



UNIVERSIDAD DE O'HIGGINS
ESCUELA DE INGENIERÍA
ING. CIVIL EN MODELAMIENTO MATEMÁTICO DE DATOS

Análisis de Series de Tiempo para la Predicción del Flujo de Ciclistas

Proyecto de datos I

Integrantes:

Gabriel Díaz, María Droguett, Ignacio Jiménez

Profesor:

Víctor Bucarey

1. Resumen ejecutivo

El proyecto se centra en el análisis del flujo de bicicletas en las calles principales de las localidades de Rancagua y Machalí, con el objetivo de proporcionar información valiosa para la toma de decisiones en materia de transporte urbano y promoción de estilos de vida saludables. Se recopilaron datos históricos sobre el flujo de bicicletas, los cuales fueron sometidos a un riguroso proceso de limpieza y posterior imputación de datos faltantes utilizando varias técnicas de imputación.

Se probaron diversos métodos de imputación, incluidos el uso del valor más frecuente, interpolación lineal, la media de toda la serie temporal y regresión lineal múltiple. Cada método fue evaluado en términos de su desempeño utilizando la Desviación Absoluta Media (MAD).

Los resultados obtenidos mostraron que la regresión lineal múltiple fue la técnica de imputación más adecuada, especialmente en las calles Alameda y Cabello, donde se lograron valores de MAD más bajos. Sin embargo, se observó una mayor variabilidad en los datos de las calles República de Chile y San Juan, lo que se reflejó en un MAD más elevado.

Adicionalmente, se implementaron modelos de predicción de series temporales para anticipar el flujo de bicicletas. Se utilizaron dos enfoques principales: el modelo Prophet y el modelo ARIMA.

El modelo Prophet, conocido por su capacidad para manejar datos con tendencias y estacionalidades, fue ajustado tanto con variables exógenas como sin ellas. Este modelo permitió capturar de manera efectiva las fluctuaciones diarias en el flujo de bicicletas, aunque su desempeño varió dependiendo de la calle y los regresores incluidos.

Por otro lado, el modelo ARIMA fue utilizado para analizar y predecir las series temporales del flujo de bicicletas. Se probaron diversas configuraciones de los parámetros para realizar las predicciones correspondientes. El modelo ARIMA mostró un buen rendimiento en general, pero al igual que con Prophet, se observó que la inclusión de variables exógenas como la temperatura mejoró significativamente la precisión de las predicciones.

En conclusión, el proyecto ofrece una metodología estructurada para la imputación de datos faltantes y la predicción del flujo de bicicletas en las localidades estudiadas. Los hallazgos tienen importantes implicaciones para la planificación urbana y la promoción del ciclismo como una alternativa de movilidad sostenible. La capacidad de imputar datos faltantes y predecir el flujo de bicicletas de manera precisa y confiable ayuda a generar un análisis más completo y preciso, lo que facilita la toma de decisiones informadas para mejorar la infraestructura ciclista y fomentar el uso del ciclismo en la comunidad.

2. Introducción

La bicicleta, como medio de transporte alternativo, desempeña un papel crucial en la movilidad urbana de estas localidades[9], no solo ofreciendo una opción sostenible y económica para desplazarse, sino también contribuyendo a la reducción de la congestión vehicular y a la promoción de estilos de vida más saludables. Donde nos enfocaremos en 4 principales calles **Alameda - Cabello - República de Chile con San Joaquín - San Juan con Escrivá de Balaguer**[8].



Figura 1: Mapa con ubicación de contadores.

Durante la creación y formación del proyecto se utilizaron las siguientes herramientas:

- Contadores de Bicicletas instalados por la municipalidad en las calles ya mencionadas.
- Lenguajes de Programación, principalmente Python, donde en Python principalmente utilizamos la librería Sklearn[10] y matplotlib[1].
- Estudios relacionados con imputación de datos ánomalos y modelos de predicción en específico Prophet y Arima.

En este proyecto, se propondrá desarrollar modelos de predicción sobre el flujo de bicicletas en estas calles, con el objetivo de proporcionar información útil para la toma de decisiones en materia de transporte urbano y promover un estilo de vida más saludable.

En este contexto, el desarrollo de un modelo predictivo basado en series de tiempo se presenta como una herramienta valiosa para comprender y anticipar las tendencias en el uso de bicicletas en la región. Este modelo permitirá a las autoridades gubernamentales y a los planificadores urbanos obtener información precisa y oportuna sobre la demanda de bicicletas en diferentes momentos y condiciones, lo que facilitará la toma de decisiones informadas sobre políticas públicas relacionadas con el transporte sostenible y la promoción del ciclismo como una opción viable de movilidad urbana.

Este informe está estructurado de la siguiente manera: en la sección siguiente, los estudios relacionados con el relleno de datos faltantes. Luego, presentaremos la metodología utilizada para abordarlos, seguida de los resultados obtenidos y finalmente, discutiremos las conclusiones y las próximas direcciones futuras de investigación.

2.1. Objetivo general y específicos

Desarrollar un modelo predictivo basado en series de tiempo para estimar la cantidad de bicicletas utilizadas en Rancagua y Machalí. El propósito de esta investigación es proporcionar información que pueda ser utilizada por las autoridades gubernamentales para comprender las tendencias en el uso de bicicletas y tomar decisiones informadas sobre políticas públicas relacionadas con el transporte sostenible y la promoción del ciclismo como medio de transporte en la ciudad, los objetivos serán los siguientes:

1. Recopilar datos históricos sobre la cantidad de bicicletas utilizadas en el área de interés y realizar un proceso de limpieza de datos para prepararlos adecuadamente para el análisis.

2. Realizar un análisis exploratorio de los datos recopilados con el fin de identificar tendencias y patrones relevantes en el uso de bicicletas a lo largo del tiempo.
3. Investigar diversos modelos de series de tiempo para seleccionar el más apropiado para predecir la cantidad de bicicletas utilizadas.
4. Elección de modelo para posteriormente, proceder a entrenar el modelo seleccionado utilizando una parte de los datos disponibles.
5. Evaluar el rendimiento del modelo entrenado utilizando métricas adecuadas para modelos de series de tiempo, como la desviación media absoluta (MAD) y el error de porcentaje absoluto medio (MAPE).
6. Realizar ajustes adicionales en el modelo según sea necesario para mejorar su rendimiento.

3. Metodología

3.1. Pre-procedimiento:

El preprocessamiento de datos es una etapa crucial en cualquier análisis de datos y modelado predictivo. Su importancia radica en garantizar la calidad y coherencia de los datos, lo cual es fundamental para obtener resultados precisos y fiables. A continuación, se detallan los pasos seguidos en el preprocessamiento, basados en el método de "Preprocesamiento de Datos" o "Preparación de Datos". Esta práctica es ampliamente reconocida y utilizada en el campo de la ciencia de datos, la minería de datos y el aprendizaje automático. [5]

- **Recopilación de datos:** Como primer instancia se recopilaron datos históricos (2017-2024) sobre el flujo de bicicletas en las calles mencionadas de Rancagua y Machalí, incluyendo información sobre la cantidad de bicicletas que transitan por las diferentes calles. Esta base de datos es denominada "**bike_rancagua.csv**", archivo que contiene 2575 observaciones, fue compartida por el académico Víctor Bucarey, el cual lo obtuvo de el Ministerio de Vivienda y Urbanismo (Minvu), en colaboración con Don Galvarino Carrasco.
- **Limpieza de datos:** Se realizó una limpieza de los datos para eliminar valores atípicos, datos incompletos o inconsistentes que podrían afectar la calidad del modelo predictivo. Esta etapa incluyó la identificación y corrección de errores en el dataset.

A continuación se puede observar la cantidad de datos faltantes en cada calle:

Calles	Datos	datos faltantes
Alameda	2509	66
Cabello	2003	572
República con San Joaquín	1386	1189
San Juan con Escrivá de Balaguer	1438	1137

- **Imputación de datos:** Dado que en la base de datos original cada calle tiene una cantidad de valores faltantes , donde la suma de todas forman un total de 2964 datos faltantes, se exploraron diferentes métodos de imputación para completarlos. Se probaron un total de cuatro métodos diferentes de relleno de datos.
 - 1) Imputación por valor más frecuente: En este método, los valores faltantes se reemplazan por el valor que ocurre con mayor frecuencia en la variable correspondiente, la moda.

- 2) Interpolación lineal: Es una técnica que se utiliza para estimar valores desconocidos dentro de un rango conocido a partir de puntos de datos conocidos. En este método, se asume que existe una relación lineal entre los puntos de datos conocidos, y se utiliza esta relación para predecir valores intermedios.
- 3) Imputación por valor promedio: En este método, los valores faltantes se sustituyen por el promedio aritmético de los valores observados en la variable correspondiente.
- 4) Regresión lineal múltiple: Técnica estadística que se utiliza para analizar la relación entre una variable dependiente y dos o más variables independientes. En contraste con la regresión lineal simple, que utiliza una sola variable independiente para predecir la variable dependiente, la regresión lineal múltiple considera múltiples variables independientes para hacer predicciones más precisas. En el contexto de llenar datos faltantes, la regresión lineal múltiple puede ser una herramienta útil cuando se dispone de varias variables independientes que están relacionadas con la variable dependiente y que pueden ayudar a estimar los valores faltantes de manera más precisa. El proceso de regresión lineal múltiple implica ajustar una línea (este hiperplano en dimensiones más altas) que mejor se ajuste a los datos observados. Esta línea se determina minimizando la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo.[12]

La ecuación general de la regresión lineal múltiple puede expresarse como:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$$

Donde:

- \hat{y} es el valor predicho para la observación i .
- $x_{i,j}$ es el valor de la variable predictora j para la observación i .
- β_0 es el intercepto.
- $\beta_1, \beta_2, \dots, \beta_p$ son los coeficientes de regresión para las variables predictoras x_1, x_2, \dots, x_p respectivamente.
- ε_i es el término de error asociado con la observación i .

Para cada una de las calles, se definirá un modelo de regresión en el cual se usen las otras calles como variables predictoras. Aquí se presentan los modelos para cada caso:

- Para la predicción de la variable Cabello:

$$Cabello_i = \beta_0 + \beta_1 \cdot Alameda_i + \beta_2 \cdot Rep\xedblica_i + \beta_3 \cdot San Juan_i + \varepsilon_i$$

- Para la predicción de la variable Alameda:

$$Alameda_i = \beta_0 + \beta_1 \cdot Cabello_i + \beta_2 \cdot Rep\xedblica_i + \beta_3 \cdot San Juan_i + \varepsilon_i$$

- Para la predicción de la variable Rep\xedblica:

$$Rep\xedblica_i = \beta_0 + \beta_1 \cdot Alameda_i + \beta_2 \cdot Cabello_i + \beta_3 \cdot San Juan_i + \varepsilon_i$$

- Para la predicción de la variable San Juan:

$$San Juan_i = \beta_0 + \beta_1 \cdot Alameda_i + \beta_2 \cdot Cabello_i + \beta_3 \cdot Rep\xedblica_i + \varepsilon_i$$

Una vez que los modelos están entrenados con los datos disponibles, se procede a utilizar estos modelos para predecir los valores faltantes en cada calle. El entrenamiento de los modelos implica ajustar los coeficientes $\beta_0, \beta_1, \beta_2$ y β_3 utilizando los datos completos. Esto se realiza para cada calle, asegurando que los modelos sean precisos y representativos.

En el caso donde faltan datos en las variables que se usan como predictores (por ejemplo, Alameda, República o San Juan para predecir Cabello), dichas observaciones se ignorarán (es decir, se eliminarán con la función de Python `dropna()`) del conjunto de datos utilizado para el entrenamiento del modelo. Esto se hace para asegurar que el modelo se entrena únicamente con observaciones completas y confiables.

Con los coeficientes de los modelos de regresión ya estimados, estos modelos se aplican a las observaciones con datos faltantes. Por ejemplo, si falta un valor en la variable Cabello, se utilizarán los valores disponibles de las variables Alameda, República y San Juan para predecir el valor faltante de Cabello. Este proceso se repite para cada observación y para cada variable con datos faltantes, utilizando los modelos correspondientes.

Este proceso permite completar el conjunto de datos original, llenando los espacios dejados por los datos faltantes. Al tener un conjunto de datos más completo, se puede realizar un análisis más profundo y detallado, ya que se dispone de más información y se reducen las incertidumbres asociadas con los datos incompletos.

- **Selección del método de relleno:** Después de probar los diferentes métodos de imputación, se evaluaron y compararon los resultados obtenidos utilizando métricas de desempeño relevantes, principalmente la Desviación Media Absoluta (MAD). Y luego, se seleccionó el método de regresión lineal múltiple como el más adecuado para imputar los datos faltantes en la base de datos.
- **Desarrollo del modelo de regresión lineal múltiple:** Se procedió a desarrollar un modelo de regresión lineal utilizando los datos imputados. Este modelo se entrenó utilizando sklearn de la librería de Python.
- **Validación del modelo:** Se validó el modelo de regresión lineal utilizando técnicas de validación cruzada. Se ajustaron los parámetros del modelo según fuera necesario para mejorar su capacidad predictiva y evitar el sobreajuste.
- **Evaluación del modelo:** Se evaluó el modelo final utilizando métricas de evaluación pertinentes, como la desviación media absoluta (MAD), para posteriormente determinar su precisión y fiabilidad en la predicción del flujo de bicicletas en Rancagua y Machalí.

3.2. Prophet

Prophet es una herramienta de código abierto desarrollada por Facebook para la previsión de series temporales. Es especialmente útil en escenarios con datos históricos diarios, semanales o mensuales y puede manejar series temporales con patrones estacionales y tendencias no lineales. [4]

Esta herramienta se basa en un modelo aditivo compuesto por varios componentes que describen la tendencia, la estacionalidad y los efectos de las vacaciones. La formulación matemática del modelo Prophet se puede expresar de la siguiente manera:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

Donde:

- $y(t)$ es el valor de la serie temporal en el tiempo t .
- $g(t)$ representa la tendencia, es decir, cómo cambia el valor medio de la serie temporal con el tiempo.
- $s(t)$ captura la estacionalidad, describiendo patrones repetitivos y periódicos (diarios, semanales, anuales, etc.).

- $h(t)$ modela los efectos de los días festivos y otros eventos especiales.
- ε_t es el término de error, que captura la variabilidad no explicada por el modelo.

3.3. ARIMA Y SARIMAX

ARIMA (AutoRegressive Integrated Moving Average) y SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors) son técnicas estadísticas de uso común y gran prestigio para la predicción de series temporales [2]. Ambos modelos se construyen a partir de tres elementos principales.

Aunque los modelos ARIMA son bien conocidos, los modelos SARIMAX amplían el marco de ARIMA al incluir patrones estacionales y variables exógenas.

En la notación de los modelos ARIMA y SARIMAX, los parámetros p , d , y q representan los componentes autorregresivos, de diferenciación y de media móvil, respectivamente. Por su parte, P , D , y Q representan los mismos componentes para la parte estacional del modelo, y m es el número de períodos en cada temporada.

- Donde p es el orden (número de lags temporales) de la parte autorregresiva del modelo, d es el grado de diferenciación (el número de veces que se han restado los valores consecutivos de la serie) y q es el tamaño de la media móvil del modelo.
- Donde P es el orden (número de lags temporales) de la parte estacional del modelo, D es el grado de diferenciación de la parte estacional del modelo y Q es el tamaño de la media móvil de la parte estacional del modelo.
- Donde m indica al número de períodos en cada temporada.

Cuando los términos P , D , Q , m son cero y no se incluyen variables exógenas, el modelo SARIMAX es equivalente a un ARIMA.

Los modelos ARIMA y SARIMAX[3] se construyen a partir de tres componentes principales:

- Componente autorregresivo (AR): Establece una relación entre el valor actual y sus valores anteriores (retardos).
- Componente de media móvil (MA): Considera que el error de predicción actual es una combinación lineal de errores de predicción anteriores.
- Componente integrado (I): Sugiere que los valores originales de la serie han sido reemplazados por las diferencias entre valores consecutivos, y este proceso de diferenciación puede repetirse varias veces.

La ecuación general para un modelo ARIMA(p, d, q) es:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Donde:

- Y_t es el valor de la serie en el tiempo t .
- c es una constante que funciona como un regularizador y representa la media de la serie temporal.

- $\phi_1, \phi_2, \dots, \phi_p$ son los coeficientes de la parte autorregresiva.
- $\theta_1, \theta_2, \dots, \theta_q$ son los coeficientes de la parte de media móvil.
- ε_t es el término de error en el tiempo t .

Para los modelos SARIMAX, la ecuación se expande para incluir los componentes estacionales en las variables exógenas X :

$$Y_t = c + \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \Phi_1 Y_{t-m} + \dots + \Phi_P Y_{t-mP} + \Theta_1 \varepsilon_{t-m} + \dots + \Theta_Q \varepsilon_{t-mQ} + \beta X_t + \varepsilon_t$$

Donde:

- $\Phi_1, \Phi_2, \dots, \Phi_P$ son los coeficientes de la parte autorregresiva estacional.
- $\Theta_1, \Theta_2, \dots, \Theta_Q$ son los coeficientes de la parte de media móvil estacional.
- X_t son las variables exógenas en el tiempo t .
- β son los coeficientes de las variables exógenas.

3.4. Variables Exógenas para ambos modelos

A continuación, se presenta el paso a paso para ajustar los modelos Prophet, ARIMA y SARIMAX utilizando Python y algunas de sus bibliotecas populares:

Sin variables exógenas:

1. Importar las bibliotecas necesarias, como pandas, Prophet (para el modelo Prophet), matplotlib, statsmodels (para los modelos ARIMA y SARIMA), etc.
2. Cargar los datos desde los archivos Excel correspondientes.
3. Renombrar las columnas según sea necesario para que coincidan con las convenciones de Prophet ('ds' para la fecha y 'y' para el valor a predecir).
4. Dividir los datos en conjuntos de entrenamiento y prueba.
5. Inicializar el modelo Prophet.
6. Ajustar el modelo Prophet sin variables exógenas.
7. Crear un DataFrame para las futuras fechas donde se harán las predicciones.
8. Realizar las predicciones utilizando el modelo ajustado.
9. Calcular las métricas de evaluación del modelo, como el Error Porcentual Absoluto Medio (MAPE) y la Desviación Absoluta Media (MAD).
10. Visualizar las últimas predicciones junto con los intervalos de confianza.
11. Visualizar los componentes del pronóstico, como tendencia y estacionalidad.
12. Ajustar el modelo ARIMA utilizando la función **auto_arima** de la biblioteca **pmdarima** para determinar automáticamente los mejores valores de los parámetros.

13. Hacer predicciones utilizando el modelo ajustado.
14. Calcular las métricas de evaluación del modelo, como la Desviación Absoluta Media (MAD) el Error Porcentual Absoluto Medio (MAPE), etc.
15. Visualizar los resultados y comparar las predicciones con los datos reales.

Con variables exógenas:

1. Cargar los datos con variables exógenas desde los archivos Excel correspondientes.
2. Renombrar las columnas según sea necesario para que coincidan con las convenciones de Prophet ('ds' para la fecha, 'y' para el valor a predecir y las columnas adicionales para las variables exógenas).
3. Inicializar el modelo Prophet y agregar las variables exógenas utilizando `model.add_regressor('nombre_de_la_variable')`.
4. Ajustar el modelo Prophet utilizando las variables exógenas proporcionadas.
5. Crear un DataFrame para las futuras fechas donde se harán las predicciones, asegurándose de incluir las variables exógenas correspondientes.
6. Realizar las predicciones utilizando el modelo ajustado.
7. Ajustar el modelo SARIMA utilizando la función `auto_arima` de la biblioteca `pmdarima`, incluyendo las variables exógenas en el argumento `exogenous`.
8. Hacer predicciones utilizando el modelo ajustado.
9. Calcular las métricas de evaluación del modelo, como la Desviación Absoluta Media (MAD) el Error Porcentual Absoluto Medio (MAPE), etc.
10. Visualizar los resultados y comparar las predicciones con los datos reales.

3.4.1. Definición de Variables

En el desarrollo del modelo de predicción de series temporales se utilizan las librerías Prophet y ARIMA (Autoregressive Integrated Moving Average). Se han considerado diversas variables para mejorar la precisión de las predicciones. A continuación, se describen las principales variables utilizadas y su tratamiento dentro de los modelos:

1. Variable Objetivo Prophet: La variable "*y*" representa las predicciones de la demanda o conteo de bicicletas. Los datos de esta variable fueron extraídos de las bases de datos `predicciones_cabello_alameda.xlsx` y `predicciones.xlsx`, luego renombrados a "*y*" para su compatibilidad con Prophet.
2. Variable Objetivo ARIMA: Similarmente, la variable "*y*" en ARIMA representa las predicciones del conteo de bicicletas. Los datos fueron extraídos de las mismas bases de datos mencionadas anteriormente.
3. Índice Temporal Prophet: La variable "*ds*" es el índice temporal que Prophet utiliza para las predicciones, representando las fechas correspondientes a cada observación de la variable objetivo. Esta columna se obtiene de las bases de datos `predicciones_cabello_alameda.xlsx` y `maximos_minimos.xlsx`.

4. **Índice Temporal ARIMA:** De manera similar, la variable temporal en ARIMA representando las fechas correspondientes a cada observación de la variable objetivo. Los datos se obtienen de las mismas bases de datos.
5. **Regresores Adicionales:** Para capturar mejor la variabilidad de la serie temporal y mejorar la precisión del modelo, se han incorporado regresores adicionales que representan condiciones externas.
 - **Temperatura Máxima:** Representa la temperatura máxima diaria, obtenida del archivo `maximos_minimos.xlsx`. Se utilizó para captar el efecto de las condiciones climáticas en la demanda de bicicletas, y se imputaron los valores faltantes.
 - **Temperatura Mínima:** Representa la temperatura mínima diaria, también obtenida del archivo `maximos_minimos.xlsx`, y fue tratada para imputar valores faltantes y eliminar outliers.

Las siguientes librerías de Python implementan modelos ARIMA-SARIMAX, que son herramientas estadísticas utilizadas para el análisis y la predicción de series temporales. Estas librerías proporcionan funciones y métodos para construir, ajustar y validar modelos ARIMA (AutoRegressive Integrated Moving Average) y SARIMAX (Seasonal ARIMA with Exogenous variables).

- **statsmodels:** Es una de las librerías más completas para modelado estadístico en Python. Su API (Interfaz de programación de aplicaciones) suele resultar más intuitiva para aquellos que provienen del entorno R que para aquellos acostumbrados a la API orientada a objetos de scikit-learn.
- **pmdarima:** Esta librería adapta el modelo SARIMAX de statsmodels a la API de scikit-learn, lo que permite a los usuarios familiarizados con las convenciones de scikit-learn sumergirse fácilmente en el modelado de series temporales.

3.5. Métricas para evaluación del modelo

El **MAPE (Error de Porcentaje Absoluto Medio)** es una medida comúnmente utilizada para evaluar la precisión de un modelo de pronóstico o predicción. Se calcula como el promedio del valor absoluto de los errores porcentuales entre las observaciones reales y las predicciones del modelo, expresado como un porcentaje[7].

La **Desviación Media Absoluta (MAD)** es una medida de dispersión que evalúa la magnitud promedio de las discrepancias absolutas entre los valores observados y los valores pronosticados en un conjunto de datos. Matemáticamente, se define como la media aritmética de las diferencias absolutas entre cada observación y su correspondiente valor pronosticado[6].

La fórmula para el cálculo del **MAPE** y **MAD** es:

$$\text{MAPE} = \frac{\sum_{t=1}^n \frac{|A_t - F_t|}{|A_t|} * 100}{n}$$

$$\text{MAD} = \frac{\sum_{i=1}^n |A_i - F_i|}{n}$$

Donde:

- n es el número total de observaciones.
- A_t es el valor observado en el tiempo t .

- F_t es el valor pronosticado en el tiempo t .
- $|A_t - F_t|$ denota el valor absoluto de $A_t - F_t$.

El **MAPE** se expresa como un porcentaje, lo que lo hace fácil de interpretar. Cuanto menor sea el valor del **MAPE**, mejor será la precisión del modelo de pronóstico o predicción. Sin embargo, se debe tener en cuenta que el **MAPE** tiene algunas limitaciones, especialmente cuando las observaciones reales son cercanas a cero, ya que puede conducir a una inestabilidad en el cálculo debido a la división por cero.

4. Resultados

4.1. Pre-procesamiento

Según se mencionó en la metodología, inicialmente se emplearán cuatro métodos: la media, la moda, la interpolación y la regresión lineal múltiple. A continuación se presentarán los resultados obtenidos mediante estos cuatro métodos.

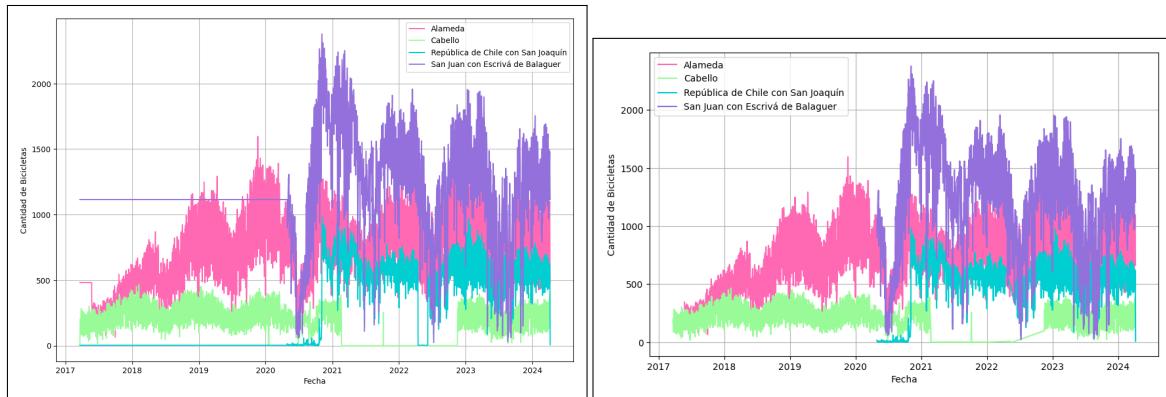


Figura 2: Imputación por valor más frecuente.

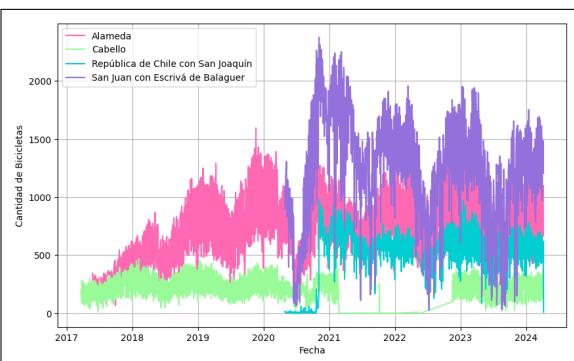


Figura 3: Interpolación lineal.

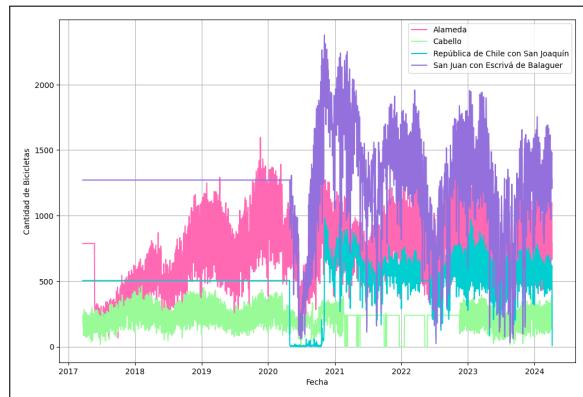


Figura 4: Imputación por valor promedio.

Luego aplicando el modelo seleccionado para la imputación de datos, en este caso la regresión lineal múltiple, obtenemos lo siguiente:

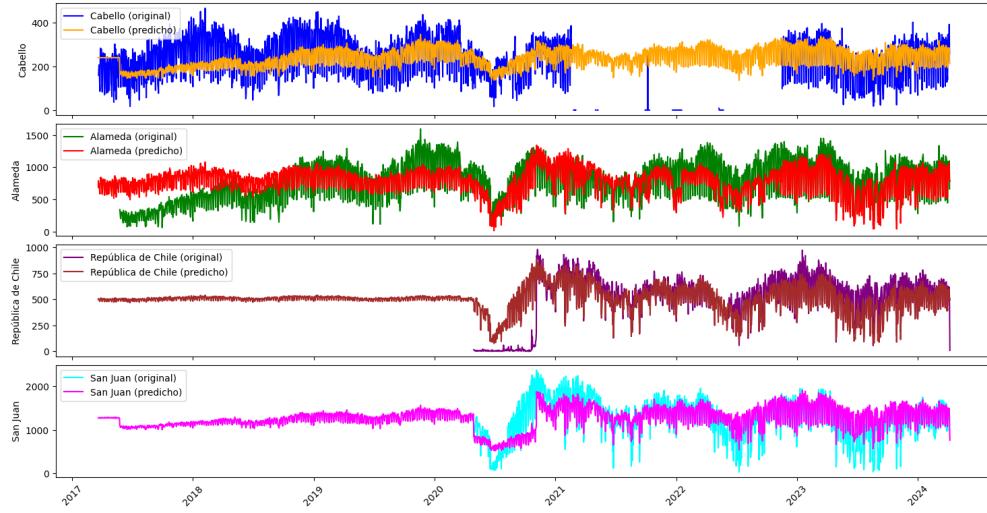


Figura 5: Regresión lineal múltiple.

Visualmente es posible notar que mejora de manera significativa la aproximación a comparación de los métodos anteriores, por lo cual creamos la comparación de todas las calles con el modelo aplicado, versus las calles sin el modelo.

Es evidente que en los sectores de República de Chile y San Juan con Escrivá de Balaguer existen numerosos datos faltantes desde 2017 hasta 2020. Por consiguiente, nos enfocaremos en visualizar la información a partir de este último año.

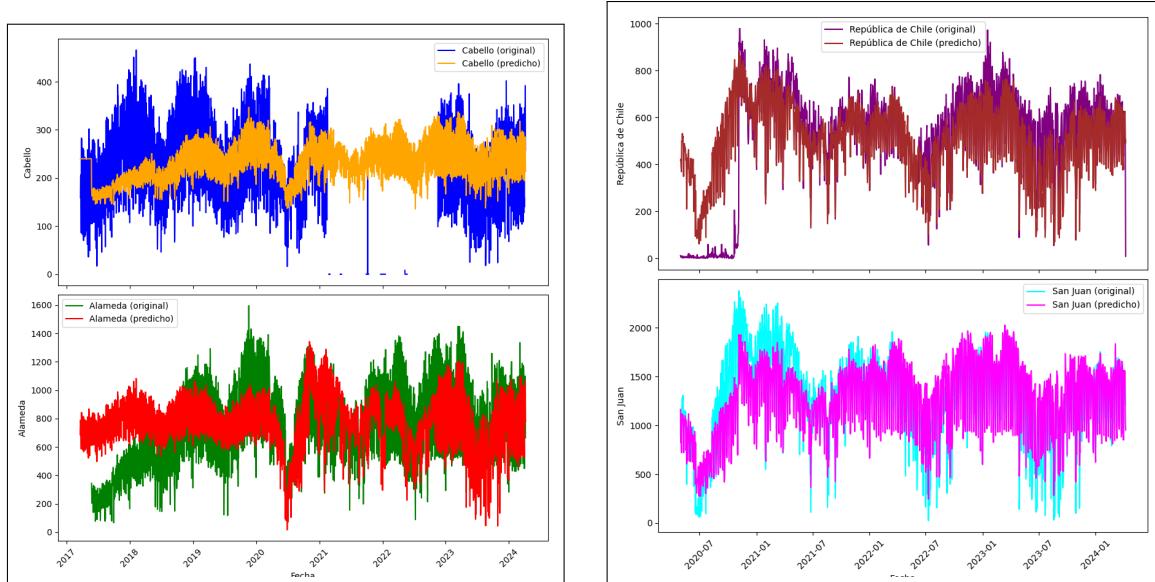


Figura 6: Regresión lineal múltiple.

Donde podemos observar que al separar las bases de datos de estas calles, por un lado Cabello y

Alameda, y por el otro República de Chile y San Juan se ve una mejoría en la visualización.

Y los valores del MAD para la imputación de los datos fueron los siguientes:

Modelo Imputación	Valor más frecuente	Interpolación Lineal	Promedio	Regresión múltiple
MAD Alameda	452.01	502.21	461.07	193.97
MAD Cabello	428.21	323.67	280.91	56.06
MAD República de Chile	443.43	315.32	222.79	70.01
MAD San Juan	532.34	704.21	642.81	167.41

4.2. Prophet

A continuación, se implementará el método Prophet para cada una de las calles respectivas, utilizando la librería **Prophet**[11]. Además, se visualizarán las tendencias correspondientes.

4.2.1. Prophet sin variables exógenas

Se hará un análisis visual sobre los resultados del modelo prophet para las distintas calles sin variables exógenas con sus respectivas tendencias:

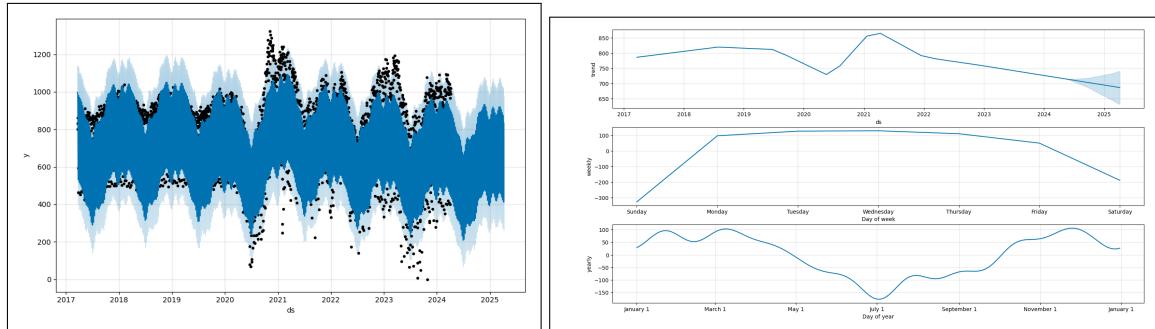


Figura 7: Visualización del Modelo Prophet en Alameda.

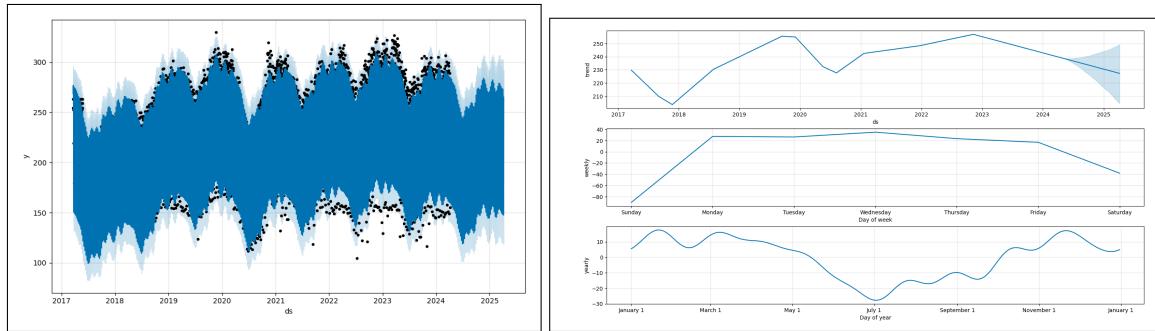


Figura 8: Visualización del Modelo Prophet en Cabello.

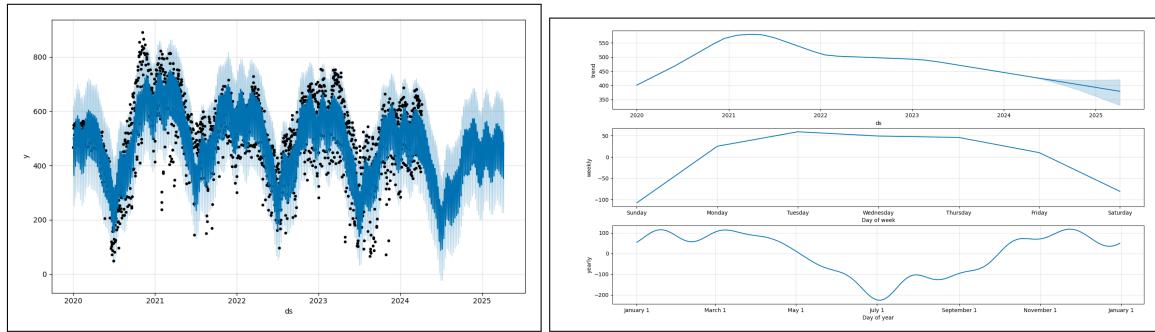


Figura 9: Visualización del Modelo Prophet en República de Chile con San Joaquín.

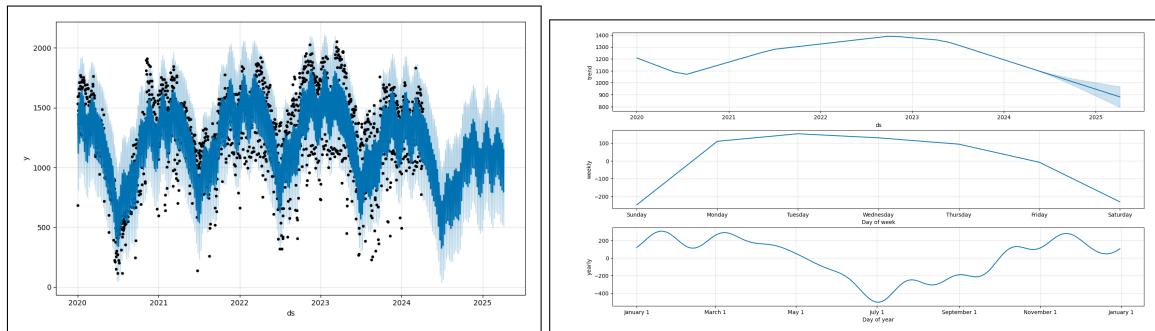


Figura 10: Visualización del Modelo Prophet en San Juan con Escrivá de Balaguer.

Análisis de Resultados con el Método Prophet:

En la primera gráfica, se muestra la predicción del conteo de ciclistas para las calles, utilizando el modelo Prophet. Las observaciones clave de esta gráfica son:

- **Datos Reales y Predicción:** Los puntos negros representan los datos reales del conteo de ciclistas, mientras que la línea azul oscura muestra la predicción del modelo. Es evidente que los modelos capturan la tendencia general y la estacionalidad del conteo de ciclistas a lo largo del tiempo.
- **Intervalos de Confianza:** Las bandas de color azul claro representan los intervalos de confianza del 95 % para las predicciones. Se observa que la mayoría de los puntos de datos reales caen dentro de estos intervalos, lo que sugiere que el modelo tiene una buena capacidad predictiva, de momento visualmente hablando.

La segunda gráfica descompone la predicción del conteo de ciclistas en sus componentes de tendencia y estacionalidad:

- **Tendencia:** La tendencia muestra un comportamiento variable a lo largo del tiempo, con un aumento gradual hasta 2021 en el caso de Alameda, un aumento hasta finales de 2020 y luego un aumento hasta 2023 en el caso de Cabello, un aumento a mitades del año 2021 para República de Chile y un aumento hacia finales de 2022 para San Juan, todo esto seguido de una ligera disminución hacia 2025 para todas las calles. Esta tendencia puede reflejar cambios en el uso de bicicletas a lo largo del tiempo debido a diversos factores como pudo ser la pandemia que duró desde finales de 2019 a inicios del 2023, también las distintas condiciones de las ciclovías, etc.

- **Estacionalidad Semanal:** La estacionalidad semanal indica variaciones en el conteo de ciclistas según el día de la semana. Se observa un aumento significativo en el conteo de ciclistas los días lunes, alcanzando un pico el miércoles y disminuyendo hacia el fin de semana, con el punto más bajo los domingos. Esto puede reflejar patrones de desplazamiento laboral y recreativo. Donde se observa el mismo comportamiento para todas las calles.
- **Estacionalidad Anual:** La estacionalidad anual muestra variaciones a lo largo del año, con picos alrededor de marzo y octubre, y disminuciones notables en los meses de invierno (julio-agosto) y a finales de año. Este patrón puede estar relacionado con condiciones climáticas y eventos estacionales que afectan el uso de bicicletas. Donde se observa el mismo comportamiento para todas las calles.

4.2.2. Prophet con variables exógenas

A continuación, se implementarán las variables exógenas, que consisten en los datos registrados de temperatura máxima y mínima diaria.

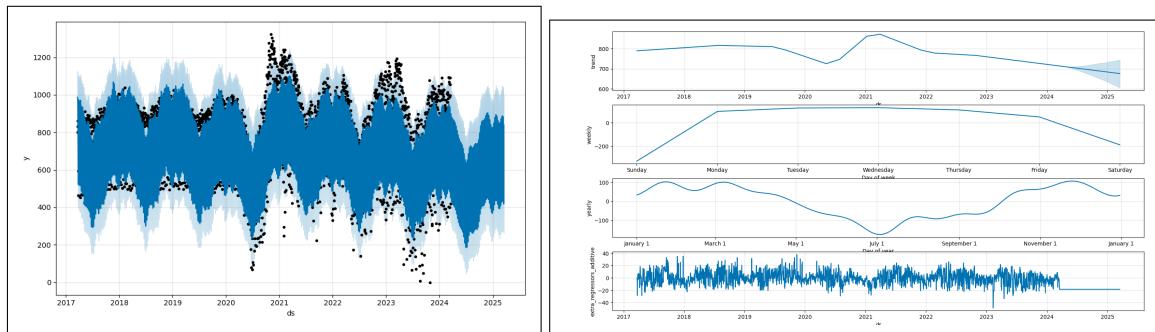


Figura 11: Visualización del Modelo Prophet en Alameda con variable exógena.

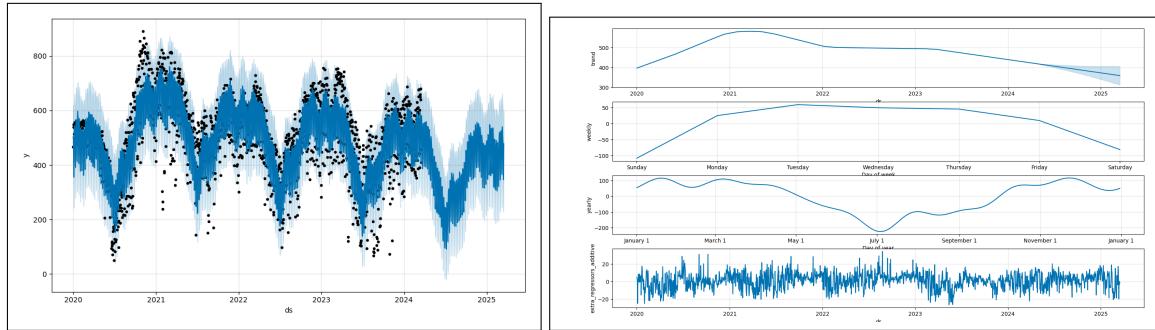


Figura 12: Visualización del Modelo Prophet en República de Chile con San Joaquín con variable exógena.

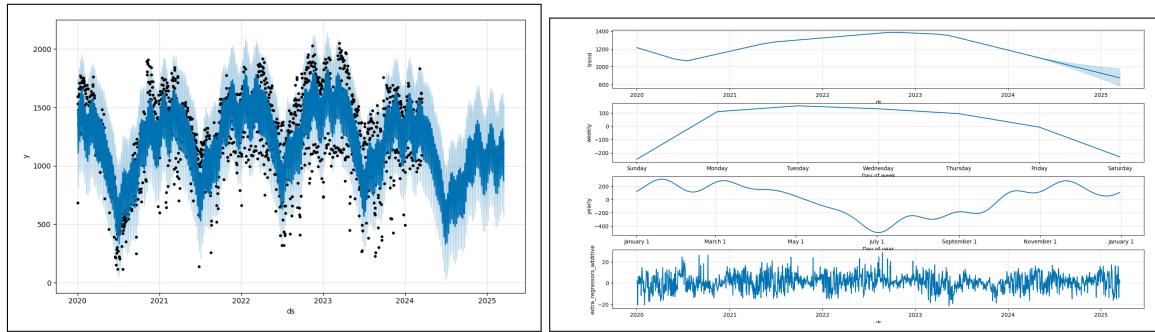


Figura 13: Visualización del Modelo Prophet en San Juan con Escrivá de Balaguer con variable exógena.

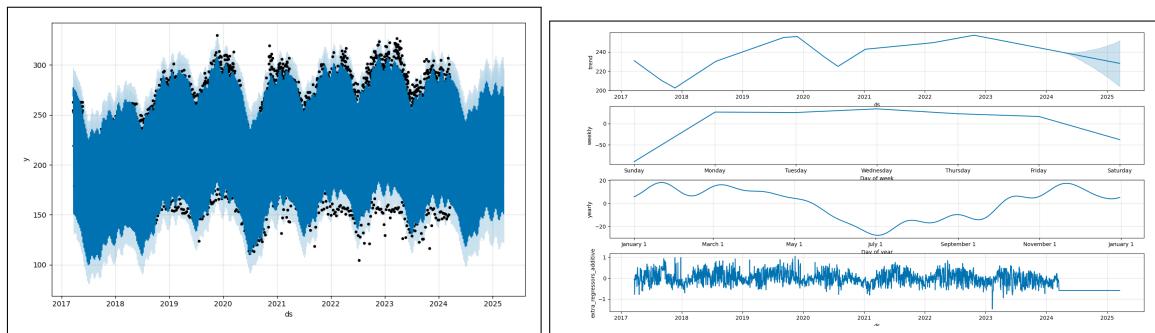


Figura 14: Visualización del Modelo Prophet en Cabello con variable exógena.

Análisis de Resultados con el Método Prophet con Variables Exógenas:

En la primera gráfica, similar a la descripción anterior se muestra la predicción del conteo de ciclistas para las calles, utilizando el modelo Prophet ahora con variables exógenas.

La segunda gráfica descompone la predicción del conteo de ciclistas en sus componentes de tendencia y estacionalidad, el análisis sigue siendo similar a sin variables exógenas, sin embargo se agrega una nueva descripción “extra_regressor_additive“:

- **Regresor extra (Temperatura Máxima y Mínima):** La gráfica muestra la contribución adicional de estas variables a lo largo del tiempo. Se puede observar que las temperaturas mínimas y máximas tienen un impacto significativo y fluctuante en el conteo de ciclistas. Estos efectos son más pronunciados durante ciertas épocas del año, lo que sugiere que las condiciones climáticas extremas (muy frías o muy cálidas) pueden afectar negativamente la cantidad de ciclistas. Donde observamos el mismo comportamiento para todas las calles.

4.3. ARIMA Y SARIMAX

A continuación, se implementarán los modelos ARIMA y SARIMAX para cada una de las calles respectivas, utilizando la librería **statsmodels**. Estos modelos permiten capturar tanto las tendencias lineales como las estacionales en series temporales, proporcionando una previsión basada en la descomposición de la serie en componentes autoregresivos e integrados. Además, se visualizarán las tendencias correspondientes.

4.3.1. Arima

El modelo ARIMA (Autoregressive Integrated Moving Average) es un modelo estándar para series temporales que no considera variables exógenas. A continuación, se detalla el análisis de los resultados obtenidos:

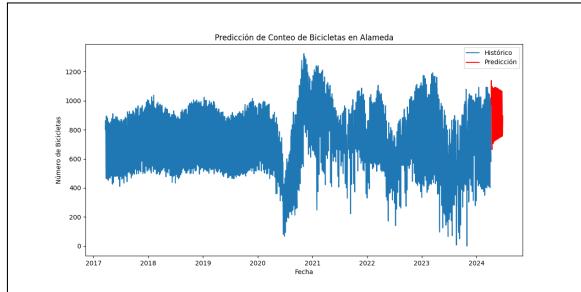


Figura 15: Modelo Arima en Alameda.

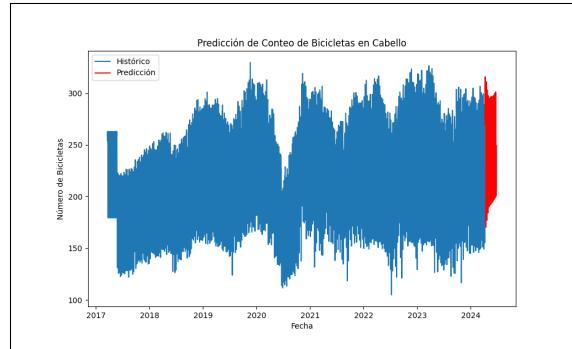


Figura 16: Modelo Arima en Cabello.

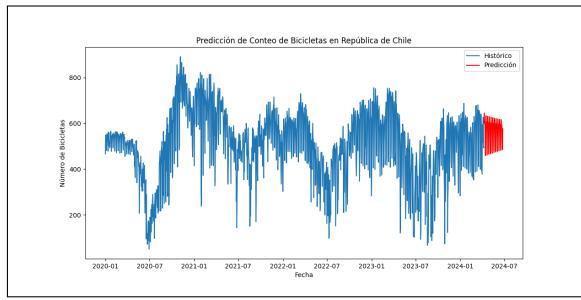


Figura 17: Modelo Arima en República de Chile con San Joaquín.

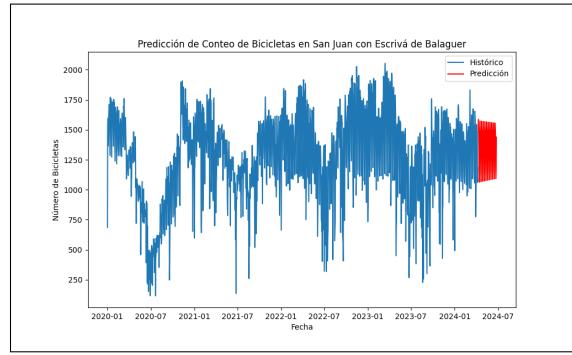


Figura 18: Modelo Arima en San Juan con Escrivá de Balaguer.

Análisis de Resultados con el Método Arima:

En cada gráfica, se muestra la predicción del conteo de ciclistas para las calles, utilizando el modelo Arima. Las observaciones clave de esta gráfica son:

- **Datos Reales y Predicción:**

Los gráficos muestran las predicciones del modelo ARIMA en rojo, comparadas con los datos históricos del conteo de ciclistas. En términos generales, el modelo ARIMA captura la tendencia global del número de ciclistas en diferentes calles, pero presenta ciertas limitaciones en la captura de la estacionalidad y las variaciones a corto plazo.

Las predicciones tienden a seguir la tendencia histórica de manera adecuada, reflejando visualmente un buen ajuste a los datos en calles como República de Chile con San Joaquín y San Juan con Escrivá de Balaguer, donde las desviaciones entre las predicciones y los datos reales son mínimas.

■ **Errores y Predicciones:**

La comparación entre las líneas rojas (predicciones) y los datos históricos permite evaluar la precisión del modelo. En algunas instancias se puede visualizar, las predicciones en rojo no coinciden perfectamente con los datos reales, indicando momentos en que el modelo ARIMA no ha logrado capturar correctamente las variaciones en los datos. La presencia de estas discrepancias sugiere la necesidad de ajustar el modelo o considerar variables adicionales para mejorar la precisión de las predicciones.

■ **Limitaciones:**

Aunque el modelo ARIMA proporciona una buena aproximación general de las tendencias, no captura adecuadamente algunas variaciones abruptas en el número de ciclistas. Esto es especialmente notable en la calle Cabello, donde las fluctuaciones bruscas no son bien predichas visualmente hablando.

4.3.2. Sarimax

En el siguiente caso, se implementarán variables exógenas, que, como se mencionó anteriormente, son las temperaturas máximas y mínimas diarias. Lo que cambia el modelo anterior ahora pasará a ser modelo Sarimax.

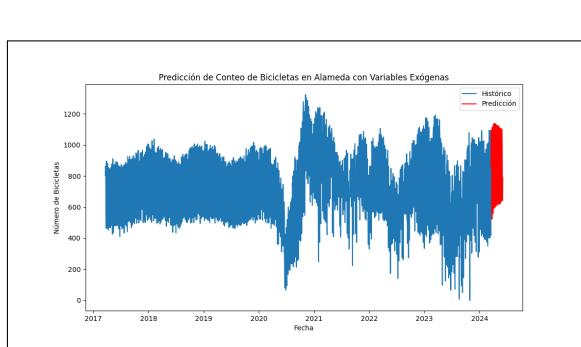


Figura 19: Modelo Sarimax en Alameda.

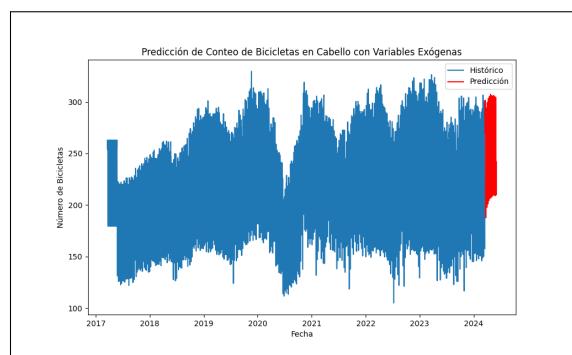


Figura 20: Modelo Sarimax en Cabello.

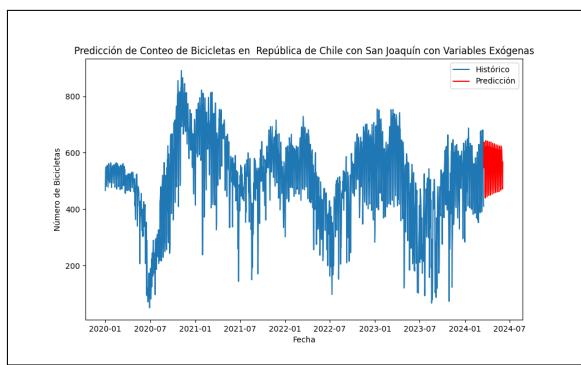


Figura 21: Modelo Sarimax en República de Chile con San Joaquín.

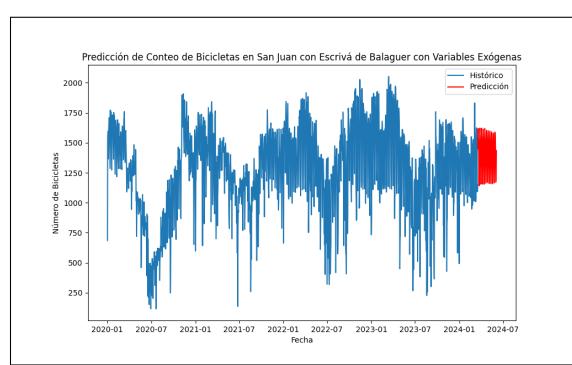


Figura 22: Modelo Sarimax en San Juan con Escrivá de Balaguer.

Análisis de Resultados con el Método Sarimax:

En cada gráfica, similar a la descripción anterior se muestra la predicción del conteo de ciclistas para las calles, esta vez utilizando el modelo Sarimax:

■ **Datos Reales y Predicción:**

La inclusión de variables exógenas mejora notablemente las predicciones visualmente. El modelo SARIMAX ajusta mejor los picos y valles observados en los datos históricos, capturando tanto las tendencias a largo plazo como las fluctuaciones estacionales con mayor precisión.

Las predicciones del modelo SARIMAX muestran una representación visual más precisa de los datos históricos en todas las calles analizadas, ajustando mejor las variaciones abruptas y los patrones estacionales.

■ **Impacto de las variables exógenas:**

La inclusión de variables exógenas, como las temperaturas máximas y mínimas diarias, tiene un impacto significativo visualmente en la mejora de las predicciones. La temperatura tienen una relación importante con el número de ciclistas, reflejando cómo las condiciones climáticas pueden influir en la cantidad de ciclistas.

En calles como Alameda y Cabello, la consideración de estas variables exógenas permite al modelo capturar mejor las fluctuaciones y proporcionar predicciones más precisas visualmente.

■ **Errores y Predicciones:**

La predicciones en rojo del modelo SARIMAX muestran visualmente una mayor precisión en comparación con las del modelo ARIMA, con menos discrepancias notables entre predicciones y los datos reales.

Esto puede indicar que la inclusión de variables exógenas mejora significativamente la capacidad del modelo para capturar las variaciones y tendencias en los datos.

Dado que es posible observar y analizar visualmente el comportamiento de los modelos, será necesario utilizar técnicas de medición de error.

5. MAPE y MAD

A continuación, se presentan las tablas con los valores de MAPE y MAD obtenidos para los modelos de prophet, arima y sarimax con y sin variables exógenas. Estos resultados permiten comparar la eficiencia de los modelos y seleccionar el más adecuado para las predicciones:

Modelo	Variable exógena	MAPE (%)			
		Alameda	República	San Juan	Cabello
Arima y Sarimax	Sin	397.12	49.23	26.44	12.97
	Con	313.52	51.4	28.25	23.93
Prophet	Sin	57.49	19.4	19.68	5.05
	Con	55.59	19.63	19.8	5.01

Modelo	Variable exógena	MAD			
		Alameda	República	San Juan	Cabello
Arima y Sarimax	Sin	184.72	121.29	193.65	28.4
	Con	154.6	122.47	196.15	47.76
Prophet	Sin	79.67	65.24	167.42	11.09
	Con	78.48	65.61	167.61	10.99

6. Conclusiones

Durante el desarrollo del proyecto, se implementaron diversos modelos para mejorar la predicción. Desde el momento de la adquisición de la información, se realizó un análisis descriptivo exhaustivo. Posteriormente, se identificaron deficiencias de datos en distintos períodos, lo que hizo necesaria la limpieza de los mismos, eliminando observaciones atípicas y completando los valores faltantes.

En este punto, fue necesario revisar la estrategia metodológica, se comenzó explorando técnicas como la media, mediana y moda, sin embargo, se constató que no eran métodos apropiados de aproximación. Fue entonces cuando se profundizó en el estudio de la imputación de datos anómalos, lo que resultó crucial al aplicar técnicas de regresión lineal para corregir los datos faltantes. Esta metodología se estructuró de manera que la estimación de las variables dependientes estuviera relacionada con otras calles circundantes. El resultado de este proceso se reflejó en los gráficos presentados en la sección de Resultados del informe. Los resultados indicaron que la regresión lineal múltiple fue la técnica más adecuada para la imputación de datos, logrando los valores más bajos de Desviación Absoluta Media (MAD).

Adicionalmente, se implementaron modelos de predicción de series temporales para anticipar el flujo de bicicletas. Se utilizaron dos enfoques principales: el modelo Prophet y el modelo ARIMA. El modelo Prophet, conocido por su capacidad para manejar datos con tendencias y estacionalidades, fue ajustado tanto con variables exógenas como sin ellas. Este modelo permitió capturar de manera efectiva las fluctuaciones diarias en el flujo de bicicletas, aunque su desempeño varió dependiendo de la calle y los regresores incluidos.

Por otro lado, el modelo ARIMA fue utilizado para analizar y predecir las series temporales del flujo de bicicletas. Se probaron diversas configuraciones de parámetros (p, d, q) para optimizar las predicciones. El modelo ARIMA mostró un buen rendimiento en general, pero al igual que con Prophet, se observó que la inclusión de variables exógenas como la temperatura mejoró significativamente la precisión de las predicciones.

Los resultados de los modelos fueron evaluados utilizando dos métricas principales: MAPE y MAD. Los valores indicaron que el modelo Prophet generalmente presenta menores errores en comparación con el modelo ARIMA, especialmente cuando se incluyen variables exógenas.

Para mejorar aún más la precisión de los modelos, se sugiere eliminar los días feriados del análisis, ya que estos pueden introducir ruido en los datos y afectar negativamente las predicciones. La eliminación de estos días puede ayudar a los modelos a concentrarse en patrones más consistentes y representativos del comportamiento regular del flujo de bicicletas.

Referencias

- [1] Alfonso Saura Sánchez. *Manual de matplotlib*. Consulta realizada el 1 de mayo de 2024. Aprende con Alf. 2024. URL: <https://aprendeconalf.es/docencia/python/manual/matplotlib/>.
- [2] V. P. Ariyanti y Tristyanti Yusnitasari. “Comparison of ARIMA and SARIMA for Forecasting Crude Oil Prices”. En: *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* 7.2 (2023), págs. 405-413. DOI: 10.29207/resti.v7i2.4895. URL: <https://doi.org/10.29207/resti.v7i2.4895>.
- [3] George EP Box, Gwilym M Jenkins y Gregory C Reinsel. “Time series analysis: forecasting and control”. En: *Holden-Day* (1970).
- [4] Facebook. *Prophet: Automatic Forecasting Procedure*. <https://facebook.github.io/prophet/>. Consultado el Fecha de acceso. Fecha de acceso.
- [5] Jiawei Han, Micheline Kamber y Jian Pei. *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann, 2024.
- [6] Rob J Hyndman y George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018. URL: <https://otexts.com/fpp3/>.
- [7] Spyros Makridakis, Steven C Wheelwright y Rob J Hyndman. *Forecasting: Methods and Applications*. Wiley, 1997.
- [8] Ministerio de Vivienda y Urbanismo de Chile. *Ubicación de contadores de bicicletas en Rancagua y Machalí*. Consulta realizada el 1 de mayo de 2024. Ministerio de Vivienda y Urbanismo de Chile. 2021. URL: [https://www.arcgis.com/home/webmap/viewer.html?url=https%3A%2F%2Fgeoide\[minvu.cl%2Fserver%2Frest%2Fservices%2FPlans_Programas%2FCiclov%25C3%25ADas_Minvu%2FFeatureServer%2F0&source=sd](https://www.arcgis.com/home/webmap/viewer.html?url=https%3A%2F%2Fgeoide[minvu.cl%2Fserver%2Frest%2Fservices%2FPlans_Programas%2FCiclov%25C3%25ADas_Minvu%2FFeatureServer%2F0&source=sd).
- [9] Municipalidad de Rancagua. *Ubicación de ciclovías Rancagua y Machalí*. Consulta realizada el 9 de mayo de 2024. Municipalidad de Rancagua. 2024. URL: <https://www.rancagua.cl/red-ciclovias/> (visitado 09-05-2024).
- [10] scikit-learn developers. *sklearn linear model LinearRegression*. Consulta realizada el 1 de mayo de 2024. scikit-learn. 2024. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.
- [11] Sean J. Taylor y Benjamin Letham. *Prophet: Automatic Forecasting Procedure*. <https://pypi.org/project/prophet/>. Accessed: 2024-06-13. 2024.
- [12] Ronald E. Walpole et al. *Introducción a la estadística*.