Time Series Forecasting of Sports Ticket Sales with Venue-Independent Fan Segments

Shun Takagi[1] [a], Haru Terashima[1] [b], Naoki Tamura[1] [c], Kazuyuki Shoji[1] [d], Kenta Urano[1] [e], Takuro Yonezawa[1] [f] and Nobuo Kawaguchi[1,2] [g]

[1] *Graduate School of Engineering, Nagoya University, Nagoya, Japan*

[2] *Institute of Innovation for Future Society, Nagoya University, Nagoya, Japan*

Keywords:
Sports Attendance Forecasting, Time Series Forecasting, Fan Segmentation, LSTM.

Abstract:
Online ticket sales in sports events accumulate detailed data such as daily sales volumes and purchaser attributes, establishing a foundation for optimizing sales strategies. Optimizing ticket sales requires time series forecasting during the sales period; however, this has not been sufficiently explored. Sports ticket sales have a short sales period, and the limited observations available as input to forecasting models make existing methods prone to reduced accuracy. Furthermore, the recent surge in new stadium construction has increased demand for forecasts at venues lacking historical data. In such situations, leveraging data from existing venues becomes crucial, yet inter-venue differences make it difficult to extract common patterns from total sales alone. In this study, we propose an input representation that decomposes total sales into fan segments based on visit frequency. This enables extraction of purchasing patterns common across venues. In our experiments, we first compared existing time series forecasting methods and confirmed that LSTM combined with fan segment decomposition achieved the highest accuracy. Furthermore, through comparison experiments varying input representations and input lengths, fan segment decomposition consistently demonstrated higher accuracy than other decomposition methods.

1 INTRODUCTION

Online ticket sales in sports events accumulate detailed time series data on ticket transactions. With the adoption of digital platforms, information such as purchaser attributes, ticket price categories, and daily sales volumes is now recorded at a fine granularity, enabling quantitative analysis of the sales process. In sports event operations, teams must adjust seat allocation (Leslie and Sorensen, 2013; Woratschek et al., 2014), promotional campaigns (Courty, 2003; Coates and Humphreys, 2005), pricing (Chang et al., 2016; Shapiro and Drayer, 2012), and discount strategies (Cisyk and Courty, 2021; Sahni et al., 2017) in anticipation of demand that will occur by the event day. This has laid the foundation for sports teams to optimize their ticket sales strategies based on sales data.

Building on this foundation, teams need to make various decisions before the event date. Examples include determining the number and timing of complimentary ticket distribution, and adjusting prices in dynamic pricing schemes. However, in practice, while diverse data is referenced, operations often rely on rules of thumb. To support these decisions, time series forecasting of future sales trends from any point during the sales period is required.

However, previous research on sports attendance forecasting has primarily focused on forecasting the final attendance on the event day. Approaches that forecast daily sales trends during the sales period have not been sufficiently explored.

When formulating sports ticket sales as a

[a] https://orcid.org/0009-0002-4375-4120
[b] https://orcid.org/0009-0001-0234-5307
[c] https://orcid.org/0000-0002-3362-697X
[d] https://orcid.org/0000-0001-5946-2646
[e] https://orcid.org/0000-0003-2906-537X
[f] https://orcid.org/0000-0001-9781-0402
[g] https://orcid.org/0000-0002-0444-2290

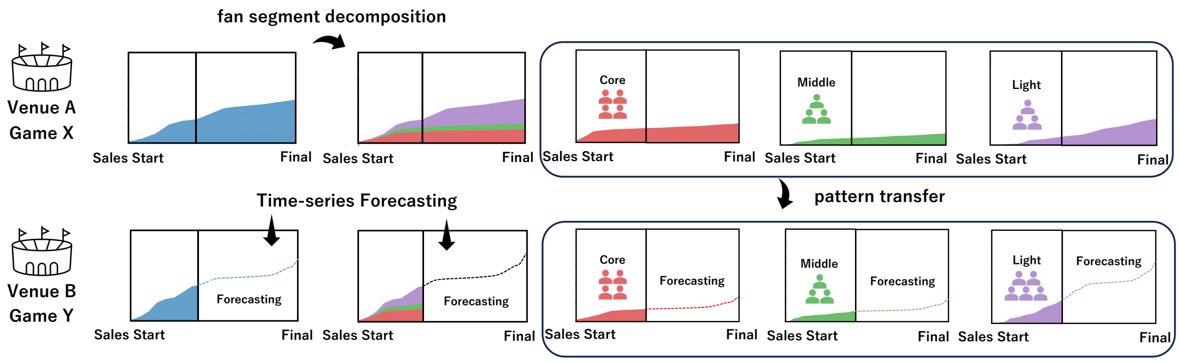

Figure 1: Overview of the proposed approach. While total sales patterns vary across venues, decomposing by fan segments reveals transferable patterns within certain segments.

time series forecasting problem, several challenges arise. First, the series length is short. In typical time series forecasting tasks such as stock prices, weather, and electricity demand, long-term historical data is often available. However, ticket sales periods are limited to approximately one month. Furthermore, when making forecasts early in the sales period, even fewer observations are available as input. For instance, when forecasting on the fifth day after sales begin, only five days of observations can be used as input. Thus, forecasts must be made from limited observations.

Additionally, recent years have seen a surge in new stadium construction and major renovations. This has increased demand for forecasting at venues with insufficient historical data. In such situations, leveraging data from existing venues becomes crucial. However, since venues differ in capacity and location, directly using total sales or occupancy rates has limitations. Due to such inter-venue differences, features based solely on total sales make it difficult to extract common patterns from existing venue data.

In this study, we formulate attendance forecasting as a daily time series forecasting problem to predict sales trends for the remaining period based on early-stage observations. In machine learning, how input data is represented significantly affects model performance, and appropriate representation design has proven effective across various domains. To address the challenges described above, we propose an approach based on input representation design, as illustrated in Figure 1. Specifically, we decompose total sales into fan segments based on past visit frequency. Since fan segments are classifications based on individual visit frequency, they are less dependent on venue characteristics. With total sales alone, patterns often differ across venues and games, making it difficult to transfer learned knowledge to new venues. However, by decomposing into fan segments, certain categories that are less venue-dependent exhibit similar patterns across venues, enabling pattern transfer for those categories. As a result, forecasting performance remains stable regardless of the target venue. This allows the model to efficiently learn common patterns even across different venues.

To evaluate the proposed approach, we conducted experiments in two stages. First, we applied existing time series forecasting methods to daily sales forecasting and compared them, then evaluated an LSTM combined with fan segment decomposition. The results showed that our approach achieved the highest accuracy. Next, we conducted comparison experiments varying input representations and input lengths. We compared decomposition methods—total sales, price category, fan segment, and their combinations—across different input lengths. Fan segment decomposition achieved the highest accuracy under almost all conditions.

The contributions of this paper are as follows:

- We propose fan segment decomposition as a venue-independent input representation that enables the extraction of transferable patterns across venues.
- We demonstrate that combining fan segment decomposition with sequential models (LSTM) achieves the highest forecasting accuracy.
- Through systematic experiments comparing multiple input representations and input lengths, we validate the effectiveness of the proposed approach for cross-venue generalization.

2 RELATED WORK

This section reviews existing research on sports attendance forecasting and clarifies the positioning of this study.

2.1 Sports Attendance Prediction

Sports event attendance forecasting has been studied for many years, with the primary focus on understanding demand levels for each game in advance (Schreyer and Ansari, 2022). These studies target final attendance or occupancy rates at the game level, emphasizing the estimation of overall event demand. Most conventional studies use external factors such as day of the week, opponent, team performance, stadium capacity, and weather as explanatory variables (Pang and Wang, 2024; Mueller, 2020). Combining these key determinants with machine learning models has enabled highly accurate forecasting of final attendance for each game.

However, these approaches have two limitations. First, they cannot capture demand fluctuations caused by unforeseen factors such as player injuries or promotional campaigns announced after sales begin. Second, they focus on final attendance and do not provide frameworks for forecasting sales trajectories during the sales period, which is essential for mid-sales decision-making such as pricing adjustments.

2.2 Forecasting the Sales Process from Early Sales Signals

The problem of forecasting future sales trajectories or final demand from early observations after sales begin has been widely studied in hotel reservations (Fiori and Foroni, 2019; Contessi et al., 2024) and airline ticket sales (Imanaka and Sakama, 2021; He et al., 2023). In these domains, frameworks using booking curves to model temporal changes in demand and estimate final demand from early booking status have been proposed. Furthermore, such forecasts are also utilized for sales optimization including dynamic pricing, and integrated research combining forecasting and decision-making has advanced (Gao et al., 2022; Chenavaz and Dimitrov, 2025; Yang et al., 2022).

Meanwhile, in sports, time series models for attendance forecasting have been reported (Pang, 2025), but most target transitions at the game level throughout the season. Research forecasting the sales process from sales start to game day at the daily level for each game remains limited. Therefore, forecasting frameworks applicable to mid-sales decision-making as described in the introduction have not been sufficiently established.

Additionally, sports ticket sales have unique challenges. The sales period is short with limited observation sequences, each game is an independent event with significantly varying demand structures, and sales patterns differ across venues. Due to these characteristics, directly applying methods that assume long-term, continuous reservation data like hotels or airlines is difficult, requiring input representation innovations for forecasting from short observation sequences.

2.3 Sales Composition as an Input Representation

To capture sales trends and make forecasts from short sequences, total sales alone are insufficient. Even with the same total sales volume, the sales process and future trajectory may differ depending on the customer segments or price tiers that comprise the sales.

Existing studies have conducted demand analysis based on spectator attributes and price categories (Popp et al., 2025). For example, studies classifying spectators into home and away fans (Humphreys et al., 2022), and studies evaluating demand by price tier (Quansah et al., 2024; Kaiser et al., 2019) have been reported. However, these studies focus on aggregate values at the game level, and research using sales composition as input to time series forecasting models is limited.

This study focuses on fan segment decomposition based on visit frequency as an input representation. Since visit frequency is an individual attribute independent of venue-specific factors, it is expected to capture purchasing patterns common across venues. We validate the effectiveness of this approach under short-sequence conditions and for cross-venue generalization.

3 METHOD

3.1 Problem Setting

This study addresses daily time series forecasting for sports ticket sales. We define the sales period as T days, with data from day 1 to day n as input and daily trends from day $n+1$ to day T as output. The model is trained on historical data from past games across multiple venues. During inference, given observed sales data from the early period of a target game, the model forecasts sales for the remaining days.

The forecasting target is the occupancy rate. The occupancy rate is defined as the number of tickets sold divided by the venue capacity:

$$d_t = \frac{s_t}{C} \quad (1)$$

where s_t is the number of tickets sold on day t and C is the venue capacity. The use of occupancy rate absorbs scale differences between venues, enabling unified handling of data from different venues.

The input includes not only the daily occupancy rate but also the cumulative occupancy rate from the start of sales. This allows for richer representation of sales progress. On the other hand, the output targets only the daily occupancy rate for each day. Furthermore, sales data is decomposed by category to capture detailed purchasing patterns. With K categories, the input has $2K$ dimensions (K categories $\times$ (daily + cumulative)), and the output has K dimensions. The input sequence $\mathbf{X}$ and output sequence $\mathbf{Y}$ are defined as:

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \in \mathbb{R}^{n \times 2K} \quad (2)$$

$$\mathbf{Y} = (\mathbf{y}_{n+1}, \mathbf{y}_{n+2}, \ldots, \mathbf{y}_T) \in \mathbb{R}^{(T-n) \times K} \quad (3)$$

The specific decomposition methods are described in Section 3.2, and the detailed design of input and output is described in Section 3.3.

3.2 Input Representations

This study compares four category decomposition methods for sales data: total sales ($K = 1$), price category ($K = K_p = 5$), fan segment ($K = K_f = 5$), and fan segment $\times$ price category ($K = K_f \times K_p = 25$).

Total sales is the simplest representation using only the overall occupancy rate ($K = 1$).

Price category is a representation that separates occupancy rate by ticket price category ($K_p = 5$). The definition of price categories is shown in Table 1. Complimentary tickets refer to free tickets distributed through campaigns or promotions. The number of price categories was set to match the number of fan segments ($K_p = K_f = 5$) to ensure a fair comparison between decomposition methods.

Table 1: Definition of price categories.

Price Category	Definition
Complimentary	Free tickets
Discount	Tickets offered at reduced prices
Low-price	3,000 JPY or less
Mid-price	3,000–5,000 JPY
High-price	5,000 JPY or more

Fan segment is a representation that classifies customers based on past visit frequency ($K_f = 5$). Specifically, each customer's fan segment is determined based on their visit frequency in the most recent 30 games prior to and including each target game. The definition of fan segments is shown in Table 2. This segmentation follows the classification used by the data provider in their marketing operations, reflecting a categorization commonly applied in practice. Since this five-category classification is grounded in practical marketing usage, we adopt it as the basis for setting the number of categories in this study.

Table 2: Definition of fan segments.

Fan Segment	Number of Past Visits
Super-core	20 or more
Core	10–19
Middle	5–9
Light	2–4
Super-light	1

Fan segment $\times$ Price category is a representation combining fan segments and price categories ($K = K_f \times K_p = 25$).

3.3 Model: Encoder-Decoder LSTM

We adopt an LSTM with Encoder-Decoder architecture as the model for time series forecasting using the input representations described in the previous section. This section describes its architecture and training methods. Figure 2 shows the overall structure of the model.

3.3.1 Architecture

This model consists of an Encoder that processes the input sequence during the observation period and a Decoder that generates the output sequence during the forecasting period.

Encoder The Encoder is an LSTM that processes the input sequence and compresses the entire sequence information into hidden states. The input $\mathbf{x}_t \in \mathbb{R}^{2K}$ on day t is constructed as a vec-

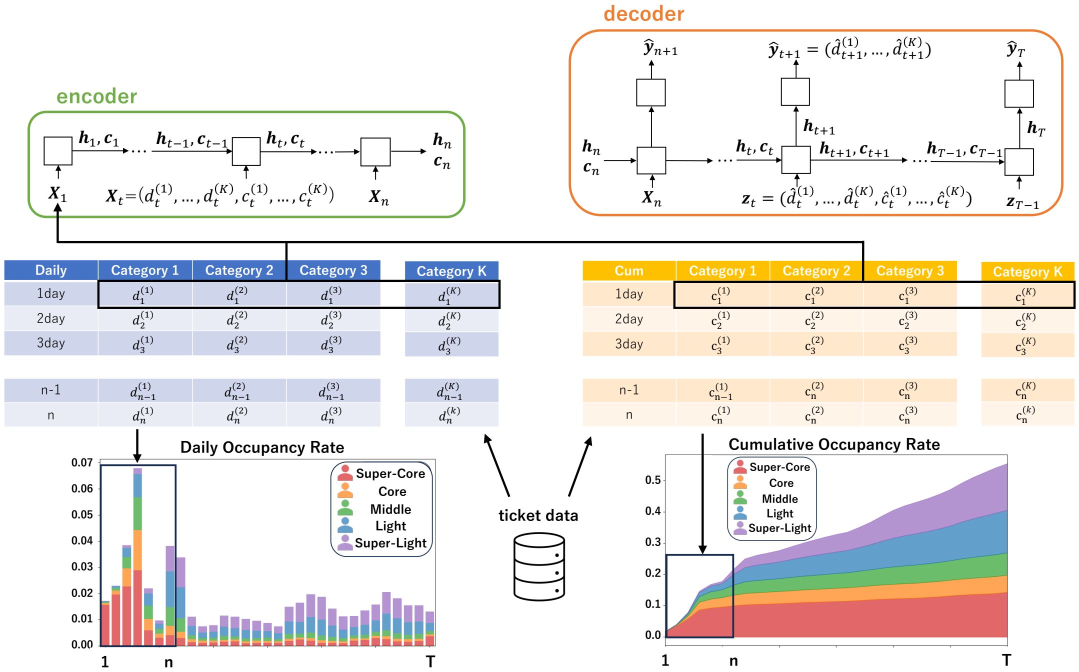

Figure 2: Architecture of the Encoder-Decoder LSTM model.

tor concatenating category-wise daily values and cumulative values:

$$\mathbf{x}_t = (d_t^{(1)}, \ldots, d_t^{(K)}, c_t^{(1)}, \ldots, c_t^{(K)}) \tag{4}$$

where $d_t^{(k)}$ is the daily occupancy rate of category k on day t, and $c_t^{(k)}$ is the cumulative occupancy rate. The cumulative occupancy rate is defined as:

$$c_t^{(k)} = \sum_{\tau=1}^{t} d_\tau^{(k)} \tag{5}$$

At each time step t, the LSTM cell performs the following update:

$$\mathbf{h}_t, \mathbf{c}_t = \mathrm{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}) \tag{6}$$

where $\mathbf{h}_t$ is the hidden state and $\mathbf{c}_t$ is the cell state. The final states $\mathbf{h}_n$ and $\mathbf{c}_n$ of the Encoder are passed to the Decoder.

Decoder The Decoder consists of an LSTM and a K-dimensional output layer, receiving the final state of the Encoder as its initial state and generating daily occupancy rates sequentially through $(T - n)$ steps of autoregression.

The input to each step of the Decoder is a $2K$-dimensional vector concatenating daily values and cumulative values, similar to the Encoder. By explicitly providing cumulative values, the model can understand the sales progress at each time point. The initial input to the Decoder uses the daily values and cumulative values of the last day of the observation period (day n).

The forecast for day t ($t > n$) is performed as follows. First, the input vector is constructed from the previous step's forecasts:

$$\mathbf{z}_t = (\hat{d}_{t-1}^{(1)}, \ldots, \hat{d}_{t-1}^{(K)}, \hat{c}_{t-1}^{(1)}, \ldots, \hat{c}_{t-1}^{(K)}) \tag{7}$$

where $\hat{d}_{t-1}^{(k)}$ is the daily occupancy rate of category k forecasted at the previous step, and $\hat{c}_{t-1}^{(k)}$ is the forecasted cumulative occupancy rate.

Next, the LSTM and output layer forecast the daily occupancy rate for day t:

$$\mathbf{h}_t, \mathbf{c}_t = \mathrm{LSTM}(\mathbf{z}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}) \tag{8}$$

$$\hat{\mathbf{y}}_t = \mathbf{W}_o \mathbf{h}_t + \mathbf{b}_o = (\hat{d}_t^{(1)}, \ldots, \hat{d}_t^{(K)}) \tag{9}$$

where $\mathbf{h}_t$ is the hidden state, $\mathbf{c}_t$ is the cell state, and $\mathbf{W}_o$ and $\mathbf{b}_o$ are parameters of the output layer.

Finally, the cumulative occupancy rate is updated from the forecasted daily occupancy rate:

$$\hat{c}_t^{(k)} = \hat{c}_{t-1}^{(k)} + \hat{d}_t^{(k)} \tag{10}$$

This cumulative value is used to construct the input $\mathbf{z}_{t+1}$ for the next step. By repeating the above procedure from $t = n + 1$ to $t = T$, daily occupancy rates for the entire forecasting period are generated sequentially.

3.3.2 Training

In autoregressive Decoders, forecasting errors propagate and accumulate through subsequent steps. This study introduces three techniques to address this challenge.

Scheduled Sampling. During training, at each step of the Decoder, the ground truth is used as the next input with probability p, and the previous step's forecast is used with probability $1-p$. This sampling is performed independently for each category. During inference, $p=0$ is used, relying entirely on forecasts. This enables the model to acquire robustness against error propagation.

Dual Loss Function. While the Decoder outputs daily occupancy rates, the cumulative occupancy rate (final sales volume) is also important in practice. Using only daily loss causes errors to accumulate in cumulative values, while using only cumulative loss weakens learning of daily patterns. Therefore, we use both daily loss and cumulative loss:

$$\mathcal{L}_{\text{daily}} = \frac{1}{T-n}\sum_{t=n+1}^{T} \|\hat{\mathbf{y}}_t - \mathbf{y}_t\|^2 \tag{11}$$

$$\mathcal{L}_{\text{cum}} = \frac{1}{T-n}\sum_{t=n+1}^{T} \|\hat{\mathbf{c}}_t - \mathbf{c}_t\|^2 \tag{12}$$

where $\hat{\mathbf{c}}_t$ and $\mathbf{c}_t$ are the forecasted and ground truth cumulative occupancy rate vectors, respectively.

Variance Normalization. Since daily values and cumulative values have different scales, a simple weighted sum would cause one loss to dominate. Therefore, we normalize each loss by the batch variance to make them dimensionless before integration:

$$\mathcal{L} = \alpha\frac{\mathcal{L}_{\text{daily}}}{\sigma^2_{\text{daily}}+\epsilon} + \beta\frac{\mathcal{L}_{\text{cum}}}{\sigma^2_{\text{cum}}+\epsilon} \tag{13}$$

where σ^2_{daily} and σ^2_{cum} are the within-batch variances, ϵ is a small constant for numerical stability, and α and β are weight parameters.

4 EXPERIMENTS

4.1 Dataset and Evaluation Setup

This study uses ticket sales data from home games of a professional sports team based in the Kanto region. The target games include only regular season games from four seasons, excluding preseason matches and postseason games (playoffs, etc.). During the target period, this team used seven venues (Venue A–Venue G) for home games. Table 3 shows the capacity of each venue and the breakdown of the number of games in each season. The total number of target games across four seasons is 105.

Table 3: Number of games and capacity for each venue.

Venue	22/23	23/24	24/25	25/26
A(10,000)	24	14	14	0
B(10,000)	0	0	0	16
C(10,000)	0	0	3	0
D(7,000)	0	0	4	0
E(3,250)	0	10	2	0
F(3,050)	5	6	5	0
G(2,500)	0	0	2	0
Total	29	30	30	16

Since the actual sales period varies across games, we aligned the number of days from the sales start date to a common time axis. All target games have sales periods longer than 20 days ($D > 20$), so we applied a piecewise rescaling to map the elapsed days d from sales start to $d' \in \{1, 2, \ldots, 30\}$. Specifically, to preserve early and late sales information: (1) the first 10 days after sales start ($d \leq 10$) are mapped as $d' = d$; (2) the last 10 days before sales end ($d > D-10$) are mapped as $d' = 20 + (d - (D-10))$; (3) the intermediate period ($10 < d \leq D-10$) is linearly rescaled to the range $d' = 11$ to 20. Since the intermediate period mapping yields continuous values, we rounded to the nearest integer for use as daily time series. This rescaling enables comparison of all games within a unified framework of "first 10 days + middle 10 days + last 10 days."

To evaluate forecasting performance at new venues, this study adopts Leave-One-Group-Out (LOGO) cross-validation. Let the set of venues in the dataset be $\mathcal{V} = \{v_1, v_2, \ldots, v_M\}$. In LOGO, the following procedure is repeated for each venue v_j: (1) use all games at venue v_j as test data; (2) use games at remaining venues $\mathcal{V} \setminus \{v_j\}$ as training data; (3) train the model on training data and evaluate on test data. This evaluation method measures generalization performance to venues never observed during training, simulating practical situations of forecasting at venues without historical data, such as new stadiums. Note that this cross-validation scheme does not account for temporal ordering across seasons; training data may include games from later seasons than the test data. This design choice prioritizes evaluating cross-venue generalization over temporal generalization.

4.2 Model Comparison

This section compares multiple methods to evaluate the combination of the proposed input representation (fan segment decomposition) and forecasting models. The input length is fixed at 5 days, with the forecasting target being the 25 days from day 6 to day 30.

4.2.1 Setup

To identify suitable models for the short-sequence conditions specific to sports ticket sales, we compare multiple forecasting methods. Prior studies have compared various neural network architectures for time series forecasting (Yamak et al., 2019; Yunita et al., 2025; He et al., 2023). Following these comparison frameworks, we evaluate statistical methods including Moving Average (MA) and Additive Pickup, as well as deep learning models including ANN (multilayer perceptron), GRU, Transformer, and LSTM. MA forecasts future values using the average of past observations. Additive Pickup is a classical method used in airline and hotel revenue management, which forecasts future sales based on historical pickup patterns. ANN processes the input as a flat vector without explicitly modeling temporal structure. In contrast, GRU and LSTM are sequential models that process input step by step while maintaining hidden states, enabling them to learn temporal dependencies. Transformer uses self-attention mechanisms to learn relationships between all positions in a sequence, but requires large amounts of data for effective training. For deep learning models, we evaluate both settings: using only total sales occupancy rate as input, and using occupancy rates decomposed by fan segments as input.

MAE and RMSE on cumulative occupancy rate are used as evaluation metrics. For models that forecast by fan segment, category-wise forecasts are summed to obtain total sales for comparison. Cumulative occupancy rates are constructed by accumulating daily forecasts from the forecasting start day onward. Errors are calculated for each venue and averaged across all venues to eliminate the influence of differences in the number of games.

4.2.2 Results

Table 4 shows the results. In addition to errors on the cumulative curve over the entire forecasting period (All), we also evaluate errors on the cumulative occupancy rate on the final day (day 30) (Final). The final day metrics represent the forecasting accuracy of the ultimate sales volume on the game day. LSTM with fan segment decomposition achieved the lowest errors for both All and Final metrics.

For LSTM and GRU, fan segment decomposition reduced errors compared to using only total sales. Sequential models can effectively learn the temporal evolution of purchasing patterns for each fan segment. Furthermore, by decomposing into fan segments, the model can efficiently learn patterns common across venues, contributing to improved generalization performance.

In contrast, Transformer showed poor accuracy regardless of the input representation. Transformer models generally require large amounts of data for effective training. However, in the sports domain, the number of games per year is inherently limited to tens or at most a few hundred, which represents a structural constraint. Therefore, sequential models such as LSTM and GRU, which can learn effectively from smaller datasets, are more suitable for sports ticket sales forecasting. Based on these results, this study adopts the combination of LSTM and fan segment decomposition, and conducts detailed analysis in the following sections.

4.3 Comparison of Input Representations and Input Lengths

This section compares four types of input representations using LSTM while varying the input length. The input representations compared are: total sales ($K = 1$), price category ($K = K_p = 5$), fan segment ($K = K_f = 5$), and fan segment × price category ($K = K_f \times K_p = 25$). The input length is varied from 1 to 29 days, and forecasting accuracy is evaluated for each condition. The evaluation metric is MAE on cumulative occupancy rate, as in Section 4.2.

4.3.1 Results

Table 5 shows the MAE for each input representation with various input lengths. Early-stage forecasting is important in practice and represents

Table 4: Comparison of forecasting models. All: metrics on cumulative curve over entire forecasting period. Final: metrics on final day cumulative occupancy rate. MAE and RMSE are shown in percentage points.

Model	Input	All		Final	
		MAE (%)	RMSE (%)	MAE (%)	MAPE (%)
MA	Total	31.87	38.19	63.32	87.31
Additive Pickup	Total	8.32	9.95	11.73	16.70
ANN	Total	7.29	8.66	9.72	14.47
ANN	Fan segment	7.68	9.21	9.87	15.10
GRU	Total	7.58	9.12	11.71	16.68
GRU	Fan segment	7.12	8.35	8.41	12.89
Transformer	Total	11.10	13.05	19.57	27.42
Transformer	Fan segment	20.13	23.24	31.59	44.17
LSTM	Total	6.80	8.17	10.67	15.34
LSTM	Fan segment	**6.65**	**7.86**	**8.07**	**11.76**

a challenging condition due to limited observed data. Fan segment decomposition achieved the best accuracy in 7 out of 9 conditions. Figure 3 shows the rank distribution of each input representation across all conditions. Fan segment decomposition achieved rank 1 in 62.1% of conditions, and the combined rate of ranks 1–2 reached 93.1%. In contrast, Price×Fan decomposition ranked 4th (last) in 89.7% of conditions, consistently showing the lowest accuracy. Price category decomposition ranked 3rd in 69.0% of conditions, showing no significant improvement compared to total sales.

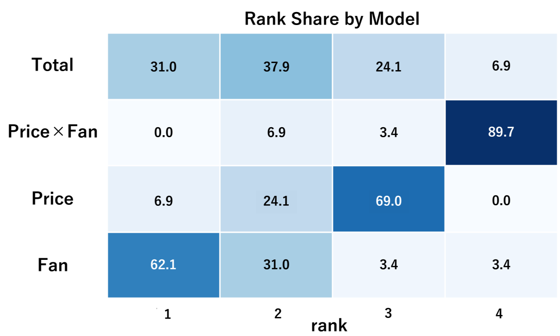

Figure 3: Distribution of ranks achieved by each input representation across input lengths 1–29 days. Fan segment decomposition achieved rank 1 in 62.1% of conditions.

These results confirm that fan segment decomposition achieves consistently high forecasting accuracy regardless of input length. Price category decomposition showed no improvement over total sales. Each venue has a different composition of price tiers, and the same price category can account for vastly different proportions of occupancy rate across venues. This makes it difficult to extract common patterns across venues. Price×Fan decomposition performed worst. While finer decomposition increases information, 25 categories are excessive for 105 games, leading to insufficient data per category. Additionally, including price tiers inherits the same venue-dependent structure, further hindering cross-venue pattern extraction.

4.3.2 Analysis

This section analyzes why fan segment decomposition achieved higher forecasting accuracy than price category decomposition.

To evaluate the variation in occupancy rates across venues for each category, we calculated the coefficient of variation (CV). Table 6 shows the results. The average CV for fan segments is 27.5%, while that for price categories is 55.6%, approximately twice as high. Notably, the CV for the Middle fan segment is only 7.7%, indicating very small variation across venues.

This difference is attributed to the nature of each decomposition method. Fan segments are classifications based on individual visit frequency and do not depend on venue characteristics. Enthusiastic fans tend to purchase early regardless of the venue, while casual fans tend to purchase closer to the game day. Such purchasing behavior patterns are based on individual attributes and are therefore common across venues.

In contrast, price categories depend heavily on the pricing strategy of each venue. Since complimentary ticket distribution policies and price tier compositions differ by venue, occupancy rate patterns vary significantly across venues even for the same price category. For example, one venue may sell many low-price tickets, while another venue may adopt a sales strategy centered on high-price tiers.

Table 5: MAE (%) for each input representation with various input lengths. Bold indicates the best performance for each input length.

Input representation	1 day	2 days	3 days	4 days	5 days	10 days	15 days	20 days	25 days
Total sales	10.8	11.6	9.2	8.5	9.0	5.7	3.5	**1.1**	0.3
Price×Fan	10.0	10.7	10.5	10.3	10.2	8.3	5.3	2.2	0.5
Price category	10.3	**9.4**	8.6	8.3	8.1	5.6	4.0	1.3	0.4
Fan segment	**9.7**	10.5	**7.9**	**7.5**	**7.5**	**5.3**	**3.3**	1.2	**0.3**

Table 6: Coefficient of variation (CV) of average occupancy rate across venues for each category.

Decomposition	Category	CV (%)
Fan segment	Super-core	47.5
Fan segment	Core	28.6
Fan segment	Middle	7.6
Fan segment	Light	18.1
Fan segment	Super-light	35.7
Fan segment average		**27.5**
Price category	Complimentary	44.7
Price category	Discount	58.2
Price category	Low-price	79.8
Price category	Mid-price	43.2
Price category	High-price	51.9
Price category average		**55.6**

In LOGO evaluation, the test venue is never observed during training. With price category decomposition, the large variation in occupancy rates across venues makes it difficult to apply knowledge learned from training data to new venues. In contrast, fan segment decomposition shows smaller variation across venues, enabling the model to learn transferable patterns and achieve high generalization performance even for unknown venues.

5 CONCLUSION

In this study, we formulated sports ticket sales as a daily time series forecasting problem. We proposed fan segment decomposition based on visit frequency as an input representation to extract purchasing patterns common across venues. The proposed approach was combined with an Encoder-Decoder LSTM and evaluated using LOGO cross-validation on data from 105 games across 7 venues.

In our experiments, we first compared existing time series forecasting methods to identify suitable models under the short-sequence conditions specific to sports ticket sales. The results confirmed that LSTM achieved the highest accuracy. Next, we compared multiple decomposition methods to verify input representations that generalize across venues. Fan segment decomposition consistently showed higher accuracy than total sales, price category, and fan segment × price category. This suggests that decomposing by fan segments enables the model to learn purchasing patterns common across venues for certain segments, resulting in stable forecasting even for unseen venues.

However, this study has several limitations. Category-wise error analysis revealed that the Light and Super-light segments had larger forecasting errors compared to other fan segments. These segments include many attendees who receive complimentary tickets, and the distribution of complimentary tickets depends on operational decisions. This makes forecasting from past sales trends alone difficult. Additionally, this study used data from a single team only. Since fan segmentation based on visit frequency is a universal concept applicable to any sport, validation across other teams and sports is a promising direction for future work.

Future research directions include incorporating time-series factors such as operational information (e.g., complimentary ticket distribution timing) and price changes during the sales period. This would enable the model to account for factors that were difficult to predict from past sales trends alone, potentially improving forecasting accuracy. In practical applications, it is important to update forecasts iteratively as observation data accumulates during the sales period. This enables high-accuracy forecasting based on demand fluctuations that cannot be captured by external factors alone. Furthermore, such an approach could contribute to practical decision-making, including determining optimal timing for complimentary ticket distribution and identifying appropriate phases for price adjustments.

ACKNOWLEDGEMENTS

This work was partly supported by JST CREST (JPMJCR21F2, JPMJCR22M4), Cabinet Office SIP3 (JPJ012495), and JSPS KAKENHI (22H03580, 22K18422). We also thank Toyota Systems Corporation for providing the data.

REFERENCES

Chang, Y.-M., Potter, J. M., and Sanders, S. (2016). Inelastic sports ticket pricing, marginal win revenue, and firm pricing strategy: A behavioral pricing model. *Managerial Finance*, 42(9):922–927.

Chenavaz, R. Y. and Dimitrov, S. (2025). Artificial intelligence and dynamic pricing: a systematic literature review. *Journal of Applied Economics*, 28(1):2466140.

Cisyk, J. and Courty, P. (2021). Stadium giveaway promotions: How many items to give and the impact on ticket sales in live sports. *Journal of Sport Management*, 35(6):551–565.

Coates, D. and Humphreys, B. R. (2005). Novelty effects of new facilities on attendance at professional sporting events. *Contemporary Economic Policy*, 23(3):436–455.

Contessi, D., Viverit, L., Pereira, L. N., and Heo, C. Y. (2024). Decoding the future: Proposing an interpretable machine learning model for hotel occupancy forecasting using principal component analysis. *International Journal of Hospitality Management*, 121:103802.

Courty, P. (2003). Some economics of ticket resale. *Journal of Economic Perspectives*, 17(2):85–97.

Fiori, A. M. and Foroni, I. (2019). Reservation forecasting models for hospitality smes with a view to enhance their economic sustainability. *Sustainability*, 11(5):1274.

Gao, J., Le, M., and Fang, Y. (2022). Dynamic air ticket pricing using reinforcement learning method. *RAIRO-Operations Research*, 56(4):2475–2493.

He, H., Chen, L., and Wang, S. (2023). Flight short-term booking demand forecasting based on a long short-term memory network. *Computers & Industrial Engineering*, 186:109707.

Humphreys, B. R., Reade, J. J., Schreyer, D., and Singleton, C. (2022). Separating the crowds: Examining home and away attendances at football matches. In *Essays on Sports Economics in Memory of Stefan Kesenne*. University of Oviedo Publishing.

Imanaka, K. and Sakama, C. (2021). Predicting air ticket demand using deep neural networks. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2901–2908. IEEE.

Kaiser, M., Ströbel, T., Woratschek, H., and Durchholz, C. (2019). How well do you know your spectators? a study on spectator segmentation based on preference analysis and willingness to pay for tickets. *European Sport Management Quarterly*, 19(2):178–200.

Leslie, P. and Sorensen, A. (2013). Resale and rent-seeking: An application to ticket markets. *The Review of Economic Studies*, 81(1):266–300.

Mueller, S. Q. (2020). Pre-and within-season attendance forecasting in major league baseball: A random forest approach. *Applied Economics*, 52(41):4512–4528.

Pang, Y. (2025). Time-series forecasting in sports: Using lstm and gru for stadium attendance prediction. *Physical Culture and Sport*, 111(1):25–35.

Pang, Y. and Wang, F. (2024). Forecasting stadium attendance using machine learning models: A case of the national football league. *Studia sportiva*, 18(2).

Popp, N., Greenwell, T. C., Cocco, A. R., and Bonney, N. (2025). What do ticket prices reveal about the growth in women's sport? the case of ncaa division i women's college basketball. *Sport Marketing Quarterly*, 34(1):3–15.

Quansah, T. K., Buraimo, B., and Lang, M. (2024). Determining the price of football: an analysis of matchday ticket prices in the english premier league. *European Sport Management Quarterly*, 24(3):764–784.

Sahni, N. S., Zou, D., and Chintagunta, P. K. (2017). Do targeted discount offers serve as advertising? evidence from 70 field experiments. *Management Science*, 63(8):2688–2705.

Schreyer, D. and Ansari, P. (2022). Stadium attendance demand research: A scoping review. *Journal of Sports Economics*, 23(6):749–788.

Shapiro, S. L. and Drayer, J. (2012). A new age of demand-based pricing: An examination of dynamic ticket pricing and secondary market prices in major league baseball. *Journal of Sport Management*, 26(6):532–546.

Woratschek, H., Horbel, C., and Popp, B. (2014). The sport value framework – a new fundamental logic for analyses in sport management. *European Sport Management Quarterly*, 14(1):6–24.

Yamak, P. T., Yujian, L., and Gadosey, P. K. (2019). A comparison between arima, lstm, and gru for time series forecasting. In *Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence*, pages 49–55.

Yang, Y., Chu, W.-L., and Wu, C.-H. (2022). Learning customer preferences and dynamic pricing for perishable products. *Computers & Industrial Engineering*, 171:108440.

Yunita, A., Pratama, M. I., Almuzakki, M. Z., Ramadhan, H., Akhir, E. A. P., Mansur, A. B. F., and Basori, A. H. (2025). Performance analysis of neural network architectures for time series forecasting: A comparative study of rnn, lstm, gru, and hybrid models. *MethodsX*, 15:103462.