Informe Assigment 2

ICAI. Machine Learning.

Álvaro Rodriguez y Pablo Sanz

Curso 2021-22. Última actualización: 2021-12-03

Índice

1. Presentación de los datos

Los datos a tratar pertenecen al número de desempleados de España a lo largo de los años. Como se puede ver en representación inferior, podemos observar diferentes épocas que ha sufrido el país en su historia más reciente.

Si empezamos por el extremo inferior que data sobre los inicios del año 2000, observamos que el empleo se mantiene más o menos constante en cuanto a tendencia, aunque se aprecia una gran estacionalidad según si estamos en los meses de invierno o verano. Este patrón se repite hasta el año 2008 donde con la gran crisis financiera mundial, el paro subió bruscamente situándose en más de 5.000.000 de personas en el año 2012. Pasado este momento, empieza a decrecer progresivamente observándose siempre aumentos y descensos según los meses de invierno o verano respectivamente hasta finales de 2019.

En este momento, a raíz de la aparción del COVID-19, la serie rompe todos los esquemas, aumentando en muy pocos meses casi 1.000.000 de desempleados debido a la crisis sanitaria y a todas las restricciones impuestas. Unos meses depués, estos altos niveles de desempleo empiezan a decrecer a gran ritmo hasta situarse en el momento actual donde el ritmo de bajada ha disminuido.

Veamos nuestra gráfica de la evolución del paro:

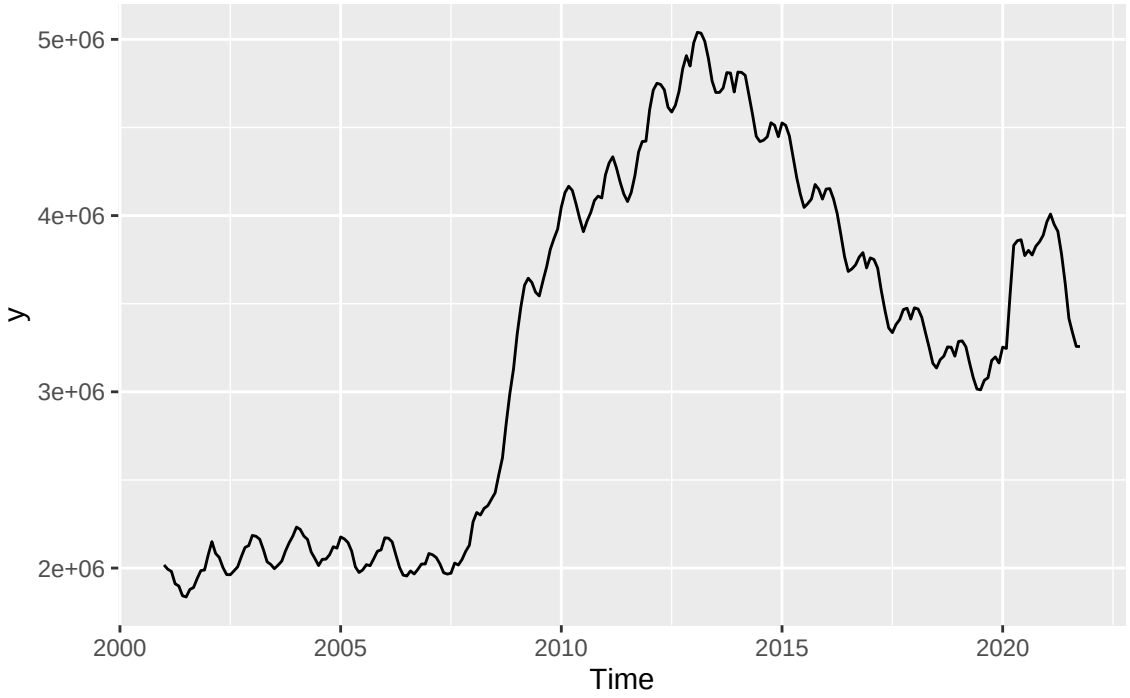

Figura 1: Evolución del paro

2. Predicción

En primer lugar explicaremos el proceso que nos llevó a dar como resultado que el paro en el mes de noviembre es de **3.240.623**.

2.1. Análisis exploratorio

Lo primero que hicimos tras ver los datos del paro fue ver si la serie es estacionaria en varianza, es decir, que independientemente del nivel la varianza es constante. Esto lo medimos a traves de un Box-Cox como se ve en el siguiente gráfico:

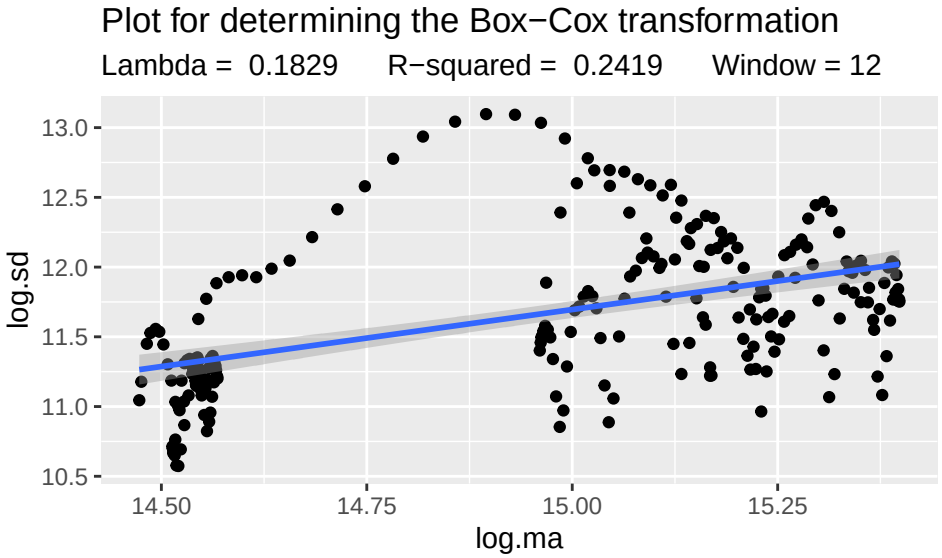

Figura 2: Box-Cox datos

Que refleja si es necesario realizar una transformación Box-Cox. Sin embargo, vemos que la línea azul no es muy creciente y que R-square no es muy alto (R-square=0.2419). Por lo tanto, no es necesario realizar una estabilización de varianza.

2.2. Seasonal ARIMA

En primer lugar representamos la serie temporal junto con su ACF y PACF para inspeccionar la serie regular y estacional. Vemos que la serie necesita una diferenciación en la parte regular ya que el ACF disminuye lentamente a lo largo del tiempo.

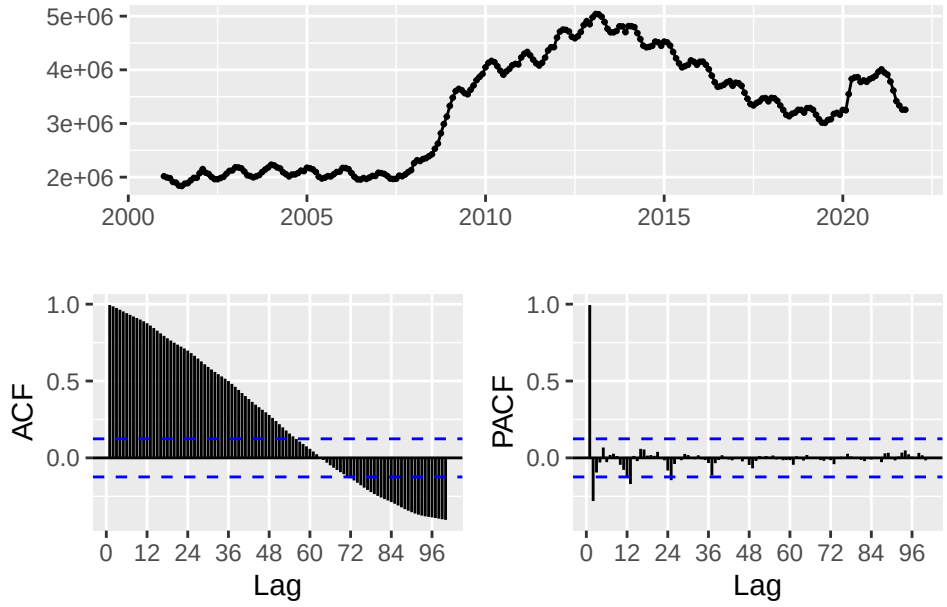

Figura 3: Gráfica datos

Tras haber diferenciado en la parte regular, volvemos a mostrar el ACF y el PACF.

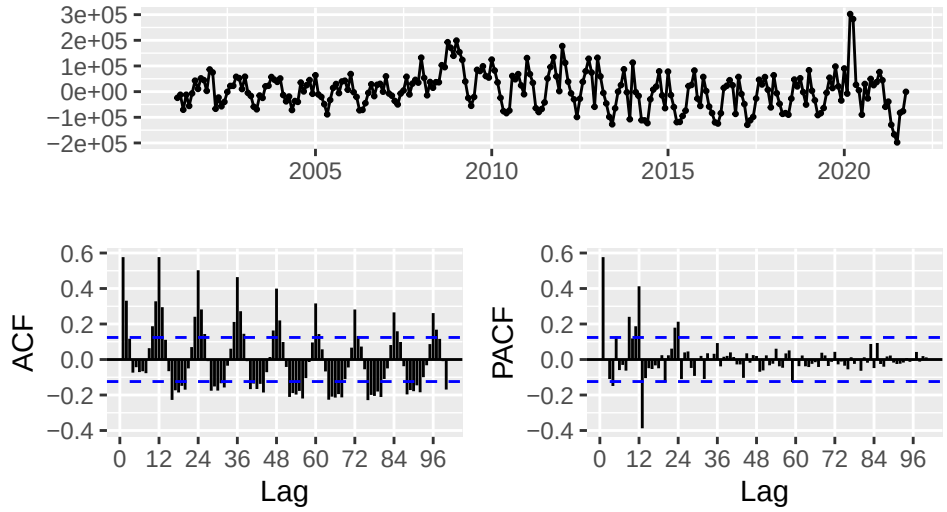

Figura 4: Gráfica tras diff regular

En la imagen superior vemos que cada 12 puntos aparece un pico y por tanto deducimos que tenemos que diferenciar en la parte estacional también.

Para posteriormente poder ajustar correctamente nuestra parte estacional del modelo, vamos a diferenciar la parte estacional y estudiar su gráfica ACF y PACF.

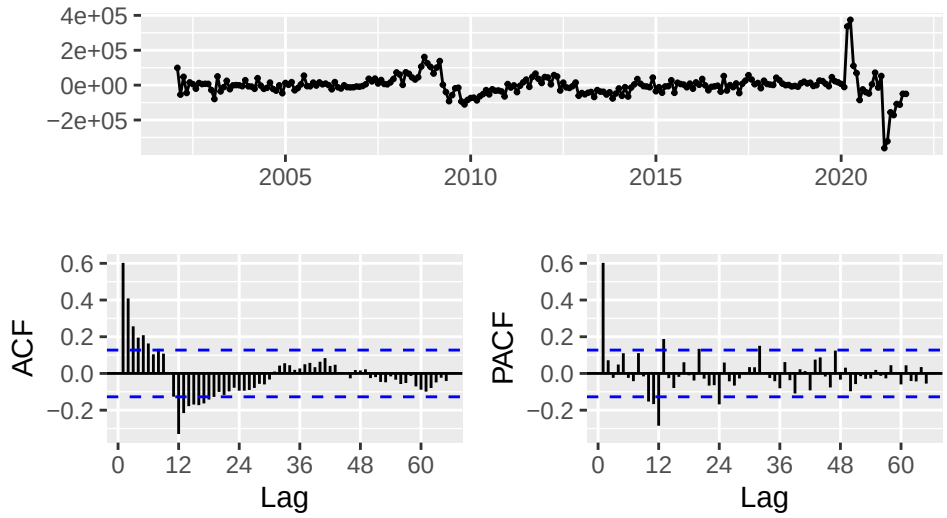

Figura 5: Gráfica tras diff estacional

Podemos ver ahora que la parte estacional ha mejorado ya que no se observan picos constantes en periodos de 12 (si es cierto que en algún múltiplo de 12 de forma puntual existe algún pico). Sin embargo, observamos cómo el ACF va decreciendo progresivamente (lo que lo asociamos a un proceso autorregresivo) y en PACF hay un coeficiente significante en el punto 12, por lo que en la parte estacional aplicamos un proceso AR(1).

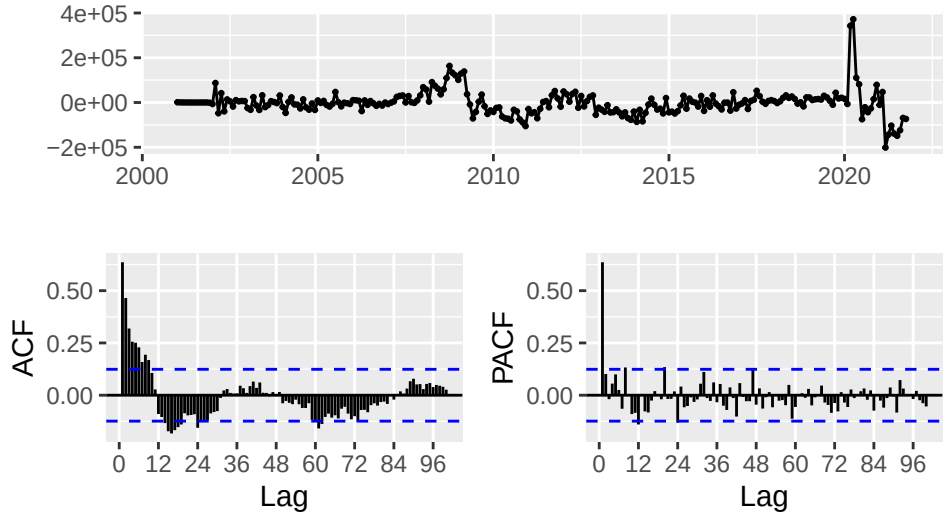

Figura 6: Residuos del modelo

Ahora en los residuos podemos ver que la parte estacional parece ajustada más o menos, sin embargo; la parte regular no esta bien modelada. Para ello miramos los residuos, como va decreciendo poco a poco en ACF y vemos un coeficiente bastante significativo en el PACF, volvemos a aplicar un AR(1). Por tanto volvemos a modelar nuestros datos con la parte regular ajustada.

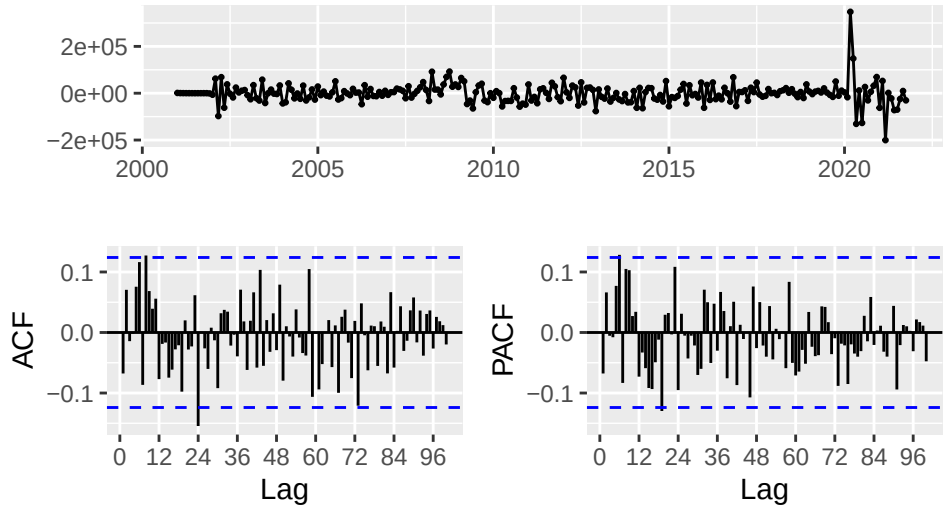

Figura 7: Residuos del modelo mejorado

Tras ver como quedan los residuos, lo que haremos será ver alguna información relevante a cerca de nuestro modelo. Alguno de estos datos son datos de control y error como el RMSE o el MAE.

```
## Series: y
## ARIMA(1,1,0)(1,1,0)[12]
##
## Coefficients:
##          ar1     sar1
##       0.6496  -0.5673
## s.e.  0.0497   0.0663
##
## sigma^2 estimated as 1.962e+09:  log likelihood=-2873.46
## AIC=5752.93   AICc=5753.03   BIC=5763.33
##
## Training set error measures:
##                     ME     RMSE      MAE        MPE      MAPE       MASE
## Training set -585.4926 42948.38 27682.11 0.02010853 0.8720577 0.09454437
##                     ACF1
## Training set -0.06769942
```

También es importante ver si nuestros coeficientes son realmente significativos, esto lo miraremos con la función coeftest.

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1   0.649599   0.049729 13.0629 < 2.2e-16 ***
## sar1 -0.567347   0.066343 -8.5517 < 2.2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obtenemos que ambos coeficientes son relevantes, por lo que podemos pasar a otro punto. Lo siguiente será ver donde caen las raices del polinomio característico. Para ver esto usaremos las inversas de las raices, las cuales si caen dentro del circulo unidad cumplirán lo que se pide para tener un proceso estacionario.

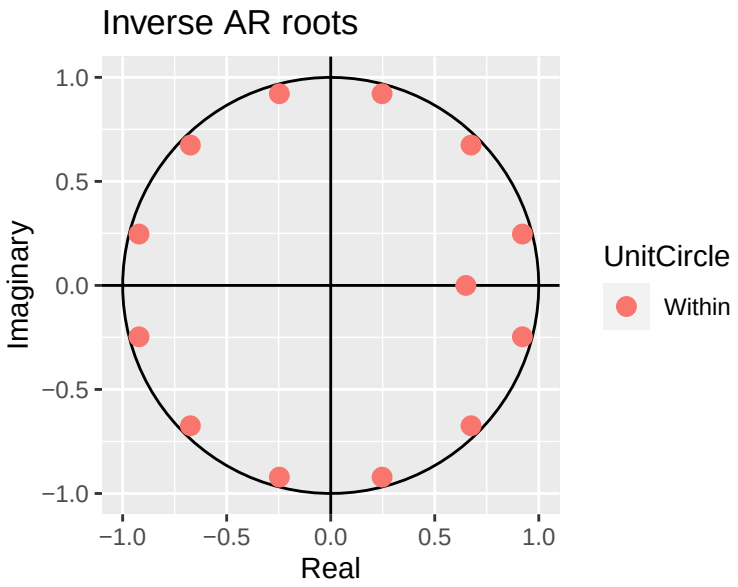

Figura 8: Inversas de raices

No obstante, en la gráfica de los residuos se podía ver que estos no son perfectamente ruido blanco en la parte estacional.

Por este motivo vamos a intentar ajustarlo de forma más precisa con un proceso MA(2).

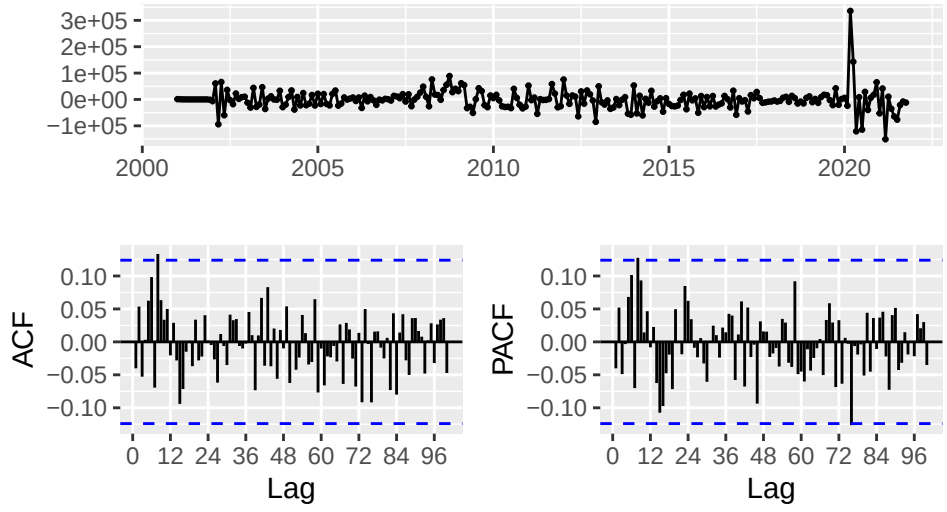

Figura 9: Residuos del modelo mejorado

Y volvemos a comprobar los mismos elementos anteriores con el modelo final. En este caso vemos también como se comportan los residuos.

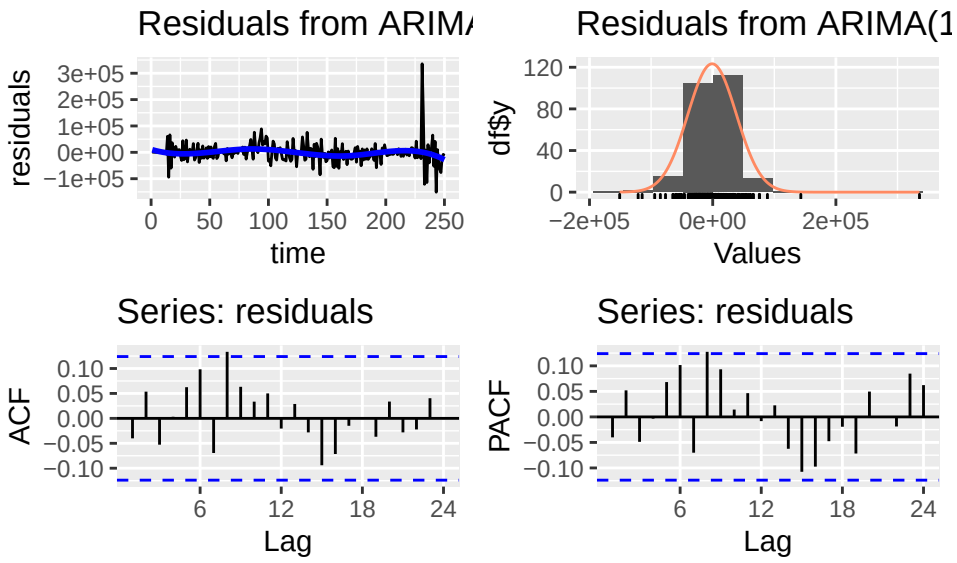

Figura 10: Análisis residuos

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,0)(1,1,2)[12]
## Q* = 19.082, df = 20, p-value = 0.5165
##
## Model df: 4.   Total lags used: 24
```

Después miramos los otros elementos restantes:

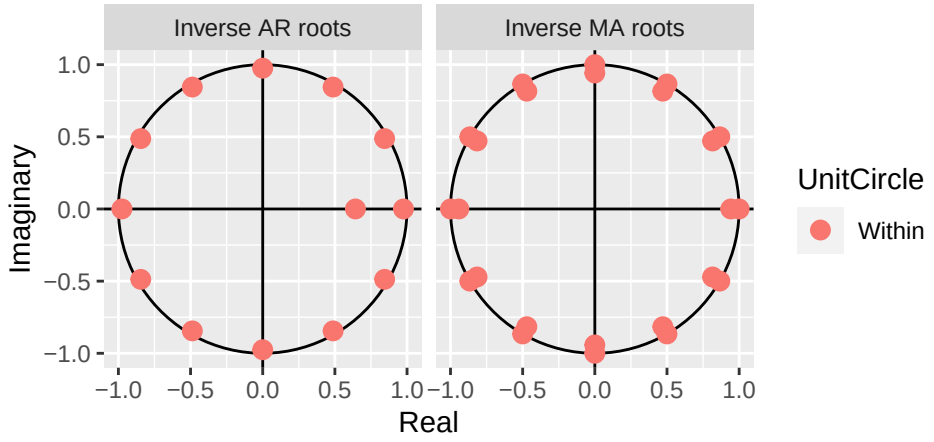

Figura 11: Inversas de raices

Podemos afirmar que de nuevo nuestras inversas de las raíces se encuentran dentro del círculo unidad, por lo tanto podemos afirmar que es estacionario.

Vemos que el ruido generado es ruido blanco, que todas las coeficientes del modelo son relevantes y que generan unos p-valores pequeños. Además, comprobamos que el modelo genera datos predichos que se adaptan

bien a los datos reales. Eso lo podemos ver en el siguiente gráfico:

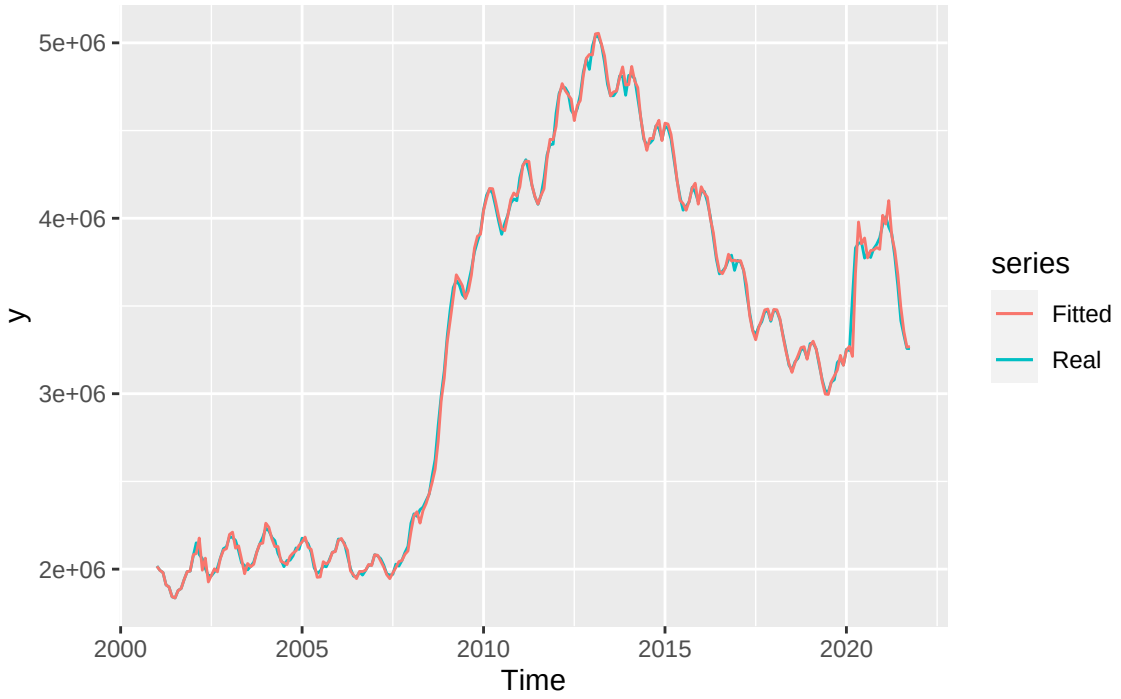

Figura 12: Comparación ajuste

Como podemos observar, la curva azul y la roja son muy parecidas y se adapta bien a todos los cambios. Lo que es muy satisfactorio. Por ello predecimos obteniendo el resultado enviado para el assigment

```
##          Point Forecast   Lo 80   Hi 80   Lo 95   Hi 95
## Nov 2021        3240623 3187368 3293877 3159177 3322068
```

2.3. Otros modelos intentados

Probamos también a crear otros modelos. Partimos de la base de haber diferenciado una vez tanto la parte estacionaria como la regular. Vemos que el ACF va disminuyendo de forma senoidal en la parte estacionaria, por lo que añadimos un MA(1). Tras ello, tenemos que ajustar la parte regular, añadiendo RA(2) y un MA(1).

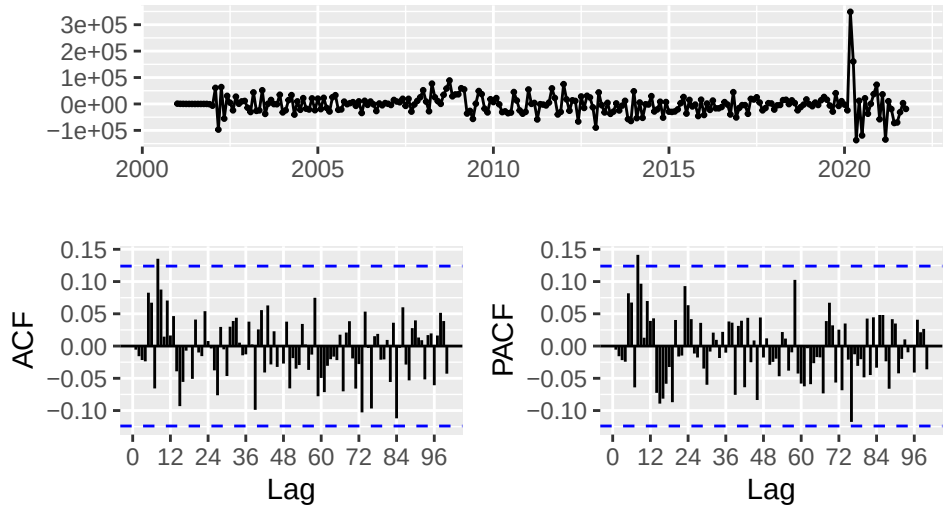

Figura 13: Modelo alternativo

Tras crear el nuevo modelo comprobamos algunas métricas relevantes como en el caso del modelo entregado.

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1  -0.230650   0.154521  -1.4927    0.1355
## ar2   0.607363   0.098842   6.1448 8.008e-10 ***
## ma1   0.856134   0.162073   5.2824 1.275e-07 ***
## sma1 -0.723409   0.068744 -10.5232 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Series: y
## ARIMA(2,1,1)(0,1,1)[12]
##
## Coefficients:
##           ar1     ar2     ma1     sma1
##       -0.2307  0.6074  0.8561  -0.7234
## s.e.   0.1545  0.0988  0.1621   0.0687
##
## sigma^2 estimated as 1.787e+09:  log likelihood=-2863.4
## AIC=5736.8   AICc=5737.06   BIC=5754.14
##
## Training set error measures:
##                      ME     RMSE      MAE         MPE      MAPE       MASE
## Training set -920.8681 40813.86 25549.44 0.007740555 0.8067537 0.08726053
##                      ACF1
## Training set -0.005591997
```

Tras ello comprobamos el comportamiento de los residuos:

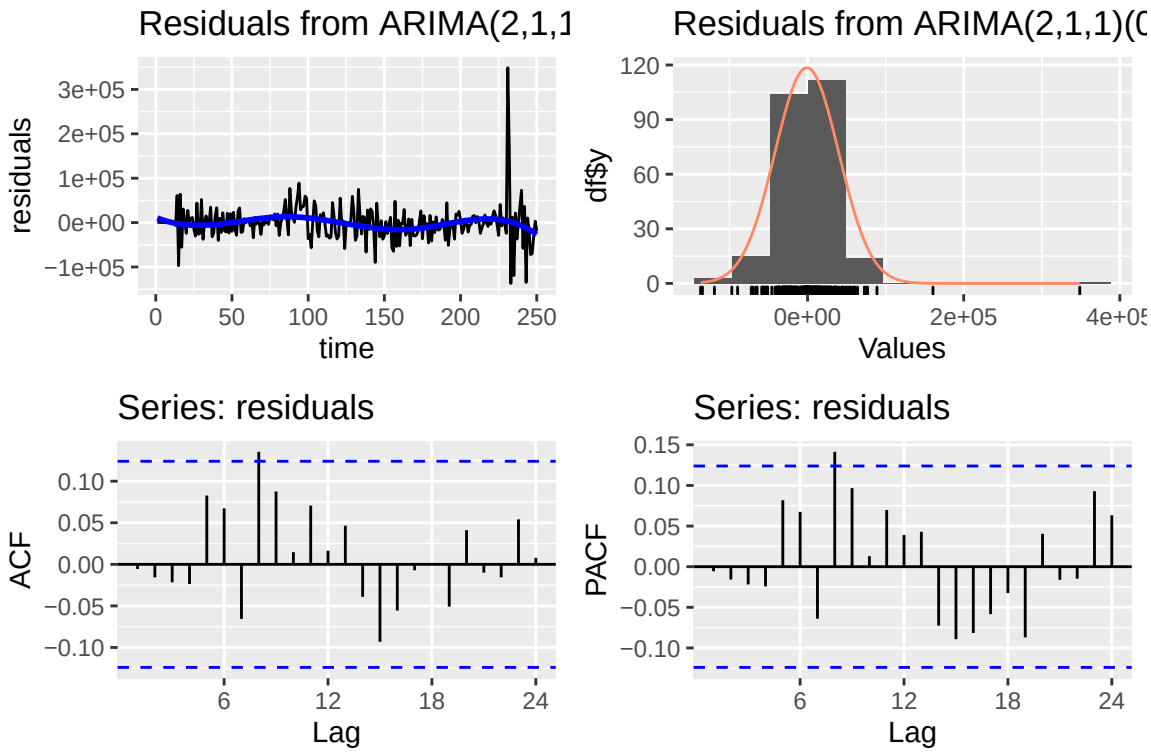

Figura 14: Análisis residuos

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,1)(0,1,1)[12]
## Q* = 18.839, df = 20, p-value = 0.5323
##
## Model df: 4.   Total lags used: 24
```

Por último comprobamos que tal ajusta nuestro modelo con los datos:

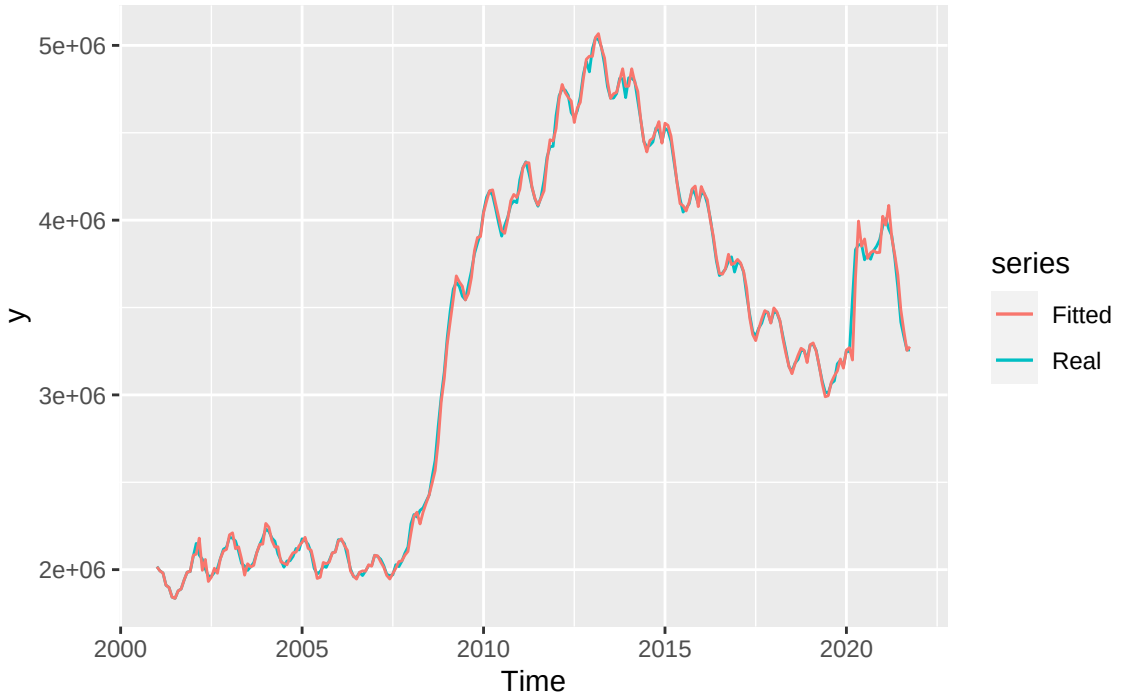

Figura 15: Comparación ajuste

Sin embargo, estos resultados a pesar de que generan ruido blanco y la mayoría de variables aparecen muy significativas, generaba resultados ligeramente peores que los del modelo anterior. Pero esto lo veremos en mayor profundidad en un apartado posterior donde compararemos todos nuestros modelos.

```
##          Point Forecast   Lo 80   Hi 80   Lo 95   Hi 95
## Nov 2021        3220246 3166067 3274426 3137386 3303107
```

3. Nuevos modelos

3.1. SARIMA con Recorte Periodo de Tiempo

Para buscar una predicción más acertada a la realidad, hemos creado varios modelos nuevos. La línea de trabajo que hemos adoptado ha sido analizar la serie temporal después de la crisis de 2008, es decir; cuando los datos del paro empiezan a decrecer. De esta forma intentamos conseguir un modelo que no esté mal influido por los datos previos donde se producía un aumento del paro.

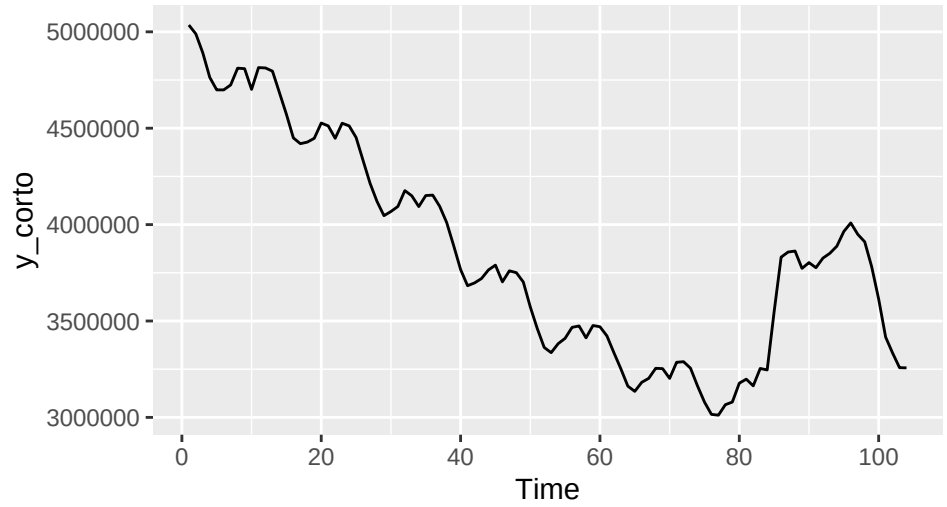

Figura 16: Datos recortados

Una vez tenemos estos datos, empezamos procediendo de igual forma. Analizamos si la serie es estacionaria en varianza.

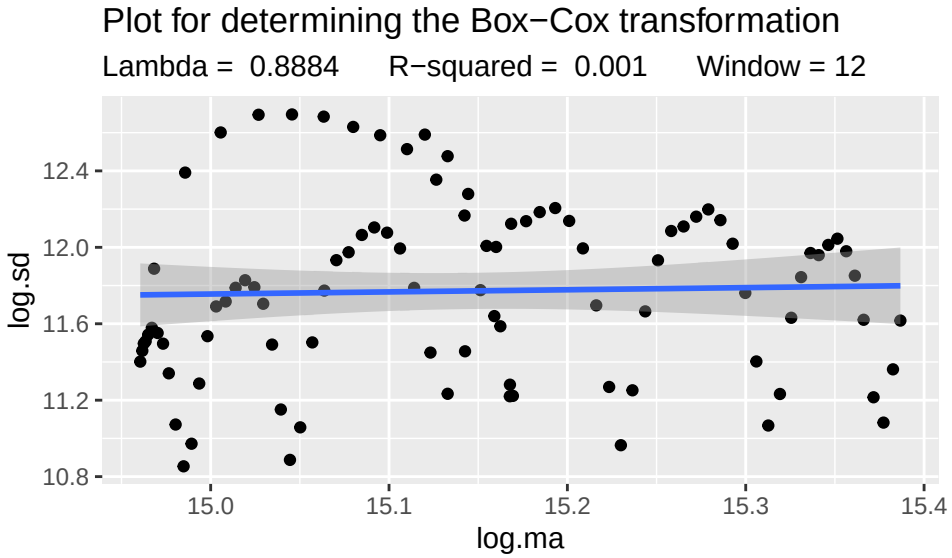

Figura 17: Box-Cox datos recortados

Que como se puede ver lo es ya que la línea azul es práctiacmente horizontal y R-squared = 0.001. Por lo tanto, no es necesario realizar una estabilización de varianza.

Empezamos por tanto representando la serie temporal junto con su ACF y PACF para inspeccionar la serie regular y estacional

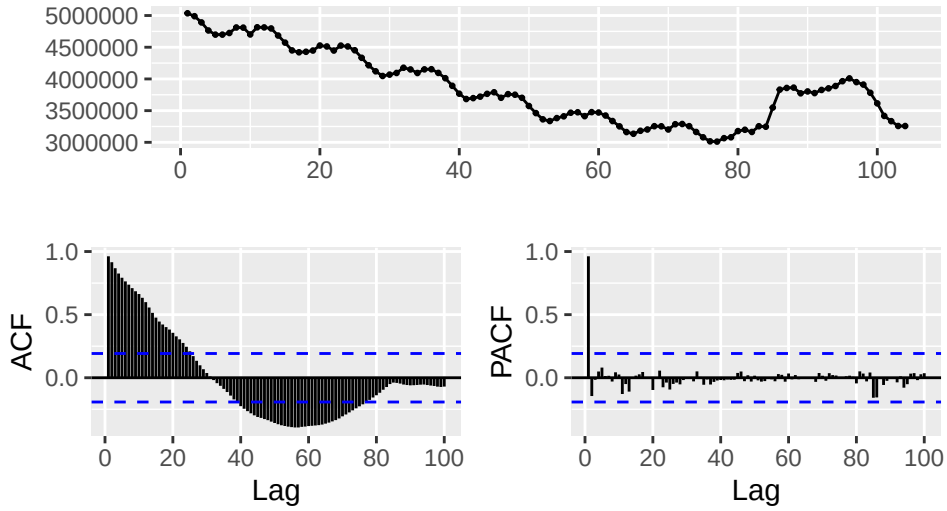

Figura 18: Modelo recortado

Vemos que la serie necesita una diferenciación en la parte regular ya que el ACF dismunye progresivamente por el tiempo y volvemos a mostrar el ACF y el PACF.

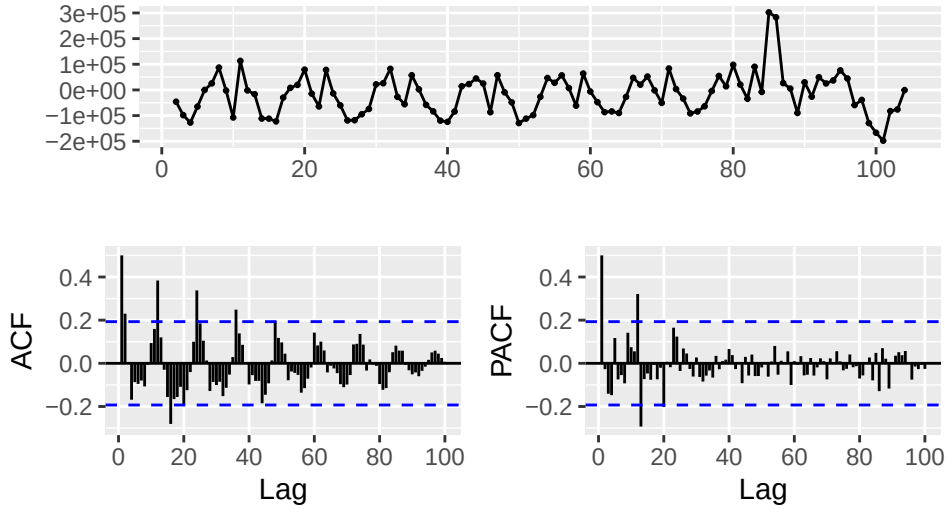

Figura 19: Modelo tras diff

Vemos que cada 12 puntos aparece un pico y por tanto deducimos que tenemos que diferenciar en la parte estacional también.

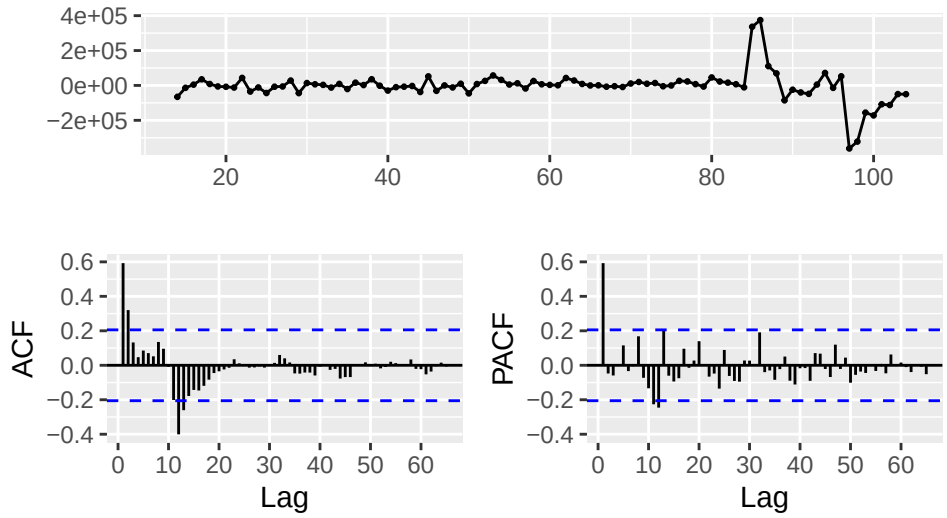

Figura 20: Modelo tras diff estacional

Ahora vemos que en el punto 12, el primero de la parte estacional se produce un residuo significativo, a su vez también observamos que en el PACF se genera un coeficiente significativo en el número 12. Por ello, aplicamos un RA(1) y volvemos a representar los ACF y PACF para poder ajustar la parte regular.

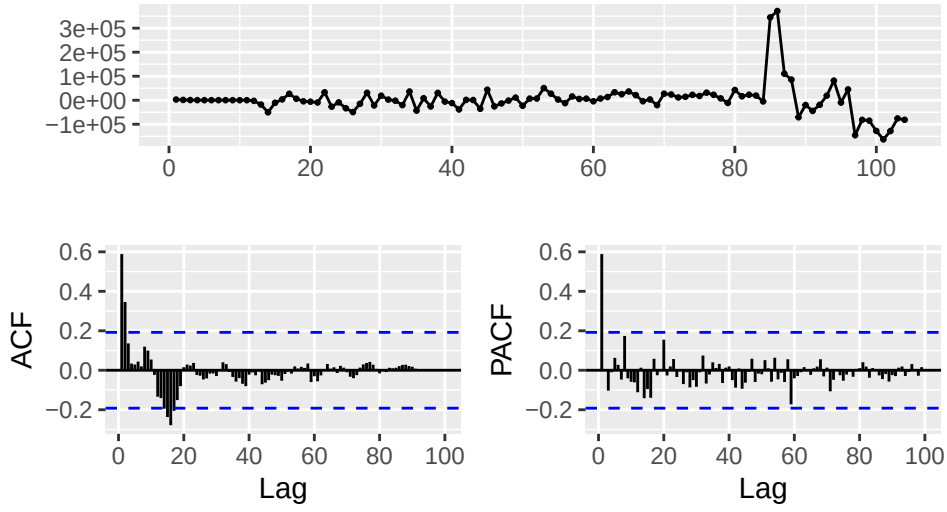

Figura 21: Modelo parte estacional

Observamos que hay dos coeficientes bastante significativos, uno en el ACF y otro en el PACF por ello aplicaremos conjuntamente un AM(1) y un AR(1).

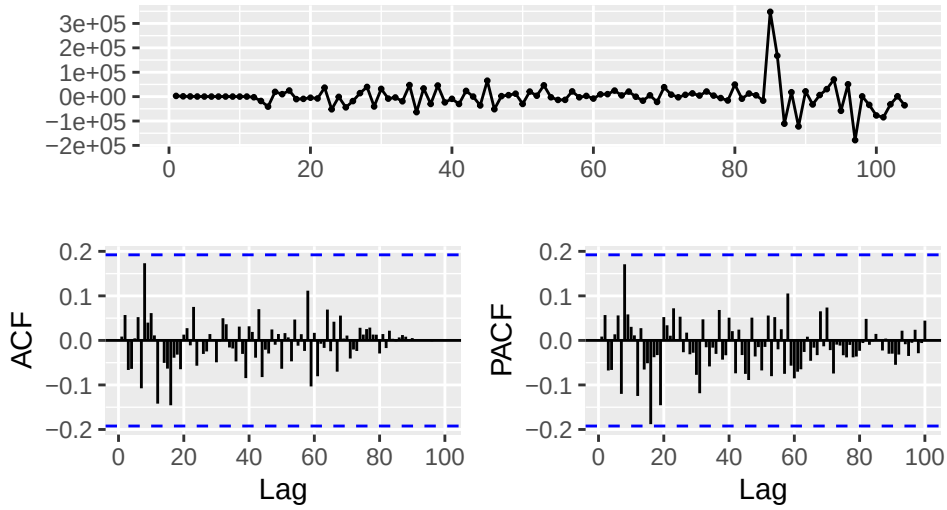

Figura 22: Modelo final

Ahora miraremos un resumen de nuestro modelo junto con la significiación de los coeficientes.

```
## Series: y_corto
## ARIMA(1,1,1)(1,1,0)[12]
##
## Coefficients:
##          ar1      ma1     sar1
##       0.6000  -0.0086  -0.6237
## s.e.  0.1309   0.1549   0.0954
##
## sigma^2 estimated as 3.248e+09:  log likelihood=-1127.22
## AIC=2262.44   AICc=2262.91   BIC=2272.49
##
## Training set error measures:
##                     ME     RMSE     MAE        MPE      MAPE      MASE      ACF1
## Training set 221.1872 52421.33 28044.7 0.01935668 0.7480002 0.4488551 0.0086689
##
## z test of coefficients:
##
##        Estimate Std. Error z value  Pr(>|z|)
## ar1   0.6000344  0.1308658  4.5851 4.537e-06 ***
## ma1  -0.0086143  0.1549499 -0.0556    0.9557
## sar1 -0.6237171  0.0954476 -6.5347 6.376e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tras ello estudiaremos y evaluaremos los residuos generados por nuestro modelo.

```
##
##  Ljung-Box test
```

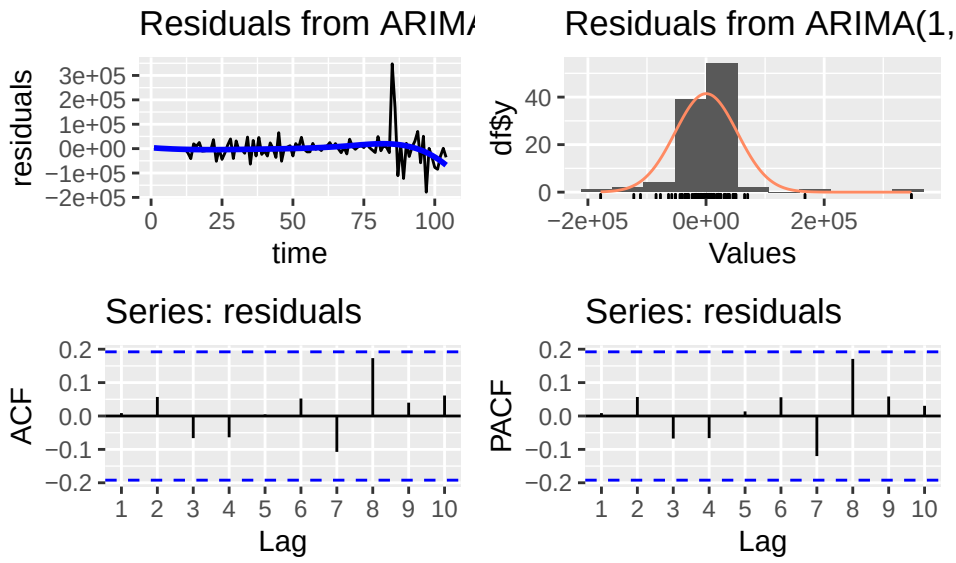

Figura 23: Revisión Residuos

```
##
## data:  Residuals from ARIMA(1,1,1)(1,1,0)[12]
## Q* = 6.9895, df = 7, p-value = 0.43
##
## Model df: 3.   Total lags used: 10
```

Tras ello miramos las inversas de las raices para comprobar que nuestro proceso es estacionario.

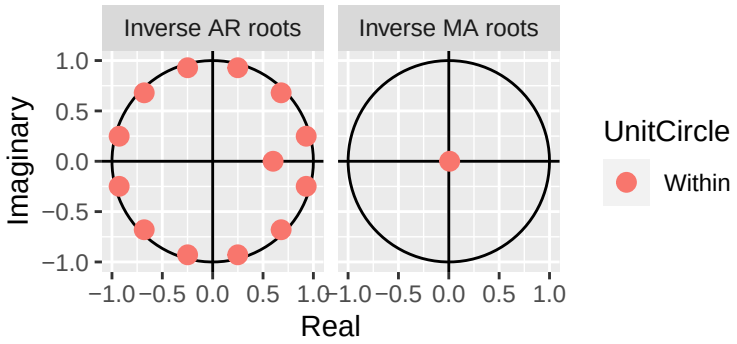

Figura 24: Inversas de raices

Por ultimo vamos a comparar que tal ajusta nuestro modelo a la serie temporal sobre la que hemos trabajado.

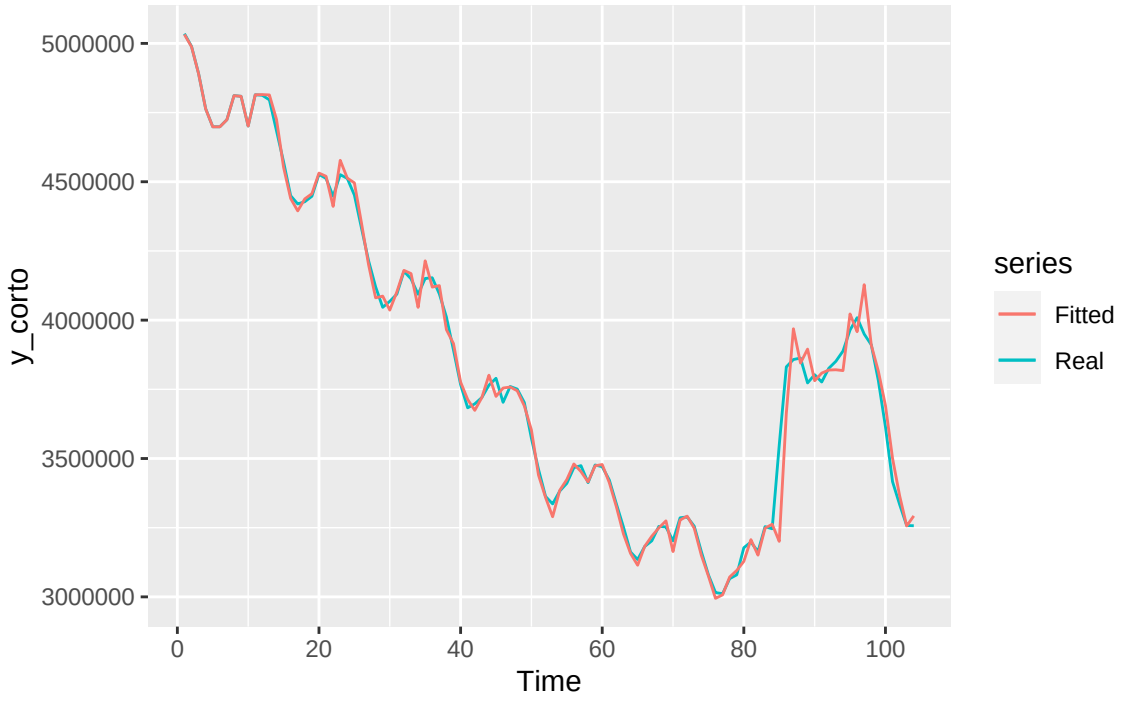

Figura 25: Comparación Modelaje

Y también veremos la predicción que nos da nuestro modelo para este mes.

```
##     Point Forecast   Lo 80   Hi 80   Lo 95   Hi 95
## 105        3231397 3158364 3304430 3119703 3343092
```

3.2. Regresión Dinámica

En este apartado lo que trataremos de hacer será realizar un modelo de regresión dinámica para poder predecir el paro de noviembre.

Para ello vamos a crear un nuevo conjunto de datos, donde tengamos la variable COVID donde el 0 indica que no fue periodo COVID y el 1 indica periodo COVID. Además ampliaremos el periodo COVID hasta hoy en día (como experimento). A su vez para ver mejor la gráfica inicial lo que vamos a hacer es expresar el paro en millones, por tanto un valor de 3 representa 3000000.

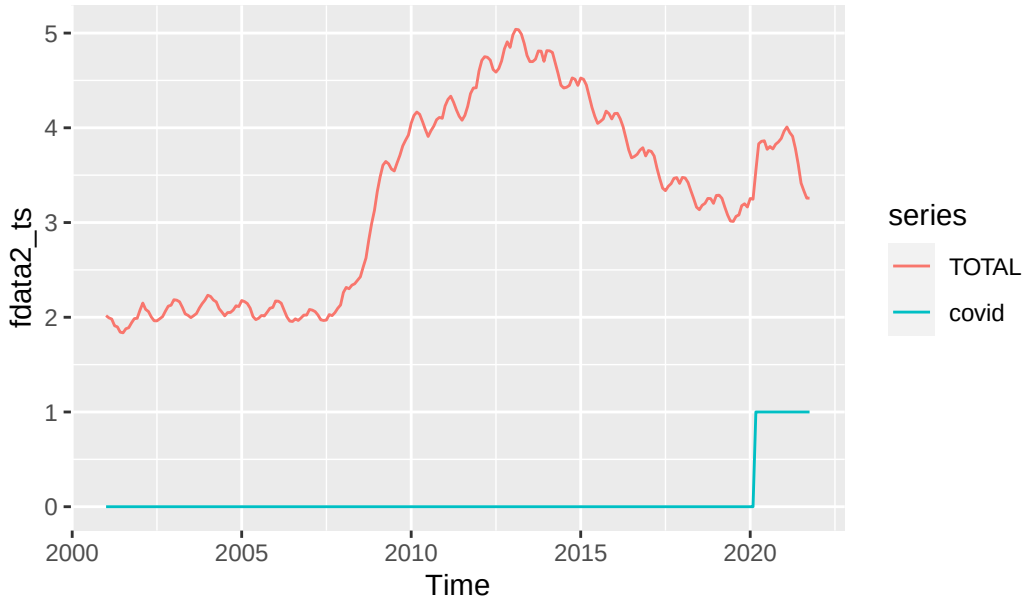

Figura 26: Gráfico de datos

Tras ello iniciaremos nuestro modelo de regresión dinámica donde tendremos como variable de salida el paro. Comenzamos con un modelo simple que usaremos para ajustar posteriormente los coeficientes.

Lo primero que veremos será la significancia de los coeficientes.

```
##
## z test of coefficients:
##
##            Estimate Std. Error  z value  Pr(>|z|)
## ar1       0.9941703  0.0046667 213.0342 < 2.2e-16 ***
## sar1      0.8268666  0.0339370  24.3648 < 2.2e-16 ***
## intercept 1.6627317  1.5689192   1.0598    0.2892
## T1-MA0    0.3477648  0.0302234  11.5065 < 2.2e-16 ***
## T1-MA1    0.3467540  0.0302255  11.4722 < 2.2e-16 ***
## T1-MA2    0.1311715  0.0302325   4.3388 1.433e-05 ***
## T1-MA3    0.1182004  0.0302344   3.9095 9.250e-05 ***
## T1-MA4    0.0091683  0.0302314   0.3033    0.7617
## T1-MA5    0.0434148  0.0302216   1.4365    0.1508
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obtenemos que una parte considerable de nuestros coeficientes son significantes.

Tras ello pasamos a ver los errores, donde en caso de no ser estacionarios tendremos que diferenciar para conseguir que sea estacionario.

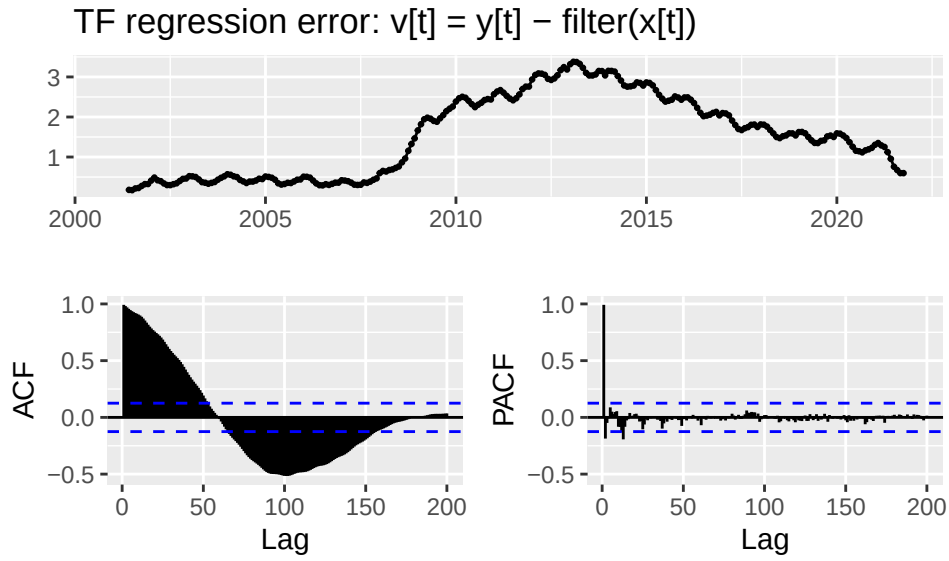

Figura 27: Estudio de los errores

Es bastante claro que los errores no son estacionarios, por tanto diferenciaremos.

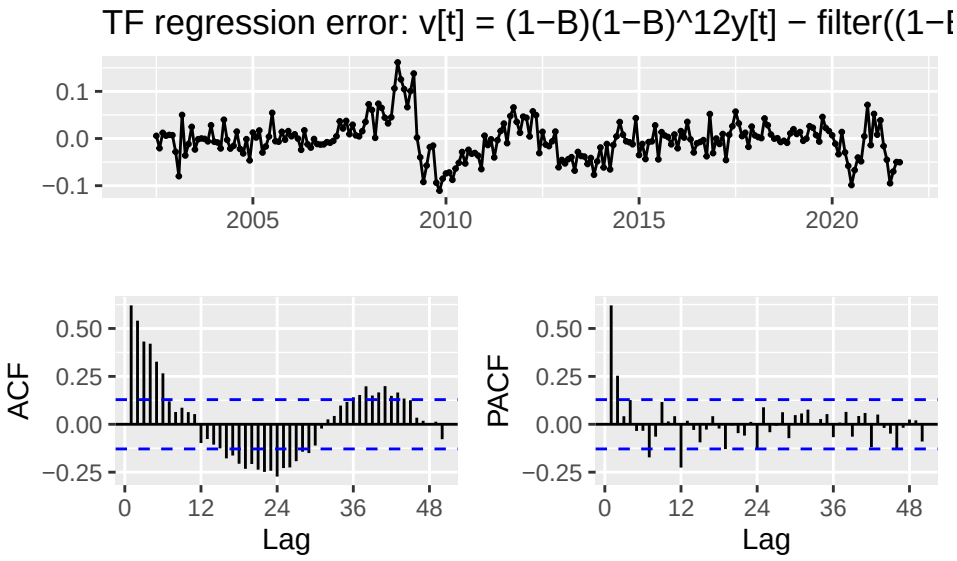

Figura 28: Estudio de los errores

Tras diferenciar tanto la parte regular como la estacional ya obtenemos unos errores que son estacionarios (o en gran medida).

A continuación lo que haremos será identificar los coeficientes de nuestro modelo.

```
##          Estimate Std. Error    z value     Pr(>|z|)
## T1-MA0 0.36948979 0.02475712 14.9245895 2.280345e-50
## T1-MA1 0.36054561 0.02944004 12.2467780 1.748036e-34
## T1-MA2 0.13998612 0.03093670  4.5249209 6.041810e-06
## T1-MA3 0.12690884 0.03092483  4.1037848 4.064456e-05
```

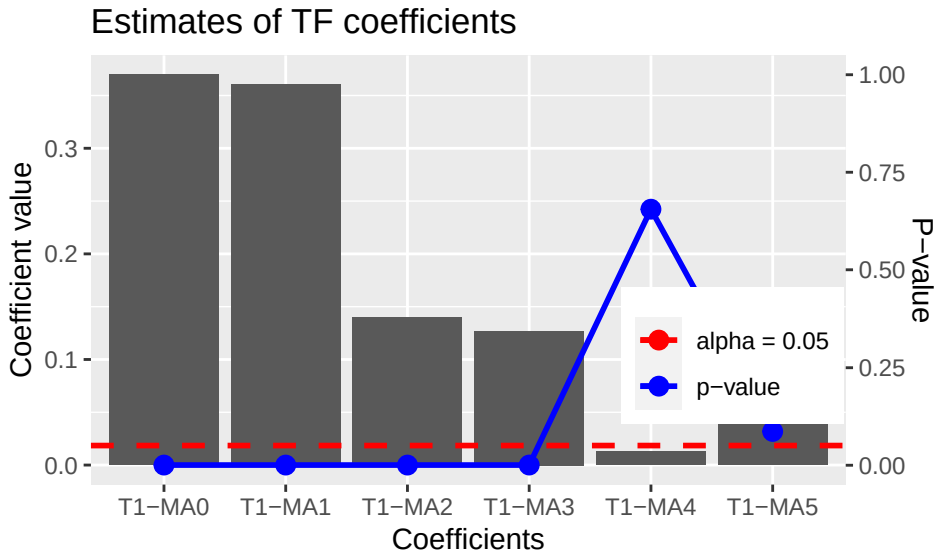

Figura 29: Estimación de coefs

```
## T1-MA4 0.01310360 0.02939991  0.4457022 6.558124e-01
## T1-MA5 0.04241736 0.02472250  1.7157392 8.620978e-02
```

Para el valor de b lo que haremos será ver cuantos coeficientes no significativo hay antes del primer coeficientes significativo, que en nuestro caso será el primero. Por tanto tendremos que b = 0. Para s lo que haremos será ver cuantos coeficientes significativos desde el primero (no incluido) hay hasta que aparece el primer coeficiente no significativo, que en nuestro caso es 3. Por último para ajustar r miramos el comportamiento de los coeficientes, donde se puede ver que la caida es de tipo exponencial, por tanto tomamos r igual a 1.

Tras realizar este ajuste lo que haremos será ajustar los ordenes de las partes estacional y regular. Esto lo haremos con los errores que hemos visto antes.

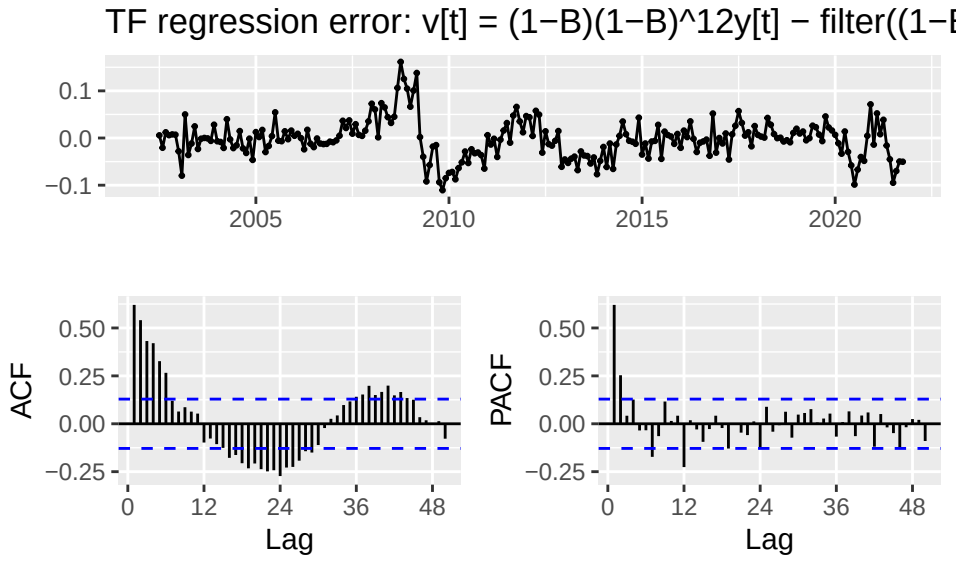

Figura 30: Estudio de los errores

Podemos ver que la gráfica ACF oscila y va decreciendo poco a poco, por lo que nos encontramos ante un proceso autorregresivo. En la parte regular tomamos un proceso AR(1) junto con un AM(1).

Ahora volvemos a realizar nuestro modelo pero ya con los coeficientes estimados y obtenemos una serie de métricas.

```
##
## Call:
## arima(x = y2, order = c(1, 1, 1), seasonal = list(order = c(1, 1, 0), period = 12),
##     include.mean = FALSE, method = "ML", xtransf = xlag, transfer = list(c(1,
##         3)))
##
## Coefficients:
##          ar1      ma1     sar1   T1-AR1  T1-MA0  T1-MA1  T1-MA2  T1-MA3
##       0.8712  -0.4121  -0.2291  -0.1584  0.3542  0.4018  0.1781  0.1257
## s.e.  0.0467   0.0848   0.0704   0.3209  0.0237  0.1086  0.1024  0.0313
##
## sigma^2 estimated as 0.0009196:  log likelihood = 485.34,  aic = -954.67
##
## Training set error measures:
##                        ME       RMSE        MAE          MPE      MAPE
## Training set -0.001052443 0.02951947 0.02188013 -0.004843168 0.6975851
##                    MASE         ACF1
## Training set 0.3921479 0.0008568904
```

Tras un breve resumen de elementos clave veamos significancia de coeficientes, residuos, etc.

```
##
## z test of coefficients:
##
##         Estimate Std. Error z value  Pr(>|z|)
## ar1     0.871195   0.046661 18.6706 < 2.2e-16 ***
## ma1    -0.412051   0.084777 -4.8604 1.171e-06 ***
## sar1   -0.229143   0.070433 -3.2533 0.0011406 **
## T1-AR1 -0.158411   0.320927 -0.4936 0.6215843
## T1-MA0  0.354215   0.023735 14.9240 < 2.2e-16 ***
## T1-MA1  0.401776   0.108615  3.6991 0.0002164 ***
## T1-MA2  0.178054   0.102429  1.7383 0.0821534 .
## T1-MA3  0.125671   0.031273  4.0185 5.857e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lo primero que miraremos serán los residuos de nuestro modelo, donde miraremos si estos son ruido blanco.

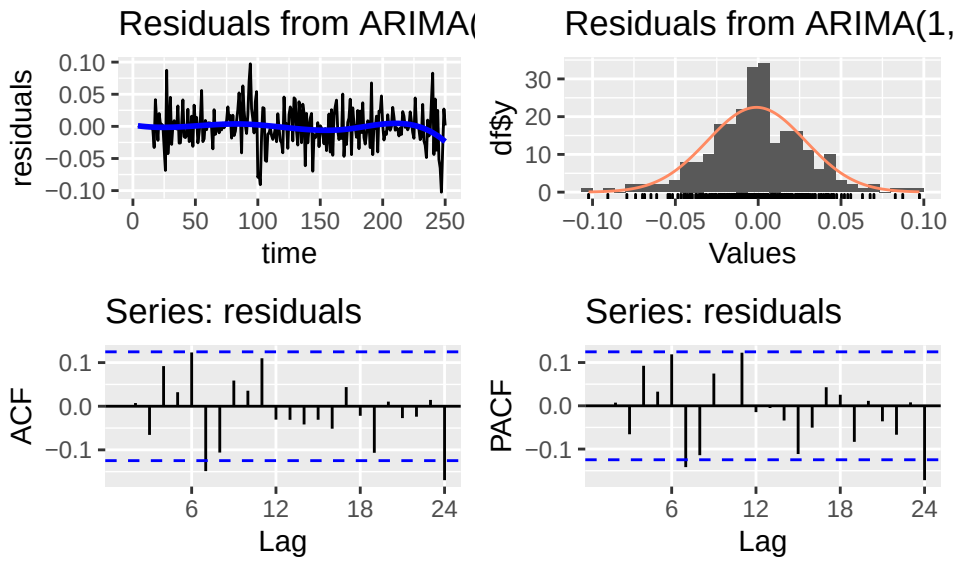

Figura 31: Estudio de los errores

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,1)(1,1,0)[12]
## Q* = 34.349, df = 16, p-value = 0.004875
##
## Model df: 8.   Total lags used: 24
```

A pesar de que un coeficiente salga un poco de la banda de confianza podemos asumir que nuestro ruido es un ruido blanco.

Veamos el gráfico CCF para comprobar correlaciones entre coeficientes.

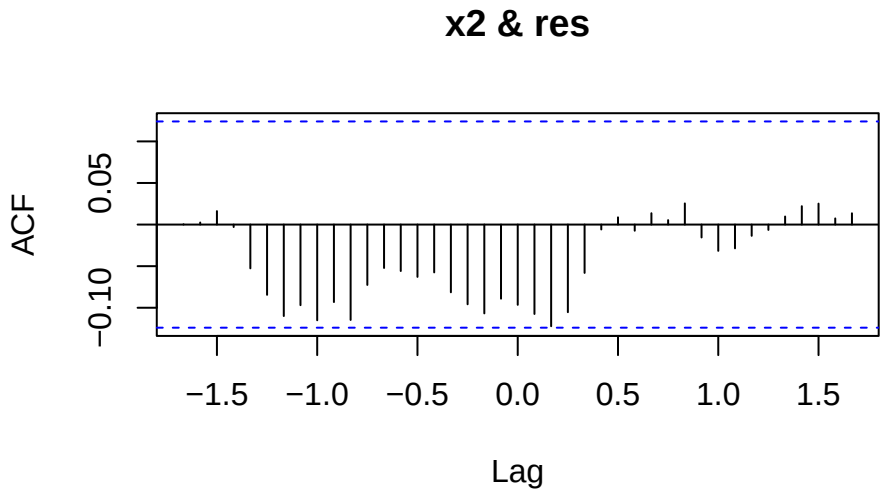

Figura 32: Gráfico CCF

Por último vamos a ver que tal ajusta nuestro modelo la serie temporal sobre la que hemos estado trabajando.

Figura 33: Diferencia Modelaje

Vemos que nuestro modelo ajusta bastante bien la serie temporal sobre la que hemos trabajado. Si es cierto que le cuesta algo más los cambios bruscos o los picos exagerados en poco tiempo.

No obstante, aquí se nos presenta un nuevo reto: saber cuándo tenemos que modificar la variable x y volver a configurarla a 0, es decir; saber cuándo el COVID-19 ha dejado de tener efecto. Para ello hemos entrenado un modelo ARIMA (1,1,1),(1,1,0)(12) (el mismo que usamos arriba) y datos hasta el inicio de la época COVID, febrero 2020. Una vez que lo tenemos entrenado predecimos los resultados hasta el mes de noviembre de este año, 21 valores nuevos, como se ven en la imagen.

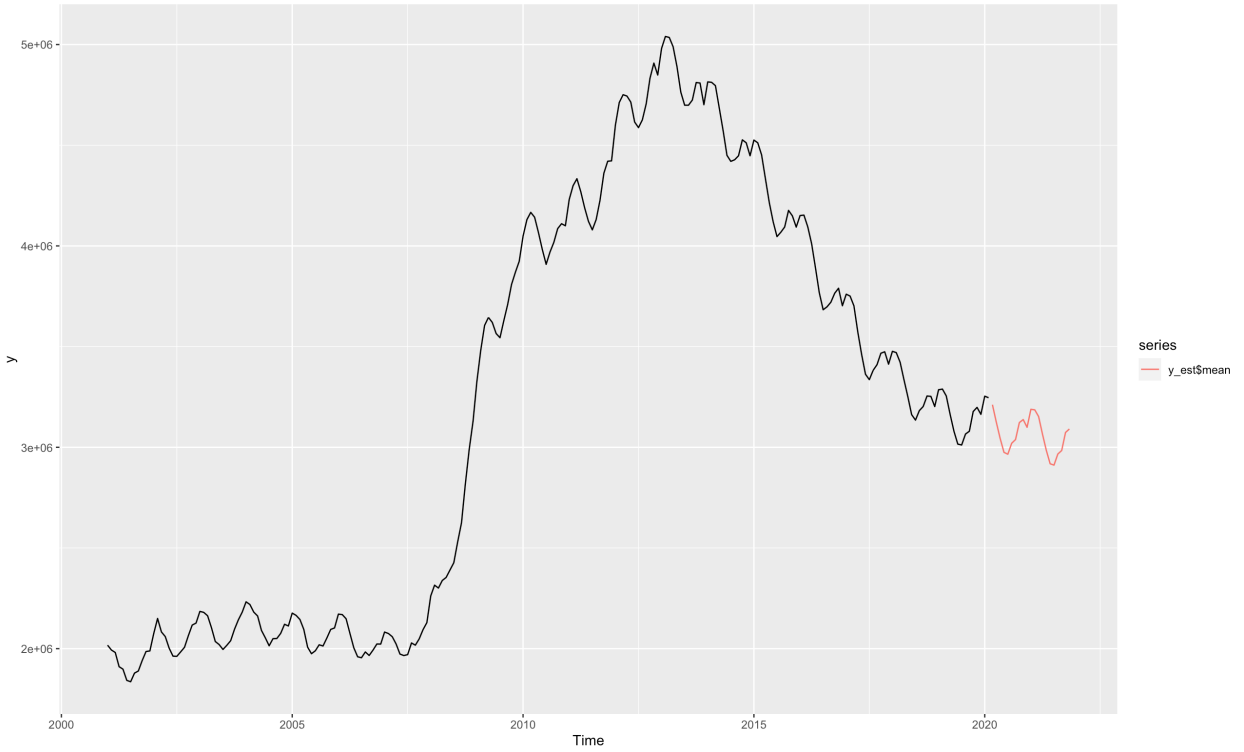

Figura 34: Predicción modelo sin Covid

El valor predicho para el mes de noviembre sin tener en cuenta el factor COVID-19 es aproximadamente 3.100.000. La idea es comparar este valor con el valor predicho teniendo en cuenta la variable COVID, que ha sido aproximadamente 3.208.000. Esto quiere decir que a día de hoy sigue influyendo el factor covid en el número de desempleados y por lo tanto, es correcto utilizar la variable $x = 1$ para modelar nuestro modelo.

4. Comparación modelos

Una vez presentados diversos modelos para predecir el paro lo que debemos hacer es entender y conocer cual es el mejor de ello, dando puntos a favor y puntos en contra a cada uno de ellos.

El primer parámetro que miraremos será el conjunto de errores que tiene cada modelos (RMSE, MSE...) para así conocer cual es el que minimiza el error en nuestros datos.

–	ME	RMSE	MAE
SARIMA	-585,4926	42948,38	27682,11
SARIMA2	-920,8681	40813.86	25549,44
REG DIN	1,18	39103,75	28012,7
ARIMA REC	221,1872	52421,33	28044,7

Otra forma de comparar los modelos es a través de sus predicciones y modelado de la serie temporal.

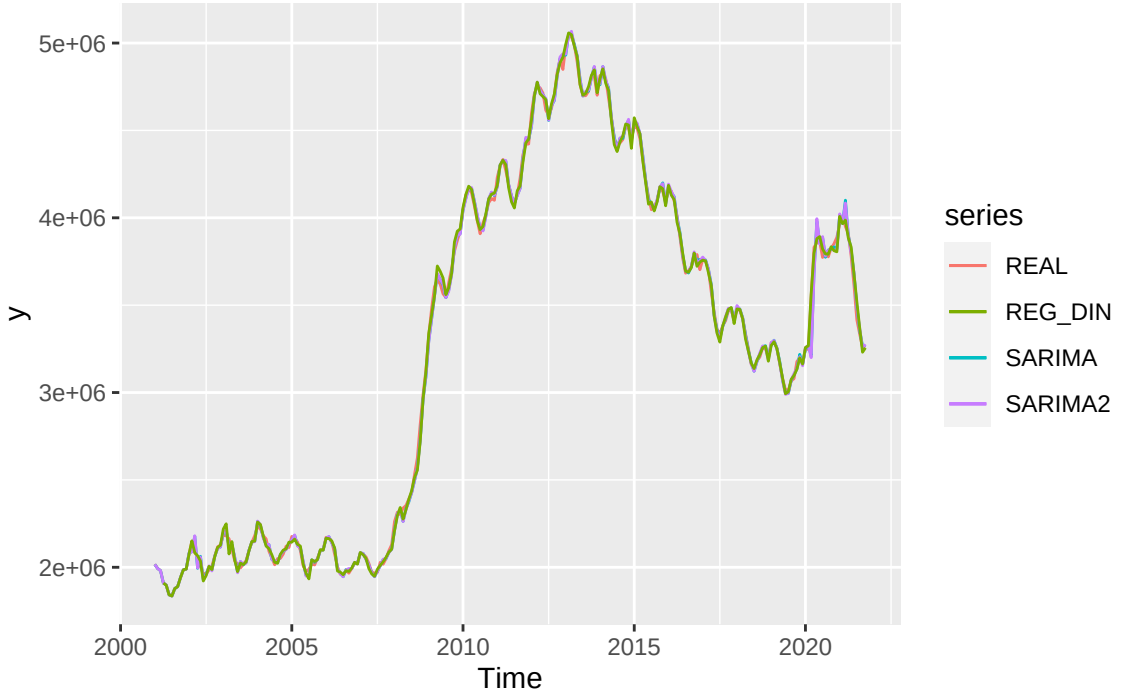

Figura 35: Diferencia Modelaje

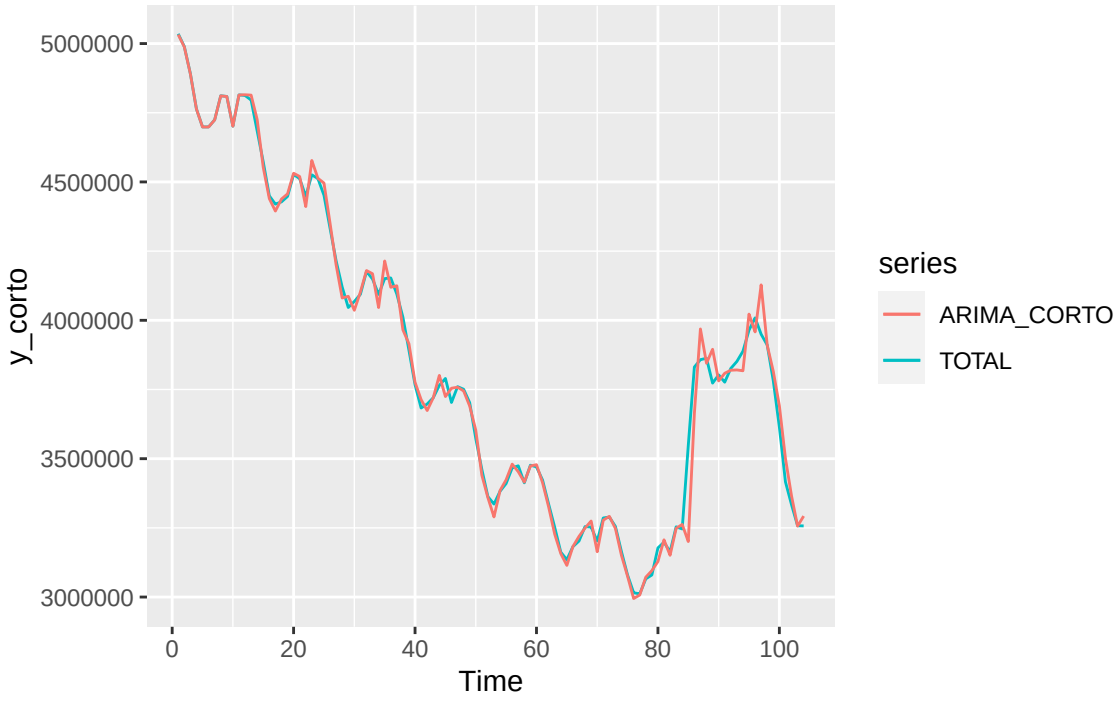

Figura 36: Diferencia Modelaje