Ajuste SARIMA serie de Nuevas Vacunaciones contra el COVID-19 en Nueva Zelanda

Por:

Jhon Alexander Bedoya Carvajal

Juan Daniel García Espinosa

Materia:

Series de Tiempo

Profesora:

Mariana Alejandra Palacio Rodríguez

UNIVERSIDAD DE ANTIOQUIA

1803

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN

2021

Ajuste SARIMA serie de Nuevas Vacunaciones en Nueva Zelanda

La serie de Nuevas Vacunaciones en Nueva Zelanda, es una serie que va desde el 24 de febrero de 2021 hasta el 10 de agosto de 2021 y cuenta con 168 observaciones.

Se decidió no tomar los datos desde el 20 de febrero porque las observaciones del 20 al 23 de febrero son valores muy pequeños comparados con el resto de la serie y afectan bastante la varianza de la serie.

Se hizo validación cruzada para dividir la base de datos en los datos de entrenamiento y los datos de prueba, con una proporción de 90 % y 10 %, respectivamente, quedando 151 observaciones para entrenamiento y 17 observaciones para prueba.

En adelante, cuando se mencione la serie, será la serie de entrenamiento.

A continuación de presenta el gráfico de la serie y el gráfico de su ACF, para estudiar el comportamiento y algunas características de esta.

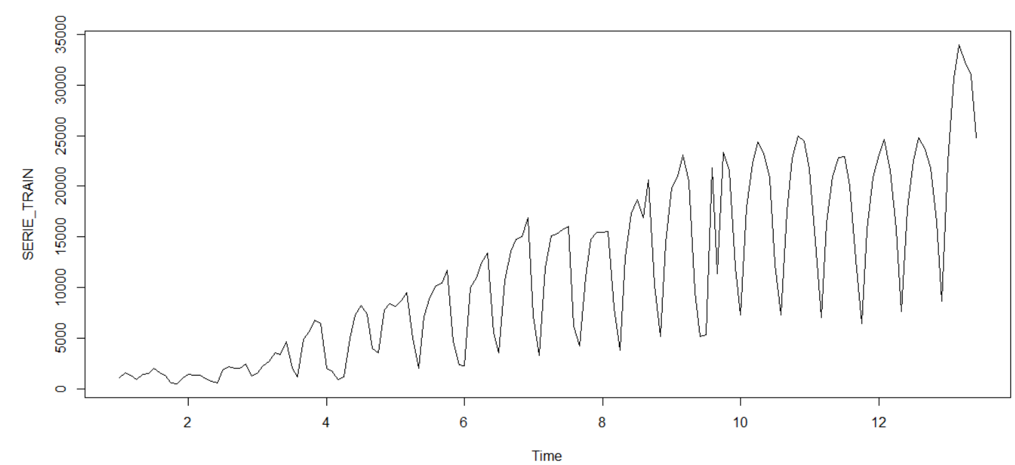


Gráfico 1. *Serie de nuevos vacunados en Nueva Zelanda*

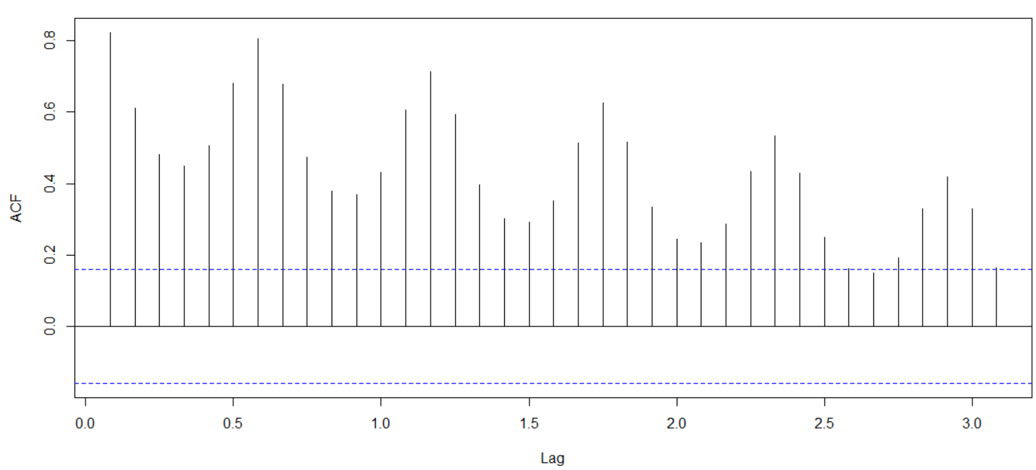


Gráfico 2. *ACF Serie de nuevos vacunados en Nueva Zelanda*

En el gráfico de la serie (Gráfico 1) se puede observar que la serie tiene una tendencia global creciente y el ACF (Gráfico 2) lo confirma con el decaimiento lento a cero que tienen los rezagos. Los datos no oscilan alrededor de una media constante y la serie tiene tendencia, por tanto, se puede afirmar que la serie no es estacionaria en media. Aparentemente, la serie tampoco es estacionaria en varianza, puesto que la amplitud de los datos inicialmente es muy pequeña y va creciendo considerablemente. El ACF (Gráfico 2) nos muestra que la serie tiene una estacionalidad de, pues cada siete periodos un rezago es mucho más significativo, y como la serie no es estacionaria en varianza, esta estacionalidad es multiplicativa y resulta conveniente aplicar logaritmo a la serie para tener una serie con varianza constante.

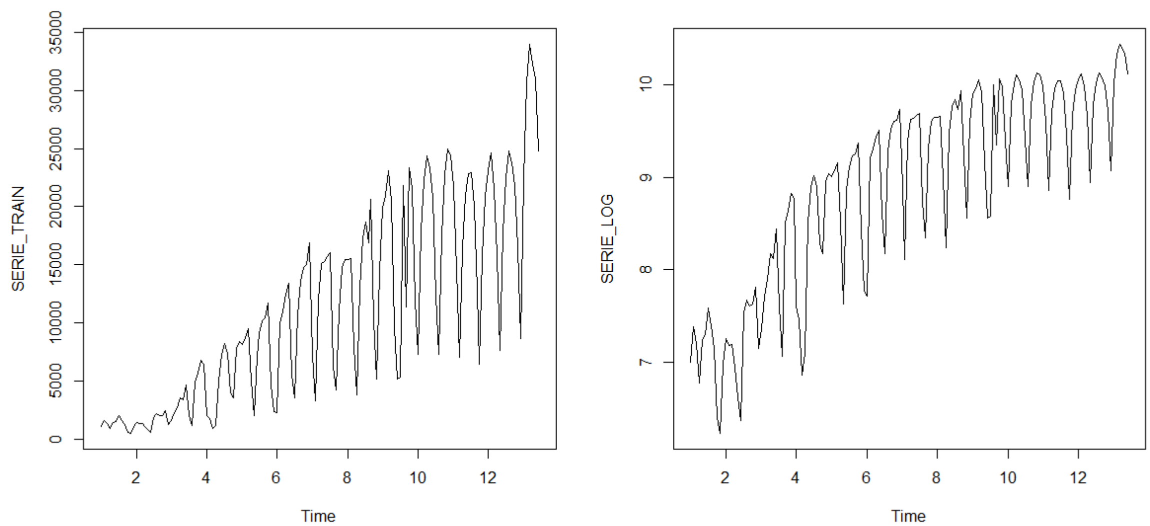


Gráfico 3. *Serie vs Ln(Serie)*

Al aplicar Logaritmo a la serie, obtenemos una serie en la que la varianza es aproximadamente constante, es decir, es estacionaria en varianza. Desde ahora se trabajará con la serie con Logaritmo.

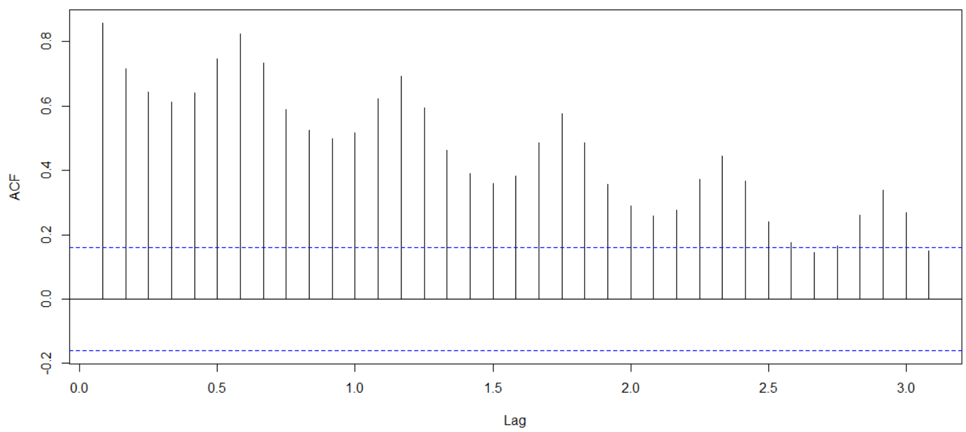


Gráfico 4. *ACF de serie con Logaritmo*

EL ACF de la serie con Logaritmo (Gráfico 4), con el decaimiento lento a cero de sus rezagos, nos muestra que esta serie tiene tendencia. Además, esta serie tiene estacionalidad, se puede apreciar que cada 7 periodos un rezago es más significativo.

Como se evidencia en las gráficas anteriormente analizadas, la serie logaritmo de nuevas vacunaciones no es estacionaria en media, por lo que es necesario diferenciarla en su parte regular. Para esto, se utiliza la prueba *Dickey – Fuller* aumentada para determinar si la serie tiene raíz unitaria.

$$H_0 = Y_t \text{ tiene raíz unitaria (no es estacionario)}$$

Vs.

$$H_1 = Y_t \text{ no tiene raíz unitaria (es estacionario)}$$

```
        Augmented Dickey-Fuller Test

data:  SERIE_LOG
Dickey-Fuller = -2.3554, Lag order = 5, p-value = 0.4283
alternative hypothesis: stationary
```

Con un valor-p de 0.4283, el cual es mucho mayor al nivel de significancia del 0.05, se concluye que no se rechaza la hipótesis nula, y, por tanto, la serie no es estacionaria y debe diferenciarse regularmente.

Se diferencia la serie y nuevamente se realiza la prueba *Dickey – Fuller* para la serie diferenciada. Con esta diferenciación se perderá un dato, por tanto, el total de datos de la serie ya no es 151, si no 150.

```
        Augmented Dickey-Fuller Test

data:  SERIE_diff
Dickey-Fuller = -15.182, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Con un valor-p muy pequeño, el cual es mucho menor al nivel de significancia del 0.05, se concluye que se rechaza la hipótesis nula, y, por tanto, la serie con una diferenciación es estacionaria.

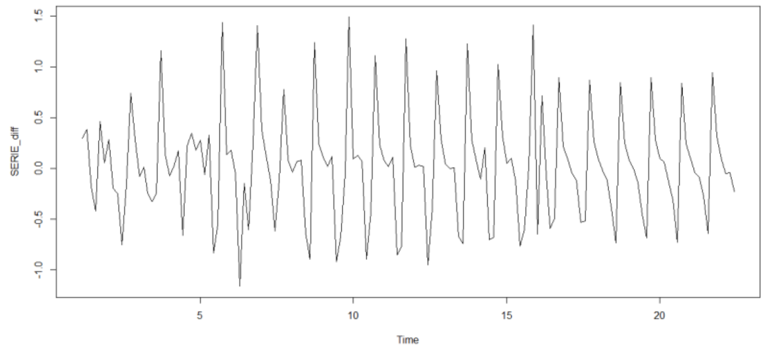


Gráfico 5. *Ln(serie) con una diferenciada*

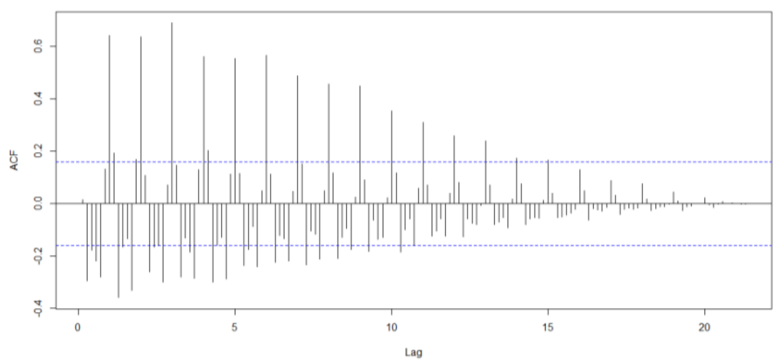


Gráfico 6. ACF de Ln(s*erie) con una diferenciación regular*

Como se evidencia en el gráfico de Ln(serie) con una diferenciada (Gráfico 5), esta serie es estacionaria en media. Pero, el ACF (Gráfico 6) nos muestra que aún existe un componente estacional que debe corregirse.

El resultado de la función "*nsdiffs()"* aplicada a Ln(serie) diferenciada una vez, da "1", por tanto, hay que aplicar una diferenciación estacional a Ln(serie) diferenciada regularmente una vez. Con esta diferenciación estacional, la cantidad de observaciones que se pierden es igual a un periodo estacional, por tanto, el total de observaciones ahora es de 143.

Luego de hacer la diferenciación, se aplica nuevamente la función "*nsdiffs()"* a Ln(serie) diferenciada regularmente una vez, dando como resultado el valor de "0", por tanto, no se diferencia más estacionalmente.

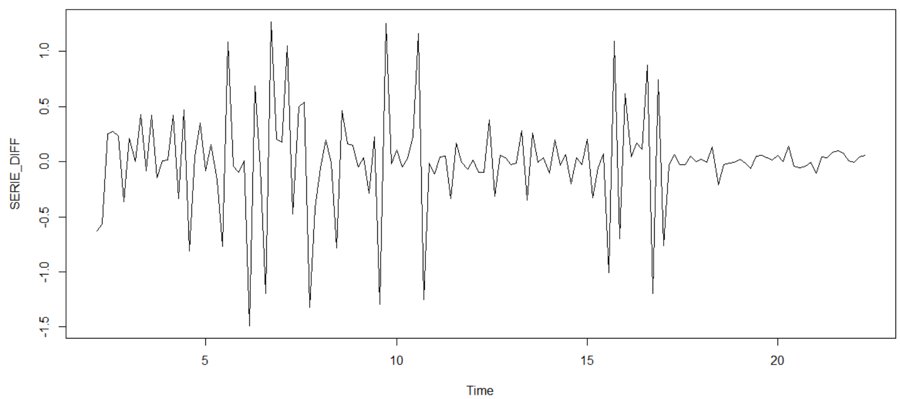


Gráfico 7. Ln(*serie) con una diferenciación regular y una diferenciación estacional*

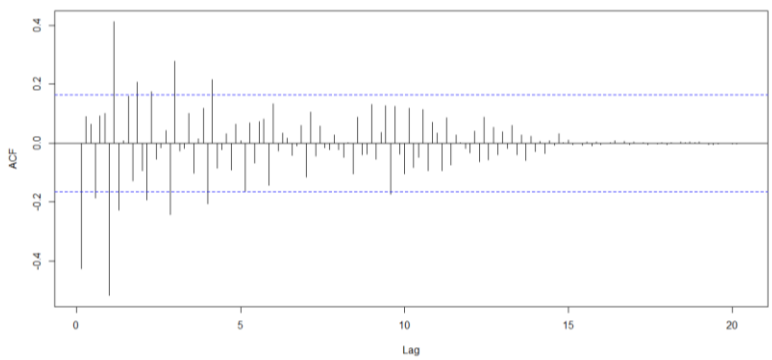


Gráfico 8. ACF de Ln(s*erie) con una diferenciación regular y una diferenciación estacional*

Se puede observar que la serie resultante es una serie estacionaria y sin estacionalidad, a la cual se le puede iniciar a estudiar modelos SARIMA. La serie Ln(serie) con una diferenciación regular y una diferenciación estacional es con la que se trabajará en adelante y a la que se hará referencia cuando se mencione "la serie".

2. Identificación de parámetros

2.1 P y Q

Identificamos modelos ARMA estacionales a la serie mediante el análisis de su ACF y PACF

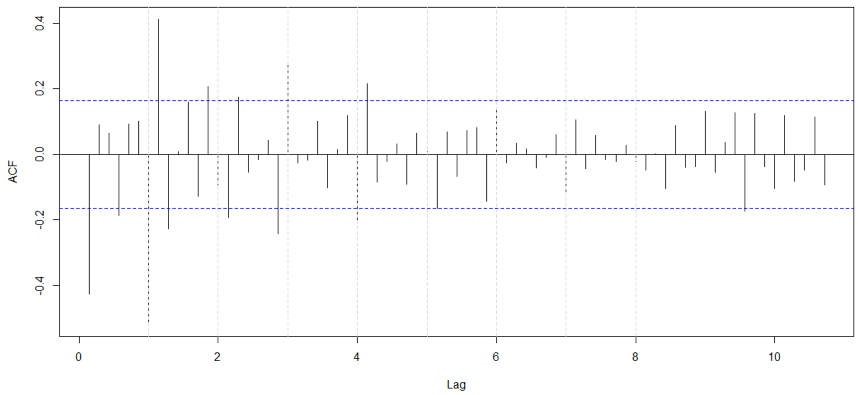


Gráfico 9. ACF estacional de la serie

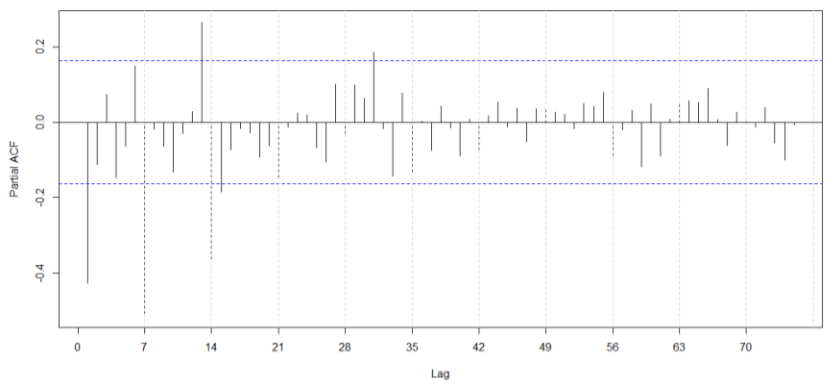


Gráfico 10. PACF estacional de la serie

Identificamos que el ACF corta en el rezago k=1 y el PACF decrece, así se obtiene un modelo ARMA(0,1) o MA(1) para el componente estacional de la serie, es decir P=0 y Q=1.

2.2 p y q

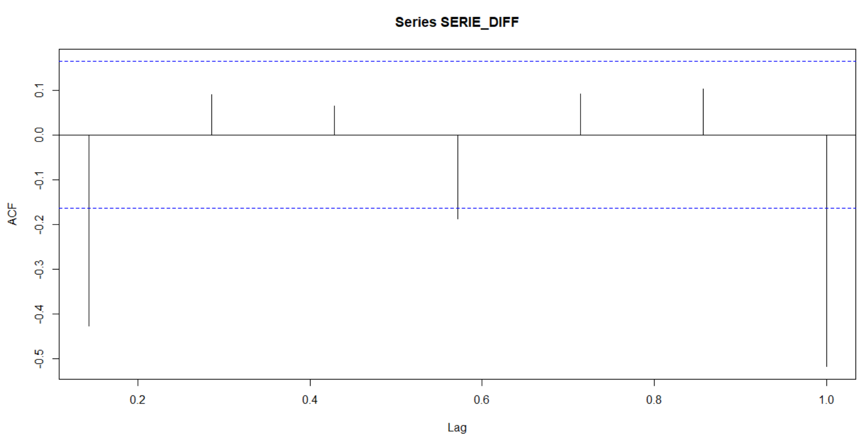


Gráfico 11. ACF de primero periodo estacional de la serie

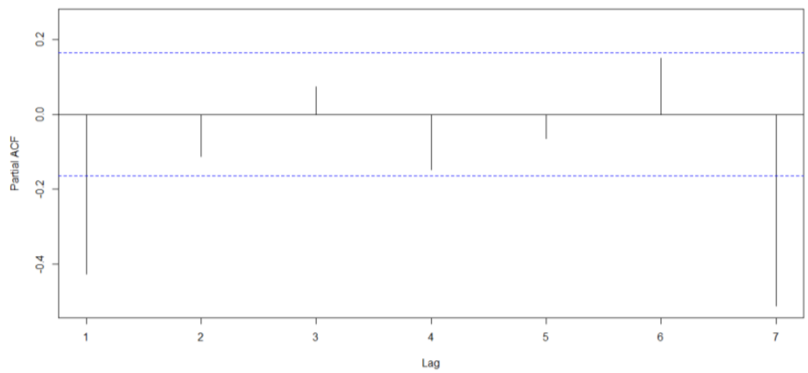


Gráfico 12. PACF de primero periodo estacional de la serie

Se identifican 2 posibles modelos, un ARMA(1,1) ya que se puede observar que, tanto el ACF con el PACF, cortan en el rezago k=1 y por tanto, p=1 y q=1. Pero, también se podría observar que el ACF decae, obteniendo así un modelo ARMA(1,0), con p=1 y q=0.

Modelos identificados:

- Modelo 1 -> ARIMA(1,1,1)(0,1,1)
- Modelo 2 -> ARIMA(1,1,0)(0,1,1)

Usando la función auto.arima() obtenemos el siguiente modelo:

- Modelo 3 -> ARIMA(1,0,2)(0,1,1)

Se iniciará ajustando el modelo 3, ya que es el de orden superior.

3 Ajuste de modelos y significancia de parámetros

3.1 Modelo 3

```
Series: SERIE_TRAIN
ARIMA(1,0,2)(0,1,1)[7]
Box Cox transformation: lambda= 1

Coefficients:
         ar1      ma1     ma2     sma1
      0.7580  -0.3879  0.4020  -0.4186
s.e.  0.0874   0.0977  0.0915   0.0719

sigma^2 estimated as 5062297:  log likelihood=-1314.95
AIC=2639.9   AICc=2640.33   BIC=2654.75

Training set error measures:
                 ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 466.46 2166.453 1521.449 -1.547741 23.97126 0.7617931 -0.07316979
```

Modelo teórico	
$(1-\varphi_1 L)(1-L^4)(1-L)Ln(Y_t) = \mu + \varepsilon_t(1-\theta_1 L - \theta_2 L^2)(1-\theta_{1s} L^4)$	
Parámetros ajustados del modelo	
$\hat{\varphi}_1 = 0.7580$	

$\hat{\theta}_1 = -0.3879$	$\hat{\theta}_2 = 0.4020$
$\hat{\theta}_{1s} = -0.4186$	

Se procede a evaluar la significancia estadística de cada uno de los parámetros del modelo para probar las siguientes hipótesis:

$$H_o: \varphi_i = 0 \; vs \;\; H_1: \varphi_i \neq 0$$

$$H_o: \theta_i = 0 \; vs \;\; H_1: \theta_i \neq 0$$

$$H_o: \theta_{is} = 0 \; vs \;\; H_1: \theta_{is} \neq 0$$

Las hipótesis mencionadas anteriormente se evalúan con en p-valor, si este es menor al nivel de significancia de 0.05, el parámetro es significativo.

P-valores de parámetros del modelo 3

```
         ar1          ma1          ma2         sma1
3.956104e-15 1.076913e-04 2.002000e-05 2.927315e-08
```

Es evidente que los parámetros del modelo 3 son significativo, pues el p-valor para todos es menor al nivel de significancia.

3.2 Modelo 1

```
Series: SERIE_TRAIN
ARIMA(1,1,1)(0,1,1)[7]
Box Cox transformation: lambda= 1

Coefficients:
          ar1      ma1     sma1
      -0.5145  -0.0922  -0.4753
s.e.   0.1242   0.1446   0.0649

sigma^2 estimated as 5230831:  log likelihood=-1308.61
AIC=2625.22   AICc=2625.51   BIC=2637.07

Training set error measures:
                   ME     RMSE     MAE      MPE     MAPE      MASE         ACF1
Training set 85.97442 2202.221 1518.75 -9.74764 25.01223 0.7604415 -0.005464061
```

Modelo teórico	
$(1-\varphi_1 L)(1-L^4)(1-L)Ln(Y_t) = \mu + \varepsilon_t(1-\theta_1 L)(1-\theta_{1S} L^4)$	
Parámetros ajustados del modelo	
$\hat{\varphi}_1 = -0.5145$	
$\hat{\theta}_1 = -0.0922$	
$\hat{\theta}_{1s} = -0.4753$	

Se procede a evaluar la significancia estadística de cada uno de los parámetros del modelo para probar las siguientes hipótesis:

$$H_o: \varphi_i = 0 \; vs \;\; H_1: \varphi_i \neq 0$$

$$H_o: \theta_i = 0 \; vs \;\; H_1: \theta_i \neq 0$$

$$H_o: \theta_{is} = 0 \ vs \ \ H_1: \theta_{is} \neq 0$$

Las hipótesis mencionadas anteriormente se evalúan con en p-valor, si este es menor al nivel de significancia de 0.05, el parámetro es significativo.

P-valores de parámetros del modelo 1

```
         ar1          ma1         sma1
5.448080e-05 5.249303e-01 1.041497e-11
```

Se evidencia que el único parámetro no significativo es el de medias móviles de orden 1, eliminando este parámetro obtenemos como resultado el modelo 2, por tanto, se procede a evaluar la significancia de parámetros para este modelo.

3.3 Modelo 2

```
Series: SERIE_TRAIN
ARIMA(1,1,0)(0,1,1)[7]
Box Cox transformation: lambda= 1

Coefficients:
          ar1     sma1
      -0.5752  -0.4760
s.e.   0.0686   0.0649

sigma^2 estimated as 5208451:  log likelihood=-1308.82
AIC=2623.63   AICc=2623.8   BIC=2632.52

Training set error measures:
                   ME     RMSE      MAE      MPE     MAPE      MASE        ACF1
Training set 80.98622 2205.339 1521.868 -9.41941 25.18067 0.7620031 -0.03716075
```

Modelo teórico
$(1-\varphi_1 L)(1-L^4)(1-L)Ln(Y_t) = \mu + \varepsilon_t(1-\theta_{1s}L^4)$
Parámetros ajustados del modelo
$\hat{\varphi}_1 = -0.5752$
$\hat{\theta}_{1s} = -0.4760$

Se procede a evaluar la significancia estadística de cada uno de los parámetros del modelo para probar las siguientes hipótesis:

$$H_o: \varphi_i = 0 \ vs \ \ H_1: \varphi_i \neq 0$$

$$H_o: \theta_{is} = 0 \ vs \ \ H_1: \theta_{is} \neq 0$$

Las hipótesis mencionadas anteriormente se evalúan con en p-valor, si este es menor al nivel de significancia de 0.05, el parámetro es significativo.

P-valores de parámetros del modelo 1

```
         ar1         sma1
2.083681e-14 9.159412e-12
```

Se evidencia que todos los parámetros del modelo 2 son significativos.

Dado que el *Modelo 3* y el *Modelo 2* cumple con la significancia de todos sus parámetros, hay que hacer una comparación de perdida de información y de los errores de ambos modelos.

Modelo	AIC	BIC	MAPE	MAE
2. SARIMA (1,0,2) (0,1,1)	2623.63	2632.52	25.18067	1521.868
3. SARIMA (1,1,0) (0,1,1)	2639.9	2654.75	23.97126	1521.449

Tabla 1. Comparación de perdida de información y errores de los modelos 2 y 3

De esta anterior tabla no se puede inferir mucho, pues la perdida de información y los errores son muy parecidos en ambos modelos. Por tanto, se deberá validar, para los errores, los supuestos de normalidad, homocedasticidad e independencia para ambos modelos.

4 Validación de supuestos

4.1 Modelo 2

Normalidad

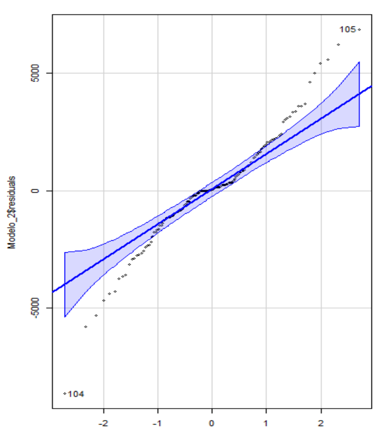


Gráfico 13. qqPlot errores del *Modelo 2*

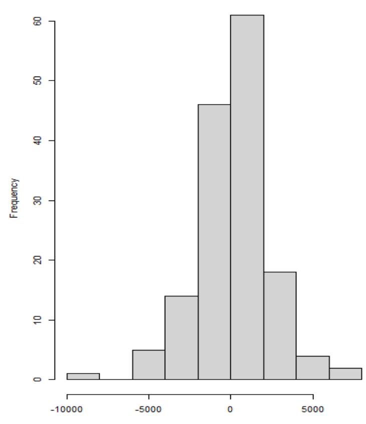


Gráfico 14. Histograma errores del *Modelo 2*

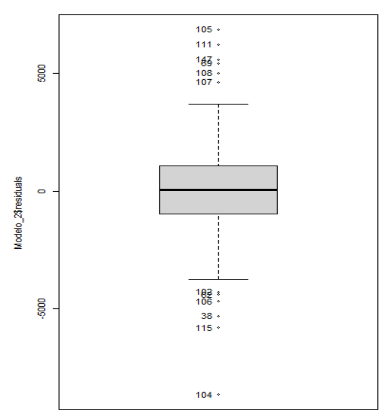


Gráfico 15. Boxplot errores del *Modelo 2*

```
        Shapiro-Wilk normality test

data:  Modelo_2$residuals
W = 0.95957, p-value = 0.0002119
```

Prueba de normalidad Shapiro-Wilk

Hipótesis para prueba de normalidad:

H_0: Los errores proceden de una distribución normal

H_1: Los errores no proceden de una distribución normal

Gráficamente, hay altas evidencias para concluir que los errores no se distribuyen normal, en el qqPlot muchas observaciones se salen de las bandas de confianza y el Boxplot muestra que hay muchos valores atípicos. Para confirmar lo dicho anteriormente, con un valor-p de 0.0002119 obtenido con la prueba de normalidad Shapiro-Wilk y menor al nivel de significancia de 0.05, se rechaza la hipótesis nula, por tanto, se concluye que los errores de la serie no son normales.

Homocedasticidad

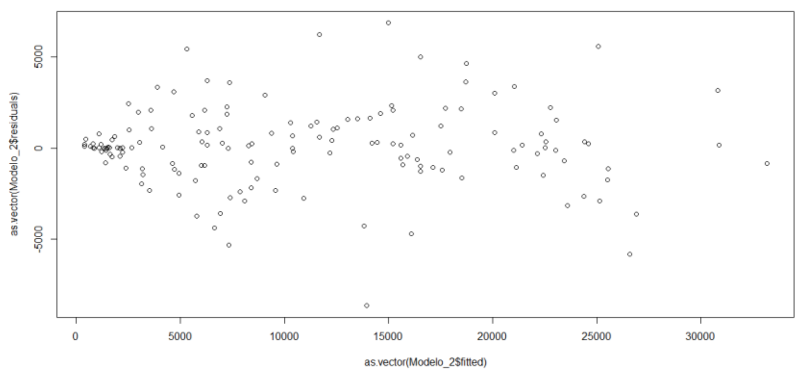


Gráfico 16. Dispersión de errores del *Modelo 2*

Prueba de homogeneidad de varianza:

H_0: Los errores tiene varianza constante

H_1: Los errores no tienen varianza constante

```
        studentized Breusch-Pagan test

data:  Modelo_2$fitted ~ Modelo_2$residuals
BP = 0.49771, df = 1, p-value = 0.4805
```

Prueba para varianza Breusch-Pagan

Como se evidencia en la el Gráfico 16, los errores tienen buena dispersión en todo el gráfico, aunque al principio del gráfico (de izquierda a derecha) se evidencia una baja dispersión, la prueba de varianza Breusch-Pagan nos confirma, con un p-valor mayor al nivel de significancia del 0.05, que los errores tienen varianza constante.

Independencia

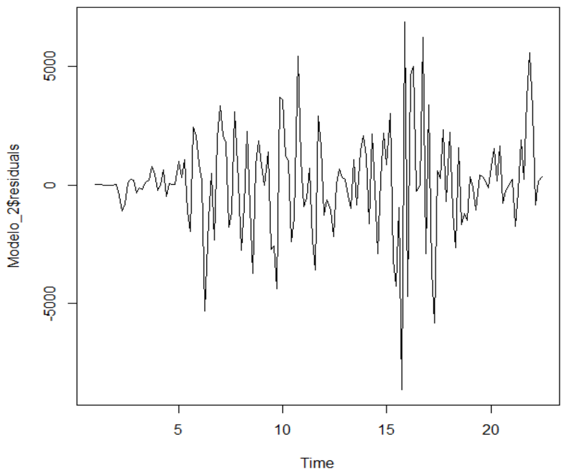


Gráfico 17. Serie de los errores del *Modelo 2*

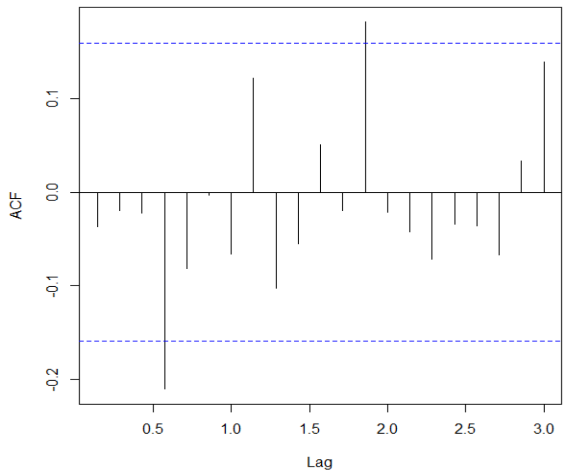


Gráfico 18. ACF de los errores del *Modelo 2*

```
            Box-Pierce test

data:  Modelo_2$residuals
X-squared = 0.20852, df = 1, p-value = 0.6479
```

Prueba Box-Pierce para independencia

En el Gráfico 17 se puede evidenciar que la serie de los errores del modelo 2 no tiene una tendencia, su ACF muestra que la mayoría de los errores se encuentran dentro de las bandas de confianza y, además, la prueba para independencia Box-Pierce nos da un p-valor mayor al nivel de significancia, por tanto, los errores del modelo 2 cumplen con el supuesto de independencia.

4.2 Modelo 3

Normalidad

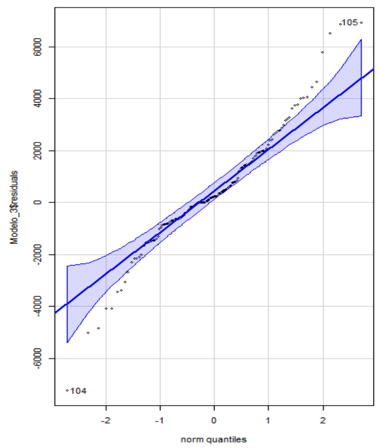


Gráfico 19. qqPlot errores del *Modelo 3*

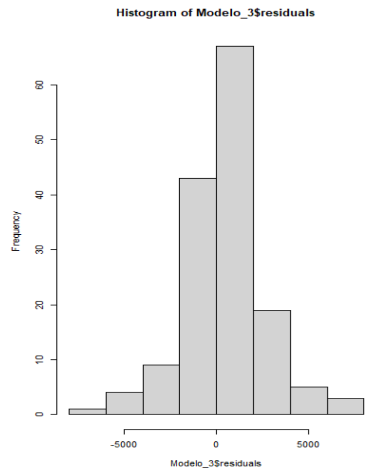


Gráfico 20. Histograma errores del *Modelo 3*

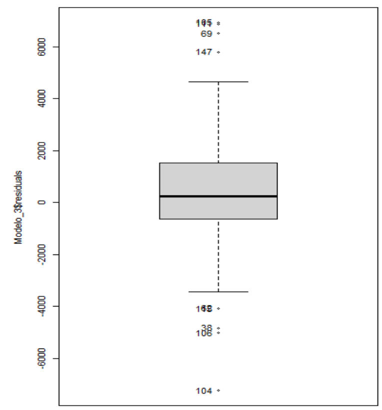


Gráfico 21. Boxplot errores del *Modelo 3*

```
        Shapiro-Wilk normality test

data:  Modelo_3$residuals
W = 0.9614, p-value = 0.0003139
```

Prueba de normalidad Shapiro-Wilk

Hipótesis para prueba de normalidad:

H_0: Los errores proceden de una distribución normal

H_1: Los errores no proceden de una distribución normal

Gráficamente, hay altas evidencias para concluir que los errores no se distribuyen normal, en el qqPlot muchas observaciones se salen de las bandas de confianza y el Boxplot muestra que hay muchos valores atípicos. Para confirmar lo dicho anteriormente, con un valor-p de 0.0003139 obtenido con la prueba de normalidad Shapiro-Wilk y menor al nivel de significancia de 0.05, se rechaza la hipótesis nula, por tanto, se concluye que los errores de la serie no son normales.

Homocedasticidad

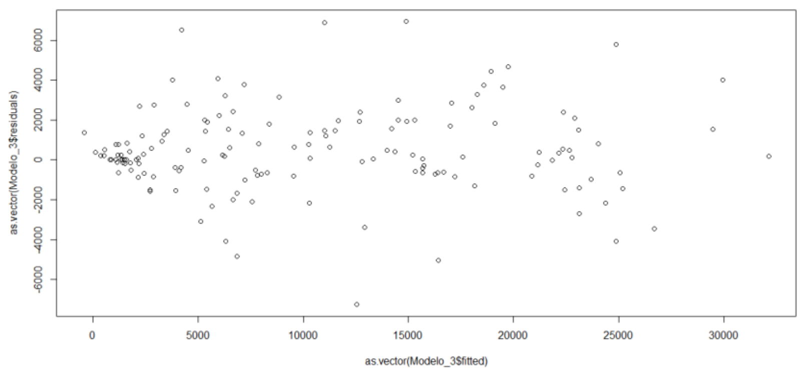


Gráfico 22. Dispersión de errores del *Modelo 3*

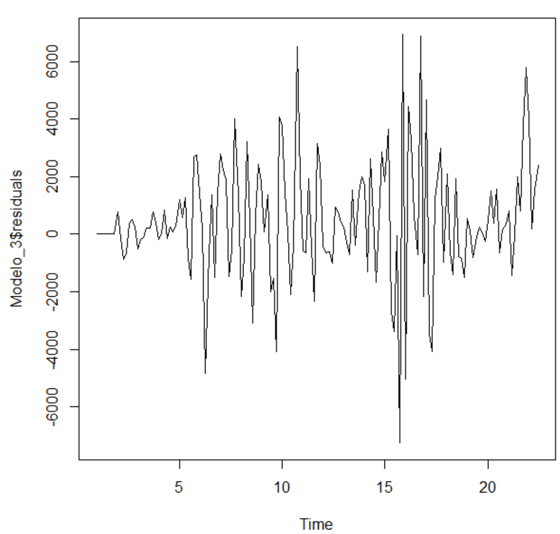
```
        studentized Breusch-Pagan test

data:  Modelo_3$fitted ~ Modelo_3$residuals
BP = 0.38835, df = 1, p-value = 0.5332
```

Prueba para varianza Breusch-Pagan

Como se evidencia en la el Gráfico 22, los errores tienen buena dispersión en todo el gráfico, aunque al principio del gráfico (de izquierda a derecha) se evidencia una baja dispersión, la prueba de varianza Breusch-Pagan nos confirma, con un p-valor mayor al nivel de significancia del 0.05, que los errores tienen varianza constante.

Independencia



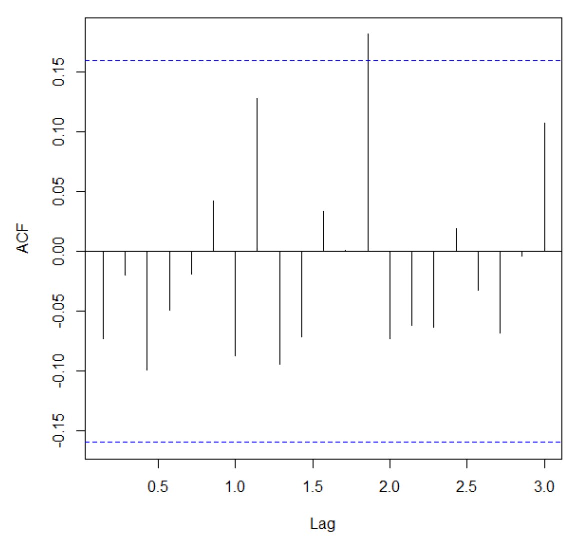


Gráfico 23. Serie de los errores del *Modelo 3* **Gráfico 24.** ACF de los errores del *Modelo 3*

```
        Box-Pierce test

data:  Modelo_2$residuals
X-squared = 0.20852, df = 1, p-value = 0.6479
```

Prueba Box-Pierce para independencia

En el Gráfico 23 se puede evidenciar que la serie de los errores del modelo 3 no tiene una tendencia, su ACF muestra que la mayoría de los errores se encuentran dentro de las bandas de confianza y, además, la prueba para independencia Box-Pierce nos da un p-valor mayor al nivel de significancia, por tanto, los errores del modelo 3 cumplen también con el supuesto de independencia.

PRONOSTICOS

Aunque lo recomendado es predecir solo un periodo estacional, es decir los próximos 7 rezagos, la validación cruzada nos dejó 17 datos para predecir y quisimos compararlos para un detalle mas completo de los pronósticos.

ARIMA (1,1,0) (2,1,1)

El modelo (1,1,0) (2,1,1) dio el MAPE mas alto de los dos modelos seleccionados, por lo que debería tener el pronóstico menos exacto, en la grafica es la serie de color rojo. Se hizo la comparación de toda la serie en una grafica y en otra se hizo únicamente la comparación de los siguientes 17 periodos (datos de testeo)

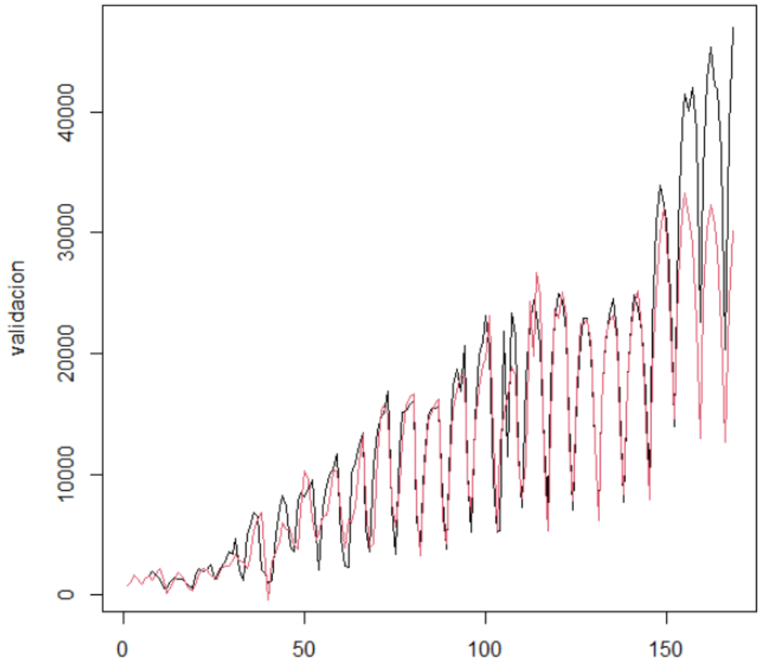


Gráfico 25. *Comparación de la serie original vs pronósticos del modelo 3*

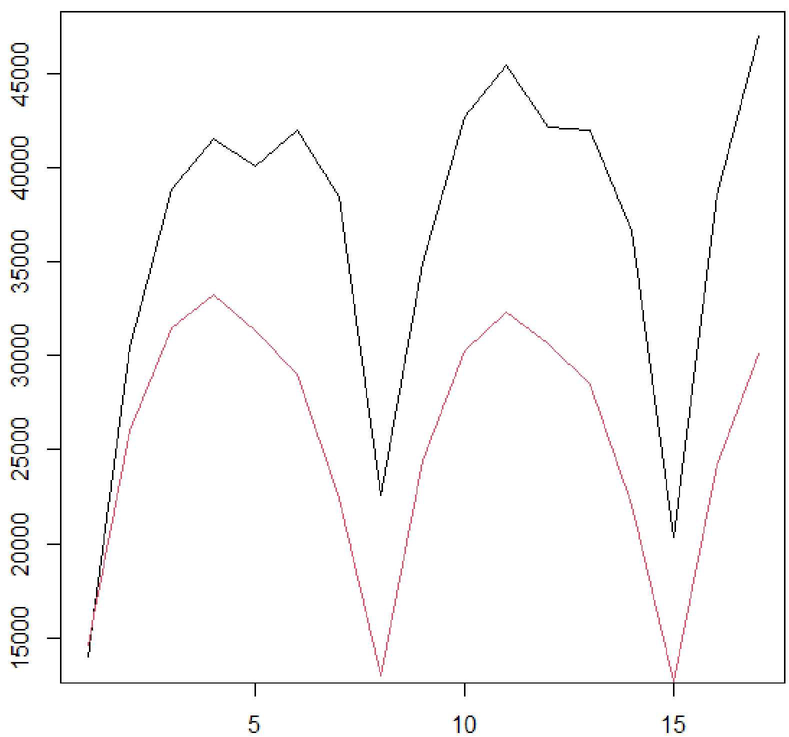


Gráfico 26. *Comparación de la serie original vs pronósticos del modelo 3 en los siguientes 17 periodos*

Conclusiones: Los pronósticos tienen un patrón de comportamiento similar a la serie original, capaz de seguir picos y bajones de la serie.

	Point Forecast	Lo 95	Hi 95
22.57143	14630.96	10221.025	19040.90
22.71429	26021.10	21318.609	30723.60
22.85714	31521.42	25937.378	37105.46
23.00000	33249.90	27217.322	39282.49
23.14286	31365.23	25089.399	37641.06
23.28571	29016.79	22605.365	35428.22
23.42857	22464.35	15976.289	28952.40
23.57143	13013.25	5726.058	20300.45
23.71429	24391.43	16947.256	31835.60
23.85714	30286.08	22528.141	38044.02
24.00000	32313.49	24380.865	40246.11
24.14286	30655.40	22624.125	38686.68
24.28571	28478.72	20391.306	36566.14
24.42857	22056.48	13936.976	30175.98
24.57143	12704.08	4009.048	21399.11
24.71429	24157.06	15355.740	32958.39
24.85714	30108.43	21068.090	39148.77

Tabla 1. *Valores pronosticados modelo 3*

#ARIMA(1,1,0)(0,1,1)

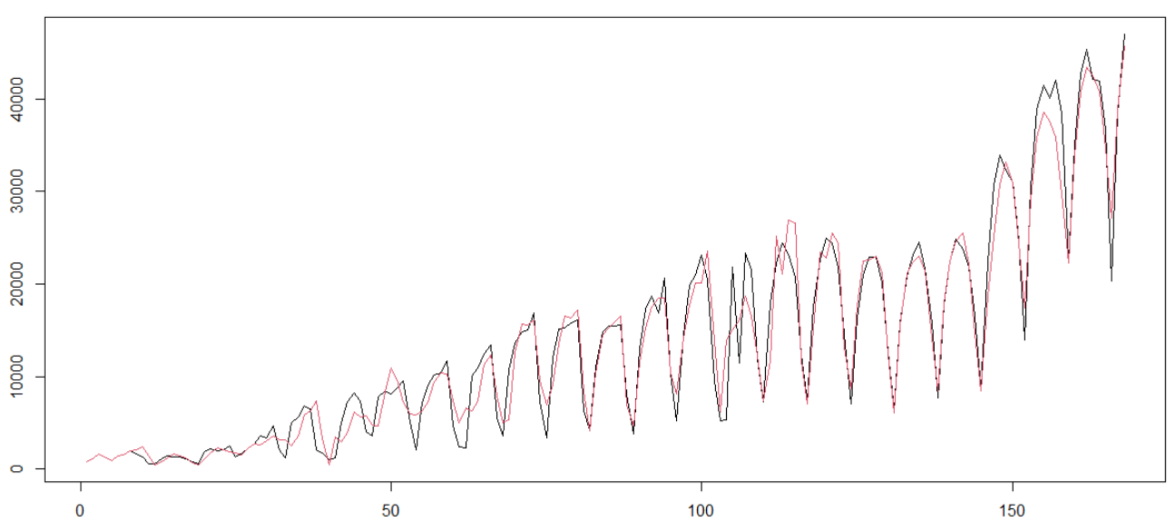


Gráfico 27. *Comparación de la serie original vs pronósticos del modelo 2.*

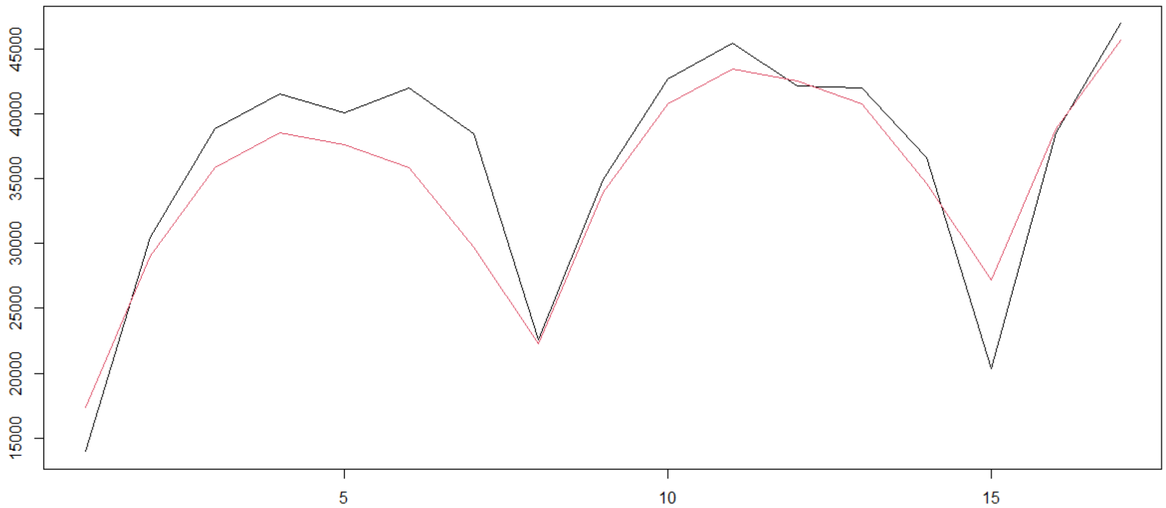


Gráfico 28. *Comparación de la serie original vs pronósticos del modelo 2 en los siguientes 17 periodos*

	Point Forecast	Lo 95	Hi 95
22.57143	17386.58	12913.38	21859.78
22.71429	29051.47	24191.26	33911.68
22.85714	35874.33	29954.28	41794.38
23.00000	38554.43	32116.74	44992.12
23.14286	37651.70	30541.43	44761.96
23.28571	35847.02	28228.11	43465.92
23.42857	29729.56	21577.75	37881.37
23.57143	22266.19	12623.68	31908.69
23.71429	33975.91	23591.15	44360.67
23.85714	40772.99	29414.94	52131.03
24.00000	43467.91	31367.90	55567.92
24.14286	42556.65	29675.23	55438.08
24.28571	40756.88	27184.10	54329.66
24.42857	34636.60	20381.32	48891.88
24.57143	27174.85	11442.39	42907.31
24.71429	38883.64	22271.46	55495.82
24.85714	45681.25	27986.51	63375.99

Tabla 2. *Valores pronosticados modelo 2*

Conclusión: como era de esperarse gracias a MAPE, el modelo 2 tiene más exactitud en el pronostico de datos, es un pronostico muy similar a la serie original, capaz de seguir el patrón de la serie e incluso interceptarla en ciertos puntos, por ende el modelo que mejor se acomoda seria el modelo numero 2.