

## PREDIKSI PENUMPANG LRT JAKARTA MENGGUNAKAN SARIMAX DAN XGBOOST DENGAN EFEK KALENDER

Muhammad Hafiz Fazli<sup>1</sup>, M.Taqy Abiyu Dzakwan<sup>2</sup>, Nada Ardelia<sup>3</sup>, Gemala Aleida Fitri<sup>4</sup>, Akbar Rizki<sup>5\*</sup>, Windi Pangesti<sup>6</sup>

Study Program in Statistics and Data Science - School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia

\*e-mail: [akbar.ritzki@apps.ipb.ac.id](mailto:akbar.ritzki@apps.ipb.ac.id)

DOI: 10.14710/j.gauss.15.1.188-199

### Article Info:

Received: 2025-12-23

Accepted: 2026-05-25

Available Online: 2026-05-30

### Keywords:

LRT Jakarta; passenger demand; SARIMAX; time series forecasting; XGBoost.

**Abstract:** The daily passenger volume of Jakarta's LRT fluctuates significantly due to weekly seasonality and calendar variations, making accurate forecasting important for operational planning and decision-making. This study aims to determine the most effective model for forecasting daily passenger demand by comparing the SARIMAX and XGBoost methods on transportation data characterized by strong seasonal patterns and external influences. SARIMAX was selected because it models seasonal and autoregressive structures alongside exogenous variables, while XGBoost captures nonlinear relationships between temporal features and external factors. The dataset covers the period from 1 January 2024 to 31 August 2025 and includes variables such as weekends, national holidays, and special events. Model evaluation was conducted using walk-forward cross-validation and hyperparameter tuning. The results show that the SARIMAX(1,0,1)(0,1,1)<sub>7</sub> model achieved the best performance, with a validation MAPE of 11.26% and a test MAPE of 8.64%, outperforming XGBoost. SARIMAX also reproduced weekly fluctuation patterns more consistently, indicating that it is more suitable for forecasting transportation demand with strong seasonal characteristics and relatively stable external influences.

## 1. PENDAHULUAN

Jakarta menghadapi tantangan mobilitas yang semakin kompleks akibat pertumbuhan penduduk dan tingginya aktivitas sosial-ekonomi. Dengan kepadatan 16.165 jiwa/km<sup>2</sup> pada 2024, Jakarta menjadi provinsi terpadat di Indonesia (Badan Pusat Statistik Provinsi DKI Jakarta, 2025). Menurut *Economic and Social Commission for Asia and the Pacific* (United Nations, 2019), kawasan Jabodetabek mencapai 49,5 juta perjalanan per hari. Didukung pula dengan laporan dari *Institute for Transportation Development Policy* (ITDP Indonesia, 2021) yang mencatat bahwa 74% perjalanan masih didominasi kendaraan pribadi sehingga menyebabkan kemacetan dan tingginya emisi karbon. Dalam mengatasi masalah tersebut, pemerintah mengembangkan transportasi publik berbasis rel seperti *Light Rail Transit* (LRT) yang dinilai lebih efisien, tepat waktu, dan ramah lingkungan. (BPS Provinsi DKI Jakarta, 2024) menunjukkan peningkatan jumlah pengguna LRT Jakarta hingga 102.707 penumpang pada Juni 2024 dan peningkatan lebih dari 10 juta penumpang LRT Jabodebek hanya dalam satu tahun (The Jakarta Post, 2025). Meski demikian, jumlah penumpang bersifat fluktuatif dan dipengaruhi oleh faktor eksternal seperti akhir pekan, libur nasional, serta *event* tertentu. Fluktuasi ini menjadikan peramalan jumlah penumpang harian penting untuk dianalisis sehingga dapat mengoptimalkan jadwal, alokasi armada, dan efisiensi operasional.

Berdasarkan penelitian yang dilakukan oleh Lee dan Bang (2024), model statistik seperti SARIMAX efektif digunakan untuk data deret waktu yang memiliki pola musiman dan dipengaruhi variabel eksternal karena mampu mengakomodasi komponen *autoregressive*, *moving average*, dan variabel eksogen secara simultan. Model ini dinilai sesuai untuk peramalan jumlah penumpang harian LRT Jakarta yang dipengaruhi pola mingguan, hari libur nasional, dan *event* tertentu. Namun, SARIMAX memiliki keterbatasan dalam menangkap hubungan non-linier. Di sisi lain, Talusan et al. (2022) menunjukkan bahwa XGBoost mampu menangani pola yang lebih kompleks dan sering menghasilkan performa prediksi yang lebih baik pada data transportasi. Meskipun demikian, penelitian terkait peramalan jumlah penumpang transportasi publik di Indonesia masih didominasi oleh penggunaan model statistik atau *machine learning* secara terpisah sehingga efektivitas masing-masing pendekatan pada data penumpang harian LRT masih belum banyak dikaji.

Penelitian Saputra et al. (2024), sebagai contoh, hanya menggunakan metode SARIMA untuk meramalkan jumlah penumpang LRT Palembang dengan data bulanan tanpa mempertimbangkan variabel eksternal maupun perbandingan dengan metode *machine learning*. Oleh karena itu, penelitian ini dilakukan untuk mengisi keterbatasan penelitian sebelumnya dengan membandingkan pendekatan statistik dan *machine learning* pada peramalan jumlah penumpang harian LRT Jakarta. Perbandingan tersebut penting dilakukan karena data transportasi publik memiliki karakteristik musiman, dipengaruhi faktor eksternal, serta berpotensi membentuk pola non-linier sehingga diperlukan evaluasi komparatif untuk memperoleh model prediksi yang lebih efektif dalam mendukung pengelolaan operasional transportasi publik.

## 2. TINJAUAN PUSTAKA

Model *SARIMAX* merupakan perluasan dari model *SARIMA* dimana pada model ini ditambahkan variabel eksogen ( $X$ ) (Neydi et al., 2025). Pada pemodelan ini faktor-faktor yang mempengaruhi peubah *dependent*  $Y$  pada waktu ke- $t$  tidak hanya oleh peubah  $Y$  dalam waktu, tetapi juga peubah-peubah *independent* lainnya dalam waktu ke- $t$ . Bentuk umum untuk pemodelan *SARIMAX* ditunjukkan pada Persamaan (1) dengan perhitungan residual pada Persamaan (2).

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \dots + \beta_k X_{k,t} + \omega_t \quad (1)$$

$$\omega_t = \frac{\theta_q(B)\theta_Q(B)^s}{\phi_p(B)\phi_P(B^s)(1-B)^d(1-B^s)^D} a_t \quad (2)$$

Komponen pada persamaan model di atas dapat dijelaskan sebagai berikut,  $Y_t$  merepresentasikan nilai pengamatan aktual dari peubah dependen yang dicatat pada waktu ke- $t$ , sementara dinamika fluktuasi acak jangka pendeknya ditangkap oleh  $\omega_t$  yang bertindak sebagai komponen *error term* atau residual model. Proses pergeseran waktu pada data runtun waktu diakomodasi secara matematis oleh operator *backshift* ( $B$ ). Selanjutnya, model ini juga melibatkan komponen galat acak atau *white noise* yang disimbolkan sebagai  $a_t$  untuk menunjukkan nilai konstanta kesalahan acak pada saat  $t$ . Seluruh kerangka persamaan tersebut kemudian disempurnakan oleh kehadiran parameter  $\mu'$  yang berfungsi sebagai konstanta penyeimbang atau intersep dasar di dalam model.

Pada Persamaan (2), parameter yang dipakai dalam model dapat dijabarkan secara rinci menggunakan serangkaian operator polinomial aljabar pembentuk sisaan yang didefinisikan sebagai berikut:

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \quad (3)$$

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q \quad (4)$$

$$\Phi_p(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_p B^{ps} \quad (5)$$

$$\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \dots - \Theta_Q B^{Qs} \quad (6)$$

Estimasi parameter dalam model ini mencakup koefisien struktural yang merepresentasikan perilaku historis data runtun waktu secara non-musiman maupun musiman. Pola ketergantungan terhadap nilai masa lalu diwakili oleh parameter  $\phi_p$  sebagai komponen *Seasonal Autoregressive* (SAR) berordo P dan parameter  $\phi_p$  yang mencerminkan komponen *Autoregressive* (AR) non-musiman berordo p. Sementara itu, pengaruh dari kejut galat atau sisaan masa lalu diakomodasi secara simultan oleh parameter  $\theta_q$  yang bertindak sebagai komponen *Seasonal Moving Average* (SMA) berordo Q, serta parameter  $\theta_q$  yang merepresentasikan komponen *Moving Average* (MA) non-musiman berordo q.

Extreme Gradient Boosting (XGBoost) merupakan algoritma *gradient boosting* yang membangun model secara aditif melalui serangkaian *decision tree* untuk meminimalkan *loss function* (Chen and Guestrin, 2016). Algoritma ini unggul karena memanfaatkan informasi second-order, regularisasi struktural, serta teknik *shrinkage* dan *subsampling* yang meningkatkan stabilitas model pada data yang fluktuatif dan non-linier, termasuk deret waktu penumpang harian. Di Indonesia, XGBoost telah digunakan untuk peramalan penumpang MRT Jakarta dan menunjukkan akurasi yang konsisten tinggi dibandingkan metode klasik (Harriz *et al.*, 2023). Model XGBoost membangun prediksi sebagai penjumlahan *tree* yang ditunjukkan pada Persamaan (7).

$$\hat{y}_i = \sum_k^K f_k(x_i), \quad f_k \in F \quad (7)$$

dengan  $F = \{ f(x) = w_{q(x)} \mid q: R^m \rightarrow \{1, 2, \dots, T\}, w \in R^T \}$  sebagai ruang fungsi *regression tree*. Setiap fungsi  $f_k \in F$  merepresentasikan satu pohon regresi, dengan  $q(x)$  sebagai struktur pohon yang memetakan observasi ke *leaf* tertentu dan  $w$  sebagai bobot *leaf*. Pada iterasi ke- $t$ , pohon baru  $f_t(x)$  dibentuk untuk mengoreksi error model sebelumnya. Proses optimasi menggunakan pendekatan Taylor orde dua pada Persamaan (8) yang membuat pemilihan node *split* lebih akurat.

$$L^{(t)} \approx \sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2] + \Omega(f_t) \quad (8)$$

Pada Persamaan (8),  $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$  menyatakan *gradient*, yaitu turunan pertama fungsi *loss* terhadap nilai prediksi pada iterasi sebelumnya, sedangkan  $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$  menyatakan hessian, yaitu turunan kedua fungsi *loss* terhadap nilai prediksi pada iterasi sebelumnya. Sementara itu,  $\Omega(f_t)$  merupakan fungsi regularisasi yang digunakan untuk mengontrol kompleksitas *tree* pada iterasi ke- $t$ .

Regularisasi digunakan untuk mencegah *overfitting* pada data musiman dan fluktuatif seperti jumlah penumpang harian LRT. Regularisasi ditunjukkan pada Persamaan (9).

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (9)$$

Pada Persamaan (9),  $T$  menyatakan jumlah *leaf* dalam *tree*,  $w_j$  merupakan bobot pada *leaf* ke- $j$ ,  $\gamma$  adalah parameter penalti terhadap jumlah *leaf*, dan  $\lambda$  merupakan parameter regularisasi L2 yang digunakan untuk mengontrol besar bobot *leaf* agar model tidak terlalu kompleks.

Bobot optimal tiap *leaf* dihitung dalam proses *tree splitting* untuk memilih node terbaik pada setiap iterasi yang ditunjukkan pada Persamaan (10).

$$w_j^* = -\frac{G_j}{H_{j+}} \quad (10)$$

Pada Persamaan (10),  $w_j^*$  merupakan bobot optimal pada *leaf* ke- $j$ . Nilai  $G_j = \sum_{i \in I_j} g_i$  menyatakan akumulasi *gradient* dari seluruh observasi yang berada pada *leaf* ke- $j$ , sedangkan  $H_j = \sum_{i \in I_j} h_i$  menyatakan akumulasi *hessian* pada *leaf* tersebut. Sementara itu,  $I_j$  merupakan himpunan observasi yang termasuk ke dalam *leaf* ke- $j$ , dan  $\lambda$  adalah parameter regularisasi L2 yang digunakan untuk mengontrol besar bobot *leaf*.

*Walk-Forward Cross-Validation* dengan *Expanding Window* adalah metode validasi data deret waktu yang memeriksa urutan data dengan melatih model pada *window* yang terus diperluas (Nurhambali *et al.*, 2024). Pada setiap iterasi, *window train* menggunakan seluruh data historis hingga hari ke- $t$ , kemudian divalidasi pada titik  $t + 1$  seperti pada Tabel 1 sehingga mencegah adanya *leakage*.

Tabel 1. *Rolling data* pada *walk-forward cross-validation*

Data historis	Data prediksi
[Days]	Day1
[Days + Day1]	Day2
[Days + Day1 + Day2]	Day3

*Hyperparameter tuning* adalah proses pencarian kombinasi *hyperparameter* yang menghasilkan performa model terbaik pada data validasi dengan mengontrol parameter model (Wicaksana and Huda, 2025). Pada model *boosting*, pemilihan parameter berupa kedalaman *tree*, *learning rate*, atau jumlah perlu dilakukan dengan secara terpat agar model tidak mengalami *underfitting* atau *overfitting*. Secara umum, *hyperparameter tuning* bertujuan mencari nilai optimal yang meminimalkan fungsi objektif.

### 3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini merupakan data sekunder berupa jumlah penumpang harian *Light Rail Transit* (LRT) Jakarta yang diperoleh dari Portal Satu Data Jakarta (Dinas Perhubungan Provinsi DKI Jakarta, 2025). Sebagai penghubung antar moda (*feeder regional*), LRT merupakan sistem angkutan massal berbasis rel dengan kapasitas sedang dan menggunakan jalur layang terpisah dari lalu lintas jalan. Dataset mencakup periode 1 Januari 2024 hingga 31 Agustus 2025 dengan total 609 observasi tanpa *missing value*. Pemilihan rentang waktu tersebut didasarkan pada fakta bahwa LRT Jakarta baru mulai beroperasi pada 28 Agustus 2023 sehingga ketersediaan data yang lengkap dan konsisten baru tersedia mulai awal tahun 2024.

Proses analisis data dilakukan dengan menggunakan *python* dengan dua metode analisis yang dilakukan untuk perbandingan model. Model pertama dengan metode klasik, yaitu SARIMAX, sedangkan model kedua menggunakan XGboost. Berikut adalah tahapan-tahapan pada penelitian ini:

1. Eksplorasi data untuk mengidentifikasi pola deret waktu mencakup:
  - a) Pemeriksaan statistik deskriptif
  - b) Dekomposisi musiman mingguan
  - c) Analisis pola khusus pada hari libur, akhir pekan, dan *event* promosi tarif LRT.

2. Penambahan variabel *dummy* ke dalam dataset, terdiri dari *dummy* akhir pekan, hari libur nasional dan cuti bersama, libur panjang (rangkaian tanggal dengan minimal tiga hari libur berturut-turut), hampir libur panjang (rangkaian tanggal yang hampir membentuk libur panjang, ditandai dengan satu hari kerja yang menyela rangkaian hari libur), serta *dummy event* promosi tarif LRT (Hari Kartini, Hari Ulang Tahun Jakarta, dan Hari Bhayangkara).
3. *Splitting* data menjadi data latih dan data uji dengan jumlah data latih sebanyak 553 observasi dan data uji sebanyak 56 observasi.
4. Pemodelan dengan metode SARIMAX
  - a. Pemodelan regresi *Ordinary Least Squares* (OLS) awal menggunakan variabel eksogen (X) yang terdiri dari variabel *dummy* yang tetera di tahapan 2.
  - b. Menganalisis sisaan regresi OLS melalui plot *Autocorrelation Function* (ACF) dan *Partial Autocorrelation Function* (PACF) untuk mengidentifikasi adanya pola autokorelasi, termasuk indikasi musiman mingguan dan potensi komponen *Autoregressive* (AR) atau *Moving Average* (MA) yang belum terjelaskan.
  - c. Melakukan uji stasioneritas menggunakan Augmented Dickey-Fuller (ADF) pada data yang telah dilakukan *differencing* musiman untuk memastikan kestasioneran rata-rata.
  - d. Melakukan *differencing* musiman sebesar 1 dengan periode 7 hari (*weekly seasonal differencing*) pada sisaan OLS untuk menghilangkan pola musiman yang masih tersisa.
  - e. Melakukan pemeriksaan kestasioneran dalam ragam dengan transformasi Box-Cox. Estimasi ( $\lambda$ ) serta interval kepercayaannya digunakan untuk menentukan apakah transformasi diperlukan sebelum pemodelan SARIMAX.
  - f. Memvisualisasikan kembali ACF dan PACF pada data yang sudah ditangani untuk menentukan kandidat orde SARIMAX yang konsisten dengan pola autokorelasi.
  - g. Melakukan pemilihan model terbaik menggunakan *walk forward CV* dengan *expanding window*, dengan metrik utama MAPE dan penilaian AIC per *fold* untuk mengevaluasi stabilitas performa tiap kandidat.
  - h. Melakukan *local grid search* untuk mengevaluasi orde SARIMAX dengan membangun kombinasi parameter  $p, q, P$ , dan  $Q$  di sekitar spesifikasi terbaik awal. Kandidat model dibentuk dari *grid* terbatas  $p \in \{0,1,2\}$ ,  $q \in \{0,1,2\}$ ,  $P \in \{0,1\}$ , dan  $Q \in \{1,2\}$  dengan menyaring kombinasi trivial maupun yang telah diuji sebelumnya. Secara prinsip, *grid search* mengevaluasi seluruh kombinasi dalam *grid* yang jumlah totalnya ditunjukkan pada Persamaan (11).

$$K_{total} = \prod_{i=1}^m k_i \quad (11)$$

dengan  $k_i$  adalah jumlah kandidat pada setiap *hyperparameter*. Meskipun bersifat deterministik, *grid search* cenderung tidak efisien pada ruang pencarian besar karena banyak kombinasi tidak memberikan kontribusi signifikan (Bergstra, Ca and Ca, 2012). Oleh karena itu, *local grid search* digunakan untuk membatasi ruang pencarian pada kombinasi yang relevan, dan setiap kandidat dievaluasi menggunakan *expanding-window cross-validation* sehingga spesifikasi SARIMAX dengan MAPE paling stabil dapat dipilih sebagai model final.

- i. Memeriksa diagnostik residual dari model terbaik (Ljung-Box, ADF, dan inspeksi ACF/PACF) untuk memastikan tidak ada autokorelasi yang tersisa dan bahwa model valid secara statistik.

- j. Menguji potensi *overfitting* dengan membandingkan model terbaik terhadap model dengan orde AR atau MA yang lebih tinggi ( $p+1$  atau  $q+1$ ) untuk memastikan bahwa peningkatan kompleksitas tidak menghasilkan perbaikan yang semu.
5. Pemodelan dengan XGBoost
- Pembentukan *feature engineering* kalender sebagai variabel eksogen yang tertera pada Tahap 2.
  - Pembentukan fitur tambahan yang meliputi fitur lag (lag 1, lag 2, lag 3, lag 7, dan lag 14), *rolling window* (*rolling mean* 7, *rolling mean* 14, dan *rolling std* 14), serta *seasonal differencing* pada lag 7. Tahapan ini bertujuan untuk menjaga keadilan perbandingan model XGBoost dengan SARIMAX yang memiliki struktur musiman.
  - Proses pemodelan dilakukan dengan *walk-forward cross-validation* menggunakan *expanding window*. Setiap *fold* menghasilkan nilai MAPE dan *hyperparameter*. Hasil tersebut digunakan sebagai dasar *hyperparameter tuning* dengan Optuna, yang mengoptimalkan pencarian menggunakan *Tree-structured Parzen Estimator* (TPE) (Akiba *et al.*, 2019). Pada pendekatan ini, distribusi hasil evaluasi dibagi menjadi dua kelompok yang tertera pada Persamaan (12) dan Persamaan (13). Proses pemodelan dilakukan dengan *walk-forward cross-validation* menggunakan *expanding window*. Setiap *fold* menghasilkan nilai MAPE dan *hyperparameter*. Hasil tersebut digunakan sebagai dasar *hyperparameter tuning* dengan Optuna, yang mengoptimalkan pencarian menggunakan *Tree-structured Parzen Estimator* (TPE) (Akiba *et al.*, 2019). Pada pendekatan ini, distribusi hasil evaluasi dibagi menjadi dua kelompok yang tertera pada Persamaan (12) dan Persamaan (13).

$$l(x) = p(x \mid y < y^*) \quad (12)$$

$$g(x) = p(x \mid y \geq y^*) \quad (13)$$

dan pemilihan hyperparameter dilakukan dengan memaksimalkan rasio dengan persamaan yang terdapat pada Persamaan (14):

$$x^* = \underset{x}{\operatorname{argmax}} \frac{l(x)}{g(x)} \quad (14)$$

sehingga proses pencarian lebih terarah pada kombinasi *hyperparameter* yang memiliki peluang terbesar menghasilkan MAPE terendah, menjadikannya lebih efisien untuk model kompleks seperti XGBoost.

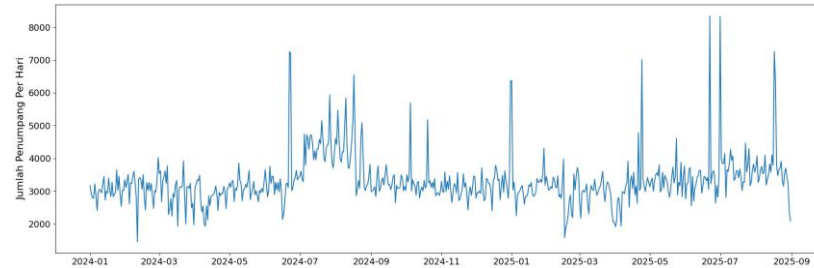
- Peramalan dilakukan secara rekursif dengan memprediksi langkah pertama ( $t+1$ ) dan memperbarui seluruh lag serta *rolling* berdasarkan hasil prediksi tersebut untuk mencegah *leakage*.
  - Prediksi langkah berikutnya dilakukan berdasarkan prediksi langkah sebelumnya dengan terus membangun ulang atau *update* seluruh lag dan *rolling* baru. Proses ini dilakukan secara iterasi hingga horizon  $h=56$  terpenuhi.
6. Melakukan evaluasi pada model terbaik dengan metrik evaluasi menggunakan *Mean Absolute Percentage Error* (MAPE) terhadap data uji yang ditunjukkan oleh Persamaan (15).

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|Y_t - \hat{Y}_t|}{Y_t} \quad (15)$$

Pada Persamaan (15),  $Y_t$  menyatakan nilai aktual pada periode ke- $t$ , sedangkan  $\hat{Y}_t$  merupakan nilai hasil peramalan pada periode ke- $t$ . Sementara itu,  $n$  menunjukkan jumlah observasi yang digunakan dalam proses evaluasi. Nilai MAPE menunjukkan rata-rata persentase kesalahan absolut antara nilai aktual dan hasil peramalan sehingga semakin kecil nilai MAPE maka semakin baik kinerja model.

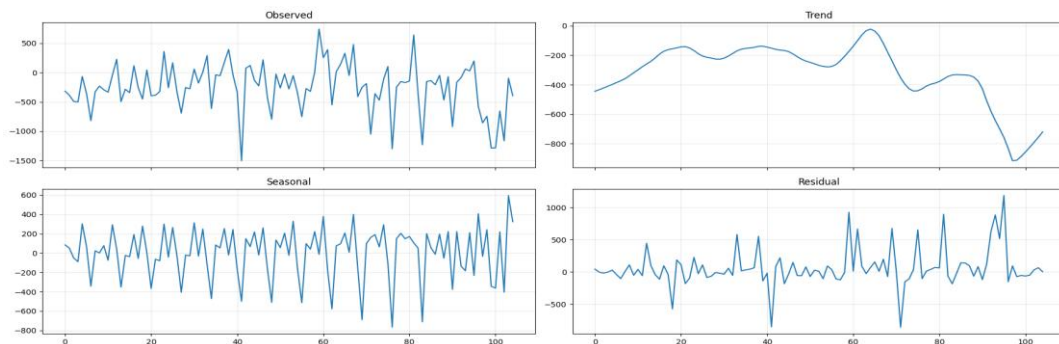
7. Perbandingan model SARIMAX dan XGBoost dalam *forecasting* jumlah penumpang harian LRT Jakarta untuk 56 hari ke depan.

#### 4. HASIL DAN PEMBAHASAN

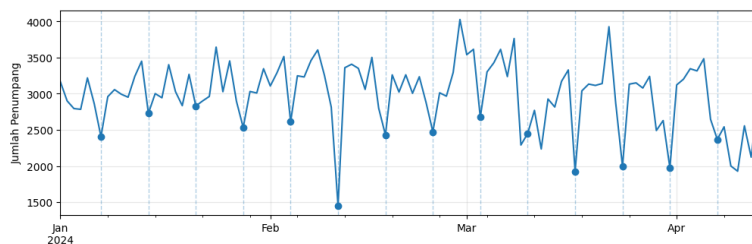


Gambar 1. Plot deret waktu jumlah penumpang LRT harian

Plot deret waktu menunjukkan pola yang berulang setiap beberapa periode tertentu. Hal tersebut menandakan adanya *seasonality* pada data. Selain itu, terlihat bahwa pada hari-hari tertentu terdapat lonjakan penumpang LRT seperti pada Gambar 1, mengindikasikan terdapat suatu faktor eksternal yang mempengaruhinya. Pola musiman yang berulang serta pengaruh variabel eksternal menjadikan SARIMAX dan XGBoost sebagai dua pendekatan yang relevan untuk dibandingkan. SARIMAX unggul dalam memodelkan struktur musiman dan autokorelasi secara eksplisit, sementara XGBoost mampu menangkap hubungan non-linier antara fitur temporal dan faktor eksternal. Meskipun indikasi musiman muncul pada visualisasi awal, pemastian terhadap pola tersebut tetap diperlukan untuk menghindari kesalahan interpretasi, sehingga analisis dapat dilengkapi dengan dekomposisi deret waktu menggunakan metode *Seasonal-Trend Decomposition using LOESS* (STL) (Rizky Mahendra *et al.*, 2025). Hal tersebut memungkinkan pemisahan komponen tren, musiman, dan residual secara sistematis.



Gambar 2a. Plot dekomposisi STL pada 15 minggu pertama tahun 2024

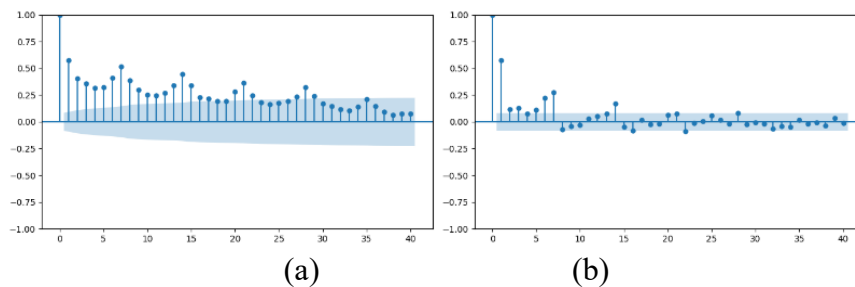


Gambar 2b. Plot deret waktu 15 minggu pertama tahun 2024

Pada Gambar 2a, pola musiman mingguan terlihat melalui dekomposisi STL, dan diperjelas dengan melihat plot jumlah penumpang secara langsung. Pemilihan rentang 15

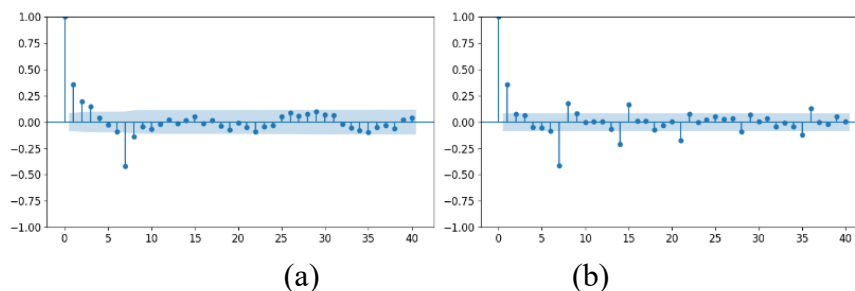
minggu dilakukan karena periode ini cukup panjang untuk menangkap beberapa siklus mingguan, namun masih cukup pendek sehingga fluktuasi musiman yang berulang dapat diamati secara jelas. Pada Gambar 2b, terlihat bahwa jumlah penumpang selalu menurun pada hari ke-7 setiap minggunya, memperkuat pola musiman mingguan yang telah terlihat pada Gambar 2a.

Tahapan awal pada pemodelan diawali dengan analisis regresi untuk melihat hubungan antara hari libur dan jumlah penumpang. Hasil residual dari pemodelan regresi menunjukkan nilai Durbin-Watson sebesar 0,853, menandakan terdapat autokorelasi positif pada residualnya. Berikutnya dilakukan pemeriksaan pola residual pada hasil regresi dengan menggunakan plot ACF dan PACF pada sisaan regresi yang dapat dilihat pada Gambar 4.



Gambar 4. (a) Plot ACF dan (b) Plot PACF residual regresi

Plot ACF dan PACF menunjukkan adanya pola musiman mingguan yang ditandai oleh signifikansi pada lag ke-7 dan kelipatannya. Oleh karena itu, dilakukan *differencing* musiman satu kali dengan periode 7 untuk menghilangkan komponen musiman sebelum penentuan kandidat parameter model. Kestasioneran dalam rata-rata kemudian diperiksa menggunakan uji Augmented Dickey-Fuller (ADF) dan diperoleh p-value < 0,05, yang menandakan data telah stasioner. Selanjutnya, kestasioneran dalam varians diperiksa melalui transformasi Box-Cox, dengan estimasi  $\lambda$  sebesar 1,2267 dan selang kepercayaan [1,0636; 1,4037]. Nilai 1 tidak termasuk dalam selang tersebut sehingga transformasi Box-Cox diterapkan untuk menstabilkan varians.



Gambar 5. (a) Plot ACF dan (b) Plot PACF residual regresi setelah ditangani

Kandidat model awal ditentukan berdasarkan plot ACF dan PACF sisaan setelah *differencing* satu kali pada komponen musiman ditunjukkan pada Gambar 5. Kinerja setiap kandidat model dinilai melalui *walk-forward cross-validation* dengan *expanding-window* dengan mempertimbangkan nilai AIC dan MAPE. Berdasarkan hasil evaluasi tersebut, dipilih model terbaik SARIMAX(1,0,1)(0,1,1)<sub>7</sub> dengan nilai AIC terkecil yang tertera pada Tabel 1.

Tabel 1. Kinerja kandidat model berdasarkan *walk-forward cross-validation*

Model SARIMAX	AIC mean	MAPE mean	Ljung-Box pass rate
SARIMAX(1,0,1)(0,1,1) <sub>7</sub>	5613,6322	11,26%	0,9552

Model SARIMAX	AIC mean	MAPE mean	Ljung-Box pass rate
SARIMAX(2,0,0)(0,1,1) <sub>7</sub>	5642,3525	11,44%	0,1045
SARIMAX(0,0,1)(0,1,1) <sub>7</sub>	5645,1071	11,27%	0
SARIMAX(1,0,0)(0,1,1) <sub>7</sub>	5650,5191	11,21%	0
SARIMAX(1,0,0)(1,1,1) <sub>7</sub>	5650,6611	11,34%	0
SARIMAX(1,0,0)(1,1,0) <sub>7</sub>	5666,4126	11,95%	0

Model SARIMAX(1,0,1)(0,1,1)<sub>7</sub> diperiksa melalui uji diagnostik berupa Ljung-Box pada beberapa lag untuk memastikan tidak adanya autokorelasi residual, serta Uji ADF untuk memastikan kestasioneran residual yang ditunjukkan pada Tabel 2. Berdasarkan hasil diagnostik, model memenuhi asumsi tidak ada autokorelasi yang tersisa. Selanjutnya model digunakan untuk melakukan peramalan pada data uji, kemudian akurasi dihitung menggunakan nilai MAPE pada skala asli setelah dilakukan *inverse* transformasi Box-Cox dengan MAPE sebesar 8,641%.

Tabel 2. Hasil uji diagnostik pada model SARIMAX(1,0,1)(0,1,1)<sub>7</sub>

Uji Diagnostik	p-value	Keterangan
Ljung-Box Lag 7	0,1780	Tidak ada autokorelasi
Ljung-Box Lag 14	0,3608	Tidak ada autokorelasi
Ljung-Box Lag 28	0,6298	Tidak ada autokorelasi
ADF	<0,0001	Stasioner

Pemeriksaan *overfitting* dilakukan dengan membandingkan model SARIMAX(1,0,1)(0,1,1)<sub>7</sub> terhadap model dengan kompleksitas sedikit lebih tinggi, yaitu SARIMAX(1,0,2)(0,1,1)<sub>7</sub> dan SARIMAX(2,0,1)(0,1,1)<sub>7</sub>. Meskipun kedua model memberikan MAPE yang lebih rendah (8,405% dan 8,331% dibandingkan 8,641%), nilai AIC keduanya meningkat (10136,983 dan 10153,319 dibandingkan 10126,710), yang menunjukkan penalti kompleksitas model yang lebih besar. Peningkatan akurasi yang diperoleh tidak sebanding dengan kenaikan kompleksitas dan penurunan efisiensi model. Oleh karena itu, model SARIMAX(1,0,1)(0,1,1)<sub>7</sub> dipilih sebagai model akhir karena memberikan keseimbangan terbaik antara akurasi dan kesederhanaan model.

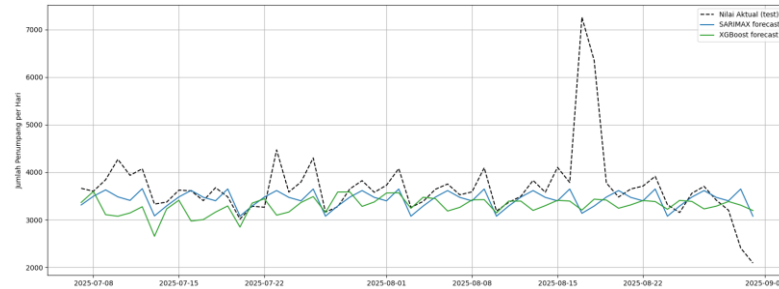
Pemodelan XGBoost menggunakan fitur-fitur deret waktu yang telah dijelaskan pada bagian metode. Proses *hyperparameter tuning* dilakukan menggunakan 50 percobaan dengan *framework* Optuna. Setiap percobaan dievaluasi melalui *65-fold walk-forward expanding-window cross-validation*. Hasil tuning menunjukkan bahwa *trial* ke-42 memberikan performa terbaik dengan nilai MAPE 13,09% pada *cross-validation*. Parameter lengkap untuk model terbaik ditampilkan pada Tabel 3.

Tabel 3. Parameter model terbaik XGBoost

Hyperparameter	Nilai
<i>N estimators</i>	245
<i>Max depth</i>	10
<i>Learning rate</i>	0,0065
<i>Subsample</i>	0,9968
<i>Colsample by tree</i>	0,9833
<i>Weight</i>	17,8451
$\gamma$	9,6417
$\lambda$	0,0039
$\alpha$	0,0068

Model final tersebut selanjutnya dievaluasi pada data *test* yang tidak digunakan selama proses *tuning* maupun *cross-validation*. Hasil evaluasi menunjukkan bahwa model XGBoost memiliki nilai MAPE sebesar 12,39% pada data *test*. Nilai ini lebih rendah dibandingkan rata-rata MAPE *cross-validation*, sehingga model dinilai mampu melakukan generalisasi dengan baik.

Hasil pada Tabel 4 menunjukkan bahwa model SARIMAX(1,0,1)(0,1,1)<sub>7</sub> menghasilkan MAPE lebih rendah dibandingkan model XGBoost. Nilai MAPE yang lebih kecil mengindikasikan bahwa SARIMAX memiliki tingkat kesalahan prediksi yang lebih rendah dan mampu memprediksi dengan lebih baik pada data jumlah penumpang harian LRT Jakarta dengan variabel eksogen.

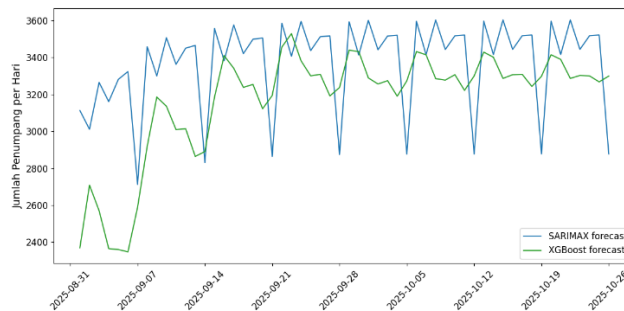


Gambar 6. Perbandingan hasil *forecasting* pada periode data uji

Tabel 4. Perbandingan model

Model	MAPE <i>Train</i>	MAPE <i>Test</i>
SARIMAX(1,0,1)(0,1,1) <sub>7</sub>	11,26%	8,64%
XGBoost	13,09%	12,39%

Performa SARIMAX yang lebih unggul menunjukkan bahwa model ini lebih mampu menangkap pola musiman mingguan serta dinamika temporal pada data. Di sisi lain, kinerja XGBoost cenderung kurang optimal karena karakteristiknya yang lebih sesuai untuk data berskala besar, sedangkan data dalam penelitian ini hanya mencakup rentang 20 bulan yang ditunjukkan pada Gambar 6. Kondisi ini memungkinkan XGBoost menangkap fluktuasi acak sebagai sinyal. Oleh karena itu, SARIMAX dinilai lebih sesuai untuk *forecasting* pada penelitian ini.



Gambar 7. *Forecasting* untuk 56 hari ke depan

Hasil prediksi untuk 56 hari ke depan menggunakan hasil terbaik dari kedua metode menunjukkan adanya perbedaan karakteristik dalam menangkap pola musiman harian LRT. Model SARIMAX mampu merepresentasikan fluktuasi mingguan lebih konsisten, sehingga pola naik–turun jumlah penumpang yang berulang setiap tujuh hari terlihat jelas dalam garis prediksinya yang ditunjukkan oleh Gambar 7. Sebaliknya, XGBoost menghasilkan estimasi yang lebih halus dan kurang responsif terhadap perubahan musiman, meskipun masih berada dalam kisaran nilai yang wajar. Kedua model memproyeksikan rata-rata jumlah penumpang harian berada pada kisaran 3.300–3.500 penumpang, yang konsisten dengan pola historis.

Temuan ini menunjukkan bahwa SARIMAX lebih mampu mempertahankan pola musiman mingguan pada data penumpang LRT, sedangkan XGBoost cenderung menghasilkan prediksi yang lebih stabil dan halus. Perbedaan tersebut mengindikasikan bahwa model statistik seperti SARIMAX lebih sesuai digunakan pada data transportasi publik yang memiliki pola musiman kuat dan fluktuasi harian yang berulang. Dengan kemampuan tersebut, SARIMAX berpotensi lebih efektif untuk mendukung perencanaan operasional sertaantisipasi perubahan jumlah penumpang harian pada layanan LRT.

## 5. KESIMPULAN

Hasil penelitian menunjukkan bahwa jumlah penumpang harian LRT memiliki fluktuasi tinggi dengan pola musiman mingguan yang kuat, serta dipengaruhi oleh faktor eksternal, sehingga memerlukan model peramalan yang mampu menangkap dinamika tersebut. Berdasarkan hasil evaluasi, model SARIMAX(1,0,1)(0,1,1)<sub>7</sub> terbukti memberikan akurasi terbaik dengan MAPE 8,64%, lebih baik dibandingkan XGBoost yang mencapai 12,39%. Peramalan 56 hari ke depan juga menunjukkan bahwa SARIMAX mereplikasi pola mingguan secara lebih stabil, sementara XGBoost menghasilkan prediksi yang lebih halus dan kurang responsif terhadap variasi musiman.

Temuan tersebut mengindikasikan bahwa model statistik seperti SARIMAX masih lebih sesuai digunakan pada data transportasi publik yang memiliki pola musiman kuat dan dipengaruhi faktor eksternal yang relatif konsisten. Kemampuan model dalam merepresentasikan fluktuasi jumlah penumpang secara lebih stabil diharapkan dapat mendukung perencanaan operasional transportasi publik, khususnya dalam mengantisipasi perubahan jumlah penumpang harian pada layanan LRT.

## DAFTAR PUSTAKA

- Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A Next-generation Hyperparameter Optimization Framework, in: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, pp. 2623–2631. DOI: <https://doi.org/10.1145/3292500.3330701>.
- Badan Pusat Statistik Provinsi DKI Jakarta, 2025. Penduduk, Laju Pertumbuhan Penduduk, Distribusi Persentase Penduduk, Kepadatan Penduduk, Rasio Jenis Kelamin Penduduk Menurut Kabupaten/Kota di Provinsi DKI Jakarta, 2024. BPS Provinsi DKI Jakarta. URL: <https://jakarta.bps.go.id>.
- Badan Pusat Statistik Provinsi DKI Jakarta, 2024. Perkembangan Transportasi Provinsi DKI Jakarta Juni 2024. BPS Provinsi DKI Jakarta. URL: <https://jakarta.bps.go.id>.
- Bergstra, J., Ca, J.B., Ca, Y.B., 2012. Random Search for Hyper-Parameter Optimization *Yoshua Bengio, Journal of Machine Learning Research*.
- Chen, T., Guestrin, C., 2016. XGBoost, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, pp. 785–794. DOI: <https://doi.org/10.1145/2939672.2939785>.
- Dinas Perhubungan Provinsi DKI Jakarta, 2025. Data Penumpang LRT di Provinsi DKI Jakarta. satudata. URL: <https://satudata.jakarta.go.id>.
- Harriz, M.A., Akbariani, N.V., Setiyowati, H., Santoso, H., 2023. Enhancing the Efficiency of Jakarta's Mass Rapid Transit System with XGBoost Algorithm for Passenger Prediction. *Jambura Journal of Informatics* 5, 1–6. DOI: <https://doi.org/10.37905/jji.v5i1.18814>.
- ITDP Indonesia, 2021. ITDP Annual Report 2021. Jakarta.

- Lee, G.C., Bang, J.Y., 2024. Forecasting Container Throughput of Singapore Port Considering Various Exogenous Variables Based on SARIMAX Models. *Forecasting* 6, 748–760. DOI: <https://doi.org/10.3390/forecast6030038>.
- Neydi, D., Pakaya, P., Achmad, N., Hasan, I.K., Wungguli, D., Abdussamad, S.N., 2025. Prediksi Wisatawan Mancanegara di Indonesia Menggunakan Metode SARIMAX dengan Efek Variasi Kalender Libur Nasional. *Riset Mahasiswa Matematika* 4, 287–300. DOI: <https://doi.org/10.18860/jrmm/v4i6.34937>.
- Nurhambali, M.R., Angraini, Y., Fitrianto, A., 2024. Implementation of Long Short-Term Memory for Gold Prices Forecasting. *Malaysian Journal of Mathematical Sciences* 18, 399–422. DOI: <https://doi.org/10.47836/mjms.18.2.11>.
- Rizky Mahendra, R., Damaliana, A.T., Diyasa, I.G.S.M., 2025. Pendekatan Time Series Decomposition (STL) Dalam Prediksi Kecelakaan Berbasis Kepadatan Lalu Lintas Sebagai Dasar Kebijakan Di Tol Surabaya-Gempol. *Jurnal Impresi Indonesia* 4, 1538–1548. DOI: <https://doi.org/10.58344/jii.v4i5.6491>.
- Saputra, A., Gustriansyah, R., Sanmorino, A., Mair, R.Z., Sartika, D., Puspasari, S., 2024. Prediction Passenger Numbers in Light Rail Transit using Seasonal Autoregressive Integrated Moving Average (SARIMA). *PRZEGLĄD ELEKTROTECHNICZNY* 1, 45–47. DOI: <https://doi.org/10.15199/48.2024.10.07>.
- Talusan, J.P., Mukhopadhyay, A., Freudberg, D., Dubey, A., 2022. On Designing Day Ahead and Same Day Ridership Level Prediction Models for City-Scale Transit Networks Using Noisy APC Data, in: *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 5598–5606. DOI: <https://doi.org/10.1109/BigData55660.2022.10020390>.
- The Jakarta Post, 2025. Greater Jakarta LRT breaks new ridership record. The Jakarta Post. URL: <https://www.thejakartapost.com/>.
- United Nations, E. and S.C. for A. and the P., 2019. Review of sustainable transport connectivity in Asia and the Pacific 2019: addressing the challenges for freight transport. United Nations Economic and Social Commission for Asia and the Pacific, Bangkok.
- Wang, X., Zhang, N., Zhang, Y., Shi, Z., 2018. Forecasting of Short-Term Metro Ridership with Support Vector Machine Online Model. *J Adv Transp* 2018, 1–13. DOI: <https://doi.org/10.1155/2018/3189238>.
- Wicaksana, H.S., Huda, K., 2025. Penerapan Word2Vec dan SVM dengan Hyperparameter Tuning untuk Deteksi Phishing. *JURIKOM (Jurnal Riset Komputer)* 12, 361–371. DOI: <https://doi.org/10.30865/jurikom.v12i3.8729>.