Trabajo Final - Series de Tiempo y pronóstico

Andrés Felipe Palomino - David Stiven Rojas
Códigos: 1922297 - 1924615
Universidad del Valle

26 de julio de 2023

1. Introducción

El presente trabajo tiene como finalidad desarrollar y aplicar los conceptos tratados en el curso de series de tiempo, para ello se trabajara con una serie de tiempo, la cual es una sucesión de datos medidos en un determinado tiempo y ordenados cronológicamente. Con base en la teoría estudiada, y la metodología planteada por Box y Jenkins* (1976), se procederá a modelar la serie de tiempo y realizar los respectivos pronósticos planteados en los objetivos.

1.1. Base de Datos

Stack Overflow es un sitio de preguntas y respuestas para programadores profesionales y aficionados, en el sitio se abordan una amplia gamma de temas de programación, como lo son las librerías que utilizan distintos software de programación. La serie temporal a trabajar, consiste en el conteo mensual de preguntas que se realizan en la web sobre la librería PyGTK, desde el año 2009 hasta el año 2019. PyGTK es un módulo para crear interfaces gráficas de usuario en Python, por lo que permite el desarrollo de aplicaciones, juegos, entre otros. Sea $\{Y_t\}$ el proceso que describe el conteo mensual de preguntas que se realizan en la web sobre la librería PyGTK.

1.2. Problema de investigación

Una compañía de programación de juegos está interesada en realizar capacitaciones sobre programas que ayuden en la creación e interfaces de juegos. Por ello, desean evaluar y observar que librerías podrían presentar una alta demanda en el mundo informático. Específicamente se hará un proceso de seguimiento sobre la librería PyGTK por lo que es necesario realizar predicciones sobre la misma para evaluar la capacidad de inversión que pueda realizarse.

1.3. Objetivos del estudio

- Identificar, estimar y validar un modelo adecuado para la serie de tiempo.
- Pronosticar el número de preguntas que se realizaran sobre la librería PyGTK para el mes enero y febrero del año 2020.

2. Formulación general de una clase de modelos

Se considera un modelo general Autorregresivo Integrado de Medias Móviles (ARIMA):

$$(1 - \phi_1 B ... - \phi_p B^p)(1 - B)^d Z_t = \theta_0 + (1 - \theta_1 B - .... - \theta_q B^q) a_t$$

3. Identificación preliminar del modelo

Como primer paso para tener una idea preliminar del modelo a tratar, se procederá a graficar la serie de tiempo.

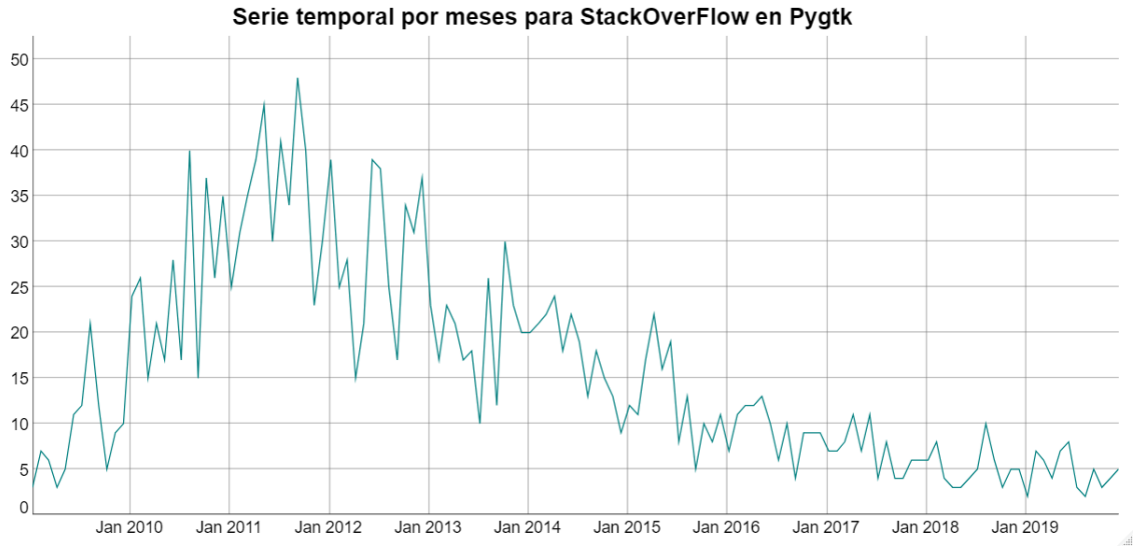

Figura 1: Comportamiento de la serie de tiempo

Como se observa en la Figura 1, se evidencia un proceso no estacionario tanto en media como en varianza, es decir, $E(Z_t) = \mu_t$, y $Var(Z_t) \neq Var(Z_{t-k})$ Para $k \neq 0$. Por ende, como primer paso, se debe encontrar una transformación que estabilicé la varianza, específicamente se usara la transformación en potencia introducida por Box y Cox (1964).

Valores de λ	Transformaciones
-1.0	$1/Z_t$
-0.5	$1/\sqrt{Z_t}$
0.0	$\ln Z_t$
0.5	$\sqrt{Z_t}$
1.0	Z_t

*Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations, *J. Roy. Stat. Soc.* Ser. B. 26, 211-252.

Tabla 1: Valores de lambda con sus respectivas transformaciones.

Realizando la estimación del λ con la función boxcox de la librería MASS, se obtiene el siguiente gráfico:

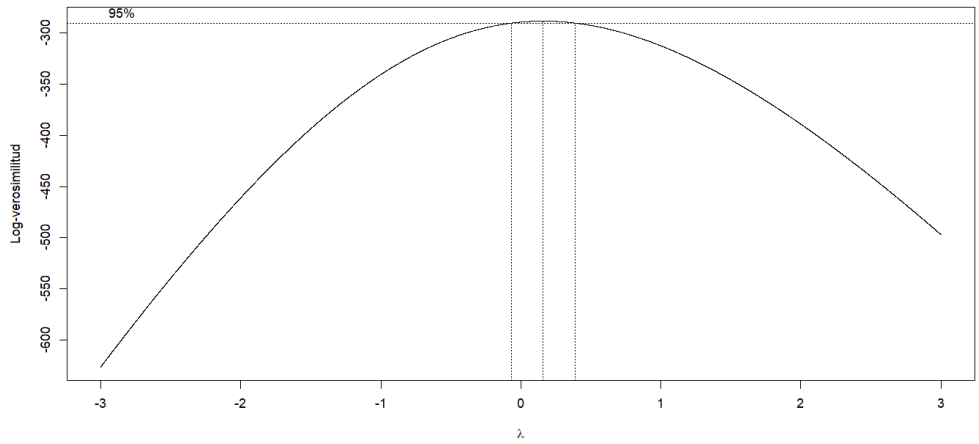

Figura 2: Perfiles de verosimilitud para λ

En la figura 2 se observa el valor λ que minimiza el error cuadrático medio residual, el cual, corresponde al valor máximo de la Figura 2 ($\lambda = 0{,}1$). Con base en la tabla 1, se procede a realizar la transformación LnY_t

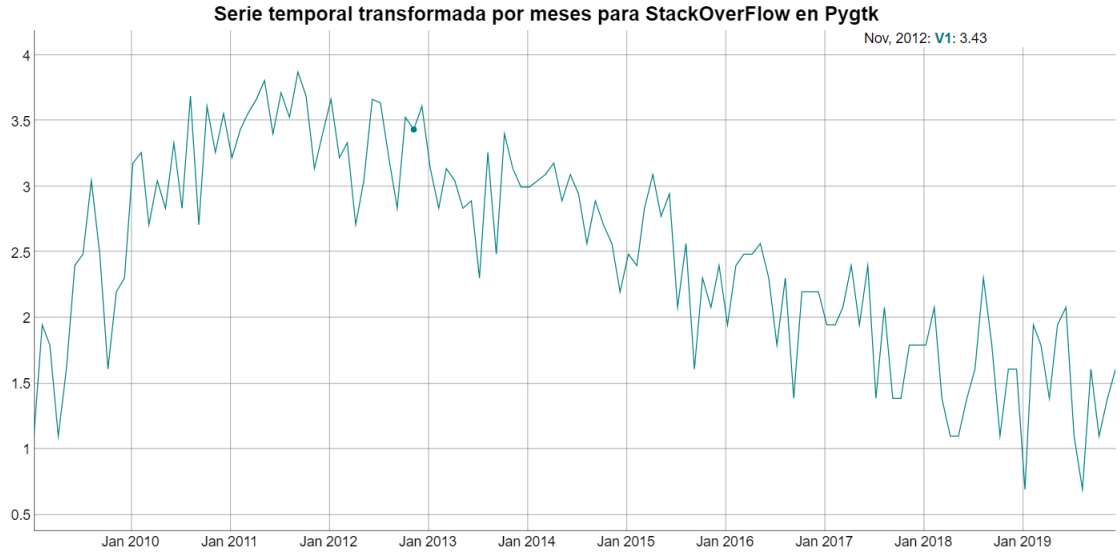

Figura 3: Comportamiento de la serie de tiempo transformada.

Como se observa en la figura 3, el conjunto de datos presenta una serie constante en varianza (Z_t), pero no constante en media, por lo que es necesario diferenciar la serie. En la siguiente figura se observará la serie diferenciada, un grado.

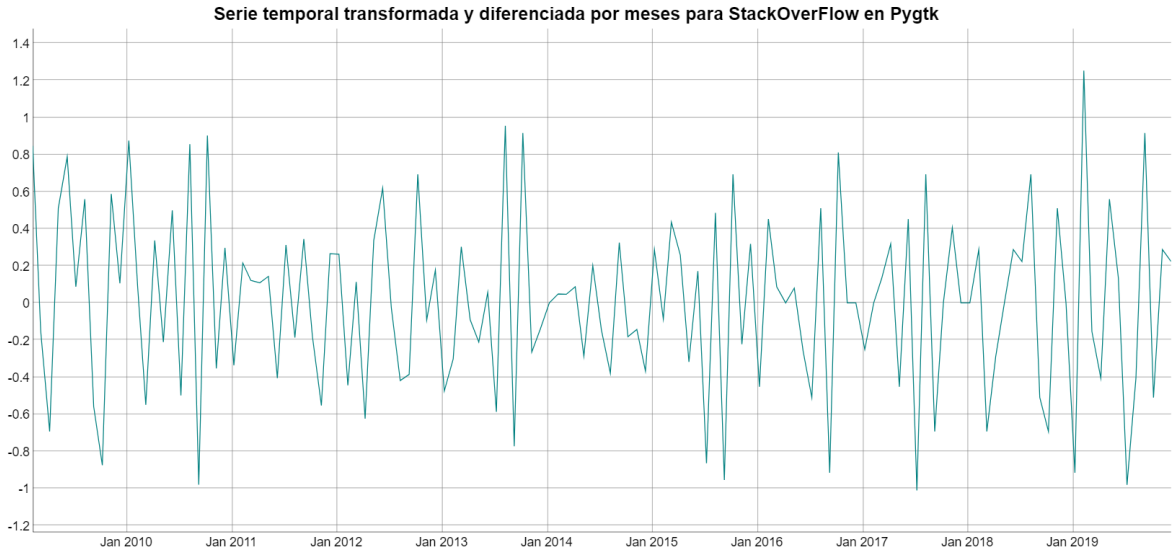

Figura 4: Comportamiento de la serie de tiempo transformada y diferenciada un grado.

En la figura 4 se observa el comportamiento de la serie una vez realizado tanto el proceso de estabilización en varianza como en media. Para rectificar que se está bajo una serie de tiempo estacionaria se procede a realizar la prueba de Dickey Fuller, la cual busca determinar la existencia o no de raíces unitarias en una serie de tiempo.
H_0: Existe una raíz unitaria en la serie
H_1: No existe ninguna raíz unitaria en la serie.

Al realizar la prueba se obtuvo un valor P menor a 0.1, por lo que con un nivel de significancia del 5 % se concluye que la serie es estacionaria, ya que, no posee raíces unitarias.

La serie se podría ajustar bajo un modelo ARIMA (p,1,q), ya que fue necesario de un proceso de diferenciación para tener un proceso estacionario. Para determinar el orden del polinomio autorregresivo (p) y del polinomio de medias móviles (q) se realiza el análisis de la función de autocorrelación (ACF) y la función de autocorrelación parcial (PACF) presentados en la Figura 5.

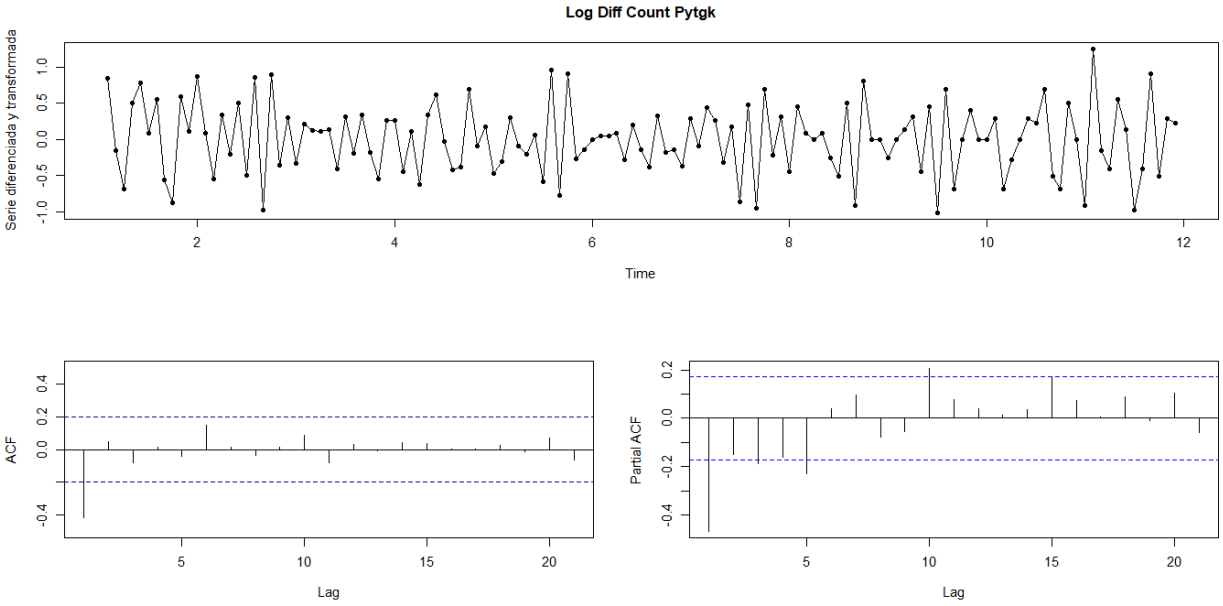

Figura 5: Función de autocorrelación y autocorrelación parcial muestral de la serie de tiempo estacionaria homogénea.

En la Figura 5 se observa que la ACF muestral se anula a partir del rezago k=2, además la PACF presenta un patrón de decrecimiento. Por tanto, al realizar la comparación de este proceso con el ACF y PACF teórico, se asume que el proceso generador de la serie es un ARIMA(0,1,1), que también se le denomina como un IMA(1,1). También cabe destacar que el componente $\theta_0 = 0$, ya que no es necesario de un componente determinístico. Entonces el modelo a estimar sería el siguiente:

$$(1 - B)Z_t = (1 - \theta B)a_t$$

Donde Z_t es él $log(Y_t)$

4. Estimación de parámetros del modelo preliminar

Para estimar el parámetro del modelo θ, se utiliza la suma condicional de cuadrados para encontrar los valores iniciales, y luego la máxima verosimilitud exacta. La estimación del parámetro es la siguiente:

$$\hat{\theta} = 0{,}66523 \qquad \hat{\sigma}_a^2 = 0{,}168$$

Donde $\hat{\sigma}_a^2$ es la varianza del componente ruido blanco a_t y el error estándar de la estimación de θ es 0.0646. Entonces el modelo preliminar propuesto para la serie es:

$$(1 - B)Z_t = (1 - 0{,}66523B)a_t$$

con $a_t \sim$N(0, 0.168)

5. Diagnóstico del modelo

Por último, se procede a validar el modelo estimado, el cual debe ser consistente con los supuestos teóricos sobre los que se basa.

5.1. Verificación de invertibilidad del modelo estimado

Debido a que el modelo ajustado es un IMA(1,1), solamente se debe verificar si el proceso es invertible, ya que $1 + \theta_1^2 < \infty$ y por ende un proceso de medias móviles finito de orden 1 es siempre estacionario.
El proceso es invertible si las raíces de $\theta(B) = 0$ caen por fuera del círculo unitario.

$$(1 - \hat{\theta}B) = 0$$

$$(1 - 0{,}66523B) = 0$$

$$|B| = |\frac{1}{0{,}66523}| > 1$$

Por lo tanto, el proceso es invertible, adicionalmente se observa gráficamente en la Figura 6.

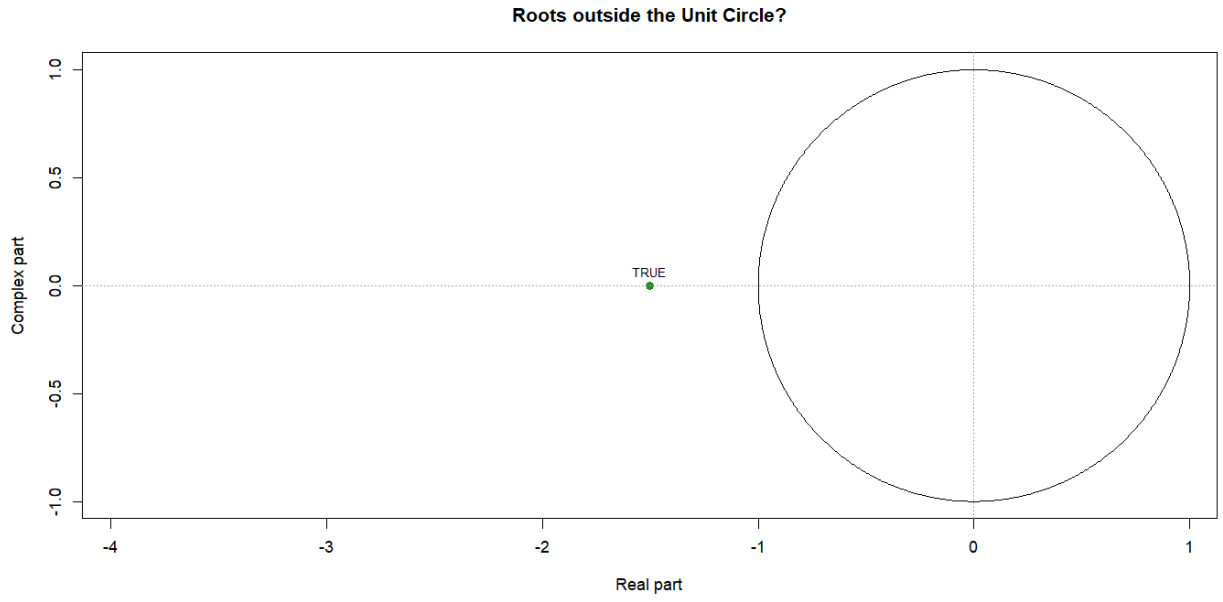

Figura 6: Raíz del polinomio IMA(1,1)

5.2. Verificación de residuales generados por un proceso ruido blanco a_t

Para evaluar si los residuales son generados por un proceso de ruido blanco de media cero, no correlacionados y varianza constante, se procede a graficar los residuales contra el tiempo, el ACF, PACF y los valores P, obtenidos secuencialmente para el test de Ljung-Box, además se presenta un gráfico Cuantil-Cuantil para evaluar la normalidad del ruido blanco a_t.

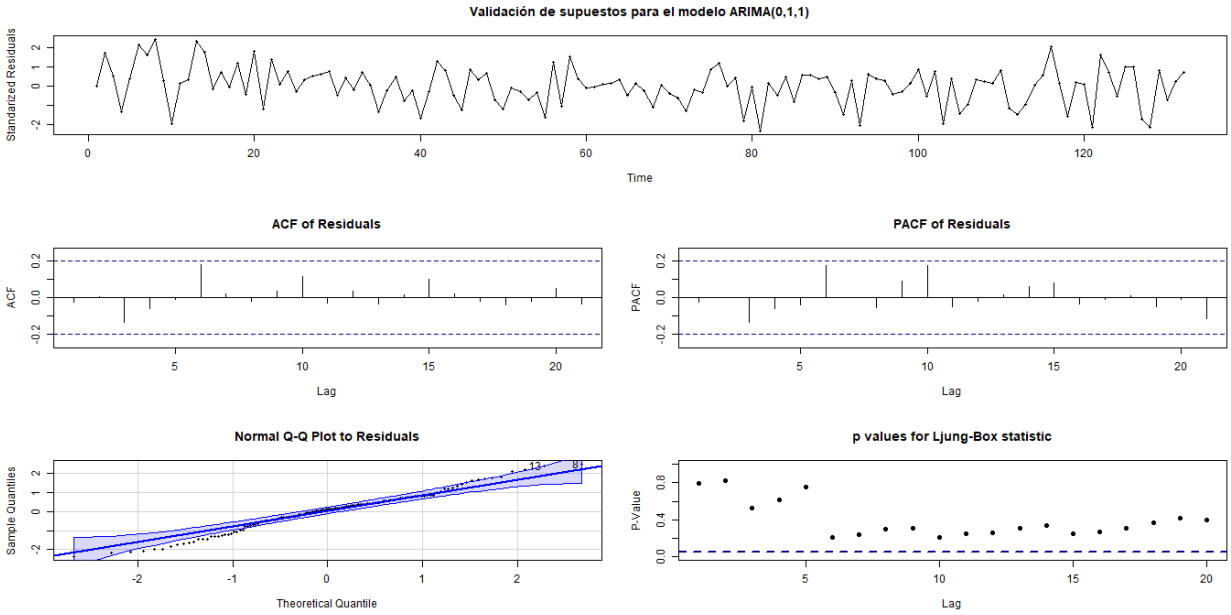

Figura 7: Comportamiento de los residuos: (A) Residuales contra el tiempo (B) ACF de los residuos (C) PACF de los residuos (D) Gráfico cuantil-cuantil normal (E) Valores-P para la prueba Ljung-Box

En la Figura 7 (A) se observa que los residuales no presentan algún patrón sistemático a abrirse o comprimirse y tampoco se evidencian puntos atípicos, ya que no superan dos desviaciones estándar, además estos oscilan alrededor de cero, lo que podría indicar que los residuales provienen de un proceso de media y varianza constante ya que se realizó la prueba de White (Busca determinar si los errores del modelo presentan varianza constante) donde se obtuvo un valor p de 0.8313, por lo cual se presenta homoscedasticidad en los residuos. Además, según la Figura 7 (B) y (C) estos residuales parecen provenir de un proceso de ruido blanco de media cero. Por último se observa la Figura 7 (E) la cual muestra los valores P del estadístico de Ljung-Box (prueba la hipótesis nula de que los coeficientes de autocorrelación hasta un desfase k son iguales a cero), en ella se observa el no rechazo de la hipótesis nula a un nivel de significancia $\alpha = 0{,}5$, por ende los primeros 20 coeficientes de correlación son conjuntamente cero.

5.3. Verificación de normalidad del ruido blanco a_t

Como se observa en la Figura 7 (D) las observaciones se encuentran dentro de las bandas de confianza, exceptuando en los cuantiles inferiores. Para tener un mayor soporte de que el proceso ruido blanco se distribuye normal se utilizara la prueba de Jarque-Bera.
H_0: Los residuos se ajustan una distribución normal
H_1: Los residuos no se ajustan a una distribución normal

```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 0.3747
  P VALUE:
    Asymptotic p Value: 0.8291
```

Se evidencia que la prueba arrojo valor-p mucho mayor al nivel de significancia del 5 %, por lo tanto, no hay evidencia suficiente para rechazar el supuesto de normalidad.

5.4. Significancia individual del parámetro estimado

Por último, se realiza una prueba t para evaluar la significancia del parámetro en el modelo, el estadístico se construye con la estimación dividida sobre el error estándar.

$H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$

```
z test of coefficients:

     Estimate Std. Error z value  Pr(>|z|)
ma1 -0.665299   0.064592   -10.3 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se evidencia que la prueba arrojo un valor-p menor al nivel de significancia del 5 %, por lo tanto, hay evidencia suficiente para rechazar la hipótesis nula y, por lo tanto, θ es significativamente distinto de cero

6. pronóstico y ajuste del modelo:

Se procede a graficar la serie transformada vs los valores ajustados con el modelo estimado para evaluar si la estimación del modelo es consistente con la realidad.

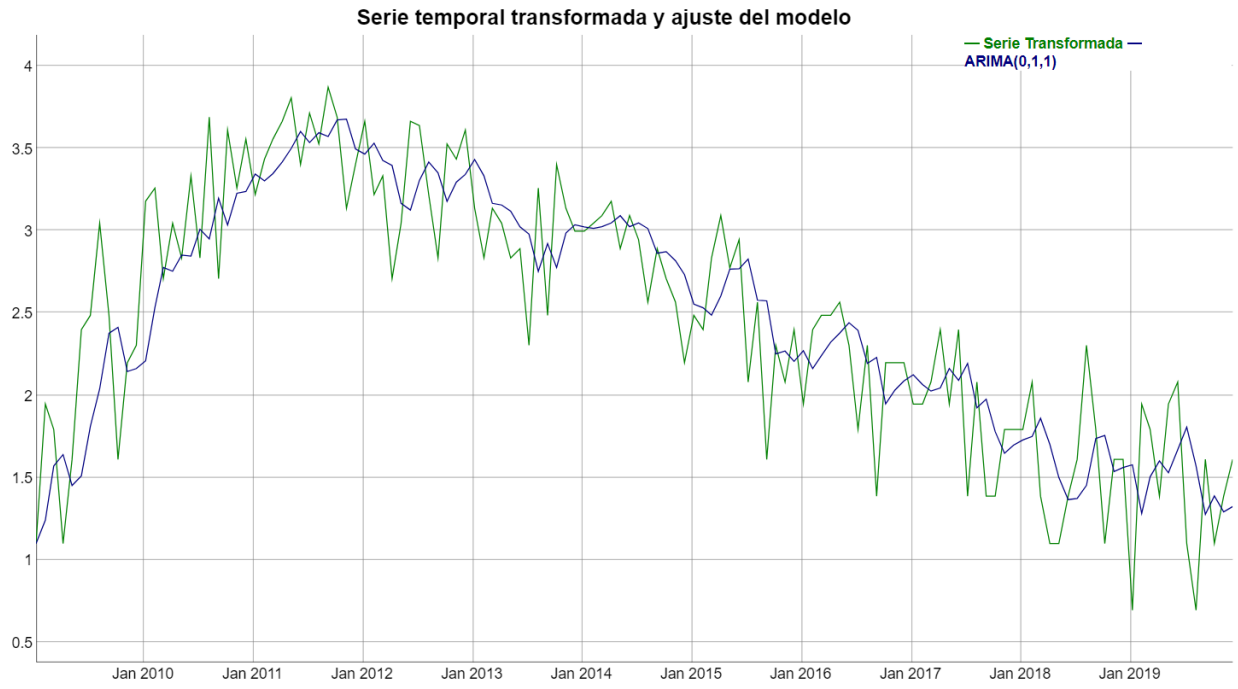

Figura 8: Comportamiento de la serie vs Valores ajustados por el modelo

Como se observa gráficamente en la Figura 8, el modelo presenta un buen ajuste, los valores ajustados por el modelo se asemejan a la serie temporal real y siguen su misma trayectoria, además se calculó el RMSE el cual dio 0.4067875, lo que indica que el modelo posee una buena capacidad predictiva, ya que dio un valor bajo.
Por último, para cumplir con los objetivos propuestos, se realizará los pronósticos para el mes de enero y febrero del año 2020 con base en la información obtenida por la serie temporal.

pronóstico:
Se tiene el modelo IMA(1,1):

$$(1-B)Z_t = (1-0{,}66523B)a_t$$

Donde $\theta = 0{,}66523$, $\sigma_a^2 = 0{,}168$, además, se cuenta con las observaciones $Z_{130} = 1{,}1$, $Z_{131} = 1{,}39$, $Z_{132} = 1{,}61$ y se quiere pronosticar Z_{133}, Z_{134} junto con sus intervalos de predicción del 95 %.

Despejando Z_t de un modelo IMA(1,1) general se tiene:

$$Z_t = Z_{t-1} + a_t - \theta_1 a_{t-1}$$

Reemplazando t por $n + l$ y sacando el valor esperado condicional se tiene:

$$\hat{Z}_n(l) = \hat{Z}_n(l-1) + \hat{a}_n(l) - \theta_1 \hat{a}_n(l-1)$$

Por lo que la forma general del pronóstico para $l \geq 1$ es:

$$\hat{Z}_n(l) = Z_n - \theta_1 a_n$$

Reemplazando:

$$\hat{Z}_{133} = 1{,}61 - 0{,}66523 a_n$$
$$\hat{Z}_{134} = 1{,}61 - 0{,}66523 a_n$$

donde $a_n = Z_n - \hat{Z}_n$ y por último sé realiza $exp(\hat{Z}_n(l))$ para obtener él $\hat{Y}_n(l)$ ya que se realizó la transformación logarítmica para estabilizar la varianza.

Límites de predicción:

Para obtener los limites de predicción, se debe calcular las ponderaciones ψ_j de manera recursiva, donde $\psi_j = \sum_{i=0}^{j-1} \pi_{j-1}\psi_i$
Y los coeficientes π_j se obtienen escribiendo el modelo en forma AR.

$$(1 - \theta_1 B) * (1 - \pi_1 B - \pi_2 B^2 - \pi_3 B^3 - .....) = (1 - B)$$

Solucionando se obtiene que:

$$\psi_j = (1 - \theta_1)$$

para $1 \leq j \leq l - 1$

Por lo tanto, los límites de predicción del 95 %son:

$$\text{Para } Z_{133} \text{ son } Z_{133} \pm 1{,}96\sqrt{\sigma_a^2}$$

$$\text{Para } Z_{134} \text{ son } Z_{134} \pm 1{,}96\sqrt{\sigma_a^2(1 + (1 - \theta_1)^2)}$$

De igual manera se realiza la transformación inversa en los límites de predicción para obtener los límites de Y_t.
Finalmente, los resultados para las predicciones e intervalos se consignan Tabla 2.

Tiempo	**Lim. inferior**	**pronóstico**	**Lim. superior**
Enero 2020	1.850277	4.131885	9.226983
Febrero 2020	1.770973	4.131885	9.640164

Tabla 2: Predicciones y límites de predicción a un nivel de confianza del 95 %

7. Conclusión

Como se observa en los resultados de los pronósticos, la función de predicción es una línea horizontal, que depende de la información que se tenga sobre el último dato de la serie (Z_n) y del residuo dado por el ajuste y el valor real (a_n), además los límites de predicción van aumentando a medida que se pronostique más hacia el futuro. Por ende, no es recomendable realizar predicciones a largo plazo.
En términos de los pronósticos, se evidencia que para el mes enero y febrero del siguiente año, no se presentara un aumento en la cantidad de preguntas que se realizaran sobre la librería PyGTK en comparación a los meses anteriores. Por lo que se recomienda no realizar una inversión significativa en programas de capacitación.

8. Bibliografía

Cryer, J. D., Chan, K.-S. (2008). Time series analysis: with applications in R (Vol. 2). Springer.

Jenkins, G. M., Box, G. E. P. (1976). Time Series Analysis Forecasting and Control (Vol. 2). Holden-Day, San Francisco.