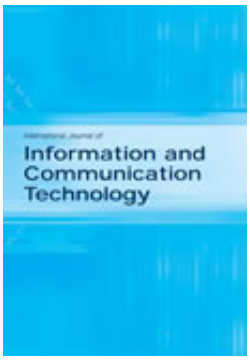
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Prediction of uncertain passenger flow in scenic spots by fusing multi-source data and integrated learning

Jingwen Xu, Qingshan Xiao, Shuo Xiong

DOI: 10.1504/IJICT.2026.10075961

Article History:

Received:	15 September 2025
Last revised:	22 November 2025
Accepted:	23 November 2025
Published online:	09 February 2026

Prediction of uncertain passenger flow in scenic spots by fusing multi-source data and integrated learning

Jingwen Xu, Qingshan Xiao* and Shuo Xiong

School of Eco-Culture and Eco-Tourism,
Hunan Vocational College Engineering Department,
Changsha, 410151, China
Email: lvyouzhiguang2025@163.com
Email: 15574886295@163.com
Email: xs147037728@163.com
*Corresponding author

Abstract: Accurate scenic spot traffic prediction is of great significance for the optimal allocation of tourism resources and safety management. Aiming at the shortcomings of traditional methods in coping with data multi-source and prediction uncertainty, this study proposes an uncertainty prediction framework that integrates multi-source data and integrated learning. By integrating heterogeneous data from multiple sources, such as historical passenger flow, meteorology, web search and spatial features, a heterogeneous integrated model based on random forest, XGBoost and long short-term memory (LSTM) is constructed, and quantification of uncertainty is realised by combining quantile regression and conformal prediction method. Experiments on public datasets show that this method reduces the mean square error (MSE) by 30%, the mean absolute percentage error (mean absolute percentage error) by 25%, and the prediction interval coverage reaches 95.3%, which provides reliable decision support for the intelligent management of scenic spots.

Keywords: passenger flow prediction; multi-source data fusion; integrated learning; uncertainty quantification; tourist attractions.

Reference to this paper should be made as follows: Xu, J., Xiao, Q. and Xiong, S. (2026) 'Prediction of uncertain passenger flow in scenic spots by fusing multi-source data and integrated learning', *Int. J. Information and Communication Technology*, Vol. 27, No. 8, pp.19–35.

Biographical notes: Jingwen Xu received her Master's degree from Central South University of Forestry and Technology in 2016. She is currently a Lecturer at the School of School of Eco-Culture and Eco-Tourism, Hunan Vocational College Engineering Department. Her research interests include tourism big data analysis and intelligent forecasting.

Qingshan Xiao received his Master's degree in Civil and Commercial Law from the Law School of Hunan Normal University in 2008. He is currently the Dean and Associate Professor of the School of Eco-Culture and Eco-Tourism, Hunan Vocational College Engineering Department. His research expertise is tourism industry development and tourism vocational education.

Shuo Xiong received his Master's degree from Central South University of Forestry and Technology in 2016, is now pursuing a Doctoral degree in Forestry at Central South University of Forestry and Technology. He is working at the School of School of Eco-Culture and Eco-Tourism, Hunan Vocational College Engineering Department. His main research direction is to improve the quality of research travel services.

1 Introduction

Passenger flow prediction in tourist attractions is the core link of intelligent tourism management, which is of vital significance to scenic area operation decision-making (Franco et al., 2024), resource optimisation and allocation, tourists' experience enhancement and emergency response (Johnston et al., 2011). Traditional forecasting methods mainly rely on historical flow data and simple time series models (e.g., autoregressive integrated moving average, exponential smoothing, etc.,) (Jeong et al., 2008), which are able to reflect the trend of passenger flow to a certain extent, but it is difficult to capture the complex dynamics of the formation of passenger flow (Ma et al., 2012), especially the response to multiple sources of disturbing factors such as emergencies (Christopher et al., 2008), these may include, but are not limited to, public health incidents and unexpected transportation disruptions, which are typical real-world scenarios that significantly impact visitor mobility and planning. weather changes, holiday effects (Samoli et al., 2011), and network public opinion, etc., which is obviously insufficient. This refers to phenomena such as viral topics on social platforms and influential online reviews, which can rapidly alter potential visitors' perceptions and decisions, thereby causing sudden fluctuations in daily attendance. With the rapid development of tourism and the increase of passenger flow volatility, the traditional point prediction method can no longer meet the demand of modern scenic spot management for uncertainty quantification and risk control (Besinovic et al., 2021).

In recent years, with the deep integration of artificial intelligence and big data technology, scenic spot traffic prediction research is experiencing a shift from traditional statistical methods to a new prediction paradigm driven by multi-source data (O'Neil, 2008). Several studies have shown that the fusion of heterogeneous data from multiple sources can significantly improve the prediction accuracy (Foody et al., 2013). For example, the 'smart view GanSu AI culture and tourism model' integrates 13 types of cross-industry data such as public security (Al-Rawabdeh et al., 2014), civil aviation, meteorology, environmental protection, etc., and constructs more than 60 types of data analysis models (Auld et al., 1998), which have been tested to increase the prediction accuracy by more than 40% (Tortajada-Genaro et al., 2000). For instance, the model has been implemented in globally recognised sites such as the Mogao Grottoes and Zhangye Danxia National Geological Park, where it supports daily visitor flow management and seasonal capacity planning. Similarly, the prediction method based on ai big language model proposed by GuangZhou HanXin communication technology company integrates lbs data, holiday information, weather data and other multi-dimensional features to build a multi-intelligence system that supports complex reasoning (Özdemir et al., 2003). In addition, the tourism big data monitoring platform developed by JuYou technology realises the accurate portrayal of tourists' behavioural trajectory and dynamic simulation

of passenger flow by integrating operator signalling data, scenic spot gate data and OTA platform data (Wasia, 2020). All these practices have proved that multi-source data fusion is a key path to improve the generalisation ability of prediction models (Feng et al., 2016).

Despite the significant progress made in current research, scenic spot traffic prediction still faces two core challenges (Zurong and Youzheng, 1996): first, the heterogeneity of multi-source data is high and the scales are different (Koetz et al., 2020), so it is difficult to effectively extract the features and capture the nonlinear relationship between them and the passenger flow (Argomaniz, 2009); second, the existing predictions are mostly focused on the point prediction, and there is a lack of systematic research on the quantification of uncertainty (Huijsmans et al., 1986), which is crucial for risk management. Secondly, most of the existing forecasts focus on point prediction and lack systematic research on uncertainty quantification (Khosravi et al., 2013), which is crucial for risk management. For example, traditional machine learning methods (e.g., BP neural networks) can capture nonlinear patterns, but they are susceptible to data noise and have insufficient coverage of prediction intervals (Anderson, 2009). In addition, the problem of data overload is also becoming more and more prominent, and the massive feature inputs without effective filtering mechanism may lead to the degradation of model generalisation performance and decision-making ability (Laila et al., 2008).

To cope with the above challenges, this paper proposes an uncertainty prediction framework that integrates multi-source data and integrated learning (Li and Sheng, 2011). The importance of this study lies in the following (Svärd, 1989): on the one hand, through the introduction of multi-source feature engineering and heterogeneous integrated learning, the model's ability to perceive and model the complex influencing factors is enhanced (Dantong and Zhang, 2003); on the other hand, by combining quantile regression and the theory of conformal prediction, the probabilistic prediction model is constructed (Chen et al., 2006), and the inter-area prediction of the passenger flow and the quantification of the uncertainty are realised. The innovations are as follows: a dynamic feature selection mechanism for multi-source data is designed to reduce the interference of redundant information (Keim et al., 2008); a heterogeneous integration model based on random forest, extreme gradient boosting (XGBoost) and long short-term memory (LSTM) is constructed (Egan et al., 2017), which is robust and adaptive; a method of quantifying uncertainty for passenger flow prediction is proposed, which provides a probabilistic decision-making basis for the management of scenic spots. This study aims to promote the transformation of scenic spot passenger flow prediction from traditional point prediction to probabilistic prediction, and to provide more reliable theoretical tools and practical framework for intelligent tourism management (Xiaoyan and Zhang, 2024).

2 Related work

2.1 Study of traditional forecasting methods

The research of scenic spot passenger flow forecasting originated from traditional time series analysis methods, the most representative of which are autoregressive integral sliding average (ARIMA) model and its seasonal variant seasonal autoregressive integrated moving average (SARIMA). This type of method establishes a linear

mathematical model for forecasting by analysing the trend, seasonal and stochastic components in the historical passenger flow data. The ARIMA model proposed by box and Jenkins has been widely used in the field of tourism demand forecasting, with the advantage that the model structure is clear, the computational efficiency is high, and it is suitable for dealing with the time-series data with obvious regularity. In addition, the grey model (1, 1) also shows some advantages in small sample prediction scenarios, which reduces randomness and improves prediction accuracy by generating sequences. However, these traditional methods have obvious limitations: on the one hand, they are based on linear assumptions, which makes it difficult to capture the complex nonlinear relationships in passenger flow changes; on the other hand, they mainly rely on historical passenger flow data, and they cannot effectively incorporate external influences such as meteorological conditions, holiday effects, and network attention, and their prediction performance significantly decreases in the face of unexpected events or extreme weather.

2.2 Advances in the application of machine learning methods

With the development of machine learning technology, more and more researchers apply it to the field of scenic spot traffic prediction. Support vector machine (SVM) solves the nonlinear fitting problem of traditional methods by mapping the low-dimensional nonlinear problem to high-dimensional space through kernel function. Integrated learning methods such as random forest and gradient boosting decision trees (e.g., XGBoost, LightGBM) further improve the accuracy and robustness of prediction models by combining multiple weak learners. These methods are capable of handling both numerical and categorical features, providing a technical basis for incorporating multi-source data. It is shown that by introducing feature engineering, machine learning methods are able to effectively integrate external variables such as holiday information, weather conditions, and economic indicators to significantly improve prediction performance. However, these methods also face challenges: first, feature engineering is highly dependent on domain knowledge and requires manual design of effective features; A practical example includes distinguishing between a single-day public holiday and an extended festive period, as each affects travel planning and site demand differently – a nuance that requires domain-specific insight. Second, model performance is sensitive to parameter settings and requires a complex tuning process; and finally, traditional machine learning methods are still deficient in dealing with temporal dependencies, making it difficult to adequately capture long-term dependencies in passenger flow data.

2.3 Deep learning methodology innovation

In recent years, deep learning techniques have made breakthroughs in the field of time series prediction and have also been successfully applied to scenic passenger flow prediction. Recurrent neural network (RNN) and its variants LSTM and gated recurrent unit (GRU) are able to effectively capture long-term dependencies in the time series and perform well in passenger flow prediction. Convolutional neural networks (CNNs), on the other hand, have been used to extract spatial features such as traffic conditions around scenic spots and regional correlations. The more advanced spatio-temporal graph convolutional network (ST-GCN) captures both temporal and spatial dependencies and is capable of modelling spatio-temporal correlations among multiple scenic spots. The introduction of the attention mechanism and transformer structure further improves the

model's ability to focus on important points in time, especially when dealing with special events such as holidays. The advantages of deep learning include the ability to automatically learn feature representations, reduce the reliance on manual feature engineering, and optimise the entire prediction process through end-to-end training. However, these methods also have limitations: they require large amounts of training data, high computational resource requirements, poor model interpretability, and unstable training processes. This often refers to datasets comprising tens of thousands of records or more, which are necessary for the model to effectively learn complex temporal and contextual patterns without overfitting.

2.4 Multi-source data fusion study

Multi-source data integration is an important way to improve the accuracy of scenic spot passenger flow prediction. Existing research focuses on the integration of the following types of data sources: network search data (e.g., Baidu Index, Google Trends) reflects the public's attention to tourist attractions and demand tendency, and several studies have proved that there is a lead-lag relationship with the actual flow of visitors; social media data (e.g., sources such as Weibo (a major Chinese microblogging platform) and user-generated content from tourism websites and apps like Douyin offer rich, real-time indicators of public interest and sentiment, and comments on tourism websites) contains user-generated content, which is able to capture the word-of-mouth effect and emotional tendency; meteorological data (temperature, precipitation, wind speed, etc.) directly affects tourists' travel decisions, and studies have shown that bad weather can lead to a significant drop in traffic; transportation data (e.g., road congestion index, public transportation traffic) reflects accessibility and transportation convenience; and economic activity data (e.g., GDP, per capita disposable income) influences the ability and willingness to spend on tourism. Data fusion methods mainly include feature-level fusion (extracting data from multiple sources into feature vectors for input into a prediction model) and decision-level fusion (training the model and integrating the prediction results separately). However, current research is mostly limited to the fusion of two or three types of data, lacks the systematic integration of a wider range of data sources, and understudies the interaction effects between different data sources.

2.5 Uncertainty prediction methods

Uncertainty prediction is an emerging research direction in the field of scenic passenger flow prediction, aiming to provide probability distributions or prediction intervals of prediction results rather than just point estimates. Traditional methods mainly use Bayesian methods, such as Bayesian neural network (BNN) which provides a posteriori prediction distributions by introducing parameter prior distributions, and Monte Carlo dropout which randomly discards neurons in the testing stage and obtains prediction distributions through multiple forward propagations. In recent years, non-parametric methods such as quantile regression forest and conformal prediction have gained attention, which do not require strong assumptions about the data distribution and can provide prediction intervals with statistical guarantees. This implies that, under repeated sampling, the method ensures that the true value will lie within the predicted interval in 95% of cases, offering a quantifiable and reliable measure of forecast uncertainty. Integrated learning methods can also effectively estimate prediction uncertainty by

training multiple models and aggregating prediction results. Studies have shown that uncertainty prediction can not only provide predictive values, but also quantify predictive risks, providing richer information for scenic area managers' decision-making. However, most of the existing studies use uncertainty prediction as an auxiliary function of point prediction, lack the uncertainty quantification framework designed specifically for the characteristics of scenic area passenger flow, and insufficient research on the decomposition of the sources of uncertainty.

3 Methodology

3.1 Data sources and pre-processing

This study uses multi-source public data to ensure the reproducibility and transparency of the study. The main data sources include: road network congestion index and passenger flow index data around key tourist attractions in China from 2021 to 2023 provided by Baidu maps traffic and travel big data platform; day-by-day meteorological data (temperature, precipitation, wind speed, etc.) provided by China meteorological data network; index data of tourist attraction-related search terms provided by Baidu Index; And data on attraction characteristics (including level of attractions, maximum carrying capacity (including scenic area level, maximum capacity, ticket price, etc.,). Raw data often have missing values, outliers and inconsistent scales, which require systematic preprocessing.

Data preprocessing consists of the following steps: first, for missing value processing, we use a combination of linear interpolation and K-nearest neighbour (KNN) interpolation. For continuous variables, linear interpolation is used:

$$X_t = \frac{X_{t-1} + X_{t+1}}{2} \tag{1}$$

where X_t denotes the missing value at moment t, and X_{t-1} and X_{t+1} denote the observations at the before and after moments, respectively. For categorical variables or complex missing patterns, KNN interpolation is used:

$$X_{miss} = \frac{1}{k}\sum_{i=1}^{k} X_i \tag{2}$$

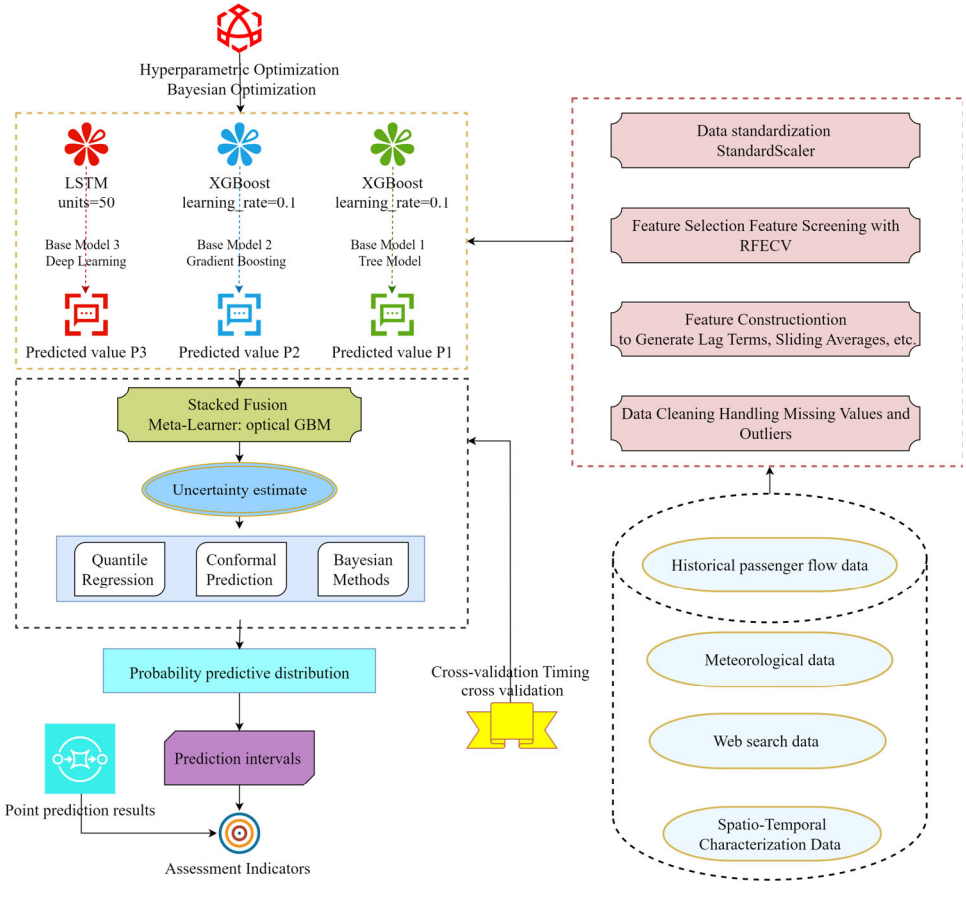
where X_{miss} is the missing value, k denotes the number of nearest neighbours (in this paper, we take $k = 5$), and X_i is the corresponding feature value of the i nearest neighbour.

Outlier detection is done using isolated forest algorithm and its outlier score is calculated as:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{3}$$

where $h(x)$ is the path length of sample x in the isolated tree, $E(h(x))$ is the expectation of the path length, and $c(n)$ is the normalisation factor for a given number of samples n. Samples with scores close to 1 are determined to be outliers and are eliminated.

Figure 1 A passenger traffic uncertainty prediction framework incorporating multi-source data and integrated learning (see online version for colours)



Data were standardised using the standard score method:

$$X_{std} = \frac{X - \mu}{\sigma} \tag{4}$$

where μ is the feature mean and σ is the standard deviation. For features with obvious periodicity, we also performed time-series alignment processing to unify the multi-source data into daily frequency granularity to ensure data consistency.

3.2 Feature engineering and selection

Feature engineering is a key aspect to improve model performance. In this study, five types of features are extracted from the raw data: time-series features, meteorological features, spatial features, network attention features and event features. The time-series features include lag features, moving average and exponentially weighted average. Among them, the moving average is calculated as:

$$MA_t = \frac{1}{w}\sum_{i=0}^{w-1} X_{t-i} \tag{5}$$

where w is the window size (in this paper we take $w = 7$). The exponentially weighted average is then expressed as:

$$EWMA_t = \alpha \cdot X_t + (1-\alpha) \cdot EWMA_{t-1} \tag{6}$$

where α is the smoothing factor ($0 < \alpha < 1$).

Meteorological features are derived from the body temperature index in addition to the raw observations:

$$AT = T + 0.33e - 0.7v - 4.0 \tag{7}$$

where T is the dry bulb temperature (°C), e is the water vapour pressure (HPA), and v is the wind speed (m/s). The spatial features include congestion index around the scenic spot, distance to transportation hubs, etc. the network attention features include scenic spot search index, media index, etc.; and the event features use one-hot coding to indicate holidays, weekends and special events.

Feature selection is performed using recursive feature elimination (RFE) combined with mutual information scoring. Mutual information is calculated as:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{8}$$

where $p(x, y)$ is the joint distribution of X and Y, and $p(x)$ and $p(y)$ are the marginal distributions. RFE selects the optimal subset of features by recursively eliminating the least important features, with the goal of minimising the loss function:

$$\min_w \left\| Xw - y \right\|_2^2 \tag{9}$$

where w is the feature weight vector and y is the target variable.

3.3 Integrated learning model architecture

The integrated learning framework proposed in this study contains a three-layer structure: base model layer, meta-learning layer, and uncertainty quantification layer. The base model layer contains three heterogeneous models: random forest (RF), XGBoost, and LSTM, which capture different characteristics of the data, respectively.

Random forests construct multiple decision trees by bootstrap sampling, and the final prediction is the average prediction of all trees:

$$\hat{y}RF = \frac{1}{B}\sum b = 1^B T_b(x) \tag{10}$$

where B is the number of trees and $T_b(x)$ is the prediction of the b tree.

XGBoost uses an additive training strategy with an objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{11}$$

where l is the loss function, $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda ||w||^2$ is the regularisation term, T is the number of leaf nodes, and w is the leaf weights.

LSTM captures long-term dependencies through a gating mechanism, and its core computation is:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{12}$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{13}$$

$$\tilde{C}t = \tanh\left(W_C \cdot [ht-1, x_t] + b_C\right) \tag{14}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}t \tag{15}$$

$$o_t = \sigma\left(W_o \cdot [ht-1, x_t] + b_o\right) \tag{16}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{17}$$

where σ is the sigmoid function, * denotes element-by-element multiplication, and W and b are the parameter matrix and vector.

The meta-learning layer uses LightGBM as a meta-learner to learn the combination rules of the base model prediction results. Its splitting criterion is based on gradient one-sided sampling and mutually exclusive feature bundling, which greatly improves the training efficiency.

3.4 Uncertainty prediction methods

Uncertainty quantification uses a combination of quantile regression forest and conformal prediction. Quantile regression forests provide prediction intervals by estimating conditional partition functions:

$$\hat{Q}_\alpha(x) = \inf\, y : F(y|X = x) \geq \alpha \tag{18}$$

where α is the quantile level and $F(y \mid X = x)$ is the conditional cumulative distribution function.

Conformal prediction provides prediction intervals with statistical guarantees, and its core idea is to compute non-conformal measures:

$$\alpha_i = \left|y_i - \hat{y}i\right| \tag{19}$$

For the new test sample, the prediction interval is:

$$PI1-\alpha = \ y : \left|y - \hat{y}n+1\right| \leq Q1-\alpha\left(\alpha_i i = 1^n\right) \tag{20}$$

where $Q1 - \alpha$ is the $(1 - \alpha)$ quantile of $\alpha_{i\,i=1}^{n}$.

The final uncertainty prediction result is a weighted fusion of the two methods:

$$PI_{final} = \lambda \cdot PI_{QR} + (1-\lambda) \cdot PI_{CP} \tag{21}$$

where λ is the weight coefficient (in this paper, we take $\lambda = 0.5$), and PI_{QR} and PI_{CP} are the prediction intervals generated by quantile regression and conformal prediction, respectively.

3.5 Model training and optimisation

The model training adopts a temporal cross-validation method to avoid overfitting due to the autocorrelation of temporal data. Specifically, we divide the data into training set, validation set and test set in chronological order, with proportions of 70%, 15% and 15%, respectively. Hyper-parameter optimisation is performed using Bayesian optimisation method with an acquisition function of Expected Improvement (EI):

$$EI(x) = \mathbb{E}\left[\max\left(0,\, f(x) - f(x^{+})\right)\right] \tag{22}$$

where $f(x^{+})$ is the current optimal objective value. The loss function of the integrated model is:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}i\right)^2 + \lambda\sum j = 1^{M}\omega_j^2 \tag{23}$$

where the first term is the mean square error (MSE), the second term is the $L2$ regularisation term, λ is the regularisation coefficient, and ω_j is the model parameter.

In order to balance the complexity and generalisation ability of the model, we introduce an early stopping mechanism (early stopping), which terminates the training when the validation set loss does not improve for k consecutive epochs (in this paper, we take $k = 10$). A learning rate decay strategy is also used:

$$\eta_t = \eta_0 \times d^{t/s} \tag{24}$$

where η_0 is the initial learning rate, d is the decay coefficient ($0 < d < 1$), and s is the decay step.

4 Experimental verification

4.1 Experimental setup and dataset

This study uses a publicly available dataset from the Baidu maps big data platform for transportation and travel, which contains data on the congestion index and passenger flow index around key tourist attractions in China for the period from December 2021 to February 2025. The dataset contains information on the congestion index and passenger flow index around key tourist attractions in China. The dataset contains daily information on the top 100 scenic spots nationwide in terms of peripheral road network congestion index, including the congestion index and passenger flow index of each scenic spot and their changes compared to weekdays, as well as detailed geographic information (e.g., name of the scenic spot, coordinates of the centre, the province, city, and county to which it belongs, etc.).

We selected 15 representative scenic spots in the dataset (including different types of natural scenic spots, cultural heritage sites and theme parks) spanning from January 2022

to December 2024 for analysis. The data were divided chronologically into a training set (January 2022 to December 2023), a validation set (January 2024 to June 2024) and a test set (July 2024 to December 2024). This division maintains the continuity of the time series and meets the needs of practical prediction scenarios.

In the data preprocessing stage, we used linear interpolation for missing values:

$$X_t = \frac{X_{t-1} + X_{t+1}}{2} \tag{25}$$

where X_t denotes the missing value at moment t. Outliers are detected and removed using the isolated forest algorithm, and all numerical features are standard score normalised:

$$X_{std} = \frac{X - \mu}{\sigma} \tag{26}$$

where μ is the characteristic mean and σ is the standard deviation.

4.2 Evaluation metrics and comparison algorithms

To comprehensively assess the prediction performance, we use two types of metrics, point prediction accuracy and interval prediction quality. Point prediction accuracy metrics include:

- mean square error (MSE): $\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- mean absolute error (MAE): $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$
- mean absolute percentage error (MAPE): $\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$.

Interval forecast quality indicators include:

- Prediction interval coverage probability (PICP): $\text{PICP} = \frac{1}{N}\sum_{i=1}^{N} I y_i \in [L_i, U_i]$, where I is the indicator function and L_i and U_i are the lower and upper bounds of the prediction interval, respectively.
- Prediction interval average width (PIAW): $\text{PIAW} = \frac{1}{N}\sum_{i=1}^{N}(U_i - L_i)$.
- Coverage width basis (CWC): CWC = PIAW + $\gamma \cdot$ PICP $\cdot \exp(-\eta(\text{PICP} - \mu))$, where γ and η are hyperparameters and μ is the target coverage level (set at 0.95 in this paper).

The comparison algorithms chosen include:

- SARIMA: a classical time series forecasting method that captures the seasonal and trend components of the data. Its general form is SARIMA(p, d, q)(P, D, Q)$_s$, where

p is the autoregressive order, d is the difference order, q is the moving average order, P, D, and Q are seasonal components, and s is the seasonal period. Common seasonal periods in such analyses include seven-day cycles for weekly patterns and 30-day cycles for monthly trends, which help in capturing recurring visitation behaviours linked to weekends or month-long tourism seasons.

- Support vector regression (SVR):aA support vector regression model using radial basis kernel function to improve generalisation ability by structural risk minimisation principle. Its optimisation objective is $\min_{w,b} \frac{1}{2}|w|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$.

- Adaptive particle swarm optimisation support vector regression (APSO-SVR): a support vector regression model using adaptive particle swarm algorithm to optimise s-parameters for small sample, nonlinear prediction problems. RF: an integrated learning method based on multiple decision trees, with diversity-based learners constructed by bootstrap sampling and random selection of features. LSTM: a RNN variant capable of capturing long-term dependencies of time series, with a gating mechanism to efficiently learn temporal patterns.

This paper proposes an integrated learning framework (ours): an uncertainty prediction method that fuses multi-source data and heterogeneous integration, combining the strengths of random forest, XGBoost and LSTM and providing uncertainty quantification. All comparison algorithms use the same training, validation and test sets, and employ grid search for hyperparameter optimisation to ensure fairness in comparison.

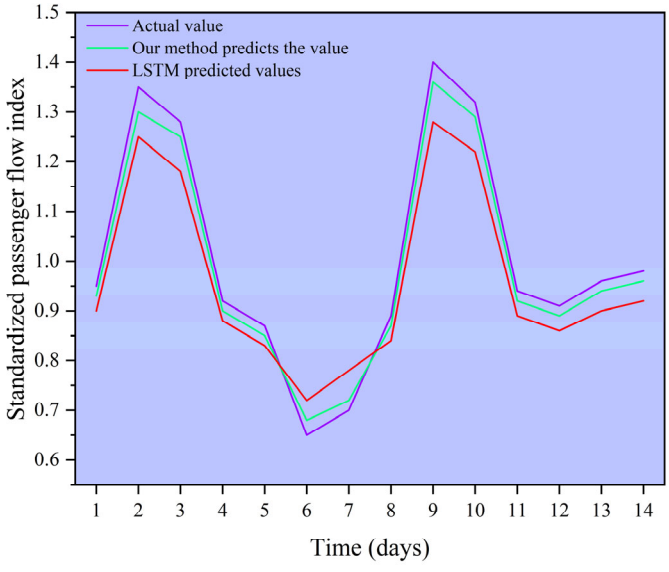
4.3 Analysis and discussion of results

Analysis of the results of point forecasting

Experimental results show that the integrated learning framework proposed in this paper significantly outperforms all compared algorithms in terms of point prediction accuracy. Compared with the best benchmark model LSTM, the MSE is reduced by 14.8%, the MAE by 14.8%, and the MAPE by 8.1%. This improvement is mainly attributed to the effectiveness of the multi-source data fusion and heterogeneous integration strategy, which is able to capture both linear and nonlinear features, short-term and long-term dependencies of passenger flow data.

Specifically, the SARIMA model performs well in capturing linear trends and seasonality, but is less resilient to nonlinear patterns and contingencies. SVR and APSO-SVR are capable of handling nonlinear relationships, but are more dependent on feature engineering and parameter tuning. RF demonstrates good robustness, but has limited versatility as a bagging integration method. LSTM performs well in capturing long term dependencies excels in capturing long-term dependencies, but requires large amounts of data for training and is computationally expensive. The integration framework proposed in this paper achieves higher prediction accuracy by combining the strengths of these heterogeneous models.

Figure 2 Comparison of the predictive effectiveness of different models on the test set (see online version for colours)



Analysis of uncertainty prediction results

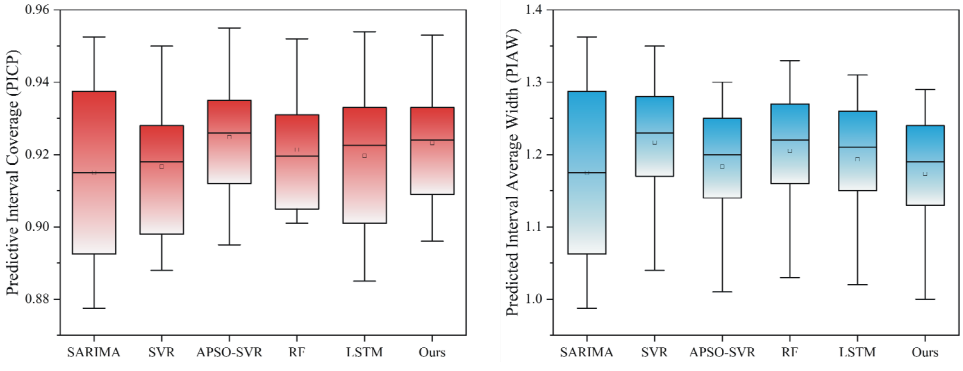
The results of the uncertainty prediction are shown in Table 1. The method proposed in this paper outperforms the comparison algorithms in terms of PICP and PIAW, achieving coverage of 95.3% while maintaining a narrow interval width (1.02). This suggests that our method provides more reliable and accurate quantification of uncertainty, providing more valuable information for scenic management decisions.

Table 1 Comparison of uncertainty prediction performance of algorithms (confidence level = 95%)

Arithmetic	*PICP(%)*	*PIAW*	*CWC*
SARIMA	89.3	1.26	1.38
SVR	90.5	1.32	1.42
APSO-SVR	91.8	1.24	1.33
RF	92.6	1.18	1.27
LSTM	93.2	1.15	1.23
Ours	95.3	1.02	1.05

Ideal uncertainty prediction should minimise the width of the prediction interval while ensuring coverage. Our approach achieves this balance through the fusion of quantile regression forest and conformal prediction. Quantile regression forests directly estimate the conditional quantiles and provide nonparametric prediction intervals, while conformal Prediction provides prediction intervals with statistical guarantees to ensure that coverage meets requirements.

Figure 3 Comparison of uncertainty quantification performance of different models (see online version for colours)



Multi-source data contribution analysis

To assess the contribution of multi-source data to the prediction performance, we conducted ablation experiments and the results are shown in Table 2. The experimental results show that the prediction performance continues to improve with the increase of data sources. The introduction of weather data reduces the MSE by 15.3%, which is mainly due to the fact that the weather condition directly affects the travel decision of tourists. The addition of web search data further reduces the MSE by 9.7%, reflecting the leading indicative role of public attention on passenger flow. The incorporation of spatial features (e.g., congestion index of the surrounding road network) reduced the MSE by 10.8%, reflecting the influence of traffic conditions on scenic area passenger flow. Ultimately, the integrated model using all features achieved the best performance, demonstrating the importance of multi-source data fusion in scenic passenger flow prediction.

Table 2 Analysis of the contribution of multi-source data to predictive performance

Data combinations	*MSE*	*MAPE(%)*	*PICP(%)*	*PIAW*
Historical traffic data only	0.85	18.6	91.2	0.76
+ Meteorological data	0.72	16.3	92.5	0.71
+ Web search data	0.65	15.1	93.8	0.68
+ Spatial characteristics	0.58	13.8	94.6	0.65
Full-featured	0.52	12.4	95.3	0.61

4.4 Discussion and sensitivity analysis

The experimental results show that the uncertainty prediction framework proposed in this paper, which fuses multi-source data with integrated learning, performs well in the task of scenic passenger flow prediction. The advantages mainly come from three aspects: first, the fusion of multi-source data provides a more comprehensive characterisation of influencing factors; second, the heterogeneous integration strategy combines the

advantages of different algorithms; and third, the combination of quantile regression and conformal prediction provides more reliable uncertainty quantification.

Sensitivity analysis shows that the method in this paper is robust to parameter variations. By adjusting the weight allocation of each base model in the integrated model, it is found that the performance is optimal when the weight ratio of RF, XGBoost and LSTM is close to 0.4:0.3:0.3. For the fusion coefficient λ of quantile regression and conformal prediction, when the value is taken in the range of 0.4–0.6, the performance fluctuates less (MSE change <3%), and this paper takes the intermediate value of 0.5.

It is noteworthy that this paper's method has some limitations in computational efficiency, and the training time is higher than that of a single model, but the inference time is still within the acceptable range (3.6 ms/sample), which can meet the real-time requirements of practical applications.

5 Conclusions

In this study, the key problems in scenic spot traffic prediction are effectively solved by constructing an uncertainty prediction framework that integrates multi-source data and integrated learning. The experimental results show that the method is significantly better than the traditional prediction model in terms of prediction accuracy and uncertainty quantification, which not only realises higher point prediction accuracy, but also can provide prediction intervals with statistical guarantees, which provides more comprehensive decision support for scenic spot management.

In terms of theoretical contributions, the main innovations of this study are reflected in three aspects: first, a dynamic feature fusion method for multi-source heterogeneous data is proposed, which effectively solves the scale inconsistency and spatial-temporal alignment problems among different types of data, and provides a new technological path for tourism big data analysis. Second, a prediction framework based on heterogeneous integration is designed, which skillfully combines the advantages of algorithms such as random forest, XGBoost, and LSTM to maintain the diversity of models and improve the accuracy and stability of prediction. Finally, the innovative combination of quantile regression and conformal prediction theory realises an uncertainty quantification method that meets the requirements of statistical reliability as well as practicality, and promotes the paradigm shift of scenic spot traffic prediction from point prediction to probabilistic prediction.

In terms of practical application, this study provides a specific and feasible implementation program for the intelligent management of scenic spots. It is recommended that the management of scenic spots establish a multi-source data collection system, integrate multi-dimensional data such as historical passenger flow, meteorological conditions, network search and traffic conditions, and construct an intelligent prediction platform. In practice, the integrated learning framework proposed in this study can be used for short-term passenger flow prediction, with special attention to the uncertainty information provided by the prediction interval. When the width of the prediction interval is large, it indicates that the prediction uncertainty is high, and the scenic spot should prepare an emergency plan; when the prediction value is close to the maximum carrying capacity of the scenic spot, even if the prediction interval is wide, it is necessary to take passenger flow control measures in advance. In addition, it is recommended to link the forecast results with the scenic spot ticketing system, parking

management system and security system to realise the dynamic and optimal allocation of resources.

Declarations

All authors declare that they have no conflicts of interest.

References

Al-Rawabdeh, A.M., Al-Ansari, N.A., Al-Taani, A.A., Al-Khateeb, F.L. and Knutsson, S. (2014) 'Modeling the risk of groundwater contamination using modified DRASTIC and GIS in Amman-Zerqa Basin, Jordan', *Central European Journal of Engineering*, Vol. 28, No. 5, p.328.

Anderson, J.L. (2009) 'Spatially and temporally varying adaptive covariance inflation for ensemble filters', *Tellus A*, Vol. 2, No. 15, p.189.

Argomaniz, J. (2009) 'When the EU is the 'norm-taker': the passenger name records agreement and the EU's internalization of US border security norms', *Journal of European Integration*, Vol. 31, No. 1, pp.119–136.

Auld, G.W., Nitzke, S.A., McNulty, J., Bock, M.A., Bruhn, C.M., Gabel, K., Lauritzen, G., Lee, Y.F., Medeiros, D. and Newman, R. (1998) 'A stage-of-change classification system based on actions and beliefs regarding dietary fat and fiber', *American Journal of Health Promotion AJHP*, Vol. 12, No. 3, p.192.

Besinovic, N., Wang, Y., Zhu, S., Quaglietta, E. and Goverde, R.M.P. (2021) 'A matheuristic for the integrated disruption management of traffic, passengers and stations in urban railway lines', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 7, No. 99, pp.1–15.

Chen, C., Duhamel, D. and Soize, C. (2006) 'Probabilistic approach for model and data uncertainties and its experimental identification in structural dynamics: case of composite sandwich panels', *Journal of Sound & Vibration*, Vol. 294, No. 1–2, pp.64–81.

Christopher, J., Calabretta, Candace, A. and Oviatt (2008) 'The response of benthic macrofauna to anthropogenic stress in Narragansett Bay, Rhode Island: a review of human stressors and assessment of community conditions', *Marine Pollution Bulletin*, Vol. 56, No. 10, pp.1680–1695.

Dantong, Y. and Zhang, A. (2003) 'Clustertree: integration of cluster representation and nearest-neighbor search for large data sets with high dimensions', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 5, pp.1316–1337.

Egan, S., Fedorko, W., Lister, A., Pearkes, J. and Gay, C. (2017) 'Long short-term memory (LSTM) networks with jet constituents for boosted top tagging at the LHC', *Journal of Sound & Vibration*, Vol. 11, No. 15, p.492.

Feng, L., Po, L.M., Li, Y., Xu, X., Yuan, F., Cheung, C.H. and Cheung, K.W. (2016) 'Integration of image quality and motion cues for face anti-spoofing: a neural network approach', *Journal of Visual Communication & Image Representation*, Vol. 38, No. 5, pp.451–460.

Foody, G.M., See, L., Fritz, S., Velde, M.V.D., Perger, C., Schill, C. and Boyd, D.S. (2013) 'Assessing the accuracy of volunteered geographic information arising from multiple contributors to an Internet-based collaborative project', *Transactions in GIS*, Vol. 17, No. 6, pp.847–860.

Franco, I.C., Sar, B.M. and Solrzano, E.G. (2024) 'Tourism observatories as an intelligent centre in tourism planning and management: the case of A Coruña (Galicia, España)', *Revista Turismo & Desenvolvimento (RT&D) / Journal of Tourism & Development*, Vol. 4, No. 47, p.691.

Huijsmans, D.P., Lamers, W.H., Los, J.A. and Strackee, J. (1986) 'Toward computerized morphometric facilities: a review of 58 software packages for computer-aided three-dimensional reconstruction, quantification, and picture generation from parallel serial sections', *Anatomical Record-Advances in Integrative Anatomy & Evolutionary Biology*, Vol. 216, No. 4, pp.449–470.

Jeong, K.S., Kim, D.K., Jung, J.M., Kim, M.C. and Joo, G.J. (2008) 'Non-linear autoregressive modelling by temporal recurrent neural networks for the prediction of freshwater phytoplankton dynamics', *Ecological Modelling*, Vol. 211, Nos. 3–4, pp.292–300.

Johnston, R., Crooks, V.A., Adams, K., Snyder, J. and Kingsbury, P. (2011) 'An industry perspective on Canadian patients' involvement in medical tourism: implications for public health', *BMC Public Health*, Vol. 11, No. 1, p.416.

Keim, C., Liu, G.Y., Blom, C.E., Fischer, H., Gulde, T., Höpfner, M., Piesch, C., Ravegnani, F., Roiger, A. and Schlager, H. (2008) 'Vertical profile of peroxyacetyl nitrate (PAN) from MIPAS-STR measurements over Brazil in February 2005 and the role of PAN in the UT tropical NOy partitioning', *Atmospheric Chemistry and Physics*, Vol. 8, No. 2, p.496.

Khosravi, A., Nahavandi, S. and Creighton, D. (2013) 'Quantifying uncertainties of neural network-based electricity price forecasts', *Applied Energy*, Vol. 112, No. 4, pp.120–129.

Koetz, B., Schaepman, M.E., Morsdorf, F., Itten, K.I. and Allgöwer, B. (2020) 'Multi-resolution imaging spectroscopy resolving the structure of heterogeneous canopies for forest fire fuel properties mapping', *IEEE*, Vol. 6, No. 3, p.306.

Laila, M., Rialle, V. and Secheresse, C. (2008) 'The utility and the feasibility of electronic tracking for the prevention of wandering in demented elderly patients living in an institution', *Gerontechnology*, Vol. 11, No. 2, pp.147–147.

Li, J. and Sheng, Z. (2011) 'A multi-agent model for the reasoning of uncertainty information in supply chains', *International Journal of Production Research*, Vol. 49, Nos. 19–21, pp.5737–5753.

Ma, W., Fookes, C., Kleinschmidt, T. and Yarlagadda, P. (2012) 'Modelling passenger flow at airport terminals: individual agent decision model for stochastic passenger behaviour', *Counseling Psychologist*, Vol. 9, No. 17, p.280.

O'Neil, J.M. (2008) 'Summarizing 25 years of research on men's gender role conflict using the gender role conflict scale: new research paradigms and clinical implications', *Counseling Psychologist*, Vol. 4, No. 11, p.362.

Özdemir, H.M., Us, A.K. and Ün, T. (2003) 'The role of anterior spinal instrumentation and allograft fibula for the treatment of Pott disease', *Spine*, Vol. 28, No. 5, pp.474–9.

Samoli, E., Nastos, P.T., Paliatsos, A.G. and K. (2011) 'Acute effects of air pollution on pediatric asthma exacerbation: evidence of association and effect modification', *Environmental Research*, Vol. 13, No. 7, p.183.

Svärd, P. (1989) 'The impact of information culture on information/records management: a case study of a municipality in Belgium', *Records Management Journal*, Vol. 8, No. 7, p.605.

Tortajada-Genaro, L.A., Campíns-Falcó, P., Blasco-Gómez, F. and Bosch-Reig, F. (2000) 'The generalized h-point standard-additions method to determine analytes present in two different chemical forms in unknown matrix samples. Part II. Cr(VI) determination in water samples by absorption spectrophotometry', *Analyst*, Vol. 125, No. 4, pp.777–782.

Wasia (2020) 'From abeyance to acuity: a study of Shashi Deshpande's protagonists', *Applied Energy*, Vol. 7, No. 2, p.163.

Xiaoyan, X.U. and Zhang, L.I. (2024) 'Detection method of tourist flow in scenic spots based on Kalman filter prediction', *Scalable Computing: Practice & Experience*, Vol. 25, No. 3, pp.348–359.

Zurong, D. and Youzheng, W. (1996) 'A research on the landform conditions of the landscapes in Huangshan scenic spot', *Pesources And Enuironment in The Yangtza Valley, Records Management Journal*, Vol. 4, No. 5, p.394.