Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física
Maestría en Ciencia de Datos

Planificación de Eventos Socioculturales Para Evitar la Afectación De Usuarios De MiBici

---

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
Maestro en Ciencia de Datos

Presenta:
José Ricardo Javier De León Flores

Director:
Mtro. Byron Michael Motta Bonilla

Tlaquepaque, Jalisco, 27 de mayo de 2024

Planificación de Eventos Socioculturales Para Evitar la Afectación De Usuarios De MiBici

José Ricardo Javier De León Flores

Resumen

Jalisco ha tenido un crecimiento poblacional acelerado, se ha convertido en una de los estados más importantes del país. Como todo estado con crecimiento acelerado, los problemas vienen, problemas de toda indole. Este trabajo esta enfocado en los problemas que causa una mala planificación de los eventos socioculturales, el cómo es afectado el traslado de la población por la ciudad cuando los eventos obstaculizan las vias de transporte, en este caso el enfoque esta en el transporte urbano denominado Mi Bici.

Mi Bici es un sistema de transporte publico que esta localizado en la zona metropolitana de Guadalajara (ZMG), diversas estaciones se encuentran ubicadas dentro de esta zona en las cuales mediante una renta tienes acceso a una bicicleta y se puede retornar ya sea en la misma estación o en una diferente. Básicamente este es el modo de operar del sistema. Este trabajo intenta demostrar, mediante un modelo de tiempo el crecimiento de la población que hace uso de este transporte y así poder llegar a demostrar que grandes cantidades de usuarios son afectados en sus actividades diarias, ya sea para trasladarse a su trabajo, a la escuela o simplemente por actividad física, cuando eventos socioculturales de cualquier magnitud obstruyen las areas en donde se mueven estas masas de ciclistas.

La pandemia juega un papel importante en este pronostico, gráficamente se puede visualizar el crecimiento de esta población de ciclistas se fue a cero dada este acontecimiento, ya que todos tuvimos que permanecer durante un largo tiempo encerrados.

Es por eso que este trabajo se acoto a un tiempo hasta antes de la pandemia y así visualizar si existía un pronostico de crecimiento o al menos una estabilización en la cantidad de población de usuarios para alertar a las autoridades del gobierno de Jalisco, ya que una eventos que interrumpieran las actividades del día a día de esta población pudiera causar descontento entre los usuarios de este transporte.

Los resultados de este trabajo, demostraron que en realidad hay un crecimiento en general y que el pronostico tiende al crecimiento, por lo cuál pudiera ser la premisa para un análisis más profundo para detectar más patrones que pudieran ayudar al gobierno de Jalisco a una buena planificación de cada evento que se libere y coincida en la zona de trafico de usuarios de MiBici.

Palabras clave: MiBici, bicicletas, transporte urbano, series de tiempo, pronostico, ARIMA, SARIMA.

Tabla de Contenidos

Índice de figuras

Página

Índice de tablas

Página

Dedicado a...

Este trabajo está dedicado a mi esposa, quien ha sido un pilar fundamental durante estos dos años de maestría. Mientras yo me concentraba en mis estudios, ella se encargaba con amor y dedicación de nuestros hijos y de sus actividades academicas y deportivas. Su apoyo incondicional fue crucial para lograr este objetivo.

Agradezco de manera especial a mis compañeros de clase, en particular a Lilivette y Oscar de Astrazeneca, por su constante disposición para resolver dudas y por mantener una comunicación fluida que

fue fundamental para sacar adelante esta maestría.

Expreso mi profundo agradecimiento a Astrazeneca, entidad que patrocinó gran parte del financiamiento de esta maestría, así como al ITESO y al Gobierno de Zapopan.

A todos y cada uno de mis maestros que ponen siempre su esfuerzo por hacer de nosotros y de nuestro país un ejemplo para un mundo mejor.

Quiero reconocer la invaluable guía y orientación de mi asesor de proyecto, el Ingeniero y Maestro Byron Michael Motta Bonilla, quien estuvo siempre dispuesto a brindarme su conocimiento y experiencia para enriquecer este trabajo.

Agradezco también a mí mismo por el esfuerzo constante y el compromiso demostrado para cumplir con los objetivos de esta maestría, así como por encontrar el equilibrio entre mi rol como padre y mis responsabilidades académicas.

Finalmente, doy gracias a Dios por otorgarme la salud y la fortaleza necesarias para alcanzar esta meta.

1 Introducción

Figura 1.1: MiBici Transporte publico urbano.

Del texto en el sitio web territorio: 'En 2003, en el techo de una casa del centro de Guadalajara surgió Popular, un fanzine que se distribuía en copias fotostáticas con contenido quisquilloso y creativo; propositivo e informativo, creado por Isaac Padilla y sus primos, Mario, Carlos y Fabián Delgado.

El primer número era sobre la movilidad urbana aunque sus intereses abarcaban el esténcil, el hip-hop, la comida, la política, el diseño, la gráfica, la patineta o el ocio.

Desde hacía tiempo la bicicleta era su principal medio de transporte, por ahorrador y por su fácil disponibilidad, pues las encontraron arrumbadas en casa de sus tías o de sus padres.

El uso de la bicicleta en Guadalajara como medio de transporte y como vehículo para actividades laborales, deportivas y de esparcimiento se remonta a la década de 1940, cuando se consideraba una buena opción para reducir gastos.

Para finales de los años noventa, Guadalajara padecía problemas ambien-

tales, sociales, económicos y de planeación muy graves, además de un alto deterioro en el espacio público.

Las políticas de movilidad privilegiaban al automóvil mientras que el transporte público, la bici o la movilidad peatonal, eran temas marginales.

En 1999, Gabriel Michel, un profesor del ITESO, junto con un grupo de estudiantes de la carrera de arquitectura, prepararon durante su servicio social un estudio, que después presentaron al gobierno, sobre la bici como medio de transporte y la falta de infraestructura en la ciudad.

El estudio incluía la propuesta de crear una red de ciclovías que fue inmediatamente adoptada por Fernando Garza (PAN), alcalde recién electo de Guadalajara.

A principios del 2002 se inicia la construcción de la primera ciclovía, pero unos meses después, un grupo de colonos que sumaban entre 100 y 120, demandan al municipio argumentando que no fueron consultados, que traerían problemas viales y que el número de personas que viaja en bicicleta por la zona era tan poca, que no se justificaba la inversión.

En respuesta, se realizó un recorrido ciclista a favor de la medida, en el que participaron cerca de 150 personas, entre funcionarios, ciudadanos y académicos, pero los vecinos ganaron la demanda y a inicios del 2003, se tuvo que destruir la ciclovía.

El movimiento ciclista no desistió y se manifestó de diferentes maneras, una de ellas a través de la vía recreativa, la cual tuvo un gran impacto al conectar diferentes municipios de la zona metropolitana de Guadalajara y al fomentar la participación de los habitantes de Jalisco. Esto propició que la bicicleta como medio de transporte adquiriera una gran importancia. La Vía RecreActiva impulsó una nueva conversación pública y cambió la relación entre los grupos organizados y el gobierno.

En pocos años, la ciudad cambió y llegó a haber más de 60 grupos promoviendo el uso de la bicicleta.

La vía recreativa reunía aproximadamente 100 mil usuarios y se organizaban más de 30 paseos nocturnos.

Los colectivos proponían políticas, ejercían presión, influían y colaboraban para encontrar soluciones', [1] la Figura 1.1 hace alusión a como la bicicleta cada ves más forma parte de de la urbanización [2].

1.1 Contexto

Tomado del sitio web del gobierno del estado de Jalisco: 'La movilidad ha sido un objeto de estudio y análisis profudo en los últimos años. Más aún en ciudades que han tenido un crecimiento poblacional y de urbanidad en estados como Jalisco.

Uno de los problemas que llega con la movilidad es en gran parte la sostenibilidad ambiental y energética mundial. El problema ambiental más grave es el asociado con la dependencia de energías fósiles no renovables, y a su vez el imapctio ambiental que genera el producir este tipo de energías.

En el año 2014, el gobierno del estado de Jalisco emprendió una iniciativa conocida como 'MiBici', MiBici es un servicio de transporte público basado en una red de bicicletas compartida, con el objetivo de proporcionar una alternativa de transporte ecológico y sostenible para los habitantes de la Zona Metropolitana de Guadalajara (ZMG). Esta iniciativa permitía a los residentes de la región acceder a bicicletas a través de estaciones ubicadas estratégicamente en tres de las principales ciudades de la ZMG.

Desde su implementación, el 1 de diciembre del 2014, MiBici ha cambiado la forma de movernos en la metrópoli, a 7 años, cada día es más común ver a los usuarios en las bicicletas de MiBici por las calles de la ciudad y en la infraestructura ciclista, que va sumando kilómetros como parte también de la apuesta en Jalisco por la movilidad activa.

El programa inició con 11 estaciones en Zapopan, 86 en Guadalajara y 1276 unidades; actualmente se cuenta con 300 estaciones en los municipios de Guadalajara (227), Zapopan (61) y Tlaquepaque (12) y 3 mil 200 bicicletas para los más de 109 mil usuarios registrados. Según datos de la encuesta a usuarios del 2020, el 58 % utiliza MiBici para llegar a sus lugares de trabajo y sus viajes son complementados caminando en un 34 % y en transporte público 25 %.

De acuerdo con esta última encuesta los usuarios han aumentado los días que usan el sistema, comparado con el 2019 creció un 10 % el uso en 5 o más días a la semana y entre los beneficios que han notado por usar la bicicleta mencionan una mejor condición física, ahorro en su economía y el sentirse más relajados.

En general el 98 % de los encuestados señalaron que MiBici ha contribuido a una mejora de la movilidad en la ciudad. Además, la tarjeta Mi Movilidad se puede vincular a MiBici y así usarla con el resto del sistema de transporte público.

Cada día se realizan en promedio cerca de 11 mil viajes con una duración promedio de 12 minutos, evitando así viajes en vehículos motorizados y contribuyendo con la reducción de CO_2 al no emitir contaminantes [3]'.

1.2 Justificación

'Jalisco es el segundo de los 32 estados de la República Mexicana donde se concentra la mayor riqueza económica, cultural y simbólica del país. Es también el estado que más empleos formales genera. La capital, Guadalajara, es una de las áreas metropolitanas más vibrantes y complejas de toda América Latina, sede de importantes festivales y ferias internacionales, y desde hace algún tiempo, un hub o centro de operaciones para los desarrollos digitales y el emprendedurismo. En 2020, Guadalajara fue nombrada Capital Mundial del Libro para 2022 por la UNESCO.

Dado que varias de las rutas del sistema de transporte MiBici comprenden las calles más transitadas de la zona metropolitana de Guadalajara (ZMG), y al hacer el analis de las rutas, ubicar las horas pico y los días de mas afluencia por zona pudiera influir para definir que días pudieran ser lo más convenientes para planificar un evento cultural o social que pueda llegar a ser afectado por el uso del transporte urbano MiBici ya que una mala planeación pudiera costar varios miles de pesos al gobierno del estado simplemente por posporner el evento o retrasarlo.

Al hacer el analis de las rutas, ubicar las horas pico y los días de mas afluencia por zona se pudiera deducir que días pudieran ser lo más convenientes para planificar un evento sociocultural y así evitar la afectación de las actividades cotidianas como ir al trabajo o ralizar algún otro tipo de actividad a usuarios que se movilizan con el transporte urbano MiBici. Esta deducción pudiera servir de sugerencia al gobierno del estado para planificar este tipo de eventos masivos' [4].

1.3 Problema

Los eventos socioculturales masivos que tienen convergencia con las rutas del transporte urbano MiBici, influyen de una manera negativa afectando las actividades cotidianas de los usuarios de MiBici, como las trayectorias de transporte al trabajo o quizás a alguna otra actividad recurrente.

1.4 Objetivos

1.4.1 Objetivo general

Recomendar la planificación de eventos socioculturales a las autoridades de Jalisco por medio de diferentes tipos de análisis de los datos recopilados de la página oficial de MiBici, así como la afectación los usuarios de este transporte, cuando este tipo de eventos empatan con las rutas más transitadas.

1.4.2 Objetivos específicos

- *Analizar la información existente y hacer una propuesta en la cual se recomiende la planificación de eventos socioculturales en la ZMG que pueda afectar la actividad diaria de los usuarios de MiBici.*
- *Analizar la zona en un mapa para contemplar la afectación de la población de usuarios de MiBici.*

2 Metodología

- *Como parte de la limpieza y arreglo de datos, es formar un solo dataset que contenga toda la información del rango de fechas antes mencionado.*
- *Analizar la integridad de los datos, es decir, que la información este completa.*
- *En caso de no estar completa, aplicar las técnicas de imputación de datos, dependiendo el tipo de variable.*
- *Analizar la distribución de los datos.*
- *Aplicar transformación y escalamiento de ser necesario.*
- *Visualizar los datos gráficamente.*
- *Definir el modelo para clasificar.*

2.1 Descripción de los datos

Los base de datos de MiBici, es una base de datos abierta, por parte del gobierno del estado de Jalisco [5]. En su página podemos descárgalos en formato csv.

Los datos están disponibles en descarga por año y a su vez, por mes, que comprende desde el año 2014 al año 2023. A su vez, existe un archivo extra que contiene información de las estaciones.

Tipos de archivos:

- *datos_abiertos_YYYY_mm.csv*
- *nomenclatura_YYYY_mm.csv*

Ahora vamos a describir lo que contiene el archivo ***datos_abiertos_YYYY_mm.csv***

Columnas:

- *Viaje_Id (El Id del viaje recorrido) agregar tipo de dato.*

- *Usuario_Id (El Id de usuario de MIBici).*
- *Genero (M: Mujer, H:Hombre).*
- *Anio_de_nacimiento (Año de nacimiento del usuario YYYY : HH:MM).*
- *Inicio_del_viaje (Fecha y Hora).*
- *Fin_del_viaje (Fecha y Hora.)*
- *Origen_Id (Id de la Estación en donde se origino el viaje).*
- *Destino_Id (Id de la Estación en donde se termino el viaje).*

Ahora vamos a describir lo que contiene el archivo ***nomenclaturas_YYYY_mm.csv***

Columnas:

- *Id (Id de la estación).*
- *Name (Nombre de la estación).*
- *Obcn (Alias de la estación).*
- *Location (Lugar de La estación, municipio).*
- *Latitude (coordenada de ubicación).*
- *Longitude (coordenada de ubicación).*
- *Status (IN_SERVICE : En servicio, NOT_IN_SERVICE: Sin Servicio).*

2.2 Análisis exploratorio

Tratamiento de CSV

Como ya se mencionó los datos son públicos y están disponibles en la página del gobierno de Jalisco de MiBici.

Los datos están publicados por año y por mes, desde Diciembre del 2014 hasta el 2023 hasta el último mes concluido completo (Octubre). Por lo tanto se tienen 107 archivos con extensiones (csv) que tienen que ser integrados en un solo conjunto de datos para su análisis.

El primer tratamiento que se le dará a los datos será el de formar un solo documento con todos los archivos descargados localmente. Los archivos se encuentran guardados por año y por mes Figura 2.1.

Figura 2.1: Datos crudos divididos por año y mes.

Para el tratamiento de datos se utlizará el lenguaje de programación en python versión 3. Se elegió este lenguaje dado que es el lenguaje con el que se tiene mayor experiencia y además tiene un amplio soporte en datos.

Se utilizó una función recursiva para concatenar todos los archivos, uno debajo del otro formando un dataframe principal. Finalmente, con el fin de continuar trabajando localmente, una vez concatenados, se realizó un respaldo de estos en un archivo nuevo con formato CSV.

Analisis de variables (Columnas)

Lamentablemente, como es muy común, los archivos no estaban guardados de forma estandarizada, es decir, el nombre de las columnas o variables en algunos casos no eran los mismos, lo que provocó duplicidad de variables haciendo referencia a lo mismo.

En este caso, se tiene como resultado, tres columnas de "Año_de_nacimiento" debido a la codificación que se usa. Muy probablemente la letra "ñ" es la que afecta esta duplicidad de columnas.

Figura 2.2: Duplicidad de columnas causadas por una mala estandarización.

Se hace el análisis del numero de filas que contiene cada columna, para crear una columna nueva que contenga el acumulado de filas de todas las columnas con nombres diferentes.

Por último se eliminan las columnas con diferentes nombres dejando un solo archivo con la nueva columna creada.

Tratamiento del tipo de dato

Una vez teniendo un archivo limpio, se comienza con el análisis meramente de datos. Para esto, por medio del comando ***dtypes*** *podemos observar el tipo de dato relacionado para cada variable o columna. La salida de este comando nos arroja que tenemos algunas variables como el 'Genero', 'anio_de_nacimiento', 'inicio_del_viaje' y 'Fin_del_viaje' son de tipo 'object' o 'float64' respectivamente, las cuales no tienen el formato apropiado para efectos del pronostico que se quiere llegar a hacer por medio de una* ***serie de tiempo*** *y algunos otros análisis que se pueden llegar hacer como la de usuarios entre determinados rangos de edad.*

Por alguna razón, el 'anio_de_nacimiento' es de tipo 'float64', pero como la columna es de tipo año, se cambia a int64 para normalizarlo con los otros datos de tipo entero.

De igual manera, las variables de inicio y fin de viaje se convierten de tipo 'object' a tipo 'date_time'.

Una vez hecha la conversión de datos al tipo de dato adecuado, procedemos a revisar los campos vacíos o ***null*** *para ver si es necesario hacer alguna otra manipulación de datos como alguna imputación o descartar columnas o filas. Por el momento, solo se revisará los nulos de las variables de tiempo, ya que estas serán el enfoque de la* ***serie de tiempo****.*

Afortunadamente ni una de estas variables de tiempo contienen nulos, por lo tanto se toma la decisión de elegir una de las dos variables de tiempo para formar la serie. Elegí como variable principal ***'inicio_del_viaje'*** *ya que aquí tenemos la certeza de cuando comenzó a usarse el transporte, de otra manera, si elegimos el 'Fin_del_viaje' no tenemos certeza de que la devolución del medio de transporte en este caso, la bicicleta.*

Ahora bien, para hacer una análisis de los viajes por las zonas tenemos las variables 'Origen_Id' y 'Destino_Id', las cuales actualmente se encuentran com tipo de variable 'int64' pero no esta como un tipo de variable categórica como tal; Las categorías se encuentran en el archivo de nomenclaturas el cual sera cargado en un nuevo dataframe, anteriormente descrito. Para hacer una referencia del nombre de la zona con el origen y el destino, es necesario hacer una trasformación mediante la relación de ambos dataframes.
El procedimiento es sencillamente ubicar la columna común en ambos datafra-mes y fusionarla en nuestro dataframe original.

*Una vez más se revisaran los datos **nulos** que tengamos en estas variables, se obtuvo un resultado de 757 datos faltantes en la columna de 'Origen_Id' de un total de 25,762,932 filas, esto representa el **.0029%** de datos perdidos. En comparación con 'Destino_Id' que es un monto de datos nulos muy similar 868, nuevamente decidí tomar el criterio de usar la columna de inicio de viaje dado que se me hace más certero trabajar con una variable que pareciera tener una información más verídica, ya que el registro de inicio del viaje nos dice más que la del final de viaje dado que no tenemos la certeza de cuando se regreso el vehículo.*

Se consideró usar imputación por medio de la moda tomando en cuenta que la imputación se hará a una variable categórica[6], dado que no es un porcenaje alto de datos nulos pudiera llegar considerar eliminarlos pero decidi en este caso reemplazrlos con la moda.

Graficas de frecuencia

En la Figura 2.3 se puede observar la frecuencia de viajes en bicicleta dividido por zonas de acuerdo al catálogo.

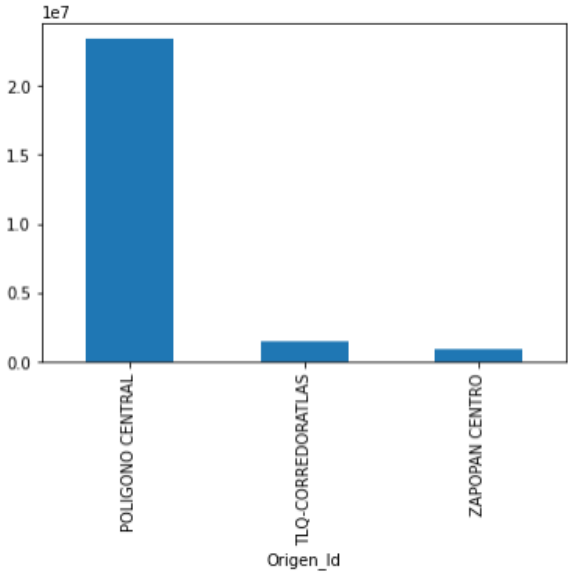

Figura 2.3: Gráfica de frecuencia de viajes por zona.

*Con ayuda de una libreria de python llamada '**folium**' se señalizará en un mapa las zonas en donde se muestra el uso del servicio de MiBici.*

*Mediante la conversión a dataframe del archivo con datos de localización **nomenclatura_YYYY_mm.cs** que contiene las columnas '**latitude**' y*

*'**longitude**', se realizo la señalización de uso por localidad. El mapa con la concentración de las tres zonas como estan divididas en los archivos se veria como en la Figura 2.4*

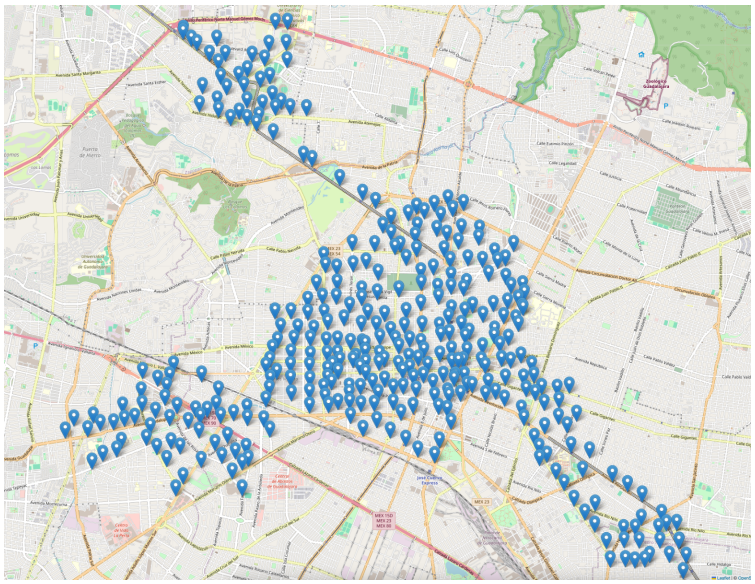

Figura 2.4: Mapa de localidades de estaciones de MiBici.

*En la Figura 2.5 estan señalizadas las zonas por bloques de colores, **Poligono central** representada en color rojo, **Zapopan** representada en color verde y **Tlaquepaque** en color azúl.*

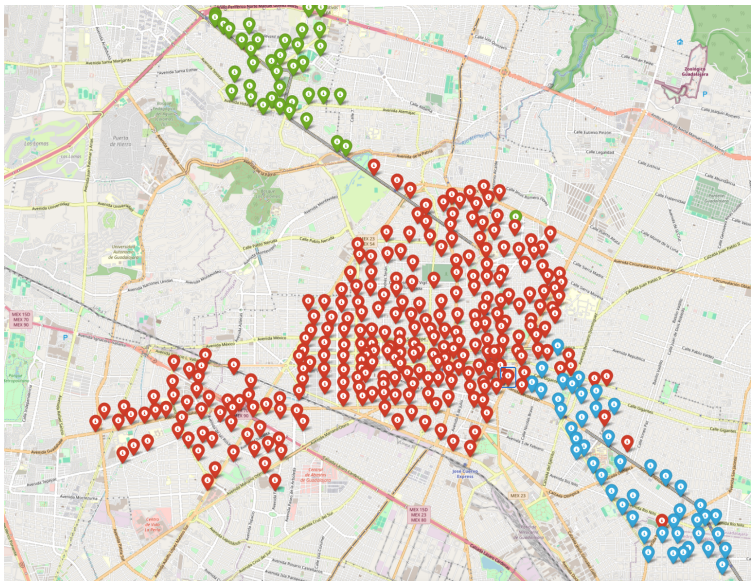

Figura 2.5: Mapa de localidades de estaciones de MiBici por bloques.

*Igual que en la primer gráfica de barras, se observa que la mayoría de localidades las preserva la zona del **polígono central** (rojo).*

La visualización de los datos como una serie de tiempo es esencial para poder hacer un pronostico del crecimiento de nuestras localidates en conjunto.

Las columna que usaremos para hacer la gráfica de la serie sería 'inicio_del_viaje' que contiene la fecha de cada viaje realizado, además es necesario crear una columna nueva.

Dado que hay multiples viajes diarios, se opto por hacer un acumulado de viajes por día, por lo que la fecha sufrirá una transformación del tipo 'YYYY-mm-dd', que representan el año, mes y día. Las variables o columnas finales para la serie de tiempo ahora con los formatos nuevos serán ***'inicio_del_viaje'*** *y* ***'num_viajes'*** *con sus respectivos tipos de datos, 'date_time' e 'int64'.*

- *inicio_del_viaje (Fecha, YYY-mm-dd)*
- *num_viajes (numero de viajes por día)*

La serie de tiempo graficada para los viajes diarios seria Figura 2.6

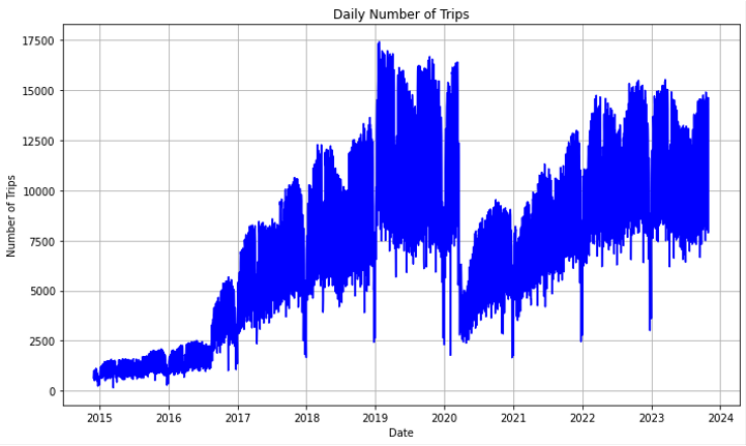

Figura 2.6: Serie de tiempo para la frecuencia de viajes día.

Notablemente hay un crecimiento desde el inicio del programa hasta el año 2020, en donde notamos que hay una bajada, hay que recordar que este año fue el año de pandemia en donde todos los servicios fueron detenidos para hacer confinamiento en casa.

Por otro lado, se puede notar una estacionalidad a finales de cada año, hay una caída, quizás por la temporada de frio o quizás por que las actividades comienzan a disminuir también en esa época. Por otro lado, al comienzo del año, nuevamente comienza a incrementar el uso del servicio.

Ahora bien, se puede hacer una mejor visualización de la gráfica, haciendo una agrupación de datos por mes, para lo cuál aplicamos a nuestro dataset una periodicidad con el parámetro 'M' para mensual y agrupamos los valores del

mes en una columna de valores nueva, el resultado se puede visualizar mejor en la Figura 2.7.

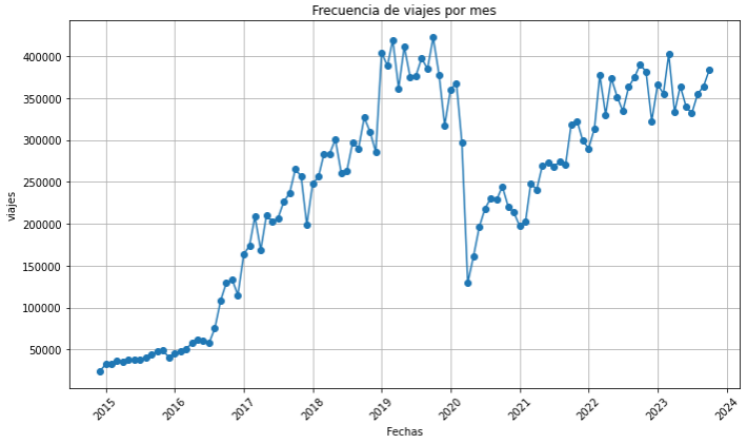

Figura 2.7: Serie de tiempo para la frecuencia de viajes por mes.

Por medio de la gráfica anterior se nota cierta tendencia, la cual es interrumpida en el tiempo por la pandemia a los inicios del año 2020.

Para comenzar a elegir un modelo que nos pueda ayudar con el pronostico del uso del transporte, se decidió solo trabajar con los datos hasta antes de la pandemia, ya que como fue explicado anteriormente, es un evento inusual y por lo tanto se decidió descartar. Se pudiera trabajar datos antes y después de la pandemia, en este trabajo se decidió trabajar con los datos antes ya que se cuenta con mayor cantidad de datos.

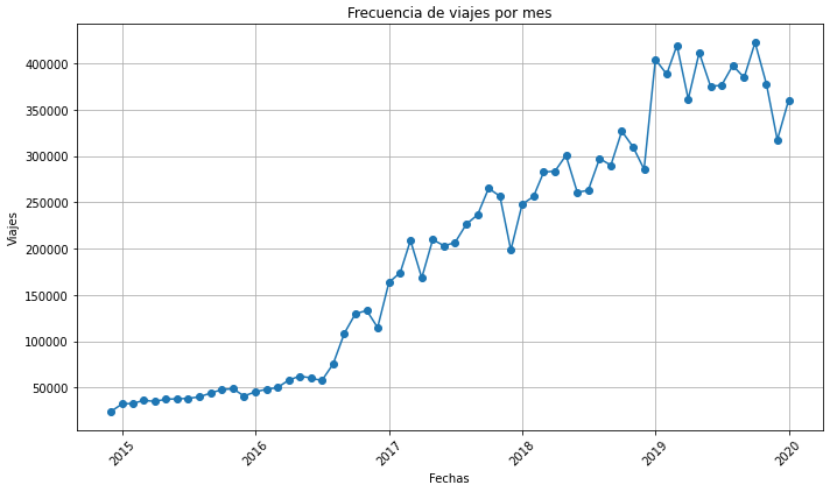

Figura 2.8: Serie de tiempo para la frecuencia de viajes por mes hasta antes de la pandemia.

Ahora se ven en la Figura 2.8 solamente los datos hasta antes de la pandemia, con una tendecia claramente creciente.

De la gráfica anterior se puede inferir que por el mes de Marzo del 2020 hubo una caída inminete pero nunca llego a cero, comenzando a recuperarse poco a poco a su uso normal.

Descomposicón de la serie

*Descomponer la serie en sus diferentes componentes como la **tendencia**, **estacionalidad** y **residuos** ayuda a visualizar con que tipo de serie estamos tratando para poder elegir a un modelo.*

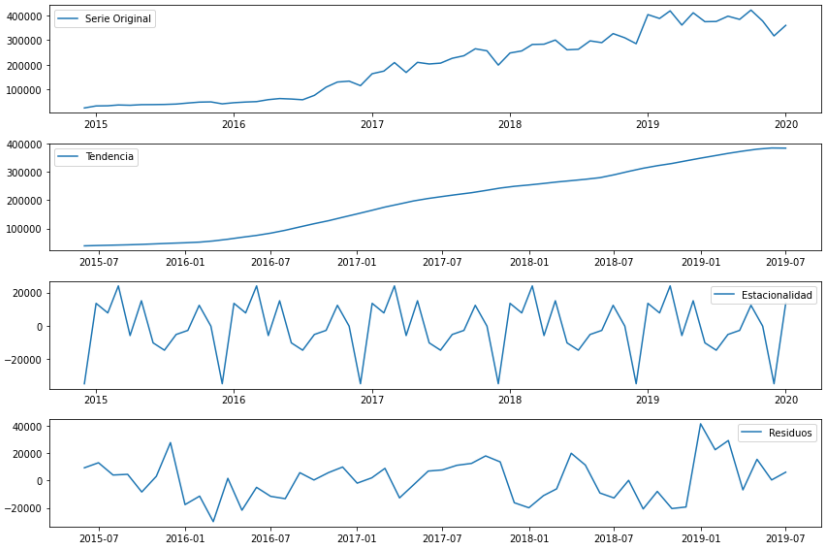

Figura 2.9: Descomposición de la serie en sus diferentes componentes.

*Se observa notablemente una tendecia **positiva** la cual era muy notable desde nuestra gráfica original. La estacionalidad no era tan obvia, pero con la descomposición se puede visualizar mejor. Se puede observar que existe una estacionalidad y se repite cada **año**, y sucede poco antes de que termine el año, posiblemente porque las épocas navideñas, ya sea por lo difícil de manejar en esas épocas de tráfico o por las vacaciones o por el frio posiblemente. Los residuos son las fluctuaciones que tienen los datos a través del tiempo.*

Estas características mencionadas anteriormente, son características que se pudieran ajustar a un modelo estacional, como lo es el SARIMA.

2.3 Descripción de los modelos

SARIMA

*'El modelo **SARIMA** es una extensión del modelo **ARIMA**, que se utiliza frecuentemente cuando sospechamos que un modelo puede tener un efecto estacional' [7].*

'El modelo SARIMA se basa en la aplicación de los modelos ARMA a una serie temporal transformada donde se ha eliminado el comportamiento estacional y no estacionario, para una serie estacional con s periodos por año se utilizan los modelos' [8].

$$SARIMA(p,d,q)x(P,D,Q)s$$

Los elementos de tendencia

'Los elementos de tendencia son identificados por:

- *p: Que es la autorregresión.*
- *d: Orden de diferencia.*
- *q: Promedio móvil.*

Los elementos Estacionales

Los elementos estacionales para un ARIMA que deben configurarse son:

- *P: Autoregresivo estacional.*
- *D: Diferencia estacional.*
- *Q: Media móvil estacional.*
- *m: El número de pasos de tiempo para un solo período estacional.*

Juntos, la notación para un modelo SARIMA se especifica como:

$$SARIMA(p,d,q)(P,D,Q)m$$

Ejemplo:

$$SARIMA(3,1,0)(1,1,0)12$$

Es importante destacar que el parámetro m influye en los parámetros P, D y Q. Por ejemplo, un valor de m de 12 para datos mensuales sugiere un ciclo estacional anual. Un $P = 1$ haría uso de la primera observación compensada

estacionalmente en el modelo, por ejemplo, $t - (m \times 1)$ o $t - 12$. Un $P = 2$ usaría las últimas dos observaciones compensadas estacionalmente $t - (m \times 1)$, $t - (m \times 2)$. De manera similar, un D de 1 calcularía una diferencia estacional de primer orden y un $Q = 1$ utilizaría errores de primer orden en el modelo (por ejemplo, media móvil)'[9].

ARIMA

Tomado del sitio web, 'El modelo ARIMA es una clase de modelos lineales que utiliza valores históricos para predecir valores futuros. ARIMA significa Autoregressive Integrated Moving Average, cada técnica de los cuales contribuye a la predicción final. Vamos a entenderlo uno por uno.

Autorregresivo (AR)

En un modelo de autorregresión, pronosticamos la variable de interés utilizando una combinación lineal de valores pasados de esa variable. El término autorregresión indica que es una regresión de la variable contra sí misma. Es decir, utilizamos valores rezagados de la variable objetivo como nuestras variables de entrada para predecir valores futuros. Un modelo de autorregresión de orden p se verá así:

$$m_t = \beta_0 + \beta_1 m_{t-1} + \beta_2 m_{t-2} + \beta_3 m_{t-3} + \ldots + \beta_p m_{t-p}$$

En la ecuación anterior, el valor observado actual de m es una función lineal de sus últimos p valores. Los coeficientes de regresión $\beta_0, \beta_1, \ldots, \beta_p$ se determinan después del entrenamiento. Hay algunos métodos estándar para determinar los valores óptimos de p, uno de los cuales es analizar los gráficos de la función de autocorrelación y autocorrelación parcial.

La función de autocorrelación (ACF) es la correlación entre los valores actuales y pasados de la misma variable. También considera el efecto translaticio que los valores llevan consigo en el tiempo, aparte de un efecto directo. Por ejemplo, los precios del petróleo hace 2 días afectarán los precios de hace 1 día y, eventualmente, los de hoy. Pero los precios del petróleo hace 2 días también podrían tener un efecto en el día de hoy, que mide ACF.

Por otro lado, la autocorrelación parcial (PACF) solo mide la correlación directa entre los valores pasados y los valores actuales. Por ejemplo, PACF solo medirá el efecto de los precios del petróleo hace 2 días en el día de hoy sin efecto translaticio.

Los gráficos de ACF y PACF nos ayudan a determinar la dependencia del valor pasado, lo que a su vez nos ayuda a deducir p en AR. Consulte aquí para comprender cómo deducir los valores de p (AR) y q (MA) en profundidad.' [10]

En otras palabras, el modelo ARIMA es un tipo de modelo de series de tiempo que se utiliza para hacer un pronostico basado en datos históricos y tendencias pasadas, pero solo funciona para datos sin componentes estacionales. Por otro lado esta el modelo SARIMA que es una extensió del ARIMA sin embargo esta diseñado para trabajar con las componentes estacionales.

2.4 Descripción de las métricas

2.5 Definición

*'El **RMSE** representa la raíz cuadrada de las diferencias promedio al cuadrado entre los resultados predichos y observados. Es una métrica utilizada predominantemente en análisis de regresión y pronóstico, donde la precisión es significativa. Cuanto menor sea el RMSE, mejor será la capacidad del modelo para predecir con precisión. Por el contrario, un RMSE más alto indica una mayor discrepancia entre los resultados predichos y reales.*

Fórmula del RMSE: El Pilar del Cálculo

Cuando se trata del RMSE, todo comienza con la fórmula, la representación matemática que da vida a este concepto. La fórmula del RMSE es elegante y sencilla:

$$RMSE = \sqrt{\left(\frac{\sum(P_i - O_i)^2}{n}\right)}$$

Aquí, P_i denota el valor predicho, O_i representa el valor observado y n es el número total de observaciones o puntos de datos. La suma de las diferencias al cuadrado entre los valores predichos y observados se divide por el número de observaciones, y se toma la raíz cuadrada del resultado para obtener el RMSE. Este cálculo sirve como medida de las diferencias entre los valores predichos por un modelo y los valores observados en la realidad.

Cálculo del RMSE

Al desglosar el cálculo del RMSE, encontramos que el proceso es metódico y sistemático. Inicialmente, se calcula la diferencia entre el valor observado y predicho para cada punto de datos. Esta diferencia, conocida como residuo, se eleva al cuadrado. Los residuos al cuadrado luego se suman para obtener una cifra acumulativa, que se divide por el número de puntos de datos para dar el error cuadrático medio (MSE). Finalmente, se calcula la raíz cuadrada del MSE, lo que resulta en el RMSE.

Esta secuencia de operaciones asegura que los errores más grandes tengan un impacto desproporcionadamente mayor en el RMSE, lo que lo hace sensible a los valores atípicos. Por lo tanto, es una medida robusta cuando los errores sustanciales son particularmente indeseables' [11].

Error Absoluto Medio (MAE)

'El MAE se define como el promedio de la diferencia absoluta entre los valores pronosticados y los valores reales.

Ecuación del error absoluto medio

$$MAE = \frac{\sum |y' - y|}{n}$$

Donde y' es el valor pronosticado y y es el valor real. n es el número total de valores en el conjunto de prueba. El MAE nos dice qué tan grande puede ser el error que podemos esperar del pronóstico en promedio. Los valores de error están en las unidades originales de los valores pronosticados y $MAE = 0$ indica que no hay error en los valores pronosticados.

Cuanto menor sea el valor de MAE, mejor será el modelo; un valor de cero significa que no hay error en el pronóstico. En otras palabras, al comparar múltiples modelos, se considera mejor el modelo con el MAE más bajo.

Sin embargo, el MAE no indica el tamaño relativo del error y puede resultar difícil diferenciar errores grandes de errores pequeños. Se puede utilizar con otras métricas (ver Error Cuadrático Medio Raíz a continuación) para determinar si los errores son más grandes. Además, el MAE puede pasar por alto problemas relacionados con el bajo volumen de datos; ver las dos últimas métricas en este artículo para abordar ese problema' [12].

'Aunque existen múltiples métricas, cada una proporciona información específica que puede o no ser adecuada para su caso de uso específico. Esto significa que se debe elegir una métrica en función del caso de uso y de la comprensión de los datos involucrados en la realización de las predicciones.

Resumen de métricas de error para pronósticos:

El MAE es útil cuando se necesita medir el error absoluto. Es fácil de entender pero no es eficiente cuando los datos tienen valores extremos.

El RMSE (NRMSE) también es útil cuando la dispersión es importante y se necesitan penalizar valores más grandes. El RMSE es más fácil de interpretar en comparación con el MSE porque el valor del RMSE está en la misma escala que los valores pronosticados' [12].

Estas son las dos métricas que se usaron en los modelos, pero en la que se enfoco más fue en el rmse ya que es un error con mejor precisión y el los datos que se usaron se comportan mejor para usar un rmse

.

La Tabla 2.1 es una comparativa de los errores utilizados en los diferentes

modelos con las diferentes pruebas que se hicieron.

Modelo	RMSE	MAE	Media	Hetero.
ARIMA (Log)	0.9795	0.9360	11.3007	3.49
ARIMA	114718	105294	199237	47.09
SARIMA	29006	22568	199237	2.31
SARIMA (Log)	0.1281	0.1053	11.3007	0.35
SARIMA (Log) Post-Split	0.7356	0.6254	11.3007	1.55

Tabla 2.1: Comparación de métricas de modelos

En la tabla Tabla 2.1, se representan los resultados de todos los experimentos que se hicieron en los diferentes modelos usados, en la cuál podemos observar que tuvieron en cuenta diferentes métricas, como el MAE y el RMSE. Como ya lo hemos descrito anteriormente estas medidas son utilizadas para evaluar el rendimiento de los modelos de pronóstico en términos de la precisión.

Se puede observar que el RMSE y MAE del modelo ***SARIMA con Transformación Logarítmica*** *resulta con los valores de predicción más bajos, con un RMSE de* ***0.1281*** *y un MAE de* ***0.1053****, por lo cual se puede concluir hasta el momento que este modelo puede resultar el mejor de entre los modelos evaluados.*

La ***heterocedasticidad*** *es una métrica que nos indica la varianza de los datos a lo largo del tiempo. Es por eso que se consideró esta medida también, ya que la presencia de heterocedasticidad amplia puede influir en los pronósticos de nuestro modelo. Por lo tanto, una heterocedasticidad menor daría un pronóstico más fiable. Al aplicar una transformación logarítmica, logramos reducir esta métrica.*

Es importante la media de los datos ya que puede proporcionar información acerca de las dimensionalidad, ya que se esta trabajando con transformaciones de datos y así nos podemos dar cuenta de que tan desproporcionados quedan nuestros datos antes y después de aplicar dicha transformación.

2.6 Descripción de los experimentos o simulaciones

Elección del modelo

Independientemene de que modelo se elegirá, se deben visualizar las medidas de autocorrelación explicadas anteriormente.

La Figura 2.10 y la Figura 2.11 muestran las gráficas de autocorrelación para la serie de tiempo de MiBici.

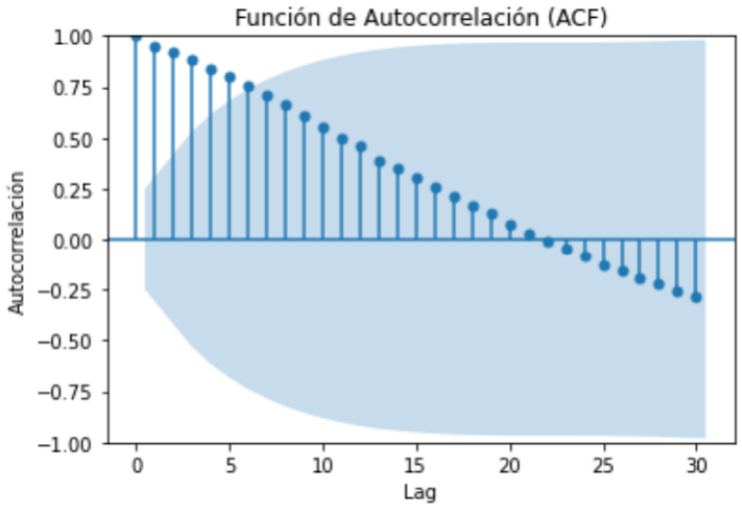

Figura 2.10: Autocorrelación ACF.

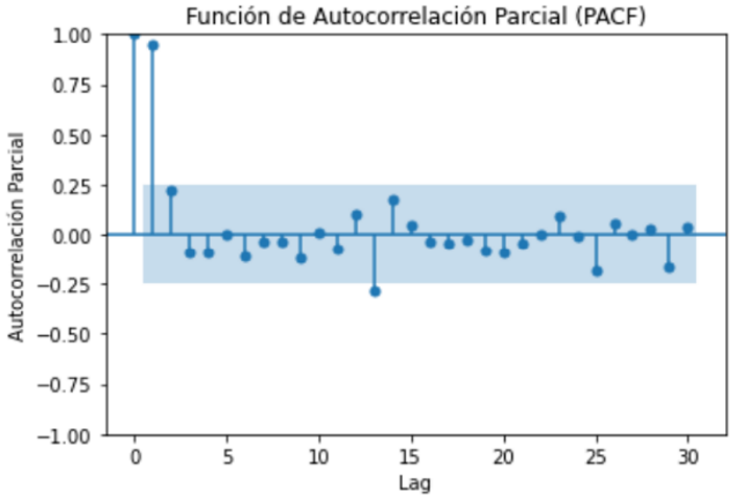

Figura 2.11: Autocorrelación PACF.

Para calcular los valores p,d,q se itero por medio de una función probado diferenes valores, obteniendo al final los mejores valores para usarlos en el modelo.

Modelo ARIMA

Dado que la descomposición de los datos en sus difentes componentes habia mostrado una cierta estacionalidad, se decidio hacer una prueba del modelo ARIMA para comparar sus resultados con el SARIMA.

El 'split' de los datos se probó en diferentes proporciones, esto debido a que el error cuadratico medio 'rmse' mostraba mejoras al hacer este tipo de prubeas en el split.

*La librería que se uso para entrenar el modelo es parte de la librería **statmodels** [13] más especificamente 'statsmodels.tsa.arima.model' y para medir el error, la metrica que se uso es 'rmse' como se descrbio ya anteriormene, esto con la paquetería sklearn.metrics de **sklearn** [14].*

Se corre el modelo, buscando los mejores parametros p, d, q, obteniendo los siguientes resultados:

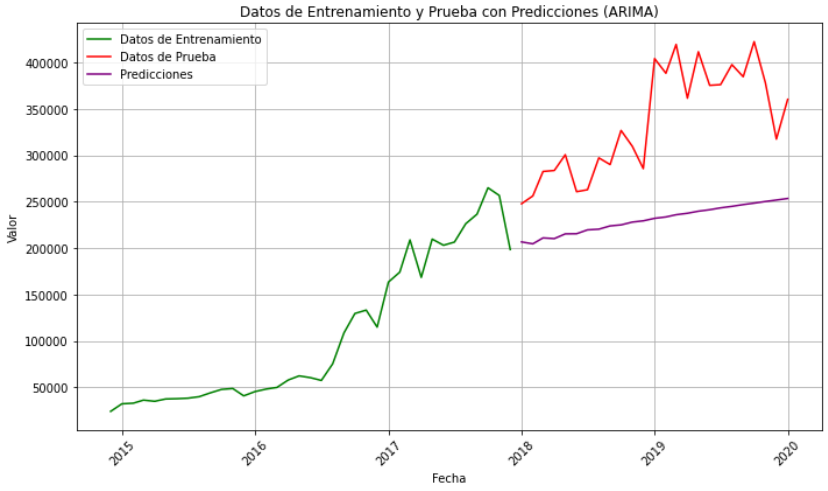

Figura 2.12: Gráfica con predicción para modelo ARIMA.

La Tabla 2.2 muestra los valores p, d,q probados:

p	d	q
0	0	0
0	0	1
0	0	2
0	1	0
0	1	1
0	1	2
0	2	0
0	2	1
0	2	2
1	0	0
1	0	1
1	0	2
1	1	0
1	1	1
1	1	2
1	2	0
1	2	1
1	2	2
2	0	0
2	0	1
2	0	2
2	1	0
2	1	1
2	1	2
2	2	0
2	2	1
2	2	2

Tabla 2.2: Valores p, d y q probados

*La función **get_forecast** nos da el pronostico con una muestra n = 30, la cual representan días de pronostico. Además en la misma Figura 2.13 se muestra el intervalo de confianza para el pronostico hecho, esta es calculada mediante la función **forecast.conf_int()**, por defecto esta función calcula un intervalo de confianza del 95 %, pero se puede configurar el porcentaje de intervalo de confianza con el parametro alpha [15].*

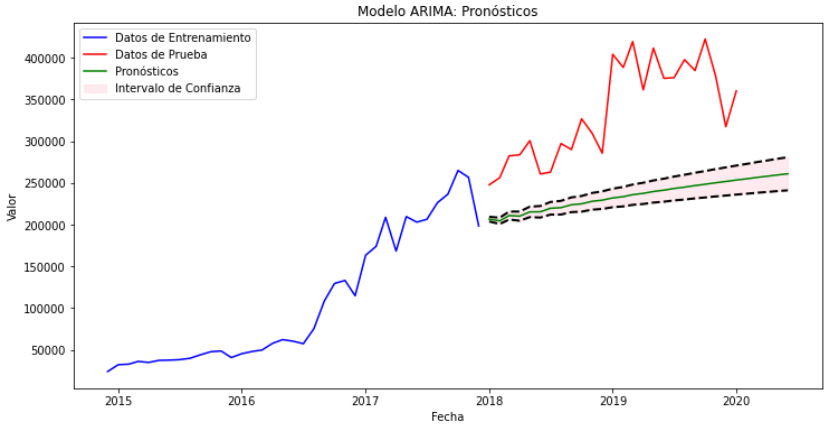

Figura 2.13: Gráfica de forecast 95 % para modelo ARIMA.

*El resumen de las estadísticas que se muestra a continuación, resaltando el valor de la heterocedasticidad de **47.09** muestra que hay una varianza en los datos la cuál debe de ser tratada por alguna técnica como la transformación de datos por medio de algún logaritmo, raíz o transformación de box-cox.*

```
RMSE en el conjunto de prueba: 114718.37841412629
                               SARIMAX Results
==============================================================================
Dep. Variable:      num_de_viajes_mes   No. Observations:                   37
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -407.111
Date:                Wed, 17 Apr 2024   AIC                            824.222
Time:                        16:41:29   BIC                            832.139
Sample:                  12-01-2014   HQIC                           826.985
                       - 12-01-2017
Covariance Type:                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1456      0.555      0.262      0.793      -0.943       1.234
ar.L2          0.8175      0.792      1.032      0.302      -0.735       2.370
ma.L1         -0.2272      0.735     -0.309      0.757      -1.669       1.214
ma.L2         -0.6854      1.016     -0.675      0.500      -2.676       1.305
sigma2      4.889e+08   9.67e-09   5.06e+16      0.000    4.89e+08    4.89e+08
===================================================================================
Ljung-Box (L1) (Q):                   0.23   Jarque-Bera (JB):                19.71
Prob(Q):                              0.63   Prob(JB):                         0.00
Heteroskedasticity (H):              47.09   Skew:                            -0.96
Prob(H) (two-sided):                  0.00   Kurtosis:                         6.08
===================================================================================
```

Figura 2.14: Resumen de datos estadisticos del modelo ARIMA.

Aplicando una transformación logarítmica se obtiene menos heterocedasticidad pero la gráfica luce diferente, se observa que el pronostico tiene una tendencia a la baja, lo contrario al anterior que la tendencia era a la alta.

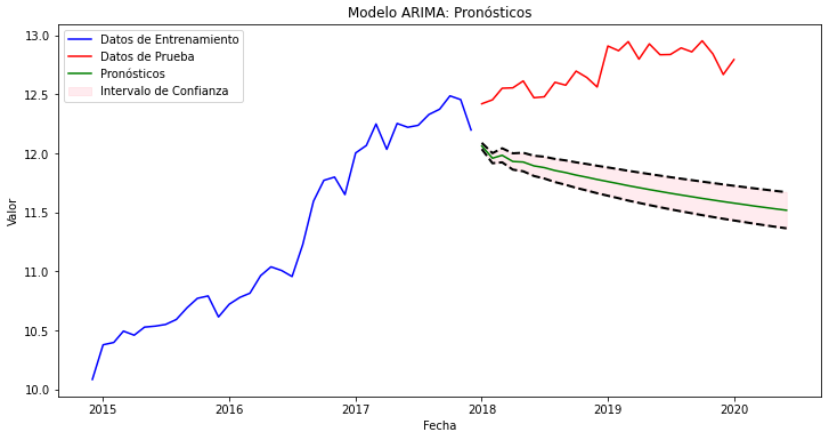

Figura 2.15: Gráfica de forecast 95 % para modelo ARIMA con logaritmo aplicado.

El nuevo resumen de las estadísticas con el logaritmo aplicado se puede observar diferentes valores de p, d, q, con los datos transformados se usan valores de 2, 0, 2, *en comparativa con el pasado que se uso* 2, 1, 2. *El rmse no puede compararse entre los modelos ya que es una escala diferente.*

RMSE en el conjunto de prueba: 0.9795354370268312

SARIMAX Results

Dep. Variable:	num_de_viajes_mes	No. Observations:	37
Model:	ARIMA(2, 0, 2)	Log Likelihood	18.081
Date:	Wed, 17 Apr 2024	AIC	-24.163
Time:	19:54:10	BIC	-14.497
Sample:	12-01-2014	HQIC	-20.755
	- 12-01-2017		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	11.1038	0.725	15.314	0.000	9.683	12.525
ar.L1	0.3661	0.264	1.387	0.165	-0.151	0.883
ar.L2	0.5909	0.238	2.481	0.013	0.124	1.058
ma.L1	0.8996	0.274	3.279	0.001	0.362	1.437
ma.L2	0.5143	0.185	2.786	0.005	0.152	0.876
sigma2	0.0196	0.004	4.816	0.000	0.012	0.028

Ljung-Box (L1) (Q):	1.56	Jarque-Bera (JB):	1.42
Prob(Q):	0.21	Prob(JB):	0.49
Heteroskedasticity (H):	3.49	Skew:	0.44
Prob(H) (two-sided):	0.04	Kurtosis:	3.38

Figura 2.16: GResumen de datos estadisticos del modelo ARIMA con logaritmo aplicado.

Modelo SARIMA

Con el modelo SARIMA tenemos un enfoque similar, pero aquí si se considera la estacionalidad con los parametros P, D, Q y m.

De igual manera se considera la librería de ***sklearn*** *para las metricas y* ***statmodels*** *para entrenar al modelo pero con una pequeña variación usando la paquetería statsmodels.tsa.statespace.sarimax [16].*

En la Figura 2.17 se muestra claramente un mejor ajuste de las predicciones a nuestros datos, tanto así que se ve un poco ajustada.

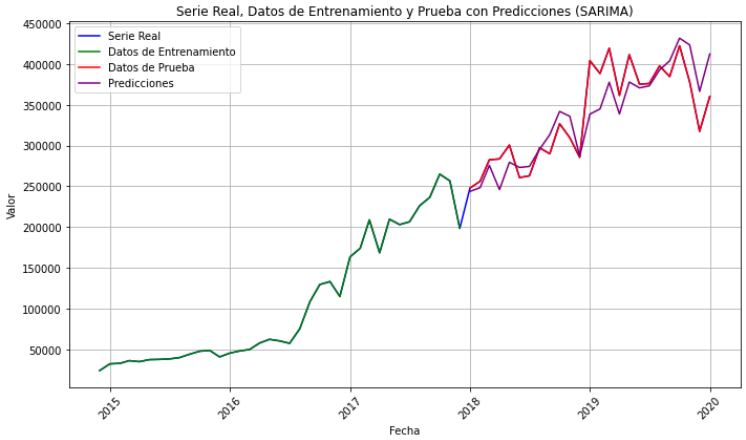

Figura 2.17: Gráfica con predicción para modelo SARIMA.

De igual manera el forecast Figura 2.18 con un intervalo de confianza del 95 % se muestra ajustado.

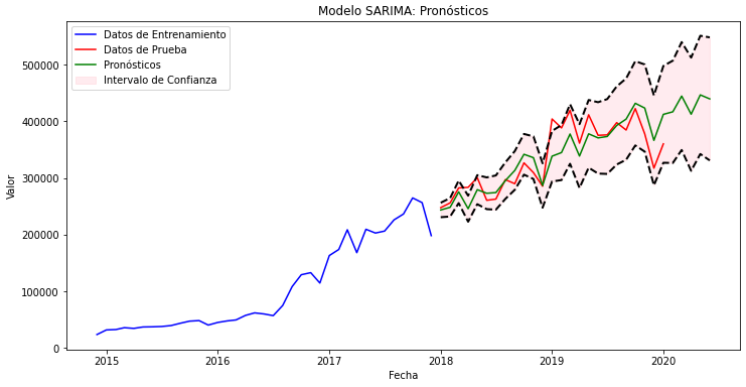

Figura 2.18: Gráfica de forecast 95 % para modelo SARIMA.

El resumen de las estadisticas del modelo SARIMA, da un nuevo valor de la heterocedasticidad de ***2.31*** *muestra que hay my poca varianza en los datos, de igualmanera puede ser tratada por medio de una transformación de datos como un logaritmo, raíz o transformación de box-cox. El rmse del conjunto de train con valor a* ***29,006.9351*** *esta en una escala entre 24,7754 y 42,2473, por lo tanto es un error esta más cercano a un error bajo que a un error alto.*

```
RMSE en el conjunto de prueba: 29006.93515716038
                                SARIMAX Results
==========================================================================================
Dep. Variable:                 num_de_viajes_mes   No. Observations:                   37
Model:             SARIMAX(1, 0, 2)x(1, 1, [1], 12)   Log Likelihood                -284.629
Date:                              Wed, 17 Apr 2024   AIC                            581.258
Time:                                      20:55:02   BIC                            588.571
Sample:                                12-01-2014   HQIC                           583.286
                                     - 12-01-2017
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9963      0.080     12.513      0.000       0.840       1.152
ma.L1         -0.2022      0.283     -0.713      0.476      -0.758       0.353
ma.L2          0.1072      0.473      0.226      0.821      -0.820       1.035
ar.S.L12      -0.8077      1.384     -0.584      0.559      -3.519       1.904
ma.S.L12       0.6150      1.886      0.326      0.744      -3.082       4.312
sigma2      5.877e+08   1.26e-08   4.67e+16      0.000    5.88e+08    5.88e+08
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                 0.69
Prob(Q):                              0.83   Prob(JB):                         0.71
Heteroskedasticity (H):               2.31   Skew:                            -0.39
Prob(H) (two-sided):                  0.26   Kurtosis:                         2.78
===================================================================================
```

Figura 2.19: Resumen de datos estadisticos del modelo SARIMA.

SARIMA con transformación (logaritmo)

Si aplicamos una transformación de un logaritmo los resultados son los siguientes:

Se alcanza a visualizar una suavización del crecimiento, posiblemente la tendencia es correcta, sin embargo, pudiera seguir habiendo sobre ajuste.

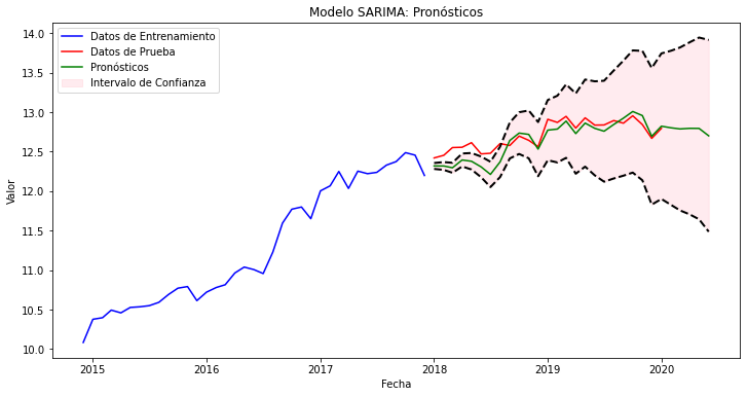

Figura 2.20: Gráfica de forecast 95 % para modelo SARIMA con logaritmo aplicado.

El resumen estadistico ahora muestra un rmse del conjunto de train entre o y 1 debido a la transformación logarítmica, un error de ***0.1281*** *muy cercano al cero.*

La sig Figura 2.22 muestra el comportamiento del error a lo largo de las

```
RMSE en el conjunto de prueba: 0.12812092148271736
                                SARIMAX Results
==========================================================================================
Dep. Variable:                  num_de_viajes_mes   No. Observations:                   37
Model:             SARIMAX(2, 2, 2)x(1, 1, [1], 12)   Log Likelihood                17.236
Date:                              Wed, 17 Apr 2024   AIC                          -20.472
Time:                                      20:19:03   BIC                          -12.524
Sample:                                12-01-2014   HQIC                         -18.473
                                     - 12-01-2017
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3428      1.191      0.288      0.773      -1.991       2.677
ar.L2          0.0221      0.661      0.033      0.973      -1.274       1.318
ma.L1         -1.7001      8.634     -0.197      0.844     -18.623      15.223
ma.L2          0.9654      9.526      0.101      0.919     -17.705      19.636
ar.S.L12      -0.7377      1.432     -0.515      0.607      -3.545       2.070
ma.S.L12      -0.6562     14.945     -0.044      0.965     -29.949      28.636
sigma2         0.0037      0.068      0.054      0.957      -0.130       0.137
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                 10.74
Prob(Q):                              0.90   Prob(JB):                          0.00
Heteroskedasticity (H):               0.35   Skew:                             -1.06
Prob(H) (two-sided):                  0.16   Kurtosis:                          5.60
===================================================================================
```

Figura 2.21: Resumen de datos estadisticos del modelo SARIMA con logaritmo aplicado.

pruebas que se hicieron con los diferentes valores p, d, q.

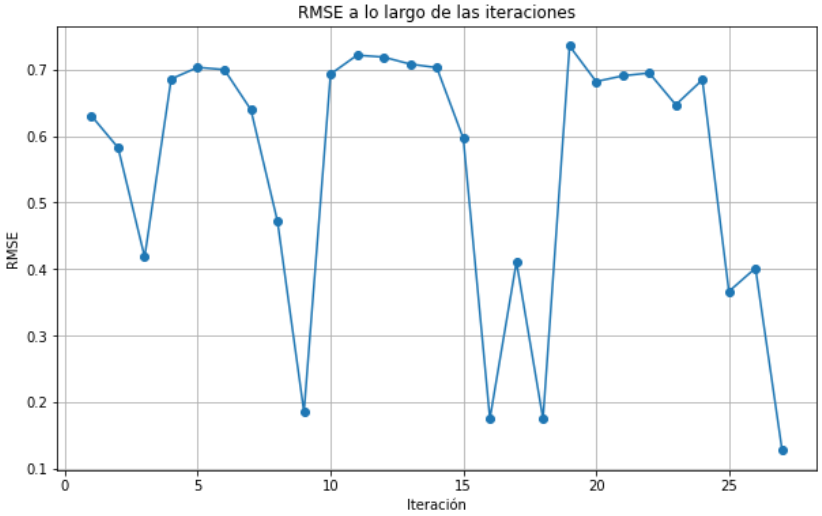

Figura 2.22: comportamiento de rmse del conjunto de test a lo largo de las iteraciones.

Se pueden observar que los errores tienden a subir y abajar dado que se hacen diferentes prubeas al mismo tiempo y cada una de ellas se gráfican hasta dar con la que da menor error.

SARIMA con transformación antes de hacer el split (logaritmo)

En un intento más por obtener un resultado, se realizó un experimento más, el cual consiste en hacer el split de los datos y después aplicar la transformación al training y al test por separado.

Esta tecnica, fué sugerida por mi mentor dado que yo no la conocía y pense que siempre las transformaciones se hacian antes de hacer un split de datos. Los resultados fueron los siguientes:

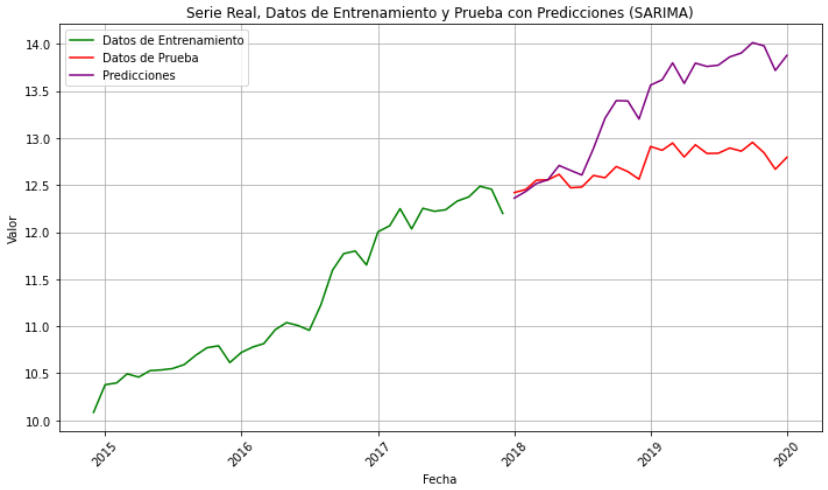

Figura 2.23: Gráfica con predicción para modelo SARIMA con transformación después del split de datos.

En comparación con hacer la transformación antes del split, aquí las predicciones hacen una tendencia creciente, que a través del timpo pareciera que no llega a una estabilidad.

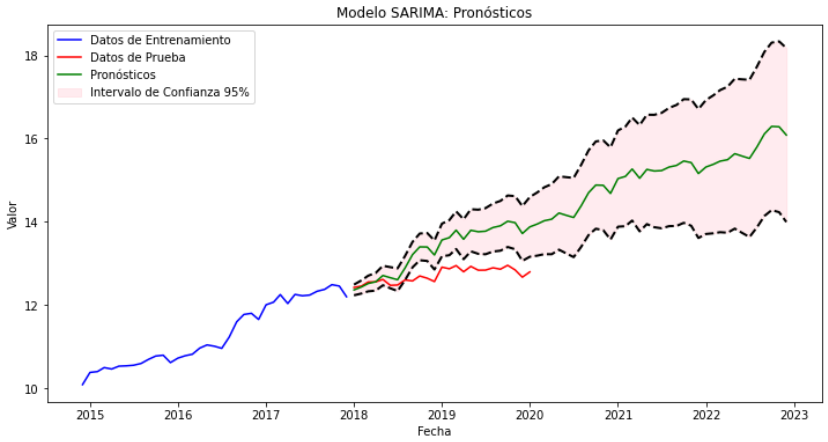

Figura 2.24: Gráfica de forecast 95 % para modelo SARIMA con transformación después del split de datos.

*Las medidas en esta ocación demuestran un poco más de varianza en los datos heterocedasticidad de **1.55** y un rmse del conjunto de train es de **0.73** comparado con la prueba anterior, esta prueba presenta mayor error en del modelo.*

```
RMSE en el conjunto de prueba: 0.7356255947898908
                                SARIMAX Results
==========================================================================================
Dep. Variable:                    num_de_viajes_mes   No. Observations:
Model:            SARIMAX(2, 0, 0)x(1, 1, [1], 12)   Log Likelihood
Date:                              Sat, 20 Apr 2024   AIC                              -
Time:                                     11:54:26   BIC                              -
Sample:                                12-01-2014   HQIC                             -
                                     - 12-01-2017
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.7212      0.222      3.248      0.001       0.286       1.156
ar.L2          0.2761      0.222      1.241      0.214      -0.160       0.712
ar.S.L12      -0.9994      1.512     -0.661      0.509      -3.962       1.964
ma.S.L12       0.9034    118.852      0.008      0.994    -232.042     233.848
sigma2         0.0032      0.370      0.009      0.993      -0.722       0.728
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 1.17
Prob(Q):                              0.96   Prob(JB):                         0.56
Heteroskedasticity (H):               1.55   Skew:                            -0.52
Prob(H) (two-sided):                  0.55   Kurtosis:                         3.16
```

Figura 2.25: Resumen de datos estadisticos del modelo SARIMA con transformación después del split de datos.

El rmse del conjunto de test a lo largo de las iteraciones es mostrado por la Figura 2.26

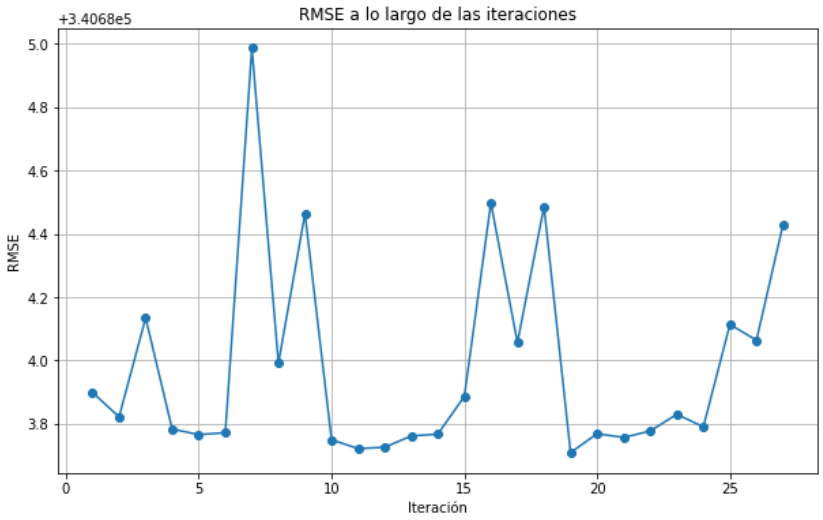

Figura 2.26: comportamiento de rmse del conjunto de test a lo largo de las iteraciones con transformación después del split de datos.

3 Resultados y discusión

3.1 Resultados y discusión

ARIMA vs SARIMA:

- *El modelo SARIMA con transformación logarítmica después del split de datos resulto ser el mejor, aún que el modelo con una transformación antes del split arroja mejores métricas, se quiere evitar un sobre ajuste, por lo tanto elegir un modelo mas holgado es una opción considerable.*

- *Al elegir el SARIMA sobre el ARIMA se comprueba que las componentes estacionales del modelo SARIMA funcionan adecuadamente de acuerdo al análisis previo al descomponer la serie.*

Transformación logarítmica:

- *La transformación logarítmica además de las componentes estacionales, mejoran la precisión de los modelos, esto se puede verificar en las métricas que RMSE, MAE y heterocedasticidad.*

Variación en los errores:

- *Se puede observar que la varianza de los errores (RMSE, MAE y heterocedasticidad) es muy significativa entre los modelos, inclusive entre el mismo modelo manejando una transformación se puede observar una variación significante.*

- *Es importante destacar el análisis de los datos previos al elegir un modelo (esto incluiría la descomposición de la serie) con eso podemos determinar un poco mejor que modelo sería el más adecuado para hacer un pronostico más acertado.*

En resumen, nuestros datos muestran una estacionalidad en la descomposición de la serie de tiempo. Dado que el modelo SARIMA es un modelo efectivo para estas cualidades, los resultados de las métricas (RMSE, MAE y heterocedasticidad) reafirmaron que es un modelo que se ajusta mejor a las condiciones presentadas en los datos.

La tendencia mostrada en las gráficas coincide con el pronostico hecho por el modelo, hay que recordar que la serie de tiempo se acoto hasta antes de la pandemia. Se cree que si este evento catastrófico no hubiera pasado, la tendencia sería positiva creciente y pareciera ser que se equilibraría (ligeramente continua) en un punto en el futuro.

Esto se traduce a un crecimiento de viajes y evidentemente de usuarios en el uso de este transporte. Una mala planeación al no tomar en cuenta los eventos socioculturales que pudieran llegar a presentarse en la zona en donde se tiene actividad de las rutas de MiBici pudiera afectar de manera significativa la gran cantidad de usuarios que se tiene actualmente y la que pronostica un crecimiento.

4 Conclusiones y trabajo futuro

4.1 Conclusiones

Guadalajara cuenta con una población en la zona metropolitana de 4.797 millones [17], 'La oferta de espectáculos artísticos del sector público, en el periodo 2014-2018, es asimétrica y muestra un considerable número de municipios ampliamente desfavorecidos con relación a su dimensión poblacional. Durante este periodo, sólo 1 % de los municipios tuvo más de dieciséis eventos por cada 10,000 habitantes, en contraste con 62 % de los municipios que tuvo entre uno y cinco eventos por cada diez mil habitantes (SCJ, 2018b).' [18]

Considerando el crecimiento poblacional en la zona metropolitana de Guadalajara y que el gobierno de Jalisco apuesta por este tipo de eventos socioculturales y sin tomar aún en cuenta los eventos deportivos que también son eventos masivos, el documento de transparencia del gobierno de Jalisco cita 'Con respecto a la infraestructura cultural, en Jalisco hay 283 bibliotecas, 136 casas de la cultura, 119 museos y 32 teatros (SCJ, 2018c). Sin embargo, se observa una distribución inequitativa y centralizada de los espacios para realización de actividades culturales.' [18], lo cual nos indica una falta de planeación en la distribución de los eventos culturales perse.

Ya identificado que este es un problema que puede afectar a las conglomeraciones a la hora de crear un evento; se aconseja al gobierno del estado de Jalisco tomar en cuenta esta población de 'bicicletistas' usuarios del transporte MiBici que si bien comenzó siendo un pequeño grupo de personas rodando, en los últimos años ha tomado una fuerza muy importante, tan importante que nuestro modelo lo representa en sus gráficas actuales y en su proyección a futuro.

Otra recomendación sería la de estar haciendo constantes proyecciones (mensuales) y digo mensuales ya que se tiene un dataset nuevo mensual en la pagina oficial de MiBici, así se podría visualizar que tan rápido es el crecimiento y que tan urgente es tomar en cuenta la planeación de estos eventos.

No solo se puede tomar en consideración el crecimiento en general, hasta

este momento se encuentra dividido por zonas y una de las propuestas de estudio que hago a futuro es la de analizar el crecimiento individual por zona, con algún mapa de calor e individualizar el modelo por cada sección, así el gobierno de Jalisco puede percatarse de que zonas 'bicicleteras' pudieran ser más afectas al no hacer una debida planeación.

Guadalajara siempre ha sido una ciudad de grandes eventos deportivos y socioculturales, ya sea para atraer turismo o por que es una ciudad grande o simplemente por ofrecer diversión a los jaliscienses. No se puede dejar de lado las afectaciones a conglomerados tan grandes y con proyección de crecimiento como lo esta siendo este sistema de transporte urbano.

Hoy por hoy, el uso de la bicicleta no es solo un fenómeno de moda mundial, cada vez el transportarse en bicicleta se esta convirtiendo en un modo de vida saludable, ecológico y de cierta manera rápido en las grandes metropolis globales.

4.2 Trabajo futuro

De este trabajo pueden resultar una serié de investigaciones a futuro, aquí se pueden enumerar algunos casos de análisis:

1. ***Análisis de Estaciones de Bicicletas**: Hoy en día se tienen identificadores por zona en donde estas ubicadas las estaciones con bicicletas, se pudiera hacer una análisis detallado por zona, para ubicar las zonas con mayor demanda de bicis.*

2. ***Visualizaciones de Calor**: Existen otras maneras para visualizar el impacto de la demanda como los mapas de calor, se pudiera delimitar las zonas con mayor actividad para comenzar un panorama inicial.*

3. ***Anticipar el abastecimiento de más bicicletas**: Este ejercicio también sirve para dar una idea del crecimiento a futuro mediante el pronostico. Una vez identificado si existe o no crecimiento, se puede hacer un plan de compra de bicicletas con anticipación.*

4. ***Exploración de Otros Modelos**: Además de experimentar con otros modelos para obtener quizás mejores métricas, se pueden utilizar otros modelos para hacer algún tipo de clasificación para identificar puntos geográficos con mayor densidad de usuarios por ejemplo.*

5. ***Aprovechar la información que arrojan las gráficas**: Se puede aprovechar la estacionalidad para planear mantenimiento preventivo a las bicicletas o a las estaciones en sí. Gracias a las gráficas obtenidas se puede comenzar a trabajar en una plan de mantenimientos, las horas pico, nos dicen claramente en que horarios los usuarios pueden ser más afectados.*

Bibliografía

[1] G. de Jalisco, "Mibici: el sistema público de bicicletas en guadalajara." https://www.territorio.mx/p/mibici-el-sistema-publico-de-bicicletas. [En línea; accedido el 26-marzo-2024].

[2] OpenAI's DALL-E, "Public Bicycle on City Street." Generated by OpenAI's Image Generator Tool, 2024.

[3] Gobierno de Jalisco, "Se cumplen siete años de historias con mibici." https://jalisco.gob.mx/es/prensa/noticias/136608. [En línea; accedido el 28-marzo-2024].

[4] Gobierno de Jalisco, "Jalisco en cifras." https://innovacioncultura.jalisco.gob.mx/contexto-cultural/. [En línea; accedido el 30-marzo-2024].

[5] "Página oficial de datos de mibici del gobierno del estado de jalisco." https://www.mibici.net/es/datos-abiertos/. [En línea; accedido el 1-abril-2024].

[6] N. Urrego, "Tratamiento de valores vacíos ii: Estrategias de imputación estadística (moda, mediana y media)." https://bit.ly/3QFbyZF. [En línea; accedido el 1-abril-2024].

[7] "Promedio movil autorregresivo integrado integrado estacional (sarima)." https://bit.ly/4buFJLe. [En línea; accedido el 04-abril-2024].

[8] C. Miranda Chinlli, "Modelos sarima en la predicción de series temporales." https://masteres.ugr.es/estadistica-aplicada/sites/master/moea/public/inline-files/TFM_MIRANDA_CHINLLI_CARLOS.pdf. [En línea; accedido el 04-abril-2024].

[9] J. Brownlee, "How to grid search sarima hyperparameters for time series forecasting in python." https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/. [En línea; accedido el 05-abril-2024].

[10] *"Modelo autorregresivo."* https://neptune.ai/blog/arima-sarima-real-world-time-series-forecasting-guide. *[En línea; accedido el 10-abril-2024].*

[11] *"Root mean square error (rmse) definition."* https://bit.ly/3y5pyp4. *[En línea; accedido el 26-abril-2024].*

[12] *"Mean absolute error definition."* https://bit.ly/3UzvNJr. *[En línea; accedido el 26-abril-20.*

[13] *"statsmodels arima model."* https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html. *[En línea; accedido el 11-abril-2024].*

[14] *"mean squared error."* https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html. *[En línea; accedido el 26-abril-2024].*

[15] *"Libreria oficial, statsmodels sarima results."* https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAXResults.conf_int.html. *[En línea; accedido el 10-abril-2024].*

[16] *"statsmodels sarimax model."* https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html. *[En línea; accedido el 26-abril-2024].*

[17] *"Población en el gdl."* https://bit.ly/3ye0TyM. *[En línea; accedido el 9-mayo-20].*

[18] *"Plan estatal de desarrollo en jalisco."* https://bit.ly/3wxeuAP. *[En línea; accedido el 9-mayo-202].*