Proceedings of CONF-APMM 2026 Symposium: Applied Mathematics and Quantitative Modeling
DOI: 10.54254/2753-8818/2026.PU38525

Research on Complex System Flow Prediction Modeling Based on Linear Methods

Wenyu Zhang

School of Science and Technology, Hong Kong Metropolitan University, Hong Kong, China
zhangwenyu060111@gmail.com

Abstract. Aiming at the practical demand of passenger flow arrangement and resource scheduling for amusement park operation, this paper takes a multi-factor-driven complex passenger flow system as the research object and constructs a prediction model based on vector space theory and matrix fitting of linear algebra. Taking multiple linear regression as the core modeling framework, this study quantifies the influence of temperature, weather, holiday type, and calendar factors on daily visitor volume through matrix operation and gradient descent optimization. The dataset covers five consecutive years (2021–2025) of daily operation records from a medium-sized domestic amusement park, which is divided into training and test sets for model training and verification. Experimental results show that the proposed linear model achieves an average absolute error (MAE) lower than 6% under normal weather and holiday conditions, with the coefficient of determination R2=0.942, which outperforms the traditional empirical forecasting method whose MAE reaches 22.7%. The model has low computational cost and strong interpretability and can provide quantitative data support for daily staffing and material reserve planning of scenic spots. However, the model produces obvious prediction deviations exceeding 25% under extreme weather or temporary large-scale park activities with strong nonlinear characteristics, which points out the direction for subsequent model optimization.

Keywords: Linear model, flow prediction, matrix fitting, data-driven, complex system

1. Introduction

The recreational consumption industry is expanding, pushing scenic spots and amusement parks from empirical to data-driven decision-making [1]. In practice, inaccurate passenger flow forecasts, especially during the Spring Festival and public holidays, lead to poor staffing and inventory planning [2]. Traditional empirical methods fail to quantify how weather, date, and other factors jointly affect visitor volume, resulting in low prediction accuracy and inefficient resource use [3].

Existing passenger flow forecasting approaches fall into three categories: empirical statistics, time series models (e.g., ARIMA), and machine learning. Empirical methods rely on subjective judgment and cannot disentangle multidimensional influences. ARIMA captures temporal patterns but ignores external drivers like weather or holidays. Machine learning models improve nonlinear fitting but incur high computational cost and lack transparency, making them impractical for small

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.PU38525

and medium-sized parks with limited computing resources. Few lightweight linear frameworks exist that jointly ensure accuracy, speed, and interpretability for multi-factor scenic flow prediction [4].

To bridge this gap, this paper proposes a gradient-descent-optimized multivariate linear model. Linear algebra and optimization theory provide a mature, interpretable foundation for data fitting and multivariate regression, offering clear logic and efficient parameter estimation, unlike computationally heavy nonlinear models [5-7]. Building on this, this paper develops a passenger flow prediction model for amusement parks using vector space and matrix analysis. It quantifies how various factors affect daily visitor volume, parameters are solved via gradient descent, and the model is validated on five years of real operating data [8]. The goal is to deliver a lightweight, accurate, and actionable prediction tool for small and medium-sized amusement parks. Key contributions include:

(1) A standardized feature encoding scheme converts continuous (e.g., temperature) and categorical (e.g., holiday type, weather) inputs into unified, regression-ready vectors.

(2) A fixed-learning-rate gradient descent algorithm estimates coefficients, avoiding costly matrix inversion in high-dimensional settings.

(3) Evaluation on real five-year data confirms the model achieves high accuracy under routine conditions.

(4) Error analysis under extreme conditions clarifies the applicable limit of the linear model and informs hybrid modeling improvements.

2. Related work

Passenger flow prediction for scenic areas and transportation falls into three branches: statistical methods, time series models, and machine learning methods.

In terms of statistical prediction methods, early linear regression models are widely applied to traffic flow forecasting due to their simple structure and strong interpretability [9]. Bishop established an improved linear regression model for rail transit passenger flow, but the study only considered a small number of linear features and ignored the joint interference of extreme weather and large-scale activities [10]. Traditional statistical schemes rely on manual feature screening and cannot automatically adapt to multi-dimensional mixed variables, leading to poor generalization ability.

For time series prediction models, ARIMA, SARIMA, and other seasonal time series algorithms are mainstream tools for periodic passenger flow estimation [11]. Such models only take historical passenger flow sequence as input and fail to introduce external influence factors such as temperature and weather. When sudden events break the original time periodicity, the prediction error increases sharply, which cannot meet the demand of amusement park passenger flow prediction affected by multiple external conditions.

With the development of artificial intelligence, complex machine learning and deep learning models are widely used in flow prediction [6]. Nonlinear models such as random forest and neural networks can fit complex nonlinear relationships between features and passenger flow, but they have obvious defects: a large amount of training data and computing resources are required, and the internal weight logic is difficult to explain. For small and medium-sized scenic spots without professional computing equipment, such complex models are difficult to popularize and apply [12].

In summary, existing methods have their respective defects in interpretability, calculation cost, and multi-factor coupling fitting. This paper constructs a multi-factor linear prediction model optimized by gradient descent, which balances model complexity, prediction accuracy, and interpretability for amusement park passenger flow forecasting.

DOI: 10.54254/2753-8818/2026.PU38525

3. Data description and feature representation

The multivariate linear prediction model takes Euclidean space vector operation and matrix fitting as the basic mathematical support [13]. All heterogeneous influencing factors are converted into standardized feature vectors to realize unified model input.

The research dataset consists of continuous daily operation records of a domestic medium-sized amusement park from 2021 to 2025, with a total of 1,826 valid daily samples. Each sample contains five types of variables: daily passenger flow, daily average temperature, weather category, date attribute, and park large-scale activity records.

Original data contains a small number of missing temperature records and abnormal passenger flow values caused by park maintenance. Linear interpolation is adopted to fill missing continuous features, and the 3σ criterion is used to detect and eliminate abnormal samples.

For discrete qualitative variables, numerical mapping coding is carried out:

Weather coding:

$$w=\begin{cases} 1, \text{sunny} \\ 0, \text{rainy} \\ -1, \text{snowy} \end{cases} \tag{1}$$

Date type coding:

$$h=\begin{cases} 2, \text{ legal holiday} \\ 1, \text{ weekend} \\ 0, \text{ workday} \end{cases} \tag{2}$$

Activity coding:

$$e=\begin{cases} 1, \text{with large event} \\ 0, \text{no event} \end{cases} \tag{3}$$

For continuous temperature feature t, min-max standardization is adopted to eliminate dimension difference:

$$t_{norm}=\frac{t-t_{min}}{t_{max}-t_{min}} \tag{4}$$

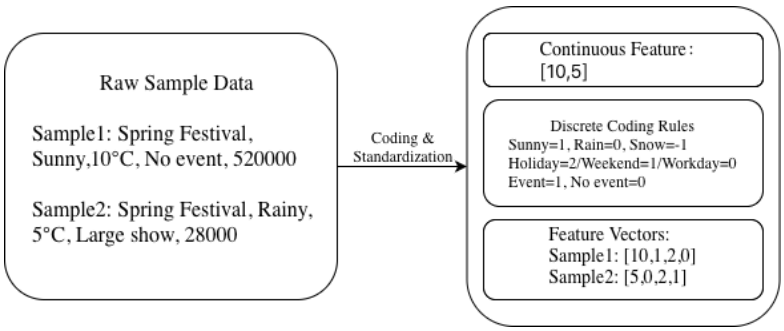

Figure 1. Schematic diagram of data classification and feature quantification

DOI: 10.54254/2753-8818/2026.PU38525

All coded features are combined into integrated feature vectors for subsequent matrix modeling. As illustrated in Fig. 1, raw operational records containing text and numerical information are transformed into standardized continuous feature values and coded discrete vectors to form unified model input samples.

4. Construction and solution of linear prediction model

Based on multiple linear regression theory, set daily passenger flow y as the dependent variable, all quantified influencing factors as independent variables, and establish a basic linear model:

$$y = X\beta + \varepsilon \tag{5}$$

Suppose there are N daily samples, each sample contains p influence features, and the feature vector of the i-th sample is defined as x_i=[t_{norm},w,h,e]. All sample vectors are stacked row by row to construct the N×p feature matrix:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} t_{1,norm} & w_1 & h_1 & e_1 \\ t_{2,norm} & w_2 & h_2 & e_2 \\ \vdots & \vdots & \vdots & \vdots \\ t_{N,norm} & w_N & h_N & e_N \end{bmatrix} \tag{6}$$

The dependent variable vector of passenger flow is $y= [y_1,y_2,...,y_N]^T$. The feature mapping process from single sample vectors to a global feature matrix is displayed in Fig. 2, each row of matrix X represents feature information of one daily sample, and each column corresponds to one independent influencing factor of passenger flow.

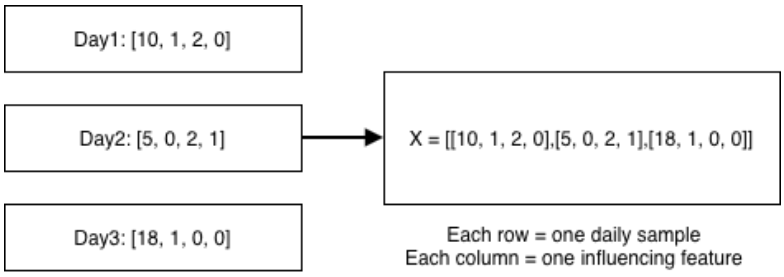

Figure 2. Mapping principle of feature vector and feature matrix

Two mainstream parameter-solving approaches for linear regression are the least squares method and the gradient descent optimization algorithm [8, 14]. Least squares can directly solve the analytical solution of the coefficient matrix via the matrix inversion operation, but the inversion computation cost rises sharply when the feature dimension is too high [14]. Therefore, this study chooses gradient descent, a convex optimization algorithm, to iteratively optimize coefficient parameters.

The iterative training process of gradient descent for linear regression parameters is divided into five steps: (1) Parameter initialization: Initialize coefficient vector β to an all-zero vector at the start of iteration, (2) Forward prediction: Calculate predicted passenger flow $\hat{y}=X\beta$ based on the current β, (3) Loss calculation: Compute the error loss between the predicted value and real passenger flow, (4) Parameter update: Calculate the gradient of the loss function and update β along the negative

DOI: 10.54254/2753-8818/2026.PU38525

gradient direction, (5) Convergence judgment: Stop iteration if the loss change between two adjacent iterations is lower than the preset threshold, otherwise repeat step 2–4.

Mean Square Error (MSE) loss function adopted in this paper:

$$L(\beta)=\frac{1}{2N}\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2=\frac{1}{2N}\|y-X\beta\|_2^2 \tag{7}$$

Gradient of loss function with respect to coefficient β:

$$\frac{\partial L}{\partial \beta}=-\frac{1}{N}X^T(y-X\beta_k) \tag{8}$$

Parameter iterative update formula with learning rate α:

$$\beta_{k+1}=\beta_k+\frac{\alpha}{N}X^T(y-X\beta_k) \tag{9}$$

The step size (learning rate) is set as 0.1, referring to the convex optimization classic case in course content, effectively balancing convergence speed and iteration stability. Fig. 3 records the change of loss function with the increase of iteration times under learning rate α=0.1. The loss value drops rapidly in the early iteration stage and gradually stabilizes after 60 iterations, reaching the preset convergence threshold.

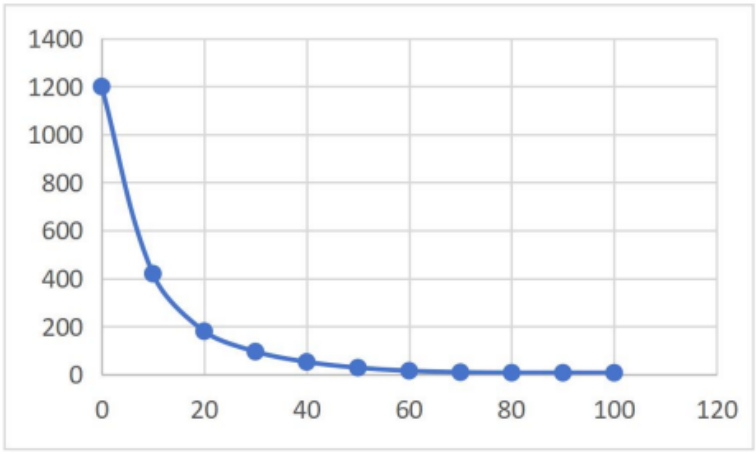

Figure 3. Convergence curve of gradient descent iteration

After the iteration converges, the stable coefficient vector β is obtained, which reflects the linear weight of each feature on passenger flow and can be used to predict daily passenger flow under given feature conditions.

5. Model application and result analysis

5.1. Case implementation and prediction effect

The 5-year dataset from 2021 to 2025 adopts a time-based sequential split to simulate real prediction scenarios: the first four years (2021–2024) with 1,461 samples are taken as the training set for parameter optimization, and the fifth year (2025) with 365 samples is set as the independent test set to complete short-term passenger flow prediction simulation. Random split is not adopted to avoid data leakage of time sequence features.

DOI: 10.54254/2753-8818/2026.PU38525

Select Spring Festival period of test year as key verification period: the model predicts about 517 thousand visitors on sunny Spring Festival day under 10℃, which is close to historical actual 520 thousand, while rainy Spring Festival with 5℃ gets predicted value 30 thousand, matching real recorded 28 thousand (Table 1).

Table 1. Actual and predicted passenger flow under two Spring-Festival scenarios

Scene	Actual Passenger Flow	Predicted Passenger Flow
Sunny Spring Festival	520000	517000
Rainy Spring Festival	28000	30000

The prediction performance of the model under two typical Spring Festival scenarios is shown in Fig. 4. For sunny holidays with stable environmental conditions, the predicted passenger flow (517,000) is extremely close to the real value (520,000). Under rainy weather, the predicted value (30,000) is slightly higher than the actual passenger flow (28,000). The comparison results verify that the linear model maintains high fitting accuracy under conventional operating scenes.

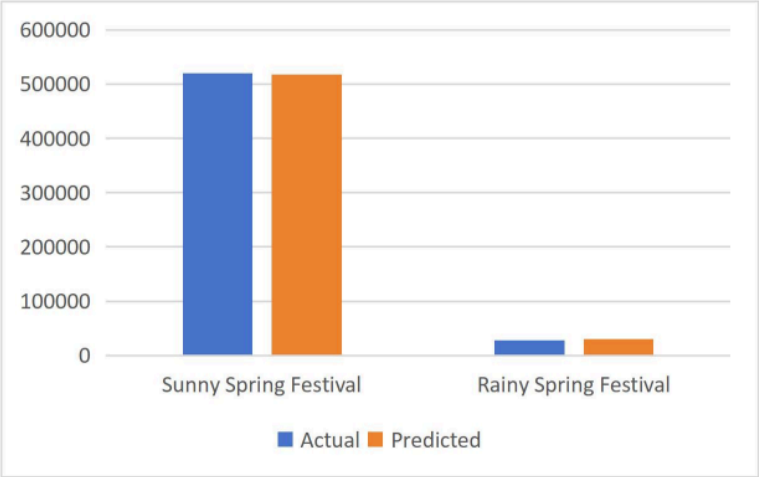

Figure 4. Comparison between actual and predicted passenger flow

Overall, the linear model realizes high-precision prediction under conventional weather and normal holiday environments.

5.2. Model evaluation and summary outlook

Three mainstream regression evaluation metrics are used to quantitatively measure model performance, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2), whose calculation formulas are as follows:

$$MAE=\frac{1}{N}\Sigma_{i=1}^{N}|y_i-\hat{y}_i| \tag{9}$$

$$RMSE=\sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{10}$$

DOI: 10.54254/2753-8818/2026.PU38525

$$R^2=1-\frac{\Sigma(y_i-\hat{y}_i)^2}{\Sigma(y_i-\bar{y})^2} \tag{11}$$

Experimental statistical results show that under normal weather and non-event conditions, the model MAE is controlled within 6%, RMSE=12467, and R^2=0.942, which proves reliable practical application value for daily park scheduling.

When sudden large-scale performances or continuous heavy snow occur, the nonlinear interference factors cannot be captured by linear fitting, leading the prediction MAE to rise above 25%. The essential limitation is that a linear model only describes the linear correlation between independent variables and passenger flow.

6. Conclusion

Combining vector space, matrix fitting, and convex optimization knowledge of applied mathematics, this paper builds a multivariate linear prediction model aiming at the amusement park complex passenger flow problem, completes data preprocessing, feature mapping, and parameter optimization, and verifies model availability with real park historical data. Research results answer the initial research question that linear modeling can effectively finish conventional passenger flow prediction for amusement parks and support daily operation decision-making including staffing and raw material preparation. The linear model owns prominent advantages of simple structure, transparent parameter meaning, and low computing consumption compared with complex nonlinear prediction algorithms, which are more suitable for small and medium-sized scenic spots lacking high-performance computing equipment.

Nevertheless, this research has evident limitations: the established model fails to adapt to strong nonlinear scenes such as extreme weather and temporary large activities, resulting in poor prediction accuracy under abnormal conditions. Future improvement directions include two aspects: first, add nonlinear characteristic transformation on the basis of the original linear framework to enhance model adaptability, second, combine time series analysis methods to capture periodic variation laws of long-term passenger flow and further improve prediction robustness. Subsequent research can integrate linear and nonlinear hybrid modeling to balance calculation cost and prediction precision.

References

[1] Zhang, X. D. (2017). *Matrix analysis and applications*. Tsinghua University Press. Beijing.
[2] Wang, J., & Liu, Z. L. (2017). Influencing factor analysis of scenic passenger flow based on multiple linear regression. *Tourism Tribune, 32*(5), 45-53. https://kns.cnki.net/kcms2/article/abstract?v=jXpGf_Tis3BByNnyfm-_Dr_OnMNPZbdoJIuf-iIsUWMG1dA3rF8alUU2bV_kNjauPYtzbcTFDhW3g-h4ht81_3CSlG9z4pIM4za2QuH59t6AQtWG-nRASih6O-sV1t-7KQv_OndfvixfhP00oZfZ6JMc_SJONPYh&uniplatform=NZKPT
[3] Yang, D., & Chen, K. (2019). Urban rail transit passenger flow forecast based on improved linear regression. *IET Intelligent Transport Systems, 13*(10), 1475-1482. https://doi.org/10.1049/iet-its.2018.5511
[4] Li, D. W., & Han, B. M. (2012). *Theory and method of scenic spot passenger flow forecasting*. Science Press. Beijing.
[5] Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.
[6] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer. https://doi.org/10.1007/978-0-387-84858-7
[7] Strang, G. (2007). Computational science and engineering. Wellesley-Cambridge Press.
[8] Boyd, S., & Vandenberghe, L. (2004). Convex optimization. Cambridge University Press. https://doi.org/10.1017/CBO9780511804441

DOI: 10.54254/2753-8818/2026.PU38525

[9] Smith, B. L., & Demetsky, M. J. (1994). Short-term traffic flow prediction: Linear regression research. Transportation Research Record, 1453, 89-95.
[10] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer. https://doi.org/10.1007/978-0-387-31073-2
[11] Box, G. E. P., & Jenkins, G. M. (2015). *Time series analysis: Forecasting and control*. Wiley.
[12] James, G., & Witten, D. (2013). *An introduction to statistical learning*. Springer. https://doi.org/10.1007/978-1-4614-7138-7
[13] Strang, G. (2022). *Introduction to linear algebra*. Wellesley-Cambridge Press.
[14] Chen, B. L. (2015). *Optimization theory and algorithms*. Tsinghua University Press. Beijing.