Modelo Estocástico y Pronóstico de la Demanda de Pasajeros del Metro de la CDMX

Chávez Romero Luis Gerardo
Escalante López Julio César
García Flores Luis Edgar
Rodríguez Mondragón Jessica Fernanda
Sánchez López Katya Pamela , *Universidad Nacional Autónoma de México*

22 de Mayo del 2020

Resumen. Por medio de técnicas de modelación estocástica de series de tiempo, se estimará un modelo adoc al comportamiento de los datos que arrojan la cantidad de usuarios del Metro de la CDMX, ahondando en las crecientes capacidades a las que ha manejado a lo largo de su historia como medio de trasporte público.

1. Introducción

El Sistema de Transporte Colectivo Metro (STC Metro) de la Ciudad de México es un transporte que pertenece al Organismo Público Descentralizado Sistema de Transporte Colectivo. Siendo el transporte con mayor alcance territorial del país.

Uno de los problemas que presenta el STC Metro es que tiene una capacidad menor de servicio menor que a la que realmente se demanda. Esto trae como consecuencia mal servicio e ineficiencia del mismo. El "Metro"tiene una cantidad predeterminada de trenes, los cuales en ocasiones no son suficientes.

Por lo anterior se desea encontrar un modelo estocástico que describa el comportamiento de la demanda del servicio de transporte y reportar pronósticos adecuados a las autoridades pertinentes, de tal forma que se pueda llevar a cabo un adecuado plan de acción.

Objetivo

Obtener un modelo estocástico que estime la demanda de usuarios que ocuparán el servicio del STC Metro de la Ciudad de México para los primeros tres meses del 2020.

Hipótesis

El modelo obtenido será suficientemente significativo para pronosticar de manera eficiente, la demanda de usuarios que ocuparán el STC Metro.

Investigación Documental y cualitativa de estudios previos

En la investigación sobre artículos anteriores al nuestro que sentaran las bases para la realización del mismo, encontramos artículos como: "¿Por qué la gente no usa el Metro? Efectos del transporte en la Zona Metropolitana de la Ciudad de México"que es un análisis de regresión logística sobre las distancias existentes entre los usuarios del metro, con relación a su hogar y trabajo ,tomando como variables: la distancia y el nivel socio-económico de las personas. Así mismo, se hace un análisis de cuál es el motivo por el cuál, la población ocupa este transporte, en donde se concluye con la sugerencia de que, descongestionar el centro de la Ciudad de México mediante la concentración de la red del Metro en las áreas centrales, no ha dado el resultado esperado.

2. Marco Teórico

El Sistema de Transporte Colectivo Metro

«*El STC es un Organismo Público Descentralizado, cuyo objetivo es la operación y explotación de un tren rápido, movido por energía eléctrica, con recorrido subterráneo, de superficie y elevación, para dar movilidad principalmente a usuarios de la Ciudad de México y la zona Metropolitana del Valle de México*»[3], dicho tren rápido es mejor conocido como Metro.

La primer línea del Sistema de Transporte Colectivo Metro se inauguró el 04 de septiembre de 1969, logrando que «*la Ciudad de México se convirtiera en una capital más del mundo con un tren subterráneo, venciendo todas las dificultades técnicas del subsuelo más difícil del mundo*»[4]

Este medio de transporte fue creado con una misión dirigida a «*Proveer un servicio de transporte público masivo, seguro, confiable y tecnológicamente limpio. Con tarifa accesible, que satisfaga las expectativas de calidad, accesibilidad, frecuencia y cobertura de los usuarios y se desempeñe con transparencia, equidad y eficacia logrando niveles competitivos mundiales.*»[3]

No hay duda de que el Metro permite una mayor movilidad masiva de personas frente a cualquier otro medio de transporte gracias a su capacidad técnica, a las distancias que recorre y al tiempo que toma para realizar un viaje. Pero, con el tiempo, deterioro de los trenes y estaciones, aunado al crecimiento poblacional en la ciudad, el STC Metro ya no cumple con lo que reporta como su misión, en parte por no tener claro la cantidad de usuarios que tendrá, en ello radica la importancia de este trabajo.

Series Temporales

Se entiende como serie de tiempo a la sucesión de observaciones de una variable cuantitativa tomada en diferentes momentos, teniendo como limite el pasado, es decir, se toman datos históricos con respecto a periodos de tiempo.

Los componentes de una serie de tiempo son: tendencia, el cambio sistemático en el patrón de los datos; variación estacional, el patrón de comportamiento de algunos fenómenos como las estaciones del año; ciclo, es el patrón repetitivo pero con longitud mayor a un año; fluctuaciones aleatorias, todo aquello que el modelo matemático no puede explicar.

Los objetivos de análisis de las series son: describir, visualizar el comportamiento de un proceso; explicar, realizar un análisis de tipo causa-efecto; controlar, modificar el comportamiento de un proceso; pronosticar, estimar valores futuros de un fenómeno, lo cual es lo que pretendemos en este trabajo.

Metodología de Box-Jenkins

La metodología de Box-Jenkis se refiere a una aplicación de practicas de pronósticos de series de tiempo creada por George E. P. Box y Gwilym M. Jenkins en 1976, «la cual consiste en extraer los movimientos predecibles de los datos observados y separarlos de la parte no predecible»[9, Pág. 23], en sí el método se refiere a descomponer la serie de tiempo en sus componentes haciéndolos pasar por un filtro de estacionariedad, filtro de integración,

filtro autorregresivo y filtro de medias móviles hasta que se obtengan residuales que se asemejen al ruido blanco.

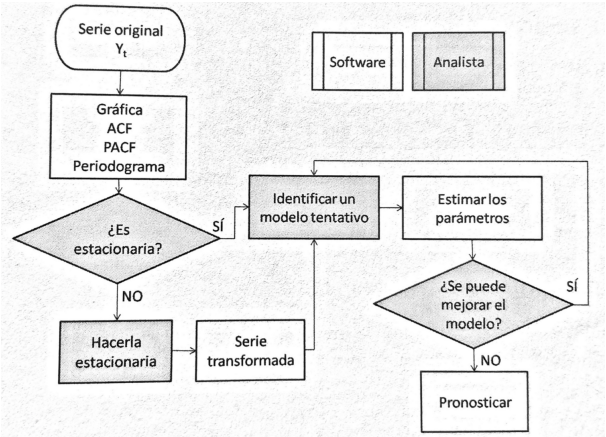

Figura 1: *Metodología de Box-Jenkins tomada de [10, Pág. 25]*

En la Figura 1 se presenta un diagrama funcional del método de Box-Jenkins [10, Pág.25], el cual, contiene los pasos ha seguir para llegar a un modelo adecuado.

Modelo SARIMA

El modelo SARIMA es una amplitud del modelo ARIMA, que se usa comúnmente cuando hay sospechas sobre la estacionalidad de un modelo. Por definición, el proceso autorregresico estacional integrados de medias móviles SARIMA (P,D,Q), es un proceso multiplicativo de dos procesos ARMA de la serie de tiempo diferenciada.

$$y_t = \delta_0 + \Phi_1 y_{t-s} + \cdots + \Phi_p y_{t-p_s} + \xi_t + H_1 \xi_{t-s} + \cdots + H_Q \xi_{t-s}$$

donde $y_t = \Delta_s^D W_t$, y W_t puede ser la serie de tiempo original o alguna transformación de ella. Para este modelo, la ACF y la PACF serán decrementos infinitos con valores significativos en múltiplos de s, con condiciones de invertibilidad y estacionariedad.

3. Análisis de los datos.

3.1. Exploración de la base de datos.

La base utilizada llamada "tabulado.csv"se tomó de la página del Instituto Nacional de Estadística y Geografía (INEGI)[9], tal base contiene 6 columnas: Periodo (meses), Longitud en servicio (km), Trenes en servicio, Kilómetros recorridos(miles de km), Pasajeros transportados (millones de pasajeros) y Energía eléctrica consumida (miles de KWH). Además contiene 447 registros, que corresponden a los mese de Enero 1986 hasta Marzo 2020.

La longitud en servicio se refiere a los kilómetros en los que se pueden transportar los pasajeros, en Enero de 1987 había 7 líneas del transporte y para Enero del 2020 el número de líneas es el actual, es decir 12 líneas. Por lo que la cantidad de trenes en servicio va ligado a la longitud de servicio.

Los kilómetros recorridos es la distancia total que recorre el metro durante un mes, en promedio el metro de la CDMX recorre 3122 miles de km por mes y ha recorrido en total 1,283,126 miles de km durante casi 30 años, lo equivalente a recorrer la circunferencia de la Tierra 32,018 veces.

Pacemos a la variable mas importante e interesante, los pasajeros. El Metro es el transporte con más alcance en una de las Ciudades mas Pobladas del mundo con más de 21 millones de habitantes. En promedio al mes el Metro recibe a 122 millones de personas. Al graficar la línea de tiempo (véase figura 4) se puede notar que la cantidad de pasajeros es periódica y cambia cada 12 meses, es decir cada año. Lo que no llevo a preguntarnos, cuál será el mes con más demanda durante el año.

En la Figura 2 podemos notar que la diferencia en la demanda de usuarios que utilizan el STC Metro no es grande, si no que durante el año varia muy poco la demanda mensual.

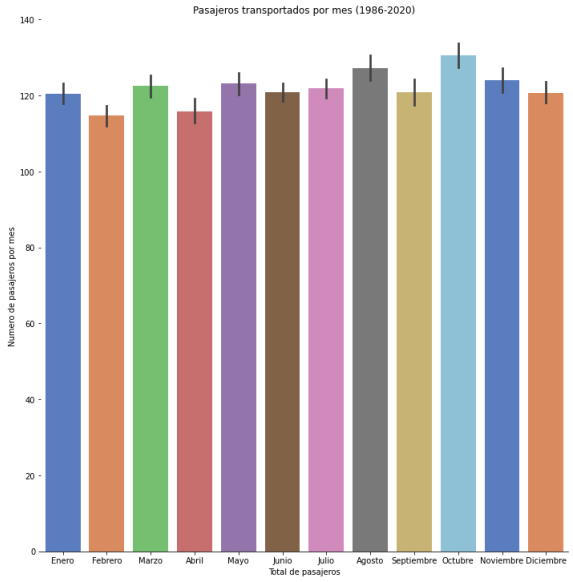

Figura 2
Meses con una mayor concentración, en nivel de pasajeros.

4. Metodología

Para efectuar este ejercicio se contó con los datos correspondientes a los pasajeros totales por mes, en millones de personas. Se obtuvieron cifras de enero 1986 a diciembre 2019, mismas que se observan en la Figura 2. Los datos recabados se procesaron utilizando Python.

Se tienen los datos a través del tiempo sobre el número de usuarios que utilizan este transporte junto con un análisis sobre artículos previos decidimos implementar un modelo de series de tiempo.

Primero gráficamos dichos datos para ver el comportamiento, a priori, que tienen nuestros datos. En este punto, nos damos cuenta que existe un dato atípico, en específico en Septiembre de 2017 ¿Por qué? Lo atribuímos al sismo ocurrido el 19 de septiembre de 2017 en la Ciudad de México.

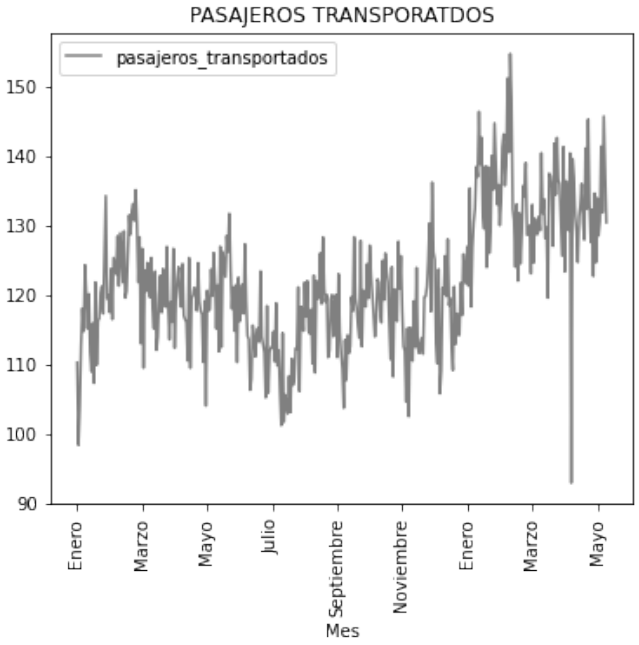

Figura 3

Al ser un dato atípico, y no un cambio estructural consideramos sustituir (para efectos del estudio de la serie de tiempo) con el promedio de los dos periodos más cercanos.

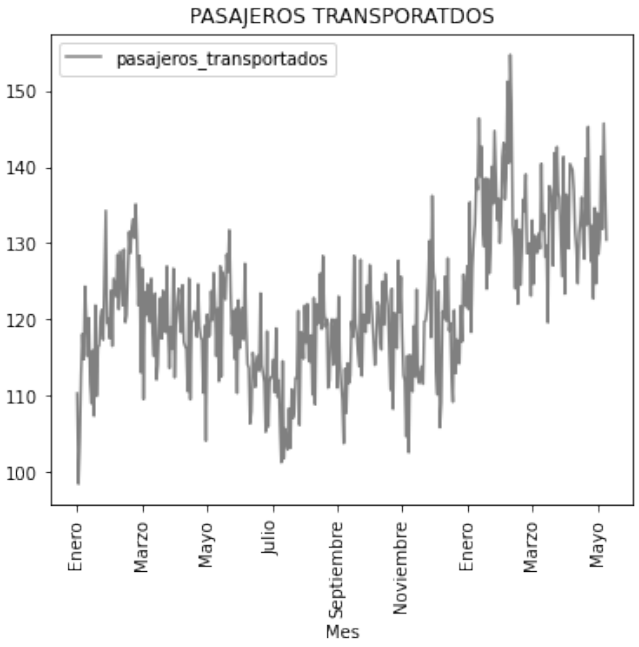

Figura 4

Una vez corroborado que la base de datos esté limpia se podrá implementar el primer intento para comprobar si existe alguna estacionariedad. Primero observaremos la función de autocorrelación.

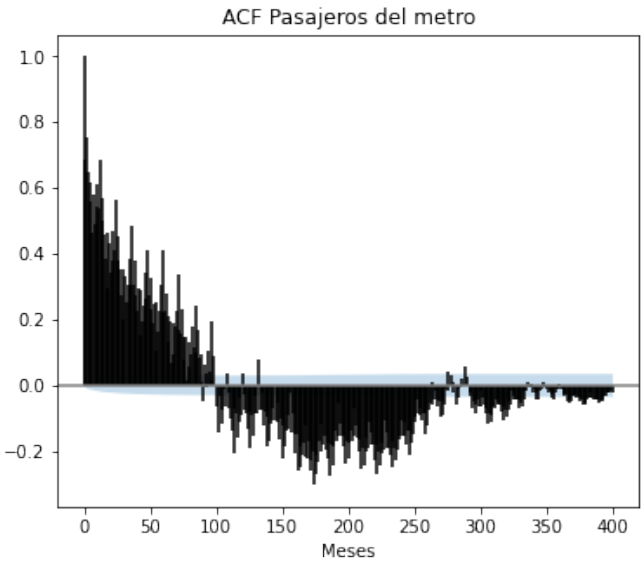

Figura 5

A priori, se observa que es necesario aplicar diferencias y es posible que nuestros no datos sean estacionarios. Para comprobarlo necesitamos pruebas como la prueba de Dickey-Fuller (ADF) para comprobar la existencia de raíces unitarias , lo que se busca es poder rechazar la hipótesis nula la cuál establece la existencia de raíces unitarias. Al poder rechazar está hipótesis aseguramos que la serie de tiempo sea estacionaria.

Se graficó la función de autocorrelación sin ningún tipo de transformación. En primera instancia, no se logro rechazar la hipótesis nula de la prueba ADF ya que el **p-value de la prueba resultó 0.4049**, demasiado alto. Por ello, aplicamos la función logaritmo natural para estabilizar la varianza, así como una diferencia a nuestros datos. Dado que ahora el p-value de la prueba resultó **0.000003**, pudimos rechazar la hipótesis nula y por tanto aceptamos que nuestros datos son estacionarios con ese nivel de probabilidad de Error Tipo I.

Una vez asegurada la estacionariedad de los datos nos asegura que conforme pasen los años la varianza no tendera a infinito provocando una incongruencia en los valores , porque si no hay varianza estable se esperaría que no haya un rango donde caería el valor.

Después fue necesario graficar la función de autocorrelación parcial (PACF) y la función de autocorrelación (ACF) para poder determinar un buen modelo que se ajuste.

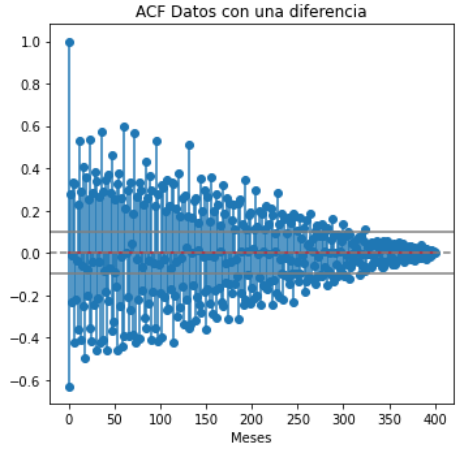

Figura 6: *ACF*

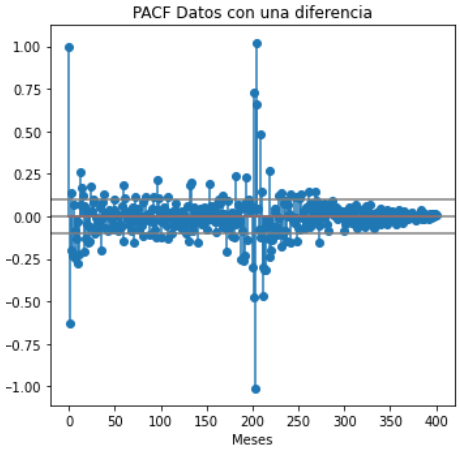

Figura 7: *PACF*

Lo que es importante en la observación de estas graficas es ver como decrecen con relación en el tiempo, un primero modelo a proponer es un ARIMA. En los modelos ARIMA(p, d, q), p representa el orden del proceso autorregresivo, d el número de diferencias que son necesarias para que el proceso sea estacionario, en este caso solo fue una diferencia por lo que $d = 1$ y q representa el orden del proceso de medias móviles. El cual será $q = 1$ esto nos dice que el choque aleatorio del que depende nuestro modelo es únicamente del ultimo.

Identificamos el orden de proceso autorregresivo ya que la función de autocorrelación no decrece de manera exponencial sino de manera proporcional hacia 0 que, la función de autocorrelación

parcial tenga solo p coeficientes distintos de cero. La expresión general de un AR(p) es una combinación lineal de p valores pasados de la variable y un ruido blanco actual. Esto es sólo un primer modelo propuesto, un ARIMA$(1, 1, 1)$.

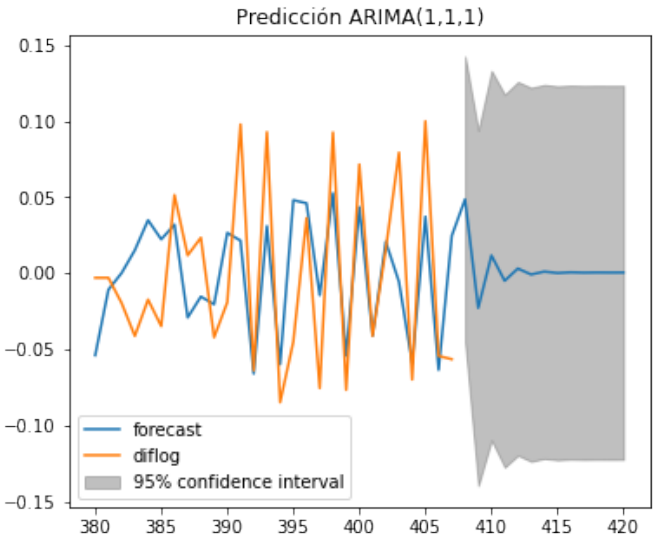

Figura 8

Donde se puede ver que el modelo no se ajusta a los datos. Por otro lado, para la comparación de modelos nos basaremos en el criterio Bayesiano, ya que empíricamente da mejores resultados con muestras grandes.

- **BIC:** -1295.793

Nosotros queremos probar el supuesto de que los residuos de nuestra serie de tiempo se comportan como ruido; por ello aplicamos la *Ljung-Box* buscando encontrar p-values muy altos para no rechazar la hipótesis de que sean ruido blanco. Sin embargo, los p-values fueron pequeños por lo que rechazamos la hipótesis de Ljung Box. Es decir, los errores no son ruido blanco. Adicionalmente, esto se puede comprobar visualmente con la función de autocorrelación de dichos residuos:

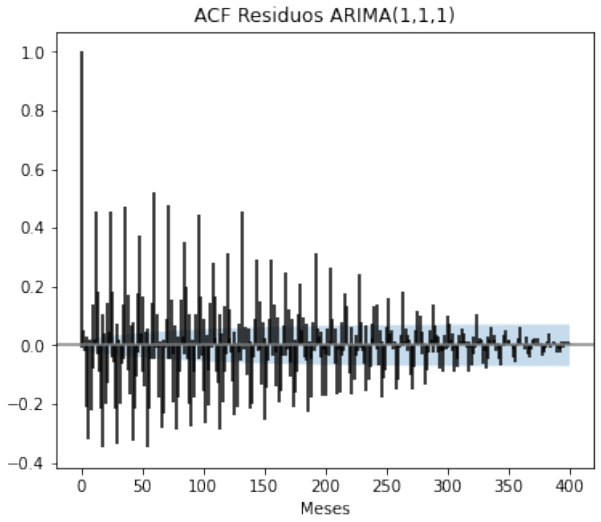

Figura 9

Esto nos llevó a proponer un segundo modelo. Un ARIMA$(2, 1, 1)$

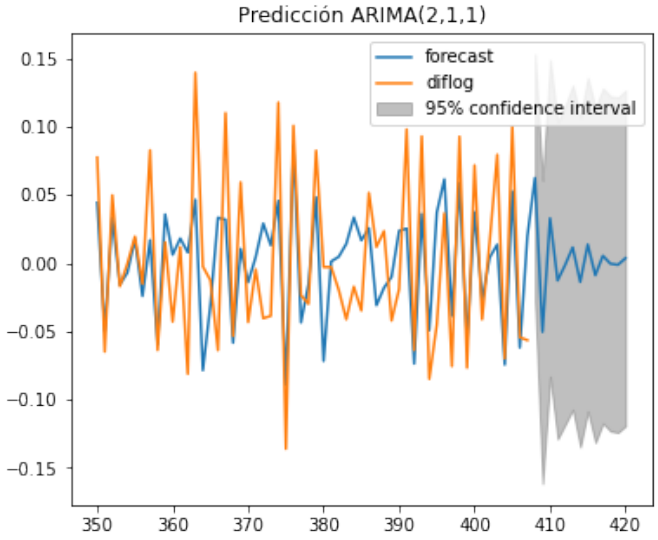

Figura 10

- **BIC:** -1320.689

Seguimos observando que el modelo no logra ajustarse a los datos. Así mismo los residuos tampoco presentan aun el comportamiento de ruido blanco.

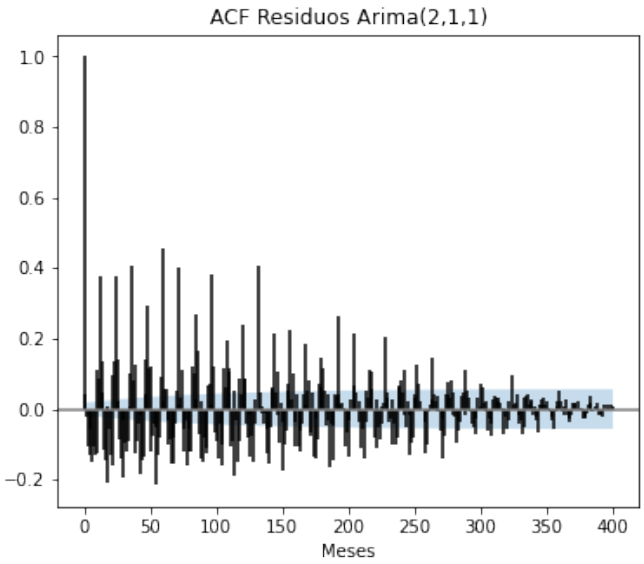

Figura 11

Por ello que descarto el modelo, a pesar de ser mejor según el críterio Bayesiano. Ahora propusimos un modelo ARIMA(2,1,2)

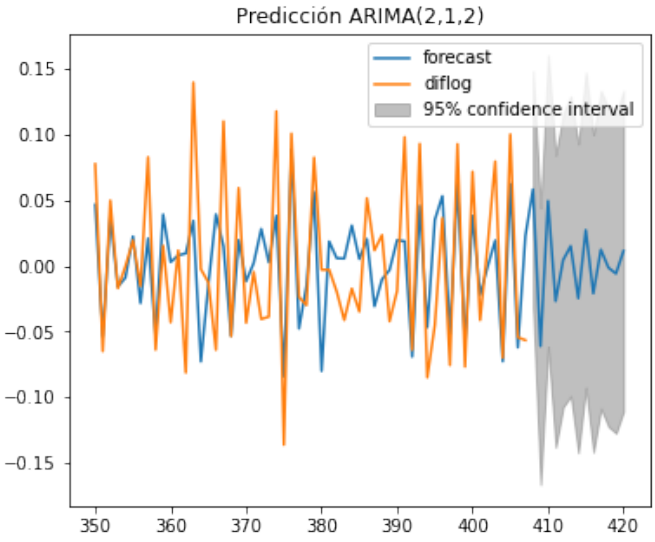

Figura 12

- **BIC:** -1319.86

Este modelo de descartó dado que el criterio Bayesiano nos indica que el modelo ARIMA(2,1,1) era ligeramente mejor. Además, resulta menos significativo el coeficiente del segundo orden de medias moviles. Dado que anteriormente el criterio Bayesiano había premiado un orden mayor en la parte autoregresivo, proponemos un ARIMA(3,1,2).

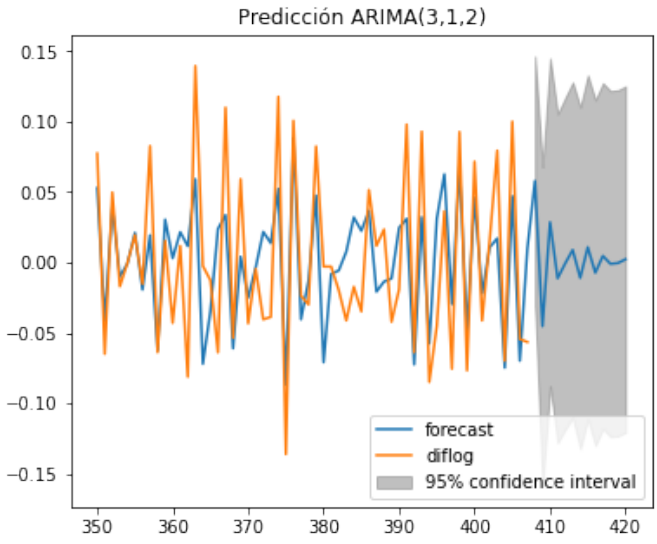

Figura 13

- **BIC:** -1327.31

Al aplicar la prueba Ljung Box y los p-values para rechazar que los residuos fueran ruido blanco eran demasiado pequeños aun. A pesar de que aun no tenemos evidencia de que los residuos sean ruido blanco, sabemos que este es un mejor modelo por el criterio Bayesiano. Con la misma lógica con la que propusimos este modelo, ahora intentamos con un ARIMA(4,1,2).

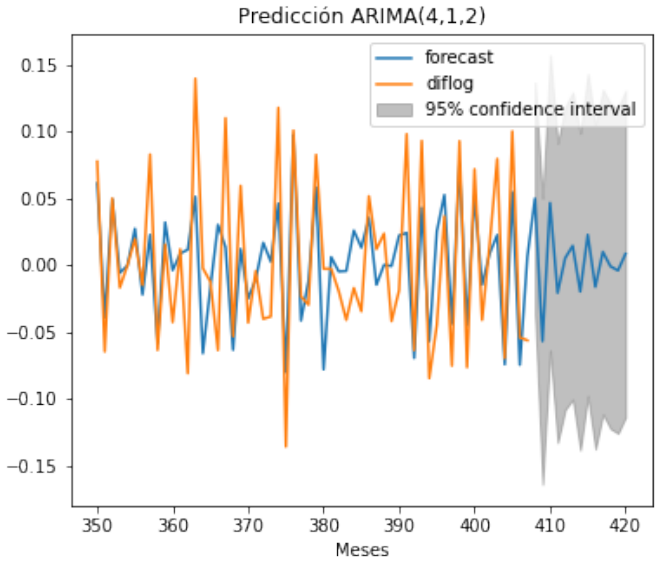

Figura 14

- **BIC:** -1336.36

El BIC de nuevo premia ese nuevo orden en la parte autoregresiva. Veremos qué pasa con un ARIMA(5,1,2)

- **BIC:** -1330.355

Vemos que ya no premia el agregar un orden mayor en la parte regresiva. Ahora, si regresamos a análisar la ACF un poco más de cerca nos damos cuenta que hay un rezago más significativo después de cada doce rezagos, por lo que proponemos un SARIMA(4,1,2) con varianza estacional cada 12 periodos.

- **BIC:** -1445.939

Hasta ahora es el mejor modelo de acuerdo al criterio BIC. Aplicando la prueba de Ljung-Box a los residuos, nos dan valores considerablemente grandes por lo que no pudimos rechazar la hipótesis nula. Es buena señal para poder comprobar que los residuos son ruido blanco. Veamos su función de autocorrelación:

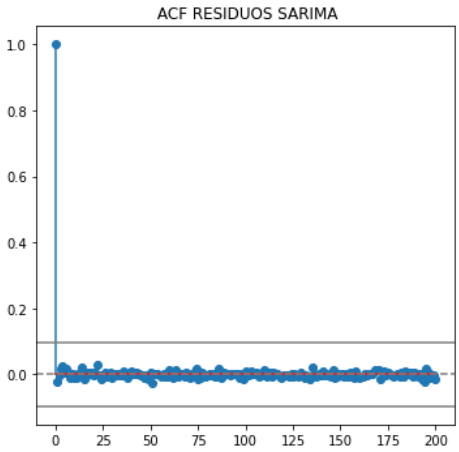

Figura 15

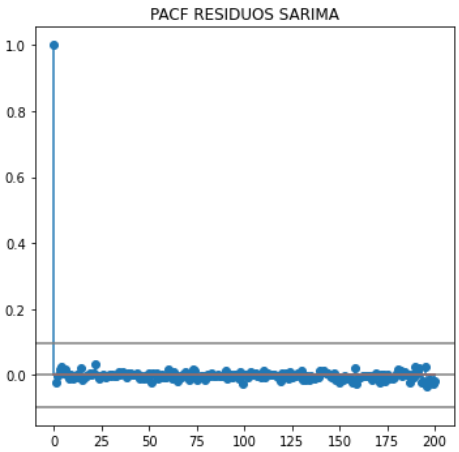

Figura 16

Nuestros residuos ya se comportan como ruido blanco. A continuación vemos un un histograma con su distribución (y una comparación con una función de densidad de una Normal con media 0 y varianza 1).

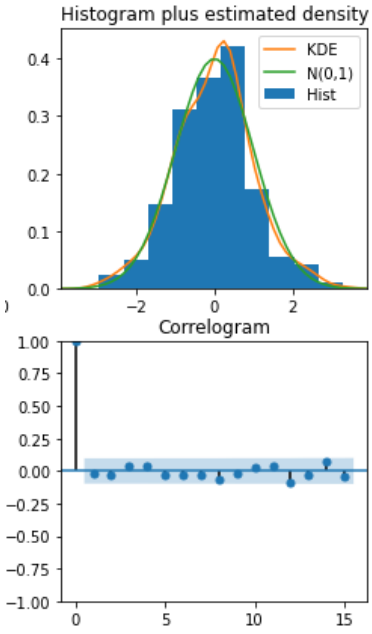

Figura 17

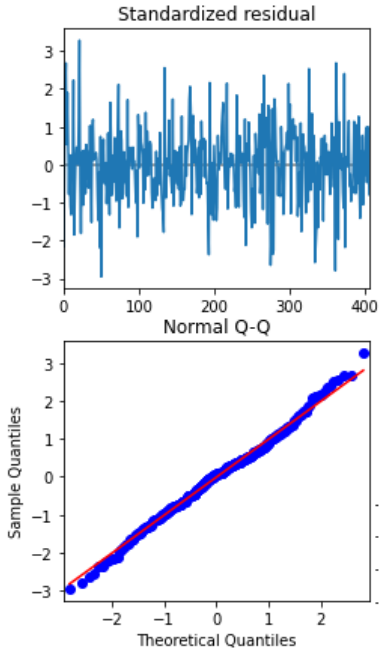

Figura 18

Por lo tanto, dado que es el mejor modelo según el criterio Bayesiano y los residuos ya se comportan como ruido banco, nos quedaremos con este modelo para hacer nuestro pronóstico.

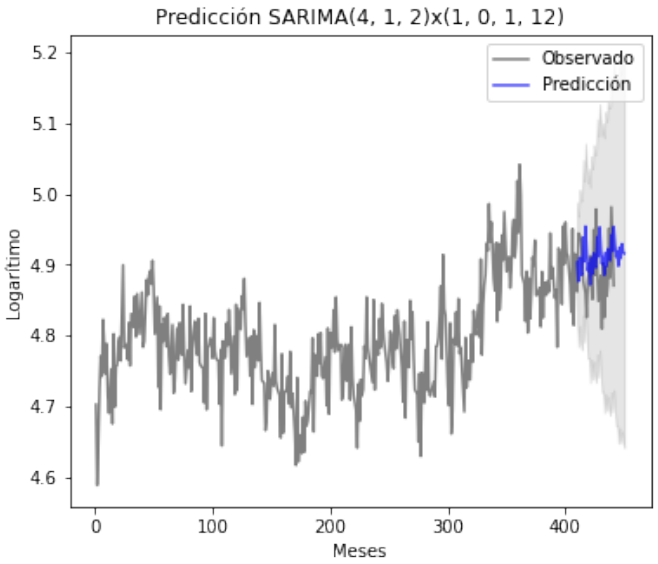

Figura 19

El pronóstico de nuestro modelo para la *cantidad de pasajeros* que transportará el metro en 2020 son los siguientes:

Mes 2020	Pronóstico (Millones de Pasajeros)
Enero	128.60
Febrero	134.76
Marzo	131.32
Abril	135.72
Mayo	132.98
Junio	132.50
Julio	139.63
Agosto	135.08
Septiembre	141.88
Octubre	137.01
Noviembre	133.39
Diciembre	134.75

Anexamos los datos observados en 2019 y los primeros tres meses de 2020

Mes 2019	Observado (Millones de Pasajeros)
Enero	132.31
Febrero	122.68
Marzo	134.61
Abril	124.67
Mayo	133.93
Junio	128.52
Julio	130.57
Agosto	141.37
Septiembre	131.82
Octubre	145.70
Noviembre	137.99
Diciembre	130.41

Mes 2020	Observado (Millones de Pasajeros)
Enero	130.70
Febrero	129.06
Marzo	107.53

Conclusiones

Pudimos demostrar la estacionariedad de nuestros datos con la prueba de Dicky Fuller, así como comprobar que los residuos de nuestros datos se comportaban como ruido blanco con el modelo SARIMA. Además, según el criterio Bayesiano dicho modelo resultó ser el mejor de todos los que exploramos. Sin embargo, a pesar de contar con datos de más de 30 años, hay eventos atípicos que no se pueden modelar con este tipo de modelos de Series de Tiempo. Al menos, hay que tenerlo en mente al utilizar estos modelos, ya que en este proyecto pudimos comprobar que para los primeros dos meses del año, los pronósticos se parecían a lo observado, sin embargo para el mes de marzo hubo una caída en el uso del metro ya que los efectos de la actual pandemia del COVID-19 empiezan a hacer efecto. Probablemente cuando el INEGI publique el mes de abril podremos comprobar que hubo una caída aun mayor y por ello no se cumplió nuestra hipótesis.

Precisamente esa fue nuestra principal **complicación** al llevar a cabo el presente proyecto,

los datos atípicos. Como mencionamos al principio de la sección de *metodología*, había un valor muy extremos en el mes de septiembre de 2017. Al graficar la función de Autocorrelación, y la de autocorrelación parcial sesgaba demasiado el comportamiento de nuestros datos que intentabamos modelar, era un dato demasiado volatil. Por otro lado, una vez solucionado lo del dato atípico, después de aplicar la transformación logarítmica para disminuir la varianza de los datos y aplicar las diferencias, en la *función de autocorrelación parcial* se ven algunos rezagos que tienen mucha correlación entre si, sin embargo ello no pudimos solucionarlo.

Por último, debemos aceptar que la base de datos que emite INEGI viene muy limpia, por lo cual la limpieza que tuvimos que hacerle a la base, fue relativamente poca.

Referencias

[1] Brockwell,P.J.y Davis, R.A,(2006).Introduction to time series and forecasting.Springer.

[2] Cabrera, A. (2005). Modelo de un pronóstico estadístico de la demanda de los principales medicamentos del I.M.S.S Puebla. Recuperado de http://catarina.udlap.mx/u_dl_a/tales/documentos/lat/cabrera_ga/

[3] Cdmx, M. (s. f.). Acerca de. Recuperado 21 de mayo de 2020, de https://www.metro.cdmx.gob.mx/organismo/acerca-de

[4] CDMX. (2017). Diagnóstico sobre el servicio y las instalaciones del sistema de transporte colectivo 2013-1018. Recuperado de https://www.metro.cdmx.gob.mx/storage/app/media/Banners/ diagnostico.pdf

[5] CETRAM(2012), Publicación oficial del transporte, SETRAVI, Ciudad de México. [http://web.archive.org/web/20071123165154/http: 2 de noviembre, 2012]. CPSR (1999), Creating Transit Station Communities, in A Transit-Oriented Development Workbook, Central Puget Sound Region, Seattle.

[6] Cowpertwait,P y Metcalfe,A (2009). Introductory Time Series with R. Springer.

[7] D. Cryer,Jonathan y Chan,Kung-Sik (2008). Time Series Analysis With Applications in R. Springer

[8] Enders, W.(2008).*Applied econometric time series*.John Wiley Sons.

[9] Geografía, E. D. N. I. Y. (s. f.). Transporte de pasajeros. Recuperado 18 de mayo de 2020, de https://www.inegi.org.mx/temas/transporteurb/default.htmlTabulados

[10] González, M. (2011). Pronósticos: Metodología de Box-Jenkins (1.a ed.). CDMX, México: FES Acatlán.

[11] Hamilton, J. D.(1994). *Time series analysis* Princeton: Princeton university press.

[12] Knight, R. y L. Trygg (1977), "Evidence of Land Use Impacts on Rapid Transit Systems, Transportation, vol. 6, pp. 231-247.

[13] Wei, W.Time series analysis: univariate and multivariate methods.Pearson Addison Wesley.

[14] Wei, W.Time series analysis: univariate and multivariate methods.Pearson Addison Wesley.