Imputación Robusta en series temporales de afluencia: Metro de Medellín

Sebastian Carvalho Salazar, ✉ scarvalhos@eafit.edu.co
Daniel Loaiza Lopez, ✉ dloaizal@eafit.edu.co
Sebastian Ramirez Escobar, ✉ sramireze1@eafit.edu.co
Asmec Duvan Urrea Uribe, ✉ adurreau@eafit.edu.co
Hernán Felipe Sánchez Cárdenas, ✉ hfsanchezc@eafit.edu.co

Asesor:

Pablo Andres Saldarriaga Aristizabal

Universidad EAFIT
Maestría en Ciencia de Datos y Analítica
Medellín
2024

TABLA DE CONTENIDOS

I. Introducción

La imputación de datos faltantes es un problema crítico en el análisis de datos, ya que los valores ausentes pueden introducir sesgos y reducir la precisión de los modelos predictivos. Los métodos robustos en estadística, que minimizan la influencia de los valores atípicos, ofrecen herramientas valiosas para abordar este problema de manera efectiva. Este trabajo se centra en la imputación robusta de datos faltantes en series temporales de la afluencia de usuarios en las diferentes líneas del Metro de Medellín, empleando tanto modelos de regresión tradicionales como métodos de regresión robusta.

A lo largo de la historia, diversos investigadores han desarrollado métodos para afrontar los datos faltantes. Desde las técnicas básicas como la eliminación de registros con datos faltantes y la sustitución por la media, hasta métodos avanzados como la imputación múltiple y algoritmos de aprendizaje automático. Cada uno de estos métodos tiene sus ventajas y limitaciones, y la elección del método adecuado depende del contexto específico y de las características de los datos.

El objetivo principal de este estudio es comparar la efectividad de un modelo de regresión tradicional con métodos de regresión robusta en la imputación de datos faltantes en series temporales que presentan valores atípicos. Utilizando el marco CRISP-DM (Cross-Industry Standard Process for Data Mining), se busca desarrollar un modelo que no solo sea preciso, sino también robusto ante la presencia de outliers.

La metodología implementada incluye la recopilación y preparación de los datos históricos de afluencia del Metro de Medellín, la implementación de diversos modelos de regresión y la evaluación de su desempeño. En este sentido, se entrenarán modelos con datos de las líneas que no presentan ausencias en los tramos faltantes, y se comparará su efectividad mediante métricas robustas.

Este trabajo contribuye al campo de la estadística no paramétrica al explorar y comparar distintas técnicas de imputación robusta, ofreciendo una visión integral y práctica para la mejora de la calidad de los análisis en presencia de datos faltantes y valores atípicos.

II. Revisión de literatura

La primera investigación sobre datos faltantes fue publicada en 1960 por Wilkinson, y en ella, el autor hace una comparación de los métodos de imputación más utilizados en la época: listwise deletion (eliminación de registros con datos faltantes) y sustitución por media [1]. Estos métodos eran fáciles de implementar pero tenían limitaciones significativas como la reducción de tamaño de la muestra y sesgo [2].

En 1976, el trabajo de Rubin representó un importante hito respecto al entendimiento de datos faltantes, proponiendo los conceptos de *Missing completely at random* (MCAR), *Missing at random* (MAR) y *Missing not at random* (MNAR) que buscan explicar sus causas. Esto dio paso al Algoritmo de Expectación-Maximización (EM) propuesto por Depster, Laid y Rubin en 1977, el cual permitió generar estimadores robustos a partir del uso del método de máxima verosimilitud, en el cual los datos faltantes se asumen como variables aleatorias y los datos imputados se generan sin que sea necesario ajustar los modelos y que también tiene aplicación cuando los datos tienen información oculta [3].

Durante la década de los 80s hubo avances considerables en este campo. Si bien se popularizaron métodos de imputación como la regresión simple, en 1983, Madow, Nisselson y Olkin desarrollaron el método de imputación no paramétrico *hot-deck*, que consiste en llenar los datos vacíos con valores observados seleccionados de forma aleatoria de casos semejantes. Algunos autores, lo han combinado con modelos de regresión para mejorar la calidad de las imputaciones [4]. Algunos años más tarde, la imputación múltiple, propuesta por Rubin en 1987, se convirtió en un método estándar para el manejo de datos faltantes debido al avance de las tecnologías computacionales de la época [5]. Este método consiste en crear múltiples (m) conjuntos de datos imputados y luego, analizar individualmente cada uno para finalmente, combinar los resultados. [6].

Con el auge de los métodos de aprendizaje automático, a partir de 1990, se empezaron a utilizar técnicas de imputación basadas en árboles de decisión y métodos de *deep learning* [7]. La robustez se volvió una característica deseable debido a la presencia inevitable de outliers que suelen traer los conjuntos de datos.

En 2006, Zhang et.al desarrollaron un método de imputación basado en kernel, denominado POP (Parameter Optimization Method), que optimiza parámetros estadísticos como la media y la función de distribución tras la imputación de datos faltantes. Este enfoque demostró ser más eficiente y robusto que la imputación por regresión determinista [8].

En 2009, Branden y Verboven desarrollaron el método SEQimpute, que se basa en una estimación secuencial de los valores faltantes en una observación incompleta, minimizando el determinante de la covarianza de la matriz de datos aumentada. Luego, se añade la observación imputada al conjunto de datos completo y el algoritmo continúa con la próxima

observación con valores faltantes. Este método demostró buen comportamiento en presencia de outliers, al utilizar estimadores robustos de ubicación y dispersión para manejar la influencia de valores atípicos [9].

En años posteriores, ha habido otros avances en este campo como el método de imputación robusta mediante regresión, propuesto por Rana et. al en 2012, una versión robusta de la imputación aleatoria por regresión clásica [10], la Imputación Robusta Basada en Optimización propuesta en 2017 por Bertsimas et. al, en la cual se integra varios modelos predictivos como k-vecinos más cercanos, máquinas de vectores soporte y árboles de decisión, mejorando la precisión fuera de muestra y reduciendo el error absoluto medio en comparación con otros métodos de imputación [11].

III. Objetivos

III-A. Objetivo general

Evaluar la efectividad de la imputación de datos faltantes en series temporales de afluencia de pasajeros del Metro de Medellín mediante el uso de métodos de regresión robusta, comparándolos con un modelo de regresión tradicional.

III-B. Objetivos Específicos

1. Implementar y comparar diferentes métodos de regresión para la imputación de datos faltantes en series temporales, incluyendo modelos de regresión lineal tradicional, RANSAC y Modelos Lineales Robustos (RLM).
2. Evaluar la robustez de los modelos frente a valores atípicos mediante la introducción controlada de outliers en el conjunto de datos y la comparación del desempeño de los modelos bajo estas condiciones.
3. Implementar métricas robustas para la evaluación de los modelos, como el Error Porcentual Absoluto Medio (MAPE) y su versión mediana (MdAPE), para proporcionar una medida más confiable del rendimiento de los modelos en presencia de valores atípicos.
4. Realizar pruebas estadísticas, como el test de Wilcoxon, para determinar si existen diferencias significativas entre los modelos robustos y el modelo de regresión lineal tradicional en términos de su capacidad para imputar datos faltantes de manera efectiva.

IV. METODOLOGÍA

Este proyecto emplea el marco CRISP-DM (Cross-Industry Standard Process for Data Mining) para la imputación robusta de datos faltantes en las series de tiempo del Metro de Medellín. El objetivo principal es comparar la efectividad de un modelo de regresión tradicional con métodos de regresión robusta, como RANSAC y Robust Linear Models, en presencia de valores atípicos. Utilizando datos de afluencia del metro en diferentes horarios (de 4 AM a 11 PM) para cada una de sus líneas, se busca imputar tramos faltantes entrenando los modelos con datos de las líneas que no presentan ausencias en dichos tramos. Los pasos metodológicos ejecutados incluyen la recopilación y preparación de datos, la implementación de los modelos de regresión y la evaluación comparativa de su desempeño en la imputación de datos faltantes, como se puede ver en la figura 1.

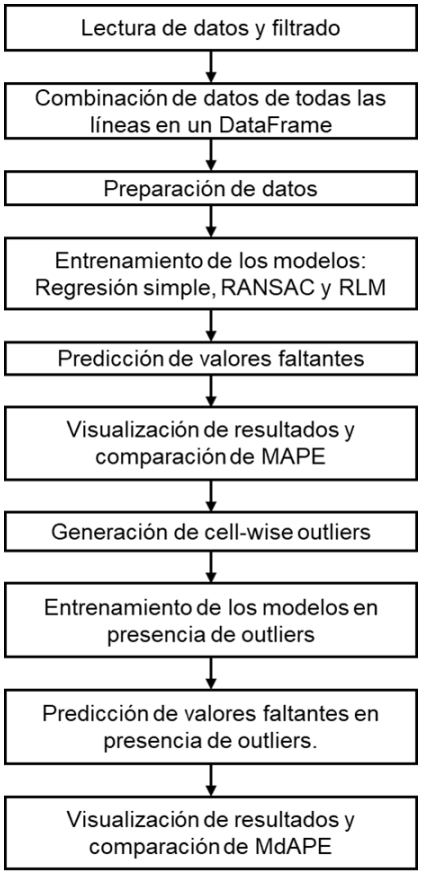


FIG. 1: Diagrama del procedimiento

A continuación se detalla el procedimiento:

IV-A. Comprensión del Negocio

El objetivo principal es imputar los datos faltantes en las series de tiempo del Metro de Medellín, específicamente en la afluencia de pasajeros para cada una de sus líneas y en

diferentes horarios desde las 4:00 AM hasta las 11:00 PM. La finalidad es desarrollar un modelo de imputación robusta que mantenga la precisión incluso en la presencia de valores atípicos y compararlo con un modelo de regresión tradicional a partir de una métrica robusta.

IV-B. Preparación de los Datos

Se inicia con la carga de los datos históricos de afluencia del Metro de Medellín, seleccionando las columnas relevantes que incluyen Línea, Horas, Fecha y Cantidad de Pasajeros. Los datos se filtran para obtener información específica de cada línea y hora indicada.

Los datos de todas las líneas se combinan en un único conjunto, utilizando la fecha como clave de combinación. Posteriormente, los datos se contaminan intencionalmente con valores atípicos para simular escenarios atípicos y probar la robustez de los modelos.

IV-C. Modelado

Para el modelado se consideran tres enfoques principales:

1. **Regresión Clásica**: Se entrena un modelo de regresión lineal tradicional.
2. **RANSAC (RANdom SAmple Consensus)**: Se entrena un modelo de regresión robusta utilizando el algoritmo RANSAC, que es capaz de manejar valores atípicos de manera efectiva a partir de un proceso iterativo.
3. **Modelos Lineales Robustos**: Se utilizan modelos de regresión lineal robusta que minimizan el impacto de los valores atípicos a partir del uso de diferentes normas robustas.

Para garantizar la validez temporal de las series, se emplea una estrategia de validación cruzada específica para series de tiempo. Se prueban diferentes combinaciones de hiperparámetros para cada modelo, optimizando (Minimizando) la métrica MAPE (Mean Absolute Percentage Error). El proceso de optimización incluye la selección de parámetros como el número máximo de ensayos, el puntaje de parada y la probabilidad de parada para el modelo RANSAC, así como la elección de diferentes normas robustas (Hampel, AndrewWave, HuberT, RamsayE, TrimmedMean) para los Modelos Lineales Robustos. Cada combinación de hiperparámetros se evalúa utilizando validación cruzada, asegurando que los modelos sean evaluados en múltiples segmentos temporales para validar su capacidad predictiva y robustez.

Además, se calcularon correlaciones de Pearson y robustas como Kendall y Spearman para la identificación de posibles relaciones no lineales entre las variables. Estas medidas de correlación permitieron evaluar la fuerza y dirección de las relaciones, considerando tanto la linearidad como la robustez frente a valores atípicos. Esto con el fin de indicar la necesidad de explorar o no la implementación de modelos no paramétricos, capaces de capturar de manera más efectiva la complejidad de las relaciones no lineales.

IV-D. Evaluación

En este proyecto, se están añadiendo *cell-wise* outliers al conjunto de datos seleccionando aleatoriamente un porcentaje de estos y aumentando significativamente sus valores. Esto se hace para simular escenarios extremos y evaluar cómo afectan a los modelos, permitiendo así entender el comportamiento de los modelos robustos frente a estos escenarios.

Finalmente, se visualizan los resultados mostrando la serie temporal original junto con los valores imputados, permitiendo evaluar gráficamente la calidad de la imputación realizada. Además, se compara el rendimiento de cada modelo utilizando métricas robustas para determinar cuál método ofrece la mejor imputación en presencia de valores atípicos.

Esta metodología permite una evaluación integral de los modelos de regresión clásicos y robustos, destacando la importancia de la robustez en la imputación de datos faltantes en series de tiempo con posibles valores atípicos.

IV-D1) MAPE y MdAPE: En este proyecto se usa el MAPE (Error Porcentual Absoluto Medio) como métrica debido a su amplia aplicación en series de tiempo. La principal ventaja del MAPE en series de tiempo es que proporciona una medida fácil de interpretar y comparativa de la precisión del modelo, independientemente de la escala de los datos. Además, se ha desarrollado una métrica robusta utilizando la mediana en lugar de la media en el cálculo del MAPE, lo que permite reducir el impacto de valores atípicos y ofrecer una evaluación más estable del rendimiento del modelo (MdAPE).

IV-D2) Intervalos de Confianza: Los intervalos de confianza se calcularon remuestreando la covariables para generar las predicciones del tramo temporal en el que se desea hacer la imputación. Luego, se tomaron el percentil 97.5 y 2.5 para el limite superior e inferior, respectivamente, de cada una de las predicciones en el tramo temporal que le corresponde, ya que se tiene una matriz de predicciones de diferentes resampleos para los diferentes días que se están llenando.

IV-D3) Test Estadístico Wilcoxon: El test de Wilcoxon es una prueba no paramétrica utilizada para comparar dos muestras emparejadas o relacionadas. En este proyecto, se emplea para evaluar si los modelos de regresión robusta (RLM y RANSAC) son significativamente mejores que el modelo de regresión lineal tradicional (RL) en la imputación de datos faltantes.

En este proyecto, se busca determinar si los modelos de regresión robusta (RLM y RANSAC) son significativamente mejores que el modelo de regresión lineal tradicional (RL) en la imputación de datos faltantes. Para ello, se plantean las siguientes hipótesis:

V. Experimentos computacionales

Análisis de correlación:

Se calculó la matriz de correlación de Pearson para determinar si la relación entre las variables es lineal, y la correlación de Spearman para medir las relaciones monótonas. Para entender más de la relación entre los datos, se realizó la diferencia entre estas, ya que, se espera que si hay una relación no lineal entre los datos, los resultados de la matriz serían mayormente negativos.

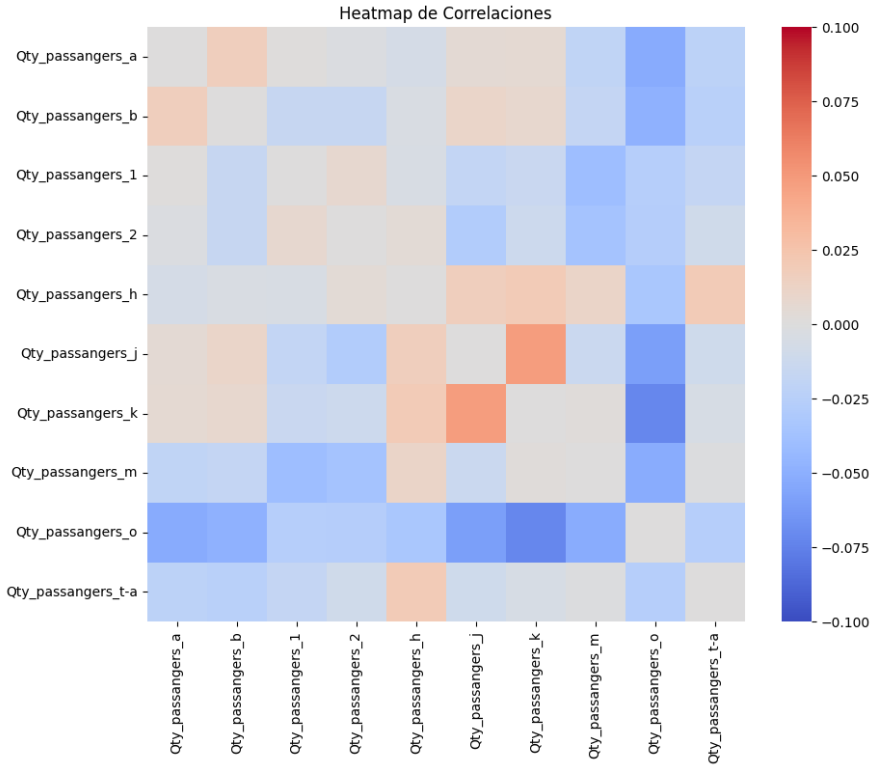
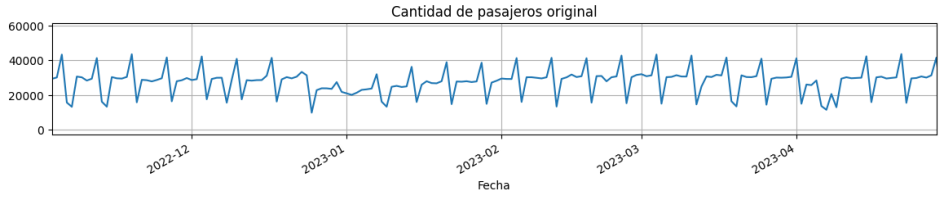


FIG. 2: Heatmap de correlaciones

De la figura 2, se puede concluir que existe una relación lineal fuerte entre las líneas, ya que, ambas correlaciones dieron muy altas y la diferencia entre las mismas es cercana a 0. Esto nos inclinaría en un principio por el uso de modelos lineales, pero se quiere validar también si es posible que en presencia de outliers el modelo RANSAC se desempeñe mejor.

V-A. Datos Limpios

En la figura 3a se observa la serie de tiempo original limpia y la serie con los valores faltantes a estimar

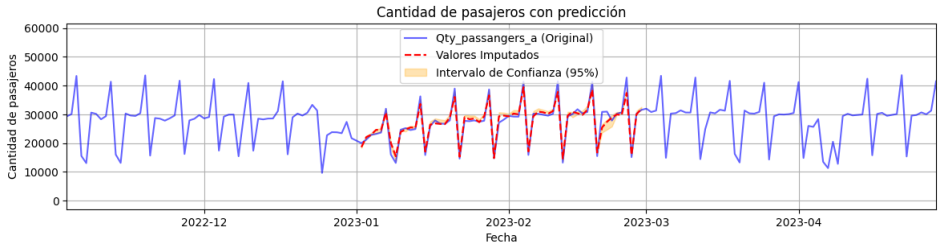


(a) Data original

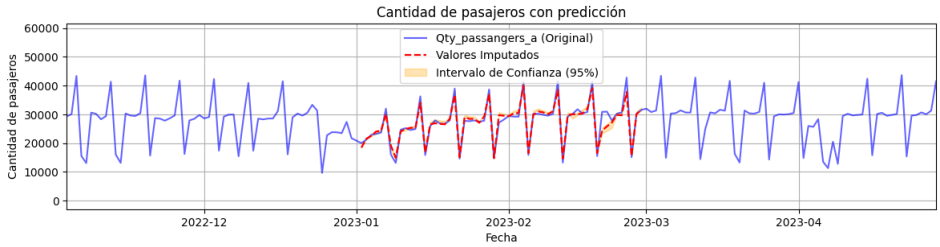
(b) Data con valores faltantes

FIG. 3: Data limpia y valores faltantes

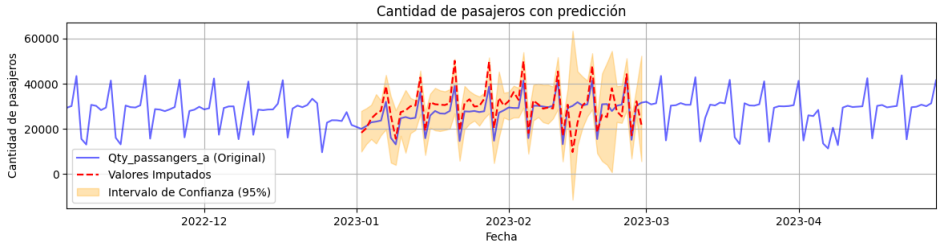
Con este conjunto de datos se realizó la estimación de los valores faltantes utilizando los modelos mencionados en la metodología. En la figura 4 se pueden ver los resultados de cada uno. Como se puede apreciar, el RANSAC manifiesta visualmente los peores resultados, con los intervalos de confianza más amplios entre los 3.:



(a) Predicciones con Regresión Lineal



(b) Predicciones con Regresión Lineal Robusta

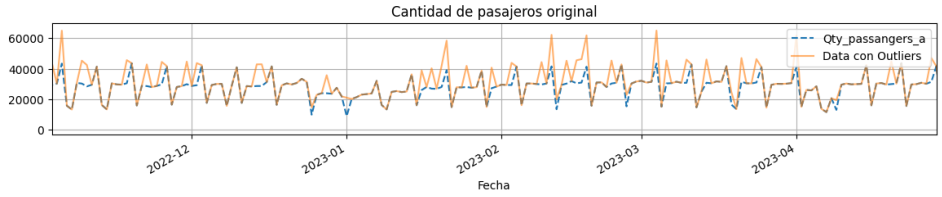


(c) Predicciones con RANSAC

FIG. 4: Predicciones de los valores faltantes usando la data limpia

V-B. Datos con Outliers:

En la figura 5 se observa la señal de tiempo original con Outliers y la señal con los valores faltantes a estimar.

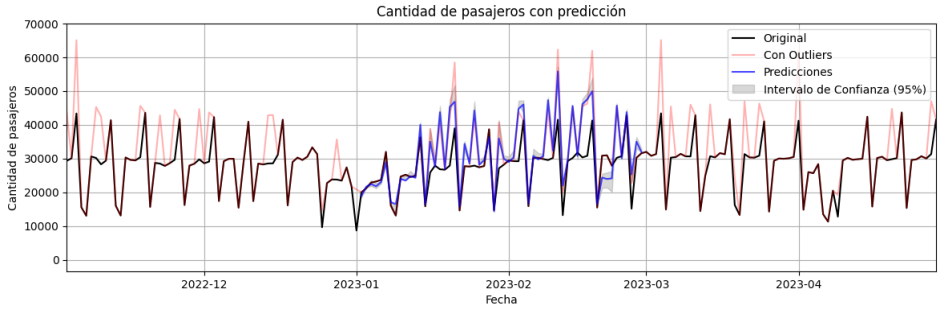


(a) Data original

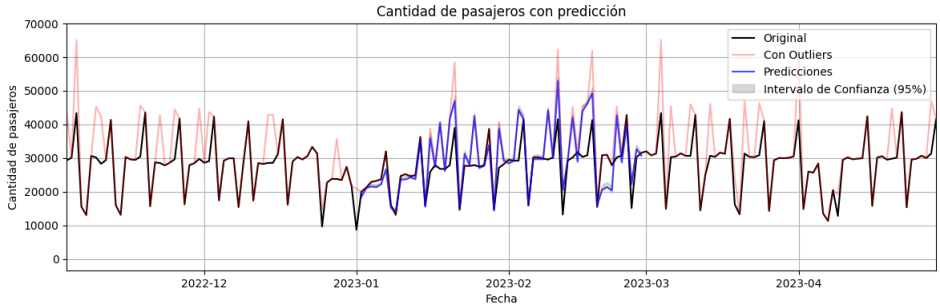
(b) Data con valores faltantes

FIG. 5: Data con Outliers y valores faltantes

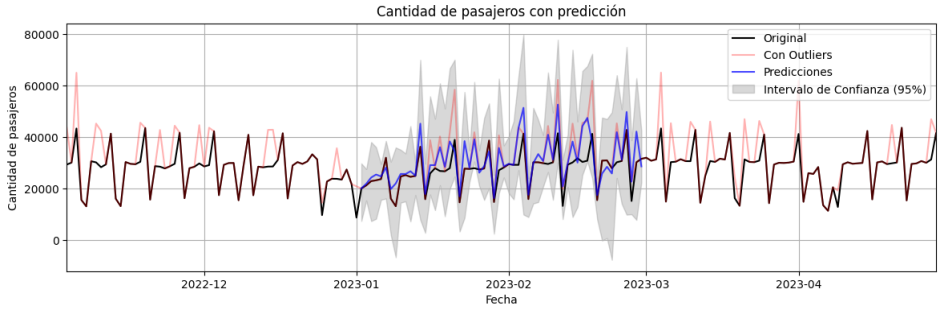
Con este conjunto de datos con Outliers se realizó la estimación de los valores faltantes utilizando los modelos mencionados en la metodología. En la figura 6 se pueden ver los resultados de cada uno. Al igual que sucedió con los datos en limpio, el modelo RANSAC presentó el peor desempeño, como se puede evidenciar en los intervalos de confianza de cada uno.



(a) Predicciones con Regresión Lineal



(b) Predicciones con Regresión Lineal Robusta



(c) Predicciones con RANSAC

FIG. 6: Predicciones de los valores faltantes usando la data con Outliers

V-C. Test de Wilcoxon

La hipótesis plantea que al menos uno de los modelos robustos supera al regresor lineal tradicional. Para evaluar esto, se proponen las siguientes hipótesis:

- Hipótesis General

Hipótesis General	
Hipótesis nula (H_0)	Ninguno de los modelos robustos (RLM y RANSAC) es mejor que el modelo de regresión lineal (RL). Es decir, el modelo RL está desplazado a la izquierda de los otros dos modelos.
Hipótesis alternativa (H_a)	Al menos uno de los modelos, ya sea el modelo de Regresión Robusta (RLM) o el modelo RANSAC, presenta diferencias significativas con respecto al modelo de Regresión Lineal (RL). Es decir, el modelo que presenta estas diferencias está más desplazado a la izquierda, en comparación con el RL.

TABLA. I: Hipótesis General

V-D. Prueba de Wilcoxon con datos limpios

- Comparación entre RLM y RL

Comparación entre RLM y RL	
Hipótesis nula (H_0)	El modelo de regresión lineal (RL) es igual o mejor que el modelo de regresión lineal robusta (RLM). Es decir, no hay diferencias significativas en las imputaciones realizadas por ambos modelos.
Hipótesis alternativa (H_1)	El modelo de regresión lineal robusta (RLM) es significativamente mejor que el modelo de regresión lineal (RL). Además, el RLM está más desplazado a la izquierda en comparación con el RL, indicando un mejor desempeño.

TABLA. II: Comparación entre RLM y RL

TABLA. III: Resultados de la prueba

Estadístico de la prueba	**Valor-p**
61.0	0.1518

Dado que el valor-p es mayor al nivel de significancia α de 0.05, no se rechaza la hipótesis nula. Es decir, no hay evidencia suficiente para afirmar que la mediana del MAPE del modelo RLM es menor que la del modelo RL y por tanto, sus predicciones seon mejores.

- Comparación entre RANSAC y RL

Comparación entre RANSAC y RL	
Hipótesis nula (H_0)	El modelo de regresión lineal (RL) es igual o mejor que el modelo RANSAC. Es decir, no hay diferencias significativas en las imputaciones realizadas por ambos modelos.
Hipótesis alternativa (H_1)	El modelo RANSAC es significativamente mejor que el modelo de regresión lineal (RL). Además, el RANSAC está más desplazado a la izquierda en comparación con el RL, indicando un mejor desempeño.

TABLA. IV: Comparación entre RANSAC y RL

TABLA. V: Resultados de la prueba

Estadístico de la prueba	**Valor-p**
160.0	0.9998

En este caso, tampoco se rechaza la hipótesis nula, es decir, no hay evidencia suficiente para afirmar que la mediana del MAPE del modelo RANSAC es menor que la del modelo RL y por tanto, que este modelo es mejor.

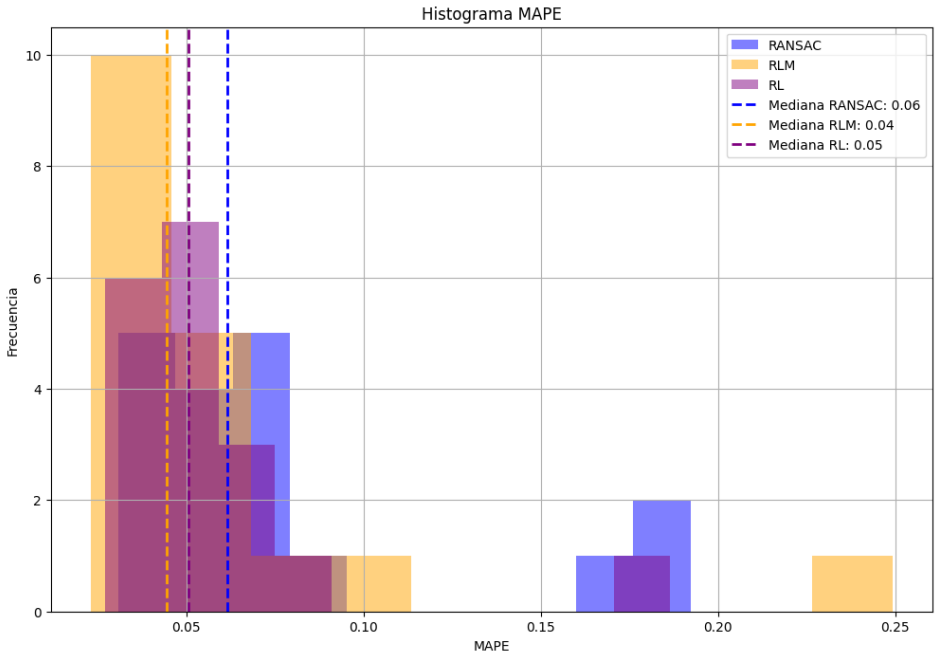


FIG. 7: Test de Wilcoxon con datos limpios

V-E. Test de Wilcoxon datos con outliers

Para los conjuntos con outliers, se utiliza la metrica MdAPE para evaluar los errores, ya que al tener presencia de Outliers el MAPE se puede ver afectado por errores muy grandes.

- Comparación entre RLM y RL

Comparación entre RLM y RL	
Hipótesis nula (H_0)	El Modelo de RLM no está desplazado a la izquierda en comparación del RL.
Hipótesis alternativa (H_1)	El Modelo de RLM presenta diferencias significativas con respecto al RL. Adicional, el RLM está mas desplazado a la izquierda, en comparación que el RL.

TABLA. VI: Comparación entre RLM y RL

TABLA. VII: Resultados de la prueba

Estadístico de la prueba	**Valor-p**
45.0	0.0407

Puesto que el valor-p es menor que 0.05, se rechaza la hipótesis nula. Por lo tanto, se puede concluir que la mediana del MdAPE del RLM es significativamente menor que la del modelo RL, es decir, este modelo tuvo mejores resultados.

- Comparación entre RANSAC y RL

Comparación entre RANSAC y RL	
Hipótesis nula (H_0)	El Modelo RANSAC no está desplazado a la izquierda en comparación del RL.
Hipótesis alternativa (H_1)	El Modelo RANSAC presenta diferencias significativas con respecto al RL. Adicional, el modelo RANSAC está mas desplazado a la izquierda, en comparación que el RL.

TABLA. VIII: Comparación entre RANSAC y RL

TABLA. IX: Resultados de la prueba

Estadístico de la prueba	**Valor-p**
140.0	0.9930

En este caso, no se rechaza la hipótesis nula: no hay evidencia suficiente para afirmar que el MdAPE del modelo RANSAC es menor que el del modelo RL.

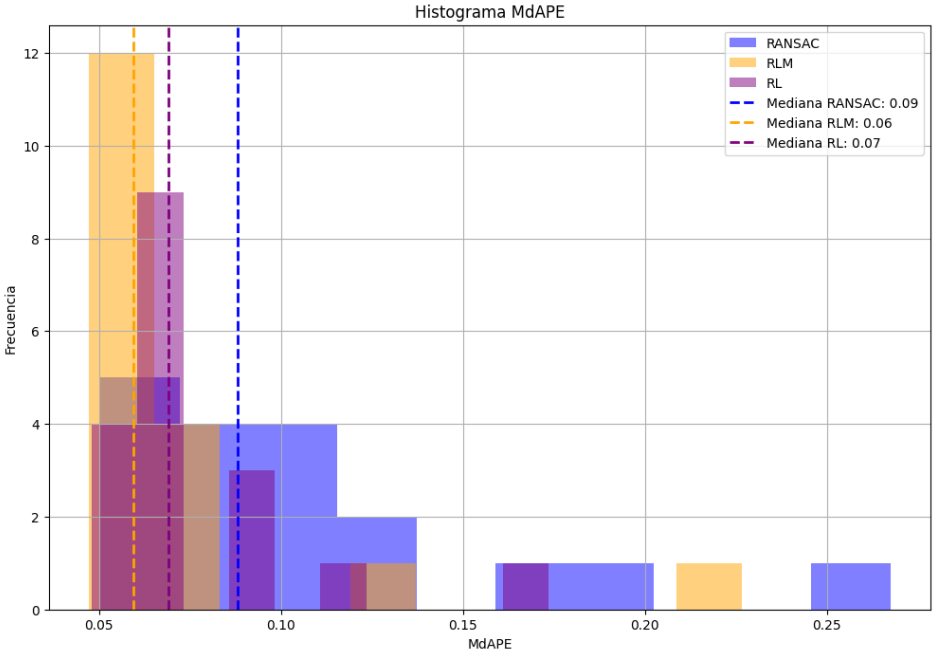


FIG. 8: Test de Wilcoxon datos con outliers

VI. Conclusiones y recomendaciones

- Con los datos limpios, el modelo cuyo MAPE tiene la menor mediana es el RLM, siendo seguido por la RL y más rezagado, el RANSAC. Sin embargo, esto no es estadísticamente significativo, con lo cual, por la prueba de Wilcoxon se concluye que el mejor modelo fue la regresión lineal. Esto hace sentido con el análisis de correlaciones, que nos dio un indicio de la linealidad de los datos.
- En presencia de cell-wise outliers, el modelo con menor mediana para el MdAPE fue el RLM, seguido por la regresión lineal y nuevamente, teniendo en último lugar la regresión RANSAC. Sin embargo, en este caso, la diferencia sí es estadísticamente significativa, con lo cual se puede concluir que en este escenario el mejor modelos es el RLM.
- Los experimentos confirmaron la importancia de emplear métricas robustas como MdAPE para evaluar la precisión de los modelos en presencia de datos atípicos, asegurando así que las evaluaciones reflejen de manera precisa el comportamiento del modelo bajo diversas condiciones.
- Si bien, a nivel visual se podía concluir que en ambos casos el mejor modelo fue el RLM, utilizar el test de Wilcoxon fue muy útil para demostrar, en términos de significancia, cuál fue el mejor modelo para cada caso. Sin embargo, el análisis visual fue útil al comparar los resultados mediante intervalos de confianza, en el cual fue muy claro que el peor desempeño lo presentó el modelo RANSAC.

VII. Anexos

VII-A. Regresión Lineal

La regresión lineal es una técnica de análisis predictivo utilizada para modelar la relación entre una variable dependiente y y una o más variables independientes X. Esta técnica es fundamental en la estadística y ha sido ampliamente estudiada en la literatura. El modelo de regresión lineal simple se puede expresar matemáticamente como:

$$y = \beta_0 + \beta_1 X + \epsilon, \tag{1}$$

donde y es la variable dependiente, X es la variable independiente, β_0 es el intercepto, β_1 es el coeficiente de regresión y ϵ es el término de error.

Para que los estimadores obtenidos mediante el método de mínimos cuadrados sean insesgados y eficientes, se deben cumplir ciertos supuestos:

1. **Linealidad:** La relación entre la variable dependiente y las variables independientes debe ser lineal.
2. **Independencia:** Las observaciones deben ser independientes entre sí.
3. **Homoscedasticidad:** La varianza de los errores debe ser constante a lo largo de todas las observaciones.
4. **Normalidad:** Los errores deben seguir una distribución normal.

Regresión lineal múltiple:

La regresión lineal múltiple es una extensión de la regresión lineal simple que permite modelar la relación entre una variable dependiente y y múltiples variables independientes $X_1, X_2, \ldots, X_p$. Este modelo se puede expresar matemáticamente como:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{2}$$

donde y es la variable dependiente, $X_1, X_2, \ldots, X_p$ son las variables independientes, β_0 es el intercepto, $\beta_1, \beta_2, \ldots, \beta_p$ son los coeficientes de regresión y ϵ es el término de error.

El método de estimación más comúnmente utilizado para ajustar un modelo de regresión lineal múltiple es el de mínimos cuadrados ordinarios (OLS, por sus siglas en inglés), que minimiza la suma de los cuadrados de los residuos (diferencias entre los valores observados y los valores predichos por el modelo):

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_p X_{ip})^2. \tag{3}$$

VII-B. RANSAC (RANdom Sample Consensus)

Es un método iterativo para calcular los parámetros de un modelo matemático a partir de un conjunto de datos observados que contiene valores atípicos. Este algoritmo no determinista produce un resultado razonable con una probabilidad mayor a medida que se permiten más iteraciones. RANSAC fue publicado por primera vez por Fischler [12] en 1981.

Los datos consisten en *inliers*, es decir, los datos cuya distribución se explica por un conjunto de parámetros del modelo, aunque pueden estar sujetos a ruido, y ”valores atípicos”, que son datos que no encajan en el modelo. Los valores atípicos pueden provenir de valores extremos del ruido, mediciones erróneas o hipótesis incorrectas sobre la interpretación de los datos. RANSAC asume que, dado un conjunto de *inliers*, existe un procedimiento que puede estimar los parámetros de un modelo que explica de manera óptima o se ajusta a esta información.

Algoritmo RANSAC

RANSAC logra su objetivo mediante la repetición de los siguientes pasos:

1. **Selección Aleatoria de Subconjuntos de Datos:** Se selecciona aleatoriamente un pequeño subconjunto de los datos originales. Este subconjunto se denomina ïnliers hipotéticos”.
2. **Ajuste del Modelo:** Se ajusta un modelo a este subconjunto de datos. Por ejemplo, si se está ajustando una línea, se calcula la línea que pasa por los puntos seleccionados.
3. **Evaluación del Modelo:** Todos los demás datos se prueban contra el modelo ajustado. Esos puntos que se ajustan al modelo estimado, de acuerdo con alguna función de pérdida específica del modelo, se consideran como parte del conjunto de consenso.
4. **Repetición del Proceso:** Los pasos 1 a 3 se repiten muchas veces (un número fijo de iteraciones). Cada iteración genera un modelo y un conjunto de consenso correspondiente.
5. **Selección del Mejor Modelo:** Después de un número fijo de iteraciones, se selecciona el modelo que tiene el mayor número de puntos en su conjunto de consenso como el mejor modelo.
6. **Reajuste del Modelo:** Se ajusta un modelo final utilizando todos los puntos del conjunto de consenso del mejor modelo encontrado.

Los valores de los parámetros t y d deben determinarse a partir de requisitos específicos relacionados con la aplicación y el conjunto de datos, posiblemente basados en la evaluación experimental. El parámetro k (número de iteraciones) se puede determinar a partir de un resultado teórico.

El valor p determina la probabilidad de que el algoritmo RANSAC seleccione solo inliers a partir del conjunto de datos de entrada cuando se eligen n puntos con los que se estiman los

parámetros del modelo. w es la probabilidad de elegir un inlier cada vez que se selecciona un solo punto:

$$w = \frac{\text{number of inliers in data}}{\text{number of points in data}} \tag{4}$$

Si w no se conoce de antemano, se puede estimar. Suponiendo que los n puntos necesarios para la estimación de un modelo se seleccionan independientemente, w^n es la probabilidad de que todos los n puntos sean inliers y $1 - w^n$ es la probabilidad de que al menos uno de los n puntos sea un outlier. Esta probabilidad a la potencia de k es la probabilidad de que el algoritmo nunca seleccione un conjunto de n puntos donde todos son inliers, lo que debe ser igual a $1 - p$:

$$1 - p = (1 - w^n)^k \tag{5}$$

Tomando el logaritmo en ambos lados, se obtiene:

$$k = \frac{\log(1 - p)}{\log(1 - w^n)} \tag{6}$$

Este resultado asume que los n puntos de datos se seleccionan independientemente. En caso de que los puntos se seleccionen sin reemplazo, el valor derivado de k debe tomarse como un límite superior. Para aumentar la confianza, se puede agregar la desviación estándar $SD(k)$ de k:

$$SD(k) = \frac{\sqrt{1 - w^n}}{w^n} \tag{7}$$

VII-C. Robust Linear Model (RLM)

Con el objetivo de reducir el impacto de valores atípicos, se desarrolló el modelo de regresión robusta (RLM). A diferencia del modelo de regresión simple, que utiliza mínimos cuadrados, este modelo utiliza diferentes funciones de pérdida para reducir el impacto de los valores atípicos sobre la estimación de los coeficientes del modelo [13]:

- Función de Huber: combina las características de la pérdida cuadrática y la pérdida absoluta, proporcionando una transición suave entre las dos. La idea es usar una pérdida cuadrática para errores pequeños (lo que da eficiencia) y una pérdida lineal para errores grandes (lo que da robustez) [14].

$$\rho_\delta(r) = \begin{cases} \frac{1}{2}r^2 & \text{si } |r| \leq \delta \\ \delta(|r| - \frac{1}{2}\delta) & \text{si } |r| > \delta \end{cases}$$

Donde:

- r es el residuo o error de predicción, $r = y - \hat{y}$.
- δ es un parámetro que determina el punto de transición entre las pérdidas cuadrática y lineal.

- Función de Hampel: es una extensión de la función de Huber que tiene tres partes: una parte cuadrática cerca del cero, una parte constante para grandes residuos, y una región de transición suave entre estas dos partes [15]. Está definida como:

$$\rho(r) = \begin{cases} \frac{1}{2}r^2 & \text{si } |r| \leq a \\ a\left(|r| - \frac{a}{2}\right) & \text{si } a < |r| \leq b \\ a\left(c - \frac{a}{2}\right) + (|r| - c)\frac{c-b}{2} & \text{si } b < |r| \leq c \\ a\left(c - \frac{a}{2}\right) + (c - b)\frac{c-b}{2} & \text{si } |r| > c \end{cases}$$

Donde:

- r es el residuo o error de predicción, $r = y - \hat{y}$.
- a, b, c son los parámetros que definen las zonas de transición.

- Función de Andrew: también conocida como Andrew's wave, es suave y periódica, lo que permite que los residuos grandes tengan una influencia limitada en la estimación de los parámetros del modelo [15].

$$\rho(r) = \begin{cases} 1 - \cos\left(\frac{r}{c}\right) & \text{si } |r| \leq \pi c \\ 2 & \text{si } |r| > \pi c \end{cases}$$

Donde:

- r es el residuo o error de predicción, $r = y - \hat{y}$.
- c es un parámetro de escala que controla la amplitud de la función.

- Función de pérdida de Ramsay: crece menos rápidamente que la cuadrática para grandes valores de los residuos [16]. Se define matemáticamente como:

$$\rho(r) = \begin{cases} \frac{r^2}{2}\left(1 - \left(\frac{r}{c}\right)^2 + \frac{1}{3}\left(\frac{r}{c}\right)^4\right) & \text{si } |r| \leq c \\ \frac{c^2}{6} & \text{si } |r| > c \end{cases}$$

Donde:

- r es el residuo o error de predicción, $r = y - \hat{y}$.
- c es un parámetro de escala que controla la amplitud de la función.

- Trimmed mean: más que una función de pérdida es un estimador robusto de la media

que se obtiene descartando una fracción c de las observaciones más grandes y más pequeñas, y luego calculando la media de las observaciones restantes [17]. Dado un conjunto de datos $x_1, x_2, \ldots, x_n$, la media recortada de orden α se calcula de la siguiente manera:

1. Ordenar los datos en orden ascendente.
2. Eliminar los primeros y últimos $\alpha\,\%$ de los datos.
3. Calcular la media de los datos restantes.

Si n es el tamaño del conjunto de datos, y k es el número de datos a recortar de cada extremo, entonces $k = \lfloor \alpha n \rfloor$. La media recortada se calcula como:

$$\text{Media recortada} = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} x_i$$

Donde:

- n es el número total de datos.
- k es el número de datos eliminados de cada extremo, calculado como $k = \lfloor \alpha n \rfloor$.
- x_i son los valores del conjunto de datos después de eliminar los extremos.

REFERENCIAS

[1] G. Wilkinson, «Comparison of missing value procedures,» *Australian Journal of Statistics*, 1960.

[2] M. I. S. Adnan Farah Zakaria, *60-year research history of missing data: A bibliometric review on Scopus database*. Institute of Engineering Mathematics, Universiti Malaysia Perlis, 2020.

[3] D. Gallardo López, *Algoritmo EM*. Universidad de Alicante, 2000.

[4] A. M. Ferreira, «Metodologías de análisis y imputación de datos faltantes en series de velocidad del viento,» *VI Congreso Galego de Estatística e Investigación de Operacións*, 2003.

[5] F. Medina y M. Galván, *Imputación de datos: Teoría y práctica*. Unidad de Estadísticas Sociales de la División de Estadística y Proyecciones Económicas de la CEPAL, 2007. dirección: https://repositorio.cepal.org/server/api/core/bitstreams/02dd479f-fae2-43c4-b5ec-5419fa7f6190/content.

[6] R. Alfaro y M. Fuenzalida, «Imputación múltiple en encuestas microeconómicas,» *Cuadernos de economía*, págs. 273-288, 2009, ISSN: 0717-6821. DOI: 10.4067/S0717-68212009000200007.

[7] A. Goicoechea, *Imputación basada en árboles de clasificación*. Eustat, 2002.

[8] S. Zhang, Y. Qin, X. Zhu, J. Zhang y C. Zhang, *Optimized parameters for missing data imputation*. In: Yang, Q., Webb, G. (eds) PRICAI 2006: Trends in Artificial Intelligence, 2006.

[9] V. S. Branden KV, «Robust data imputation,» *Comput Biol Chem*, 2009.

[10] S. Rana, A. H. John y H. Midi, «Robust regression imputation for analyzing missing data,» en *2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE)*, 2012, págs. 1-4. DOI: 10.1109/ICSSBE.2012.6396621.

[11] D. Bertsimas, C. Pawlowski e Y. D. Zhuo, «From predictive methods to missing data imputation: an optimization approach,» *Journal of Machine Learning Research*, vol. 18, n.º 196, págs. 1-39, 2018. dirección: http://jmlr.org/papers/v18/17-073.html.

[12] M. A. Fischler y R. C. Bolles, «Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,» *Commun. ACM*, vol. 24, n.º 6, págs. 381-395, jun. de 1981, ISSN: 0001-0782. DOI: 10.1145/358669.358692. dirección: https://doi.org/10.1145/358669.358692.

[13] D. Huang, R. Cabral y F. De la Torre, «Robust Regression,» vol. 38, oct. de 2012, págs. 616-630, ISBN: 978-3-642-33764-2. DOI: 10.1007/978-3-642-33765-9_44.

[14] P. J. Huber, «Robust Estimation of a Location Parameter,» *The Annals of Mathematical Statistics*, vol. 35, n.º 1, págs. 73-101, 1964. DOI: 10.1214/aoms/1177703732. dirección: https://doi.org/10.1214/aoms/1177703732.

[15] M. A. Alam, K. Fukumizu e Y.-P. Wang, «Influence Function and Robust Variant of Kernel Canonical Correlation Analysis,» *Neurocomputing*, vol. 304, mayo de 2017. DOI: 10.1016/j.neucom.2018.04.008.

[16] A. D. Deria, A. Hoyyi y M. Mustafid, «REGRESI ROBUST ESTIMASI-M DENGAN PEMBOBOT ANDREW, PEMBOBOT RAMSAY DAN PEMBOBOT WELSCH MENGGUNAKAN SOFTWARE R,» *Jurnal Gaussian*, 2019. DOI: 10.14710/j.gauss.v8i3.26682.

[17] «Trimmed Mean: Definition, Calculating Benefits.» (), dirección: https://statisticsbyjim.com/basics/trimmed-mean/. (accessed: 15.05.2024).