UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO
ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA

TESIS

IMPUTACIÓN POR INTERVALOS BASADA EN PRONÓSTICO Y DETECCIÓN DE CAMBIOS ABRUPTOS EN EL MODELAMIENTO DE SERIES TEMPORALES FALTANTES Y SU APLICACIÓN

PARA OPTAR AL GRADO ACADÉMICO DE MAESTRO EN ESTADÍSTICA

AUTOR
Br. MIGUEL ANGEL QUINTANA YAURIS

ASESOR:
Dr. ALFREDO VALENCIA TOLEDO
CÓDIGO ORCID: 0000-0001-6505-9634

CUSCO – PERÚ
2026

Universidad Nacional de San Antonio Abad del Cusco

INFORME DE SIMILITUD

(Aprobado por Resolución Nro.CU-321-2025-UNSAAC)

El que suscribe, el **Asesor** ALFREDO VALENCIA TOLEDO quien aplica el software de detección de similitud al trabajo de investigación/tesistitulada: IMPUTACIÓN POR INTERVALOS BASADA EN PRONÓSTICO Y DETECCIÓN DE CAMBIOS ABRUPTOS EN EL MODELAMIENTO DE SERIES TEMPORALES FALTANTES Y SU APLICACIÓN

Presentado por: MIGUEL ANGEL QUINTANA YAURIS DNI N° 75015878;

presentado por: ..... DNI N°: .....

Para optar el título Profesional/Grado Académico de MAESTRO EN ESTADÍSTICA

Informo que el trabajo de investigación ha sido sometido a revisión por 3 veces, mediante el Software de Similitud, conforme al Art. 6° del ***Reglamento para Uso del Sistema Detección de Similitud en la UNSAAC*** y de la evaluación de originalidad se tiene un porcentaje de 6 %.

Evaluación y acciones del reporte de coincidencia para trabajos de investigación conducentes a grado académico o título profesional, tesis

Porcentaje	Evaluación y Acciones	Marque con una (X)
Del 1 al 10%	No sobrepasa el porcentaje aceptado de similitud.	X
Del 11 al 30 %	Devolver al usuario para las subsanaciones.	
Mayor a 31%	El responsable de la revisión del documento emite un informe al inmediato jerárquico, conforme al reglamento, quien a su vez eleva el informe al Vicerrectorado de Investigación para que tome las acciones correspondientes; Sin perjuicio de las sanciones administrativas que correspondan de acuerdo a Ley.	

Por tanto, en mi condición de Asesor, firmo el presente informe en señal de conformidad y **adjunto** las primeras páginas del reporte del Sistema de Detección de Similitud.

Cusco, 05 de Julio de 2026

Firma

Post firma ALFREDO VALENCIA TOLEDO

Nro. de DNI 43162177

ORCID del Asesor 0000-0001-6505-9634

Se adjunta:

1. Reporte generado por el Sistema Antiplagio.
2. Enlace del Reporte Generado por el Sistema de Detección de Similitud: **oid:** 27259:604998044

MIGUEL ANGEL QUINTANA YAURIS

IMPUTACIÓN POR INTERVALOS BASADA_EN_PRONÓSTICO_Y_DETECCIÓN_DE_CAMBIOS_A...

Universidad Nacional San Antonio Abad del Cusco

Detalles del documento

Identificador de la entrega
trn:oid:::27259:604998044

Fecha de entrega
5 jul 2026, 10:29 p.m. GMT-5

Fecha de descarga
5 jul 2026, 10:45 p.m. GMT-5

Nombre del archivo
IMPUTACIÓN POR INTERVALOS BASADA_EN_PRONÓSTICO_Y_DETECCIÓN_DE_CAMBIOS_ABRUPTO....pdf

Tamaño del archivo
2.1 MB

158 páginas

34.223 palabras

182.441 caracteres

6% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado
- Coincidencias menores (menos de 10 palabras)

Exclusiones

- N.º de coincidencias excluidas

Fuentes principales

5% Fuentes de Internet
2% Publicaciones
4% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO
ESCUELA DE POSGRADO

INFORME DE LEVANTAMIENTO DE OBSERVACIONES A TESIS

Dr. LEONCIO SOLIS QUISPE, Director de la Escuela de Posgrado, nos dirigimos a usted en condición de integrantes del jurado evaluador de la tesis intitulada **IMPUTACIÓN POR INTERVALOS BASADA EN PRONÓSTICO Y DETECCIÓN DE CAMBIOS ABRUPTOS EN EL MODELAMIENTO DE SERIES TEMPORALES FALTANTES Y SU APLICACIÓN** del **Br. MIGUEL ANGEL QUINTANA YAURIS.** Hacemos de su conocimiento que el (la) sustentante ha cumplido con el levantamiento de las observaciones realizadas por el Jurado el día **TREINTA DE JUNIO DE 2026.**

Es todo cuanto informamos a usted fin de que se prosiga con los trámites para el otorgamiento del grado académico de MAESTRO EN ESTADÍSTICA.

Cusco, 03 julio de 2026

DRA. KARLA ZELMIRA APARICIO ARENAS
Primer Replicante

MTRO. JULIO CESAR HUAMAN CUSIHUAMAN
Segundo Replicante

DR. HENRY LAZO CHUQUIHUAYTA
Primer Dictaminante

MTRO. JOEL GRIMALDO OLARTE ESTRADA
Segundo Dictaminante

DEDICATORIA

A Dios y a mi Señor de Huanca por fortalecer mi espíritu en los momentos de mayor debilidad y por permitirme culminar esta etapa bajo su bendición.

A mi Mamita Ana Yauris Centeno, quien hoy me acompaña desde el cielo. Hace apenas unos meses que partiste al cielo y tu ausencia duele, seguirás siendo la luz que ilumina mi camino, mi ángel, dedico cada página de esta tesis a tu memoria. No puedo olvidarte, no voy a olvidarte te dije, es una promesa. El dolor a veces pienso que es tan grande **"CUANDO UN HIJO PIERDE A UNA MADRE"** que es incomparable con nada... es como si uno perdiera una parte de su alma.

A Julio Mayorga Challco, por su apoyo incondicional y por ser guía fundamental en el aspecto profesional y personal, Su aporte ha sido un elemento esencial y estructural en mi persona para lograr los éxitos alcanzados hasta ahora.

Miguel Angel Quintana Yauris.

AGRADECIMIENTO

A Dios por darme la fuerza necesaria para continuar, por la fuerza y sabiduría por ayudarme a cuminar esta etapa importante en mi vida.

A mi padre Carlos Quintana Hurtado, a mis hermanos Carlos Quintana Yauris y Alfredo Robinson Quintana Yauris, quienes son parte de estos logros sin el apoyo de ustedes no seria posible.

A mi asesor, Dr. Alfredo Valencia Toledo por su invaluable apoyo, paciencia y su conocimiento, para la elaboración de esta tesis, sus consejos, sugerencias fueron fundamentales para alcanzar este logro.

A mis estimados amigos y colegas: Ever Rojas Rayme, Tony Godofredo Ticona Flores y Charles Jhon Mamani Mayta. Los que me ofrecieron su amistad sin condiciones, su comprensión y consejos en tiempos difíciles, contribuyeron a mi formación; sus risas y respaldo me han servido y continúan sirviéndome.

Al Dr. Epifanio Puma Huañec y a la Dra. Isabel Corbacho Carpio, por creer en mi persona y capacidad, su confianza , sus recomendaciones fueron y son fundamentales en mi vida.

Al Dr. Santiago Soncco Tumpi por creer en mi, darme la oportunidad como persona y tambien en el aspecto profesional.

A una persona muy especial Ruth Gianela en mi vida, por ser parte de todos los logros hasta el dia de hoy por ser un apoyo fundamental.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

La presente investigación tiene como objetivo analizar la imputación por intervalos basada en pronósticos y detectar cambios abruptos para el modelamiento de una serie temporal faltante. Para ello, se integran diversos métodos y modelos de series temporales, incluyendo los enfoques clásicos de Metodología Box-Jenkins (ARMA, ARIMA y SARIMA), modelos robustos (Holt-Winters, STL(Seasonal-Trend decomposition using LOESS)) y TBATS(Trigonometric, Box-Cox, ARMA, Trend, Seasonal), el modelo de pronóstico Prophet fue aplicado para la imputación, y la técnicas de detección de cambios abruptos *"Interval Forecast Imputation"*(IFI). Como resultado, el autor propone la metodología MAQY-TS, la cual integra dos etapas: Etapa 1 imputación y detección de cambios abruptos y la etapa 2 modelamiento. Esta metodología es formalizada mediante un algoritmo que permite la detección de datos faltantes, imputación, detección de cambios abruptos, reconstrucción de la serie con datos faltantes hasta la selección del modelo óptimo que mejor se ajusta a la serie. El análisis y procesamiento de los datos se realizó con el software R. Asimismo, se identificaron y analizaron series temporales reales en las cuales se aplicó la metodología. Los resultados en las series Número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025) y Número pasajeros internacionales en el Aeropuerto Internacional Velasco Astete (2004-2025), Entre todos los modelos analizados, el TBATS fue el que se adaptó mejor y tuvo la capacidad de predecir más alta. Además, se ratifica la adecuación y robustez de la metodología, que se basa en un enfoque adaptativo e integral.

Palabras clave: Series de tiempo, Datos faltantes, Imputación, Cambios abruptos, Modelamiento.

ABSTRACT

This research aims to analyze interval imputation based on forecasts and detect abrupt changes for modeling a missing time series. To this end, various time series methods and models are integrated, including classical Box-Jenkins methodologies (ARMA, ARIMA, and SARIMA), robust models (Holt-Winters, STL (Seasonal-Trend decomposition using LOESS)), and TBATS (Trigonometric, Box-Cox, ARMA, Trend, Seasonal). The Prophet forecasting model was applied for imputation, and the Interval Forecast Imputation (IFI) technique was used for detecting abrupt changes. As a result, the author proposes the MAQY-TS methodology, which integrates two stages: Stage 1, imputation and detection of abrupt changes, and Stage 2, modeling. This methodology is formalized through an algorithm that allows for the detection of missing data, imputation, detection of abrupt changes, reconstruction of the series with missing data, and selection of the optimal model that best fits the series. Data analysis and processing were performed using R software. Real-world time series were also identified and analyzed, and the methodology was applied to them. The results for the series "Number of Visitors to the Inca City of Machu Picchu (2002–2025).ªnd "Number of International Passengers at Velasco Astete International Airport (2004–2025)"showed that, among all the models analyzed, TBATS was the best fit and had the highest predictive capacity. Furthermore, the suitability and robustness of the methodology, which is based on an adaptive and comprehensive approach, are confirmed.

Keywords: Time series, Missing data, Imputation, Abrupt changes, Modeling.

INTRODUCCIÓN

La presente investigación aborda el manejo de datos incompletos en un intervalo de tiempo en series temporales, el objetivo central fue desarrollar y abordar una metodología robusta capas de capturar incertidumbre en los valores falta, abordar la detección de cambios abruptos o anomalías en series temporales, y posteriormente modelar la serie temporal.

En una primera parte se implementa la técnica Interval Forcasting Imputation (IFI) que es una estrategia innovadora que integra técnicas de pronóstico con la cuantificación de incertidumbre mediante intervalos de predicción. Este enfoque no solo permite imputar valores faltantes, sino también evaluar la consistencia entre los datos observados antes y después de un intervalo ausente, facilitando la identificación de posibles cambios abruptos que podrían haber ocurrido durante dicho periodo. metodología, que fue seleccionada por su capacidad de modelar la incertidumbre basado en su estructura probabilística, la técnica (IFI) utiliza una estructura probabilística para generar estimaciones para preservar coherencia en la tendencia y estacionalidad para minimizar sesgo y dependencia temporal.

Una vez realizado la imputación, detección de cambios abruptos, se procedió al modelamiento y una evaluación comparativa entre modelos, enfocando en tres modelos robustos: Holt-Winters, STL(Loess) y TBATS. La elección de estos modelos se fundamento en la capacidad para descomponer señales complejas con estacionalidad múltiple, la resiliencia ante la presencia de valores atípicos (Outliers). Los resultados obtenidos, fueron evaluados mediante métricas de error por ejemplo RMSE, MAE y MAPE. El modelo TBATS, en particular mostro una eficiencia superior para capturar interacciones no lineales de la serie.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción de la situación problemática

En el entorno de la filosofía de la ciencia, Karl Popper propuso el falsacionismo como principio para distinguir la ciencia de la no ciencia: una teoría es científica solo si puede ser contradicha por la evidencia empírica. Para él, el progreso científico no se basa en confirmar teorías, sino en someterlas a pruebas que puedan demostrarlas falsas. Así, toda teoría es provisional y debe permanecer abierta a la crítica racional (Popper, 2005).

En contraste con el modelo popperiano, Thomas Kuhn cuestionó esta visión al señalar que la ciencia real no progresa solo por refutaciones, sino mediante revoluciones científicas. Según su tesis, los científicos trabajan dentro de paradigmas que orientan la investigación y solo cambian cuando las anomalías acumuladas provocan una crisis y el surgimiento de un nuevo paradigma (Kuhn, 1997).

En esta línea, la presente investigación se sitúa dentro del enfoque epistemológico de Thomas Kuhn, al considerar que la evolución metodológica en el tratamiento de series temporales, particularmente en la imputación de datos faltantes, refleja procesos de cambio paradigmático.

La falta de datos es un problema común en la adquisición de datos. Dado que los datos rara vez se pueden recopilar a la perfección y que pueden ocurrir muchos problemas, como fallos del sensor y la transmisión, errores humanos e incluso diferencias en la velocidad de recopilación, aprender a imputar datos es un paso crucial en el análisis de datos. En los casos en los que es necesario unir datos secuenciales de diferentes fuentes, el conjunto de datos obtenido con frecuencia se llena de datos faltantes. Cuando se producen datos faltantes en datos de series temporales, el problema puede comprender series temporales univariadas o multivariadas, diferentes mecanismos de datos faltantes y diferente naturaleza de las series temporales. Una buena comprensión de estos factores puede ser útil al elegir el método de imputación adecuado. También es importante elegir métodos que tengan en cuenta la información temporal intrínseca

a los datos de series temporales. (Ribeiro & Castro, 2022)

Cuando los datos de series temporales contienen valores faltantes, es una práctica común sustituirlos utilizando técnicas de imputación de valores perdidos. Sin embargo, si existe un cambio abrupto no observado dentro de los valores faltantes, las técnicas de imputación estándar resultan insuficientes, ya que tienden a sesgarse hacia un comportamiento normal de los datos. Del mismo modo, los detectores estándar no pueden identificar cambios abruptos que "se esconden" entre los datos ausentes. Para abordar estas limitaciones, se propone la técnica denominada Imputación por Intervalos de Pronóstico (Interval Forecast Imputation, IFI), la cual constituye una combinación simple e intuitiva de intervalos de incertidumbre, pronóstico y detección de anomalías. Esta técnica permite imputar un intervalo completo de tiempo con valores faltantes y, al mismo tiempo, detectar cambios abruptos en los datos de series temporales con valores ausentes. Una ventaja adicional de IFI es su compatibilidad con cualquier técnica de pronóstico de última generación. (Toller et al., 2025)

Los cambios abruptos en las series temporales son un foco de interés creciente en múltiples campos como la ecología, el clima, la medicina y la ingeniería. Estos cambios pueden originarse por diversas dinámicas subyacentes. Por un lado, pueden ser el resultado de una bifurcación matemática subyacente, lo que a menudo se describe como un "punto de inflexión"y que implica una modificación en los parámetros de control del sistema. Por otro lado, también pueden ser causados por transiciones inducidas puramente por la variabilidad estocástica (ruido) entre diferentes atractores, o por un forzamiento suficientemente rápido del sistema sin necesidad de una bifurcación. Finalmente, un proceso estocástico fuertemente autocorrelacionado (conocido como ruido blanco) puede generar lo que parecen ser cambios abruptos, si bien su clasificación como tales sigue siendo objeto de debate. (Boulton & Lenton, 2019)

Los datos de series temporales suelen contener patrones estacionales complejos que evolucionan a lo largo de diversos marcos temporales, como tendencias diarias, semanales o anuales. Los modelos tradicionales, como ARIMA y el suavizado exponencial, suelen capturar solo una estacionalidad, lo que limita su eficacia en series complejas. El modelo TBATS aborda esta limitación considerando patrones estacionales múltiples, no anidados e incluso no enteros, lo que resulta ideal para tareas de pronóstico complejas y a largo plazo. TBATS

significa Estacionalidad trigonométrica, transformación de Box-Cox, errores ARIMA, tendencia y componentes estacionales. (Despot & Arif, 2023)

1.2 Formulación del problema

1.2.1 Problema general

¿Cómo imputar por intervalos basada en pronósticos y detectar cambios abruptos para el modelamiento de una serie temporal faltante y su aplicación?.

1.2.2 Problemas específicos

- ¿Cómo imputar datos faltantes de una serie temporal en un intervalo de tiempo y detectar cambios abruptos?.
- ¿Qué modelo tiene mejor en ajuste y pronóstico para una serie temporal faltante imputada con presencia de cambios abruptos?.
- ¿Existen series temporales basadas en situaciones reales con las características estudiadas para aplicar el mejor modelo seleccionado?.

1.3 Justificación de la investigación

En primer lugar la investigación nos indica el manejo de datos faltantes en series temporales es un paso crucial, aunque a menudo ignorado, en el análisis de datos. En el trabajo, se describe la naturaleza de los datos de series temporales y los mecanismos de datos faltantes para distinguir el método de imputación adecuado, junto con una revisión de dichos métodos y su funcionamiento. Se utilizan métodos recomendados para imputar datos de distinta naturaleza (Ribeiro & Castro, 2022).

Según Ahn et al. (2022), la predicción de series temporales se ha convertido en un aspecto importante del análisis de datos y tiene numerosas aplicaciones prácticas. Sin embargo, a menudo se presentan valores faltantes indeseables, lo que puede afectar negativamente a

muchas tareas de predicción. en este aspecto nuestra investigación se centra en usar una técnica adecuada para realizar una imputación adecuada en las series temporales.

Para Andiojaya & Demirhan (2019), los métodos clásicos de análisis de series temporales no son fácilmente aplicables a series con observaciones faltantes. Para tratar la falta de datos en las series temporales, el enfoque común es utilizar técnicas de imputación para completar los huecos y obtener una serie con intervalos regulares. Sin embargo, este enfoque tiene varios inconvenientes, como sesgo de información y temporal, causalidad de relaciones y no es adecuado para series con una alta tasa de datos faltantes. En lugar de imputar directamente los valores faltantes, se propone la técnica (IFI) para mejorar la precisión de los datos faltantes.

Según Haile et al. (2024), los valores ausentes son omnipresentes en los conglomerados de datos del mundo real. Introducen incertidumbre durante el análisis de datos, lo que puede afectar las propiedades del estimador estadístico y reducir la eficiencia, llevando a conclusiones incorrectas. Por lo general, la eliminación de observaciones faltantes resulta en una valiosa pérdida de información importante, lo que sesga la estimación y la inferencia estadística.

De acuerdo con Corradin et al. (2022), una serie temporal es un tipo de datos comúnmente observado y analizado de diversas maneras en aplicaciones reales. Dentro de los posibles análisis, la detección de puntos de cambio es uno de los objetivos inferenciales cruciales para estudiar el comportamiento de una serie temporal.

La determinación de un modelo adecuado en el análisis de series de temporales es importante para describir el comportamiento de los datos. Sin embargo, debido a la diversidad de comportamientos que tienen los datos que pueden presentar, no existe un único modelo que sea optimo en todos los casos, por lo que es necesario evaluar y comparar distintos modelos mediante métricas como el RMSE, MAE, MAPE.

1.4 Objetivos de la investigación

1.4.1 Objetivo general

Analizar la imputación por intervalos basada en pronósticos y detectar cambios abruptos para el modelamiento de una serie temporal faltante y su aplicación.

1.4.2 Objetivos específicos

- Imputar datos faltantes de una serie temporal en un intervalo de tiempo y detectar cambios abruptos.
- Evaluar el modelo que tiene mejor en ajuste y pronóstico para una serie temporal faltante imputada con presencia de cambios abruptos.
- Identificar series temporales basadas en situaciones reales con las características estudiadas para aplicar el mejor modelo seleccionado.

CAPÍTULO II

MARCO TEÓRICO CONCEPTUAL

2.1 Bases Teóricas

2.1.1 Series temporales

De acuerdo con Mauricio (2007), se entiende como un conjunto de N observaciones organizadas en el tiempo de manera secuencial y con intervalos regulares las cuales describen el comportamiento de variables (serie multivariante o vectorial) correspondientes a una misma unidad de análisis a lo largo de distintos periodos.

2.1.1.1. Representación de series temporales univariantes

Para Mauricio (2007), una serie de una sola variable, las representaciones matemáticas frecuentes son:

$$x_1, x_2, \ldots, x_N; \quad (x_t)_{t=1}^{N}; \quad (x_t : t = 1, \ldots, N)$$

Donde x_t es la observación en el instante t $(1 \leq t \leq N)$ y N es el número total de observaciones que conforman la serie. Los datos pueden organizarse en un vector columna $\mathbf{x}$ con dimensión $N \times 1$:

$$\mathbf{x} \equiv [x_1, x_2, \ldots, x_N]'$$

2.1.1.2. Representación de series temporales multivariantes

Donde cada $\mathbf{x}_t \equiv [x_{t1}, x_{t2}, \ldots, x_{tM}]'$ $(M \geq 2)$ representa el vector de observaciones en el tiempo t. La serie completa de N observaciones sobre M variables se organiza en una matriz $\mathbf{X}$ de orden $N \times M$:

$$\mathbf{Y} \equiv \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix} \equiv \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{bmatrix}$$

Aquí, x_{tj} representa la observación en el momento t sobre la característica o variable j ($1 \leq j \leq M$), la cual se mantiene constante en todo momento t (Mauricio, 2007).

2.1.2 Definición de proceso estocástico

Mauricio (2007), conceptúa un proceso estocástico como una sucesión de variables aleatorias dispuestas en el tiempo de forma ordenada y con intervalos regulares, las cuales describen una sola característica o múltiples de una misma unidad de análisis a lo largo de distintos periodos.

2.1.2.1. Representaciones de procesos estocásticos univariantes

Las representaciones matemáticas más comunes para el caso univariante son:

$$\dots, X_{-1}, X_0, X_1, X_2, \dots; \quad (X_t : t = 0, \pm 1, \pm 2, \dots); \quad (X_t)$$

Donde Y_t es una variable aleatoria escalar asociada a la unidad de analisis en el instante t.

2.1.2.2. Representaciones de procesos estocásticos multivariantes

Para procesos multivariantes, la secuencia se denota:

$$\dots, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \dots; \quad (\mathbf{X}_t : t = 0, \pm 1, \pm 2, \dots); \quad (\mathbf{X}_t)$$

Donde $\mathbf{X}_t \equiv [X_{t1}, X_{t2}, \dots, X_{tM}]'$ ($M \geq 2$) representa una variable aleatoria vectorial. El componente j ($1 \leq j \leq M$) de cada $\mathbf{X}_t$ hace referencia a una característica genérica dada, la cual se mantiene constante para todo t (Mauricio, 2007).

2.1.2.3. Relación entre proceso estocástico y serie temporal

Desde una perspectiva estadística, una serie temporal puede interpretarse como una realización o muestra analizada de un proceso estocástico dentro de un periodo de tiempo determinado (Mauricio, 2007):

- Proceso estocástico: $\ldots, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N, \mathbf{X}_{N+1}, \ldots$
- Serie temporal (observada): $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$

Proceso estocástico: $\ldots, X_{-1}, X_0, \underbrace{X_1, X_2, \ldots, X_N}, X_{N+1}, \ldots$

$\downarrow$

Serie temporal: $\boxed{x_1, x_2, \ldots, x_N}$

De este modo, mientras que el proceso estocástico constituye el marco teórico de variables aleatorias, la serie temporal representa el conjunto de valores numéricos efectivamente capturados desde $t = 1$ hasta $t = N$.

2.1.3 Modelos de series de tiempo

Según Ríos (2008), los modelos de series de tiempo presentan un enfoque predictivo, dado que los pronósticos se fundamentan en el comportamiento histórico de la variable de interés. En este marco, se distinguen dos tipos principales de modelos:

- **Modelos deterministas:** Se basan en métodos de extrapolación que no consideran la naturaleza aleatoria de la serie. Su aplicación es sencilla, aunque suele implicar menor precisión. Un ejemplo son los modelos de promedio móvil, donde el pronóstico se obtiene a partir del promedio de los últimos "n"valores observados.
- **Modelos estocásticos:** Parten de la ideas simplificada del proceso aleatorio subyacente en la serie. En término sencillos, se asume que la serie observada $X_1, X_2, \ldots, X_T$ se extrae de un grupo de variables aleatorias con una cierta distribución conjunta difícil de determinar, por lo que se construyen modelos aproximados que sean útiles para la generación de pronósticos Ríos (2008).

Una serie $\{X_t\}_{t=1}^{T}$ puede ser clasificada en los siguientes:

- **Serie no estacionaria:** Se caracteriza por presentar cambios en su media, varianza y covarianza a lo largo del tiempo, lo que dificulta su modelación. No obstante, en muchos casos, al aplicar diferenciaciones una o más veces, es posible obtener una serie estacionaria..

- **Serie estacionaria:** Se define por mantener constantes su media y varianza en el tiempo, mientras que su covarianza depende únicamente del rezago. Estas propiedades permiten representar el proceso mediante modelos con parámetros constantes estimados a partir de la información histórica.

    - Media: $E(X_t) = E(X_{t+m})$ para todo t, m
    - Varianza: $\sigma(X_t) = \sigma(X_{t+m})$ para todo t, m
    - Covarianza: $cov(X_t, X_{t+k}) = cov(X_{t+m}, X_{t+m+k})$ para todo t, m, k

2.1.4 Componentes de una serie de tiempo

Ríos (2008) plantea que una serie de tiempo puede descomponerse en cuatro componentes principales, considerando además un nivel constante no directamente observable, el cual se obtiene mediante estimaciones. Estos componentes son los siguientes:

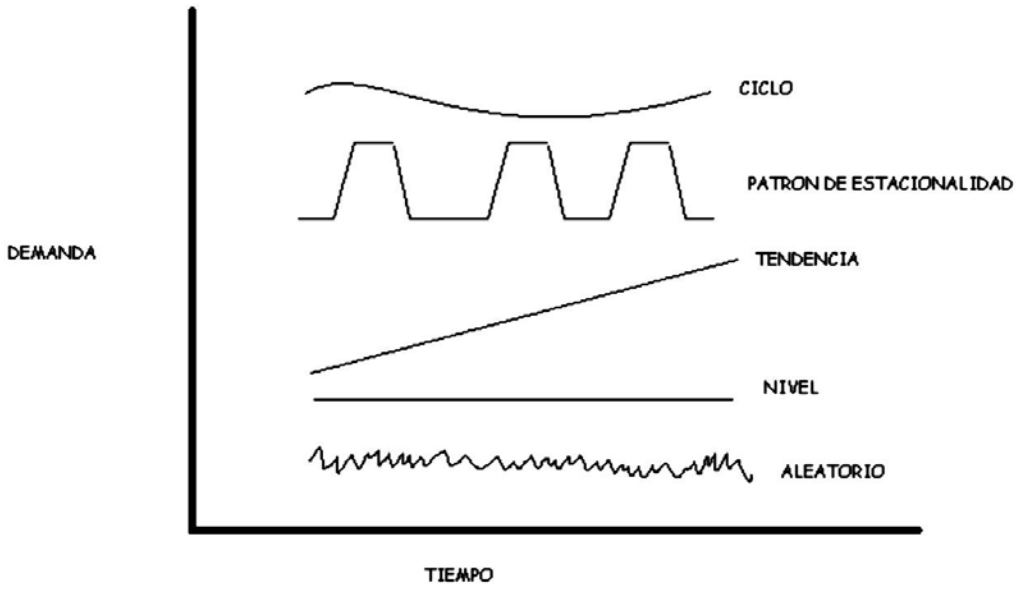
- **Tendencia (T):** Representa la dirección general de la serie a lo largo del tiempo y puede interpretarse como la evolución de su media en periodos extensos.

- **Ciclo (C):** Comprende fluctuaciones de largo plazo alrededor de la tendencia, sin una duración fija y con factores explicativos que no siempre son evidentes. Este comportamiento suele vincularse a fenómenos económicos o naturales que se desarrollan durante varios años.

- **Estacionalidad (E):** Corresponde a variaciones periódicas que se repiten en intervalos cortos y conocidos, influenciadas por factores institucionales, sociales o climáticos.

- **Aleatorio (A):** Agrupa variaciones irregulares que no responden a un patrón definido y se originan en factores impredecibles. Este componente recoge los movimientos residuales de la serie que no son explicados por la tendencia, el ciclo ni la estacionalidad.

Figura 1
Componentes de una serie de tiempo



Fuente: (*Ríos,* 2008)

2.1.5 Descomposición de una serie de tiempo

Das & Barman (2025), señalan que la descomposición de una serie de tiempo puede realizarse mediante distintos enfoques, los cuales varían en función de las características de los datos y de los objetivos planteados en el análisis. En este contexto, se reconocen diversos tipos principales de descomposición de series temporales.

2.1.5.1. Descomposición aditiva

En la descomposición aditiva, se asume que una serie temporal (TS) es la suma de sus componentes:

$$X(t) = T(t) + S(t) + R(t)$$

Tendencia $T(t)$: indica la progresión a largo plazo o la dirección de los datos.
Estacionalidad $S(t)$: captura patrones regulares y repetitivos dentro de intervalos específicos, a saber, ciclos diarios, semanales o anuales.
Residual $R(t)$: explica el ruido aleatorio o las variaciones irregulares no explicadas por la tendencia o la estacionalidad.

La descomposición aditiva es particularmente útil para TS donde las fluctuaciones estacionales e irregulares son consistentes a lo largo del tiempo. Este método se aplica a menudo en escenarios donde los datos no exhiben un crecimiento exponencial o donde el impacto de las variaciones estacionales es independiente del nivel general de la serie (Das & Barman, 2025).

2.1.5.2. Descomposición multiplicativa

En la descomposición multiplicativa, se asume que la TS es el producto de sus componentes:

$$Y(t) = T(t) \times S(t) \times R(t)$$

Tendencia $T(t)$: al igual que en la descomposición aditiva, esto refleja la dirección o el patrón subyacente.
Estacionalidad $S(t)$: aquí, las variaciones estacionales son proporcionales al nivel de la tendencia, lo que significa que aumentan o disminuyen en magnitud a medida que la tendencia cambia.
Residual $R(t)$: representa las fluctuaciones aleatorias o irregulares en los datos.

La descomposición multiplicativa se utiliza cuando la magnitud de las variaciones estacionales e irregulares está relacionada con el nivel de la tendencia. Este método es muy adecuado para TS donde las fluctuaciones aumentan o disminuyen proporcionalmente a la tendencia (Das & Barman, 2025).

2.1.5.3. Transformación logarítmica y descomposición

A veces, los datos de TS no son ni puramente aditivos ni puramente multiplicativos. En tales casos, se puede aplicar una transformación logarítmica para convertir un modelo multiplicativo en uno aditivo:

$$\log(X(t)) = \log(T(t)) + \log(S(t)) + \log(R(t))$$

En los casos en que una TS no se ajusta perfectamente a un modelo aditivo o multiplicativo, se puede emplear una transformación logarítmica. Aplicar un logaritmo a los datos transforma una serie multiplicativa en una aditiva, facilitando su descomposición y análisis. Este enfoque es fundamental en datos financieros, como precios de acciones o cifras de ventas, donde los datos pueden exhibir efectos aditivos y multiplicativos. Después de la descomposición de los datos transformados logarítmicamente, los componentes pueden elevarse a una potencia exponencial para volver a la escala original (Das & Barman, 2025).

2.1.5.4. Descomposición STL

STL es un método robusto y versátil para desagregar una TS en componentes de tendencia, estacionalidad y residual. Es beneficioso para manejar TS complejas con múltiples ciclos estacionales o patrones irregulares.

- **Estacional:** extraída usando LOESS (Suavizado de Diagramas de Dispersión Estimado Localmente), que puede adaptarse a patrones estacionales cambiantes.
- **Tendencia:** también suavizada usando LOESS, proporcionando un enfoque flexible para capturar el movimiento a largo plazo.
- **Residual:** el componente restante después de eliminar los efectos estacionales y de tendencia.

La descomposición STL ofrece una forma flexible y robusta de desglosar los datos de TS, particularmente cuando se trata de patrones estacionales complejos o irregulares. A diferencia de los métodos clásicos, STL no asume que la estacionalidad y la tendencia sean constantes a lo largo del tiempo, lo que lo hace adaptable a cambios en la estructura de datos subyacente. STL es particularmente efectivo para analizar TS con múltiples componentes estacionales, como ciclos diarios y anuales en datos climáticos, o para datos con tendencias no lineales. Personalizar los parámetros de suavizado permite a los analistas capturar variaciones sutiles

en los datos, convirtiendo a la descomposición STL en una herramienta poderosa tanto para el análisis exploratorio de datos como para el pronóstico (Das & Barman, 2025).

2.1.5.5. Descomposición clásica

Los métodos de descomposición clásica se basan en promedios móviles simples para estimar tendencias y estacionalidad.

- **Tendencia:** estimada suavizando la serie mediante un filtro de promedio móvil, eliminando los comportamientos a corto plazo.
- **Estacionalidad:** determinada restando la tendencia de la serie original e aislando el patrón estacional.
- **Residual:** representa la diferencia entre la serie original y los componentes combinados de tendencia y estacionalidad.

Los métodos de descomposición clásica se basan en promedios móviles y se encuentran entre las primeras técnicas desarrolladas para el análisis de TS. Aunque son simples, son efectivos para TS con patrones estacionales claros y estables. En la descomposición clásica, la tendencia se extrae típicamente usando un promedio móvil centrado, el cual suaviza las fluctuaciones a corto plazo.

El patrón estacional se identifica calculando el promedio de los datos sin tendencia (detrended) para cada estación. La parte restante, la diferencia entre los datos originales y los componentes combinados de tendencia y estacionalidad, se asume el residual. Este método es útil por su simplicidad y facilidad de interpretación, lo que lo transforma en un buen inicio para el análisis de TS (Das & Barman, 2025).

2.1.5.6. Descomposición X-11 y X-13-ARIMA-SEATS

Estas son técnicas estadísticas avanzadas utilizadas principalmente para TS económicas y financieras. Proporcionan ajustes estacionales detallados y son ampliamente utilizadas por agencias gubernamentales para estadísticas oficiales.

- **X-11:** un método anterior que ajusta por efectos estacionales y días comerciales.
- **X-13-ARIMA-SEATS:** un enfoque más avanzado que integra el modelado ARIMA con el proceso de descomposición para mejorar la precisión del pronóstico.

Estos métodos son particularmente valiosos cuando se trata de patrones estacionales complejos e irregularidades en los datos. Estos métodos son especialmente valiosos en las estadísticas oficiales, donde el ajuste estacional preciso resulta clave en la toma de decisiones de política informadas y análisis económicos (Das & Barman, 2025).

2.1.6 Imputación de datos faltantes

2.1.6.1. Mecanismos de datos faltantes MCAR, MAR y NNAR

Van Buuren & Van Buuren (2012), señalan que los datos faltantes pueden clasificarse según distintos mecanismos de no respuesta, entre los cuales destacan MCAR, MAR y MNAR.

MCAR(Missing Completely At Random)

Se presenta cuando la probabilidad de ausencia de datos es uniforme en todos los casos, es decir, no depende de las variables observadas ni de las no observadas. En este escenario, las causas de los datos faltantes no guardan relación con la información disponible. No obstante, en la práctica, este supuesto suele ser poco frecuente. Formalmente, los datos cumplen la condición MCAR cuando:

$$\mathrm{P}(R = 0|X_{\mathrm{obs}}, X_{\mathrm{mis}}, \phi) = \mathrm{P}(R = 0|\ \phi)$$

MAR(Missing At Random)

Se presenta cuando la probabilidad de ausencia de datos es constante dentro de grupos definidos a partir de las variables observadas. En este caso, la falta de información depende de los datos observados, pero no de los valores no observados. Este supuesto es más amplio y realista que el de MCAR, por lo que constituye la base de la mayoría de los métodos modernos de tratamiento de datos faltantes. Los datos se consideran MAR si se cumple que:

$$P(R = 0|X_{obs}, X_{mis}, \phi) = P(R = 0|X_{obs}, \phi)$$

MNAR(Missing Not At Random)

Se presenta cuando no se cumplen los supuestos de MCAR ni de MAR, lo que implica que la ausencia de datos no es aleatoria. En este caso, la probabilidad de que un dato esté ausente depende de factores no observados, lo que dificulta su tratamiento. Este tipo de mecanismo es el más complejo, ya que las causas de la falta de información no pueden identificarse directamente. Para abordarlo, se requiere recopilar información adicional sobre los motivos de la ausencia o aplicar análisis de sensibilidad que permitan evaluar la estabilidad de los resultados bajo distintos supuestos. Los datos se consideran MNAR cuando:

$$P(R = 0|X_{obs}, X_{mis}, \phi)$$

La matriz R almacena las ubicaciones de los datos faltantes en Y. La distribución de R puede depender de $X = (X_{obs}, X_{mis})$, ya sea por diseño o por casualidad, y esta relación se describe mediante el modelo de datos faltantes. Sea ψ el que contenga los parámetros del modelo de datos faltantes, entonces la expresión del modelo es $P(R|X_{obs}, X_{mis}, \phi)$ Van Buuren & Van Buuren (2012).

2.1.7 Métodos de imputación

Goicoechea (2002), señala que las técnicas de imputación pueden clasificarse en distintas categorías:

2.1.7.1. Técnicas basadas en información externa

se sustentan en variables relacionadas provenientes de otras bases de datos o en criterios previamente establecidos.

- Métodos deductivos: consisten en estimar los datos faltantes a partir de información disponible del mismo registro, con un grado razonable de certeza (Goicoechea, 2002).

2.1.7.2. Técnicas determinísticas

- **Imputación de la media:** implica reemplazar el dato ausente con la media de los valores disponibles cuando se trata de variables cuantitativas, o con la moda en el caso de variables cualitativas (Goicoechea, 2002).

  Para variables cuantitativas (media)

  $$x_i = \bar{x} = \frac{1}{n'} \sum_{j=1}^{n'} x_j$$

  Donde: x_i : valor faltante n' es el número de observaciones no faltantes.

  Para variables cualitativas (moda)

  $$x_i = \text{moda}(X)$$

  Donde: x_i : valor faltante

- **Imputación de media de clases:** consiste en agrupar las observaciones en clases mutuamente excluyentes, cada una con su propia media. Los valores faltantes se sustituyen utilizando la media correspondiente al grupo al que pertenece cada registro (Goicoechea, 2002)..

  Sea C una clase y X una variable con datos faltantes, la imputación se expresa como:

  $$x_i = \mathbb{E}[X \mid C = c_i] = \frac{1}{n_c} \sum_{j \in G(c_i)} x_j$$

  donde $G(c_i)$ es el grupo de observaciones que pertenecen a la clase c_i, y n_c su tamaño.

- **Imputación por regresión:** se basa en la estimación de un modelo lineal y, en el que la variable con datos faltantes se explica en función de un conjunto X de variables auxiliares disponibles (Goicoechea, 2002). La forma general del modelo es:

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

- **Imputación por el vecino más cercano:** este método identifica la unidad más similar a la observación con datos faltantes y, a partir de una medida de distancia entre variables x. El valor faltante se reemplaza con el de la unidad más próxima a y, considerando las variables auxiliares disponibles (Goicoechea, 2002).

$$j^* = \arg\min_j \|x_i - x_j\| \implies y_i = y_{j^*}$$

Donde:

- x_i: es el vector de variables auxiliares de la observación con dato faltante.
- x_j: es el vector de variables auxiliares de las observaciones completas.
- $\|x_i - x_j\|$: es la distancia (por ejemplo, euclidiana) entre la observación incompleta i y otra observación j.
- j^*: es el índice de la observación más cercana a i.
- y_{j^*}: es el valor observado de la variable que se quiere imputar, en el vecino más cercano.

k vecinos más cercanos:

$$x_i = \frac{1}{k} \sum_{j \in \text{vecinos}(i)} x_j$$

Donde:

- k: es el número de vecinos considerados.
- vecinos(i): es el conjunto de índices de los k vecinos más cercanos a i.
- x_j: es el valor de la variable a imputar para el vecino j.

- **Algoritmo EM (Expectation Maximization):** se fundamenta en la maximización de la función de verosimilitud y permite obtener estimaciones de máxima verosimilitud (MV) de los parámetros en presencia de datos incompletos, especialmente cuando estos presentan una estructura definida (Goicoechea, 2002).

2.1.7.3. Técnicas aleatorias o estocásticas

se caracterizan porque, al aplicar el mismo procedimiento bajo condiciones similares, pueden generar resultados distintos para una misma unidad (Goicoechea, 2002).

- **Imputación aleatoria de un caso seleccionado:** Consiste en asignar un valor faltante seleccionando de manera aleatoria un donante dentro del conjunto de datos disponibles.
- **Imputación secuencial Hot-Deck:** Cada caso es procesado secuencialmente. Si el primer registro tiene un dato faltante, este es reemplazado por un valor inicial para imputar, pudiendo ser obtenido de información externa, si el valor no está perdido, éste será el valor inicial y es usado para imputar el subsiguiente dato faltante.
- **Imputación jerárquica Hot-Deck:** Se basa en una extensión del método secuencial, en la que los datos se agrupan en clases utilizando variables auxiliares, organizadas bajo una estructura jerárquica.
- **Imputación por regresión aleatoria:** Implica, en una primera etapa, el ajuste de un modelo de regresión para estimar la variable con datos faltantes y. Posteriormente, se incorpora un término aleatorio a las predicciones, el cual puede obtenerse, por ejemplo, a partir de los residuos del modelo estimado con datos completos, seleccionados de forma aleatoria.
- **Imputación por regresión logística:** Sigue el mismo principio que la regresión anterior, pero está orientada a la imputación de variables de naturaleza binaria.
- **Imputación simple:** Consiste en sustituir los valores faltantes por estimaciones puntuales, como la media, mediana, moda o valores seleccionados aleatoriamente. Este enfoque se apoya en patrones identificables dentro de los datos observados para estimar los valores ausentes.
- **Imputación múltiple:** En lugar de asignar un único valor a cada dato faltante, este método genera varios conjuntos de datos completos, en los cuales los valores imputados varían, permitiendo incorporar la incertidumbre asociada a la imputación.

2.1.8 Imputación y detección de cambios abruptos

El método propuesto, denominado "*Interval Forecast Imputation* (IFI)", de acuerdo con Toller et al. (2025) permite superar las limitaciones tanto de las técnicas tradicionales de imputación como de los métodos clásicos de detección de cambios abruptos. Este enfoque incorpora explícitamente la incertidumbre asociada al proceso de predicción, permitiendo no solo estimar valores faltantes, sino también evaluar la consistencia estructural de la serie temporal.

2.1.8.1. Serie temporal con dato faltante

Para Toller et al. (2025), sea $\mathbf{x} = \{x_1, x_2, \ldots, x_T\}$ una serie temporal.El método IFI introduce una representación estructural basada en la segmentación.

Segmentación de la serie temporal con dato faltante

La segmentación permite distinguir datos antes y después de $\mathbf{x}_{\text{gap}}$, la serie se descompone como:

$$\mathbf{x} = \left[\mathbf{x}_{\text{before}}, \mathbf{x}_{\text{gap}}, \mathbf{x}_{\text{after}}\right]$$

donde:

- $\mathbf{x}_{\text{before}}$: observaciones previas al intervalo faltante
- $\mathbf{x}_{\text{gap}}$: intervalo completo de datos faltantes
- $\mathbf{x}_{\text{hidden}}$: valores individuales dentro del gap
- $\mathbf{x}_{\text{after}}$: observaciones posteriores al gap

2.1.8.2. Estimación del modelo de pronóstico

El objetivo es estimar $\mathbf{x}_{\text{gap}}$ mediante un modelo de pronóstico M. Se propone el uso de un modelo de pronostico M para capturar la dinamica de la serie temporal. Este modelo depende de un conjunto de parámetros θ (Toller et al., 2025).

Sea M un modelo de predicción con parámetros θ, tal que:

$$x_t = M(x_{t-1}, x_{t-2}, \ldots; \theta) + \varepsilon_t$$

donde ε_t representa un término de error aleatorio tal que:

$$\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

2.1.8.3. Imputacion del intervalo faltante

El primer paso del método consiste en la estimación de los parámetros del modelo utilizando exclusivamente la información disponible previa al intervalo faltante:

$$\hat{\boldsymbol{\theta}} = \mathrm{MLE}_M(\mathbf{x}_{\text{before}})$$

Este enfoque evita la introducción de sesgos derivados de datos faltantes y garantiza que el modelo refleje la dinámica observada del proceso.

Tras la estimación del modelo, se efectúa la imputación del intervalo faltante:

$$\hat{\mathbf{x}}_{\text{gap}} = \hat{\mathbf{x}}_{\text{hidden}} = \mathrm{forecast}_M(\mathbf{x}_{\text{before}}, h)$$

donde h representa la longitud del intervalo $\mathbf{x}_{\text{gap}}$.

Desde una perspectiva teórica, la imputación se interpreta como la estimación de la trayectoria más probable del proceso dentro del intervalo no observado, condicionada a la información previa:

$$\hat{x}_{t+h} = \mathbb{E}[X_{t+h} \mid \mathbf{x}_{\text{before}}]$$

Este enfoque modifica el problema de datos faltantes en un problema de predicción secuencial.

Además, se obtiene la predicción puntual para el siguiente valor observado:

$$\hat{x}_{t+h+1}$$

junto con la evaluación de la varianza del error de pronóstico:

$$\hat{\sigma}^2_{t+h+1} = [\hat{\sigma}^2_{t+1}, \ldots, \hat{\sigma}^2_{t+h+1}]$$

Estas cantidades permiten caracterizar completamente la distribución predictiva del proceso (Toller et al., 2025).

2.1.8.4. Interval forecast imputation (IFI)

El método IFI construye intervalos de predicción que delimitan la región de valores plausibles del proceso:

$$I_M \mid \mathbf{x}_{\text{before}} = \hat{x}_{t+h} \pm z_\alpha \hat{\sigma}_h$$

Alternativamente, el intervalo puede expresarse mediante la función cuantil de la distribución normal estándar:

$$I_{M_{\text{gap}}} = \hat{x}_{t+h+1} \pm \left| Q_{0,1}\left(\frac{\alpha}{2}\right) \right| \hat{\sigma}_{t+h+1} \tag{2.1}$$

Este intervalo define la región de valores que son consistentes con el modelo ajustado y permite evaluar la plausibilidad de las observaciones futuras (Toller et al., 2025).

2.1.8.5. Detección de cambios abruptos

Una vez imputado el intervalo $\mathbf{x}_{\text{gap}}$, el método evalúa la consistencia del modelo comparando las predicciones con los valores observados posteriores. Se define el criterio:

$$x_{t+h+1} \notin I_M \mid \mathbf{x}_{\text{before}} \Rightarrow \text{cambio estructural}$$

donde:

- x_{t+h+1}: primer valor observado después del gap
- I_M: intervalo esperado según el modelo

El método IFI evalúa la consistencia del proceso comparando las predicciones con las observaciones posteriores $\mathbf{x}_{\text{after}}$.

Se plantea el contraste de hipótesis:

- H_0: $\mathbf{x}_{\text{before}}$ y $\mathbf{x}_{\text{after}}$ provienen del mismo proceso.
- H_1: Existe un cambio estructural.

El criterio de decisión se basa en verificar si:

$$x_{t+h+1} \in I_M \mid \mathbf{x}_{\text{before}}$$

En caso contrario:

$$x_{t+h+1} \notin I_M \mid \mathbf{x}_{\text{before}} \Rightarrow \text{rechazo de } H_0$$

lo cual indica la existencia de un ruptura estructural de la serie temporal (Toller et al., 2025).

2.1.8.6. Extensión mediante modelos de pronóstico

El modelo de pronóstico M puede ser sustituido por enfoques más flexibles. En este trabajo, se propone el uso de:

$$\hat{\mathbf{x}}_{\text{gap}} = f_{\text{Prophet}}(\mathbf{x}_{\text{before}})$$

lo cual permite capturar estructuras complejas en la serie temporal, lo cual permite capturar patrones no lineales, estacionalidades múltiples y cambios estructurales de manera más robusta (Toller et al., 2025).

2.1.8.7. Pasos y análisis del algoritmo

Los pasos precisos del IFI se enumeran en el Algoritmo 1. Las líneas 1-3 describen el paso de imputación, mientras que las líneas 4-8 describen el paso de detección. El Algoritmo

1 solo devuelve $\hat{\mathbf{x}}_{\text{hidden}}$ y $\hat{\sigma}^2_{h+1}$ si detecta que no hay un cambio abrupto en $\mathbf{x}_{\text{gap}}$, ya que, de lo contrario, los valores y los intervalos imputados no son válidos.

Algorithm 1 Algoritmo 1: IFI — **Interpretación del procedimiento**

Require: $\mathbf{x}, M, \alpha$

1: $\hat{\boldsymbol{\theta}} \leftarrow \text{MLE}_M(\mathbf{x}_{\text{before}})$ ▷ Estimación de Máxima Verosimilitud

2: $\hat{\mathbf{x}}_{\text{hidden}} \leftarrow \text{forecast}_M(\hat{\boldsymbol{\theta}}, h+1)$ ▷ Pronóstico $h+1$ pasos

3: $\hat{\sigma}^2_{h+1} \leftarrow \text{var}_M(\hat{\mathbf{x}}_{\text{hidden}})$ ▷ Varianza del pronóstico

4: $I_{M|x_{\text{before}}} \leftarrow \hat{x}_{t+h+1} \pm \left|Q_{0,1}\left(\frac{\alpha}{2}\right)\right| \hat{\sigma}_{h+1}$ ▷ Ecuación

5: **if** $x_{t+h+1} \notin I_{M|x_{\text{before}}}$ **then**

6: **return** (True, [ ], [ ]) ▷ Cambio detectado, sin imputación

7: **end if**

8: **return** (False, $\hat{\mathbf{x}}_{\text{hidden}}, \hat{\sigma}^2_h$) ▷ No detectado, imputación válida

En términos de complejidad computacional, el IFI requiere recursos similares a una imputación tradicional basada en pronóstico con el modelo M. El único sobrecarga computacional relevante adicional es causada por el cálculo de $\hat{\sigma}^2_{h+1}$ en la línea 4 (Toller et al., 2025).

2.1.9 El Modelo de pronóstico prophet

Para Taylor & Letham (2018), el modelo está estructurado para poseer parámetros accesibles que pueden ser adecuados sin necesidad de conocer los aspectos profundos del modelo fundamental.

A diferencia de los modelos ARIMA, que pueden incluir covariables estacionales pero requieren tiempos de ajuste extremadamente largos y una gran experiencia técnica, este modelo se plantea como un modelo aditivo generalizado (GAM).

2.1.9.1. Componentes del Modelo

Se utiliza un modelo de serie temporal descomponible con tres componentes fundamentales: tendencia, estacionalidad y efectos de tiempo. Estos se combinan en la siguiente expresión:

$$x(t) = g(t) + s(t) + h(t) + \epsilon_t$$

Donde:

- $g(t)$ es la función de tendencia que modela cambios no periódicos.
- $s(t)$ representa cambios periódicos (ej. estacionalidad semanal y anual).
- $h(t)$ representa los efectos de los días festivos o eventos irregulares.
- ϵ_t es el término de error que representa cambios idiosincrásicos no acomodados por el modelo, asumiendo una distribución normal.

2.1.9.2. El Modelo prophet

De acuerdo con Taylor & Letham (2018) Se implementan dos modelos principales:

Crecimiento Saturado No Lineal

Para pronósticos de crecimiento que saturan en una capacidad de carga (C), se utiliza el modelo logístico básico:

$$g(t) = \frac{C}{1 + \exp(-k(t - m))}$$

Donde C es la capacidad de carga, k es la tasa de crecimiento y m es un parámetro de compensación. El modelo permite que tanto la capacidad $C(t)$ como la tasa de crecimiento sean variables en el tiempo mediante la definición explícita de puntos de cambio (changepoints).

Tendencia Lineal con Puntos de Cambio

Para casos en los que no se observa saturacion en el crecimiento, se utiliza un modelo de crecimiento lineal por tramo:

$$g(t) = (k + a(t)^\top \delta)t + (m + a(t)^\top \gamma)$$

Donde k es la tasa de crecimiento, δ contiene los ajustes de la tasa y m es el parámetro de compensación.

2.1.9.3. Estacionalidad

Para modelar efectos periódicos, el modelo depende de las series de Fourier, lo que proporciona flexibilidad para efectos estacionales suaves. La estacionalidad se aproxima mediante (Taylor & Letham, 2018):

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right)$$

Donde P es el periodo regular. Para estacionalidad anual y semanal, El número de ciclo o armónico que se va sumando para captar las sub-fluctuaciones en un periodo principal P es denotado como n.

El modelo incorpora una lista personalizada de eventos pasados y futuros. La componente estacional entonces es asumiendo que son independientes:

$$s(t) = X(t)\beta$$

Donde $X(t)$ es una matriz de regresores indicadores y $\beta \sim \text{Normal}(0, \sigma^2)$ es el parametro de cambio correspondiente en el pronóstico.

2.1.10 Ruptura estructural o cambio estructural

2.1.10.1. Bai-Perron test

Bai & Perron (1998), al usar su versión más simple (solo cambio de intercepto). Sin embargo, es importante notar que el procedimiento no se limita de ninguna manera a este caso sencillo. Puede haber otras variables en el modelo, y sus coeficientes pueden estar sujetos a cambios o permanecer constantes en la muestra. Se presenta este caso simple para facilitar la exposición utilizando la siguiente regresión lineal múltiple con m puntos de quiebre (m + 1 regímenes):

$$
\begin{aligned}
x_t &= \beta_1 Z_{t1} + E_t; \qquad && t = 1, 2, \ldots, T_1 \\
x_t &= \beta_2 Z_{t2} + E_t; \qquad && t = T_1 + 1, \ldots, T_2 \\
&\vdots \\
x_t &= \beta_j Z_{tm+1} + E_t; \qquad && t = T_{m+1}, \ldots, T
\end{aligned}
$$

De acuerdo con esta especificación, x_t es la variable dependiente observada en el tiempo t, Z_t es una columna de unos, β_j ($j = 1, 2, \ldots, m + 1$) son coeficientes (el valor de la constante para cada régimen) y E_t es el término de perturbación. Los puntos de quiebre $(T_1, \ldots, T_m)$ se tratan explícitamente como desconocidos. Buscamos estimar los coeficientes de regresión desconocidos junto con los puntos de quiebre. Esto nos permite realizar pruebas de cambios estructurales en la media de la serie.

Asi mismo (Bai & Perron, 1998), presentan estadísticos de prueba para detectar múltiples quiebres. Para cada partición m, $(T_1, \ldots, T_m)$, las estimaciones de mínimos cuadrados asociadas de β_j se obtienen minimizando la suma de los residuos al cuadrado:

$$\sum_{i=1}^{m+1} \sum_{t}^{Ti} (X_T - Z_t' \beta_j)^2$$

Esto nos proporciona $\hat{\beta}_j(T_1 \ldots T_M)$, que son las estimaciones asociadas con la partición m $(T_1 \ldots T_M)$. Al sustituirlos en la función objetivo se obtienen los puntos de quiebre estimados:

$$(\hat{T}_1, \ldots, \hat{T}_M) = \underset{T_1 \ldots T_M}{\text{mín}} \; S_T(T_1, \ldots, T_M)$$

donde $S_T(T_1, \ldots, T_M)$, es la suma de los residuos al cuadrado. Utilizando los puntos de quiebre estimados $(\hat{T}_1, \ldots, \hat{T}_M)$, las estimaciones de los parámetros encontrados son $\hat{\beta}_j(\hat{T}_1, \ldots, \hat{T}_M)$.

Bai y Perron sugieren primero un estadístico F $SupF_t(L)$ para probar la hipótesis nula de que no hay quiebres estructurales ($L = 0$) frente a la hipótesis alternativa de que existen $m = k$ quiebres. Esto prueba si $\beta_1 = \beta_2 = \cdots = \beta_{m+1}$. El procedimiento busca todas las posibles fechas de quiebre y minimiza la diferencia entre la suma de cuadrados restringida y no restringida sobre todos los quiebres potenciales.

2.1.11 Autocorrelación

En una serie de tiempo, es común que los valores de una variable no sean independientes entre sí, sino que presenten dependencia respecto a observaciones previas. En este contexto, existen distintas formas de medir dicha relación temporal entre los datos (Villavicencio, 2010).

Función de autocorrelación (ACF)

Villavicencio (2010),que mide el grado de relación entre los valores de una misma variable en distintos momentos del tiempo, separados por k periodos.

$$\rho_j = corr(X_j, X_{j-k}) = \frac{cov(X_j, X_{j-k})}{\sqrt{V(X_j)}\sqrt{V(X_{j-k})}}$$

La cual contiene las siguientes propiedades:

- $\rho_0 = 1$
- $-1 \leq \rho_j \leq 1$
- Simetría $\rho_j = \rho_{-j}$

Función de Autocorrelación Parcial (PACF)

La autocorrelación parcial evalúa la relación entre los valores de una misma variable en distintos momentos del tiempo, separados por k rezagos, controlando la influencia de los rezagos intermedios (Villavicencio, 2010).

$$\pi_j = corr(X_j, X_{j-k}/X_{j-1}X_{j-2}\dots X_{j-k+1})$$

$$\pi_j = \frac{cov(X_j - \hat{X}_j, X_{j-k} - \hat{X}_{j-k})}{\sqrt{V(X_j - \hat{X}_j)}\sqrt{V(X_{j-k} - \hat{X}_{j-k})}}$$

2.1.12 *Jarque-Bera Test*

Borja & Sánchez (2008), señalan que el supuesto de normalidad del término de error puede evaluarse mediante distintos métodos, tanto gráficos como analíticos.

Entre estos, destaca una prueba que presenta propiedades óptimas en términos de potencia asintótica, la cual se expresa de la siguiente manera:

$$JB = n\left(\frac{SK^2}{6} + \frac{(KU-3)^2}{24}\right)$$

Con $SK = \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}}$ y $KU = \frac{\hat{\mu}_4}{\hat{\mu}_2^2}$ donde $\hat{\mu}_j = \frac{\sum_{t=1}^{n}(\hat{w}_t - \bar{w})^j}{n}$, $j = 2, 3, 4$ y $\bar{w} = \frac{\sum_{t=1}^{n}\hat{w}_t}{n}$. Siendo SK y KU, representan los coeficientes muestrales de asimetría y curtosis, respectivamente. Bajo la hipótesis nula de normalidad del ruido blanco del proceso, esta estadística sigue asintóticamente una distribución chi-cuadrado con dos grados de libertad $\chi^2_{(2)}$ en el caso de series completas(serie sin datos faltantes).

2.1.13 *Dickey Fuller Test*

De acuerdo con Santra (2023), es una prueba estadística que se utiliza para verificar la estacionariedad en series temporales. Este es un tipo de prueba de raíz unitaria, a través de la cual determinamos si la serie temporal tiene alguna raíz unitaria.

La raíz unitaria es una característica de las series temporales que indica si existe alguna tendencia estocástica en la serie que la aleja de su valor medio. La presencia de una raíz unitaria hace que una serie temporal sea no estacionaria y, como resultado, genera dificultades para derivar inferencias estadísticas de la serie y predicciones futuras.

La prueba asume un modelo de serie temporal de tipo AR(1) y se representa matemáticamente como:

$$x_t = \mu + \varphi_1 x_{t-1} + \varepsilon_t$$

Después de restar x_{t-1} de ambos lados, obtenemos:

$$\Delta x_t = \mu + \delta x_{t-1} + \varepsilon_t$$

donde,

μ: Constante

δ: Coeficiente

x_{t-1}: Valor en la serie temporal con un rezago de 1

ε_t: Componente de error

La fórmula del estadístico de prueba es:

$$t_{\hat{\delta}} = \frac{\hat{\delta}}{SE(\hat{\delta})}$$

2.1.14 Mann-Kendall Test

Se efectua para determinar si una serie temporal tiene una tendencia monótona ascendente o descendente. No necesita que los datos se distribuyan normalmente o que sean lineales. Sin embargo, sí demanda que no haya autocorrelación (Zaiontz, 2012).

La hipótesis nula para esta prueba es que no hay tendencia, y la hipótesis alternativa es que existe una tendencia en la prueba de dos colas o que hay una tendencia ascendente (o descendente) en la prueba de una sola cola. Para la serie temporal $x_1, \ldots, x_n$, la prueba MK utiliza el siguiente estadístico:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign(x_j - x_i)$$

Tenga en cuenta que si $S > 0$, entonces las observaciones posteriores en la serie temporal tienden a ser mayores que las que aparecen antes, mientras que lo contrario es cierto si $S < 0$.

La varianza de S viene dada por:

$$var = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_t f_t(f_t-1)(2f_t+5)\right]$$

donde t varía sobre el conjunto de rangos ligados y f_t es el número de veces (es decir,

la frecuencia) que aparece el rango t. La prueba MK utiliza el siguiente estadístico de prueba:

$$z = \begin{cases} (S-1)/se, & S > 0 \\ 0, & S = 0 \\ (S+1)/se, & S < 0 \end{cases}$$

donde se es la raíz cuadrada de var. Si no existe una tendencia monótona (la hipótesis nula), entonces para una serie temporal con más de 10 elementos, $z \sim N(0,1)$, es decir, z tiene una distribución normal estándar.

2.1.15 Prueba de Kruskal-Wallis

Meléndez Surmay et al. (2024), describen la prueba como un método no paramétrico efectuando para comparar las medianas de dos o más muestras independientes. La hipótesis nula establece que todas las muestras provienen de una misma población, mientras que la hipótesis alternativa plantea que al menos uno de los grupos procede de una población con una mediana distinta. Este procedimiento se fundamenta en el ordenamiento de los datos mediante rangos dentro de cada grupo y constituye una alternativa al ANOVA cuando no se cumple el supuesto de normalidad. La formulación de las hipótesis se presenta a continuación:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_k \quad vs \quad H_1 : \tau_1, \ldots, \tau_k, \text{ no todos son iguales}$$

Lo cual establece que no existen diferencias significativas en los efectos de los tratamientos. La hipótesis nula define que las siguientes distribuciones $F1 = F2 = \cdots = Fk$ son iguales. Para determinar el estadístico de Kruskal-Wallis, todas las observaciones de las k muestras se combinan y se ordenan ascendentemente. Sea r_{ij} el rango de r_{ij} en esta clasificación conjunta, y Rj definido como:

$$R_j = \sum_{i=1}^{n_j} r_{ij}, \quad R_{\cdot j} = \frac{R_j}{n_j}, \quad j = 1, 2, \ldots, k$$

Así, por ejemplo, R_1 es la suma de los rangos recibidos por las observaciones del grupo 1 y $R_{\cdot 1}$ es el rango promedio para estas mismas observaciones. El estadístico H de Kruskal-Wallis

viene dado por:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left[R_{\cdot j} - \frac{N+1}{2} \right]^2$$

A un nivel de significancia α, H_0 se descarta si $H \geq h\alpha$; de lo contrario, no se rechaza. Los valores de $h\alpha$ se dan en tablas estadísticas. Cuando H_0 es verdadera, el estadístico H tiene, a medida que el mínimo de $(n_1, \ldots, n_k)$ se aproxima a infinito, una distribución chi-cuadrado χ^2 asintótica con $k-1$ gl. Bajo este supuesto, la regla de rechazo es:

$$\text{Rechazar } H \geq \chi^2_{k-1,\alpha}; \text{ de lo contrario, no rechazar.}$$

Cuando se rechaza la hipótesis nula y se deduce que al menos una muestra proviene de una población con una mediana diferente.

2.1.16 Lagrange Multiplier Test

Oliva (2019), describe la prueba de multiplicadores de Lagrange (LM) como un método utilizado para evaluar la presencia de determinadas restricciones en un modelo. Por ende la hipótesis nula se definiría de la siguiente manera:

H_0 : Los residuales no están autocorrelacionados.

H_a : Los residuales muestran alguna autocorrelación de forma autoregresiva o de medias móviles.

La prueba se fundamenta en la estimación de una regresión auxiliar, en la cual los residuos se modelan en función de las variables explicativas del modelo original y de sus propios rezagos hasta el orden m. A partir de este procedimiento, se obtiene una estadística de contraste que sigue una distribución chi-cuadrado con m grados de libertad, χ^2 , la cual se expresa de la siguiente manera:

$$LM = T \times R^2$$

Donde R^2 es el resultante de la regresión auxiliar y T es la cantidad de observaciones.

2.1.17 Conjuntos de entrenamiento y prueba

Según Hyndman & Athanasopoulos (2018), al seleccionar modelos, es práctica común dividir los datos disponibles en dos partes: *training data* y *test data* . Los *training data* se efectuan para estimar los parámetros del método de pronóstico, mientras que los *test data* se utilizan para evaluar su precisión. Dado que los *test data* no se utilizan para evaluar los pronósticos, deberían brindar una indicación fiable de la probabilidad de que el modelo realice pronósticos con datos nuevos.

Figura 2
Conjuntos de entrenamiento y prueba

Fuente: (*Hyndman & Athanasopoulos,* 2018)

El tamaño de un conglomerado de prueba suele ser cercano al 20 % de la muestra total, aunque este valor depende de la longitud de la muestra y del horizonte de pronóstico deseado. Idealmente, el conglomerado de prueba se espera que sea al menos de esa magnitud. Cabe destacar los siguientes puntos.

- Un modelo que se ajusta bien a los datos de entrenamiento no necesariamente hará buenas predicciones.
- Siempre se puede obtener un ajuste perfecto efectuando un modelo con un nivel adecuado de parámetros.
- El sobreajuste del modelo es tan problemático como no detectar un patrón sistemático en los datos.

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

Hernández (2024), señala que el modelo $SARIMA$ (Seasonal AutoRegressive Integra-

ted Moving Average) constituye una extensión del modelo $ARIMA$, diseñada para incorporar y modelar patrones estacionales presentes en las series temporales. Este enfoque resulta especialmente pertinente cuando los datos evidencian una periodicidad definida. La incorporación de componentes estacionales permite representar dichos patrones de manera explícita, lo que contribuye a mejorar la precisión de los pronósticos en contextos donde la estacionalidad es significativa.

En este modelo, se integran conjuntamente los componentes no estacionales y estacionales de la siguiente manera:

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t$$

en el que B es el operador de retardo y ε_t es el ruido blanco. donde:

- p: orden del componente autorregresivo AR.
- d: orden de diferenciación no estacional.
- q: orden del componente de medias móviles MA.
- P: orden del componente autorregresivo estacional SAR.
- D: orden de diferenciación estacional.
- Q: orden del componente de medias móviles estacional SMA.
- s: periodicidad estacional.

Este modelo incorpora componentes específicos para modelar patrones periódicos, introduciendo términos como P, D, Q y s. Puede capturar tanto dinámicas no estacionales como estacionales en la misma serie temporal. Requiere una diferenciación estacional $(1 - B^s)^D$ cuando los datos exhiben estacionalidad. Para tomar en cuenta los órdenes P, D, Q depende de la periodicidad s y la estructura propia en los datos. La elaboración de un modelo sigue una estrategia similar a $ARIMA$, pero con las siguientes adaptaciones:

1. *Identificación de parámetros*: usar ACF y $PACF$ para identificar p, d, q y P, D, Q, considerando la periodicidad s.

2. *Estimación*: ajustar el modelo con los parámetros elegidos.

2.1.18.1. Criterios Adicionales para la Identificación

Además del análisis de los correlogramas, es posible emplear herramientas complementarias para la identificación del modelo:

- **Pruebas de independencia de residuos:** Se aplican contrastes estadísticos, como la prueba de Ljung-Box, con el fin de comprobar que los residuos no presentan autocorrelación.
- **Evaluación de residuos:** Se realiza un análisis gráfico para verificar la ausencia de patrones sistemáticos en los residuos.

Una adecuada identificación del modelo resulta fundamental, ya que permite representar correctamente la dinámica temporal de los datos y obtener pronósticos consistentes (Hernández, 2024).

2.1.19 Modelo Holt-Winters

Sanz (2023), indican que el método Holt-Winters resulta apropiado para el análisis de series temporales que presentan tanto tendencia como estacionalidad marcada. En este enfoque, la componente de tendencia se modela mediante el método de Holt, el cual constituye una extensión del alisado exponencial simple, mientras que la estacionalidad se incorpora a través de la ampliación propuesta por Winters sobre dicho método.

Para un comportamiento aditivo la expresión matemática de Holt-Winters para una serie con k periodos y h número de periodos a pronosticar sería:

$$\hat{X}_{t+h} = a_t + h.b_t + S_{t+k+1+(h-1) \text{ mód } k}$$

Donde a_t, b_t y S_t se obtienen como:

$$a_t = \alpha(X_t - S_{t-k}) + (1-\alpha)(a_{t-1} + b_{t-1})$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}$$

$$S_t = \gamma(X_t - a_t) + (1 - \gamma)S_{t-k}$$

Para el modelo multiplicativo:

$$\hat{X}_{t+h} = a_t + h.b_t.S_{t+k+1+(h-1) \text{ mód } k}$$

Donde a_t, b_t y S_t vienen dados por:

$$a_t = \alpha\left(\frac{X_t}{S_{t-k}}\right) + (1 - \alpha)(a_{t-1} + b_{t-1})$$

$$b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}$$

$$S_t = \gamma\left(\frac{Y_t}{a_t}\right) + (1 - \gamma)S_{t-k}$$

Donde α representa el coeficiente de suavizamiento del nivel, β corresponde al coeficiente asociado a la tendencia y (γ) al componente estacional. De manera similar a los modelos ARIMA, el método Holt-Winters genera pronósticos a partir de la combinación del nivel, la tendencia y la estacionalidad. Estas componentes se incorporan en las ecuaciones mediante los parámetros (α), (β) y (γ). Cuando β=(γ)=0 (o false en el entorno R), el modelo se reduce al alisado exponencial simple. Por su parte, si γ=0, no se considera la componente estacional en la estimación.

2.1.20 Definición del método STL (Seasonal and trend decomposition using loess)

En Cleveland et al. (1990), el método STL utiliza el suavizamiento LOESS para descomponer una serie temporal en componentes. A continuación, se describe el suavizamiento local que constituye la base del método.

2.1.20.1. Suavizamiento LOESS

Sea un conjunto de datos $\{(x_i, y_i)\}_{i=1}^{n}$. El objetivo es estimar una función suave $g(x)$ que aproxime la relación entre x e y.

Para cada punto x, se seleccionan los q valores de x_i más cercanos y se define $A_q(x)$ como la distancia al q-ésimo vecino más lejano (Cleveland et al., 1990). Los pesos se obtienen mediante la función tricúbica:

$$W(u) = \begin{cases} (1-u^3)^3, & 0 \leq u < 1 \\ 0, & u \geq 1 \end{cases}$$

Los pesos locales son:

$$v_i(x) = W\left(\frac{|x_i - x|}{A_q(x)}\right)$$

Luego, se ajusta un polinomio local de grado d (generalmente $d = 1$ o $d = 2$) mediante mínimos cuadrados ponderados. El valor suavizado se obtiene como:

$$g(x) = \hat{y}(x)$$

$$g(x) = \hat{y}(x) = \underset{\beta_0,\beta_1}{\arg\min} \sum_{i=1}^{n} w_i(x)\,[y_i - \beta_0 - \beta_1(x_i - x)]^2$$

El parámetro q controla el nivel de suavizamiento: valores grandes producen curvas más suaves. Además, si cada observación tiene un peso p_i, estos se incorporan como:

$$w_i(x) = p_i \cdot v_i(x)$$

permitiendo considerar la confiabilidad de los datos.

2.1.20.2. El Algoritmo STL: Bucles interno y externo

El algoritmo STL opera mediante un bucle interno anidado dentro de un bucle externo. Los componentes de la serie temporal, estacional $S_v^{(k)}$ y de tendencia $T_v^{(k)}$, se definen en todos los puntos $v = 1$ a N, incluso si el valor de la serie original Y_v falta (Cleveland et al., 1990).

2.1.20.3. El bucle interno.

Cada pasada del bucle interno consiste en un suavizado estacional seguido de un suavizado de tendencia, lo que actualiza los componentes a $S_v^{(k+1)}$ y $T_v^{(k+1)}$ para la $(k+1)$-ésima iteración. El proceso consta de seis pasos:

- **Paso 1: Destendencia.** Se calcula la serie destendenciada:

$$X_v - T_v^{(k)}$$

  Si X_v está ausente, la serie destendenciada también lo está. En la pasada inicial ($k = 0$), se utiliza un valor de inicio $T_v^{(0)} \equiv 0$.

- **Paso 2: Suavizado de ciclo-subseries ($v_i(x)$).**

  Las $n_{(p)}$ **ciclo-subseries** (por ejemplo, todos los valores de enero para una serie mensual) de la serie destendenciada se suavizan separadamente usando Loess con grado $d = 1$ y parámetro de vecindario $q = n_{(s)}$. Los valores suavizados se combinan para formar la serie estacional temporal $C_v^{(k+1)}$. El suavizado se extiende $n_{(p)}$ posiciones en cada extremo en anticipación de la pérdida de datos en el filtrado de paso bajo.

- **Paso 3: Filtrado de paso bajo.**

  Se aplica un filtro de paso bajo a $C_v^{(k+1)}$ para obtener $L_v^{(k+1)}$. El filtro típicamente consiste en: media móvil de longitud $n_{(p)}$, seguida de otra de longitud $n_{(p)}$, seguida de una de longitud 3, y finalmente un suavizado Loess con $d = 1$ y $q = n_{(l)}$.

- **Paso 4: Destendencia del suavizado.**

  La componente estacional de la iteración actual se obtiene restando el componente de baja frecuencia $L_v^{(k+1)}$:

$$S_v^{(k+1)} = C_v^{(k+1)} - L_v^{(k+1)}$$

  Esto previene que la potencia de baja frecuencia (tendencia) ingrese al componente estacional.

- **Paso 5: Desestacionalización.** Se calcula la serie desestacionalizada:

$$X_v - S_v^{(k+1)}$$

- **Paso 6: Suavizado de tendencia ($T_v^{(k+1)}$).**

  La serie desestacionalizada se suaviza mediante Loess con $d = 1$ y $q = n_{(l)}$ para obtener el nuevo componente de tendencia $T_v^{(k+1)}$. Este suavizado se computa en todas las posiciones v, incluyendo las faltantes.

Los Pasos 2, 3 y 4 constituyen la porción de suavizado estacional; el Paso 6 constituye la porción de suavizado de tendencia. En las pasadas posteriores a la inicial, se utiliza el componente de tendencia $T_v^{(k)}$ de la pasada anterior, en lugar de $T_v^{(0)} \equiv 0$ (Cleveland et al., 1990).

2.1.20.4. El Bucle externo.

El bucle externo realiza $n_{(o)}$ iteraciones de robustez. Después de una pasada inicial del bucle interno, se estima el **residuo** (R_v) y se calculan los **pesos de robustez** (ρ_v):

$$R_v = X_v - T_v - S_v$$

1. **Cálculo del peso de robustez (ρ_v):** Los pesos reflejan lo extremo que es el residuo R_v. Se establece $h = 6 \cdot \text{median}(|R_v|)$. El peso de robustez en el tiempo v es:

$$\rho_v = B(|R_v|/h)$$

donde B es la función de peso **bisquare**:

$$B(u) = \begin{cases} (1-u^2)^2 & \text{para } 0 \leq u < 1 \\ 0 & \text{para } u > 1 \end{cases}$$

Un valor atípico con $|R_v|$ grande tendrá un peso pequeño o cero.

2. **Aplicación de robustez:** En las pasadas subsiguientes del bucle interno (Pasos 2 y 6), el peso de vecindario se multiplica por ρ_v. El bucle externo se repite un total de $n_{(o)}$ veces (Cleveland et al., 1990).

2.1.20.5. Post-suavizado del componente Eestacional.

Para componentes estacionales localmente rugosos, se aplica un **post-suavizado** (post-smoothing) final mediante Loess. Cleveland et al. (1990), recomiendan usar:

- Ajuste **localmente cuadrático** ($d = 2$) debido a los picos y valles inherentes al componente estacional.
- El parámetro de vecindario q puede ser pequeño o moderado.
- No se requieren iteraciones de robustez en el post-suavizado.

2.1.20.6. Parámetros de STL.

STL tiene seis parámetros clave:

- $n_{(p)}$: Número de observaciones en cada ciclo estacional.
- $n_{(i)}$: Número de pasadas a través del bucle interno.
- $n_{(o)}$: Número de iteraciones de robustez del bucle externo.
- $n_{(l)}$: Parámetro de suavizado (vecindario q) para el filtro de paso bajo y el componente de tendencia.
- $n_{(t)}$: Parámetro de suavizado (vecindario q) para el componente de tendencia (generalmente $n_{(t)} = n_{(l)}$ en la práctica).
- $n_{(s)}$: Parámetro de suavizado (vecindario q) para el componente estacional.

El parámetro $n_{(s)}$ es el más crítico y debe ser ajustado cuidadosamente para cada aplicación (Cleveland et al., 1990).

2.1.21 Modelo TBATS (Trigonometric, Box-Cox transform, ARMA errors, Trend and Seasonal component)

Está diseñado para modelar series temporales que presentan **múltiples componentes estacionales** independientes, un desafío que los modelos de suavizado exponencial tradicionales no pueden resolver eficazmente. Su estructura combina una serie de elementos clave para capturar la complejidad de los datos (Suárez et al., 2022).

2.1.21.1. Transformación de Box-Cox

El modelo inicia con la **Transformación de Box-Cox** para abordar problemas de **no linealidad** y **valores no negativos** en la serie temporal $\{X_t\}$. La serie transformada $X_t^{(\omega)}$ se define como:

$$X_t^{(\omega)} = \begin{cases} \frac{X_t - 1}{\omega} & \text{si } \omega \neq 0 \\ \log X_t & \text{si } \omega = 0 \end{cases} \tag{3}$$

donde $\omega \in \mathbb{R}$ es el parámetro de la transformación (Suárez et al., 2022).

2.1.21.2. Actualización de la Serie Transformada

La **serie transformada** en el instante t, $X_t^{(\omega)}$, se calcula como la suma de las estimaciones de sus componentes principales en el instante anterior $(t-1)$, más un término de error (d_t).

$$X_t^{(\omega)} = \mu_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{j,t}^{(i)} + d_t \tag{4}$$

donde:

- μ_{t-1} y b_{t-1} son el nivel y la pendiente de la tendencia.
- ϕ es un **factor atenuante** que modera la influencia de la pendiente.

- $\sum_{i=1}^{T} s_{j,t}^{(i)}$ es la suma de las T componentes estacionales.
- d_t es el error modelado por un proceso ARMA.

El error residual d_t se modela mediante un proceso **ARMA**(p, q), para eliminar la correlación que no haya sido capturada por las componentes de nivel, tendencia o estacionalidad:

$$d_t = \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \tag{5}$$

donde φ_i y θ_i son los coeficientes del modelo ARMA, y $\{\epsilon_t\}$ es un ruido blanco normal (Suárez et al., 2022).

2.1.21.3. Actualización de la Tendencia (Nivel y Pendiente)

Las componentes de la **tendencia** (nivel μ_t y pendiente b_t) se actualizan mediante ecuaciones de **suavizado exponencial atenuado**:

$$\mu_t = \mu_{t-1} + \phi b_{t-1} + \alpha d_t \tag{6}$$

$$b_t = (1 - \phi) b_{t-1} + \phi b_{t-1} + \beta d_t \tag{7}$$

donde α y $\beta \in [0, 1]$ son los **parámetros de suavizado**. El término ϕb_{t-1} incorpora el concepto de que la pendiente futura puede estar **atenuada** (disminuir a largo plazo), a menos que $\phi = 1$ (Suárez et al., 2022).

2.1.21.4. Cálculo y Actualización de la Componente Estacional

La i-ésima componente estacional, $s_{j,t}^{(i)}$, se representa mediante una **Serie de Fourier** para capturar su periodicidad:

$$s_{j,t}^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)} \tag{8}$$

donde k_i es el número de armónicos necesarios para la componente i. Las actualizaciones de

los términos de Fourier $s_{j,t}^{(i)}$ y $s_{j,t}^{*(i)}$ son:

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin \lambda_j^{(i)} + \gamma_1^{(i)} d_t \tag{9}$$

$$s_{j,t}^{*(i)} = -s_{j,t-1}^{(i)} \sin \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos \lambda_j^{(i)} + \gamma_2^{(i)} d_t \tag{10}$$

Aquí, $\lambda_j^{(i)} = \frac{2\pi j}{m_i}$ es la frecuencia angular del j-ésimo armónico, y m_i es el **período** de la i-ésima componente estacional. $\gamma_1^{(i)}$ y $\gamma_2^{(i)}$ son los **parámetros de suavizado** para la componente estacional (Suárez et al., 2022).

2.1.21.5. Formulación Compacta del Espacio de Estados

Finalmente, el modelo se formula de manera compacta como un **modelo de espacio de estados** para facilitar la implementación algorítmica y la iteración para la previsión.

$$X_t^{(\omega)} = \mathbf{w}'\mathbf{x}_{t-1} + \epsilon_t \tag{11}$$

$$\mathbf{y}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{g}\epsilon_t \tag{12}$$

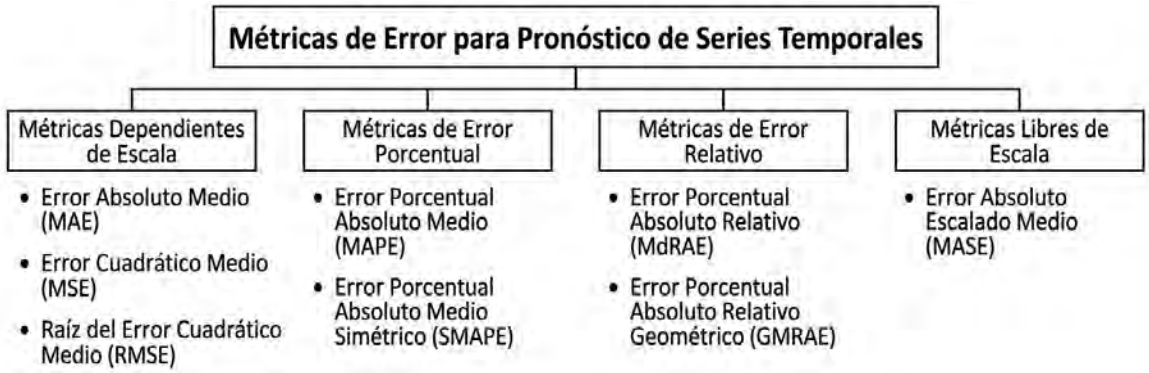
- $\mathbf{x}_t$ es el **vector de estado**, que contiene todas las componentes del modelo (μ_t, b_t, y los términos de Fourier $s_{j,t}^{(i)}$ y $s_{j,t}^{*(i)}$).
- $\mathbf{w}$ es el vector que extrae la información necesaria del estado para calcular la **ecuación de observación** (11).
- $\mathbf{F}$ es la **matriz de actualización o transición**, que define cómo el vector de estado evoluciona de $t-1$ a t.
- $\mathbf{g}$ es un vector columna que recoge el impacto del error ϵ_t en la **ecuación de transición** (12) (Suárez et al., 2022).

2.1.22 Métricas de precisión del modelo o métricas de error de ajuste.

Según Rink (2021), la evaluación de modelos de series temporales se realiza mediante métricas que cuantifican la diferencia de los valores observados y pronosticados. Estas métricas permiten determinar la precisión de los modelos y comparar su desempeño bajo distintos criterios.

Figura 3

Métricas de error de pronóstico de series temporales



Fuente: (*Rink,* 2021)

2.1.22.1. Error absoluto medio (MAE)

Cuantifica el promedio de las diferencias en valor absoluto entre los valores observados y los estimados.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|x_t - \hat{x}_t|$$

Esta métrica es sencilla de interpretar, ya que mantiene las mismas unidades de la variable analizada.

2.1.22.2. Error cuadrático medio (MSE)

Este indicador mide la media de los errores al cuadrado entre los valores observados y los estimados por el modelo.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{x}_t)^2$$

Penaliza en mayor medida los errores grandes por la elevación al cuadrado.

2.1.22.3. Raíz del error cuadrático medio (RMSE)

Es la raíz cuadrada del MSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{x}_t)^2}$$

Permite expresar el error en las mismas unidades de la variable original.

2.1.22.4. Error porcentual absoluto medio (MAPE)

El MAPE expresa el error en términos porcentuales:

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{x_t - \hat{x}_t}{x_t}\right|$$

Es útil para comparar modelos en diferentes escalas, aunque presenta limitaciones cuando los valores reales son cercanos a cero.

2.1.22.5. Error porcentual absoluto medio simétrico (sMAPE)

El sMAPE corrige algunas limitaciones del MAPE al utilizar el promedio entre valores reales y estimados:

$$sMAPE = \frac{100}{n}\sum_{t=1}^{n}\frac{|x_t - \hat{x}_t|}{(|x_t| + |\hat{x}_t|)/2}$$

Reduce la asimetría en la medición del error porcentual.

2.1.22.6. Error relativo absoluto mediano (MdRAE)

El MdRAE se basa en el error relativo absoluto mediano respecto a un modelo de referencia:

$$MdRAE = \text{mediana}\left(\frac{|x_t - \hat{x}_t|}{|x_t - x_{t-1}|}\right)$$

Es robusto frente a valores extremos.

2.1.22.7. Error relativo absoluto geométrico medio (GMRAE)

El GMRAE mide el promedio geométrico de los errores relativos:

$$GMRAE = \left(\prod_{t=1}^{n} \frac{|x_t - \hat{x}_t|}{|x_t - x_{t-1}|} \right)^{\frac{1}{n}}$$

Permite evaluar el desempeño relativo de los modelos considerando la variabilidad de la serie.

2.1.22.8. Error absoluto medio escalado (MASE)

El MASE compara el error del modelo con un modelo ingenuo de referencia:

$$MASE = \frac{\frac{1}{n} \sum_{t=1}^{n} |x_t - \hat{x}_t|}{\frac{1}{n-1} \sum_{t=2}^{n} |x_t - x_{t-1}|}$$

Un valor menor a uno indica que el modelo propuesto supera al modelo ingenuo.

El uso conjunto de estas métricas permite una evaluación global del desempeño de los modelos, con base en la magnitud absoluta del error como su comportamiento relativo. Por ello, la selección del modelo óptimo debe basarse en un análisis comparativo de múltiples indicadores.

2.1.23 Métricas capacidad predictiva

2.1.23.1. Theil's U

Patiño & Alonso (2005), indican que la capacidad predictiva de un modelo puede evaluarse mediante el coeficiente de desigualdad de Theil, el cual se basa en la raíz del error cuadrático medio (RMS) del pronóstico. Este coeficiente se define de la siguiente manera:

$$U = \frac{\sqrt{\frac{1}{T} \sum (X_t^s - X_t^a)^2}}{\sqrt{\frac{1}{T} \sum_{t=1}^{T} (X_t^s)^2} + \sqrt{\frac{1}{T} \sum_{t=1}^{T} (X_t^a)^2}}$$

expresa la raíz del error cuadrático medio en términos relativos

2.1.24 Ruido blanco ("white noise")

Villavicencio (2010), define el ruido blanco como un caso particular de proceso estocástico, caracterizado por una secuencia de variables aleatorias independientes e idénticamente distribuidas en el tiempo, con media cero y varianza constante. Este proceso se representa comúnmente como ε_t.

$$\varepsilon_t \sim N(0, \sigma^2) \qquad cov(\varepsilon_{t_i}, \varepsilon_{t_j}) = 0 \quad \forall t_i \neq t_j$$

2.2 Marco conceptual

- **Serie temporal** se define como una secuencia de observaciones registradas en distintos momentos, organizadas cronológicamente y separadas por intervalos regulares. En este tipo de datos, es común que exista dependencia entre las observaciones. El objetivo principal del análisis de una serie temporal X_t , con $t = 1, 2, \ldots, n$ es comprender su comportamiento y generar pronostico (Villavicencio, 2010).

- **Imputación** La imputación es un método general y flexible para manejar problemas de datos faltantes. Consiste en sustituir valores faltantes en un conglomerado de datos por valores estimados mediante métodos estadísticos (Little & Rubin, 2019).

- **Pronóstico** Se busca generar predicciones con el mayor nivel de precisión posible, considerando la información disponible, tanto histórica como prospectiva, que pueda incidir en los resultados del pronóstico (Hyndman & Athanasopoulos, 2018).

- **Cambio abrupto** Se define como un cambio repentino en la estructura estadística de una serie temporal, que se muestra a través de alteraciones importantes en parámetros como la varianza, la media, la tendencia o la distribución probabilística de los datos. (Aminikhanghahi & Cook, 2017).

- **Modelado** modelado de series de tiempo incluye métodos para analizar observaciones a lo largo del tiempo, con el fin de crear modelos que expliquen la comportamiento de la serie y posibiliten hacer pronósticos.(Box & Jenkins, 1976).

2.3 Antecedentes

2.3.1 Antecedentes internacionales

Toller et al. (2025), en su investigación titulada ***"Detecting abrupt changes in missing time series data"***, propone la Imputación de Pronóstico por Intervalos (IFI), que es una combinación simple e intuitiva de intervalos de incertidumbre, pronóstico y detección de anomalías que detecta cambios abruptos en los datos faltantes de series temporales. Se llegó a la conclusión de que el método (IFI) puede detectar cambios abruptos que faltan de forma realista, tiene un rendimiento consistentemente mejor que las líneas de base comparadas, tiene una caída de rendimiento favorablemente lenta a medida que se hace más largo, y conduce a un mejor rendimiento de imputación.

Dosantos (2020), en su tesis titulada *"Comparación de Modelos de Predicción para Series Temporales"*, realiza la comparación de modelos de series temporales: AR, MA, ARMA, ARIMA, Holt-Winters y el modelo TBATS. Los hallazgos indican que, para la generación de pronósticos a corto plazo, el modelo TBATS resulta adecuado independientemente de la presencia de tendencia o estacionalidad en la serie, con excepción de las series estacionarias, en las que el modelo ARIMA ofrece un mejor desempeño. En cuanto a las predicciones a largo plazo en series con tendencia acotada, se sugiere el uso del modelo TBATS cuando existe estacionalidad, mientras que, en ausencia de esta, pueden emplearse otros modelos, excluyendo el de Holt-Winters.

Paredes (2020), en su investigacion *"Predicción de múltiples series de tiempo univariadas a través de diversos modelos predictivos y meta-learning aplicado en la industria del retail"* examina el rendimiento de varios modelos, incluyendo STLF, SARIMA, Holt, TBATS, NNetAR, media móvil y un modelo de ensamble. El objetivo de elegir estos elementos fue el de captar una variedad de comportamientos en las series, como la autorregresión, la tendencia, la estacionalidad, los procesos de suavizamiento, las funciones trigonométricas y las transformaciones Box-Cox. Su finalidad es reconocer un modelo predictivo que supere la precisión de la media móvil en los pronósticos del vestuario y juguetes en los supermercados, ayudando de esta manera a disminuir los gastos logísticos que surgen a partir de fallos en la predicción. Al

comparar los resultados durante un período de cuatro semanas, se determina que el ensamble que consiste en los métodos TBATS y Holt es el más estable.

Moro (2023), en su investigación *"Análisis y predicción de series temporales para consumo eléctrico y presión en el supercomputador Finis Terrae II empleando Spark"*, analizan una variedad de enfoques predictivos, entre ellos, modelos tradicionales fundamentados en autorregresión; Prophet; el método de suavizado exponencial (Holt-Winters); y un modelo mixto que fusiona suavizado exponencial, autorregresión y media móvil, con la inclusión de términos extra para captar múltiples patrones estacionales (TBATS). Los resultados muestran que este modelo más reciente tiene un desempeño superior al predecir series de consumo energético que se caracterizan por la existencia de diversas estacionalidades.

Xian et al. (2023), es su investigación titulado *"Comparison of SARIMA model, Holt-winters model and ETS model in predicting the incidence of foodborne disease"*, Estudia la incidencia de enfermedades transmitidas por los alimentos, se obtuvieron del Centro de Prevención y Control de Enfermedades del Distrito de Nanan, Chongqing. se utilizaron para la predicción y validación. La incidencia se ajustó utilizando el modelo SARIMA (media móvil integrada autorregresiva estacional), el modelo de Holt-Winters y el modelo de suavizado exponencial (ETS). Además, utilizamos MSE, MAE y RMSE para determinar qué modelo se ajusta mejor.El MSE, MAE y RMSE del modelo de Holt-Winters, valores inferiores a los modelos SARIMA y ETS.

2.3.2 Antecedentes nacionales

Alcántara et al. (2023), en su estudio *"Modelos de series temporales predictivos para la demanda turística en Perú"*, buscan determinar el modelo predictivo más exacto para calcular la demanda turística del país, a través de la comparación entre diferentes técnicas de series temporales. Con este fin, emplean información histórica sobre la llegada mensual de turistas internacionales, que el Ministerio de Comercio Exterior y Turismo (MINCETUR) proporciona. Tres enfoques se tienen en cuenta en el análisis: Holt-Winters, ARIMA y regresión lineal, que son analizados con respecto a su capacidad para hacer predicciones. Los hallazgos indican que el modelo de regresión lineal tiene el mejor rendimiento, luego le sigue ARIMA y, finalmente,

Holt-Winters, con errores del 5.26 %, 5.56 % y 6.34 %, en ese orden, lo que demuestra un buen ajuste.

Tudela et al. (2022), en su estudio titulado *"Impacto del COVID-19 en la demanda de turismo internacional del Perú. Una aplicación de la metodología Box-Jenkins"*, analizan y proyectan la demanda de turismo internacional en el país utilizando datos mensuales correspondientes al periodo enero de 2003 a diciembre de 2020. Para ello, aplican un modelo ARIMA estacional bajo el enfoque Box-Jenkins, específicamente un SARIMA $(1,1,1)(0,1,1)_{12}$. No obstante, considerando las medidas adoptadas a nivel internacional para priorizar la salud pública durante la pandemia de COVID-19, así como la reactivación progresiva de los viajes en algunos países, el modelo proyecta una recuperación gradual y de carácter cíclico en la llegada de turistas internacionales al Perú.

Sernaque et al. (2022), en su investigación titulada *"Comparison of Arima and Holt-Winters forecasting models for time series of cereal production in Peru"*, presenta un estudio de los productos agrícolas presentan una notable volatilidad en sus niveles de producción, lo que afecta gravemente a los agricultores. La dinámica variacional de los precios de los insumos utilizados y las constantes variaciones en las condiciones climáticas tienen una influencia significativa en la cadena de producción de cereales en Perú; por lo tanto, comparado con el modelo ARIMA, el modelo de pronóstico aditivo de Holt-Winters presentó un mejor ajuste según el Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiano (BIC).

CAPÍTULO III

HIPÓTESIS Y VARIABLES

3.1 Hipótesis

3.1.1 Hipótesis general

Existe un método para Imputar por intervalos basada en pronósticos y detectar cambios abruptos que mejora en la precisión en el modelamiento de una serie temporal faltante y su aplicación.

3.1.2 Hipótesis específicas

- La imputación mediante Interval Forescat Imputation (IFI) es un método eficaz para la imputación de datos faltantes en un intervalo de tiempo y detectar cambios abruptos con prophet en una serie temporal.
- El modelo TBATS es el modelo que tiene mejor en ajuste y pronóstico para una serie temporal faltante imputada con presencia de cambios abruptos.
- Existen series temporales basadas en situaciones reales con las características estudiadas, como: llegadas de visitantes a la Ciudad Inka de Machu Picchu (2002 – 2025) y Movimiento Internacional de pasajeros en el aeropuerto Internacional Alejandro Velasco Astete (2004 – 2025).

3.2 Identificación de variables

La imputación mediante prophet es un método eficaz para la imputación de datos faltantes en un intervalo de tiempo y detectar cambios abruptos con Interval Forescat Imputation (IFI) en una serie temporal.

3.2.1 Variables dependientes

Representa la sucesión de observaciones ordenadas cronológicamente cuyo comportamiento futuro se desea estimar a partir de sus valores históricos y patrones internos.

- Número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).
- Número de pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004–2025).

3.2.2 Variables independientes

En este campo, el tiempo se considera la variable explicativa fundamental.

- Tiempo (2002-2025).
- Tiempo (2004-2025).

3.3 Operacionalización de variables

Tabla 1
Matriz de operacionalización de variables

Variable	Definición	Indicador	Tipo	Escala	Unidad de medida
Dependiente					
Número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025)	Corresponde a la cantidad total de personas que ingresan al santuario histórico de Machu Picchu durante un periodo determinado.	Total de visitantes nacionales y extranjeros registrados en el sitio arqueológico por mes o año.	Discreto	Razón	Personas
Número de pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004–2025)	Se refiere al número total de pasajeros que realizaron vuelos internacionales de llegada o salida a través del aeropuerto en un periodo determinado.	Total mensual o anual de pasajeros internacionales registrados por la autoridad aeroportuaria.	Discreto	Razón	Pasajeros

Variable	Definición	Indicador	Tipo	Escala	Unidad de medida
Independiente					
Tiempo (2002–2025).	Corresponde a la secuencia cronológica de observaciones de la serie temporal asociada al número de visitantes de Machu Picchu.	Orden temporal de los registros (meses o años).	Continua	Razón	Mes
Tiempo (2004–2025).	Representa la secuencia temporal de los registros correspondientes al número de pasajeros internacionales del Aeropuerto Alejandro Velasco Astete.	Orden temporal de los registros (meses o años).	Continua	Razón	Mes

CAPÍTULO IV

METODOLOGÍA

4.1 Ámbito de estudio: localización política y geográfica

Según Hernández-Sampieri & Mendoza (2018), el ámbito de estudio delimita el espacio geográfico, temporal y contextual en el que se desarrollará la investigación, proporcionando los límites necesarios para situar el fenómeno dentro de un entorno definido y coherente con los objetivos del estudio. Sin embargo, en el presente trabajo no se establece un ámbito geográfico específico, dado que su propósito no es intervenir en una zona determinada, sino desarrollar y fundamentar una técnica de imputación de datos faltantes aplicable a distintas series temporales.

4.2 Tipo de investigación

De acuerdo con Hernández-Sampieri & Mendoza (2018), el tipo de investigación se determina según el propósito y alcance del estudio, pudiendo ser básica, aplicada o tecnológica. En este caso, la presente investigación es de tipo aplicada, ya que busca desarrollar y fortalecer el conocimiento teórico-metodológico relacionado con la imputación de datos faltantes en series temporales. Su objetivo es aplicar y adaptar técnicas estadísticas y de modelamiento a contextos reales, con el fin de obtener resultados que contribuyan a mejorar la precisión y robustez del análisis de este tipo de datos.

4.3 Nivel de investigación

De acuerdo con Hernández-Sampieri & Mendoza (2018), el nivel o alcance de la investigación se determina por el grado de profundidad con que se aborda el fenómeno estudiado, y puede clasificarse en exploratorio, descriptivo, correlacional o explicativo. En el presente estudio, el nivel corresponde a una investigación explicativa, dado que busca analizar y fundamentar los principios teóricos y estadísticos que sustentan el desarrollo de una técnica de imputación de datos faltantes en series temporales como: llegadas de visitantes a la Ciudad Inka de Machu

Picchu (2002 – 2025) y Movimiento Internacional de pasajeros en el aeropuerto Internacional Alejandro Velasco Astete (2004 – 2025).

4.4 Unidad de análisis

Según Hernández-Sampieri & Mendoza (2018), la unidad de análisis se refiere al elemento o conjunto de elementos sobre los cuales se recolectan los datos y se busca obtener conclusiones; puede ser un individuo, grupo, institución, registro o fenómeno. No obstante, en esta investigación no se define una unidad de análisis concreta, ya que el objetivo central es aplicar la metodología MAQY-TS a las series temporales con las características estudiadas.

4.5 Población de estudio

Según Hernández-Sampieri & Mendoza (2018), la población de estudio comprende el conjunto de elementos que poseen características comunes y sobre los cuales se pretende generalizar los resultados. No obstante, en esta investigación no se define una población específica, dado que el propósito es desarrollar la metodologia MAQY-TS.

4.6 Tamaño de muestra

De acuerdo con Hernández-Sampieri & Mendoza (2018), el tamaño de la muestra se refiere al número de unidades seleccionadas de una población para el análisis, determinado generalmente por criterios estadísticos de precisión y representatividad. Sin embargo, en esta investigación no se establece un tamaño de muestra en sentido poblacional, puesto que el objetivo no es estudiar un grupo delimitado de observaciones, sino probar y validar una técnica de imputación en series temporales con datos faltantes.

4.7 Técnicas de selección de muestra

De acuerdo con Hernández-Sampieri & Mendoza (2018), las técnicas de selección de muestra permiten determinar los elementos o casos que serán analizados dentro de un estudio, en

función de los objetivos y el tipo de investigación. En el presente trabajo, la técnica de selección de muestra se basa en la inclusión de todos los registros pertenecientes a series temporales que presentan datos faltantes, ya que el propósito es evaluar la eficacia del método de imputación en la totalidad de los casos donde ocurre la ausencia de información.

4.7.1 Técnicas de recolección de información

Según Hernández-Sampieri & Mendoza (2018), la recolección de información consiste en obtener los datos necesarios para responder al problema de investigación, empleando procedimientos sistemáticos, precisos y confiables. En el presente estudio, la recolección de información se realiza a partir de series temporales que contienen registros incompletos, obtenidas de fuentes estadísticas oficiales o generadas de forma simulada mediante herramientas computacionales.

4.7.2 Técnicas de análisis e interpretación de la información

La metodología que usamos para analizar y entender la información se basó en el estudio de series temporales. Siguiendo la metodología MAQY-TS (propuesta por el autor, MAQY que vienen a ser las iniciales de su nombre de su nombre y las siglas TS viene de las series temporales), empezamos identificando los valores que faltaban en toda la serie entre estos tenemos a los intervalos con datos faltantes y a los datos puntuales que faltan. Luego, usamos un modelo de pronóstico llamado Prophet para rellenar esos espacios a esto llamaremos imputación y asegurarnos de que los datos siguieran una estructura lógica. Después, aplicamos un criterio llamado (IFI) para comprobar si los valores que habíamos observado después de la imputación de los datos faltantes eran coherentes. Esto nos permitió detectar cualquier cambio repentino en la serie. Con la serie completa, hicimos un análisis estadístico que incluía descomponer la serie en tendencia, estacionalidad y un componente residual. También analizamos la autocorrelación y la autocorrelación parcial. Además, realizamos pruebas para ver si la serie era estable(estacionariedad), si seguía una distribución normal y si los datos eran independientes. Los resultados de estas pruebas nos ayudaron a elegir el modelo más adecuado, decidiendo entre modelos clásicos(ARMA, ARIMA, SARIMA) y modelos robustos(Holt Winters, STL(Loess),

ETS, BATS, TBATS) dependiendo de si cumplían con ciertos supuestos y si había anomalías. Finalmente, para entender mejor los resultados, evaluamos el desempeño de los modelos usando medidas como RMSE, MAE y MAPE. También usamos el estadístico de Theil y analizamos los residuos. Esto nos permitió identificar el modelo óptimo y asegurarnos de que los resultados que obtuvimos eran confiables.

CAPÍTULO V

RESULTADOS Y DISCUSIÓN

5.1 Propuesta metodología MAQY-TS

5.1.1 Propuesta de la metodología MAQY-TS para la imputación y detección de cambios abruptos series temporales con datos faltantes.

La investigación presente propone la metodología **MAQY-TS**, que representa una ampliación del enfoque tradicional, incorporando un mecanismo de imputación y validación mediante intervados de pronostico para la detección de cambios abruptos en series temporales con datos faltantes. La metodología esta compuesta por dos etapas: Etapa 1 imputación y detección de cambios abruptos, la cual responde al objetivo 1, Imputar datos faltantes de una serie temporal en un intervalo de tiempo y detectar cambios abruptos. La etapa 2 modelamiento, la cual responde al objetivo 2, Evaluar el modelo que tiene mejor en ajuste y pronóstico para una serie temporal faltante imputada con presencia de cambios abruptos. Continuación se presenta las fases:

ETAPA 1: Imputación y detección de cambios abruptos.

Fase 1: Identificación de la serie

Sea X_t una secuencia temporal. Se reconocen los valores ausentes (NA) y se establecen los intervalos continuos de falta de datos (gaps):

$$G_i = (t_{\text{inicio}}, t_{\text{fin}})$$

Fase 2: Imputación mediante Prophet

Para cada vacío G_i, se utiliza un modelo de pronóstico M (Prophet) (Taylor & Letham, 2018), calibrado con los datos observados:

$$\hat{\theta} \leftarrow M(X_t)$$

Luego, se imputan los valores faltantes:

$$\hat{X}_{gap} \leftarrow forecast(\hat{\theta}, h)$$

Nota: La imputación es necesaria para asegurar la continuidad estructural de la serie.

Fase 3: Detección de cambios abruptos (IFI)

Se analiza el primer valor registrado después del gap (X_{t+1}), estableciendo un intervalo de confianza (Toller et al., 2025):

$$I_t = \hat{X}_{t+1} \pm Z_\alpha \cdot \hat{\sigma}_{t+1}$$

Regla de decisión:

$$X_{t+1} \notin I_t \Rightarrow \text{Anomalía}$$

En este caso, la imputación se conserva, pero el punto es designado como cambio estructural.

Fase 4: Reconstrucción de la serie

Se obtiene una serie completa $\tilde{X}_t$ sin datos ausentes, incluyendo indicadores de anomalías identificadas.

ETAPA 2: Modelamiento de la serie .

Fase 5: Detección de cambios estructurales

Se utiliza la prueba de (Bai & Perron, 1998), para detectar varios puntos de cambio estructural en la serie.

Fase 6: Análisis preliminar y modelamiento

Descomposición

$$X_t = T_t + S_t + R_t$$

Identificación

Se examinan las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF).

Validación de supuestos

Se llevan a cabo las siguientes evaluaciones:

- Dickey-Fuller aumentado (estacionariedad)
- Jarque-Bera (normalidad)
- Ljung-Box (independencia de residuos)

Fase 7: Selección del modelo

La selección del modelo se basa en una regla de decisión conjunta:

- Si **no existen anomalías** y se **cumplen los supuestos** (ADF, JB, Ljung-Box):
  - Modelos clásicos: ARMA, ARIMA, SARIMA
- En caso contrario (presencia de anomalías o incumplimiento de supuestos):
  - Modelos robustos: Holt-Winters, STL (LOESS), ETS, BATS, TBATS

Fase 8: Validación

Se separa la serie en grupos de entrenamiento y validación. Se emplean métricas:

- RMSE, MAE, MAPE

Fase 9: Capacidad predictiva

Se utiliza el estadístico de Theil:

$$U = \frac{\text{RMSE}_{modelo}}{\text{RMSE}_{naive}}$$

Fase 10: Diagnóstico final

Se comprueba que los residuos del modelo elegido actúan como ruido blanco a través de la prueba de Ljung-Box:

$$H_0 : \text{Los residuos son independientes}$$

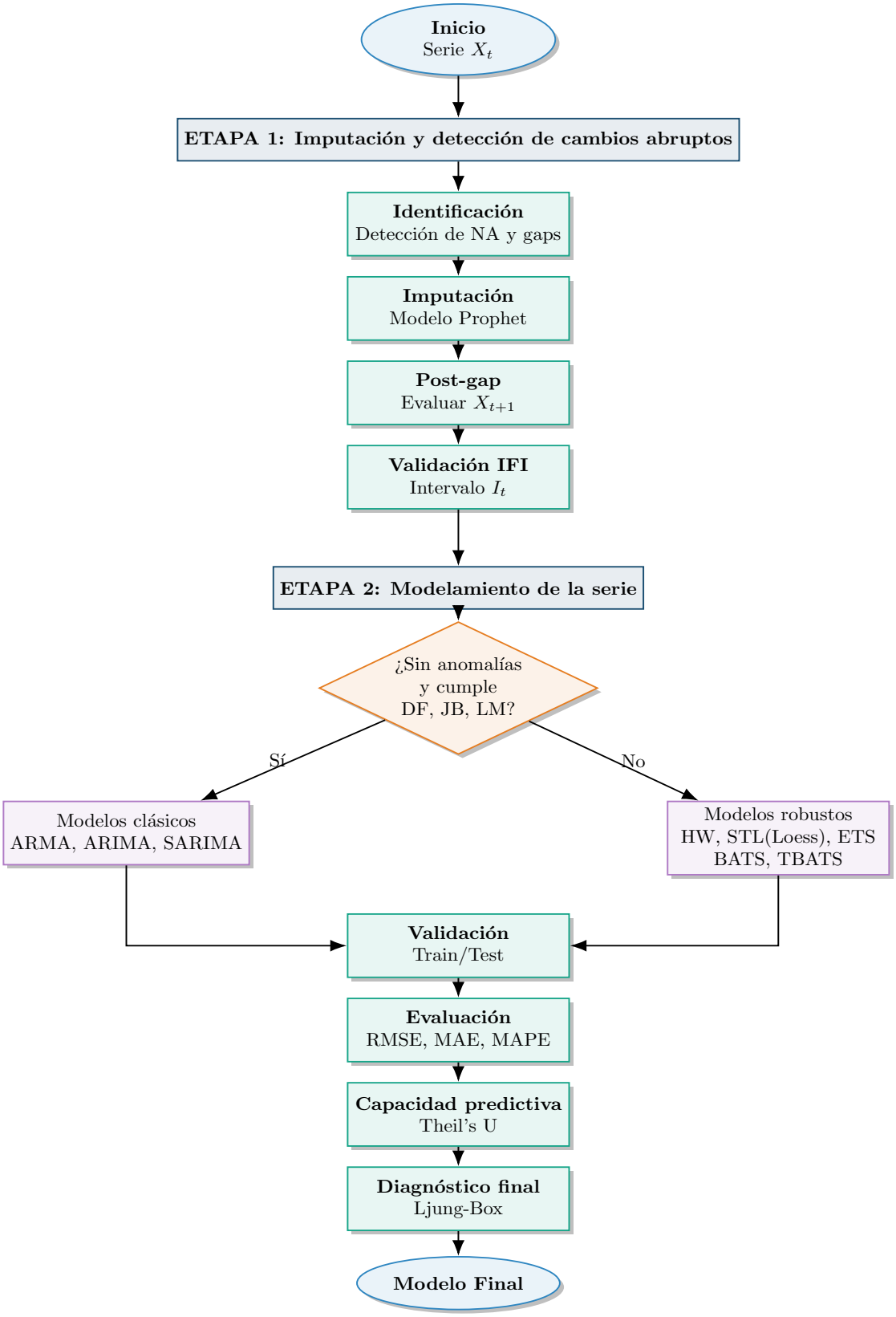
Asi mismo se formaliza en el siguiente algoritmo:

5.1.2 Algoritmo MAQY-TS para imputación y detección de cambios abruptos en series temporales.

Algorithm 1: Metodología MAQY-TS para imputación, detección de cambios abruptos y modelamiento de series temporales

Input: Serie temporal X_t con valores faltantes
Output: Modelo óptimo seleccionado y validado
ETAPA 1: Imputación y detección de cambios abruptos
Identificar índices de valores faltantes (gaps) G_i en X_t;
Ajustar modelo Prophet M sobre la serie observada;
Imputar valores faltantes obteniendo $\hat{X}_t$;
for *cada gap* G_i **do**
    Calcular intervalo de confianza I_t usando M;
    Evaluar el primer punto posterior al gap X_{t+1};
    if $X_{t+1} \in I_t$ **then**
        Marcar como comportamiento normal;
    else
        Marcar como anomalía;
ETAPA 2: Modelamiento de la serie
Aplicar prueba Dickey Fuller(DF);
Aplicar prueba Jarque-Bera(JB);
Aplicar prueba Lagrande Multiplier(LM);
if *(No existen anomalías)* ***y*** *(DF, JB, LM cumplen)* **then**
    Modelos clásicos:
    ARMA, ARIMA, SARIMA;
else
    Modelos robustos:
    Holt-Winters, STL(Loess), ETS, BATS, TBATS;
Dividir en train/test;
Calcular RMSE, MAE, MAPE;
Evaluar Theil's U;
Verificar Ljung-Box;
return Modelo final óptimo;

5.1.3 *Flujograma MAQY-TS para imputación y detección de cambios abruptos en series temporales.*



5.2 Aplicacion metodología MAQY-TS a series reales.

5.2.1 Número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Tabla 2
Resumen estadístico de la serie de número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

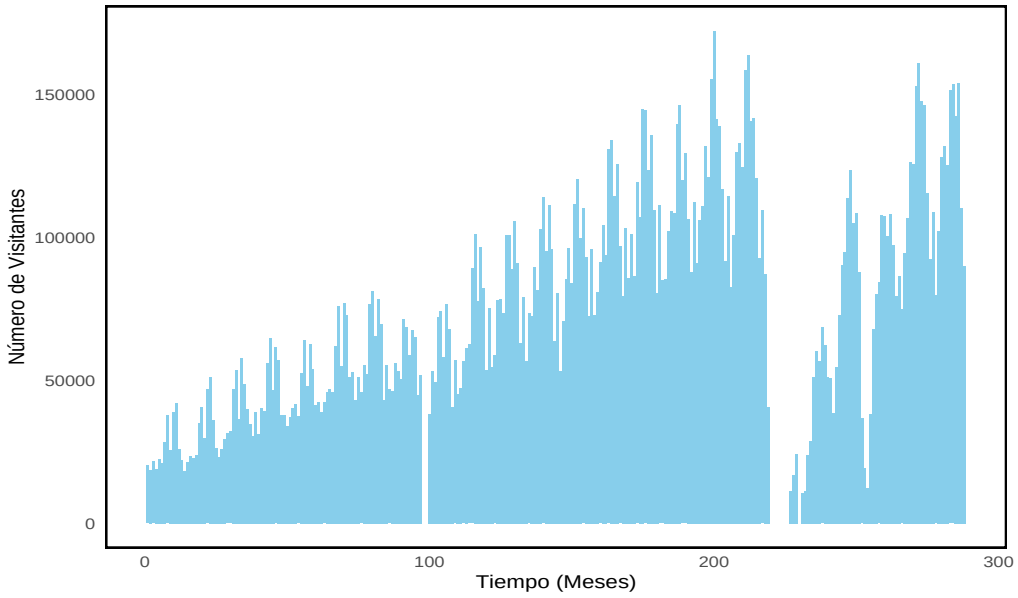
Medida Estadística	**Valor**
Mínimo	11
Primer Cuartil (Q1)	47
Mediana	74
Tasa media geométrica	0.5215709
Tercer Cuartil (Q3)	105
Máximo	173
Valores Faltantes (NA)	10

Nota. El valor NA indica el número el número de datos faltantes.

En la Tabla 2 se observa el número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025). presenta un rango de operación que oscila entre un mínimo de 11 y un máximo de 173 personas. La mediana se sitúa en 74 visitantes, mientras que la distribución de los datos muestra que el 50 % central de las observaciones se encuentra entre los 47 (Q1) y los 105 (Q3) visitantes. El cálculo de la tasa media geométrica arroja un valor de 0.5215709, reflejando el promedio de las variaciones relativas de la serie.

Figura 4

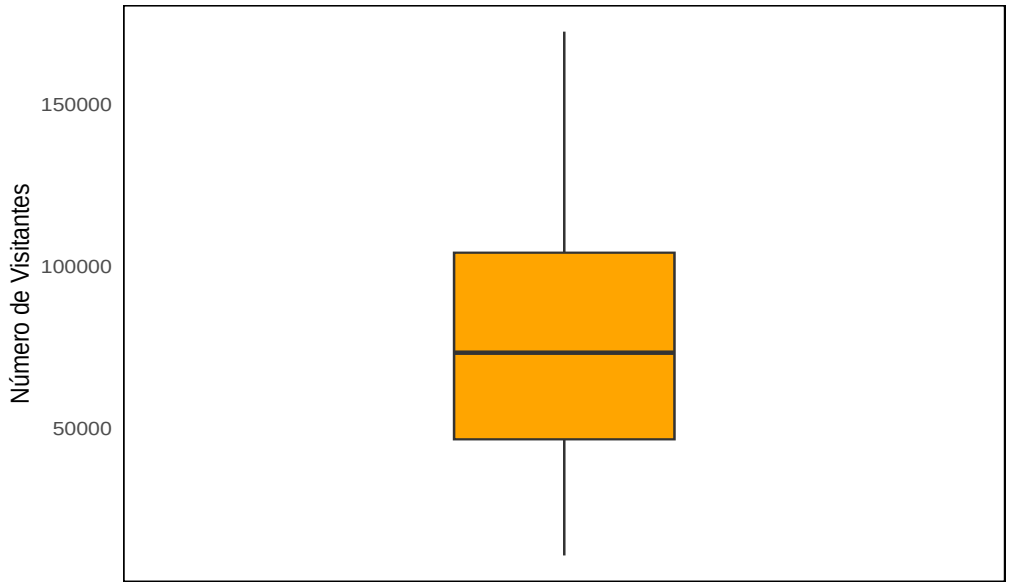
Diagrama de barras del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 4 se muestra la progresión del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025), que muestra un patrón claramente en aumento. Al comienzo, se observan cifras de entre 500 y 2,000 pasajeros. Luego, existe un aumento gradual que lleva a una cifra aproximada de 10,000 hacia el mes número 170. La serie, aunque muestra notables oscilaciones con caídas y recuperaciones abruptas, concluye con un máximo de cerca de 18,000 pasajeros.

Figura 5

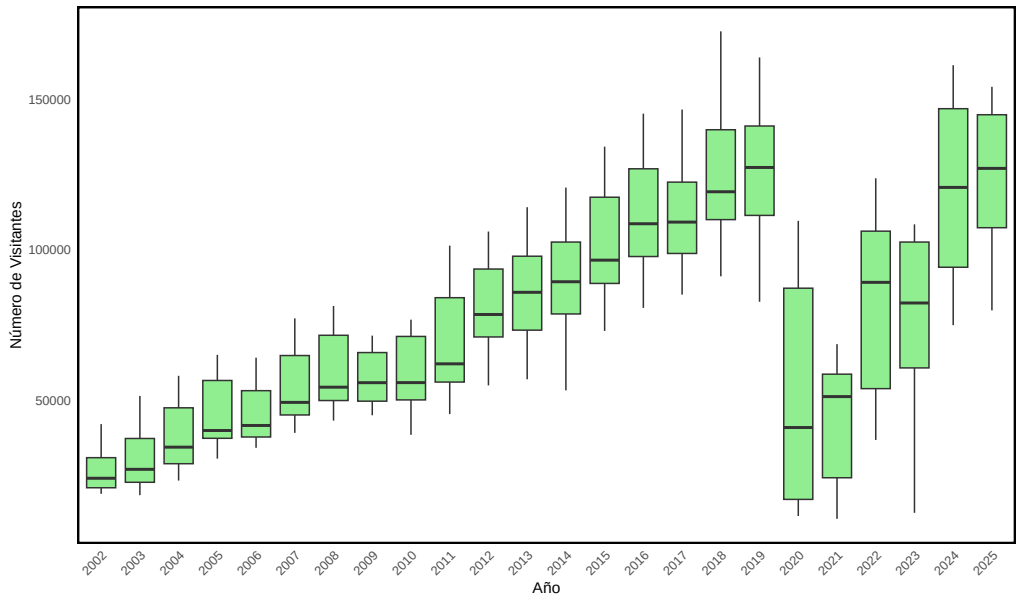
Boxplot del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 5 se muestra el diagrama de caja para el número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025). muestra una distribución con una mediana ligeramente desplazada hacia la base de la caja, lo que confirma la asimetría positiva observada en la serie histórica. El rango intercuartílico se concentra aproximadamente entre los 46,000 y los 104,000 visitantes, mientras que los bigotes se extienden para cubrir la totalidad de los datos sin evidenciar la presencia de valores atípicos (outliers).

Figura 6

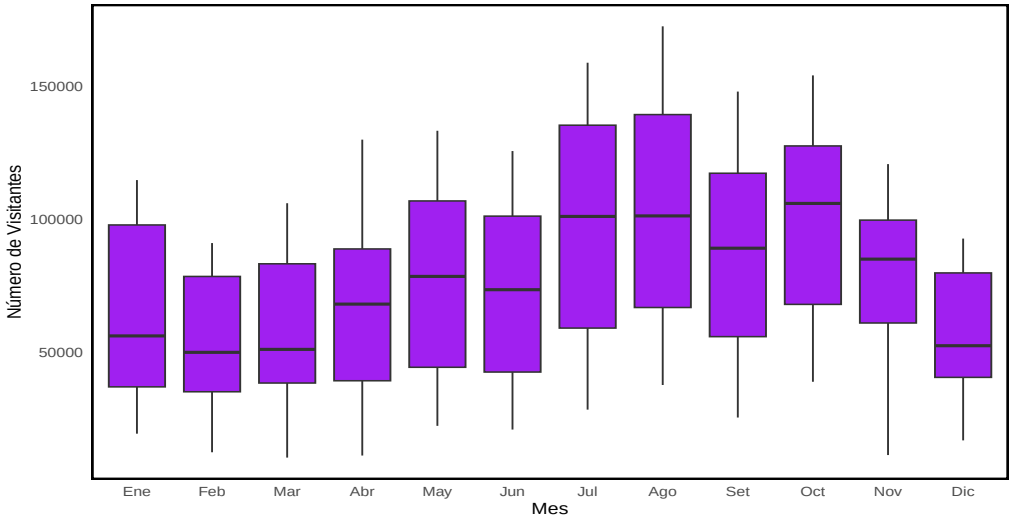
Boxplot por año del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 6 se muestra el diagrama de cajas anual del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).muestra una tendencia creciente sostenida entre 2002 y 2019, interrumpida por una caída drástica y alta variabilidad en 2020. A partir de 2024, la serie evidencia una recuperación sólida, alcanzando niveles de afluencia y medianas similares a los máximos históricos previos a la perturbación, no se observa la presencia de valores atípicos.

Figura 7

Boxplot por mes del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 7 se muestra el diagrama de cajas mensual del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025). Evidencia un marcado patrón estacional en el flujo de visitantes, con una temporada alta definida entre los meses de julio y octubre, donde se registran las medianas y niveles de dispersión más elevados del año. En contraste, el periodo comprendido entre diciembre y marzo presenta los niveles de afluencia más bajos, consolidando una estructura cíclica anual donde los picos de demanda máxima ocurren durante el tercer trimestre, no se observa la presencia de valores atípicos.

Tabla 3

Resumen de valores faltantes del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

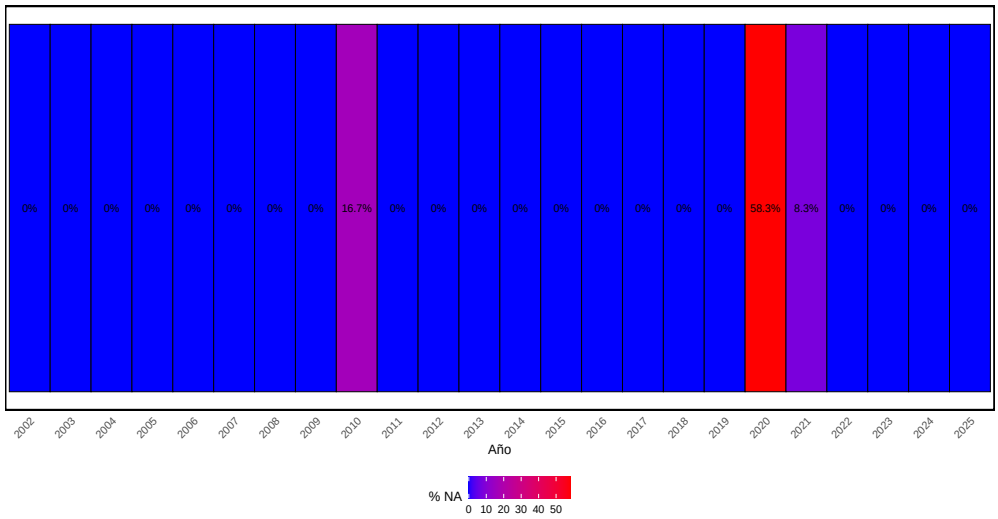
Indicador	Valor
Total de datos faltantes (NA)	19
Porcentaje de NA (%)	7.22 %

Nota. El porcentaje de valores faltantes es 10 %.

En la Tabla 3 se observa el número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025). presenta un total de 19 observaciones faltantes, lo que representa el 7.22 % del conjunto total de datos. Lo que nos permite la aplicación de técnicas de imputación para restablecer la continuidad de la serie.

Figura 8

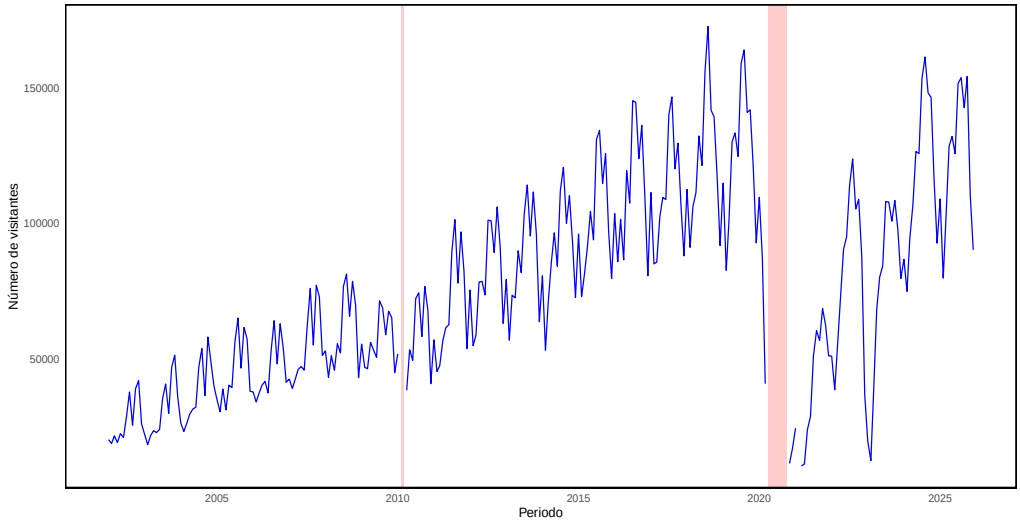
Mapa de calor de datos faltantes del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 8 se observa el mapa de calor de datos faltantes por año revela que la serie del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025). Mantuvo una integridad del 100 % en la mayoría de los periodos, con excepciones críticas localizadas. Se identifica una pérdida moderada de información en el año 2010 (16.7 %) y una interrupción severa en el año 2020, donde el porcentaje de datos faltantes alcanzó un máximo de 58.3 %. Lo que justifica el empleo de algoritmos de imputación específicos para tratar las brechas temporales significativas detectadas durante dichos años.

Figura 9

Serie con intervalo de datos faltantes del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 9 se observa La serie de tiempo mensual del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025). Interrumpida por periodos de ausencia de información resaltados mediante franjas rojas. Estos intervalos de datos faltantes, localizados en el año 2010 y de forma más extensa durante el 2020, representan vacíos de meses consecutivos que rompen la continuidad de la serie. Debido a la existencia de estos intervalos, se propone la aplicación de una técnica de imputación específica (IFI) diseñada para reconstruir la estructura temporal y recuperar la dinámica de la serie sin generar sesgos en la serie.

Tabla 4

Identificación de Gaps (Vacíos) detectados en la serie número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

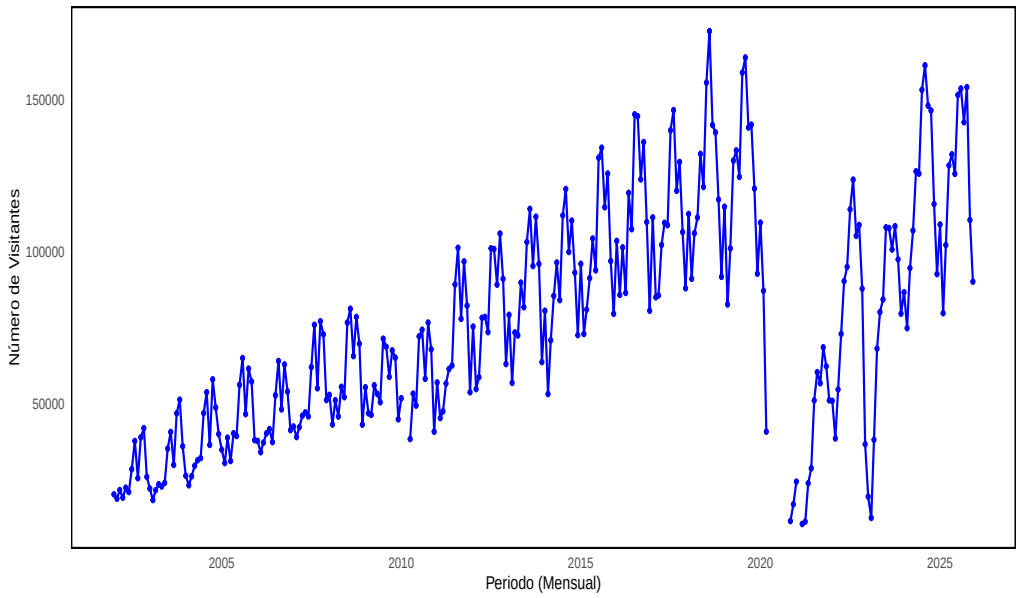
Index interval (L_i, L_s)	**Longitud**
(98, 99)	2
(220, 226)	7
(230, 230)	1

Nota. El intervalo indica la presencia de datos faltantes con (NA).

En la Tabla 4 se observa la ubicación y magnitud de los intervalos de datos faltantes identificados en la serie. Se registran tres brechas temporales específicas: la primera entre los meses 98 y 99 con una longitud de 2 periodos; la segunda, de mayor extensión, entre los meses 220 y 226 con una longitud de 7 periodos; y una observación aislada en el mes 230. La presencia de estos bloques de omisión, especialmente el intervalo de 7 meses consecutivos, justifica la implementación de una técnica de imputación robusta que considere la estructura de la serie.

Figura 10

Serie temporal con datos faltantes del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 10 se muestra la serie temporal con datos faltantes del Número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).Presenta aparentemente una tendencia ascendente y estacionalidad marcada, interrumpida por una caída drástica y vacíos de información críticos en 2020. Esta discontinuidad impide el análisis directo de sus componentes, haciendo indispensable una técnica de imputación que reconstruya los valores omitidos. Para lo cual se propone la técnica (IFI) dicha imputación es necesaria para estabilizar la tendencia y permitir una recuperación de la dinámica de crecimiento, que hacia 2025 ya alcanza niveles similares a sus máximos históricos.

Tabla 5

Identificación de intervalos con datos faltantes (Gaps) en la serie del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

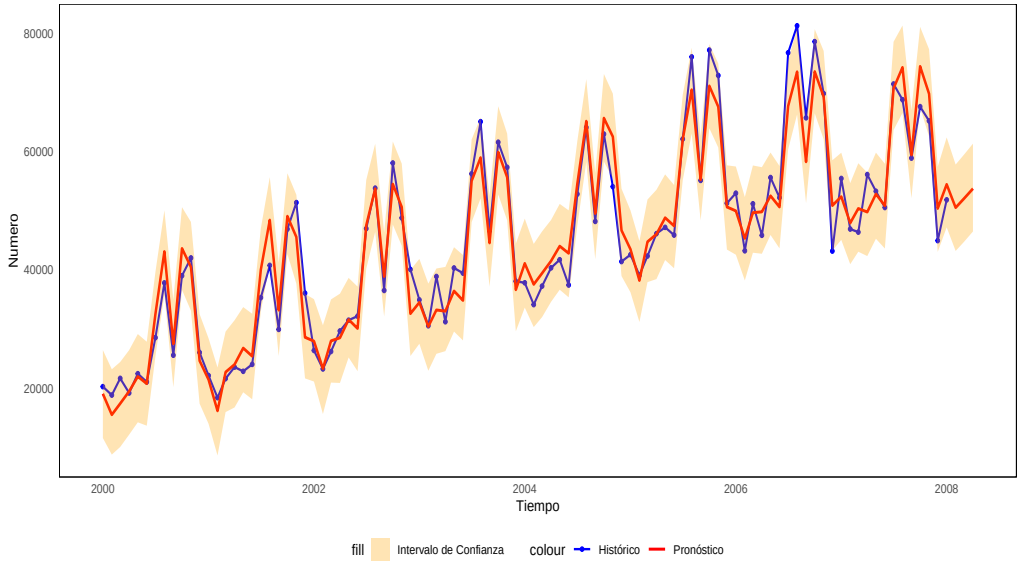
ID	Intervalo (Índices)	Longitud
1	(98, 99)	2
2	(220, 226)	7
3	(230, 230)	1

Nota. Los ID son los gap.

En la Tabla 5 se observa la Identificación de intervalos con datos faltantes (Gaps) en la serie del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025). identifica tres intervalos de datos faltantes en la serie, con longitudes de 2, 7 y 1 periodos respectivamente. Destaca el segundo intervalo por su extensión significativa (7 meses consecutivos), lo que representa la brecha de información más crítica. La precisión en los índices y fechas de inicio y fin de estos vacíos justifica la necesidad de aplicar un método de imputación (IFI) que considere la estructura secuencial de los datos para garantizar la continuidad del análisis temporal.

Figura 11

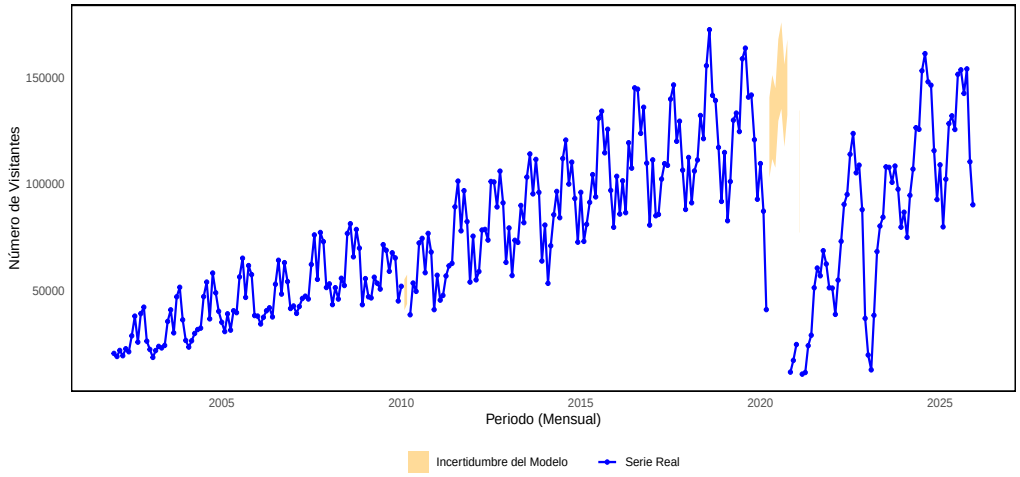
Delimitación de intervalos de confianza mediante el método IFI y el algoritmo Prophet en la serie número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 11 se muestra la fase de entrenamiento del modelo Prophet utilizando exclusivamente los datos históricos disponibles. En este proceso, el algoritmo parametriza los componentes de tendencia y estacionalidad para generar una base predictiva (línea roja) que se ajusta a la dinámica real observada. Este ajuste es un paso procedimental crítico para el método (IFI), ya que el modelo entrenado proporciona los valores estimados y los intervalos de confianza necesarios para realizar la imputación efectiva en los periodos donde se identificaron los vacíos de información.

Figura 12

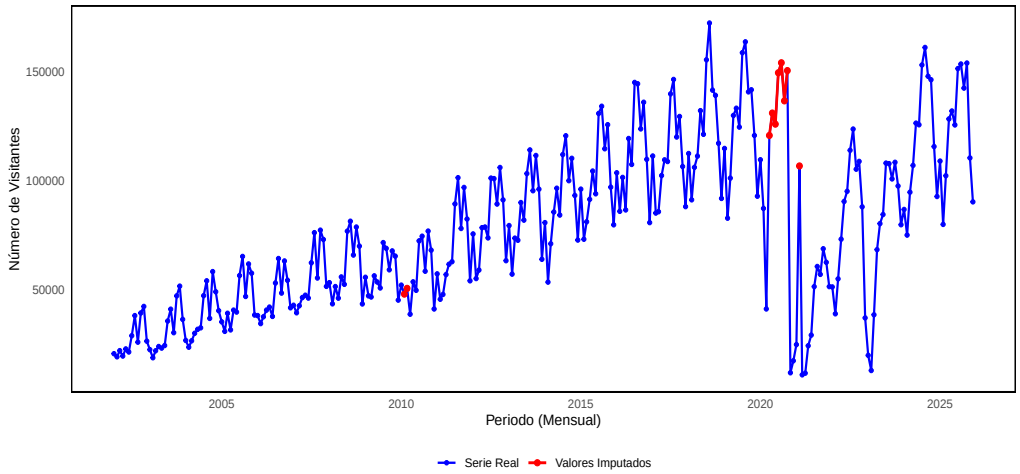
Serie temporal con datos faltantes del Número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025)



La Figura 12 muestra la fase de delimitación del método (IFI), donde las franjas naranjas representan los intervalos de confianza probabilísticos generados por Prophet para cubrir los vacíos de la serie real. Según el marco teórico propuesto, estas bandas de incertidumbre capturan la tendencia y estacionalidad histórica, estableciendo el soporte estadístico necesario para la reconstrucción de la serie. Este enfoque permite que, incluso en periodos de alta volatilidad como el año 2020, se preserve la coherencia dinámica y la memoria estructural de los datos antes de proceder con la imputación final.

Figura 13

Reconstrucción de la serie de tiempo mediante la imputación de valores estimados con el método Prophet + (IFI) del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



La Figura 13 presenta la reconstrucción final de la serie mediante el método (IFI), donde los valores imputados (puntos rojos) completan los vacíos de información detectados. De acuerdo con la base teórica del estudio, Prophet + IFI logra una integración coherente al proyectar la estacionalidad y tendencia histórica dentro de los intervalos de confianza previamente establecidos. Este ajuste minimiza el error cuadrático y preserva la memoria de la serie, permitiendo recuperar la continuidad del flujo del número de visitantes especialmente tras la ruptura del año 2010, 2020 de forma estadísticamente consistente con el comportamiento observado.

Tabla 6

Valores imputados e intervalos de confianza mediante el método Prophet + (IFI) por periodos de discontinuidad para la serie número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

ID	Fecha	Valor Imputado	Límite Inferior	Límite Superior	Gap ID
1	2010-02-01	47,708.83	40,677.07	54,730.70	1
2	2010-03-01	50,344.56	43,621.50	57,042.82	1
3	2020-04-01	120,768.29	101,917.55	139,815.36	2
4	2020-05-01	131,081.37	111,753.70	148,864.36	2
5	2020-06-01	125,982.50	108,398.61	145,421.95	2
6	2020-07-01	149,626.35	130,218.44	167,164.41	2
7	2020-08-01	154,237.72	135,716.33	174,066.88	2
8	2020-09-01	136,701.02	118,225.82	156,215.46	2
9	2020-10-01	150,629.20	131,982.91	170,269.53	2
10	2021-02-01	106,719.09	76,661.90	136,115.04	3

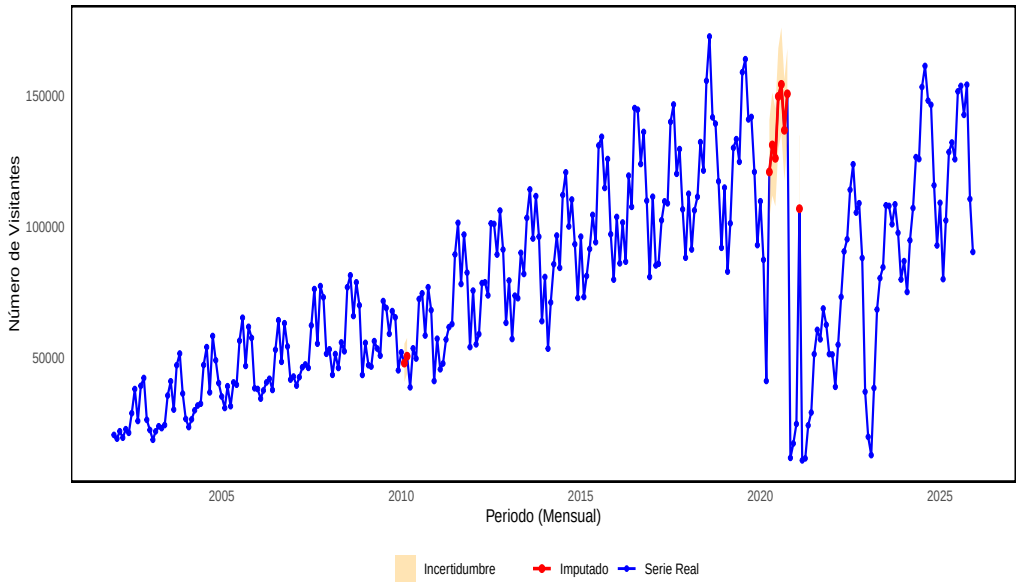
Nota. Los valores corresponden a la imputación final basada en el modelo de pronostico Prophet.

La Tabla 6 presenta los resultados numéricos de la reconstrucción de la serie, destacando tres periodos de discontinuidad con dinámicas distintas. Resulta relevante observar que en el Gap ID 2, los valores imputados alcanzan un máximo de 154,237.72 visitantes en agosto de 2020, lo que representa un volumen significativamente superior a los niveles registrados en el Gap ID 1 (2010), donde las estimaciones promedian los 49,000 visitantes. Asimismo, la amplitud de los

intervalos en el Gap ID 3 (febrero de 2021) muestra una dispersión de aproximadamente 59,453 unidades entre el límite inferior y superior, reflejando una mayor variabilidad en la estimación comparada con los periodos anteriores. Estos valores establecen la escala de magnitud necesaria para dar continuidad a la serie, respetando los niveles de crecimiento histórico proyectados por el modelo.

Figura 14

Serie de tiempo con imputación completa y intervalo de confianza, mediante el método Prophet + IFI número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 14 se observa la serie de tiempo integrada, donde la línea roja representa los valores imputados en los periodos de discontinuidad detectados, principalmente durante el año 2020. Bajo la metodología Prophet + IFI, se observa cómo la reconstrucción técnica permite dar continuidad a la trayectoria, vinculando los registros previos a la caída con la fase de recuperación observada a partir de 2021. Este proceso asegura que la tendencia ascendente y los ciclos estacionales se mantengan coherentes a lo largo de todo el horizonte temporal hasta 2025, eliminando los vacíos de información que impedían el análisis estructural y garantizando la estabilidad estadística de la serie completa.

Tabla 7

Identificación de anomalías detectadas en la serie número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

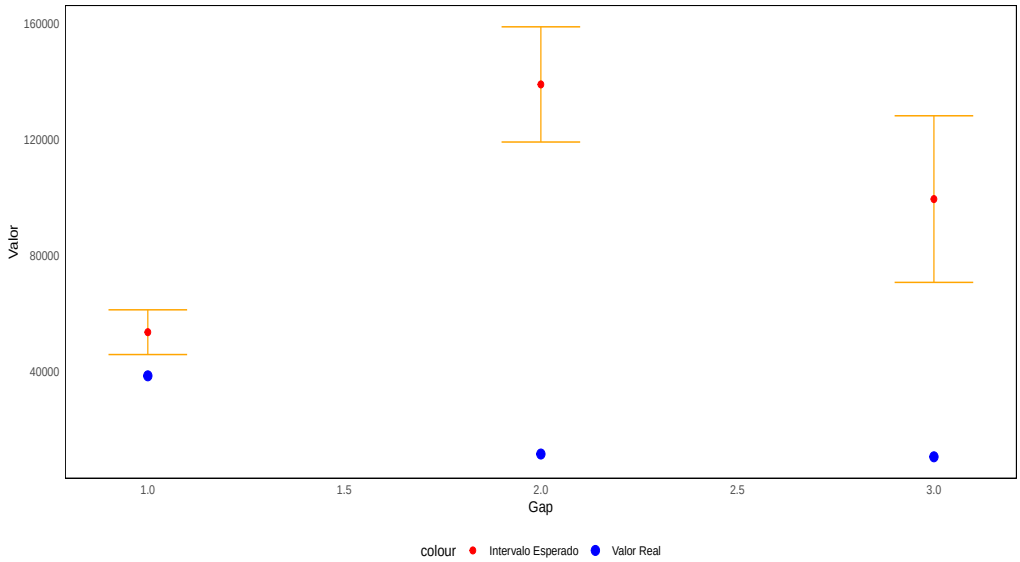
ID	Fecha (ds)	Valor Observado (y)
1	2010-04-01	38454
2	2020-11-01	11485
3	2021-03-01	10532

Nota. Valores que se desvían del comportamiento esperado.

La Tabla 7 se observa La metodología implementada consiste en el análisis del primer valor real observado inmediatamente después del intervalo de imputación. Este procedimiento utiliza dicho registro como punto de control para confrontarlo con el soporte probabilístico generado por el método (IFI); de este modo, se determina formalmente si existe una anomalía detectada o si la serie mantiene su consistencia estructural. Al centrar la evaluación en el comportamiento posterior al vacío, el modelo puede discernir si el retorno de la información sigue la memoria estadística proyectada o si, por el contrario, se ha producido un cambio abrupto en la dinámica de los datos.

Figura 15

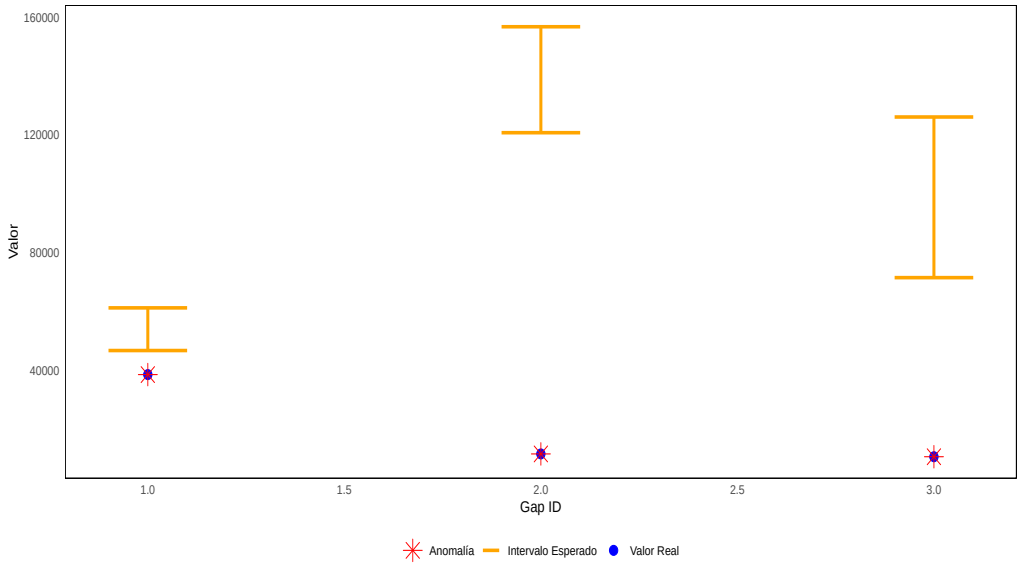
Detección de anomalías mediante intervalos de confianza IFI por bloque de discontinuidad del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 15 se observa el diagnóstico de anomalías mediante el contraste de los valores reales frente a los intervalos de confianza del método IFI. En los tres casos analizados, el registro observado se sitúa fuera del soporte estadístico esperado, lo que permite al modelo detectar formalmente rupturas estructurales. Esta divergencia confirma que la serie divergió de su memoria histórica tras los periodos de discontinuidad, validando la eficacia del método para identificar cambios abruptos de régimen.

Figura 16

Identificación de cambios abruptos basados en el criterio de exclusión del soporte estadístico número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 16 muestra el diagnóstico final de anomalías tras el proceso de imputación IFI. Se observa que, al contrastar el valor real observado con el intervalo esperado, los tres puntos de control (Gap 1, 2 y 3) se sitúan fuera de los límites de confianza proyectados, siendo identificados automáticamente como anomalías (marcadas con una estrella roja). Esta divergencia cuantitativa demuestra que la serie experimentó cambios abruptos en cada uno de estos periodos, validando la capacidad del método para detectar rupturas estructurales donde la realidad de la serie de visitantes divergió significativamente de su memoria estadística histórica.

Tabla 8

Resultados de (IFI) para los Gaps detectados en la serie número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

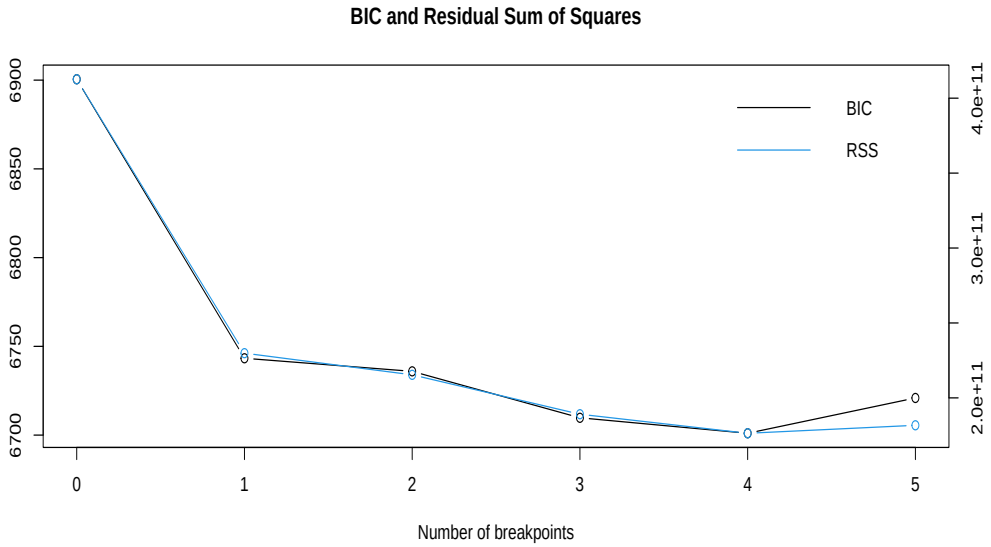
Gap ID	Inicio	Fin	Valor Real	Límite Inf.	Límite Sup.	IFI (Validación)
1	98	99	38454	46150.95	61064.98	TRUE
2	220	226	11485	119694.24	158205.87	TRUE
3	230	230	10532	71712.45	128413.11	TRUE

Nota. El resultado TRUE indica que la metodología IFI ha procesado correctamente el intervalo de datos faltantes basándose en los límites calculados.

En la Taba 8 se observa la validación final del método IFI para cada bloque de discontinuidad. Al contrastar el "Valor Realçon los umbrales calculados, se observa que en todos los casos el registro observado se sitúa significativamente por debajo del "Límite Inf.", confirmando numéricamente la ruptura de la memoria estadística. Especialmente en el Gap ID 2, la diferencia entre el valor real (11,485) y el límite inferior esperado (119,694.24) evidencia una anomalía de gran magnitud. El resultado "TRUE.[en] la columna de validación ratifica que el algoritmo procesó correctamente estos intervalos, identificando con precisión técnica los puntos donde la serie divergió de su comportamiento.

Figura 17

Optimización de rupturas mediante la prueba de Bai-Perron y criterios BIC del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



La Figura 17 se observa la evolución del Criterio de Información Bayesiano (BIC) en función del número de rupturas. El BIC alcanza su punto mínimo global en $m = 4$, señalando la estructura óptima para la serie. El repunte del BIC en $m = 5$ penaliza el sobreajuste, validando que cuatro quiebres estructurales son suficientes y necesarios para describir los cambios de régimen históricos. Esta convergencia estadística fundamenta la alta volatilidad estructural que justifica el uso posterior del método IFI.

Tabla 9

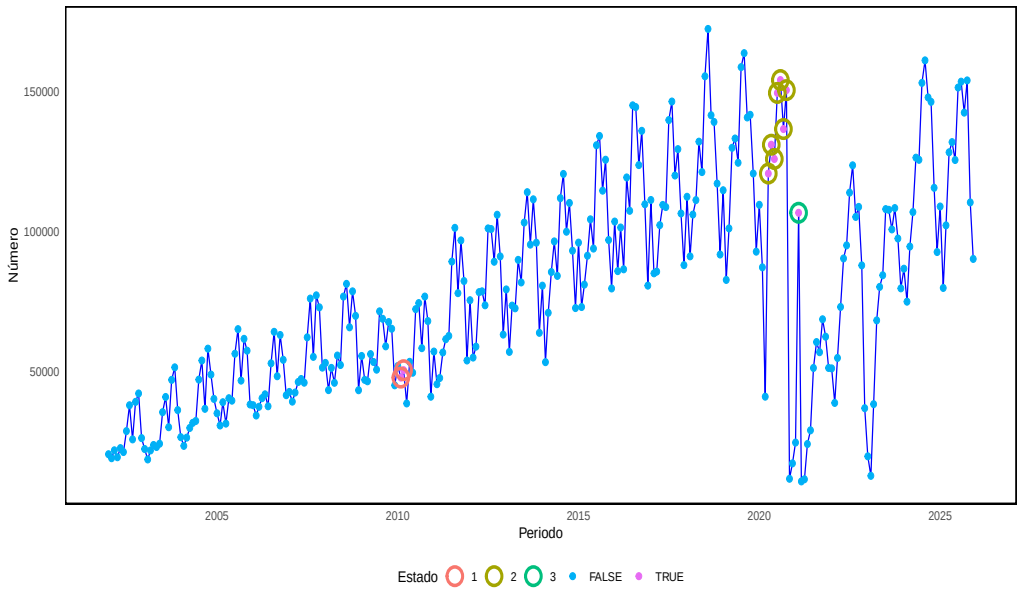
Bai-Perron para la determinación de los puntos de ruptura estructural del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Número de rupturas (m)	**BIC**	**Selección**
4	6701	Óptimo

En la Tabla 9 se observa los resultados de la aplicación de la prueba de Bai-Perron identifica que el modelo óptimo para explicar la variabilidad de la serie requiere $m = 4$ rupturas estructurales, minimizando el Criterio de Información Bayesiano (BIC) en 6701. Este resultado estadístico confirma que la serie no es un proceso homogéneo, sino que ha sufrido cuatro cambios permanentes en su estructura media a lo largo de todo el horizonte histórico analizado.

Figura 18

Optimización de rupturas mediante la prueba de Bai-Perron y criterios BIC del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



La Figura 18 ilustra la serie de tiempo completa con la identificación de los tres periodos de discontinuidad (Gaps). Los puntos resaltados con círculos de colores (Estados 1, 2, 3, 4) representan los segmentos donde se aplicó la imputación y posterior validación diagnóstica. Se observa que el Gap 1 corresponde a una fluctuación temprana, mientras que los Gaps 2 y 3 se localizan en la zona de mayor inestabilidad de la serie (post 2020), coincidiendo con las rupturas estructurales detectadas previamente. Esta visualización confirma la capacidad del modelo para integrar datos imputados manteniendo la coherencia con la dinámica global.

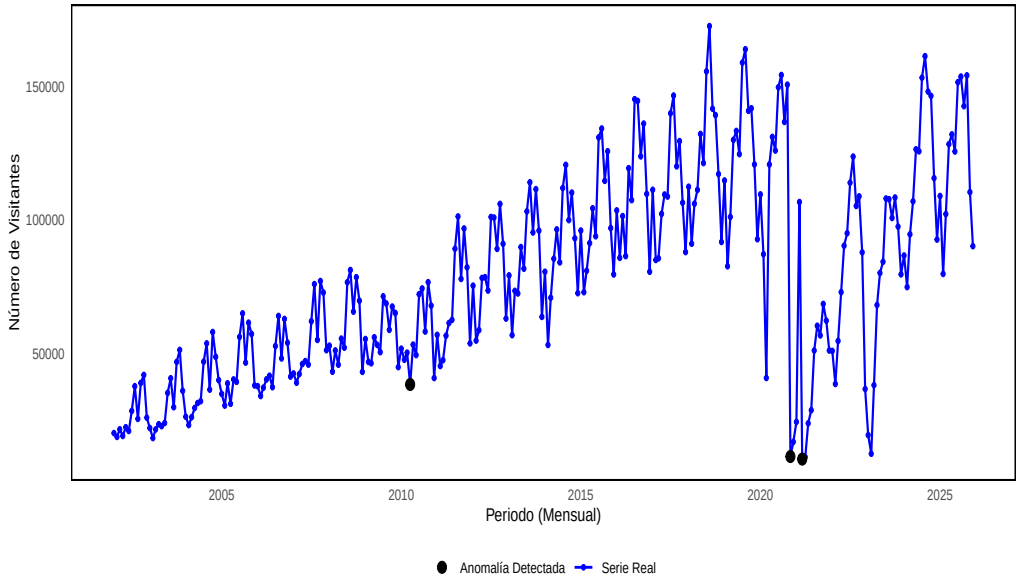
Tabla 10
Observaciones identificadas como anomalías en la serie número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Fecha (ds)	Valor observado
2010-04-01	38454
2020-11-01	11485
2021-03-01	10532

En la tabla 10 muestra las observaciones que se destacan como anomalías. Estas observaciones tienen desviaciones muy grandes respecto a lo que se espera de la serie. Los valores de abril de 2010, noviembre de 2020 y marzo de 2021 muestran cambios grandes en la tendencia usual.

Figura 19

Identificación de anomalías estructurales mediante el método IFI en el número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 19 se observa la identificación de las anomalías estructurales detectadas mediante el método (IFI). Los puntos negros señalan los momentos exactos donde el valor real observado, tras un periodo de datos faltantes, divergió significativamente del soporte probabilístico calculado. Se destaca la detección de anomalías en el periodo 2010, 2020 y 2021, lo cual coincide con la ruptura estructural global de la serie y valida la capacidad del método para señalar puntos de quiebre.

Tabla 11

Detección de anomalías por bloque de discontinuidad mediante el método IFI del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Bloque (ID)	Inicio (Índice)	Fin (Índice)	Detección de Anomalía
1	100	105	Detectada
2	227	232	Detectada
3	231	236	Detectada

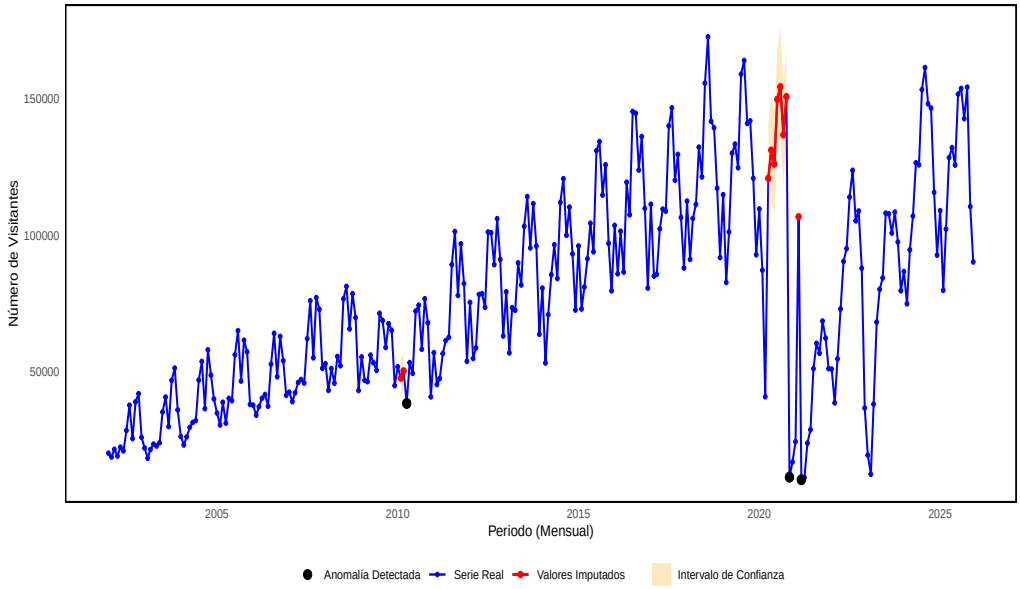
Nota. La columna Detección de Anomalía indica si el comportamiento tras el gap es irregular.

La Tabla 11 se observa los resultados del diagnóstico tras la imputación por intervalos. Para los tres bloques analizados, el método IFI clasificó el comportamiento post discontinuidad

como Anomalía Detectada. Este resultado sistemático confirma que, en cada punto de control, el valor real observado divergió de los límites de confianza esperados, evidenciando una ruptura en la continuidad estadística de la serie tras los intervalos de datos faltantes.

Figura 20

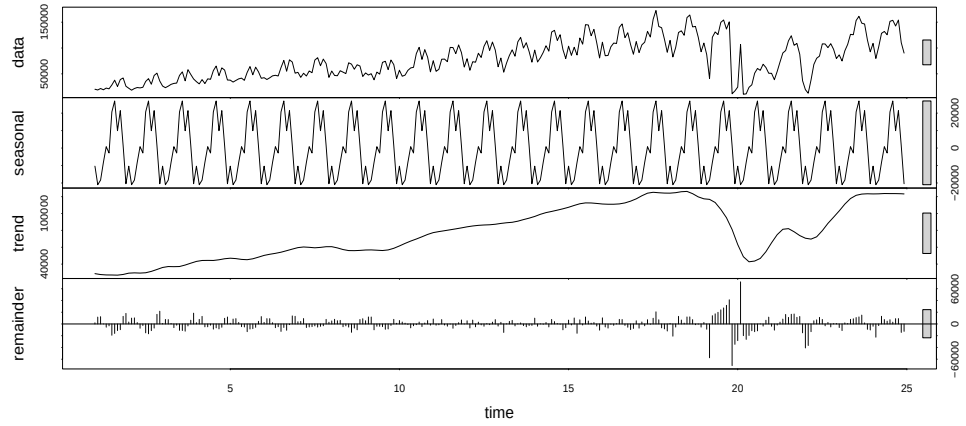
Reconstrucción de la serie temporal y diagnóstico de anomalías estructurales número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 20 se observa la reconstrucción integral de la serie de visitantes mediante el método IFI, destacando la interacción entre los valores imputados (segmentos rojos) y la validación de anomalías (puntos negros). Se observa que, tras los periodos de falta de información, el primer valor real observado se sitúa fuera de las bandas de confianza proyectadas, lo que permite al modelo diagnosticar formalmente una anomalía estructural. Este hallazgo es particularmente crítico en el periodo post 2020, donde la serie real no solo divergió del soporte estadístico, sino que inició un nuevo régimen de comportamiento, validando la capacidad del método IFI para identificar quiebres en la memoria de la serie de tiempo.

Figura 21

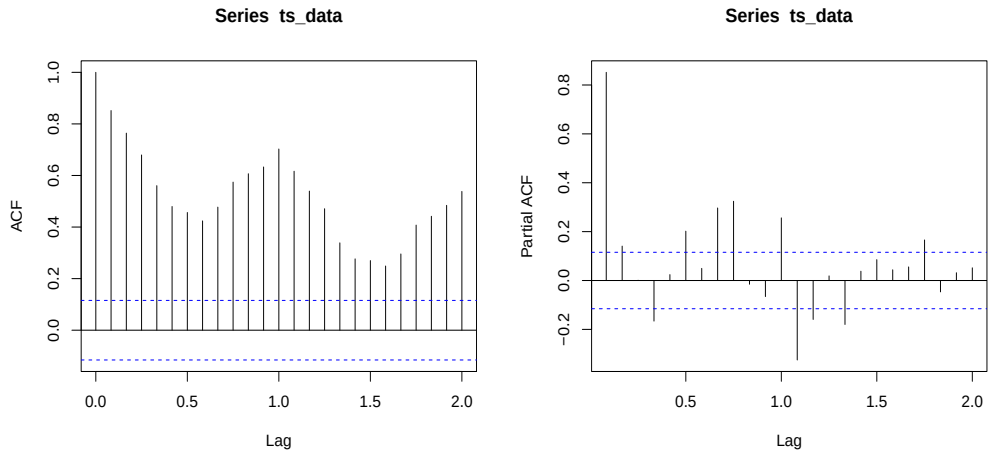
Descomposición estructural STL de la serie del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 21 se muestra los componentes de la serie imputada mediante el algoritmo STL. El componente de tendencia muestra un crecimiento sostenido desde los 40,000 hasta superar los 100,000 visitantes aproximadamente, antes de presentar una contracción marcada y su posterior recuperación. La estacionalidad revela un patrón mensual constante con oscilaciones que fluctúan aproximadamente entre mas o menos 20,000 unidades aproximadamente. Finalmente, el residuo muestra una variabilidad controlada en la mayor parte de la serie, con mayores fluctuaciones en el tramo final que alcanzan magnitudes cercanas a las 60,000 unidades aproximadamente, reflejando la dispersión no capturada por los componentes previos.

Figura 22

Funciones de autocorrelación (*ACF*) *y autocorrelación parcial* (*PACF*) *del número de visitantes a la Ciudad Inka de Machu Picchu* (2002–2025).



En la Figura 22 se muestra las funciones de autocorrelación simple (ACF) y parcial (PACF) para la serie imputada. El ACF muestra un decrecimiento lento y un patrón sinusoidal con picos significativos en los retardos (lags) múltiples de 1.0 (12 meses), lo que confirma una fuerte dependencia estacional y la no estacionariedad de la serie. Por su parte, el PACF exhibe picos prominentes en los primeros retardos (superando 0.8) y en torno al retardo estacional, sugiriendo la necesidad de aplicar diferenciación y términos autorregresivos (AR). La persistencia de valores fuera de las bandas de confianza (líneas punteadas azules) valida que la serie conserva memoria de largo plazo tras la imputación.

Tabla 12
Prueba de Jarque-Bera del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Jarque-Bera	
p-valor	0.001676

- H_0: Los datos siguen una distribución normal.
- H_1: Los datos no siguen una distribución normal.

En la tabla 12 se observa la prueba de Jarque-Bera arroja un p-valor de 0,001676, el cual es inferior al nivel de significancia estandar ($\alpha = 0{,}05$). En consecuencia, se rechaza la hipótesis nula, concluyendo que los datos no se distribuyen de forma normal.

Tabla 13
Prueba de raíz unitaria Augmented Dickey-Fuller (ADF) del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Test Dickey - Fuller	
P-Value	0.01

- H_0: La serie no es estacionaria.
- H_1: La serie es estacionaria.

En la tabla 13 se observa prueba de Dickey-Fuller arroja un p-valor de 0,01, el cual es

inferior al nivel de significancia estandar ($\alpha = 0{,}05$). En consecuencia, se rechaza la hipótesis nula, concluyendo que la serie de visitantes es estacionaria.

Tabla 14
Resultado de la prueba de tendencia Mann-Kendall del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Mann-Kendall Test	
P-Value	$2{,}22 \times 10^{-16}$

- H_0: No existe una tendencia monótona en la serie.
- H_1: Existe una tendencia monótona en la serie.

En la tabla 14 se observa prueba de Mann-Kendall arroja un p-valor de $2{,}22 \times 10^{-16}$, lo cual es significativamente inferior al nivel de significancia estandar ($\alpha = 0{,}05$). En consecuencia, se rechaza la hipótesis nula, confirmando la presencia de una tendencia monótona significativa en la serie de visitantes.

Tabla 15
Resultado de la prueba de Kruskal-Wallis para estacionalidad del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Kruskal-Wallis test	
P-Value	$8{,}897 \times 10^{-07}$

- H_0: No hay estacionalidad en la serie.
- H_1: Hay estacionalidad en la serie.

En la tabla 15 se observa prueba de Kruskal-Wallis reporta un p-valor de $8{,}897 \times 10^{-07}$, valor que resulta inferior al nivel de significancia estandar ($\alpha = 0{,}05$). Por lo tanto, se rechaza la hipótesis nula, confirmando la existencia de diferencias estadísticamente significativas en el número de visitantes según el mes del año.

Tabla 16
Resultado de la prueba Breusch-Godfrey del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

LM test	
P-Value	$2{,}2 \times 10^{-16}$

- H_0: No existe correlación serial en los residuos (independencia).
- H_1: Existe correlación serial en los residuos.

En la tabla 16 se observa prueba de Breusch-Godfrey presenta un p-valor de $2{,}2 \times 10^{-16}$, lo cual permite rechazar la hipótesis nula con un nivel de confianza superior al 95 %. Este resultado indica la presencia de autocorrelación serial en los residuos de la serie, sugiriendo que el modelo actual aún no ha capturado toda la estructura de dependencia temporal.

Tabla 17
División de la serie temporal en conjuntos de entrenamiento y prueba del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Conjunto	**Cantidad (n)**	**Porcentaje (%)**
Data de entrenamiento (Training data)	$n_{train} = 230$	80 %
Data de prueba (Test data)	$n_{test} = 58$	20 %
Total	$n = 288$	100 %

La Tabla 17 muestra cómo se divide la serie temporal en dos partes para probar el modelo. La data de entrenamiento es el 80 % de todos los datos ($n_{train} = 230$) y se utiliza para ajustar los parámetros y encontrar patrones en la historia. La data de prueba es el 20 % restante ($n_{test} = 58$) y se utiliza para ver cuán bien puede predecir el modelo datos que no ha visto antes. Esto se hace con un total de 288 observaciones para asegurarse de que el modelo tenga una base sólida.

Tabla 18

Resultados del modelo SARIMA$(1,1,1)(0,1,1)[12]$ *del número de visitantes a la Ciudad Inka de Machu Picchu* (2002–2025).

Parámetro	Estimación	Error estándar
AR(1)	0.2085	0.1049
MA(1)	-0.8340	0.0549
SMA(1)	-0.3361	0.1236

Nota. Modelo ajustado a la serie train.

En la Tabla 18 se observa las estimaciones de los parámetros para el modelo SARIMA $(1,1,1)(0,1,1)_{12}$. El componente autorregresivo no estacional AR(1) presenta un coeficiente de 0,2085, mientras que el término de media móvil MA(1) es de −0,8340. Por su parte, el coeficiente estacional de media móvil SMA(1) se estima en −0,3361, lo que confirma la persistencia de la dependencia estacional anual en los residuos de la serie.

Tabla 19

Parámetros de Suavizado del Modelo Holt-Winters (A, A, A) del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Parámetro	**Valor**
Alpha (Nivel)	0.1442034
Beta (Tendencia)	0.1817673
Gamma (Estacionalidad)	0.6775134

En la Tabla 19 se observa los coeficientes de suavizado para el modelo Holt-Winters aplicado a la serie de visitantes. El parámetro alpha ($\alpha = 0{,}1442$) indica que el nivel de la serie se ajusta lentamente, dando mayor peso a las observaciones históricas. El parámetro beta ($\beta = 0{,}1817$) sugiere una actualización moderada de la tendencia monótona identificada previamente por la prueba de Mann-Kendall. Por último, el parámetro gamma ($\gamma = 0{,}6775$) presenta el valor más elevado, lo que revela que el modelo otorga una importancia considerable a los patrones estacionales más recientes para realizar los pronósticos.

Tabla 20

Resultados del modelo STL basado en LOESS y parámetros de suavizamiento del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Parámetros del modelo STL (LOESS)			
Parámetro	**Símbolo**	**Valor**	**Interpretación**
Período estacional	n_p	12	Frecuencia mensual de la serie
Ventana estacional	n_s	∞	Estacionalidad fija (periodic)
Ventana de tendencia	n_t	19	Suavizamiento de la tendencia
Grado de Loess	degree	1	Ajuste local lineal
Salto de Loess	jump	2	Optimización computacional
Componente	**Descripción**		
Estacionalidad	Baja variabilidad		
Tendencia	Crecimiento moderado (estimado con LOESS)		
Residuo	Menor dispersión		

En la Tabla 20 se observa el modelo STL (LOESS) se configura con un período estacional (n_p) de 12 para capturar la frecuencia mensual de la serie y una ventana estacional (n_s) infinita que establece una estacionalidad fija de tipo periódica. La estructura se completa con una ventana de tendencia (n_t) de 19 para el suavizado del crecimiento moderado de la serie y un grado de Loess de 1 que define un ajuste local lineal, optimizado computacionalmente mediante un salto (jump) de 2. Finalmente, la descripción de los componentes resultantes destaca una baja variabilidad en la estacionalidad y una menor dispersión en el residuo.

Tabla 21

Parámetros Técnicos del Modelo TBATS(1, {5,1}, 1, {<12,5>}) *del número de visitantes a la Ciudad Inka de Machu Picchu* (2002–2025).

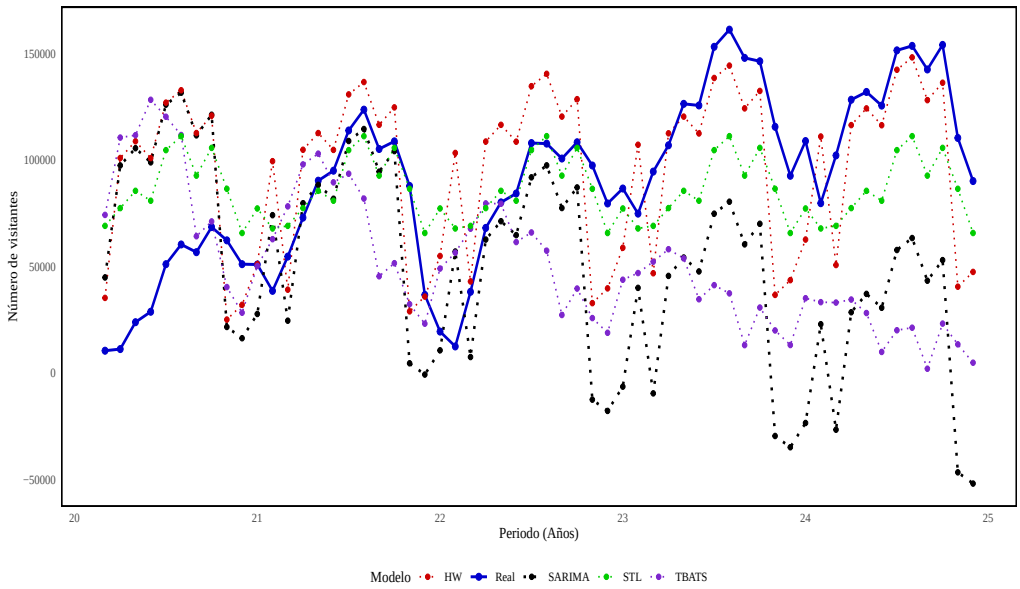
Componente	**Parámetro**	**Valor Estimado**
Box-Cox	λ	1
Suavizamiento	α	0.1796
Tendencia	β	0.1431
Amortiguamiento	ϕ	1
AR (p=5)	$\phi_1, \ldots, \phi_5$	1.015, -0.272, -0.339, 0.242, -0.379
MA (q=1)	θ_1	-0.9373
Estacionalidad	Fourier ($K = 5$)	Período 12

Nota: La configuración TBATS indica el uso de 5 términos de Fourier.

En la Tabla 21 se observa los parámetros estimados para el modelo TBATS(1, {5,1}, 1, {<12,5>}) aplicado a la serie de número de visitantes. En ella se observa que la ausencia de transformación Box-Cox ($\lambda = 1$) indica que la varianza de la serie es estable. El nivel se ajusta de forma moderada mediante el parámetro $\alpha = 0{,}1797$, mientras que la tendencia positiva ($\beta = 0{,}1431$) se proyecta de manera lineal y constante al no presentar amortiguamiento ($\phi = 1$). Asimismo, la dinámica de corto plazo es capturada por un componente autorregresivo de orden 5 (AR=5) que describe la inercia de los datos, junto con un término de media móvil (MA=1) que corrige las desviaciones inmediatas. Finalmente, la estacionalidad de período 12 se modela con 5 pares de armónicos de Fourier, lo que permite representar con precisión los ciclos anuales de la serie.

Figura 23

Comparación de modelos del Número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 23 se observa que aparentemente el modelo STL es el que mejor captura la tendencia y estacionalidad de la serie real, manteniendo una trayectoria más cercana a los valores observados en comparación con las demás alternativas. Mientras que los modelos SARIMA y TBATS presentan un deterioro crítico en su precisión hacia el final del periodo al proyectar incluso valores negativos inconsistentes, el modelo STL logra seguir de forma más estable la fluctuación de visitantes. La brecha visible entre la serie real y todas las estimaciones hacia el año 2025 explica los elevados niveles de error reportados.

Tabla 22

Comparación de métricas de precisión por modelo estadístico del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Modelo (Test)	ME	RMSE	MAE	MPE	MAPE	MASE
SARIMA (Test)	40720.30	74713.62	61859.94	6.69	95.26	6.80
Holt-Winters (Test)	-5137.31	41231.82	33187.85	-47.69	76.39	3.65
STL using Loess (Test)	4143.65	34697.51	29001.04	-38.81	65.88	3.19
TBATS (Test)	38177.78	73183.81	61672.49	-11.49	98.32	6.77

En la Tabla 22 se observa que el modelo STL using Loess (Test) es el que presenta

el mejor ajuste para la serie del número de visitantes, al reportar los valores más bajos en las principales métricas de error. Específicamente, este modelo minimiza el error cuadrático medio (RMSE = 34697.51) y el error absoluto medio (MAE = 29001.04), además de registrar un error porcentual (MAPE) de 65.88 %, el cual, aunque elevado, resulta ser el más óptimo frente a los demás competidores. Asimismo, al evaluar el indicador de escala MASE, el valor de 3.19 obtenido por el modelo STL siendo inferior a los registrados por Holt-Winters (3.65), TBATS (6.77) y SARIMA (6.80), confirma su relativa superioridad predictiva en este conjunto de datos. Por lo tanto, bajo los criterios de magnitud de error analizados, el modelo basado en la descomposición STL se consolida como la alternativa más adecuada entre las evaluadas para representar el comportamiento de la serie.

Tabla 23

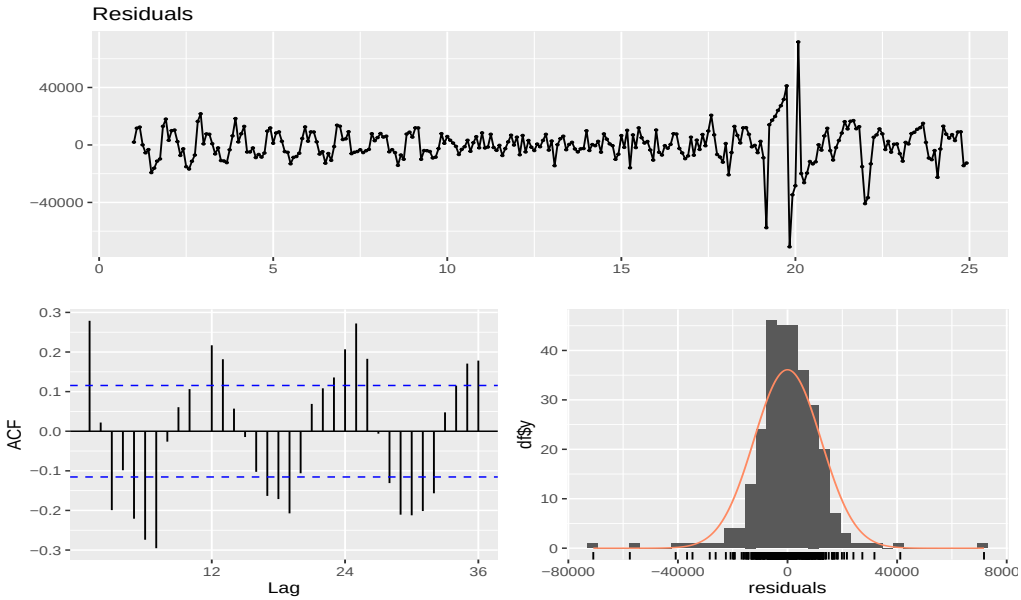
Coeficiente U de Theil para evaluar la capacidad predictiva del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Modelo (Test)	**Theil's U**
SARIMA (Test)	4.425924
Holt-Winters (Test)	4.487363
STL using Loess (Test)	3 .344486
TBATS (Test)	0.92495279

En la Tabla 23 se observa que el modelo STL registra un Coeficiente U de Theil de 3.34, siendo el valor más bajo en comparación con SARIMA (4.43) y Holt-Winters (4.49). Aunque estos resultados superan la unidad lo que sugiere una alta volatilidad en la serie de visitantes que dificulta superar a un pronóstico ingenuo, el modelo STL se posiciona como el de mejor desempeño relativo dentro de este grupo. No obstante, el modelo TBATS destaca con un coeficiente de 0.92, siendo el único que logra una capacidad predictiva superior a la caminata aleatoria para este conjunto de prueba.

Figura 24

Diagnóstico estadístico de los residuos del modelo de pronóstico STL using Loess del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).



En la Figura 24 se observa el diagnóstico de los residuos del modelo, donde el gráfico de serie temporal muestra una varianza no constante con fluctuaciones marcadas hacia el final del periodo. El correlograma (ACF) revela múltiples rezagos que superan las bandas de significancia, lo que confirma la presencia de autocorrelación serial. Finalmente, el histograma y la curva de densidad muestran una distribución con colas pesadas y valores atípicos, lo que indica que los residuos no siguen una distribución normal perfecta. En conjunto, estos resultados sugieren que, aunque el modelo captura parte de la estructura, aún existen patrones sistemáticos y volatilidad en la serie de visitantes que no han sido totalmente extraídos por el modelo. El modelo STL no requiere el supuesto de normalidad debido a su naturaleza no paramétrica. Al utilizar suavizamiento Loess, la estimación se basa en regresiones locales que no dependen de una distribución de probabilidad específica para los errores, permitiendo modelar la serie de visitantes sin las restricciones de los métodos paramétricos tradicionales.

Tabla 24
Test Ljung-Box para la autocorrelación de residuos del número de visitantes a la Ciudad Inka de Machu Picchu (2002–2025).

Ljung-Box test	
P-Value	$2{,}2 \times 10^{-16}$

- H_0: No existe autocorrelación serial (son ruido blanco).
- H_1: Existe autocorrelación serial (no son ruido blanco).

En la Tabla 24 se presenta los resultados de la prueba de Ljung-Box aplicada a los residuos del modelo. Dado que el p-valor obtenido ($2{,}2 \times 10^{-16}$) es significativamente inferior al nivel de significancia de 0.05, se rechaza la hipótesis nula de ausencia de autocorrelación. Este resultado indica que los residuos no se comportan como ruido blanco, sugiriendo que aún existe información sistemática en la serie del número de visitantes que el modelo no ha logrado capturar completamente.

5.2.2 Número pasajeros internacionales en el Aeropuerto Internacional Velasco Astete (2004-2025).

Tabla 25
Resumen estadístico de la serie de número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Alejandro Astete (2004-2025).

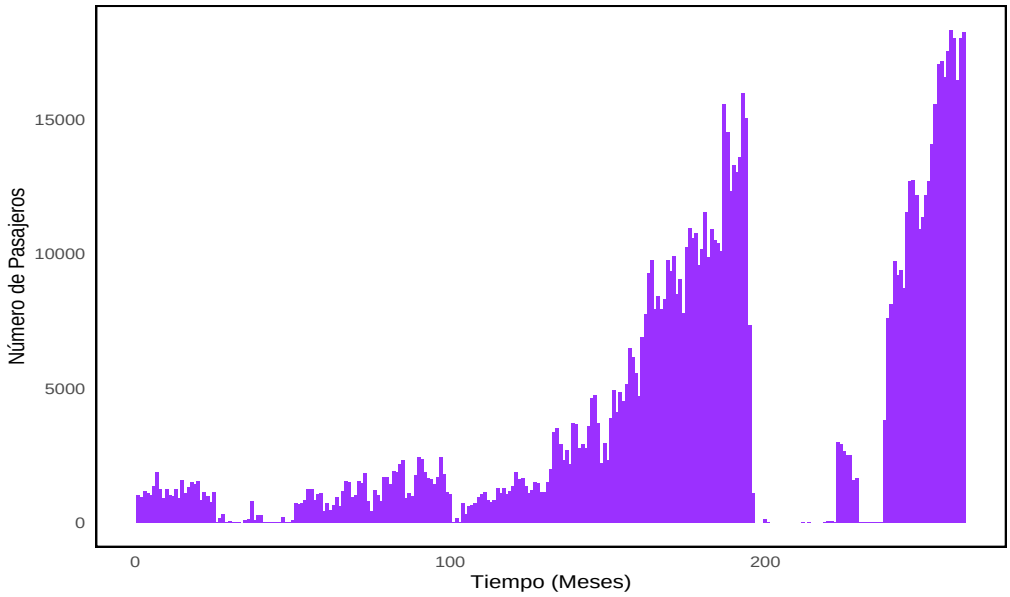
Medida Estadística	**Valor (Número de Pasajeros)**
Valor Mínimo (Min.)	2
Primer Cuartil (1st Qu.)	822
Mediana (Median)	1,520
Tasa media (geométrica)	1,11268
Tercer Cuartil (3rd Qu.)	6,240
Valor Máximo (Max.)	18,319
Valores Faltantes (NA's)	19

Nota. Resumen estadístico del flujo de pasajeros internacionales en el Aeropuerto.

En la Tabla 25 se ve que el número de pasajeros presenta un valor mínimo de 2 y un máximo de 18, 319. Se observa que el primer cuartil del número de pasajeros es de 822, la mediana es de 1, 520 y el tercer cuartil es de 6, 240. Además, en la tabla se ve una tasa media geométrica del número de pasajeros de 1,11268 y la existencia de 19 valores faltantes (NA's) en el registro.

Figura 25

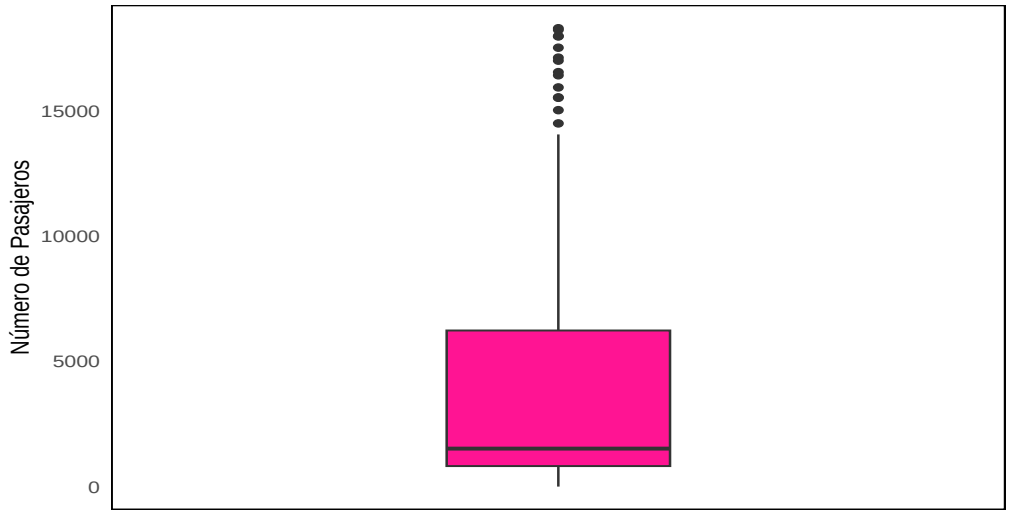
Diagrama de barras del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 25 se observa cómo ha cambiado el número de pasajeros durante los meses. Se constata que, en los primeros lapsos de tiempo, la cifra de pasajeros fue relativamente baja, con cifras cercanas a los 500 y 2.000 pasajeros aproximadamente. Después, el número de pasajeros muestra un aumento gradual, llegando a cerca de 10,000 pasajeros aproximadamente alrededor del mes 170. Además, se observan variaciones significativas en ciertos periodos, donde la cifra de pasajeros cae drásticamente y luego vuelve a crecer. Por último, en los últimos meses de la serie, se llega a un máximo de alrededor de 18,000 pasajeros aproximadamente. Esto muestra que el flujo de pasajeros ha ido creciendo con el tiempo.

Figura 26

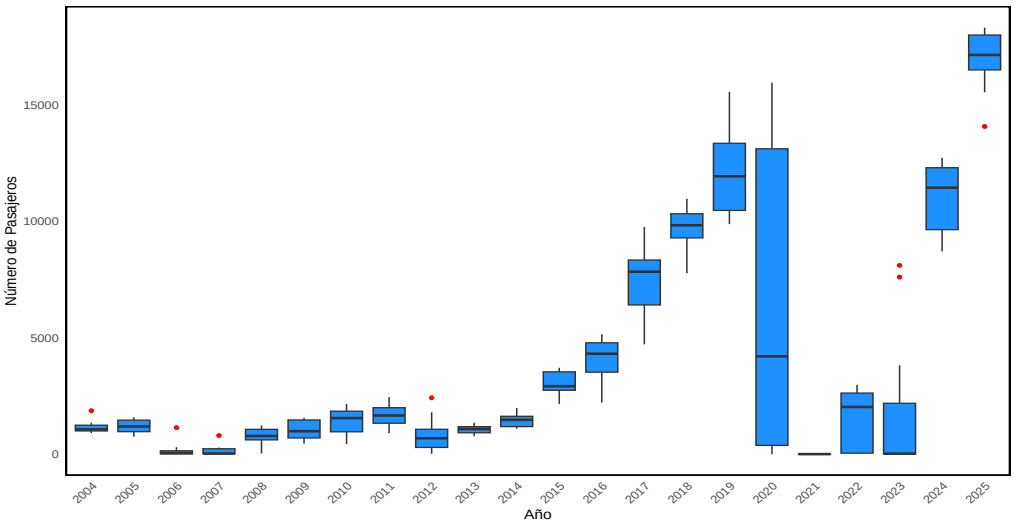
Boxplot del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 26 se observa el boxplot del número de pasajeros presenta una mediana posicionada en la parte inferior de la caja, lo que indica una distribución asimétrica hacia los valores bajos. Se observa que el rango intercuartílico concentra la mayor densidad de datos por debajo de los 7, 000 pasajeros aproximadamente, mientras que el bigote superior se extiende hasta alcanzar niveles cercanos a los 14, 000 aproximadamente. Finalmente, en la figura se ve una presencia significativa de valores atípicos (outliers) representados por puntos negros, los cuales superan los 15, 000 pasajeros y llegan hasta aproximadamente los 18, 000.

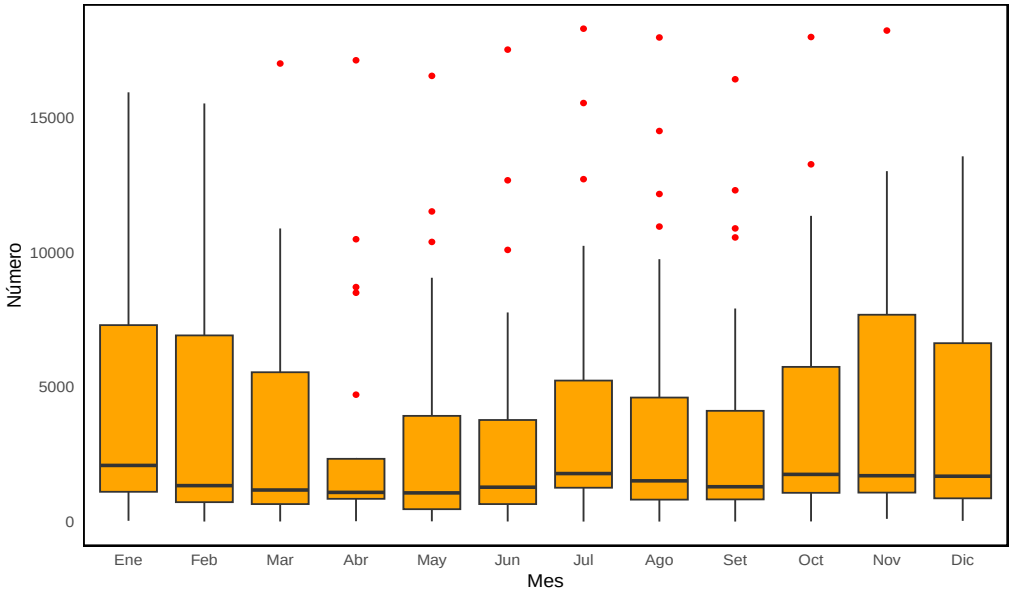
Figura 27

Boxplot por año del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 27 se observa el boxplot por año del número de pasajeros muestra una tendencia creciente desde el año 2004 hasta el 2019, alcanzando medianas superiores a los 10, 000 pasajeros al final de dicho periodo. Se observa una caída drástica y una alta variabilidad en el año 2020, seguida de una recuperación marcada a partir del 2022 que culmina en el valor máximo de la serie en 2025 con más de 15, 000 pasajeros. Finalmente, en la figura se ve la presencia de valores atípicos (puntos rojos) en diversos años..

Figura 28
Boxplot por mes del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 28 se observa el boxplot por mes del número de pasajeros presenta variaciones mensuales a lo largo del año, con las medianas más altas en los meses de enero y julio, superando los 2, 000 pasajeros. Se observa que el mes de noviembre muestra la mayor dispersión en el rango intercuartílico, mientras que los meses de julio y agosto presentan una cantidad significativa de valores atípicos (puntos rojos) que exceden los 15, 000 pasajeros. Finalmente, en la figura se ve que el número de pasajeros mantiene una base constante cercana a los 1, 500 en la mayoría de los meses, pero con una alta volatilidad reflejada en los extensos bigotes superiores de cada periodo.

Tabla 26

Análisis de valores faltantes (NA) detectados del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

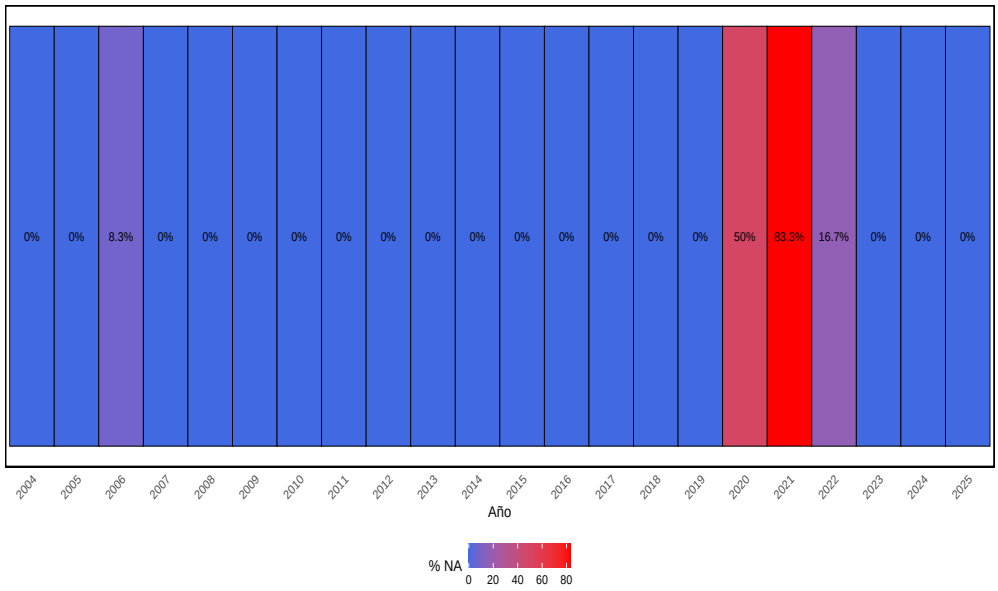
Descripción	Resultado
Total de datos faltantes (NA)	19
Porcentaje de (NA)	7.2 %

Nota. El porcentaje se calculó sobre el total de la muestra analizada.

En la Tabla 26 se observa que la variable correspondiente al número de pasajeros presenta un total de 19 valores faltantes (NA), lo que representa el 7.2 % de la serie histórica analizada. Estos resultados cuantifican la omisión de registros en dicha serie temporal.

Figura 29

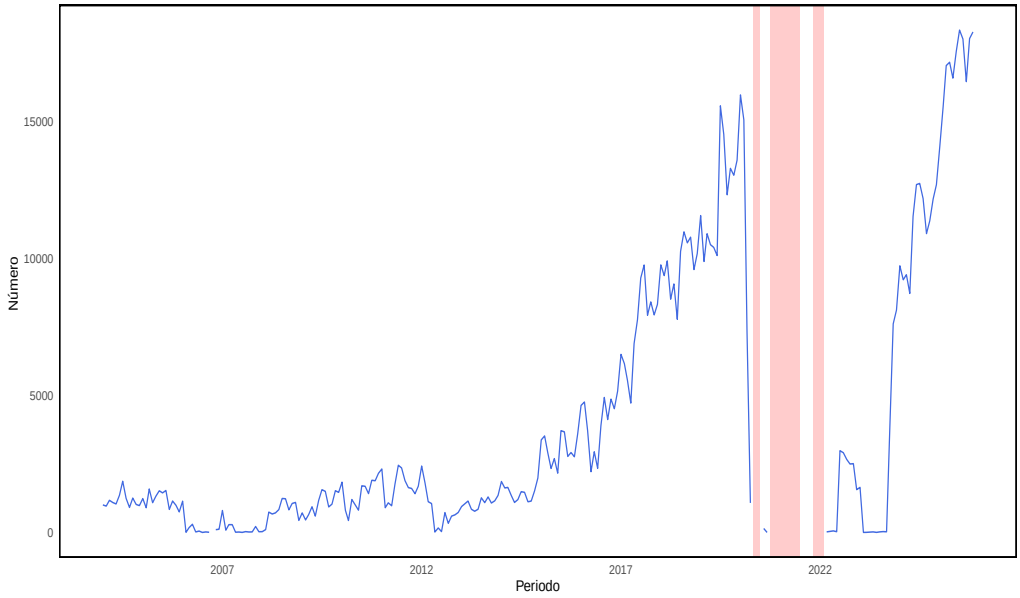
Mapa de calor de datos faltantes del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 29 se observa el mapa de calor que el porcentaje de valores faltantes (% NA) del número de pasajeros es nulo (0 %) en la mayoría de los años, lo que indica una alta integridad de datos entre 2004 y 2019, así como de 2023 a 2025. Se observa que el año 2021 presenta la mayor pérdida de información con un 83,3 %, seguido por el año 2020 con un 50 % y el 2022 con un 16,7 %. Finalmente, en la figura se ve que antes de este periodo crítico, solo el año 2006 registró una ausencia menor de datos correspondiente al 8,3 %.

Figura 30

Serie con intervalo de datos faltantes del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 30 se observa que el número de pasajeros presenta interrupciones significativas en su registro, representadas por las bandas de color rojo que indican intervalos de datos faltantes. Se observa que estas bandas se concentran principalmente entre los años 2020 y 2022, periodo donde la continuidad de la serie se rompe de manera prolongada. Finalmente, en la figura se ve que, tras el último intervalo de datos faltantes marcado en rojo, el número de pasajeros reinicia su registro con una tendencia de crecimiento acelerado.

Tabla 27

Identificación de Gaps (Vacíos) detectados en la serie número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

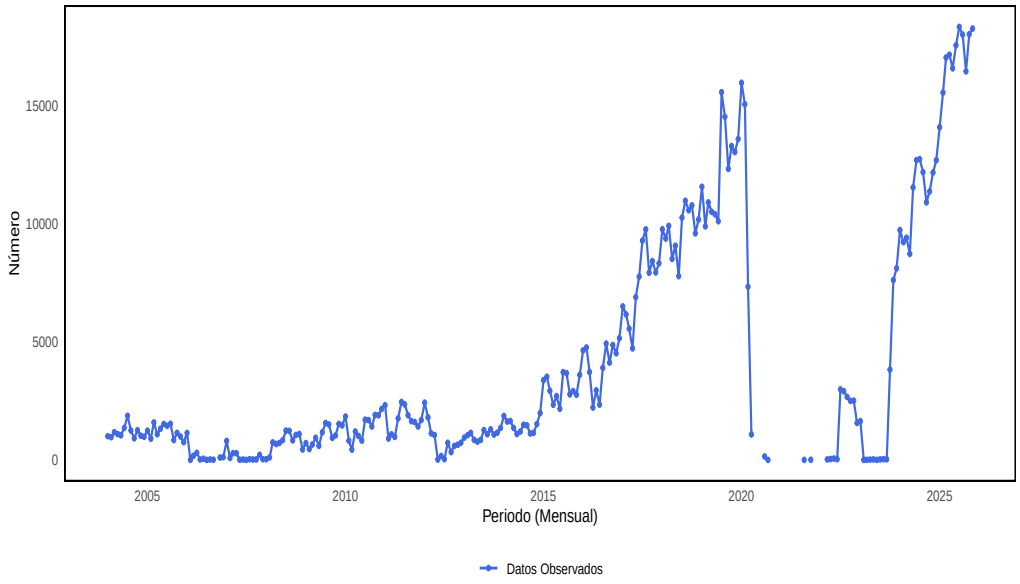
Intervalo (L_i, L_s)	**Longitud**
(34, 34)	1
(197, 199)	3
(202, 211)	10
(213, 213)	1
(215, 218)	4

Nota. La longitud indica el intervalo de datos faltantes con (NA).

En la Tabla 27 se observa se observa la ubicación y extensión de los vacíos (gaps) detectados en la serie histórica de visitantes, identificándose cinco intervalos específicos de datos faltantes. El intervalo de mayor longitud se registra entre los puntos 202 y 211 con un total de 10 registros ausentes, seguido por vacíos menores de 4, 3 y 1 unidades de longitud en los periodos restantes de la serie.

Figura 31

Serie temporal con datos faltantes del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 31 se observa que el número de pasajeros mantuvo un crecimiento sostenido desde 2015 hasta alcanzar un pico cercano a los 15,000 usuarios en 2020; sin embargo, tras una caída abrupta y periodos de inactividad, la serie muestra una recuperación que supera los 18,000 registros hacia 2026. Ante la presencia de vacíos detectados en este comportamiento, se requiere la aplicación de la técnica (IFI) para efectuar la imputación necesaria que permita dar continuidad al análisis estadístico.

Tabla 28

Cronología e identificación de intervalos con datos faltantes (Gaps) en la serie número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

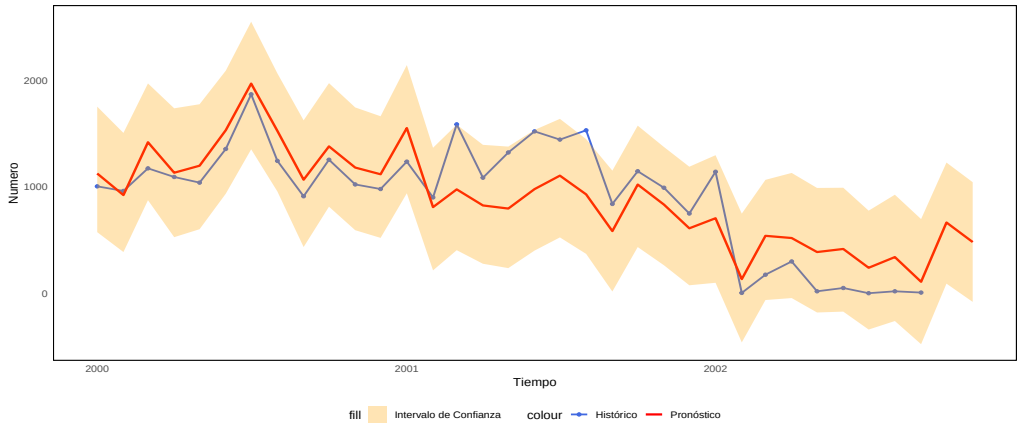
ID	Intervalo (L_i, L_s)	Fecha Inicio
Fecha Fin	**Longitud**	
1	(34, 34)	1
2	(197, 199)	3
3	(202, 211)	10
4	(213, 213)	1
5	(215, 218)	4

Nota. Los ID son los Gaps

En la Tabla 28 se observa la identificación de cinco intervalos de vacíos (gaps) en la serie histórica de visitantes, destacando el periodo comprendido entre los puntos 202 y 211 con la mayor longitud de 10 datos faltantes. Con base en la propuesta teórica de detectar cambios abruptos en series con datos ausentes, estos vacíos requieren la aplicación de la técnica (IFI) para una imputación que preserve la estructura de la serie, permitiendo así un análisis de impacto global preciso a pesar de las discontinuidades detectadas.

Figura 32

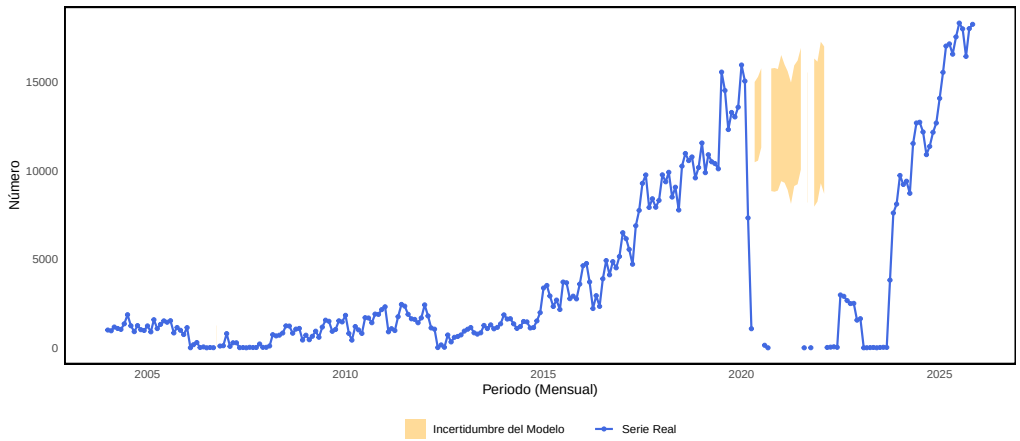
Delimitacion de intervalos de confianza mediante el metodo (IFI) y el algoritmo Prophet en la serie número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 32 se observa el proceso de entrenamiento del modelo Prophet aplicado sobre los datos históricos previos a la interrupción de la serie. El modelo utiliza una estacionalidad anual para generar un pronóstico (línea roja) que estima el comportamiento esperado de los pasajeros durante el intervalo de datos faltantes, acompañado de un intervalo de confianza al 95 % (área naranja) que cuantifica la incertidumbre de la predicción. Esta fase de entrenamiento es fundamental para la técnica IFI, ya que permite detectar cambios estructurales y asegurar que la imputación de los vacíos sea estadísticamente consistente con la tendencia y los ciclos detectados antes del quiebre.

Figura 33

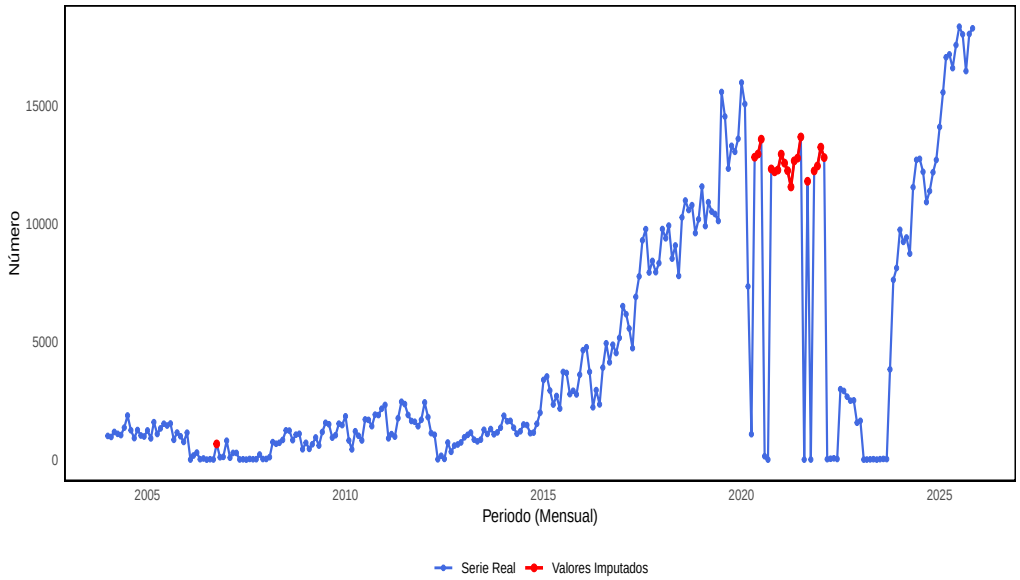
Delimitación del intervalo de predicción mediante el modelo Prophet para la imputación de vacíos en la serie número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 33 se observa cómo el modelo Prophet delimita el intervalo específico de datos faltantes, iniciando la predicción a partir del periodo histórico que se sitúa cerca del año 2002, donde el número de pasajeros oscilaba entre 1,000 y 2,000 unidades. Se proyecta una tendencia estable que se enmarca dentro de un intervalo de confianza (área naranja) cuyos límites superior e inferior se mantienen aproximadamente entre los 7000 y 16000 pasajeros aproximadamente.

Figura 34

Reconstrucción de la serie de tiempo mediante la imputación de valores estimados con el método Prophet + (IFI) del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 34 se observa la serie de tiempo completada tras integrar el modelo Prophet como motor de estimación dentro del algoritmo IFI. Los puntos rojos representan los valores imputados que, numéricamente, se estabilizan en torno a los 12,000 pasajeros aproximadamente, manteniendo la coherencia con la tendencia histórica previa. La validez de esta reconstrucción se fundamenta en la capacidad del modelo para proyectar el comportamiento esperado, permitiendo que el marco IFI realice la detección de cambios abruptos justo en la frontera posterior al vacío; es decir, al contrastar estos puntos rojos con el retorno de los datos reales observados. Dado que la transición entre la predicción de Prophet y los nuevos datos registrados es fluida, se confirma que la estructura de la serie no sufrió un cambio de régimen durante el periodo de ausencia de información.

Tabla 29

Valores imputados e intervalos de confianza mediante el método Prophet + (IFI) por periodos de discontinuidad para la serie número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

Fecha	Valor Imputado	I. Inferior	I. Superior	ID
2006-10-01	665.82	127.91	1,287.84	1
2020-05-01	12,795.49	10,366.37	15,151.49	2
2020-06-01	12,929.29	10,739.14	15,342.72	2
2020-07-01	13,558.43	11,441.79	15,758.24	2
2020-10-01	12,299.30	8,647.98	15,806.46	3
2020-11-01	12,178.63	8,862.22	15,446.40	3
2020-12-01	12,250.61	8,744.29	15,624.75	3
2021-01-01	12,926.21	9,298.95	16,395.86	3
2021-02-01	12,543.83	8,887.37	15,971.61	3
2021-03-01	12,225.97	8,644.67	15,731.85	3
2021-04-01	11,538.00	8,089.52	14,849.43	3
2021-05-01	12,640.95	9,331.74	16,175.27	3
2021-06-01	12,756.74	9,435.34	16,273.92	3
2021-07-01	13,652.17	10,259.08	17,129.38	3
2021-09-01	11,777.27	8,038.09	15,519.77	4
2021-11-01	12,216.50	7,893.75	16,424.60	5
2021-12-01	12,422.79	8,259.91	16,562.94	5
2022-01-01	13,220.48	9,185.98	17,726.88	5
2022-02-01	12,778.92	8,467.60	16,897.49	5

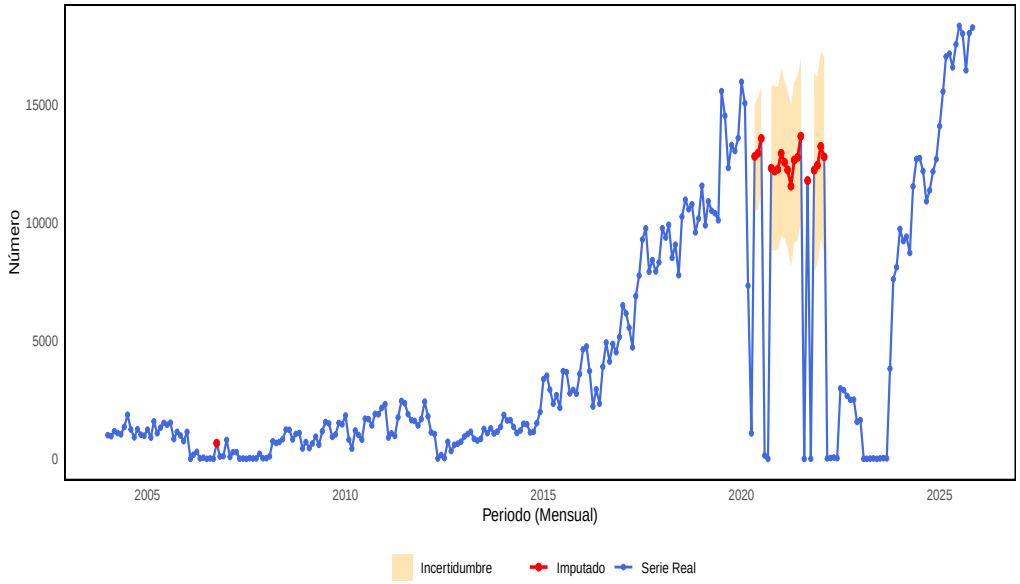
Nota. Valores calculados mediante el método de imputación.

En la Tabla 29 se observa los resultados numéricos de la fase de imputación para los cinco intervalos de datos faltantes identificados en la serie. Se observa que el modelo Prophet, actuando como motor de estimación, genera valores que guardan una estrecha relación con la magnitud de la serie en cada periodo; por ejemplo, para el primer gap (ID 1) en 2006, se imputa un valor de 665.82 pasajeros, mientras que para los vacíos registrados entre 2020 y 2022 (IDs 2 al 5), los valores estimados ascienden significativamente, situándose mayoritariamente el

rango varia de los 12,795.49 a 13,652.17 pasajeros. Cada punto imputado incluye su respectivo intervalo de confianza (inferior y superior), los cuales delimitan la variabilidad esperada y sirven como base para que el algoritmo IFI, en su etapa posterior.

Figura 35

Serie de tiempo con imputación completa e intervalo de confianza, mediante el método Prophet + IFI número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 35 se observa la serie de tiempo en una etapa donde el modelo Prophet ya ha generado los valores de reemplazo para el vacío detectado, representados por los puntos rojos que se sitúan en un nivel de aproximadamente 1,000 pasajeros. Es importante precisar que, en este punto de la visualización, el algoritmo IFI aún no ha iniciado la fase de detección de cambios abruptos; actualmente se encuentra en la etapa de imputación técnica, proporcionando la trayectoria necesaria para la continuidad de la serie. Esta fase es preparatoria y fundamental, ya que establece la línea base que el marco IFI utilizará posteriormente para contrastar con los datos reales una vez que la observación se reanude, permitiendo validar si la estructura del sistema permaneció estable durante la interrupción.

Tabla 30

Identificación de anomalías detectadas en la serie número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

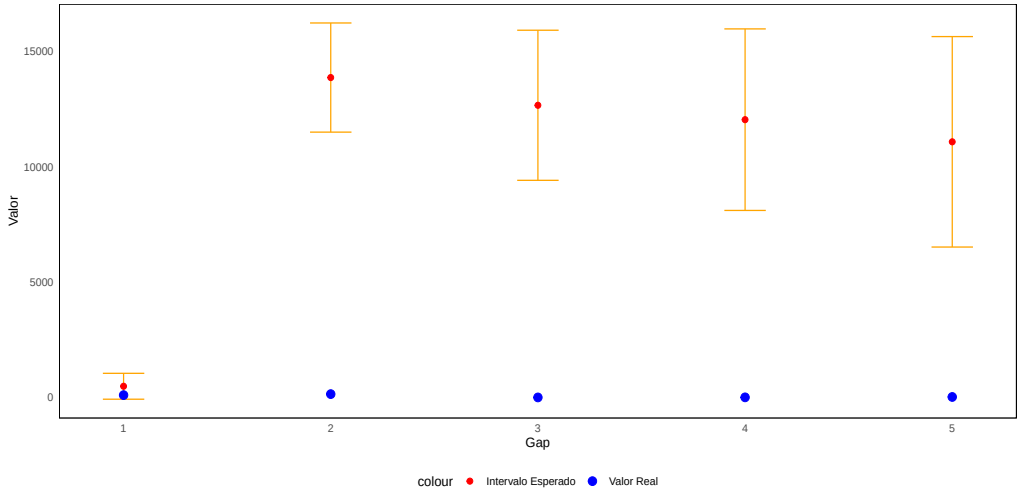
Fecha (ds)	Valor Observado (y)
2020-08-01	146
2021-08-01	2
2021-10-01	6
2022-03-01	22

Nota. Los valores representan puntos identificados como cambios abruptos.

En la Tabla 30 se observa los valores registrados inmediatamente después de la finalización de los intervalos de datos faltantes (gaps), los cuales han sido seleccionados por el algoritmo IFI para la fase de evaluación. Estos registros, que corresponden a fechas entre agosto de 2020 y marzo de 2022 con valores numéricos muy bajos (entre 2 y 146 pasajeros), representan la frontera posterior donde el sistema reanuda la observación. El objetivo de identificar estos puntos específicos es permitir que el marco IFI realice el contraste estadístico frente a lo que el modelo Prophet había proyectado; de esta manera, se podrá determinar formalmente si estos valores constituyen un cambio abrupto de régimen o si representan el inicio de una nueva dinámica en el flujo de pasajeros internacionales tras los periodos de inactividad.

Figura 36

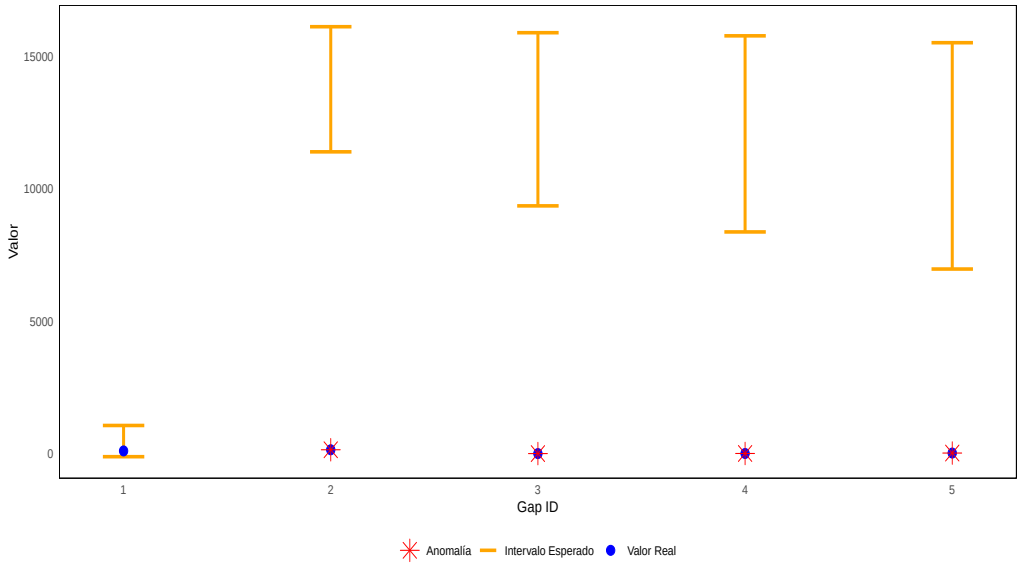
Detección de anomalías mediante intervalos de confianza IFI por bloque de discontinuidad del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 36 se observa la validación final del algoritmo IFI para cada intervalo de la serie. Los puntos azules representan los valores reales observados tras los vacíos, mientras que las barras naranjas delimitan el rango de normalidad estadística esperado. Se observa que en el primer intervalo el valor real es consistente con la serie histórica; sin embargo, en los casos 2, 3, 4 y 5, los datos reales se sitúan drásticamente fuera del margen esperado, posicionándose cerca de cero. Esta separación numérica confirma formalmente la existencia de cambios abruptos de régimen, evidenciando que la serie no retomó su comportamiento habitual, sino que sufrió una ruptura estructural definitiva tras los periodos de interrupción.

Figura 37

Identificación de cambios abruptos basados en el criterio de exclusión del soporte estadístico número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 37 se observa la identificación de anomalías en la serie tras el proceso de validación. Los puntos azules representan los valores reales registrados al finalizar cada vacío, contrastados con los intervalos esperados (barras naranjas). Para los casos 2, 3, 4 y 5, se observa que los valores reales caen fuera del rango de normalidad, siendo marcados con un símbolo de asterisco rojo como anomalías confirmadas. Esta discrepancia extrema entre la expectativa estadística y la realidad observada permite al algoritmo clasificar estos puntos como cambios abruptos de régimen, evidenciando una ruptura total en la dinámica de pasajeros internacionales en dichos periodos.

Tabla 31

Resultados de la Imputación por Intervalos de Pronóstico (IFI) para los Gaps detectados en la serie número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

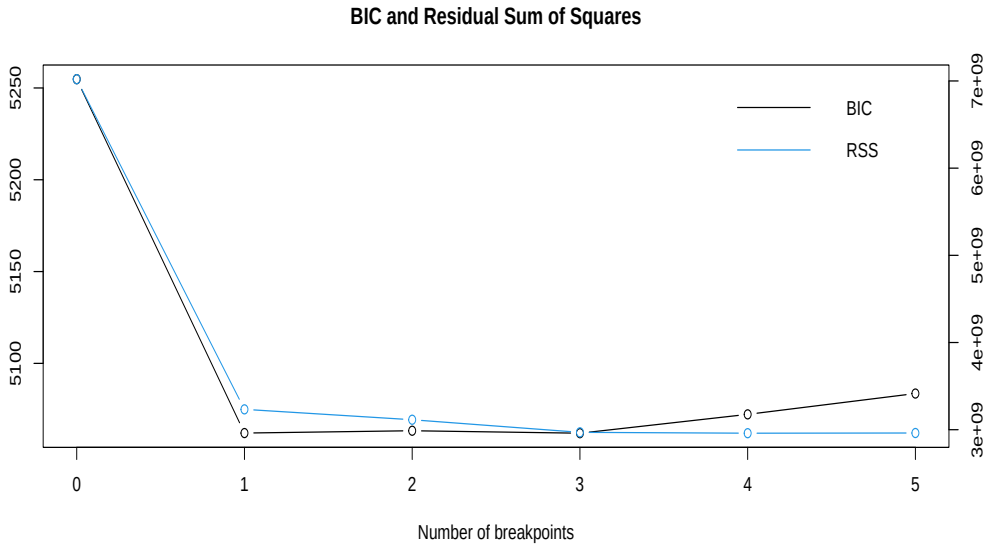
ID	Valor Real	Límite Inferior	Límite Superior	Resultado IFI
1	102	-75.1188	1046.5010	FALSE
2	146	11492.2763	16219.9130	TRUE
3	2	9405.4089	15905.7140	TRUE
4	6	8100.6030	15963.4990	TRUE
5	22	6515.1661	15630.2650	TRUE

Nota. El resultado TRUE indica la detección de un cambio abrupto fuera del intervalo.

En la Tabla 31 se observa la evaluación estadística del algoritmo IFI para cada intervalo analizado. Se observa que en el primer registro (ID 1), el valor real de 102 pasajeros se encuentra dentro del rango de normalidad, resultando en una detección negativa (FALSE). Por el contrario, para los registros del 2 al 5, los valores observados (que oscilan entre 2 y 146 pasajeros) se sitúan drásticamente por debajo de los límites inferiores calculados, que superan los 6515.1661 pasajeros. En consecuencia, el algoritmo arroja un resultado positivo (TRUE) para estos casos, confirmando formalmente la presencia de cambios abruptos. Estos hallazgos validan que la serie experimentó una ruptura estructural en su dinámica habitual tras los periodos de inactividad, imposibilitando una transición estadística convencional.

Figura 38

Optimización de rupturas mediante la prueba de Bai-Perron y criterios BIC del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 38 se observa el comportamiento de los criterios de selección para determinar las rupturas estructurales en toda la serie de tiempo. La curva del Criterio de Información Bayesiano (BIC) muestra un descenso pronunciado hasta alcanzar su punto mínimo en el valor de 3 quiebres, lo que lo define como el modelo más parsimonioso y óptimo. Por su parte, la Suma de Cuadrados de los Residuos (RSS) continúa disminuyendo ligeramente, pero sin mejoras significativas después de este punto. Este análisis gráfico confirma que dividir la serie completa en tres quiebres proporciona el mejor equilibrio entre la precisión del ajuste y la complejidad del modelo, validando estadísticamente los cambios de régimen identificados en el flujo de pasajeros.

Tabla 32

Bai-Perron para la determinación de los puntos de ruptura estructural del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

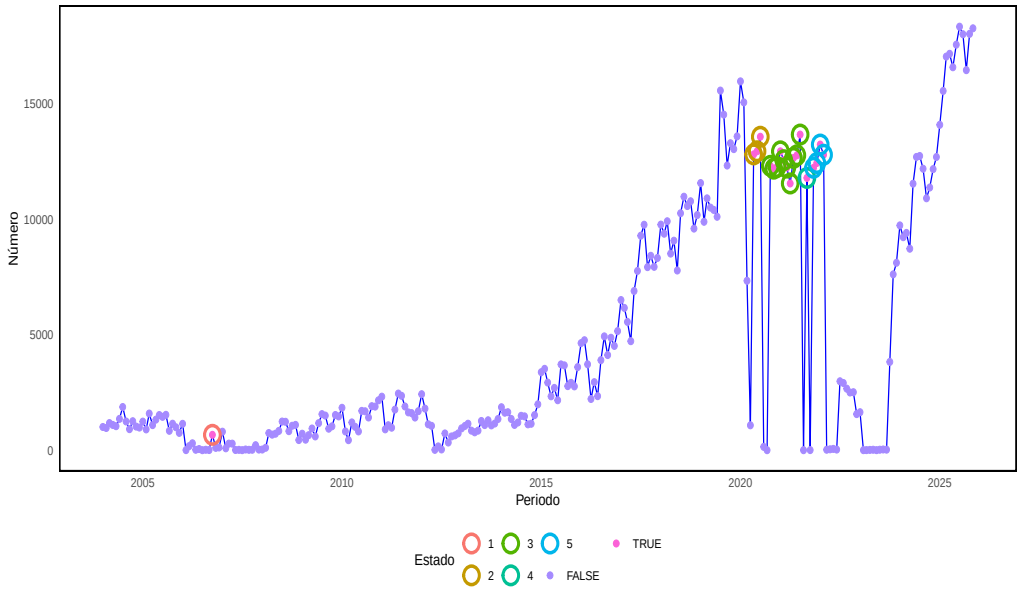
Número de rupturas (m)	**BIC**	**Selección**
3	5063	Óptimo

En la Tabla 32 se observa los resultados del análisis de estabilidad aplicado a la serie completa de pasajeros internacionales. Mediante el método de Breakpoints, se determinó de

manera estadística que existen 3 rupturas estructurales significativas a lo largo de toda la trayectoria temporal. La selección de este número como el óptimo se fundamenta en el valor del Criterio de Información Bayesiano (BIC) de 5063, el cual minimiza el error de ajuste. Este hallazgo confirma que la serie completa no se comporta de manera uniforme, sino que está dividida en regímenes diferenciados, validando que el flujo de pasajeros ha atravesado cambios permanentes en su dinámica global.

Figura 39

Identificación de anomalías estructurales mediante el método IFI en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 39 se observa la aplicación integral del algoritmo IFI sobre la serie de pasajeros. Cada círculo de color identifica uno de los cinco intervalos de datos faltantes (gaps) analizados. Los puntos internos de color fucsia (TRUE) señalan los momentos donde el algoritmo confirmó formalmente un cambio abrupto de régimen, especialmente evidentes en los periodos de 2020 a 2022, donde la serie real se desploma y se aleja drásticamente de la tendencia esperada. Por el contrario, el punto azul (FALSE) en el primer intervalo indica que, en esa fecha, el retorno de los datos fue consistente con el comportamiento histórico, sin presentar una ruptura estructural. Esta representación espacial permite localizar con precisión los eventos de inestabilidad que afectaron la dinámica global del flujo aeroportuario.

Tabla 33

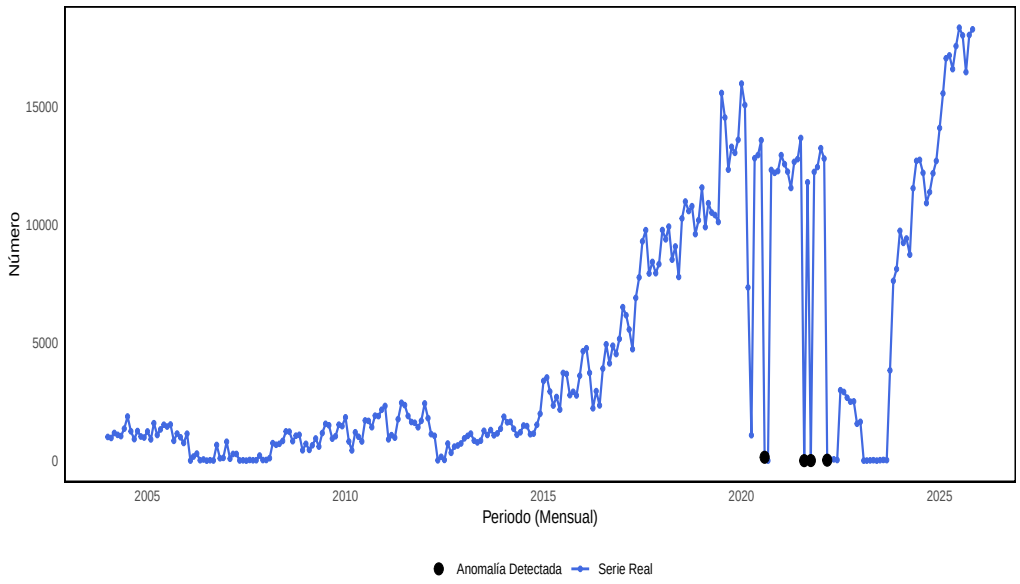
Observaciones identificadas como anomalías en la serie temporal en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

Fecha (ds)	Valor observado
2020-08-01	146
2021-08-01	2
2021-10-01	6
2022-03-01	22

En la Tabla 33 número presenta las observaciones identificadas como anomalías en la serie temporal del número de pasajeros. Las fechas de agosto de 2020, agosto de 2021, octubre de 2021 y marzo de 2022 tienen cifras muy bajas que no siguen la tendencia y el patrón habitual de la serie. Es importante detectar estos puntos extraños para poder hacer predicciones precisas en el futuro. Si no se detectan, pueden afectar la exactitud de los pronósticos que se hagan más adelante.

Figura 40

Identificación de anomalías estructurales mediante el método IFI en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 40 se observa los puntos de la serie real que han sido clasificados como anomalías tras la evaluación del algoritmo. Estos puntos, representados por círculos negros de mayor tamaño, se localizan exclusivamente en el periodo comprendido entre 2020 y 2022.

La importancia de esta visualización radica en que identifica los momentos exactos donde el flujo de pasajeros se reanudó en niveles excepcionalmente bajos tras los vacíos de información, rompiendo con la estructura estadística previa. Al ser marcados como anomalías detectadas, se confirma que en dichas fechas existió un cambio abrupto de régimen, validando que factores externos alteraron permanentemente el comportamiento normal de la serie durante esos intervalos específicos.

Tabla 34

Detección de anomalías por bloque de discontinuidad mediante el método IFI del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

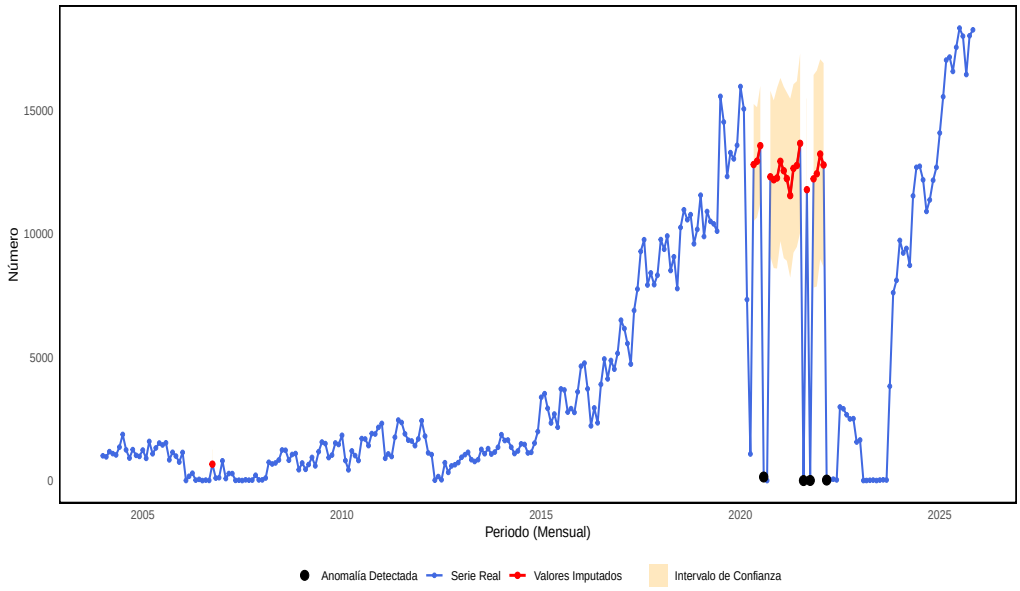
Bloque (ID)	Inicio	Fin	Detección de Anomalía
1	34	34	No detectada
2	197	199	Detectada
3	202	211	Detectada
4	213	213	Detectada
5	215	218	Detectada

Nota. La columna Detección de Anomalía indica si el comportamiento tras el intervalo (gap) es irregular según el umbral del método IFI.

En la Tabla 34 se observa los resultados de la detección de anomalías para los bloques analizados en la serie. Se observa que el Bloque 1 no presenta anomalías detectadas, lo que indica que en ese punto la serie mantuvo su comportamiento estructural. Por el contrario, los bloques del 2 al 5 muestran una detección positiva de anomalías. Estos bloques, que abarcan secuencias de puntos específicos (como el intervalo del 202 al 211 o del 215 al 218), identifican las zonas donde el flujo de pasajeros sufrió cambios de régimen significativos. Esta clasificación por bloques permite delimitar con exactitud las ventanas temporales donde la estabilidad de la serie se vio interrumpida, confirmando la existencia de múltiples periodos con comportamientos atípicos tras los vacíos de información.

Figura 41

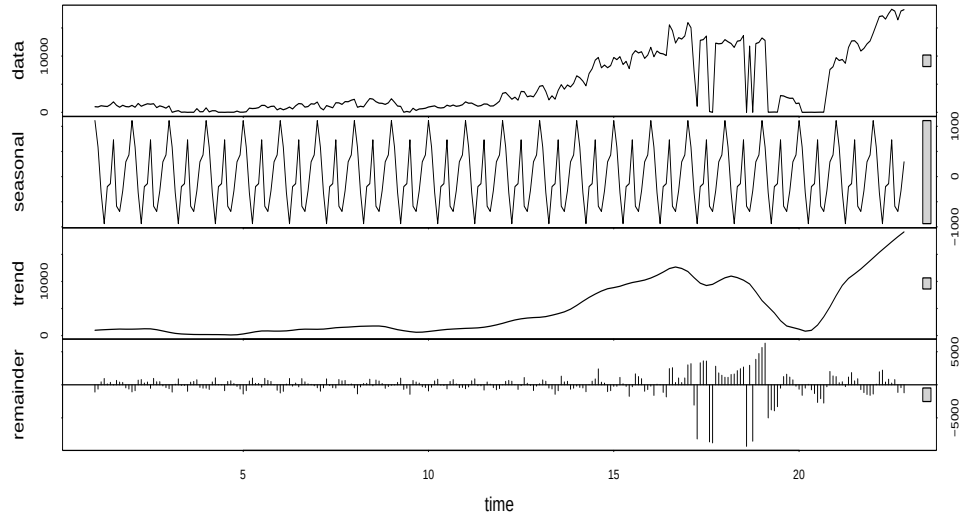
Reconstrucción de la serie temporal y diagnóstico de anomalías estructurales número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 41 se observa el resultado consolidado del proceso de reconstrucción y análisis de la serie temporal. Los puntos rojos y el área sombreada naranja representan los valores imputados junto a su intervalo de confianza, los cuales cubren los vacíos de información manteniendo la coherencia con la tendencia de aproximadamente 13,000 pasajeros alcanzada antes de las interrupciones. No obstante, la serie real (línea azul) muestra caídas drásticas inmediatamente después de estos periodos, puntos que han sido identificados con círculos negros como anomalías confirmadas. Esta visualización permite contrastar la trayectoria que el flujo de pasajeros debería haber seguido según su comportamiento histórico frente a la ruptura estructural real ocurrida entre 2020 y 2022, validando mediante el algoritmo la existencia de cambios abruptos de régimen que desviaron la serie de su dinámica normal.

Figura 42

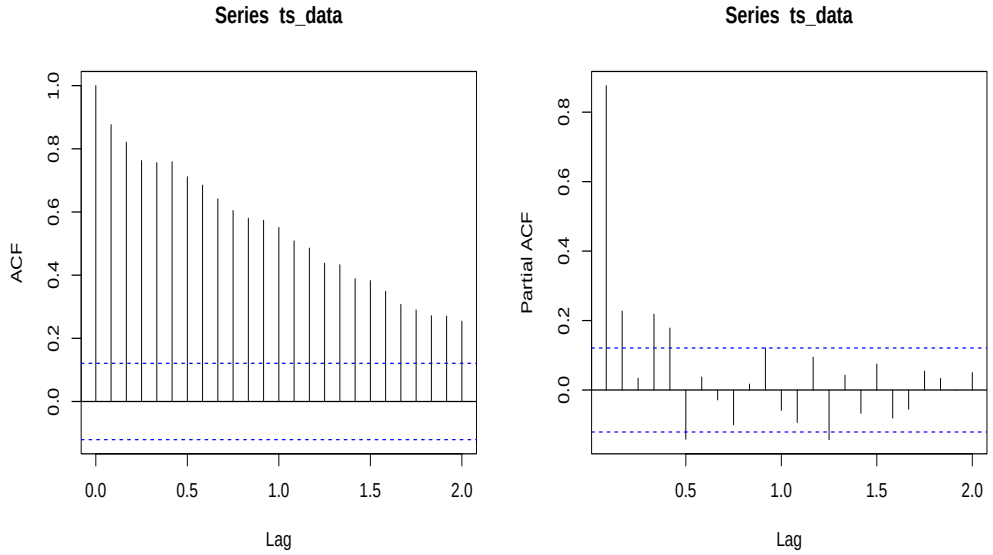
Comparación de modelos del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 42 se observa la descomposición estructural de la serie original en sus tres componentes fundamentales: estacionalidad, tendencia y residuo. El panel de la tendencia muestra un crecimiento sostenido a largo plazo que se ve interrumpido de manera drástica alrededor del periodo 20, coincidiendo con la caída severa observada en los datos reales. Por su parte, el componente estacional exhibe un patrón cíclico constante y regular, lo que confirma la existencia de fluctuaciones periódicas propias de la demanda aeroportuaria. Finalmente, el panel de residuos revela una alta volatilidad y picos negativos pronunciados en la etapa final de la serie, lo que indica que los eventos disruptivos recientes no pueden ser explicados por la tendencia ni por la estacionalidad, subrayando la naturaleza excepcional de los cambios de régimen identificados previamente.

Figura 43

Funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) del número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 43 se observa las funciones de autocorrelación aplicadas a la serie del número de pasajeros internacionales. El gráfico (ACF) muestra una persistencia extremadamente alta, con coeficientes que decrecen de forma muy lenta y lineal, lo cual es una evidencia clásica de no estacionariedad y de la presencia de una tendencia alcista dominante en el flujo de viajeros. Por su parte, el gráfico (PACF) muestra un primer rezago altamente significativo (cercano a 0.9), tras el cual la correlación cae abruptamente, sugiriendo que el comportamiento actual del número de pasajeros está fuertemente condicionado por su valor inmediatamente anterior. Estos resultados indican que, para un modelamiento adecuado para eliminar la tendencia, permitiendo así capturar la verdadera estructura estocástica de la serie.

Tabla 35

Prueba de Jarque-Bera en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

Jarque-Bera	
p-valor	1.065e-11

- H_0: Los datos siguen una distribución normal.
- H_1: Los datos no siguen una distribución normal.

En la Tabla 35 se observa un p-valor de $1{,}065 \times 10^{-11}$, valor que es significativamente inferior al nivel de significancia estandar del 5 % (0.05). Ante este resultado, existe evidencia estadística suficiente para rechazar la hipótesis nula (H_0) en favor de la hipótesis alternativa (H_1). En consecuencia, se concluye que los residuos de la serie del número de pasajeros no se distribuyen de forma normal, presentando un comportamiento con exceso de curtosis y asimetría. Este diagnóstico es fundamental para la fase de modelamiento, ya que sugiere que las estimaciones deben considerar métodos robustos o transformaciones adicionales para corregir la falta de normalidad en las perturbaciones del modelo.

Tabla 36
Prueba de Dickey-Fuller Aumentada en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

Dickey-Fuller	
p-valor	0.6139

- H_0: La serie no es estacionaria.
- H_1: La serie es estacionaria.

En la Tabla 36 se observa un p-valor de 0.6139, el cual es considerablemente superior al nivel de significancia estandar del 5 % (0.05). En consecuencia, no existe evidencia estadística suficiente para rechazar la hipótesis nula (H_0), confirmando que la serie original del número de pasajeros es no estacionaria. Este hallazgo valida técnicamente las observaciones previas de los gráficos ACF y PACF, indicando que la serie posee una tendencia que debe ser eliminada mediante procesos de diferenciación antes de proceder con el modelamiento autorregresivo.

Tabla 37
Prueba de Mann-Kendall en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

Mann-Kendall	
p-valor	2.22e-16

- H_0: No existe tendencia en la serie de tiempo.

- H_1: Existe una tendencia significativa en la serie de tiempo.

En la Tabla 37 se observa un p-valor de $2{,}22 \times 10^{-16}$, el cual es extremadamente inferior al nivel de significancia estandar de 0.05. Este resultado proporciona evidencia estadística suficiente para rechazar la hipótesis nula (H_0), confirmando de manera concluyente la presencia de una tendencia significativa en el flujo de pasajeros internacionales. Este hallazgo es coherente con los resultados de la prueba Dickey-Fuller y el análisis de autocorrelación previos, ratificando que el comportamiento de la serie está dominado por un crecimiento (o decrecimiento) persistente en el tiempo que debe ser considerado en el modelamiento final.

Tabla 38
Prueba de Kruskal-Wallis en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

Kruskal-Wallis	
p-valor	0.8279

- H_0: No existen diferencias significativas entre los grupos analizados.
- H_1: Al menos un grupo presenta una mediana significativamente distinta a los demás.

En la Tabla 38 se observa un p-valor de 0.8279, el cual es muy superior al nivel de significancia estandar de 0.05. Ante este resultado, no se puede rechazar la hipótesis nula (H_0), lo que sugiere que, bajo esta prueba específica, no se detectan diferencias significativas en el número de pasajeros entre los periodos comparados. Este hallazgo es relevante para el modelamiento, ya que indica que, a pesar de las fluctuaciones observadas visualmente en la serie, el componente estacional podría no tener un impacto tan dominante en la variabilidad de las medianas como lo tiene la tendencia identificada en las pruebas anteriores.

Tabla 39
Prueba de Breusch-Godfrey en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

LM test	
p-valor	2.2e-16

- H_0: No existe autocorrelación serial en los residuos (independencia).
- H_1: Existe autocorrelación serial en los residuos.

En la Tabla 39 se observa un p-valor de $2{,}2 \times 10^{-16}$, el cual se encuentra muy por debajo del nivel de significancia estandar del 5 % (0.05). Ante este resultado, se rechaza la hipótesis nula (H_0), confirmando que los residuos de la serie del número de pasajeros presentan una fuerte dependencia serial. Este hallazgo es crucial para el diagnóstico del modelo, ya que indica que la información histórica no ha sido capturada en su totalidad por la estructura actual, sugiriendo la necesidad de integrar términos autorregresivos adicionales o ajustar la especificación del modelo para corregir el sesgo en las estimaciones.

Tabla 40
División de la serie temporal en conjuntos de entrenamiento y prueba en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

Conjunto	**Cantidad (n)**	**Porcentaje (%)**
Data de entrenamiento (Training data)	$n_{train} = 210$	80 %
Data de prueba (Test data)	$n_{test} = 53$	20 %
Total	$n = 263$	100 %

La Tabla 40 cómo se dividió la serie temporal del número de pasajeros en datos de entrenamiento y prueba para validar el modelo. El conjunto de datos de entrenamiento está formado por el 80 % de las observaciones, lo que equivale a $n_{train} = 210$. Este conjunto se usa para encontrar patrones y ajustar los parámetros del modelo. Por otro lado, el conjunto de datos de prueba se compone del 20 % restante, es decir, $n_{test} = 53$. Este conjunto es crucial para evaluar qué tan bien puede predecir el modelo. En total, se utilizaron $n = 263$ registros. Esta división permite evaluar de manera justa cómo de bueno es el modelo para generalizar.

Tabla 41

Estimación de parámetros y errores estándar del modelo SARIMA(4,1,1)(0,0,2)[12] *en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete* (2004-2025).

Parámetro	Estimación	Error estándar
AR(1)	-0.8005	0.1572
AR(2)	-0.6946	0.0715
AR(3)	-0.6864	0.0978
AR(4)	-0.4093	0.0720
MA(1)	0.5449	0.1647
SMA(1)	0.2755	0.0831
SMA(2)	0.3367	0.1233

Nota. Modelo ajustado a la serie train.

En la Tabla 41 se observa los coeficientes estimados para el modelo $SARIMA(4,1,1)(0,0,2)_{12}$, seleccionado para capturar la dinámica de la serie del número de pasajeros. Los componentes autorregresivos (AR) presentan coeficientes negativos significativos, donde el $AR(1) = -0{,}8005$ sugiere una fuerte corrección de los valores inmediatos superiores a la media. El término de media móvil $MA(1) = 0{,}5449$ indica que el modelo incorpora de manera efectiva los choques aleatorios recientes para ajustar las predicciones. Asimismo, la presencia de términos estacionales de media móvil ($SMA(1)$ y $SMA(2)$) con valores positivos confirma que la serie mantiene una memoria de eventos pasados en ciclos de 12 meses. Dado que los errores estándar son relativamente bajos en comparación con las estimaciones, se valida la precisión de los parámetros para representar la estructura estocástica del flujo de pasajeros, integrando tanto la tendencia no estacionaria (mediante la diferenciación $d = 1$) como el comportamiento cíclico observado.

Tabla 42

Parámetros de Suavizamiento del Modelo Holt-Winters (*A, A, A*) *en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete* (2004-2025).

Parámetro	Valor Estimado
Alfa (Nivel)	0.4011
Beta (Tendencia)	0.0019
Gamma (Estacionalidad)	0.2817

En la Tabla 42 se observa que el modelo Holt-Winters aditivo presenta un parámetro Alfa de 0.4011, lo que indica que el nivel de la serie se actualiza considerando tanto la información reciente como el comportamiento histórico de manera equilibrada. Destaca especialmente el valor de Beta (0.0019), el cual, al ser cercano a cero, revela que la tendencia de la serie de pasajeros es sumamente estable y no presenta variaciones abruptas en su pendiente a corto plazo. Por último, el parámetro Gamma de 0.2817 sugiere una presencia moderada de estacionalidad que se ajusta de forma sistemática a los ciclos anuales. Estos valores definen la configuración interna del modelo y permiten caracterizar matemáticamente los componentes de nivel, tendencia y estacionalidad para este ajuste específico.

Tabla 43

Resultados del modelo STL basado en LOESS y parámetros de suavizamiento en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

Parámetros del modelo STL (LOESS)			
Parámetro	**Símbolo**	**Valor**	**Interpretación**
Período estacional	n_p	12	Frecuencia mensual de la serie
Ventana estacional	n_s	∞	Estacionalidad fija (periodic)
Ventana de tendencia	n_t	19	Suavizamiento de la tendencia
Grado de Loess	degree	1	Ajuste local lineal
Salto de Loess	jump	2	Optimización computacional
Componente		**Descripción**	
Estacionalidad		Alta variabilidad	
Tendencia		Creciente pronunciada	
Residuo		Alta dispersión	

En la Tabla 43 se observa las especificaciones del modelo STL basado en el suavizado local de regresión (Loess) aplicado a la serie de pasajeros. Se observa que el periodo estacional (n_p) se ha establecido en 12, lo que permite capturar la dinámica mensual de la serie, mientras que la ventana estacional (n_s) configurada como periódica (infinita) indica que el patrón estacional se mantiene constante a lo largo del tiempo. Por otro lado, la ventana de tendencia (n_t) de 19 periodos y el grado de Loess lineal (degree = 1) aseguran un suavizado de la tendencia que logra capturar el crecimiento pronunciado sin verse afectado excesivamente por el ruido. Finalmente, el modelo describe un componente residual de alta dispersión, lo que refleja la complejidad del flujo de pasajeros y establece las bases técnicas para la evaluación de este modelo frente a otras alternativas de pronóstico.

Tabla 44

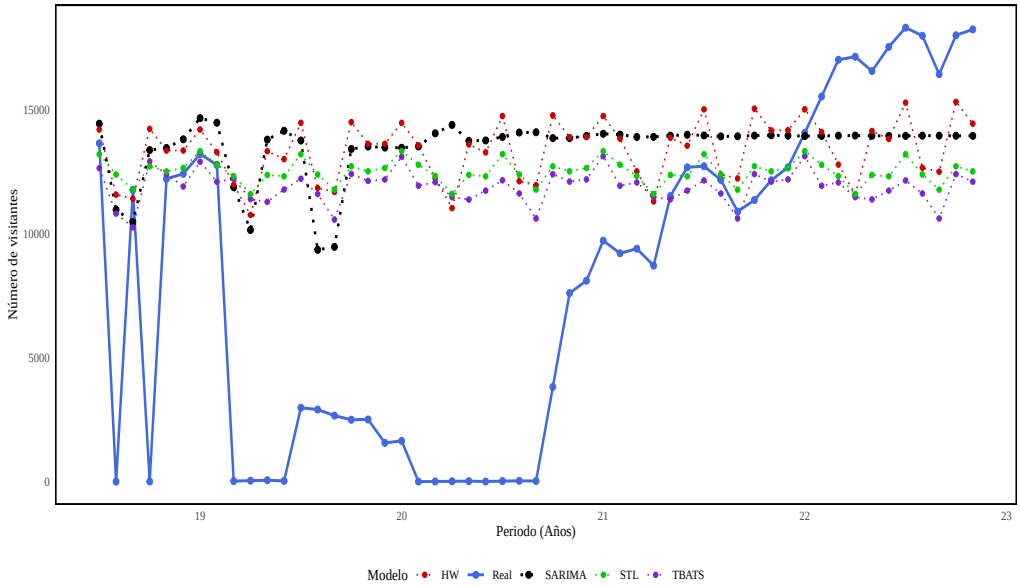
Parámetros técnicos del modelo TBATS(0.389, {2,2}, -, {<12,5>}) en el número pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

Componente	**Parámetro**	**Valor estimado**
Box-Cox	λ	0.3887
Suavizamiento	α	0.3185
Tendencia	β	No incluida
Amortiguamiento	ϕ	No incluido
AR (p=2)	ϕ_1, ϕ_2	0.0973, -0.6315
MA (q=2)	θ_1, θ_2	0.2653, 0.5561
Estacionalidad	Fourier ($K = 5$)	Período 12

En la Tabla 44 se observa los parámetros estimados del modelo TBATS, el cual integra transformaciones y componentes dinámicos para el ajuste de la serie. Se destaca el parámetro de Box-Cox ($\lambda = 0{,}3887$), que indica la aplicación de una transformación no lineal para estabilizar la varianza, junto con un parámetro de suavizamiento ($\alpha = 0{,}3185$) que pondera el nivel de la serie. El modelo excluye componentes de tendencia y amortiguamiento, optando por una estructura que combina términos autorregresivos (AR) y de media móvil (MA) de segundo orden ($\phi_1 = 0{,}0973, \phi_2 = -0{,}6315$ y $\theta_1 = 0{,}2653, \theta_2 = 0{,}5561$), lo que captura la dependencia serial interna. Finalmente, la estacionalidad se modela mediante términos de Fourier con $K = 5$ para un periodo de 12 meses, ofreciendo una representación flexible de los patrones cíclicos del flujo de pasajeros sin depender de parámetros estacionales tradicionales.

Figura 44

Comparación de modelos en el número pasajeros internacionales del Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 44 se observa el desempeño de los modelos SARIMA, Holt-Winters (HW), STL y TBATS frente a la serie real. Aunque la serie original presenta una volatilidad extrema y periodos de interrupción, el modelo TBATS (línea morada) parece mostrar, visualmente, una mayor capacidad de adaptación a las fluctuaciones de la serie en comparación con sus contrapartes. Mientras que modelos como SARIMA o HW tienden a generar un patrón más rígido o estandarizado, el ajuste de TBATS sugiere un seguimiento más flexible de la dinámica local de los datos. Esta percepción preliminar de un mejor ajuste por parte del modelo TBATS podría deberse a su capacidad intrínseca para manejar estacionalidades complejas mediante términos de Fourier y estabilizar la varianza con la transformación de Box-Cox; no obstante, esta suposición deberá ser contrastada y validada rigurosamente mediante las métricas de error correspondientes.

Tabla 45

Comparación de métricas de precisión por modelo estadístico del Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).

Modelo (test)	ME	RMSE	MAE	MAPE	MASE
SARIMA (test)	-5352.08	8282.26	6661.52	44044.84	4.8001
Holt-Winters (test)	-5206.47	8209.91	6583.62	43281.97	4.7440
STL (test)	-4338.49	7868.01	6238.68	42486.01	4.4954
TBATS (test)	-3723.64	7560.68	6111.78	39740.82	4.4040

Nota. Evaluación del desempeño basada en el conjunto de prueba (Test set).

En la Tabla 45 se observa los resultados de la evaluación de desempeño para los modelos SARIMA, Holt-Winters, STL y TBATS aplicados al conjunto de prueba. Al contrastar las métricas de error, se confirma de manera cuantitativa la suposición previa: el modelo TBATS presenta el mejor ajuste global, obteniendo los valores más bajos en todas las dimensiones analizadas. Específicamente, el TBATS registra un RMSE de 7560.68 y un MAE de 6111.78, cifras significativamente menores a las reportadas por el modelo SARIMA (RMSE: 8282.26; MAE: 6661.52). Asimismo, el error porcentual absoluto medio (MAPE) y el error escalado medio absoluto (MASE) más bajos corresponden también al modelo TBATS (39740.82 y 4.4040, respectivamente), lo que ratifica su mayor capacidad para minimizar las desviaciones frente a la serie real del número de visitantes. Estos resultados permiten concluir que la arquitectura flexible del modelo TBATS es la que mejor captura la complejidad y las fluctuaciones del flujo analizado en comparación con los métodos tradicionales.

Tabla 46

Coeficiente U de Theil de los modelos evaluados (*Test Set*) *del Aeropuerto Internacional Alejandro Velasco Astete* (2004-2025).

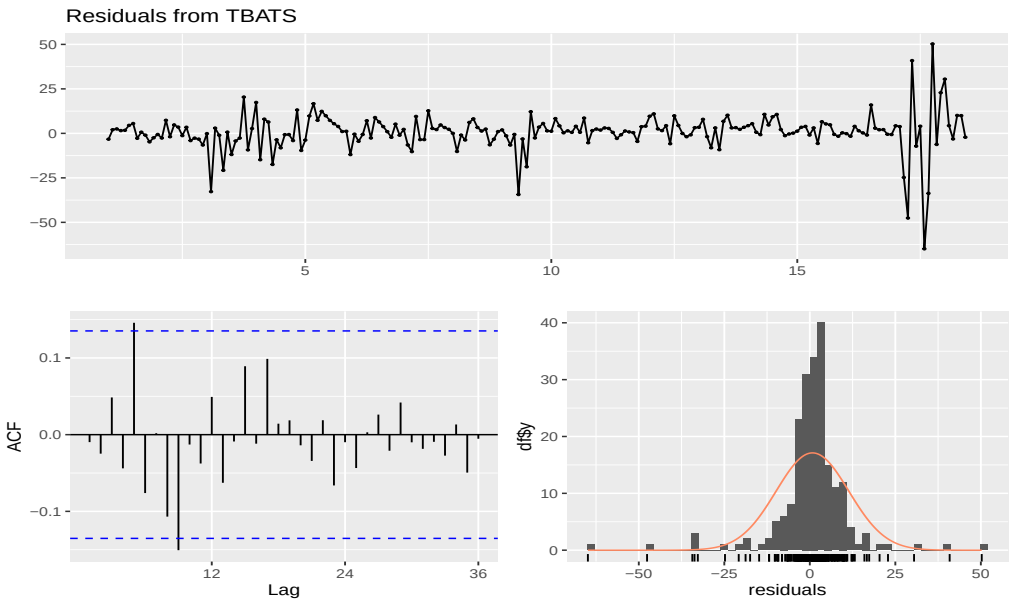
Modelo (test)	Theil's U
SARIMA (test:)	1.08125
Holt-Winters (test:)	0.98478
STL (test:)	0.95416
TBATS (test:)	0.92473

La Tabla 46 presenta el coeficiente U de Theil para los modelos evaluados en el conjunto

de prueba de la serie del número de pasajeros. Este indicador permite comparar el desempeño de los modelos frente a un pronóstico ingenuo (naive); un valor menor a 1 indica que el modelo supera al método ingenuo, mientras que un valor mayor a 1 sugiere lo contrario. En este sentido, el modelo TBATS obtuvo el coeficiente más bajo (0,92473), demostrando la mejor capacidad predictiva entre las opciones analizadas. Por el contrario, el modelo SARIMA registró un valor superior a la unidad (1,08125), indicando una menor precisión en comparación con la referencia base para esta serie de datos específica.

Figura 45

Diagnóstico estadístico de los residuos del modelo del Aeropuerto Internacional Alejandro Velasco Astete (2004-2025).



En la Figura 45 se observa el coeficiente U de Theil para cada modelo analizado en el conjunto de prueba, métrica que permite evaluar la capacidad predictiva. Los resultados muestran que el modelo TBATS obtiene el valor más bajo (0.92473), posicionándose como el modelo con mayor precisión relativa. Es importante resaltar que los modelos TBATS, STL (0.95416) y Holt-Winters (0.98478) presentan valores inferiores a la unidad, lo que indica que sus pronósticos son superiores al método de referencia. Por el contrario, el modelo SARIMA registra un coeficiente de 1.08125, sugiriendo que, para este caso particular y bajo esta métrica, su desempeño no logra superar al de un pronóstico simple. Estos hallazgos refuerzan la selección del modelo TBATS como la herramienta más eficiente entre las evaluadas para capturar la estructura de la serie y generar proyecciones del número de visitantes.

Tabla 47
Prueba de autocorrelación de los Residuos (*Ljung-Box*) *del Aeropuerto Internacional Alejandro Velasco Astete* (2004-2025).

Ljung-Box test	
P-Value	0.579

- H_0: No existe autocorrelación serial (son ruido blanco).
- H_1: Existe autocorrelación serial (no son ruido blanco).

En la Tabla 47 se observa el p-valor es significativamente menor al nivel de significancia estandar de 0.05, se rechaza la hipótesis nula (H_0). Por lo tanto, se acepta la hipótesis alternativa (H_1), confirmando la presencia de autocorrelación serial en los residuos. Aunque los residuos no sean ruido blanco, esto es consecuencia de la alta volatilidad y los valores atípicos de la serie original. No obstante, al obtener los menores errores de pronóstico y un coeficiente U de Theil menor a 1, el modelo TBATS se confirma como el más robusto y con mejor capacidad predictiva de todos los evaluados.

5.3 Discusión de resultados

1. Con respecto a la imputación de datos ausentes y la identificación de cambios repentinos, los resultados muestran que la primera parte de la metodología MAQY-TS facilitó la reconstrucción de la serie temporal de forma coherente, asegurando su continuidad estructural. La estimación de valores ausentes a través de modelos de predicción, en este caso Prophet, permitió calcular valores perdidos sin modificar de manera significativa la dinámica de la serie. De igual manera, la inclusión del criterio IFI para la identificación de anomalías facilitó reconocer alteraciones drásticas justo después de los períodos con datos ausentes. Estos resultados coinciden con lo mencionado porToller et al. (2025), quien indica que la fusión de métodos de imputación con intervalos de confianza potencia la exactitud en la identificación de valores atípicos, en series con intervalos de datos incompletos.

2. En el modelado, la elección fundamentada en anomalías y el cumplimiento de supuestos facilitó la correcta selección entre modelos tradicionales (ARMA, ARIMA, SARIMA) y robustos (Holt-Winters, STL(Loess), BATS, TBATS). Este enfoque coincide con lo planteado por Dosantos (2020), realizo una comparación de modelos donde resulto tener mejores métricas el modelo TBATS,Paredes (2020), nos indica que el modelo predictivo mas estable es el modelo Holt y el modelo TBATS, Igualmente,Moro (2023) concluyo que el modelo predictivo mas edecuado en caso de la prediccion el el modelo TBATS, Moro (2023), concluye en su estudio recomendando usar el modelo TBATS, Xian et al. (2023), hace una comparación de modelos donde concluyo que el modelo Holt-Winters es el adecuado, Alcántara et al. (2023), identifica el modelo más preciso, donde llega a la conclusión que el modelo adecuado es el Holt-Winters, Tudela et al. (2022), donde indico que el modelo SARIMA es el adecuado, (Sernaque et al., 2022), llego a la conclusión de recomendar el modelo Holt-Winters. En este contexto, la segunda parte de la metodología MAQY-TS presenta un criterio de decisión más completo, superando métodos tradicionales que se basan únicamente en pruebas estadísticas, y facilitando un modelado más flexible y exacto.

3. Respecto al tercer objetivo, la implementación de la metodología en un conjunto real de datos demuestra que la falta de información y la existencia de cambios estructurales son situaciones comunes, lo que restringe la aplicación directa de métodos convencionales. En este marco, la combinación de imputación, identificación de anomalías y modelado facilita enfrentar estas limitaciones en conjunto, aunque también genera una interdependencia entre las fases, donde fallos en la imputación pueden extenderse al modelado

4. Finalmente, a nivel del objetivo general, los resultados sugieren que la metodología MAQY-TS amplía el enfoque de tradicional al incorporar una fase previa de reconstrucción y validación de la serie. No obstante, su efectividad depende de la correcta especificación del modelo de imputación y de los criterios de detección utilizados, lo que abre la posibilidad de explorar mejoras en futuros estudios.

CONCLUSIONES

Según los resultados obtenidos mediante el trabajo de investigación se obtienen las siguientes conclusiones:

1. Se determina que la fase inicial de la metodología MAQY-TS, relacionada con la imputación y la identificación de anomalías, facilitó la reconstrucción de la serie temporal de forma coherente a través del uso de Prophet, asegurando así su continuidad. De igual manera, la utilización del criterio IFI facilitó la identificación de alteraciones repentinas después de los períodos con datos ausentes, validando la consistencia de los valores imputados y mejorando la calidad de la serie restaurada.

2. Se determina que la segunda etapa de la metodología MAQY-TS, enfocada en el modelado, facilitó la selección adecuada de los modelos de acuerdo a la existencia de anomalías y el respeto de los supuestos estadísticos. En este aspecto, se lograron implementar modelos clásicos cuando la serie satisfacía condiciones óptimas, y modelos robustos en situaciones con irregularidades, logrando un ajuste acorde a la dinámica de los datos.

3. Se determina que la implementación completa de la metodología MAQY-TS en series temporales reales facilitó el manejo efectivo de problemas de datos ausentes y cambios estructurales, alcanzando modelos con buena capacidad predictiva y residuos que satisfacen el supuesto de ruido blanco.

4. Se determina que la metodología MAQY-TS, al combinar sus dos etapas imputación y detección, y modelamiento, optimiza el análisis de series temporales en comparación con metodologías tradicionales, al añadir una fase anterior de reconstrucción y validación que fortalece la solidez y fiabilidad del modelo final.

RECOMENDACIONES

1. Es aconsejable evaluar el rendimiento del modelo de pronostico Prophet en comparación con otros modelos de pronostico en la imputación en un intervalo de tiempo con datos faltantes en series temporales, como interpolación spline, modelos con imputación interna o enfoques de aprendizaje automático. Esto permitiría determinar si Prophet es siempre la mejor opción o si hay métodos más exactos según el tipo de serie.

2. Se recomienda comparar el criterio (IFI) con otros enfoques de identificación de anomalías, como técnicas basadas en residuos, modelos para la detección de outliers o métodos robustos. Esta comparación permitiría verificar la efectividad del IFI y, eventualmente, sugerir una versión optimizada o híbrida

3. Se propone investigar la incorporación de modelos híbridos durante la etapa de modelado, fusionando métodos tradicionales y sólidos, para optimizar la capacidad de predicción en series que presentan comportamientos complejos.

4. Se sugiere implementar la metodología MAQY-TS en diversas clases de series temporales (financieras, climáticas, sociales), con el fin de analizar su generalización y solidez en diferentes contextos.

5. Se sugiere evaluar cómo los valores imputados afectan los resultados del modelado final, a través de comparaciones entre series originales completas (cuando estén accesibles) y series imputadas.

6. Es aconsejable que las próximas investigaciones se aborden no solo la imputación de valores faltantes, sino también el manejo de datos duplicados, inconsistencias y anomalías, así como otros problemas relacionados con la calidad de los datos. Además, se sugiere comparar diversas técnicas de limpieza e imputación para determinar cómo afectan la fiabilidad y exactitud de los modelos estadísticos y de series temporales.

BIBLIOGRAFÍA

Ahn, H., Sun, K., & Kim, K. P. (2022). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials, & Continua*, 70(1), 767.

Alcántara, O. R., Sánchez-Ticona, R. J., Tenorio-Cabrera, J. L., Torres-Villanueva, M., & Santos-Fernández, J. P. (2023). Modelos de series temporales predictivos para la demanda turística en el Perú.

Aminikhanghahi, S. & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2), 339–367.

Andiojaya, A. & Demirhan, H. (2019). A bagging algorithm for the imputation of missing values in time series. *Expert Systems with Applications*, 129, 10–26.

Bai, J. & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, (pp. 47–78).

Borja, J. G. & Sánchez, F. H. N. (2008). Distribución de la estadística de jarque y bera para la prueba de normalidad en una serie temporal estacionaria con datos faltantes. *Entre ciencia e ingeniería*, (4), 99–114.

Boulton, C. A. & Lenton, T. M. (2019). A new method for detecting abrupt shifts in time series.

Box, G. & Jenkins, G. (1976). Analysis: Forecasting and control. *San francisco*, 10.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., Terpenning, I., et al. (1990). Stl: A seasonal-trend decomposition. *J. off. Stat*, 6(1), 3–73.

Corradin, R., Danese, L., & Ongaro, A. (2022). Bayesian nonparametric change point detection for multivariate time series with missing observations. *International Journal of Approximate Reasoning*, 143, 26–43.

Das, P. & Barman, S. (2025). Perspective chapter: An overview of time series decomposition and its applications. *Applied and Theoretical Econometrics and Financial Crises*, (pp.5̃9).

Despot, I. & Arif, A. (2023). What is time series forecasting? Accessed: 2025-10-12.

Dosantos, P. S. (2020). Comparación de modelos de predicción para series temporales. Tesis de maestría, Universidad de Oviedo.

Goicoechea, A. P. (2002). Imputación basada en árboles de clasificación. *Eustat. Available in: http://www. eustat. es/documentos/datos/ct*, 4.

Haile, T. T., Tian, F., AlNemer, G., & Tian, B. (2024). Multiscale change point detection for univariate time series data with missing value. *Mathematics*, 12(20), 3189.

Hernández, P. U. (2024). Análisis comparativo entre métodos clásicos de forecasting y modelos de aprendizaje automático.

Hernández-Sampieri, R. & Mendoza, C. (2018). *Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta*. McGraw Hill México.

Hyndman, R. J. & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Kuhn, T. S. (1997). *The structure of scientific revolutions*, volume 962. University of Chicago press Chicago.

Little, R. J. & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Mauricio, J. A. (2007). Análisis de series temporales. *Universidad Complutence de Madrid*.

Meléndez Surmay, R., Giraldo Henao, R., & Rodríguez Cortes, F. (2024). Kruskal-wallis test for functional data based on random projections generated from a simulation of a brownian motion. *TecnoLógicas*, 27(59).

Moro, L. (2023). Análisis y predicción de series temporales para consumo eléctrico y presión en el supercomputador finis terrae ii mediante el uso de spark.

Oliva, B. (2019). Análisis de series de tiempo.

Paredes, M. Á. (2020). Predicción de múltiples series de tiempo univariadas a través de diversos modelos predictivos y meta-learning aplicado en la industria del retail. Tesis de Magíster, Universidad de Chile.

Patiño, C. I. & Alonso, J. C. (2005). Evaluación de pronósticos para la tasa de cambio en colombia. *Estudios Gerenciales*, 21(96), 13–29.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Ribeiro, S. M. & Castro, C. L. (2022). Missing data in time series: A review of imputation methods and case study. *Learning and Nonlinear Models*, 20(1), 31–46.

Rink, K. (2021). Time series forecast error metrics you should know. *Towards Datascience*.

Ríos, G. (2008). Series de tiempo. universidad de chile. facultad de ciencias fisicas y matematicas.: 52.

Santra, R. (2023). Tests for stationarity in time series—dickey fuller test & augmented dickey fuller (adf) test. *Medium, USA*.

Sanz, V. (2023). *Metodologías y herramientas para la planificación y gestión integrada de los recursos pesqueros en el estrecho de Gibraltar en un contexto de cambio climático*. PhD thesis, Universidad de Huelva.

Sernaque, Saul, H., Dayana Meca Barreto, M., Daniel Zapata Tavara, E., Nicol Marchan Domador, B., Eduardo Medina Peña, J., Alexis Nole Villalobos, D., Nicolas Aldana Yarleque, C., Saavedra Navarro, Y., Ramón Trelles Pozo, L., Chuquihuanca Yacsahuanca, N., & Mendoza, G. (2022). Comparison of arima and holt-winters forecasting models for time series of cereal production in peru. In *Proceedings of the 5th International Conference on Intelligent Human Systems Integration (IHSI 2022) Integrating People and Intelligent Systems, February 22–24, 2022, Venice, Italy*, IHSI 2022: AHFE International.

Suárez, P. et al. (2022). Comparación de métodos de predicción para series temporales. Master's thesis.

Taylor, S. J. & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37–45.

Toller, M. B., Geiger, B. C., & Kern, R. (2025). Detecting abrupt changes in missing time series data. *Information Sciences*, (pp. 122322).

Tudela, J. W., Cahui-Cahui, E., & Aliaga-Melo, G. (2022). Impacto del covid-19 en la demanda de turismo internacional del perú. una aplicación de la metodología box-jenkins. *Revista de Investigaciones Altoandinas*, 24(1), 27–36.

Van Buuren, S. & Van Buuren, S. (2012). *Flexible imputation of missing data*, volume 10. CRC press Boca Raton, FL.

Villavicencio, J. (2010). Introducción a series de tiempo. *Puerto Rico*.

Xian, X., Wang, L., Wu, X., Tang, X., Zhai, X., Yu, R., Qu, L., & Ye, M. (2023). Comparison of sarima model, holt-winters model and ets model in predicting the incidence of foodborne disease. *BMC Infectious Diseases*, 23(1), 803.

Zaiontz, C. (2012). Mann-kendall test. *Real Statistics Using Excel.[Online 25.11. 2022] Available: https://www. real-statistics. com/time-series-analysis/time-series-miscellaneous/mann-kendalltest*.

ANEXOS

A. Matriz de consistencia

Problemas	Objetivos	Hipótesis	Variables	Metodología
General			**Dependientes**	◦ **Tipo:** Aplicada. ◦ **Nivel:** Explicativa. ◦ **Población:** no se define una población específica, dado que el propósito es desarrollar y validar una técnica de imputación
¿Cómo imputar por intervalos basada en pronósticos y detectar cambios abruptos para el modelamiento de una serie temporal faltante y su aplicación?.	Analizar la imputación por intervalos basada en pronósticos y detectar cambios abruptos para el modelamiento de una serie temporal faltante y su aplicación.	Existe un método para Imputar por intervalos basada en pronósticos y detectar cambios abruptos que mejora en la precisión en el modelamiento de una serie temporal faltante y su aplicación.	◦ Número de visitantes a la ciudad inka de machu picchu (2002-2025). ◦ Número de pasajeros internacionales en el Aeropuerto Internacional Alejandro Velasco Astete (2004–2025)	
Específicos			**Independientes**	
◦ ¿Cómo imputar datos faltantes de una serie temporal en un intervalo de tiempo y detectar cambios abruptos?. ◦ ¿Qué modelo tiene mejor en ajuste y pronóstico para una serie temporal faltante imputada con presencia de cambios abruptos?. ◦ ¿Existen series temporales basadas en situaciones reales con las características estudiadas para aplicar el mejor modelo seleccionado?.	◦ Imputar datos faltantes de una serie temporal en un intervalo de tiempo y detectar cambios abruptos. ◦ Evaluar el modelo que tiene mejor en ajuste y pronóstico para una serie temporal faltante imputada con presencia de cambios abruptos. ◦ Identificar series temporales basadas en situaciones reales con las características estudiadas para aplicar el mejor modelo seleccionado.	◦ La imputación mediante Interval Forescat Imputation (IFI) es un método eficaz para la imputación de datos faltantes en un intervalo de tiempo y detectar cambios abruptos con prophet en una serie temporal. ◦ El modelo TBATS es el modelo que tiene mejor en ajuste y pronóstico para una serie temporal faltante imputada con presencia de cambios abruptos. ◦ Existen series temporales basadas en situaciones reales con las características estudiadas, como: llegadas de visitantes a la Ciudad Inka de Machu Picchu (2002 – 2025) y Movimiento Internacional de pasajeros en el aeropuerto Internacional Alejandro Velasco Astete (2004 – 2025).	◦ Tiempo (2002-2025). ◦ Tiempo (2004-2025).	

B. Datos

Datos(muestra del total de datos) número de llegadas de visitantes a la Ciudad Inka de Machu Picchu (2002 – 2025)

Año	Mes	Doméstico	Internacional	Total
2002	Enero	2480	17786	20266
2002	Febrero	2017	16824	18841
2002	Marzo	2465	19200	21665
2002	Abril	2192	17003	19195
2002	Mayo	3085	19370	22455
2002	Junio	3233	17828	21061
2002	Julio	5423	23126	28549
2002	Agosto	9351	28451	37802
2002	Septiembre	6782	18799	25581
2002	Octubre	16676	22339	39015
2002	Noviembre	21236	20787	42023
2002	Diciembre	11680	14362	26042
2003	Enero	4945	17197	22142
2003	Febrero	2659	15742	18401
2003	Marzo	3207	18416	21623
2003	Abril	3746	19824	23570
2003	Mayo	4957	17900	22857
2003	Junio	4883	19150	24033
2003	Julio	8553	26758	35311
2003	Agosto	9728	31035	40763
2003	Septiembre	7543	22401	29944
2003	Octubre	21988	24909	46897
2003	Noviembre	26518	24857	51375
2003	Diciembre	18384	17684	36068

Datos(muestra del total de datos) número Internacional de pasajeros en el aeropuerto Internacional Alejandro Velasco Astete (2004 – 2025).

Año	Mes	Doméstico	Internacional	Total
2004	Enero	55043	1005	56048
2004	Febrero	55340	962	56302
2004	Marzo	61227	1174	62401
2004	Abril	65470	1093	66563
2004	Mayo	67837	1040	68877
2004	Junio	67938	1356	69294
2004	Julio	78965	1869	80834
2004	Agosto	87245	1244	88489
2004	Septiembre	71026	912	71938
2004	Octubre	77886	1255	79141
2004	Noviembre	64497	1023	65520
2004	Diciembre	59178	980	60158
2005	Enero	63085	1237	64322
2005	Febrero	54182	901	55083
2005	Marzo	72508	1587	74095
2005	Abril	63959	1086	65045
2005	Mayo	73425	1323	74748
2005	Junio	74389	1521	75910
2005	Julio	94958	1444	96402
2005	Agosto	96554	1531	98085
2005	Septiembre	77774	840	78614
2005	Octubre	83030	1147	84177
2005	Noviembre	69166	992	70158
2005	Diciembre	58506	750	59256
2006	Enero	62133	1142	63275

C. Codigo del procesamiento en R (muestra)

```r
# ============================================================
# FUNCIÓN PARA DETECTAR TODOS LOS GAPS
# ============================================================

find_all_gaps <- function(x){

  na_flag <- is.na(x)        # TRUE donde hay NA
  r <- rle(na_flag)          # Agrupa secuencias

  ends <- cumsum(r$lengths)
  starts <- ends - r$lengths + 1

  gaps <- list()
  id <- 1

  for(i in seq_along(r$values)){
    if(r$values[i] == TRUE){
      gaps[[id]] <- c(starts[i], ends[i])
      id <- id + 1
    }
  }

  return(gaps)
}

gaps <- find_all_gaps(df$y)

# tramos con NA
#gaps

# ============================================================
# TABLA DE GAPS
# ============================================================

tabla_gaps <- data.frame(
  Gap_ID = seq_along(gaps),
  Inicio_idx = sapply(gaps, function(g) g[1]),
  Fin_idx    = sapply(gaps, function(g) g[2]),
  Longitud   = sapply(gaps, function(g) g[2] - g[1] + 1),
```

```r
  Longitud   = sapply(gaps, function(g) g[2] - g[1] + 1),

  Fecha_inicio = sapply(gaps, function(g) df$ds[g[1]]),
  Fecha_fin    = sapply(gaps, function(g) df$ds[g[2]])
)

print(tabla_gaps)


# ============================================================
#  FUNCIÓN PROPHET PARA IFI
# ============================================================

forecastProphet_IFI <- function(x_before, h, certainty=0.95){

  df_temp <- data.frame(
    ds = seq.Date(as.Date("2000-01-01"),
                  by="month",
                  length.out = length(x_before)),
    y = x_before
  )

  m <- prophet(df_temp,
               yearly.seasonality = TRUE,
               weekly.seasonality = FALSE,
               daily.seasonality = FALSE,
               interval.width = certainty)

  future <- make_future_dataframe(m, periods = h, freq="month")
  forecast <- predict(m, future)

  return(forecast)
}

#Genera predicciones + intervalos
# ============================================================
# GRAFICO PROPHET (PREDICCIÓN + INTERVALOS)
# ============================================================
```

```r
# Tomar datos antes del primer gap (ejemplo)
gap_ejemplo <- gaps[[1]]

x_before <- df$y[1:(gap_ejemplo[1]-1)]

# Crear dataframe como en la función
df_hist <- data.frame(
  ds = seq.Date(as.Date("2000-01-01"),
                by = "month",
                length.out = length(x_before)),
  y = x_before
)

# Asegurar formato Date
df_hist$ds <- as.Date(df_hist$ds)

# Horizonte
h <- gap_ejemplo[2] - gap_ejemplo[1] + 2

# Usar tu función
forecast <- forecastProphet_IFI(x_before, h)

# Crear dataframe forecast limpio
df_forecast <- data.frame(
  ds = as.Date(forecast$ds),
  yhat = forecast$yhat,
  lower = forecast$yhat_lower,
  upper = forecast$yhat_upper
)

# ============================================================
# GRAFICO
# ============================================================

p_prophet <- ggplot() +
  geom_line(data = df_hist, aes(x = ds, y = y, color = "Histórico"), linewidth = 0.8) +
  geom_point(data = df_hist, aes(x = ds, y = y, color = "Histórico"), size = 1.2) +

  geom_line(data = df_forecast, aes(x = ds, y = yhat, color = "Pronóstico"), linewidth = 1) +
```

```r
  geom_line(data = df_forecast, aes(x = ds, y = yhat, color = "Pronóstico"), linewidth = 1) +

  geom_ribbon(data = df_forecast,
              aes(x = ds, ymin = lower, ymax = upper, fill = "Intervalo de Confianza"),
              alpha = 0.3) +

  scale_color_manual(values = c("Histórico" = "royalblue", "Pronóstico" = "red")) +
  scale_fill_manual(values = c("Intervalo de Confianza" = "orange")) +

  labs(title = "",
       x = "Tiempo", y = "Numero") +
  # Modelo Prophet: Predicción e Intervalos
  theme_minimal() +
  theme(
    panel.grid = element_blank(),
    panel.border = element_rect(color = "black", fill = NA),
    legend.position = "bottom"
  )

p_prophet

#ggsave("p_prophet.pdf", p_prophet, width = 12, height = 6)

# TABLA

tabla_forecast <- data.frame(
  Fecha = forecast$ds,
  Prediccion = round(forecast$yhat, 2),
  Intervalo_Inferior = round(forecast$yhat_lower, 2),
  Intervalo_Superior = round(forecast$yhat_upper, 2)
)

print(tabla_forecast)
```

```r
# FUNCIÓN IFI (RESULTADO TRUE/FALSE AQUÍ)
# ============================================================

ifi_prophet_gap <- function(x, gap, certainty=0.95){

  x_before <- x[1:(gap[1]-1)]
  x_real <- x[gap[2]+1]

  if(is.na(x_real)){
    return(NA)
  }

  h <- gap[2] - gap[1] + 2

  forecast <- forecastProphet_IFI(x_before, h, certainty)

  lower <- tail(forecast$yhat_lower, h)[h]
  upper <- tail(forecast$yhat_upper, h)[h]

  resultado <- (x_real < lower) | (x_real > upper)

  if(resultado){
    cat(" Resultado IFI: TRUE → anomalía\n")
  } else {
    cat(" Resultado IFI: FALSE → normal\n")
  }

  return(resultado)
}

# IFI evalúa cambio abrupto después del gap

# ============================================================
#  IMPUTACIÓN COMPLETA + IFI
# ============================================================

impute_prophet_full_with_IFI <- function(df, certainty=0.95){

  df_result <- df

  df_result$lower <- NA
```

```r
  df_result$upper <- NA

  gaps <- find_all_gaps(df$y)
  ifi_results <- rep(NA, length(gaps))

  for(i in seq_along(gaps)){

    gap <- gaps[[i]]
    start <- gap[1]
    end   <- gap[2]

    cat("\n🔧 Imputando gap:", start, "-", end, "\n")

    x_before <- df_result[1:(start-1), c("ds","y")]
    x_before <- x_before[!is.na(x_before$y), ]

    h <- end - start + 1

    m <- prophet(x_before,
                 yearly.seasonality = TRUE,
                 weekly.seasonality = FALSE,
                 daily.seasonality = FALSE,
                 interval.width = certainty)

    future <- make_future_dataframe(m, periods = h+1, freq="month")
    forecast <- predict(m, future)

    yhat  <- tail(forecast$yhat, h+1)
    lower <- tail(forecast$yhat_lower, h+1)
    upper <- tail(forecast$yhat_upper, h+1)

    # IMPUTACIÓN
    df_result$y[start:end] <- yhat[1:h]

    # INTERVALOS SOLO EN GAPS
    df_result$lower[start:end] <- lower[1:h]
    df_result$upper[start:end] <- upper[1:h]

    # IFI (AQUÍ SE GENERA TRUE/FALSE)
    ifi_results[i] <- ifi_prophet_gap(df$y, gap, certainty)
```

```
    # IFI (AQUÍ SE GENERA TRUE/FALSE)
    ifi_results[i] <- ifi_prophet_gap(df$y, gap, certainty)
  }

  return(list(data=df_result, ifi=ifi_results, gaps=gaps))
}

# ============================================================
# GRAFICO IFI (EVALUACIÓN DEL CAMBIO)
# ============================================================
result <- impute_prophet_full_with_IFI(df)

df_final <- result$data
ifi_flags <- result$ifi
gaps <- result$gaps

# ============================================================
# GRAFICO IFI (VALIDACIÓN DEL CAMBIO)
# ============================================================

df_ifi_plot <- data.frame(
  Gap_ID = seq_along(gaps),
  Valor_Real = sapply(gaps, function(g) df$y[g[2]+1]),
  Lower = sapply(seq_along(gaps), function(i){
    g <- gaps[[i]]
    x_before <- df$y[1:(g[1]-1)]
    h <- g[2] - g[1] + 2
    forecast <- forecastProphet_IFI(x_before, h)
    tail(forecast$yhat_lower, h)[h]
  }),
  Upper = sapply(seq_along(gaps), function(i){
    g <- gaps[[i]]
    x_before <- df$y[1:(g[1]-1)]
    h <- g[2] - g[1] + 2
    forecast <- forecastProphet_IFI(x_before, h)
    tail(forecast$yhat_upper, h)[h]
  }),
  IFI = ifi_flags
)
```

```
# gráfico
p_IFI <- ggplot(df_ifi_plot, aes(x = Gap_ID)) +

  # Intervalo
  geom_errorbar(aes(ymin = Lower, ymax = Upper, color = "Intervalo Esperado"),
                width = 0.2, linewidth = 1.2) +

  # Valor real (azul)
  geom_point(aes(y = Valor_Real, color = "Valor Real"), size = 3) +

  # Anomalía (estrella
  geom_point(data = subset(df_ifi_plot, IFI == TRUE),
             aes(y = Valor_Real, color = "Anomalía"),
             shape = 8, size = 5) +

  # Colores
  scale_color_manual(
    name = NULL,    #  elimina "colour"
    values = c(
      "Valor Real" = "blue",
      "Intervalo Esperado" = "orange",
      "Anomalía" = "red"
    )
  ) +

  labs(
    title = "",# Detección de Cambios Abruptos (IFI)
    x = "Gap ID",
    y = "Valor"
  ) +

  theme_minimal() +
  theme(
    legend.position = "bottom",    #  leyenda abajo
    panel.grid = element_blank(),
    panel.border = element_rect(color = "black", fill = NA)
  )

p_IFI
```