PREDICCION DEL PRECIO SPOT CON MODELOS ARIMA

Nicolas Vega Munoz

2024-04-10

El análisis de series temporales es una técnica fundamental en el campo de la estadística y la ciencia de datos, especialmente en situaciones donde los datos están correlacionados en el tiempo. En este documento, exploraremos el modelado de series temporales utilizando el método ARIMA (Autoregressive Integrated Moving Average).

El objetivo principal de este análisis es desarrollar un modelo predictivo robusto que pueda capturar y predecir patrones en los datos de una serie temporal específica. Utilizaremos datos históricos de precios SPOT, los cuales están sujetos a fluctuaciones estacionales y estocásticas, lo que los hace adecuados para el modelado con ARIMA.

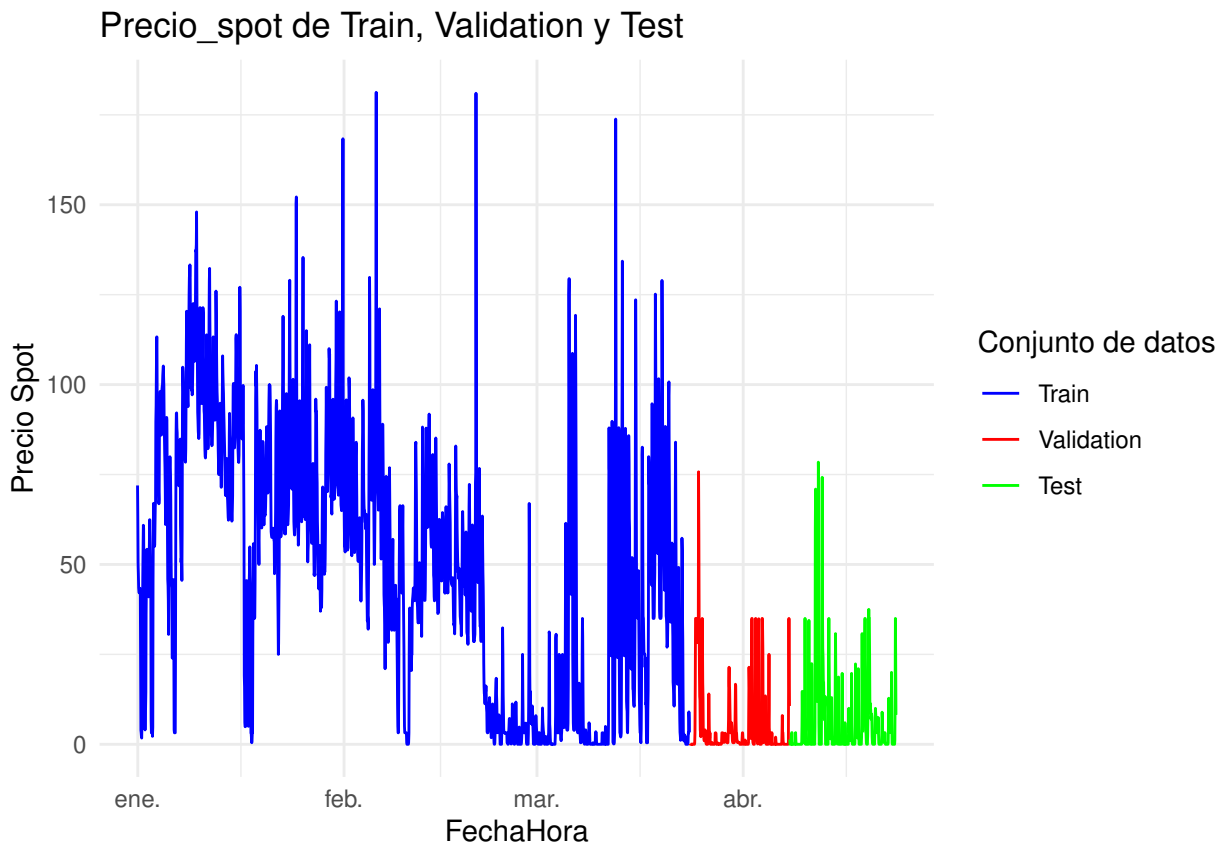
A lo largo de este informe, abordaremos los siguientes pasos:

1. **Preparación de datos**: Cargaremos y preprocesaremos los datos de la serie temporal, asegurándonos de que estén en un formato adecuado para el modelado.
2. **Diagnóstico de estacionalidad y tendencia**: Analizaremos la serie temporal para identificar componentes de tendencia y estacionalidad, lo que nos permitirá seleccionar los parámetros apropiados para el modelo ARIMA.
3. **Ajuste de modelos ARIMA**: Ajustaremos varios modelos ARIMA a los datos, utilizando diferentes configuraciones de parámetros para encontrar el modelo que mejor se ajuste a la serie temporal.
4. **Evaluación del modelo**: Evaluaremos el desempeño de los modelos ARIMA utilizando métricas de evaluación adecuadas, como el error cuadrático medio (MSE) y el error absoluto medio (MAE).
5. **Selección del mejor modelo**: Basándonos en las métricas de evaluación y en el diagnóstico de los residuos, seleccionaremos el modelo ARIMA más adecuado para realizar predicciones futuras.

A lo largo de este proceso, nos centraremos en aplicar los conceptos teóricos detrás del modelo ARIMA, así como en utilizar herramientas prácticas de programación en R para implementar y evaluar los modelos. Este análisis nos proporcionará información valiosa sobre la dinámica subyacente de la serie temporal estudiada y nos permitirá realizar predicciones precisas.

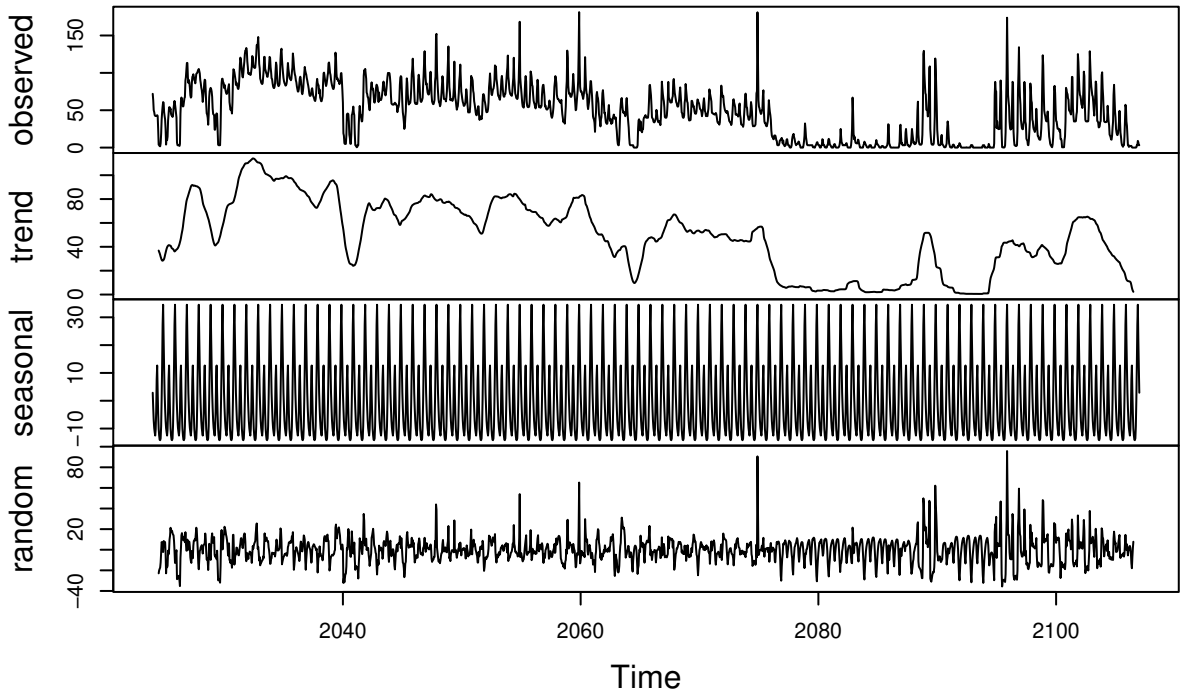
En primer lugar leemos nuestro dataset y configuramos el numero de días que usaremos para evaluar el modelo. Puesto que disponemos de un histórico grande con datos muy alejados de los valores actuales decidimos entrenar el modelo con datos a partir de este año 2024.

Procedemos a dividir el dataset en train, validation y test. Usamos 15 dias para validacion y 15 para test



Como describimos en el apartado de Fundamentos Teóricos el modelo ARIMA tiene como supuesto que nuestra serie temporal es estacionaria en media en varianza, i.e, que la media y la varianza son estables en el tiempo.

Por ello visualizaremos la serie temporal, descomponiéndola y realizaremos los contrastes de hipótesis oportunos para determinar si nos encontramos ante una serie temporal estacionaria y si se presenta una componente estacional.



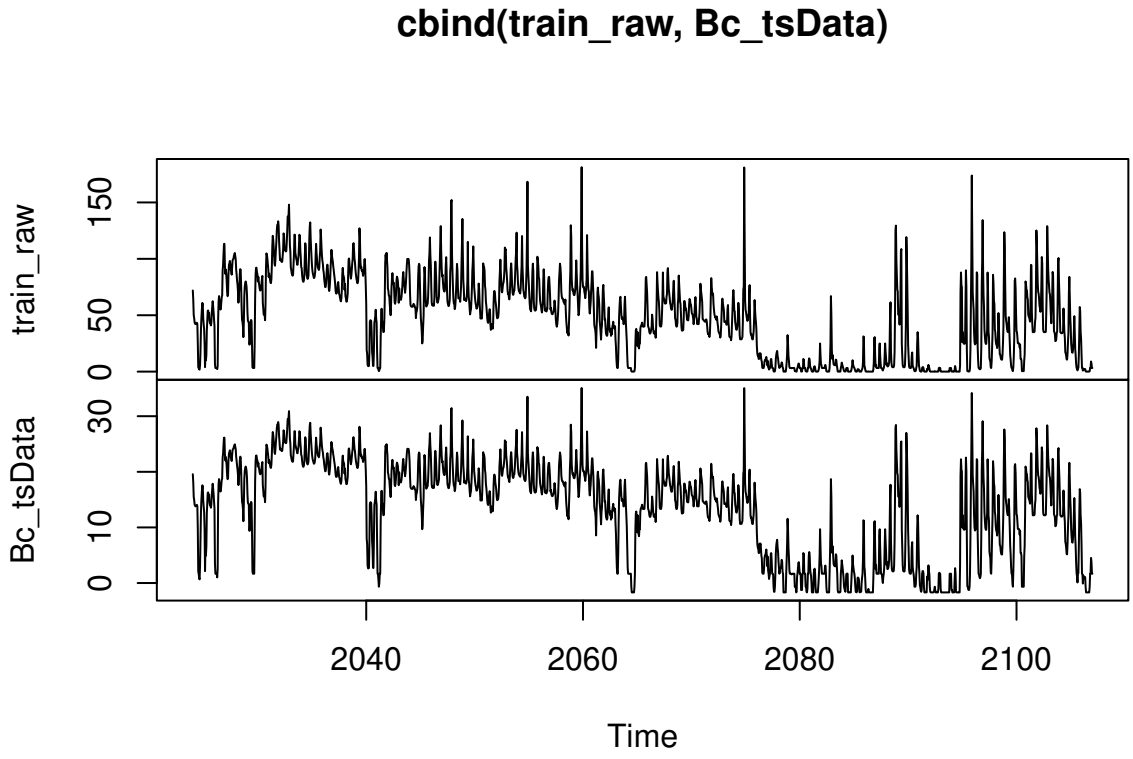
Al descomponer la serie temporal observamos que esta no es ni estacionaria en media ni en varianza. Por ello deberemos realizar las transformaciones oportunas para disponer de una serie temporal estacionaria adecuada para el modelo ARIMA.

Parece que si existe una componente estacional, pero es difícil de determinar a partir de este gráfico por lo agrupados que están los datos. Evaluaremos la existencia de componente estacional posteriormente mediante las ACF y PACF.

EL modelo ARIMA tiene como supuesto tratar con una serie estacionaria en media y en varianza. Puesto que nuestra serie temporal no lo es, debemos realizar las transformaciones necesarias.

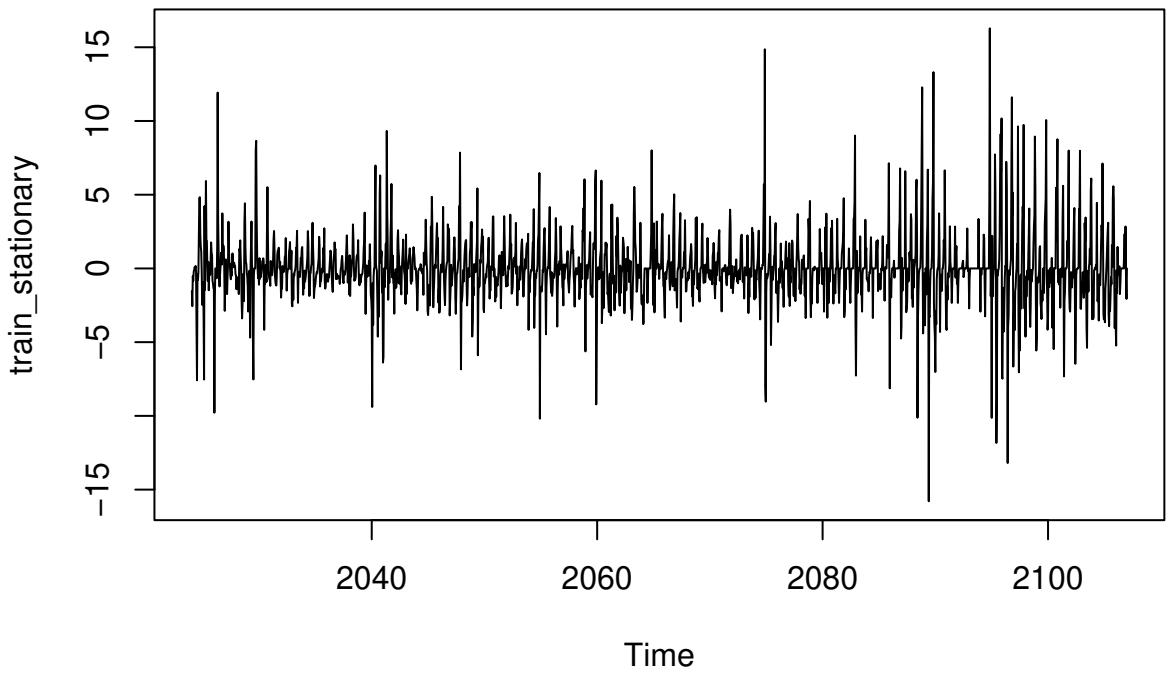
Para hacer la serie temporal estacionaria en varianza realizaremos una transformación de Box-Cox. En primer lugar calculamos el valor de lambda.

```
## [1] 0.5925395
```



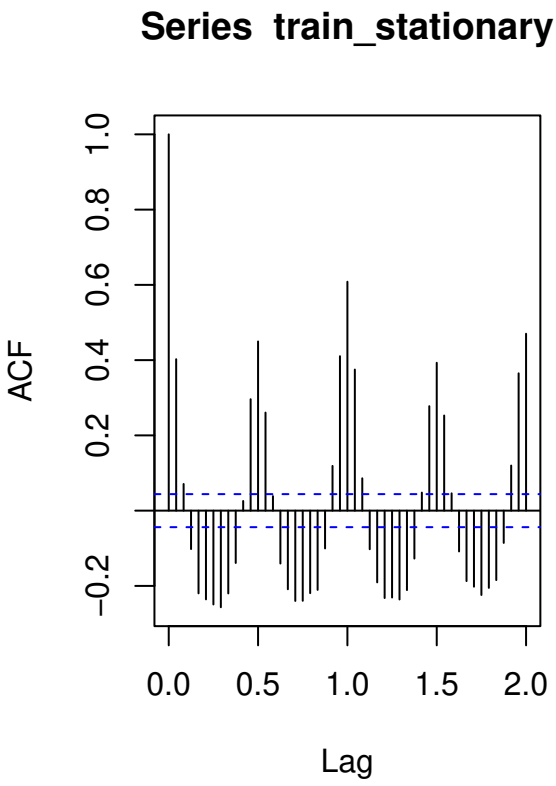
Observamos como ahora la serie es estacionaria en varianza. Ahora deberemos hacerla estacionaria en media, por lo que diferenciamos las veces necesarias la serie. En nuestro caso tan solo sera necesario diferenciar una vez. El parámetro 'd' de nuestro modelo SARIMA sera igual a 1.

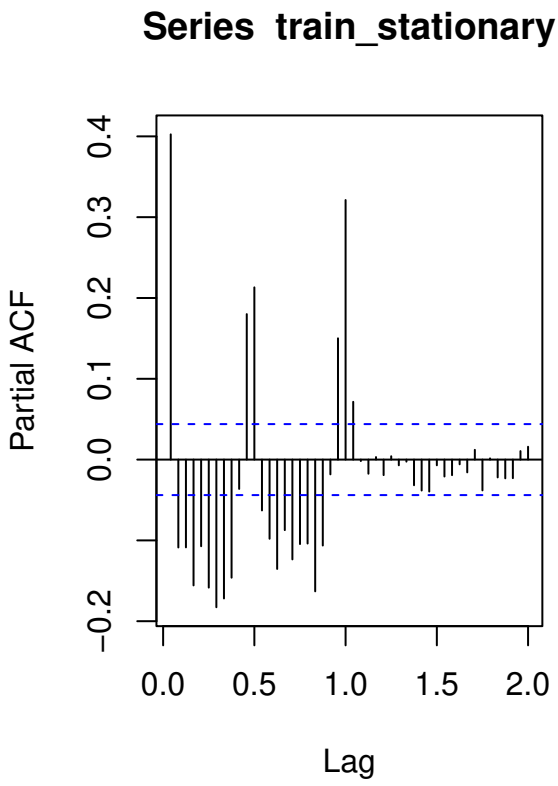
```
## [1] 1
```



Como podemos observar la serie ahora es estacionaria en media y varianza.

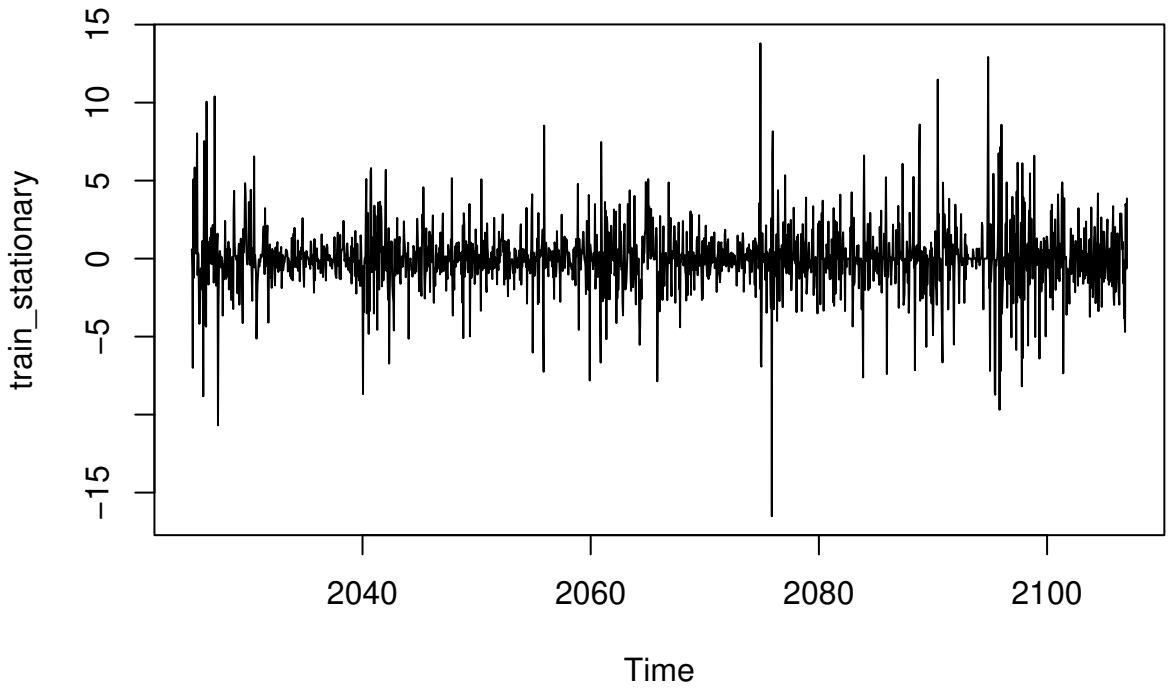
Procedemos ahora a estudiar la estacionalidad.





Como podemos ver en la PACF la serie temporal tiene el valor mas significativo en el retardo (lag) 24 (valor de la frecuencia), por lo tanto la serie tiene componente estacional diaria.

Realizamos una diferenciación estacional. El parámetro 'D' de ARIMA sera igual a 1.

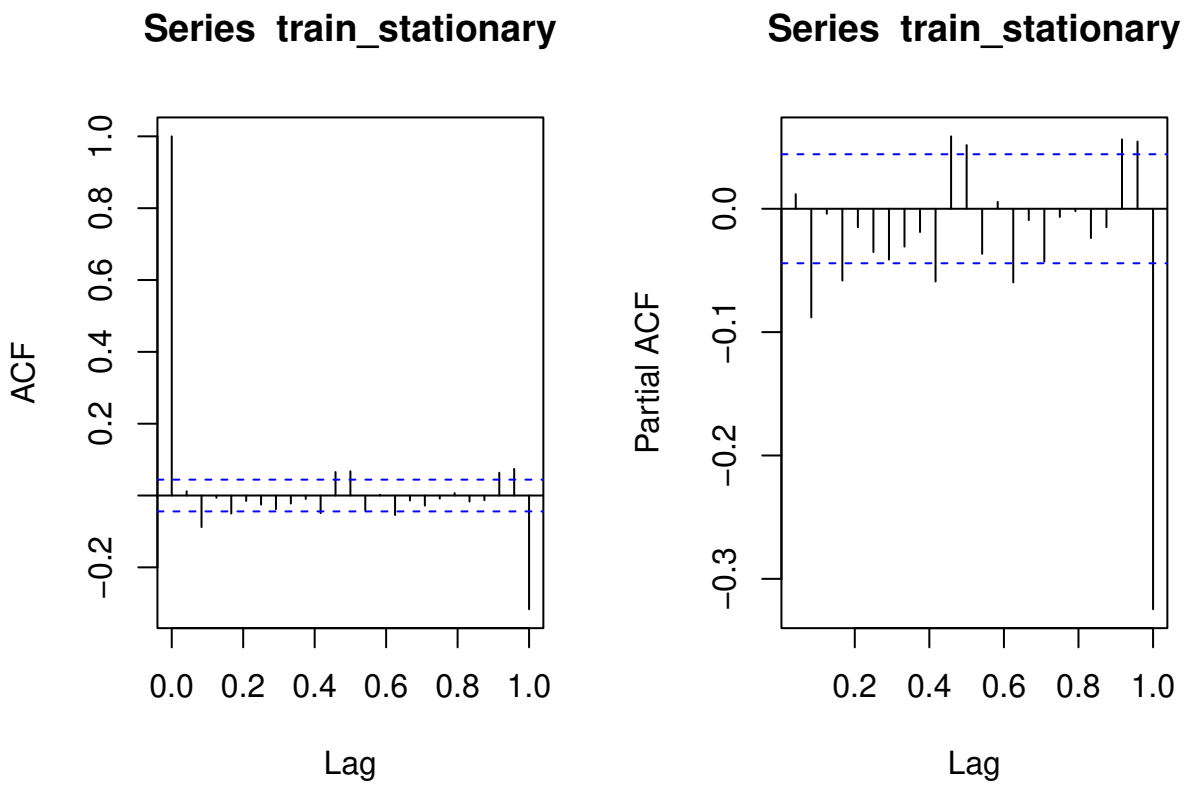


En principio ya hemos transformado nuestra serie a una estacionaria en media y varianza y sin componente estacional. Aun así, realizamos el test de Dickie-Fuller aumentado para estar seguros de que no debemos realizar mas transformaciones. La hipótesis nula es que nuestra serie no es estacionaria.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  train_stationary
## Dickey-Fuller = -13.489, Lag order = 12, p-value = 0.01
## alternative hypothesis: stationary
```

Como el p-valor es menor que 0.05 podemos rechazar la hipótesis nula de que la serie no es estacionaria, por tanto podemos concluir que nuestra serie es estacionaria.

Una vez tenemos nuestra serie temporal estacionaria podemos comenzar con la modelización. En primer lugar graficaremos las ACF y PACF para hallar los parámetros del modelo ARIMA(p,d,q)x(P,D,Q)s.



Observamos en estas gráficas como no hay ningún termino significativo (ni autorregresivo ni de medias móviles) consecutivo, por lo que los valores p y q serán igual a 0. Los parámetros 'd' y 'D' tendrán valor 1 por las diferenciaciones realizadas a la serie temporal y 's' sera 24 por su componente estacional.

Tras haber realizado este estudio estadístico podemos concluir que nuestro modelo SARIMA sera de la forma (0,1,0)x(0,1,0)24

Una vez disponemos de los parámetros de nuestro modelo SARIMA entrenamos el modelo con nuestros datos de train.

```
## Series: train_stationary
## Regression with ARIMA(0,1,0)(0,1,0)[24] errors
##
## Coefficients:
##       demanda      EUA  precio_gas  prod_eolica  prod_solar  demanda_residual
##         3e-04  -0.1127      0.3175       -3e-04      -3e-04             0e+00
## s.e.    1e-04   0.1355      0.1871        1e-04       1e-04             1e-04
##       rampa
##       1e-04
## s.e.  1e-04
##
## sigma^2 = 23:  log likelihood = -5799.51
## AIC=11615.02   AICc=11615.09   BIC=11659.6
##
## Training set error measures:
##                       ME     RMSE      MAE MPE MAPE     MASE       ACF1
## Training set 0.006341458 4.756403 3.097877 NaN  Inf 1.418735 -0.4817001
```

A priori no parece que tengamos un mal modelo, en general las métricas tienen valores bajos y un AIC=11631.238578.26

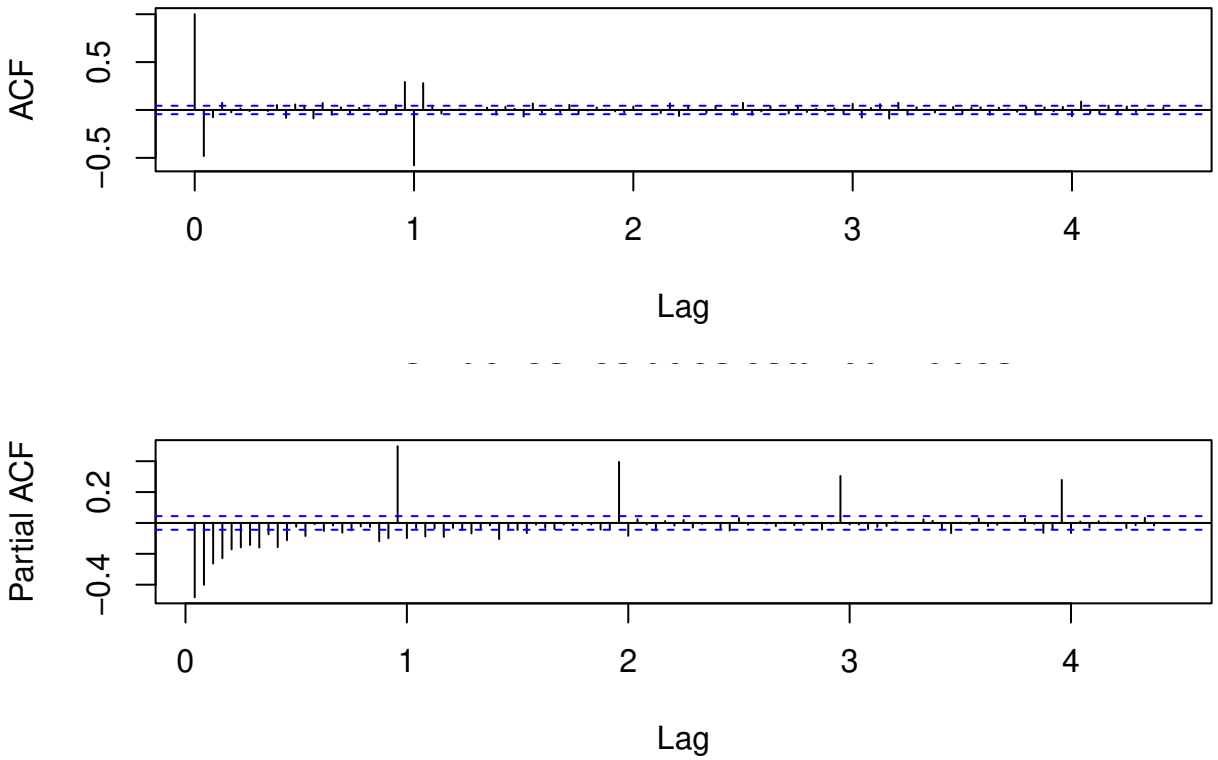
Una vez entrenado el modelo SARIMA debemos diagnosticar los residuos verificando que cumple los supuestos ARIMA:

-Autocorrelación: Los residuos no deben mostrar autocorrelación significativa, es decir, no deben exhibir patrones discernibles en sus autocorrelogramas

-Normalidad: Los residuos del modelo deben seguir una distribución normal.

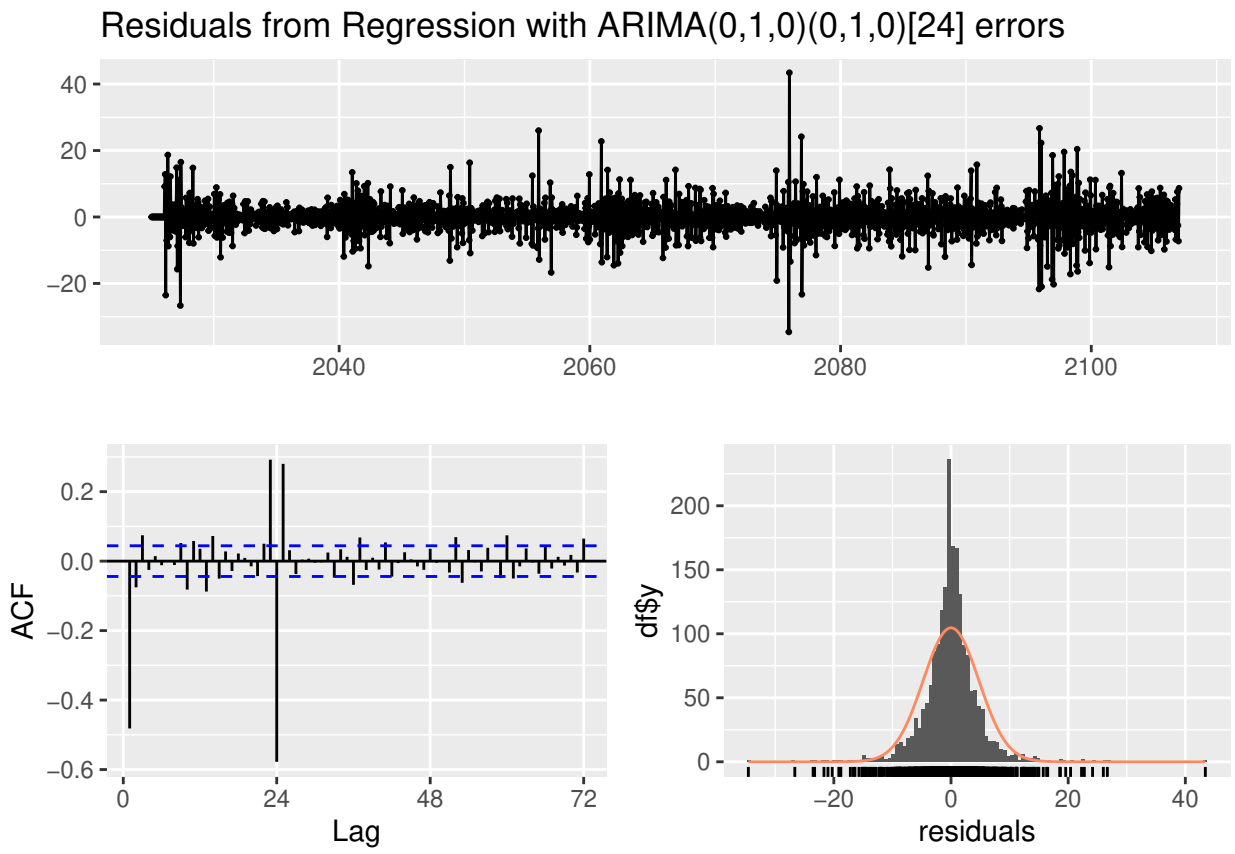
-Estacionariedad: Los residuos deben ser estacionarios, lo que significa que su media y varianza deben ser constantes a lo largo del tiempo

Verificamos la autocorrelación.



La ACF de los residuos no parece mostrar estructura y tiene casi todos los valores dentro de las bandas de confianza. Aun así en la PACF los retardos multiplos de 24 se salen de las bandas de confianza. Todo apunta a que los residuos no son aleatorios. Aun así continuamos con la diagnosis.

Realizamos el test de Ljung-Box.

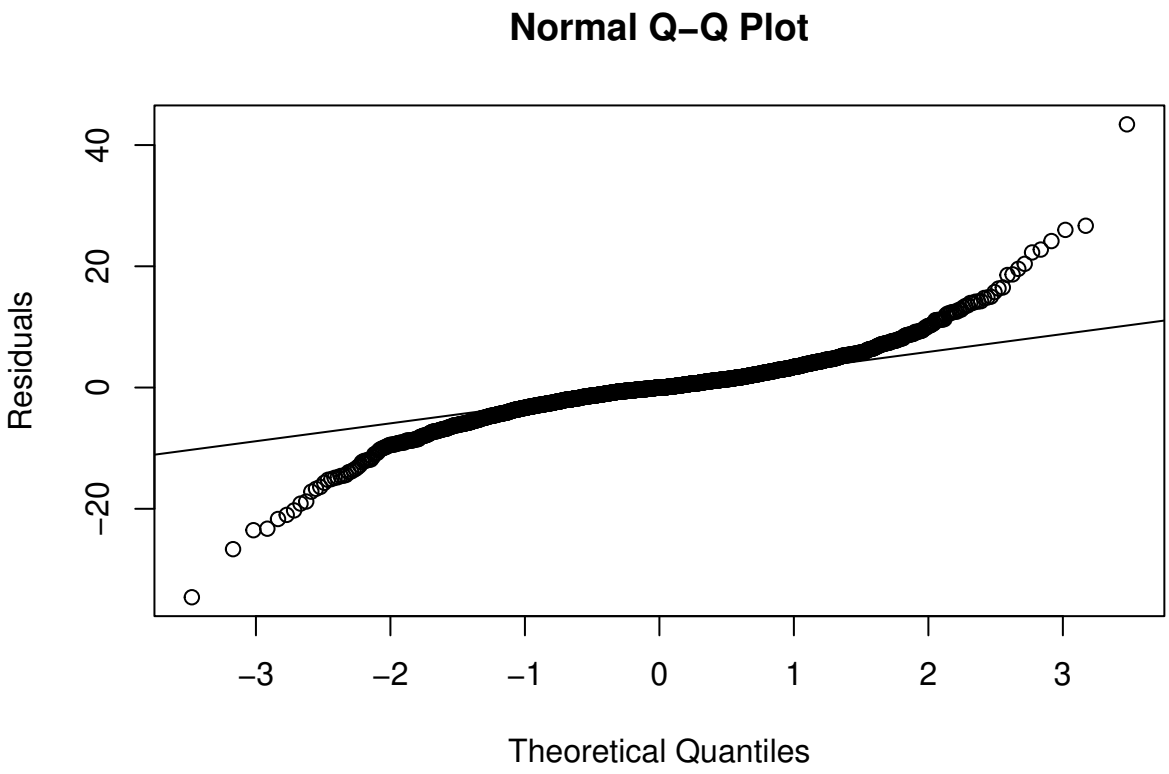



```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,1,0)(0,1,0)[24] errors
## Q* = 1595.6, df = 48, p-value < 2.2e-16
##
## Model df: 0.   Total lags used: 48
```

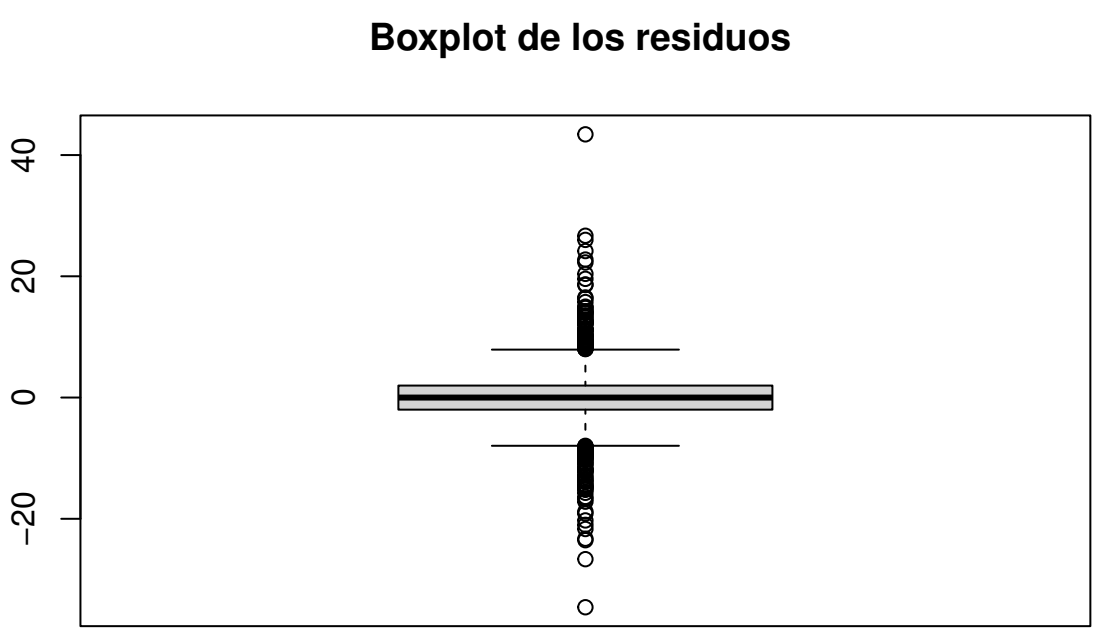
El p-valor del test de Ljung-Box es menor que 0.05 luego se puede rechazar que las primeras autocorrelaciones sean nulas, y no se puede asumir que los residuos sean ruido blanco.

Ahora si, en la ACF, podemos ver una notable correlación con lag igual a 24. Los residuos sí parecen ajustarse a una distribución normal. En la gráfica de los residuos podemos observar que si parecen ser estacionarios.

Visualizamos el QQ Plot de los residuos.



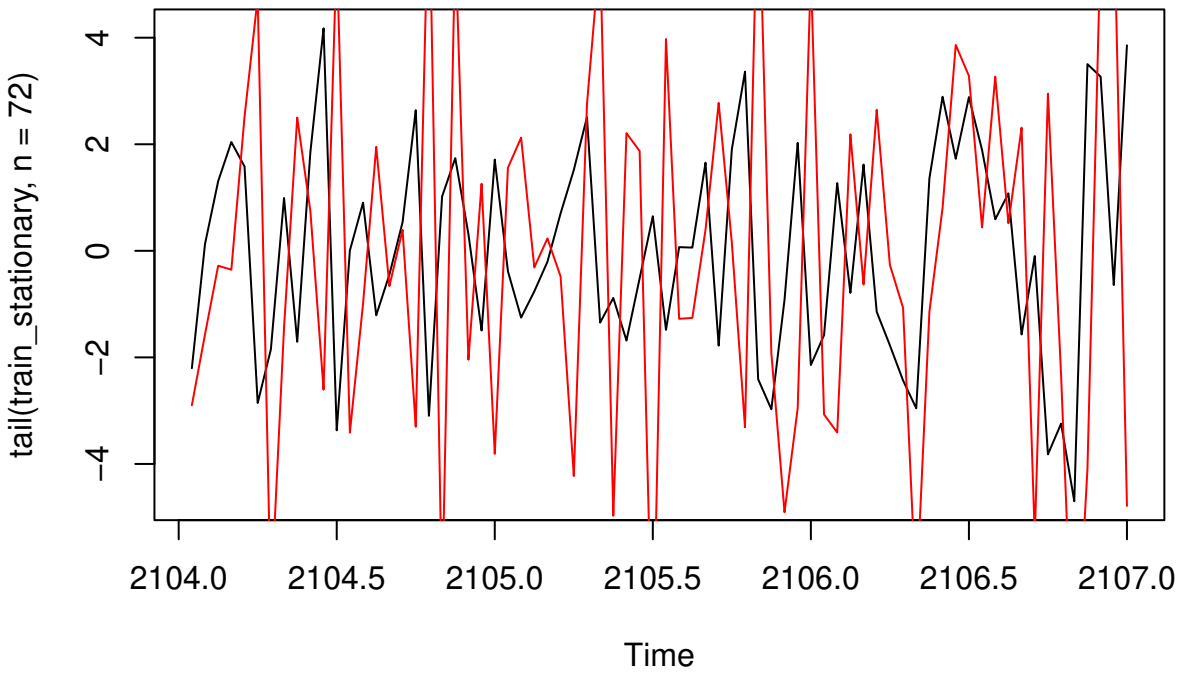
Podemos ver que los residuos aproximadamente siguen una distribución normal, a pesar de la gran cantidad de valores atípicos.



En este box plot observamos mas fácilmente la cantidad de valores atípicos presentes. Aun así la mayoría de residuos están entorno al 0.

Tras analizar los residuos podemos concluir que el modelo no supera la diagnosis. Aun así, si empíricamente el modelo tiene buena capacidad predictiva no lo descartaremos por razones prácticas.

Veamos gráficamente la diferencia entre la serie original y el modelo ajustado (en rojo)



Como se puede apreciar el modelo no se ajusta correctamente a nuestra serie.

Realizamos predicciones con los datos de validación para evaluar el modelo:

```
##                         ME      RMSE       MAE  MPE MAPE      MASE       ACF1
## Training set   0.006341458  4.756403  3.097877  NaN  Inf  1.668222 -0.4817001
## Test set     -35.381709379 40.217503 35.406744 -Inf  Inf 19.066706         NA
```

Como podemos ver nuestro modelo tiene unos errores altos y muy superiores a los obtenidos en train.Por ello haremos uso de la función AutoArima que hara una busqueda de parametros con el objetivo de obtener el mejor modelo ARIMA desde un enfoque empírico, en vez de teórico.

El mejor modelo hallado es con los parámetros ARIMA(2,0,0)(2,1,0)[24].

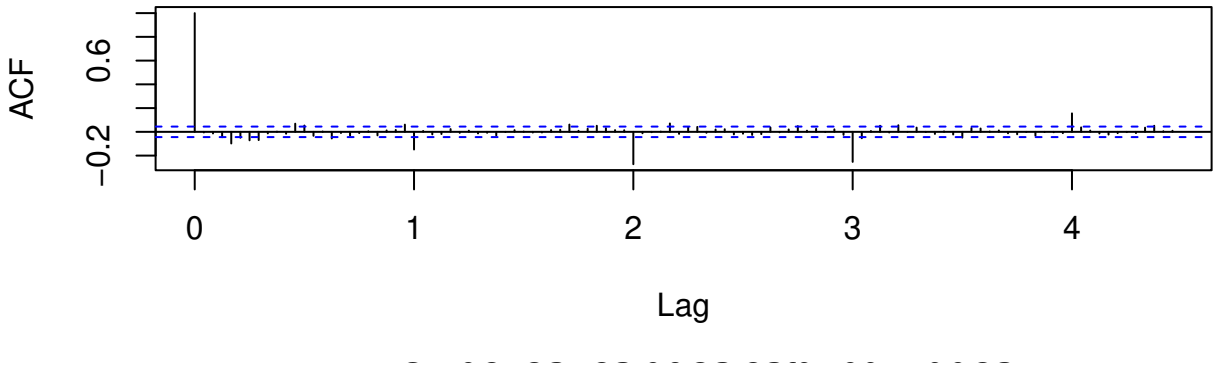
Análogo al entrenamiento del modelo teórico.

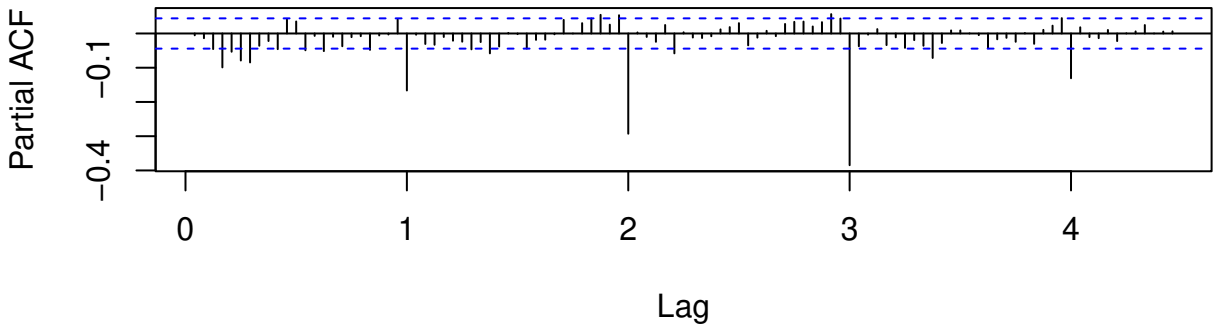
```
## Series: train_stationary
## Regression with ARIMA(2,0,0)(2,1,0)[24] errors
##
## Coefficients:
##           ar1      ar2     sar1     sar2  demanda     EUA  precio_gas
##       -0.0553  -0.1495  -0.8499  -0.4626    5e-04  0.0589      0.6635
## s.e.   0.0226   0.0225   0.0206   0.0209    1e-04  0.1207      0.1728
##       prod_eolica  prod_solar  demanda_residual  rampa
##            -5e-04      -6e-04             2e-04  1e-04
## s.e.        1e-04       1e-04             1e-04  1e-04
```

```
##
## sigma^2 = 5.394:  log likelihood = -4401.68
## AIC=8827.36   AICc=8827.52   BIC=8894.23
##
## Training set error measures:
##                       ME     RMSE      MAE MPE MAPE      MASE         ACF1
## Training set -0.006693363 2.301674 1.506854 NaN  Inf 0.6900941 -0.005332612
```

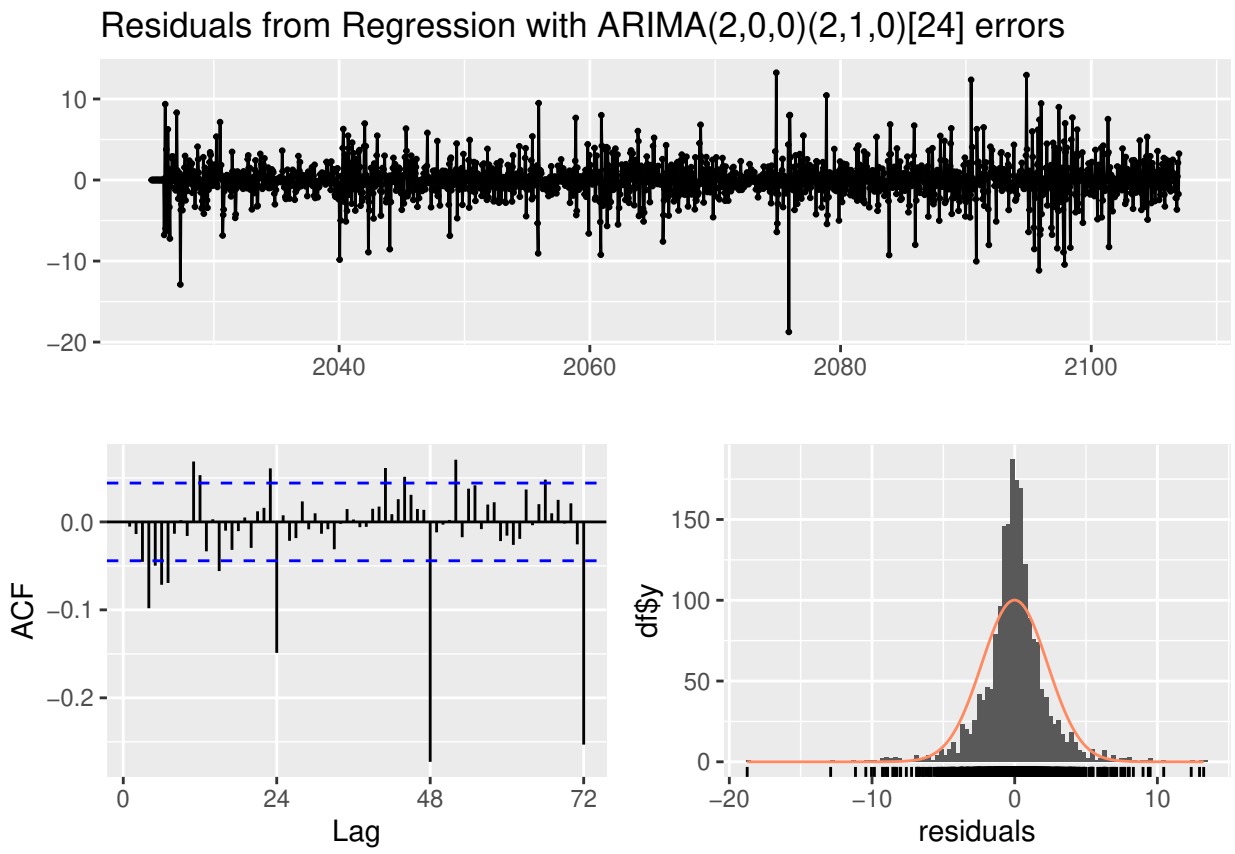
En primera instancia se observa una ligera mejora en entrenamiento respecto al modelo anterior. EL AIC ha disminuido notablemente.

Realizamos la diagnosis del modelo de forma análoga al anterior.



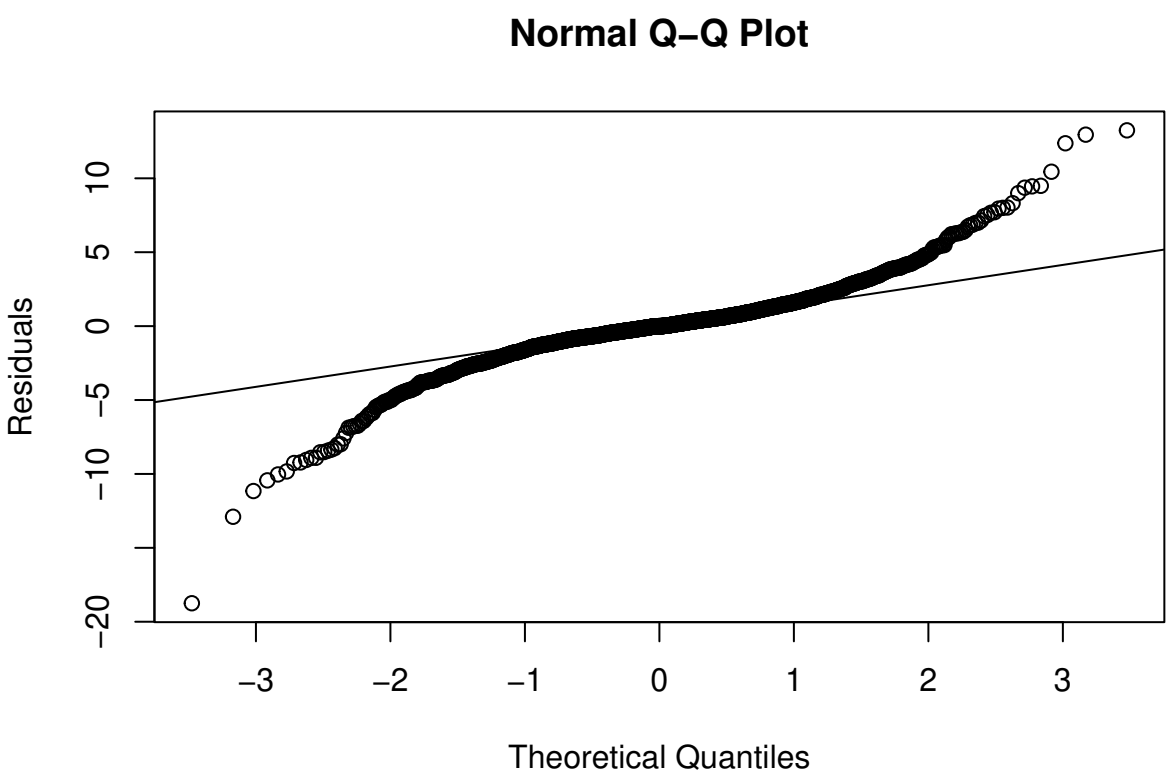


En esta ocasión si que hay una notable correlación en los retardos múltiplos de 24. Lo vemos mas claramente a continuacion:

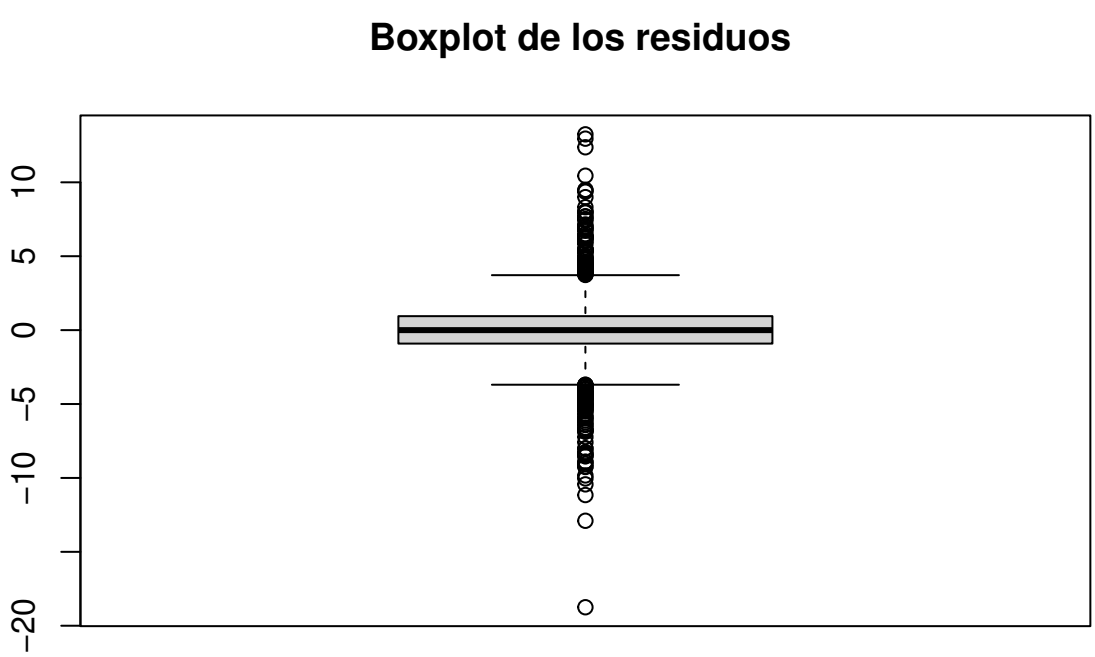



```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(2,0,0)(2,1,0)[24] errors
## Q* = 303.34, df = 44, p-value < 2.2e-16
##
## Model df: 4.   Total lags used: 48
```

El p-valor de test de Ljung-Box es menor que 0.05 luego se puede rechazar que las primeras autocorrelaciones sean nulas, y no se puede asumir que los residuos sean ruido blanco. En la ACF podemos ver una notable correlación con lag igual a 24 y múltiplos de él. Los residuos sí parecen ajustarse a una distribución normal. En la gráfica de los residuos podemos observar que si parecen ser estacionarios.



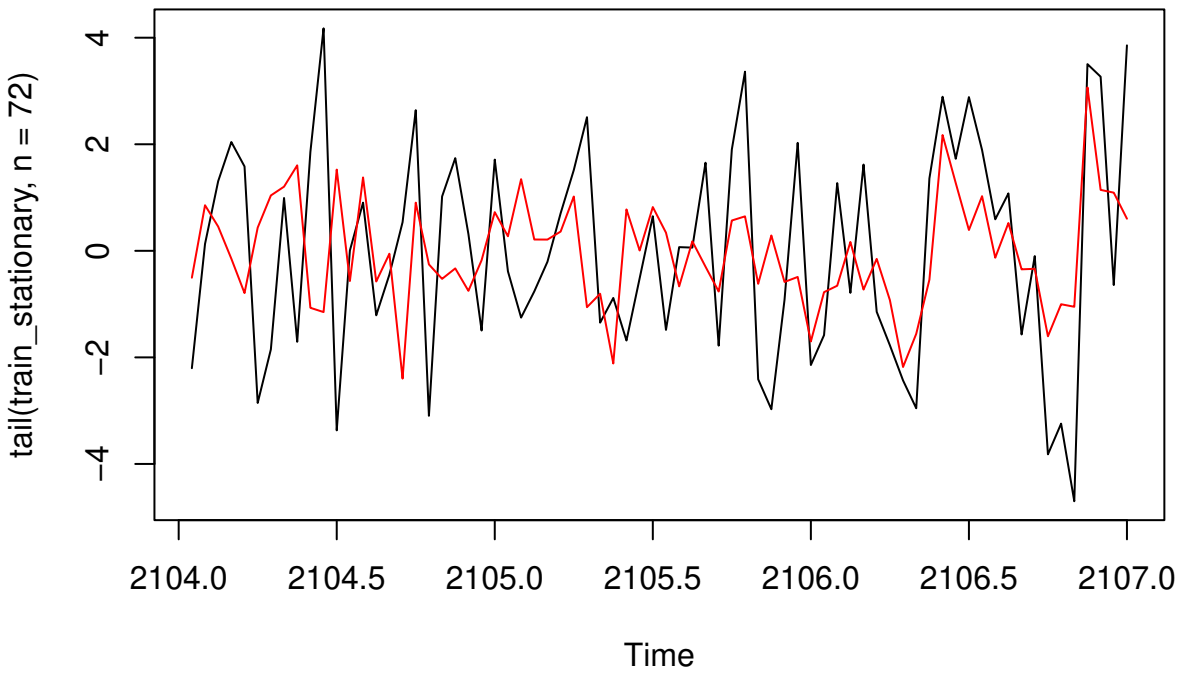
Podemos ver que los residuos aproximadamente siguen una distribución normal, a pesar de la gran cantidad de valores atípicos.



En este box plot observamos mas fácilmente la cantidad de valores atípicos presentes. Aun así la mayoría de residuos están entorno al 0.

Tras analizar los residuos podemos concluir que este modelo ARIMA tampoco supera la diagnosis. Aun así, al igual que antes, si el modelo tiene buena capacidad predictiva lo consideraremos valido.

Veamos gráficamente la diferencia entre la serie original y el modelo ajustado (en rojo)



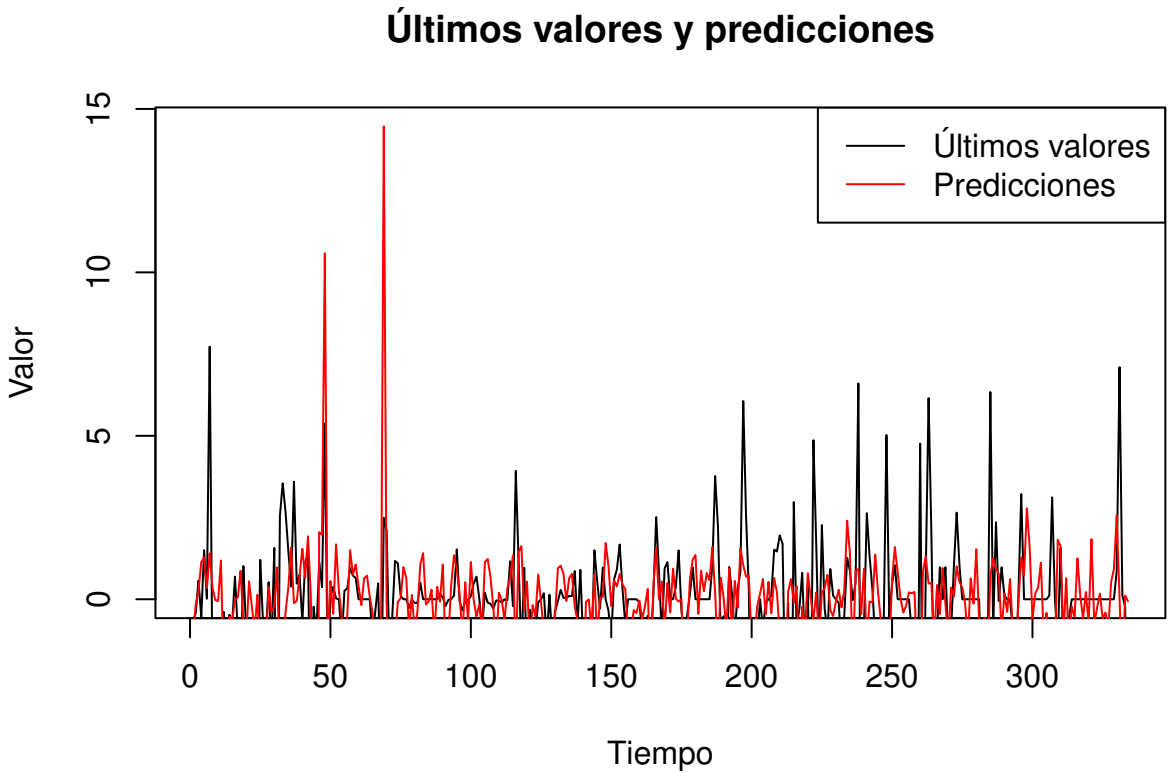
El modelo da unos resultados bastante mejores que los del anterior modelo. Aun así no se acerca a lo esperado en este trabajo.

Realizamos predicciones con los datos de validación para evaluar el modelo:

```
##                       ME     RMSE      MAE MPE MAPE      MASE         ACF1
## Training set -0.006693363 2.301674 1.506854 NaN  Inf 0.8114482 -0.005332612
## Test set      0.014231453 2.317620 1.474850 NaN  Inf 0.7942138           NA
```

Observamos como efectivamente obtenemos unos resultados mejores que con el anterior modelo. Las métricas de error tanto en train como en validación son muy buenas y prácticamente no hay diferencia entre ellas, lo cual es algo muy positivo.

Procedemos a visualizar las predicciones vs valores reales en validación.



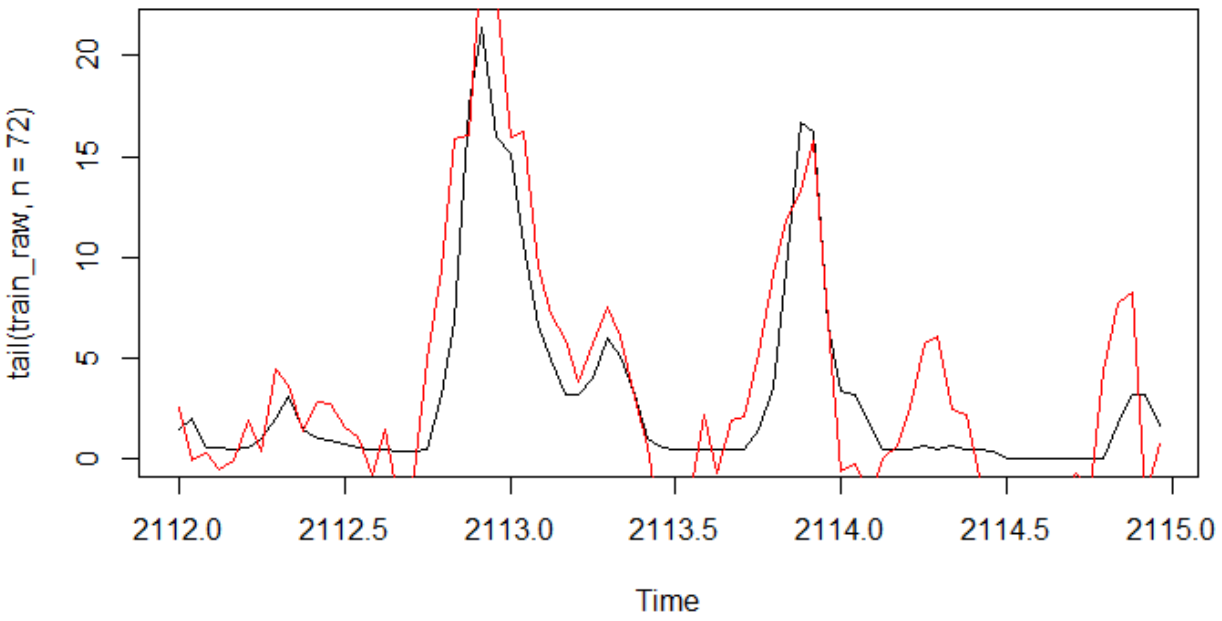
Como se puede observar, pese a que las métricas en si son buenas, las predicciones no lo son. Parece que un modelo SARIMAX no es la mejor aproximación a nuestro problema. Aun así, a modo de ultimo intento, trataremos de entrenar un modelo con la serie original, sin sufrir transformación alguna. Hacemos uso de autoarima. Esta aproximacion obvia toda la base teorica de los modelos ARIMA.

El mejor modelo hallado es ARIMA(1,1,4)(1,0,0)[24]

```
## Series: train_raw
## Regression with ARIMA(1,1,4)(1,0,0)[24] errors
##
## Coefficients:
##          ar1     ma1      ma2     ma3     ma4    sar1  demanda      EUA
##       0.8754  -1.074  -0.1192  0.1463  0.0562  0.2860   0.0031  -0.0794
## s.e.  0.0335   0.042   0.0337  0.0347  0.0305  0.0222   0.0003   0.4991
##       precio_gas  prod_eolica  prod_solar  demanda_residual   rampa
##           2.5647      -0.0032     -0.0022             6e-04  -5e-04
## s.e.      0.7388       0.0003      0.0003             3e-04   2e-04
##
## sigma^2 = 61.36:  log likelihood = -6921.85
## AIC=13871.71   AICc=13871.92   BIC=13950.06
##
## Training set error measures:
##                       ME     RMSE      MAE MPE MAPE      MASE        ACF1
## Training set -0.08447855 7.805974 4.812318 NaN  Inf 0.3239711 0.001803556
```

Las métricas han empeorado notablemente, el AIC ha aumentado a casi el doble respecto al anterior modelo.

Puesto que en esta implementacion estamos obviando toda la base teorica tampoco consideramos oportuno realizar la diagnosis del modelo, tan solo observaremos su capacidad predictiva.

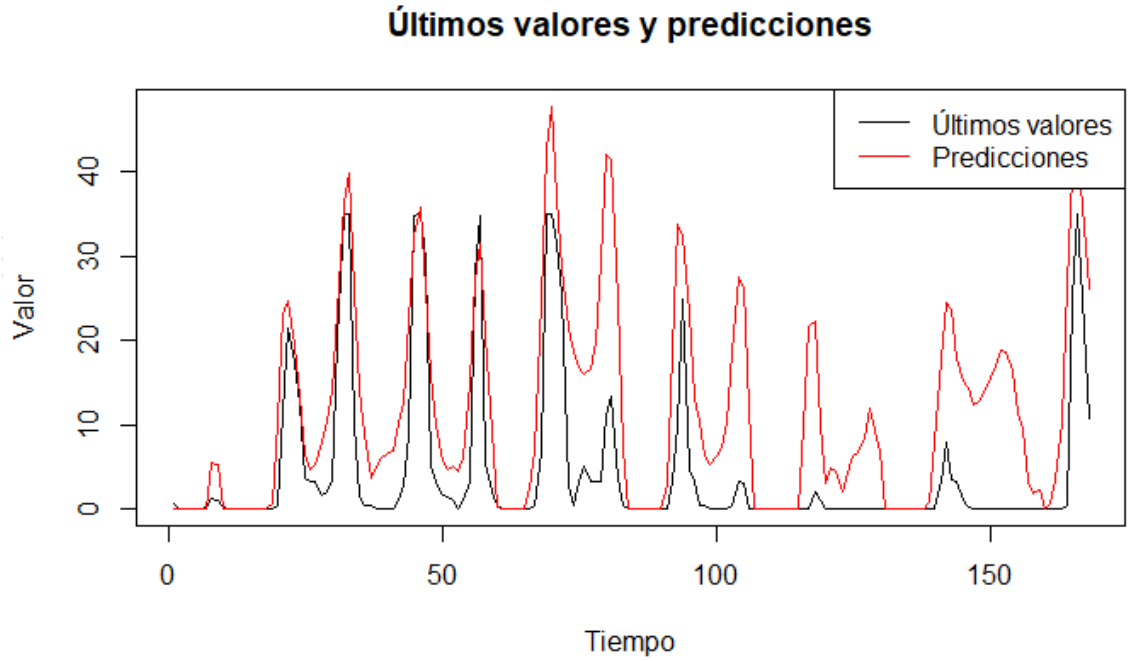


Sorprendentemente en esta ocasión si que parece ajustarse bien a nuestra serie temporal. Realizamos predicciones con el conjunto de validación y evaluamos los resultados.

```
##                       ME      RMSE       MAE  MPE MAPE      MASE        ACF1
## Training set  -0.08447855  7.805974  4.812318  NaN  Inf 0.7107139 0.001803556
## Test set     -12.16307290 15.539450 12.948720 -Inf  Inf 1.9123497          NA
```

Las métricas obtenidas han empeorado, sin embargo acabamos de ver que estas no reflejan bien la capacidad predictiva del modelo. Procedemos a visualizar las predicciones con el conjunto de validación.

Últimos valores y predicciones



Como podemos ver el modelo ajusta muy bien la forma pero no la magnitud. Puesto que este modelo es el que mejor se ajusta a nuestra serie lo seleccionaremos como modelo ganador.

Curiosamente sera este ultimo modelo ARIMA que obvia toda la base teórica nuestro modelo ganador. Sera este modelo y no cualquiera de los otros 2, pese a tener peores métricas, ya que visualmente observamos que es el que mejor modela nuestra serie temporal.

Los resultados obtenidos aun así no son satisfactorios, continuaremos estudiando modelos mas avanzados con el objetivo de realizar mejores predicciones que las obtenidas aquí.