New Explorations in Electrical Engineering (NECE)
Department of Electrical Engineering,
Faculty of Engineering
Universiti Malaya, MALAYSIA

Vol. 2, No. 1 (2026)
pp. 12–20

A SARIMA-Based Rolling One-Step-Ahead Framework for Monthly Forecasting of Prefecture-Level Guest Nights in Japan

Chang Xu, Wong Jee Keen Raymond*, Hazlie Mokhlis, and Hazlee Azil Illias
Department of Electrical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur, Malaysia
* Corresponding author. E-mail: wjkraymond@um.edu.my

Abstract. Accurate monthly guest-night forecasting supports tourism planning and accommodation management. This study compares local seasonal autoregressive integrated moving average models with global multilayer perceptron, support vector regression, and extreme gradient boosting models for 47 Japanese prefectures using monthly observations from January 2007 to January 2026. The benchmark considers model type and training scope, although their individual effects are not isolated. Performance is evaluated over January to December 2024 using a rolling one-step-ahead protocol. Each forecast uses only observations available up to the end of the preceding month; the actual value is incorporated after prediction. The local seasonal model achieves the lowest normalized mean absolute error, normalized root mean squared error, and symmetric mean absolute percentage error of 5.12%, 9.42%, and 6.16%, respectively. These results support local seasonal modeling for this dataset, which exhibits seasonality, regional heterogeneity, and pandemic-related disruption. Supplementary projections for February to December 2026 use models refitted on observations through January 2026 and are not treated as an accuracy evaluation. Overall, the local seasonal model outperforms the evaluated global machine-learning configurations, highlighting the importance of matching forecasting approaches to data characteristics.

Keywords. Tourism demand forecasting, rolling one-step-ahead forecasting, SARIMA, Japanese prefectures, sustainable tourism planning

1. Introduction

Forecasting guest nights is important for tourism demand analysis, regional planning, and accommodation management (Song et al., 2019). In Japan, monthly guest-night forecasting at the prefecture level is particularly meaningful because tourism demand exhibits clear seasonality and substantial regional heterogeneity. For this task, model selection is not straightforward. Monthly guest-night series usually contain a strong annual cycle, while different prefectures vary in tourism scale, fluctuation pattern, and recovery speed after the pandemic. These characteristics increase forecasting difficulty and make it necessary to evaluate models under realistic conditions. Although machine-learning and deep-learning methods have been widely used in forecasting (Law et al., 2019), higher model complexity does not always guarantee better performance. For forecasting problems with limited monthly samples and strong seasonal structure, simpler statistical models may still be highly effective.

Motivated by these issues, this study develops a rolling one-step-ahead forecasting framework for monthly guest-night prediction in 47 Japanese prefectures. The benchmark considers two dimensions: model type (statistical versus machine learning) and training scope (local versus global). Seasonal autoregressive integrated moving average (SARIMA) is fitted independently to each prefecture, whereas a multilayer perceptron (MLP), a support vector machine (SVM) for regression, and extreme gradient

NECE. All rights reserved.

boosting (XGBoost) are trained globally using pooled observations from all prefectures. All methods are evaluated on the same test period under the same forecasting information constraint. Local SARIMA achieves the best performance among the evaluated configurations. Because model type and training scope vary together, their individual contributions cannot be isolated in this comparison.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the methodology. Section 4 describes the experimental setup. Section 5 reports the results and discussion. Section 6 concludes the paper.

2. Related work

Tourism demand forecasting and related time-series problems have been studied using both classical statistical models and machine-learning methods. Classical models such as autoregressive integrated moving average (ARIMA) and SARIMA remain important because they can explicitly represent autocorrelation and seasonality (Goh & Law, 2002), which are common in monthly demand series. Machine-learning models such as SVM (Chen & Wang, 2007) and tree-based methods (He & Qian, 2026) have also been widely used in forecasting. These models can learn nonlinear relationships from lagged observations and engineered temporal features, and they are often applied when the data contains more complex cross-sectional patterns.

However, model complexity does not always lead to better forecasting performance. When the data are monthly, the sample size is limited, and seasonality is strong, simpler models may still be competitive. Another important issue is evaluation methodology. A fair forecasting comparison requires a strict time-order constraint so that future information is not used during model training or testing (Hewamalage et al., 2023). Based on these considerations, this study conducts a fair rolling one-step-ahead comparison between SARIMA and several machine-learning baselines for prefecture-level guest-night forecasting in Japan.

3. Methodology

3.1. Problem definition

Let $y_{i,t}$ denote the number of guest nights for prefecture i in month t, where $i \in \{1,2,\ldots,47\}$. The objective is to forecast future guest nights from historical observations.

This study considers two related tasks:

1) A fair rolling one-step-ahead evaluation for January to December 2024. When forecasting month t, only information available up to month $t-1$ is allowed.
2) A future forecasting task for February to December 2026. In this task, each model is first retrained or refitted by using all real observations up to January 2026. Recursive forecasts are then generated, and the direction of annual change in 2026 relative to 2025 is estimated.

3.2. Data preprocessing

The guest-night data were obtained from the official statistical source of Japan (Statistics Bureau, Ministry of Internal Affairs and Communications, n.d.). The raw file contained 11,904 records, including annual and monthly observations for Japan and its 47 prefectures. Preprocessing excluded 912 annual aggregate records and 229 national monthly records. Month labels were converted to month-start timestamps, commas and spaces were removed from the guest-night field before numeric conversion, and

the records were sorted by prefecture and month. The resulting panel contained 10,763 observations, with 229 consecutive months for each prefecture from January 2007 to January 2026. A data-quality check identified no missing months, missing target values, duplicate prefecture-month records, or non-positive guest-night values in the retained panel. No imputation, smoothing, or statistical outlier removal was applied.

Large fluctuations during the coronavirus disease 2019 (COVID-19) pandemic were retained as recorded demand variations rather than removed as outliers. Based on the sum across prefectures, annual guest nights in 2020 and 2021 were 44.35% and 46.68% below the 2019 level, respectively. Figure 1 compares monthly observations for the five prefectures with the largest percentage declines in annual guest nights from 2019 to 2020. The curves show a sharp contraction in 2020, followed by fluctuations and differing recovery trajectories.

3.3. Feature engineering for baseline models

For the global supervised-learning baseline models, the following input features are constructed:

1) Time features: year and month.
2) Cyclical encoding features: the calendar month $m \in \{1,2,\ldots,12\}$ is represented by $sin(2\pi m/12)$ and $cos(2\pi m/12)$. This encoding preserves the annual cycle and the adjacency of December and January.
3) Stage dummy variables: the pandemic indicator equals one from January 2020 to December 2021 and zero otherwise. The recovery indicator equals one from January 2022 onward and zero otherwise. These calendar-based indicators approximate broad demand regimes rather than exact policy boundaries or prefecture-specific recovery dates.
4) Lag features: guest nights in the previous 1, 2, 3, 6, and 12 months.
5) Rolling mean features: rolling averages over the previous 3, 6, and 12 months.
6) Prefecture dummies: 47 binary indicators encode prefecture identity. For each observation, the indicator corresponding to its prefecture equals one and all others equal zero. These variables allow the global models to distinguish between prefectures while sharing model parameters across the pooled dataset.

The resulting input comprises 14 numerical features and 47 prefecture indicators. All features are constructed strictly from information available before the current forecast origin, thereby preventing future information leakage. When some lagged features are unavailable at a forecast step, a seasonal naive forecast is used as a fallback to ensure that the rolling forecasting procedure can continue.

3.4. Proposed model and baseline models

SARIMA is fitted locally, with a separate model for each prefecture. In contrast, MLP, SVM, and XGBoost are trained globally using pooled observations from all 47 prefectures. Thus, the benchmark compares a local statistical approach with three global machine-learning approaches.

For each prefecture, four SARIMA specifications are attempted in a fixed order: $(1,1,1)(1,1,1)_{12}$, $(0,1,1)(0,1,1)_{12}$, $(1,1,0)(1,1,0)_{12}$, and $(0,1,1)(1,1,0)_{12}$. In SARIMA$(p,d,q)(P,D,Q)_{12}$, p, d, and q denote the non-seasonal autoregressive, differencing, and moving-average orders, respectively, while P, D, and Q denote the corresponding seasonal orders.

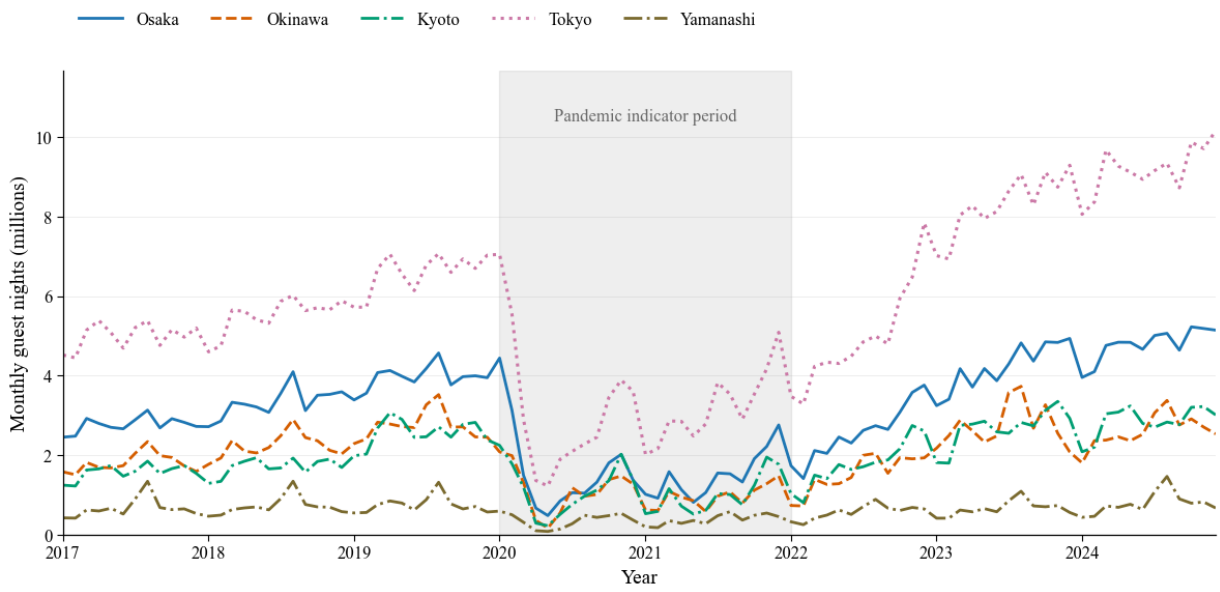


Fig. 1. Monthly guest nights in Osaka, Okinawa, Kyoto, Tokyo, and Yamanashi from January 2017 to December 2024. These prefectures recorded the five largest percentage declines in annual guest nights from 2019 to 2020. The shaded interval corresponds to the pandemic indicator used in this study, covering January 2020 to December 2021.

The seasonal period is 12 months. The first specification whose fitting procedure returns without an exception is retained; the candidates are not ranked using an information criterion. Models are fitted without a deterministic trend, with stationarity and invertibility enforcement disabled. If all four specifications fail, a seasonal naïve forecast based on the corresponding month of the previous year is used.

The MLP consists of two hidden layers with 128 and 64 hidden units, respectively. The rectified linear unit (ReLU) is used as the activation function, and Adam is adopted as the optimizer. Input features are standardized before training. This model is trained globally using samples from all prefectures, while prefecture dummies are introduced to capture cross-sectional differences.

The SVM baseline is implemented using support vector regression (SVR) with a radial basis function (RBF) kernel. Input features are standardized, while the target remains on its original scale. Grid search evaluates 18 combinations of $C \in \{1,10,30\}$, $\varepsilon \in \{0.01,0.1\}$, and $\gamma \in \{scale, 0.01,0.05\}$. Here, C controls regularization, ε specifies the width of the insensitive loss region, and γ determines the kernel coefficient. For the 2024 evaluation, January to December 2023 serves as the validation period, with preceding observations used to fit the candidate models. The configuration with the lowest validation means absolute error is selected and refitted using all training observations through December 2023.

XGBoost is implemented as a global model based on an ensemble of regression trees with squared error as the optimization objective. The model learns nonlinear relationships from time features, lagged statistics, rolling means, and prefecture indicators.

3.5. Proposed testing framework

To evaluate the proposed SARIMA model under realistic forecasting conditions, a strict one-step-ahead testing framework is adopted in the 2024 test stage. The initial history spans from January 2007 to December 2023. When forecasting a month in 2024, only observations up to the previous month are used.

After the forecast for the current month is generated, the true observed value of that month is added to the history and becomes available for the next step.

The same testing rule is applied to all baseline models so that the comparison remains fair. This framework is intended to reflect actual deployment conditions more closely than a static one-shot test, because each monthly prediction is produced under the information constraint that would exist in practice. Figure 2 shows the workflow of the proposed testing framework.

3.6. Evaluation metrics

Forecasting performance is evaluated using normalized mean absolute error (NMAE), normalized root mean squared error (NRMSE), and symmetric mean absolute percentage error (sMAPE). Their definitions are given as follows:

$$\mathrm{NMAE} = \frac{\frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n|}{\frac{1}{N}\sum_{n=1}^{N}|y_n|} \times 100\% \tag{1}$$

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}}{\frac{1}{N}\sum_{n=1}^{N}|y_n|} \times 100\% \tag{2}$$

$$\mathrm{sMAPE} = \frac{100\%}{N}\sum_{n=1}^{N}\frac{2|y_n - \hat{y}_n|}{\max(|y_n| + |\hat{y}_n|, \varepsilon)} \tag{3}$$

Here, y_n and $\hat{y}_n$ denote the observed and predicted guest nights for sample n, respectively. For the overall 2024 evaluation, $N = 47 \times 12 = 564$, corresponding to 47 prefectures over 12 test months. The metrics are calculated jointly over these prefecture-month observations rather than averaged from prefecture-level metric values. NMAE and NRMSE are normalized by the mean absolute observed value across all N observations. All three metrics are expressed as percentages, with lower values indicating better forecasting performance.

In Eq. (3), the numerical safeguard is set to $\varepsilon = 10^{-8}$. The max operator selects the larger of $|y_n| + |\hat{y}_n|$ and ε, ensuring that the denominator is at least ε and preventing division by zero. If both the observed and predicted values are zero, the corresponding contribution to sMAPE is zero.

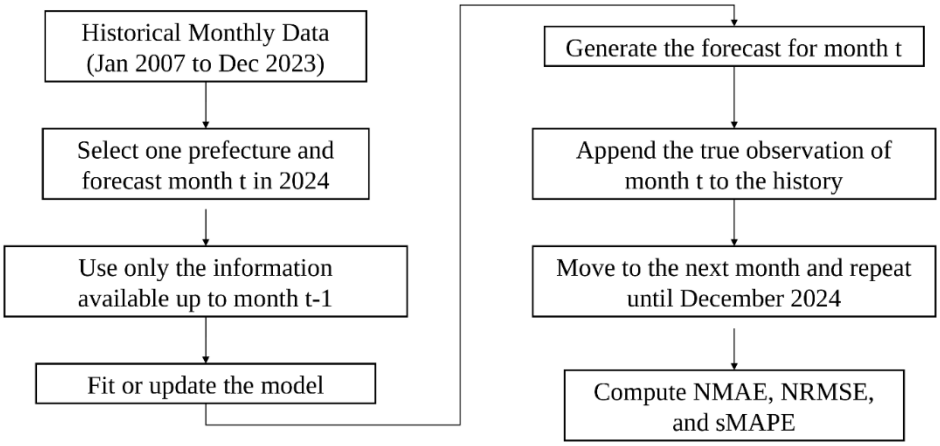


Fig. 2. Flowchart of the proposed rolling one-step-ahead testing framework.

4. Experimental setup

4.1. Data range and split

The dataset covers 47 Japanese prefectures. Each prefecture contains 229 monthly observations from January 2007 to January 2026. In the 2024 rolling evaluation stage, the initial training history spans January 2007 to December 2023, while the test period covers January 2024 to December 2024. The number of effective samples for the supervised-learning baseline models in this stage is 9,024.

In the future forecasting stage, all models are retrained or refitted using all real observations up to January 2026. At this point, the number of effective training samples for the supervised-learning baseline models increases to 10,199. The models then generate recursive forecasts for February to December 2026.

4.2. Future trend analysis setting

The 2026 results are not interpreted as a comparison of forecast accuracy because observations after January 2026 are outside the dataset used in this study. Instead, they are used only as directional projections from the models. Specifically, the actual value of January 2026 is combined with the predicted values from February to December 2026 to obtain the estimated annual total for 2026. This total is then compared with the actual annual total of 2025. When the absolute change rate does not exceed 1.0%, the result is labeled projected stable. Otherwise, it is labeled projected increase or projected decrease.

5. Results and discussion

5.1. Rolling one-step-ahead forecasting results for 2024

The overall results on the 2024 rolling one-step-ahead test set are reported in Table 1.

Table 1. Overall forecasting results on the 2024 test set.

Model	**NMAE (%)**	**NRMSE (%)**	**sMAPE (%)**
MLP	8.72	14.71	10.31
SVM	69.04	143.54	64.83
SARIMA	5.12	9.42	6.16
Global XGBoost	9.26	16.69	9.14

Table 1 shows that the proposed SARIMA model achieves the best performance among all compared methods across all three reported metrics. In particular, it attains the lowest NMAE, NRMSE, and sMAPE values of 5.12%, 9.42%, and 6.16%, respectively. These results indicate that SARIMA provides the most accurate forecasts within the present experimental setting.

The MLP and Global XGBoost exhibit broadly comparable performance. Both are substantially better than SVM, yet both remain inferior to SARIMA. This pattern suggests that global nonlinear machine-learning models can produce reasonable forecasts when supplied with engineered time-series features and prefecture indicators. However, for the present task, these models still do not match a local seasonal time-series model that explicitly captures the structure of monthly demand.

By contrast, SVM performs considerably worse than the other models. Its NMAE, NRMSE, and sMAPE reach 69.04%, 143.54%, and 64.83%, respectively, far exceeding the corresponding values of the other methods. This result indicates that the evaluated SVM configuration did not adequately capture the variation in guest nights under the present experimental setting.

5.2. Discussion

The observed performance differences can be interpreted from several perspectives. First, guest nights form a typical monthly time series with pronounced annual seasonality. SARIMA explicitly models a 12-month seasonal cycle, making it naturally appropriate for such data. By contrast, MLP, SVM, and XGBoost can represent seasonality only indirectly through month encodings, lagged variables, and rolling statistics. Their representation of seasonality therefore depends on the selected features and fitted relationships.

Second, Japanese prefectures differ substantially in tourism scale, fluctuation amplitude, peak-season pattern, and recovery speed. In this study, MLP, SVM, and XGBoost share model parameters across prefectures, while prefecture indicators distinguish regional identities. In contrast, separate SARIMA models allow prefecture-specific parameter estimates, which may help accommodate differences in local temporal dynamics.

Third, the pandemic shock and the subsequent recovery introduced heterogeneous structural changes across regions. The global baseline models incorporate pandemic and recovery dummies, but the magnitude and persistence of these effects are not uniform across prefectures. A shared set of stage variables may therefore constrain model flexibility. Local fitting may help accommodate prefecture-specific dynamics, although its contribution cannot be separated from model type in the present comparison.

The RBF-SVM produced a highly compressed prediction range in 2024, approximately 0.489 to 0.510 million guest nights, compared with an observed range of 0.145 to 10.153 million, indicating substantial underfitting. Although its inputs were standardized, the target remained on its original scale and C was searched only over $\{1,10,30\}$, making the restricted parameter range a plausible contributor to this behavior. Its poor performance therefore reflects the evaluated configuration and should not be interpreted as a general limitation of SVM-based forecasting.

More broadly, these findings suggest that forecasting performance should not be judged solely by model complexity. For some problems, especially those characterized by limited monthly sample size, clear seasonality, and strong local heterogeneity, a simpler non-deep-learning model may be more appropriate than a more complex machine-learning approach. The present results do not imply that simple models are always superior. Rather, they indicate that model choice should be guided by the structure of the data and by the realism of the testing protocol.

5.3. Supplementary trend projection for 2026

After refitting the models using observations up to January 2026, forecasts were generated for February to December 2026. The national-level projections are reported in Table 2. For reference, the recorded total number of guest nights in 2025 was 653,476,960, and the corresponding value for January 2026 was 45,462,790. Minor arithmetic discrepancies among the displayed forecasts, totals, and changes arise from rounding these quantities independently.

Table 2. Projected change in national guest nights in 2026 relative to 2025.

Model	2026 Feb-Dec Forecast	2026 Estimated Total	Absolute Change vs. 2025	Change Rate (%)	Direction
MLP	647,110,200	692,573,000	39,096,050	5.98	Projected Increase
SVM	264,070,500	309,533,300	-343,943,600	-52.63	Projected Decrease
SARIMA	577,830,600	623,293,400	-30,183,540	-4.62	Projected Decrease
Global XGBoost	584,327,600	629,790,400	-23,686,560	-3.62	Projected Decrease

As shown in Table 2, all models except the MLP project a decline in national guest nights in 2026 relative to 2025. SARIMA and Global XGBoost produce similar directional projections, with decreases of 4.62% and 3.62%, respectively. SVM projects a much larger decline of 52.63%. Given its poor performance in the 2024 evaluation, this projection should be interpreted cautiously. By contrast, the MLP projects an increase of 5.98%, indicating that the models do not fully agree on the future direction.

The results in Table 2 represent model-based projections rather than an accuracy evaluation, as observations after January 2026 were not included in this study. SARIMA achieved the best performance on the 2024 test set, but this does not establish its accuracy for the 2026 projection. These projections require validation against subsequent observations.

6. Conclusion

This study developed a SARIMA-based rolling one-step-ahead framework for monthly guest-night forecasting in 47 Japanese prefectures. Under a unified and realistic evaluation setting, SARIMA was compared with MLP, SVM, and Global XGBoost. The results showed that SARIMA achieved the best overall performance on the 2024 test set in terms of NMAE, NRMSE, and sMAPE. This suggests that a local seasonal time-series model is well suited to prefecture-level guest-night series with strong seasonality and regional heterogeneity.

More broadly, the findings indicate that, for this forecasting task, model appropriateness and evaluation realism are more important than model complexity alone. A supplementary projection analysis for 2026 was also provided, although it should be interpreted only as a directional reference because observations after January 2026 were not included in the evaluation.

Future work may consider exogenous variables, additional local and global forecasting models, and hybrid forecasting strategies under the same fair rolling evaluation protocol.

Acknowledgements

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors also acknowledge the support of the International Transdisciplinary Research Conference (INTRAC) 2026 organized by Universiti Malaya.

References

Chen, K.-Y., & Wang, C.-H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management, 28*(1), 215–226.

Goh, C., & Law, R. (2002). Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tourism Management, 23*(5), 499–510.

He, M., & Qian, X. (2026). Forecasting tourist arrivals using STL-XGBoost method. *Tourism Economics, 32*(2), 408–436.

Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Mining and Knowledge Discovery, 37*(2), 788–832.

Law, R., Li, G., Fong, D. K. C., & Han, X. (2019). Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research, 75*, 410–423.

Song, H., Qiu, R. T. R., & Park, J. (2019). A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting. *Annals of Tourism Research, 75*, 338–362.

Statistics Bureau, Ministry of Internal Affairs and Communications. (n.d.). *Statistics Dashboard: Number of guest nights (Total), Accommodation Survey*. Retrieved May 5, 2026, from https://dashboard.e-stat.go.jp/en/timeSeriesResult?indicatorCode=1003010201000010000