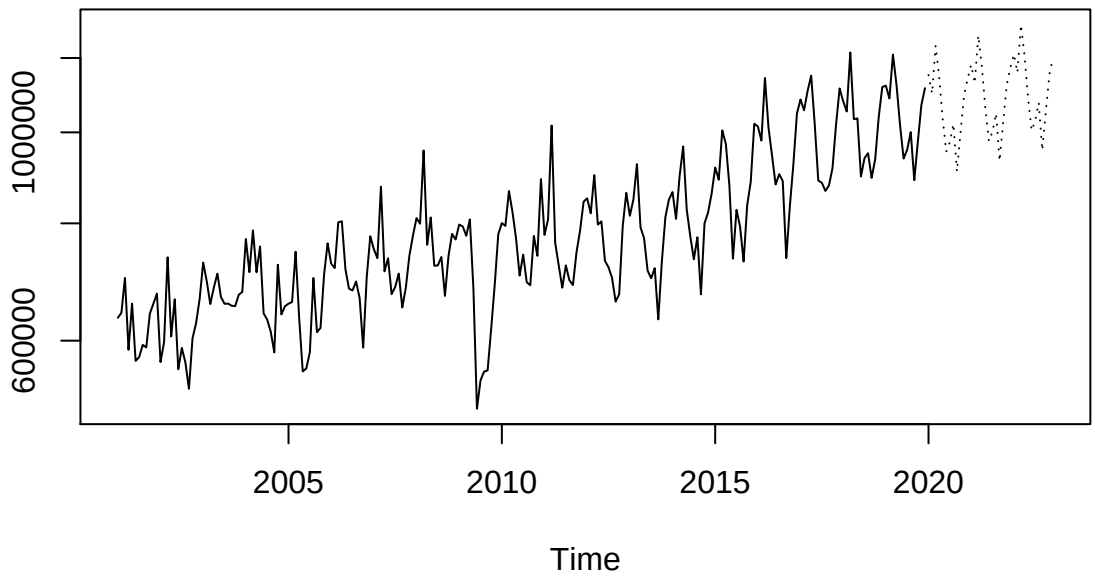



1e+05
5e+05
2005
2010
2015
2020
Time

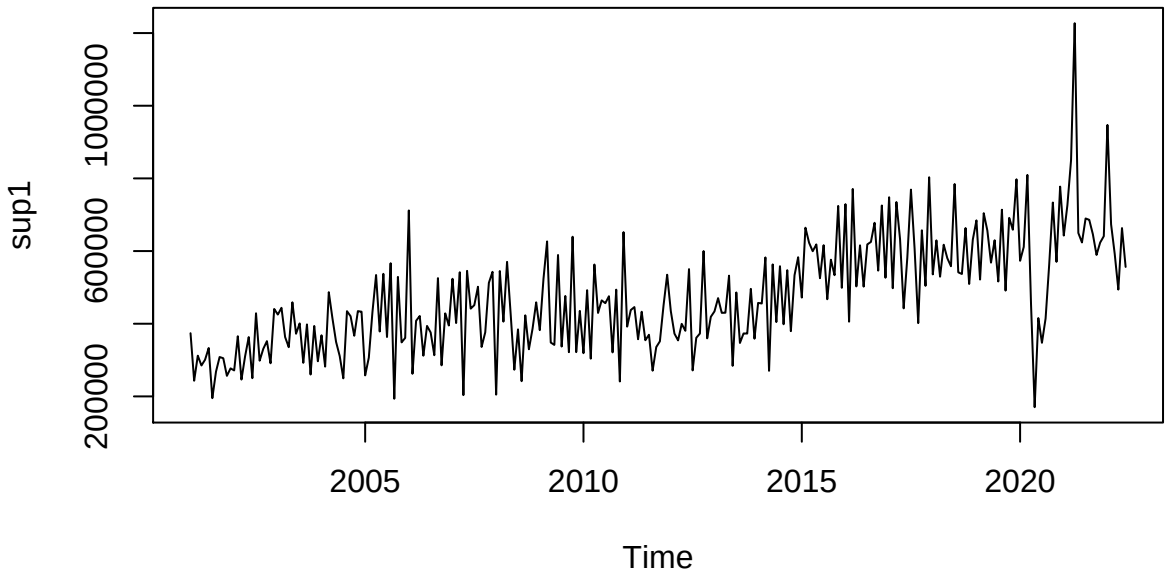

500000
1500000
2005
2010
2015
2020
Time


3e+05
4e+05
6e+05
2005
2010
2015
2020
Time

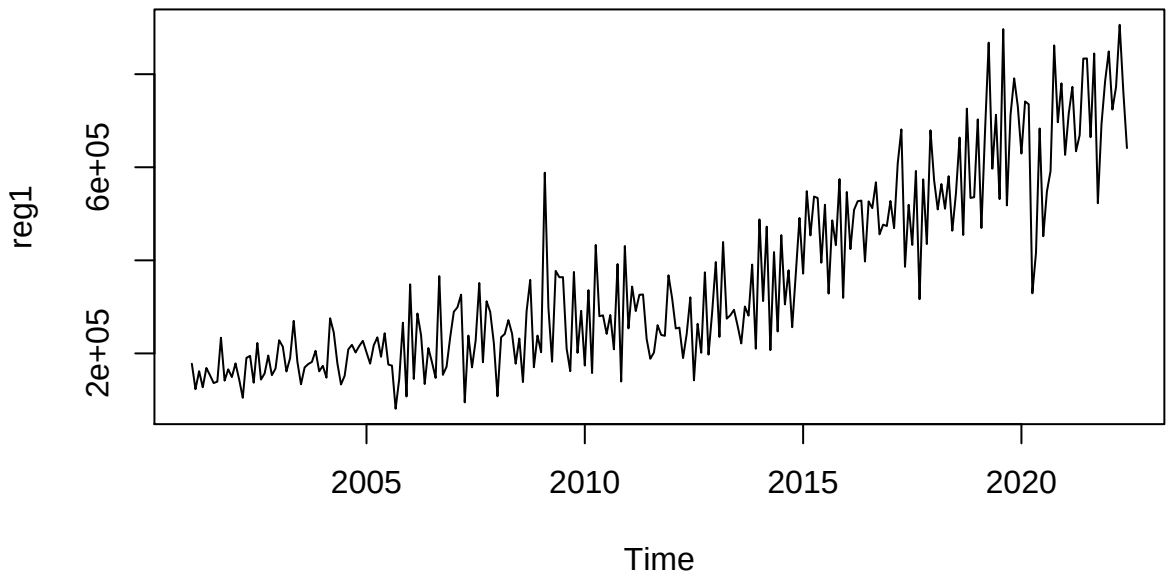

2e+05
4e+05
7e+05
2005
2010
2015
2020
Time

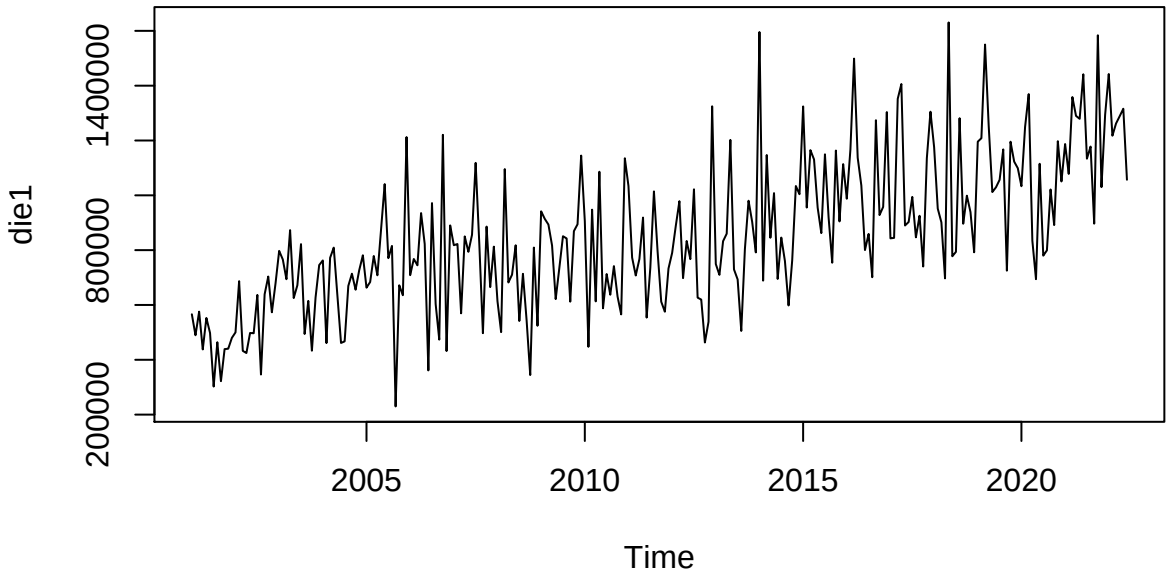


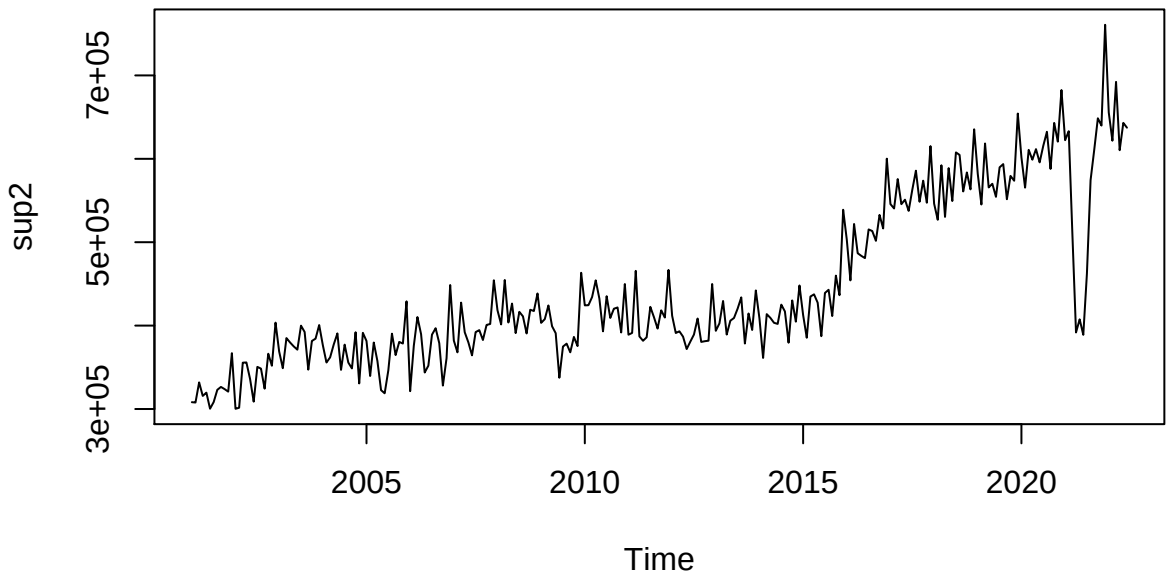
Los modelos se ven convincentes, pues la varianza y la media parecen coincidir con la tendencia observada en los pasados 20 años. Sin embargo, al contrastar con los datos reales, vemos que la predicción para los años 2020-2022 no fue correcta. Los gráficos reales de la importación y consumo de gasolina de 2001 a 2022 son los sigu-

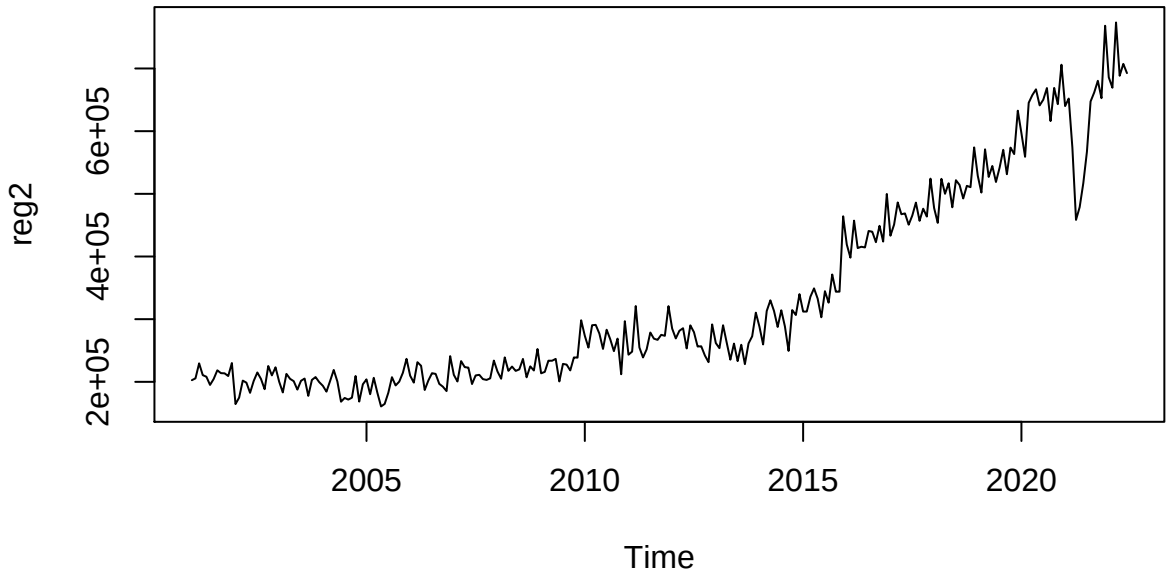


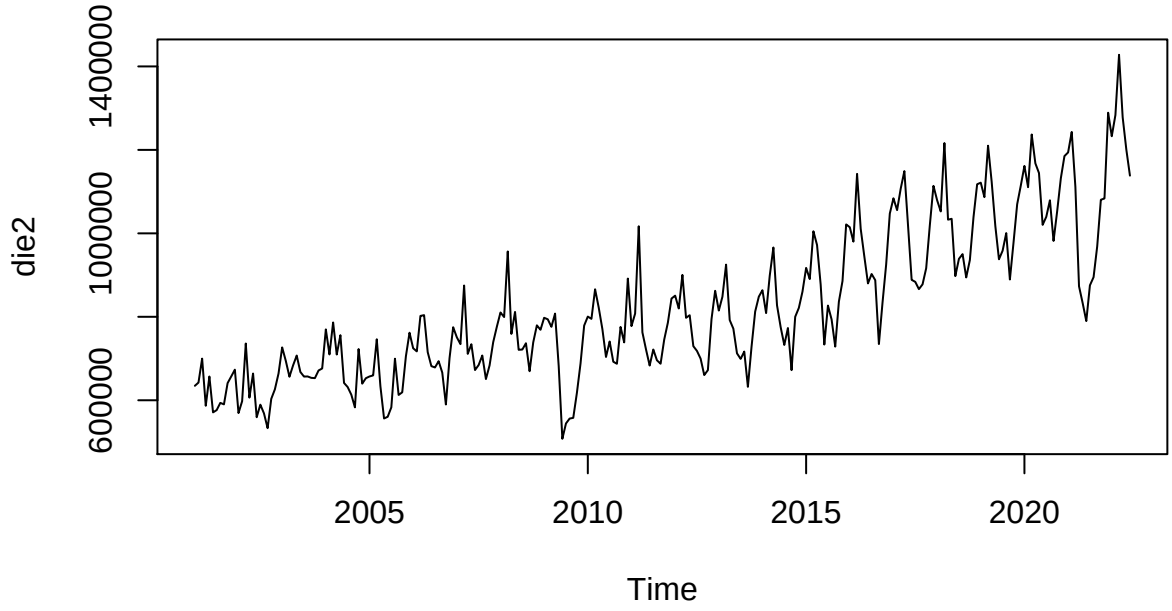
ienes:


reg1
2e+05
6e+05
2005
2010
2015
2020
Time


die1
200000
800000
1400000
2005
2010
2015
2020
Time


sup2
3e+05
5e+05
7e+05
2005
2010
2015
2020
Time


reg2
2e+05
4e+05
6e+05
2005
2010
2015
2020
Time



Los datos atípicos observados en los años 2020 en adelante son muy probablemente causados por la pandemia del Covid-19, donde el tránsito humano se vio severamente impactado y reducido. Debido a los protocolos de distanciamiento y de virtualidad, tiene sentido que la demanda por combustibles se haya reducido a inicios de 2020. El consumo e importación de combustibles parecen recuperar su curso original una vez los efectos de la pandemia fueron aminorados.

Evaluación

Claramente el modelo ARIMA no puede predecir un fenómeno externo como la pandemia del Covid-19, pero este puede darnos una idea de cual era la tendencia y que valores hubieramos esperado ver si el ambiente nacional actual hubiera sido un tanto más homogéneo a las primeras dos décadas del siglo.

Finalmente, para comparar el desempeño de estos algoritmos, empleamos los criterios de AIC y BIC.

```
AIC(fitSup1, fitReg1, fitDie1, fitSup2, fitReg2, fitDie2)
```

```
##         df       AIC
## fitSup1  6  132.8035
## fitReg1  7  204.3059
## fitDie1  7  165.4965
## fitSup2  6 -577.2146
## fitReg2  7 -442.0281
## fitDie2  8 -458.1944
```

```
BIC(fitSup1, fitReg1, fitDie1, fitSup2, fitReg2, fitDie2)
```

```
##         df       BIC
## fitSup1  6  153.0274
## fitReg1  7  227.9004
## fitDie1  7  189.0909
## fitSup2  6 -556.9908
```

```
## fitReg2  7 -418.4337
## fitDie2  8 -431.2293
```

Como esperabamos, los modelos que lidian con los volumenes de consumo, los tres últimos, poseen índices más satisfactorios de acuerdo a los criterios de información. Esto lo atribuimos a la no estacionariedad de la varianza de las primeras tres series de tiempo, la cual no pudo cancelarse por completo mediante transformaciones en las primeras etapas del laboratorio. Sin embargo, los índices obtenidos son relativamente satisfactorios, y confiamos en la capacidad de los modelos para predecir precios de gasolina en condiciones no extraordinarias.