Journal of Innovation & Knowledge 11 (2026) 100858

Contents lists available at ScienceDirect

Journal of Innovation & Knowledge

journal homepage: www.elsevier.com/locate/jik

Innovative knowledge-based system for forecasting daily hotel operations amid external events using multi-source data: A time-varying parameter state-space model

Ji Chen[a,b,d], Kang Tong[a], Qinglin Yu[a], Sichao Chen[c,d], Tomas Balezentis[e,f,*], Dalia Streimikiene[e]

[a] *School of Statistics and Data Science, Zhejiang Gongshang University, Hangzhou, China*
[b] *Collaborative Innovation Center of Statistical Data Engineering, Technology & Application, Zhejiang Gongshang University, Hangzhou, China*
[c] *College of Economics, Hangzhou Dianzi University, Hangzhou, China*
[d] *Economic Forecasting and Policy Simulation Laboratory, Zhejiang Gongshang University, Hangzhou, China*
[e] *Lithuanian Centre for Social Sciences, Vilnius, Lithuania*
[f] *Sustainability Competence Centre, Széchenyi István University, Egyetem tér 1, 9026, Győr, Hungary*

ARTICLE INFO

JEL codes:
C53
L83
Z30

Keywords:
Forecasting
Hotel occupancy
TVP-SSM
Multisource data
External events

ABSTRACT

Forecasting hotel occupancy during external shocks is particularly challenging due to their disruptive effects. This study develops a forecasting framework that integrates multisource data using a time-varying parameter state-space model (TVP-SSM). In this framework, search engine data (SED) are used to construct exogenous variables, intervention variables are used to reflect the severity of external shocks, and holiday and weekend dummy variables are used to capture the seasonal effect. The empirical study used a dataset from the hospitality industry in Hangzhou, China, covering the period from October 1, 2019, to October 28, 2021, and identified the COVID-19 pandemic as an external shock. The results show that TVP-SSM can effectively simulate the dynamic impact of external events and the periodical effect on hotel occupancy. Additionally, the prediction accuracy of TVP-SSM with intervention variables and periodical variables (TVP-SSM-1) exceeds that of competitive models. Specifically, compared to the naïve model and TVP-SSM without intervention variables and periodic variables (TVP-SSM-2), the prediction accuracy, measured by the root mean square error (RMSE) and mean absolute percentage error (MAPE), increased by 86 % and 87 %, respectively, and by 74 % and 76 %, respectively. These results indicate that the forecasting framework proposed in this study exhibits superior forecasting performance and demonstrates its capability for dynamic impact analysis of hotel occupancy at the industry level under external shocks.

Introduction

In the hospitality industry, enhancing the accuracy of occupancy forecasting can improve resource allocation and support the development of effective contingency plans (Wu et al., 2017; Zhang et al., 2021b; Gao et al., 2024; Waris & Mohd Suki, 2025). Previous studies have made many attempts to forecast tourism demand accurately. Song and Li (2008) classified tourism demand forecasting methods into three groups: time series models (Assaf et al., 2019; Dong et al., 2023), cross-section and panel econometric models (Hu & Song, 2019; Wen et al., 2019; Nicholas, 2021), and artificial intelligence technologies (He et al., 2021; Bi et al., 2022; Zhao et al., 2022). However, when tourism is affected by exceptional external shocks, such as natural disasters, terrorist attacks, or public health crises, the stability of these models is undermined (Zhang & Lu, 2022).

In response to external shocks, scholars have sought to enhance the accuracy of tourism forecasting models. Among them, the optimization of the model is the most direct path. Intervention analysis (Ozdogan & Ozdogan, 2023) and dynamic regression (Prilistya et al., 2021) are introduced into the tourism demand forecasting model. Additionally, expanding the diversity of data sources is also considered an effective means. Some literature integrates search engine data (SED; Liu et al.,

* Corresponding author.
E-mail addresses: chenji810404@zjgsu.edu.cn (J. Chen), 20020040079@pop.zjgsu.edu.cn (K. Tong), 23020040169@pop.zjgsu.edu.cn (Q. Yu), chensichao@hdu.edu.cn (S. Chen), tomas.balezentis@ekvi.lt (T. Balezentis), dalia@mail.lei.lt (D. Streimikiene).

https://doi.org/10.1016/j.jik.2025.100858
Received 17 February 2025; Accepted 15 October 2025
Available online 25 October 2025

2019; Hu et al., 2021b; Chen et al., 2024), web traffic data (Yang et al., 2013; Gunter & Önder, 2016; Emili et al., 2020; Li et al., 2023; Zhang et al., 2019), social media data (Pezenka & Weismayer, 2020; Xue et al., 2023; Zhan et al., 2024; Zhu et al., 2024; Sun et al., 2025; Tan et al., 2025), and other types of tourist flow data to discuss the construction of tourism demand forecasting models.

However, most previous studies assumed constant model parameters and could only reflect the overall impact of external events on tourism demand. However, in reality, the impact of external intervention would gradually change in real-time, according to policy and other factors. Therefore, this study draws on the concept of time-varying parameter (TVP) models to construct a TVP state-space model (TVP-SSM)-based forecasting framework utilizing multisource heterogeneous data, which aims to capture the dynamic impact of external shocks on hotel occupancy. The TVP models were transferred into the state-space model (SSM) in this framework. The dynamics of the impact of external events can be characterized by the different forms of state equations in the SSM.

To address this issue, this study proposes a new forecasting framework that leverages multiple sources of data. First, the COVID-19 pandemic is regarded as an external shock, and holiday variables are treated as periodic variables. Then, multisource data such as SED and official statistics are integrated to perform this study. Second, by constructing the TVP-SSM, the impact of the COVID-19 pandemic at different stages is incorporated into the state equations, capturing the dynamic effects of external events on tourism demand forecasting. Further, empirical analyses are conducted based on the daily data of Hangzhou and Zhoushan (China). Then comparative analyses and sensitivity analyses are performed to verify the performance of the proposed TVP-SSM model.

The remainder of the study is organized as follows. Section 2 presents the literature review. Section 3 introduces the methodology, while Section 4 describes the dataset and variables. Sections 5 and 6 provide the empirical analysis and robustness checks, respectively. Finally, Section 7 summarizes the conclusions and outlines directions for future research.

Literature review

Tourism demand forecasting with SED

Internet data, especially SED, have become increasingly popular for forecasting tourism demand (Li et al., 2020; Sun et al., 2022). Customers are prone to using online search engines, such as Baidu and Google, to plan their trips. The volume of search keywords serves as an early indicator of tourist activity, making it possible to use SED to anticipate fluctuations in tourism demand (Dergiades et al., 2018; Sun et al., 2019). Xiang and Pan (2011) analyzed the interplay between tourists' online queries and demand for a particular destination. They confirmed a direct relationship between growth in search queries and tourism demand.

Moreover, previous literature has shown that SED can significantly contribute to gains in prediction accuracy (Li & Law, 2019; Zhang et al., 2020; Wickramasinghe & Ratnasiri, 2021). For instance, Pan et al. (2012) directly introduced five search keywords into models to analyze the demand for hotel rooms. They found that including Google search data in the model could significantly enhance prediction accuracy. Similarly, Xie et al. (2021) and Zhang and Tian (2022) incorporated the Baidu Index into a least squares support vector regression model and gated recurrent unit network, respectively, showing that Baidu search data could effectively boost the prediction accuracy of tourism inflow. Zhang et al. (2023a) proposed a gated recurrent unit network-based deep learning framework that can learn the disturbance of events and the regular pattern of tourist arrival volume. To compare SED across different search engines, Yang et al. (2015) used data from Google and Baidu to model tourism inflows. They found that Google search data were more suitable for forecasting international tourism demand, whereas Baidu search data were more suitable for forecasting tourism demand within China.

Tourism demand forecasting under interventions

Tourism demand is easily influenced by exceptional external shocks, such as pandemics and terrorist attacks, resulting in poor prediction accuracy for time series models. Based on time series models, Box and Tiao (1975) pioneered an intervention analysis model to reflect the influence of exceptional external events, and this method is widely used to forecast tourism demand (Chen et al., 2007; Jiao & Chen, 2019). For example, Lai and Lu (2005) applied intervention analysis models to examine the impact of 9/11 on airline passenger flow in the United States. They found that passenger flow declined sharply within one to two months after 9/11, but the effect was sudden and temporary. Fan et al. (2023) explored how COVID-19 changed Chinese residents' travel behaviors and factors that influenced Chinese residents' actual travel and willingness to travel. Seong and Lee (2021) extended intervention analysis beyond the autoregressive integrated moving average (ARIMA) models to exponential smoothing models. They found that exponential smoothing models can more easily incorporate complex seasonality into the analysis. However, intervention analysis models are limited by the pattern and number of exceptional external events.

To address the limitations related to both the pattern and the frequency of exceptional external events, researchers have sought to incorporate their impact directly into models using dummy variables. Prilistya et al. (2021) used ARIMAX and seasonal dynamic regression models to analyze the influence of the COVID-19 pandemic on foreign tourists to Indonesia. Wu et al. (2022) proposed the mixed data sampling models (MIDAS) to monitor and forecast the hotel occupancy rates under the impact of the pandemic. Additionally, Yang et al. (2022) evaluated the usefulness of online search queries in boosting forecasting accuracy during the COVID-19 pandemic. They found that the usefulness of search queries in forecasting is associated with pandemic severity. However, these models, which assume constant parameters, can only reflect the average impact of the pandemic on tourist arrivals.

Tourism demand forecasting with the state-space model

Parameter estimation of TVP models can be realized by transforming the model form into a SSM (Athanasopoulos et al., 2011; Song et al., 2011). SSM can adjust variable coefficients in real time using the Kalman filter, which exhibits strong robustness (Kalman, 1960). For that reason, SSM has been applied to gross domestic product forecasting (Aruoba et al., 2016), stock forecasting (Monache, Petrella & Venditti, 2020), international trade (Srdelić & Dávila-Fernández, 2024), and currency price fluctuation forecasting (Shafiqah et al., 2022).

In the tourism field, Song and Wong (2003) constructed a TVP model by SSM to analyze the dynamic relationship between tourist income and tourism demand. Similarly, Wu et al. (2012) constructed an almost ideal demand system model with TVPs based on an SSM. They verified the dynamic relationship between the prices for tourism-related goods and services and their demand. Xiong et al. (2018) used SSMs to forecast China's inbound tourist arrivals from six countries. They found that income level exerts the most significant influence on tourist arrivals, and that SSMs outperform both linear regression and ARIMA models over longer time horizons.

An SSM can also effectively capture the dynamic impacts of exceptional external events on tourism demand. For instance, Jorge-Gonzalez et al. (2020) constructed a structural time series model with TVPs using an SSM to reflect the impact of five different exceptional external events on tourism flow.

Concluding remarks from the literature review

Table 1 summarizes the forms of models and the types of data included in previous studies. From Table 1, one can note that there has

Table 1
Structure of models in previous literature on tourism demand forecasting.

Literature	Time-varying parameter	Official statistics	SED	Periodical variables	Intervention variables
Li et al. (2020); Xie et al. (2021); Zhang and Tian (2022)			√		
Seong and Lee (2021)		√			√
Prilistya et al. (2021); Wu et al. (2022)		√			√
Yang et al. (2022); Zhang et al. (2023a)		√	√		√
Liu et al. (2021)	√	√	√		
Wu et al. (2012); Lee et al. (2020)	√	√			
Jorge-Gonzalez et al. (2020); Zhang et al. (2022b)	√	√			√
Our study	√	√	√	√	√

been progress in the sense of the variables and models applied. However, there are still open questions regarding tourism demand forecasting under external shocks. Therefore, specific literature gaps require extension of the existing methods.

First, the application of SED extends the scope of information that describes changes in tourism demand, improving the timeliness and accuracy of forecasting. However, although the search trends correlate with the occurrence of external shocks, SED cannot fully describe the effects of external events on tourism demand. Yang et al. (2022) have verified that SED was of little use in improving the accuracy of forecasting when the COVID-19 pandemic was severe. This study aims to use SED, together with variables capturing the impact of external events, to provide a more comprehensive description of the business environment.

Second, to accurately reflect the impact of external events on tourism demand, models such as the intervention analysis model, the ARIMAX model, and the MIDAS model have been proposed with relatively good prediction accuracy. However, the intervention analysis models are only used in conjunction with univariate models such as ARIMA and can only reflect the impact of a single external shock. The coefficients in ARIMAX and MIDAS models are fixed, which makes it impossible to simulate a flexible trend. This paper constructs TVP models using SSMs to capture the varying influence of extreme events. TVP-SSM can reflect the dynamics in the influence of external shocks according to changes in the variable parameters.

Third, existing studies have demonstrated that SSMs can better capture the dynamic dependence relationship between variables, making them suitable for reflecting the influence of exceptional external shocks on tourism demand. However, few studies have simultaneously utilized SSMs, SED, official statistics, intervention variables, and periodic variables to forecast tourism demand when external shocks influence tourism demand. In this study, a TVP-SSM is constructed to reflect the correlation between hotel occupancy, SED, and official statistics and the dynamic impact of intervention and periodical variables on hotel occupancy.

Research objectives

In recent years, research on tourism demand forecasting has made significant progress, and related studies have verified the effectiveness of SED (Zhang, 2023b; Wu et al., 2024), holiday variables (Hu et al., 2021a), etc., on tourism demand forecasting. However, most of the existing studies use univariate models for tourism forecasting (Apergis et al., 2016; Semenoglou et al., 2023), which negatively affects the accuracy of the forecasting model when external events impact it. The objective of this paper is to develop a tourism demand forecasting framework with strong dynamic forecasting capabilities under external event shocks by integrating multiple data sources. The main contributions of this study are as follows.

First, the prediction model integrates multisource data. In this study, different variables are integrated into the model in the form of explanatory variables or dummy variables, including external intervening variables, tourist flows, SED, holiday variables, weekend variables, etc., which reflect the impact of different factors on hotel occupancy comprehensively, and provide a firm support to improve the predictive ability of the model.

Second, a TVP model was constructed in state space form. Existing studies only capture the average impact of external events on forecasting results, without accounting for the dynamic evolution of these impacts. However, in real life, the impact of external events is multistage, and the impacts at each stage are dynamically changing (Sun et al., 2023). In this situation, governmental prevention and control as an intervention is also time-varying. The proposed TVP-SSM model can effectively reflect the dynamic impact of external events on hotel occupancy, further improving the accuracy of the forecasting model.

Third, the constructed prediction framework is scalable and generalized. The multisource data used in this study include official statistics, SED, intervention variables, and periodic variables. A time-varying parametric state-space model-based dynamic forecasting framework is constructed, which can effectively quantify the impact of various external events on tourism demand. The framework can be extended to other fields such as macroeconomic forecasting, traffic flow forecasting, etc. Based on this framework, other multisource data can be further considered and applied to other fields, such as macroeconomic forecasting, traffic flow forecasting, etc.

Methodology

State-space model

SSMs are generally applied to multivariate time series analysis. The multisource data used in this study include official statistics, SED, intervention variables, and periodic variables. SSMs estimate parameters according to the Kalman filter algorithm, which has accuracy and robustness (Kalman, 1960). Compared with time series models, such as ARIMA and Autoregressive Conditional Heteroskedasticity model, SSMs can reflect the dynamic dependence relationship between variables, and often have higher prediction accuracy (Shafiqah et al., 2022). The form of SSM is given as

$$y_t = z_t\alpha_t + d_t + u_t, t = 1, 2, ...T, \tag{1}$$

$$\alpha_t = \phi_t\alpha_{t-1} + \varepsilon_t, \tag{2}$$

$$u_t \sim NID(0, H_t), \varepsilon_t \sim NID(0, Q_t), \tag{3}$$

where Eqs. (1) and (2) are the measurement equation and state equation, respectively; y_t and α_t are an observable vector and an unobservable state vector, respectively; z_t is a measurement matrix, and d_t is a matrix of exogenous variables; u_t and ε_t are unrelated disturbance terms that are normally and independently distributed (mean is 0, and variances are H_t and Q_t respectively); Trepresents the number of periods covered.

From the expression in Eq. (2), it can be observed that the change follows a first-order autoregressive process. The TVP model can be transformed into SSMs based on Eqs. (1)–(3). The form of the TVP model is given as

$$y_t = \beta_0 + \sum_{i=1}^{m} x_{i,t}\beta_{i,t} + \sum_{j=1}^{n} z_{j,t}\gamma_j + u_t, t = 1, 2 \ldots T, \tag{4}$$

$$\beta_{i,t} = \phi_i \beta_{i,t-1} + \varepsilon_t, \tag{5}$$

where β_0 is the intercept; $\beta_{i,t}$ and γ_j denote time-varying and fixed coefficients to be estimated, respectively; $x_{i,t}$ and $z_{j,t}$ denote the exogenous variables with time-varying and fixed coefficients, respectively; m and n are the number of exogenous variables with time-varying and fixed coefficients, respectively. If the elements of the matrix ϕ_i in Eq. (5) are equal to unity, the state equation reduces to a random walk process (Song & Wong, 2003):

$$\beta_{i,t} = \beta_{i,t-1} + \varepsilon_t. \tag{6}$$

If the state equation can be expressed as a random walk, the parameter vector $\beta_{i,t}$ is considered nonstationary.

LASSO regression model

To handle big data, the LASSO regression can be used for dimensionality reduction (Uniejewski et al., 2019; Tian et al., 2021). The LASSO model contracts the regression coefficients by constructing a penalty function (Tibshirani, 1996). When the coefficient of a variable decreases to zero, that variable is excluded from the model. Based on the regression function $y_t = \beta_0 + \sum_{j=1}^{N} \beta_j x_{j,t}$, the estimator of the LASSO regression model is given as

$$\widehat{\beta}_j = \underset{\beta}{\operatorname{argmin}} \left\{ \underbrace{\frac{1}{n}\sum_{t=1}^{n}\left(y_t - \beta_0 - \sum_{j=1}^{N}\beta_j x_{j,t}\right)^2}_{I} + \underbrace{\frac{\lambda}{n}\sum_{j=1}^{N}|\beta_j|}_{II} \right\} \tag{7}$$

On the right-hand side of Eq. (7), terms I and II represent the residual sum of squares and penalty term, respectively; λ denotes the penalty level, N stands for the number of variables and n is the sample size. The value of λ is critical in LASSO regression. According to Schwarz (1978), the Bayesian information criterion is used to determine the appropriate value of λ.

Data transformation

Multiple variables have different orders of magnitude and dimensions. Directly incorporating these variables into models will affect the accuracy (Zhang & Lu, 2022). Therefore, normalization should be performed prior to analysis. This study employs linear min–max normalization to enhance training efficiency. The final retained keywords are indexed by $j = 1, 2 \ldots N$and periods by $t = 1, 2 \ldots T$. Then, x^*_{jt} is the number of queries with the j-th keyword during period t. The normalization is carried out as follows:

$$x_{jt} = \frac{x^*_{jt} - \underset{j}{min} x^*_{jt}}{\underset{j}{max} x^*_{jt} - \underset{j}{min} x^*_{jt}}, \tag{8}$$

where $min_j x^*_{jt}$ and $max_j x^*_{jt}$ denote operators returning the smallest and largest values of x^*_{jt}, respectively. The combined keyword variable (x_t) is then obtained as

$$x_t = \sum_{j=1}^{N} x_{jt}. \tag{9}$$

Evaluation metrics

When evaluating the predictive accuracy of forecasting models, the root mean square error (RMSE) and mean absolute percentage error (MAPE) are the two most commonly used measures. The calculation of the two measures proceeds as follows:

$$RMSE\left(y_t, \widehat{y}_t\right) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(y_t - \widehat{y}_t\right)^2}, \tag{10}$$

$$MAPE\left(y_t, \widehat{y}_t\right) = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{y_t - \widehat{y}_t}{y_t}\right| \times 100\%, \tag{11}$$

where y_t and $\widehat{y}_t$ represent the actual value and predicted value during the forecast period, respectively; T is the length of the forecast period.

To avoid prediction contingency arising from data sampling, the Diebold and Mariano (DM) test is employed to assess whether there is a significant difference between the two forecasting models. The DM test relies on the loss differential (D_t) to compare prediction ability of the two models (Zhang et al., 2021a):

$$D_t = L(e_{1,t}) - L(e_{2,t}), \tag{12}$$

where $L(e_{i,t})$ is a loss function (this study uses RMSE and MAPE) and $e_{i,t}$ represents the residual at time t calculated based on model i.

Based on Eq. (12), the DM statistic is obtained as follows:

$$DM = \frac{\overline{D}_t}{\sigma_{D_t}}. \tag{13}$$

where $\overline{D}_t$ and σ_{D_t} represent the mean and standard deviation of D_t, respectively. The standard deviation is calculated following Diebold and Mariano (2002), and the resulting sampling distribution follows a standard normal distribution. If $DM < 0$, then Model 1 outperforms Model 2 in terms of prediction ability.

Forecasting framework

This study proposes an integrated approach (Fig. 1) that incorporates SED and intervention variables into hotel occupancy forecasts.

This framework includes the following five steps:

Step 1: Selection of search keywords. The basic keywords are determined based on the four elements of tourism (food, accommodation, transportation, and attractions). Then, automatic recommendation technology (Baidu Index) is used to obtain extended keywords.
Step 2: Constructing the variables. First, the search keywords in Step 1 are selected by considering the Pearson correlation coefficient and LASSO regression (where hotel occupancy is used as a covariate or the dependent variable) and then aggregated to obtain SED variables. Second, the stationarity and cointegration tests are performed on the SED variables. Finally, the intervention and periodic variables are specified in the form of an impulse function.
Step 3: Estimation of models. After constructing the TVP-SSM with the variables specified in Step 2, the Maximum Likelihood estimation and Kalman filter are applied to estimate its parameters. The significance is also tested.
Step 4: Comparison of forecasting performance. Based on RMSE, MAPE, and DM statistics, the prediction results of TVP-SSM with intervention and periodical variables (TVP-SSM-1) are compared with those of benchmark models to verify the superiority of TVP-SSM-1.
Step 5: Robustness analysis. Based on the Hangzhou and Zhoushan data sets, three cases are designed to further validate the robustness of the TVP-SSM model: Case 1 changes the ratio of the training set and testing set, Case 2 replaces the response variables, and Case 3 changes the data set.

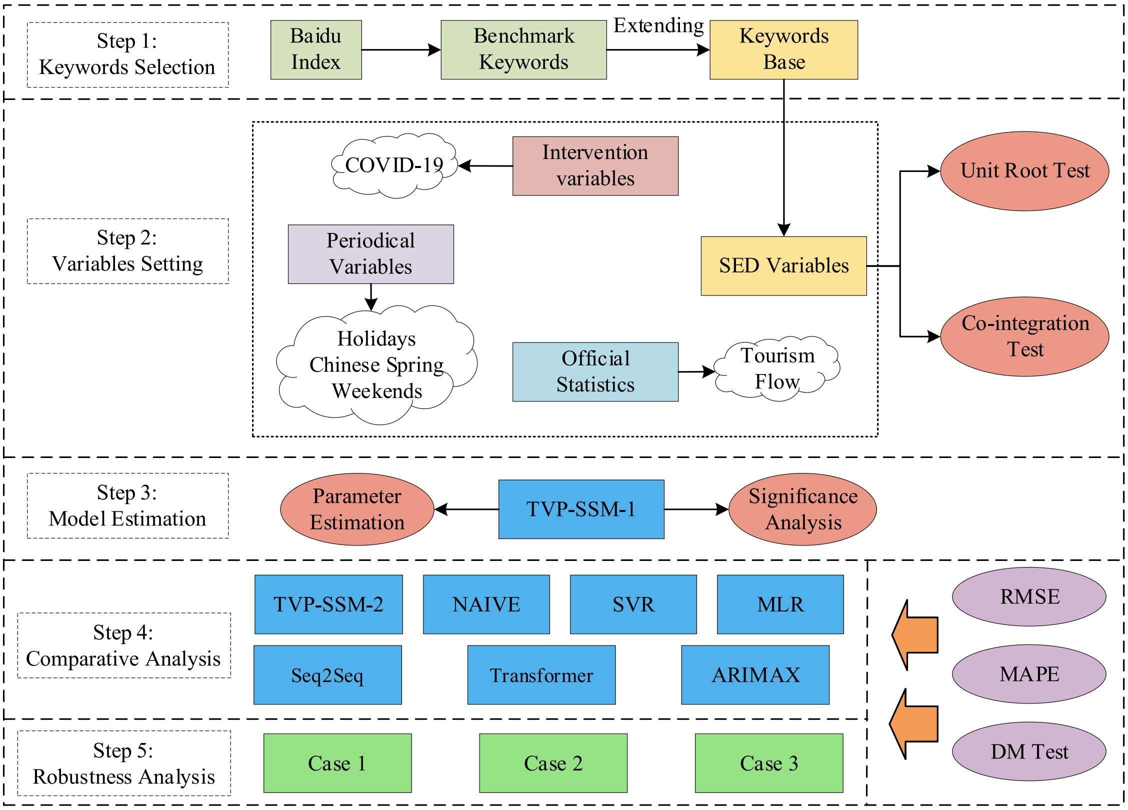

Fig. 1. Forecasting framework for hotel occupancy.

Data and variables

This study is based on multisource data, including official statistics, SED, intervention variables, and periodic variables. Official statistics, including daily hotel occupancy in Hangzhou, tourist flow in Hangzhou Hubin Road (the core scenic spot of West Lake), and passenger arrivals at Hangzhou Xiaoshan International Airport, come from routine monitoring of the hospitality industry by the Hangzhou government and Hangzhou Hotel Association from October 1, 2019, to October 28, 2021 (759 days in total). SED were collected from the Baidu Index (http://index.baidu.com), and the specific process is shown in Appendix A. Intervention variables and periodical variables are dummy variables that reflect the impact of the COVID-19 pandemic and holidays on daily hotel occupancy in Hangzhou.

Intervention variables

To reflect the impact of the intervention variables on hotel occupancy, we introduced COVID-19 state variables as intervention variables. Since the impact of the pandemic was closely tied to the trends in virus transmission in China, the different stages of the pandemic should be taken into account when defining the COVID-19 state variables. Based on the level of the local government's response to the pandemic, domestic pandemic transmission can be generally divided into two stages: the comprehensive outbreak stage and the sporadic outbreak stage (including stages I and II; the details are shown in Table 2). The COVID-19 state variables take the form of an impulse function, i.e., their value is set to one during the influence period and zero during other periods. The specific expression is shown in Eq. (14):

Table 2
Intervention and periodical variables for forecasting hotel occupancy in Hangzhou.

Variable Category	Variable	Notation	Value	Period	Remark
Intervention variables	COVID-19 state variables	$D_{1,t}$	1	2020.01.20–2020.06.30	Comprehensive outbreak stage
				2021.01.16–2021.02.28	Sporadic outbreak stage I
				2021.07.26–2021.08.31	Sporadic outbreak stage II
			0	Otherwise	——
Periodical variables	Holidays and festivals	$D_{2,t}$	1	2019.10.01–2019.10.06, 2020.09.30–2020.10.06, 2021.09.30–2021.10.06	National Day
				2019.12.31–2020.01.01, 2020.12.31–2021.01.01	New Year's Day
				2020.05.01–2020.05.04, 2021.05.01–2021.05.04	Labor Day
				2020.06.24–2020.06.26, 2021.06.11–2021.06.13	Dragon Boat Festival
				2021.04.02–2021.04.04	Tomb-Sweeping Day
				2021.09.19–2021.09.21	Mid-Autumn Festival
			0	Otherwise	——
	Chinese Spring Festival	$D_{3,t}$	1	2020.01.17–2020.01.31	Chinese New Year
				2021.02.4–2021.02.18	
			0	Otherwise	——
	Weekend	$D_{4,t}$	1	Friday and Saturday	——
			0	Otherwise	——

$$D_{i,t} = \begin{cases} 1, \text{influenceperiods} \\ 0, \text{otherperiods} \end{cases} . \tag{14}$$

Periodical variables

In this study, periodical variables include holiday variables and weekend variables. Holiday variables refer to seven legally acknowledged holidays in China, including New Year's Day, Spring Festival, Tomb-Sweeping Day, Labor Day, Dragon Boat Festival, Mid-Autumn Festival, and National Day. During these days, two to seven days off are allowed, and residents often opt for traveling. Therefore, holidays and weekends are key factors influencing tourism demand, with hotel occupancy typically expected to rise during these periods compared to regular days. The holiday and weekend variables in the forecasting model are introduced in the form of periodical variables, which are set by the same rules as Eq. (14).

Additionally, the setup of holiday variables should be done with consideration of the following two aspects. First, hotel occupancy in Hangzhou during Labor Day and National Day is markedly higher compared to that on a typical day. Second, the Spring Festival is one of the most important traditional festivals in China, and most people choose to visit their hometowns for family gatherings. Therefore, certain holidays can also lead to reduced hotel occupancy. In this light, the Spring Festival effect is set separately from other holiday's variables.

Weekend variables refer to Friday and Saturday, not Saturday and Sunday, because people's hotel check-in arrangements often depend on the day ahead. Saturday is a weekend, and most people check in early on Friday, while Monday is a weekday and most people do not stay at a hotel on Sunday. This can be observed during holidays and weekends when hotel demand fluctuates, so it is essential to set up periodic variables one day in advance for the time of impact. Table 2 presents the periods associated with intervention variables and periodic variables.

Model specification

This study models the state equation of SED variables as a first-order autoregressive process and the state equation of intervention variables as a random walk process. The TVP-SSM can be defined as follows:

$$lnY_t = \beta_0 + \gamma_1 lnX_{1,t} + \gamma_2 lnX_{2,t} + \sum_{i=1}^{4} \beta_{i,t} D_{i,t} + \beta_{5,t} lnX_{3,t} + u_t \tag{15}$$

$$\beta_{i,t} = \beta_{i,t-1} + \eta_t, i = 1, 2...4 \tag{16}$$

$$\beta_{5,t} = \phi\beta_{5,t-1} + \varepsilon_t \tag{17}$$

where Y_t is the hotel occupancy in Hangzhou, β_0 is the intercept; $X_{1,t}$ is tourist flow on Hangzhou Hubin Road, $X_{2,t}$ is passenger arrivals of Hangzhou Xiaoshan International Airport, $X_{3,t}$ is the combined SED variable, $D_{i,t}(i = 1\cdots 4)$ are all dummy variables, including intervention variables and periodical variables. $\beta_{i,t}(i = 1, 2, 3, 4)$ is the dynamic impact of the COVID-19 pandemic and holiday on hotel occupancy; $\beta_{5,t}$ is the relationship between hotel occupancy and the composite SED variable; u_t, η_t and ε_t are uncorrelated disturbance terms. γ_1, γ_2 are fixed parameters, reflecting the influence between $X_{1,t}$, $X_{2,t}$ and Y_t, respectively.

Empirical analysis

Model setup

The empirical analysis was carried out on a PC integrated with a Core I9 CPU, 16 G RAM, and the Windows 11 system, and the development environment was EViews 9. The Hangzhou hotel occupancy data set was divided into training and testing sets with a ratio of 743:16. This division ensures that the characteristics of COVID-19 can be captured in both the training and testing phases, allowing the influence of intervention variables on the results to be reflected in the model. Then, 4-step, 8-step, 12-step, and 16-step forward forecasting models were constructed using TVP-SSM-1, ARIMAX, Sequence To Sequence (Seq2Seq), Transformer, multiple linear regression (MLR), support vector regression (SVR), and naïve (NAIVE) models, respectively, to analyze the performance of different models for hotel occupancy prediction. Additionally, to verify the importance of multisource data in hotel occupancy prediction, the TVP-SSM-2, TVP-SSM-3, and TVP-SSM-4 models were constructed. Among them, TVP-SSM-2 only considers official statistics and SED, while TVP-SSM-3 only considers periodic variables, and TVP-SSM-4 only considers SED.

Descriptive analysis

As shown in Fig. 2, there are three troughs (A, B, and C) in the dynamics of daily hotel occupancy in Hangzhou during the covered period. Troughs A and B are due to the influence of the Spring Festival and the COVID-19 pandemic, and trough C is due to the outbreak of the COVID-19 pandemic.

According to Appendix A, keywords with correlation coefficients greater than 0.6 related to hotel occupancy are listed in Table 3. It can be concluded that there is a strong correlation between the retained query keyword volumes and hotel occupancy. The lag order of keyword variables ranges from zero to two, indicating that most relevant queries occur at most two days before the trip. Finally, the volumes of query keywords in Table 3 were normalized and aggregated in the combined keyword variable based on Eqs. (8)–(9). As shown in Fig. 3, the hotel occupancy and combined keyword variable generally exhibited similar fluctuation patterns, particularly in terms of peaks and troughs.

Model estimation

For econometric models, the results of the stationarity test are shown in Appendix B, and common parameter estimation methods include the least squares method, the maximum likelihood estimation method, etc. Compared to them, the Kalman filter posits that the observed value at the previous time will have an impact on the next time, which is more suitable for the real-time processing of dynamic data (Kalman, 1960). The Kalman filter is an iterative algorithm that continuously estimates and corrects parameters in the state equation based on the Kalman gain according to the difference between the observed value and the actual value.

In this study, the Kalman filter is used to estimate the parameters of the TVP-SSM-1. The algorithm is implemented by EViews 9, and convergence is achieved after 118 iterations. Ultimately, the optimal

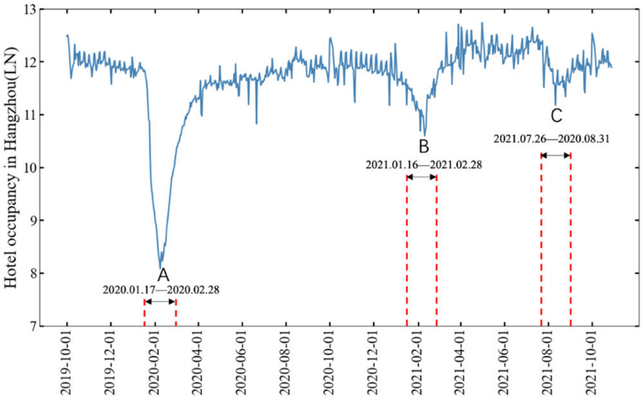

Fig. 2. Daily hotel occupancy in Hangzhou from October 1, 2019, to October 28, 2021.

Table 3
Correlation coefficients between query keyword volumes and hotel occupancy.

Keyword	Correlation Coefficient	Lag Order	Keywords	Correlation Coefficient	Lag Order
Hangzhou food	0.8024	1	Hangzhou tourism	0.6568	1
Hangzhou specialty	0.7648	1	Boat of West Lake	0.6530	2
Hangzhou snacks	0.7573	2	Hangzhou attractions	0.6388	2
Hangzhou Song City	0.6942	1			

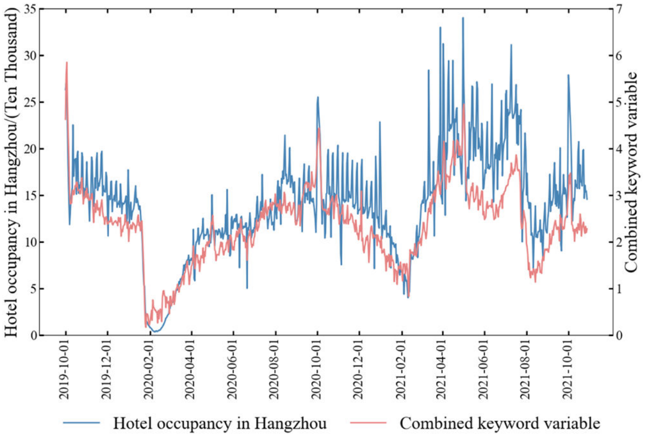

Fig. 3. Hotel occupancy and combined keyword variable in Hangzhou.

estimated coefficients for the various variables, including SED, COVID-19 state, and holiday variables, are obtained and presented in Table 4.

According to Table 4, all variables are significant at the 10 % level, suggesting that they are important determinants of hotel occupancy. Among them, official statistics (X_{1t} and X_{2t}), SED (X_{3t}), holiday variables (D_{2t}) and weekends (D_{4t}) show positive effects on hotel occupancy in Hangzhou, while COVID-19 state variables (D_{1t}) and Chinese Spring Festival (D_{3t}) show adverse effects.

During the COVID-19 pandemic, tourism was severely restricted, leading to a significant decline in hotel occupancy and a negative overall impact on the industry. A similar effect was obtained by Ozdemir et al. (2022).

As expected, the impact on hotel occupancy varied by holiday, with occupancy during Labor Day and National Day increasing significantly compared with regular days. On the contrary, hotel occupancy during the Chinese Spring Festival was significantly lower compared with typical days. These patterns were also noted by Liu et al. (2022); however, they have specific differences depending on the destination.

Table 4
Results of parameter estimation of TVP-SSM-1.

Variable	Coefficient	Standard error	Z-Statistic	P-value
Intercept	0.4765	0.2759	1.7274	0.0841*
$\ln(X_{1,t})$	0.3599	0.0203	17.6992	0.0000***
$\ln(X_{2,t})$	0.6087	0.0164	37.2285	0.0000***
$\ln(X_{3,t})$	0.4918	0.1223	4.0232	0.0001***
$D_{1,t}$	−0.0830	0.0243	−3.4153	0.0006***
$D_{2,t}$	0.0318	0.0420	1.9577	0.0503**
$D_{3,t}$	−0.1821	0.0466	−3.9064	0.0001***
$D_{4,t}$	0.0223	0.0195	1.6523	0.0985*

Note: All coefficient estimates are end-state estimates. ***, **, and * represent 1 %, 5 %, and 10 % significance levels, respectively.

Simulating the impact of dummy variables on hotel occupancy

TVP-SSM-1 updates the coefficients in real time using a Kalman filter, which allows it to capture the dynamic impacts of intervention and periodic variables on hotel occupancy in Hangzhou at different stages, ensuring that these effects align with real-world conditions.

(1). Impact of intervention variable on hotel occupancy

The impact of the COVID-19 pandemic on hotel occupancy in Hangzhou during the three phases is presented in Fig. 4. Among them, subfigure (a) in Fig. 4 indicates that during the comprehensive outbreak stage, the pandemic's influence on Hangzhou's hotel occupancy rate initially increased and then decreased. As the pandemic began to spread rapidly across the country, people's travel was severely restricted, leading to a significant decline in hotel occupancy rates, with the impact steadily increasing over time. From mid to late February 2020, the impact reached its peak, indicating the most substantial effect on hotel occupancy rates. Subsequently, with the effectiveness of a series of domestic prevention and control measures, the influence began to decline and gradually stabilized after March 23, 2020.

Subfigures (b) and (c) in Fig. 4 show that the impact gradually wanes during the sporadic outbreak stages I and II. In both stages, although local outbreaks initially had an impact on hotel occupancy, the timely implementation of prevention and control measures led to a continuous reduction in the impact. There were some minor fluctuations during these periods, but the overall tendency was a decrease in the impact. This suggests that in the face of sporadic outbreaks, preventive and control measures can effectively mitigate the adverse effects of the pandemic on the hotel occupancy rate, reducing both the scope and degree of impact.

(2). Impact of periodical variables on hotel occupancy

Given that the holiday effect of longer vacations on hotel occupancy is more realistic, this study analyzes the impact of two of China's most important traditional holidays: National Day and the Spring Festival.

Subfigure (a) in Fig. 5 shows that during National Day in 2020 and 2021, the impact of holiday effects on hotel occupancy in Hangzhou first increased and then decreased. Among them, October 3 was the most affected, with the highest hotel occupancy. Additionally, the impact of National Day on hotel occupancy in 2020 was significantly higher than that of 2021, mainly because transmission of the COVID-19 pandemic was controlled during National Day in 2020. It was more convenient for tourists to travel, which led to a "revenue spending" trend in the tourism industry. However, the pessimistic situation of the pandemic during the National Day of 2021 reduced hotel occupancy in Hangzhou.

Subfigures (b) and (c) in Fig. 5 present the impact of the Spring Festival effect on hotel occupancy in Hangzhou for the Spring Festival during 2020 and 2021, respectively. The impact of the Chinese New Year effect generally shows a decreasing trend. In these two figures, the impact stays at a high level in the week before the Chinese New Year and falls back quickly in the week after the Chinese New Year. Comparing 2020 and 2021, it is found that there is no significant difference in the impact of the Spring Festival on hotel occupancy in Hangzhou.

Strategies for forecasting and comparative experiments

The strategy for prediction is to re-estimate the model by adding one day to the forecasting horizon. Ultimately, we obtain 13 four-step-ahead forecasts, nine eight-step-ahead forecasts, five twelve-step-ahead forecasts, and one sixteen-step-ahead forecast. For the out-of-sample prediction, the lagged actual values of SED were used.

To verify the performance ability of the proposed model, two groups of comparative experiments were constructed. In the first experiment, the forecasting results of the TVP-SSM-1, TVP-SSM-2, TVP-SSM-3, and

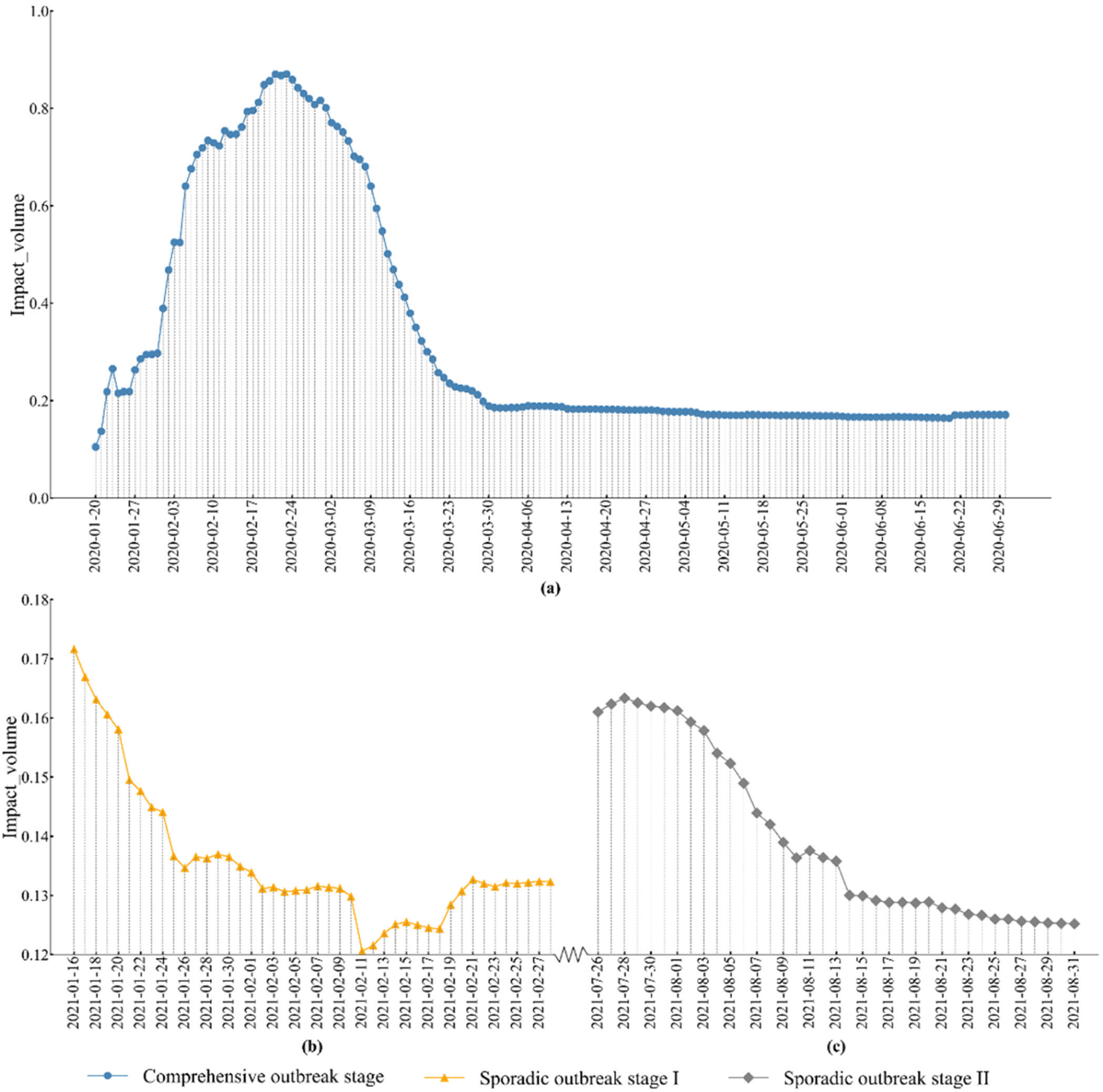

Fig. 4. Impact of COVID-19 on hotel occupancy in Hangzhou during three stages.

TVP-SSM-4 were compared to demonstrate whether the inclusion of multiple source data, such as intervention variables and periodic variables, could improve prediction accuracy. In the second experiment, to verify the higher prediction accuracy of TVP-SSM-1, the forecasting results of TVP-SSM-1 with ARIMAX, Seq2Seq, Transformer, MLR, SVR, and NAIVE models are compared. In these two experiments, the treatment of other parts remained the same, except for the use of different forecasting models.

Error analysis

The results of RMSE and MAPE for different models, presented in Table 5, show that the four-step-ahead forecast ($h = 4$) of TVP-SSM-1 has the smallest MAPE of 0.038, and the sixteen-step-ahead ($h = 16$) forecast of TVP-SSM-1 has the smallest RMSE of 0.6967. Compared to TVP-SSM-2, the RMSE and MAPE for TVP-SSM-1 are lower across all forecasting horizons. The inclusion of intervention variables increases the average RMSE and MAPE by 74 % and 76 %, respectively. Compared with TVP-SSM-3 and TVP-SSM-4, the average RMSE and MAPE rise by 93 % and 92 %, respectively. This suggests that the use of multisource data, such as intervention variables, may significantly improve the prediction accuracy. Additionally, the RMSE and MAPE for the TVP-SSM-1 are smaller than those for the ARIMAX, Seq2Seq, Transformer, MLR, SVR, and NAIVE models across different forecasting horizons. Especially compared with the NAIVE model, the average prediction accuracy of TVP-SSM-1 (as measured by the RMSE and MAPE) is higher by 86 % and 87 %, respectively.

The DM statistics in Table 6 indicate that the DM test rejects the null hypothesis at varying levels of significance. Compared with the ARIMAX model, the DM test is positive when $h = 8$. The reason is that the RMSE

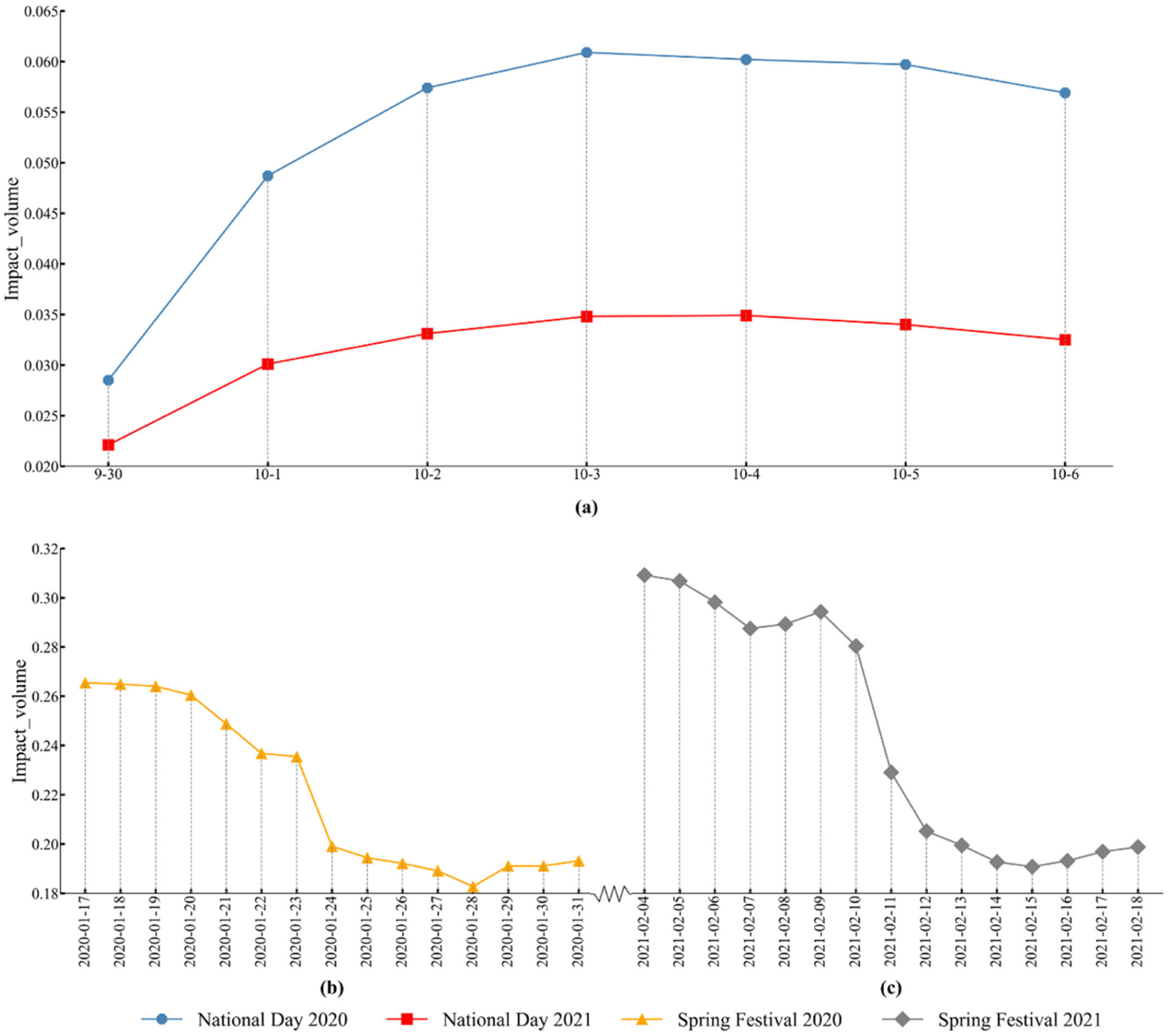

Fig. 5. Impact of holiday effects on hotel occupancy in Hangzhou.

Table 5
RMSE and MAPE for different models and forecasting horizons.

Model	h = 4		h = 8		h = 12		h = 16	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
TVP-SSM-1	**0.7804**	**0.0388**	1.0228	0.0523	**0.7420**	**0.0389**	**0.6967**	**0.0339**
ARIMAX	0.8614	0.0453	**0.9335**	**0.0492**	0.9244	0.0489	0.8636	0.0464
Seq2Seq	5.2250	0.3080	5.2197	0.3076	5.2228	0.3079	5.2193	0.3076
Transformer	6.9519	0.4125	6.9465	0.4122	6.9530	0.4126	6.9414	0.4118
MLR	3.0154	0.1655	3.0020	0.1652	3.0492	0.1657	2.9698	0.1623
SVR	2.7914	0.1440	2.7724	0.1435	2.7832	0.1436	2.8068	0.1463
NAIVE	6.3988	0.3238	5.6382	0.3005	5.7615	0.3125	5.2666	0.2800
TVP-SSM-2	3.0707	0.1658	3.2791	0.1847	2.9292	0.1618	2.9144	0.1634
TVP-SSM-3	13.9031	0.7955	13.9026	0.7954	13.903	0.7955	13.9015	0.7952
TVP-SSM-4	7.7029	0.3417	8.4239	0.3494	10.197	0.3915	8.5094	0.3552

Note: The best statistical scores are marked in boldface.

and MAPE values of the TVP-SSM-1 model are higher than those of the ARIMAX model (see Table 5); however, the difference may be negligible. Whereas the RMSE and MAPE values of the TVP-SSM-1 model are lower than those of the ARIMAX model when h = 4,12,16. Compared with the other five models, the TVP-SSM-1 shows significantly superior prediction results. Hence, it is reasonable to conclude that TVP-SSM-1 performs better.

Several factors may explain the high prediction errors observed in

Table 6
DM test based on RMSE and MAPE for comparing TVP-SSM-1 against competing models with different forecasting horizons.

Competing model	$h = 4$		$h = 8$		$h = 12$		$h = 16$	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
ARIMAX	−2.2401**	−1.9015*	2.1900*	1.8601*	−2.1899**	−1.8586*	−2.1862**	−1.8545**
Seq2Seq	−2.5055**	−3.5600***	−2.3868**	−3.1345***	−2.6330***	−4.6009***	−2.5813***	−4.1168***
Transformer	−3.2841***	−5.0372***	−3.2120***	−4.6004***	−3.3944***	−5.9728***	−3.3584***	−5.6343***
MLR	−4.8874***	−3.9122***	−4.5374***	−2.9850***	−4.2554***	−4.3588***	−3.9578***	−4.1624***
SVR	−2.9902***	−3.9158***	−2.6503**	−2.7344***	−2.9367***	−3.4671***	−3.0304***	−3.6619***
NAIVE	−2.3499**	−3.1836***	−2.4427**	−3.4810***	−2.2922**	−2.8275***	−2.0856**	−2.5881***
TVP-SSM-2	−3.7130***	−2.3675**	−4.1948***	−2.7199***	−2.4399**	−1.8468*	−2.5880***	−1.8466**
TVP-SSM-3	−4.7922***	−6.1003***	−4.7917***	−6.0992***	−4.7921***	−6.1004***	−4.7905***	−6.0964***
TVP-SSM-4	−2.5030***	−3.4181***	−2.8504***	−3.8787***	−3.4108***	−3.7904***	−2.8603***	−3.8948***

Note: ***, **, and * represent 1 %, 5 %, and 10 % significance levels, respectively.

the TVP-SSM-2, TVP-SSM-3, TVP-SSM-4, ARIMAX, Seq2Seq, Transformer, MLR, SVR, and NAIVE models. First, the TVP-SSM-2 does not contain intervention variables, and the SED variables alone cannot fully reflect the impact of the pandemic on hotel occupancy. TVP-SSM-3 relies solely on periodic variables; TVP-SSM-4 incorporates only SED data. Second, the ARIMAX model predicts the current value based on historical data, and a lack of shocks in the historical data is likely to inflate the predicted value under the pandemic restrictions. Third, the MLR model is a fixed-parameter model, which cannot dynamically describe the impact of interventions on hotel occupancy. Fourth, although the SVR model can reflect the nonlinear relationship between variables, it cannot accurately learn such a nonlinear relationship because the intervention variables are dummy variables and contain less information than continuous variables. Fifth, the NAIVE models are univariate time series models, making predictions based solely on historical data without the ability to reflect the influence of unexpected external shocks. Sixth, Seq2Seq (Lu et al., 2024) and Transformer (Ma et al., 2023) models are deep learning models that are computationally intensive and time-consuming. These models require massive training data for optimal forecasting performance. Additionally, the impacts of external shocks are unexpected, which may render the patterns learned from the training data inapplicable to the observed samples influenced by external shocks, affecting the accuracy of predictions.

Robustness analysis

In order to further validate the robustness of the TVP-SSM model, three cases are designed: Case 1 changes the ratio of the training set and testing set; Case 2 replaces the response variables; and Case 3 changes the data set. The data set for Hangzhou is utilized in Cases 1 and 2, while Case 3 uses the data set for Zhoushan. The variables and data sets in each case are shown in Table 7.

Table 7
Variables and data set for robustness testing.

Data set	Case 1 Hangzhou Data set	Case 2	Case 3 Zhoushan Data set
Response variable	Hotel occupancy in Hangzhou	Revenue of the hotel industry in Hangzhou	Revenue of the hotel industry in Zhoushan
Official statistics	Tourist flow on Hangzhou Hubin Road Passenger arrivals at Hangzhou Xiaoshan International Airport	Passenger arrivals at Hangzhou Xiaoshan International Airport	Passenger arrivals at Zhoushan Putuoshan Airport
SED	Baidu Index	Baidu Index	Baidu Index
Intervention variable	COVID-19 state variables	COVID-19 state variables	COVID-19 state variables
Periodical variable	Holidays and festivals Weekend	Holidays and festivals Weekend	Holidays and festivals Weekend

Case 1: robustness analysis based on modifying the ratio for data splitting

To verify the robustness of the model, the ratio of the training set to the testing set was categorized into four types (739:20, 743:16, 747:12, and 751:8), and predictions were performed using the settings outlined in Section 5.1.

As shown in Table 8, from a horizontal perspective, there are differences in the prediction results of the same model under varying data splitting ratios. Taking TVP-SSM-1 as an example, when the split ratio is 739:20, the MAPE is 0.1306; while when the ratio changed to 743:16, the MAPE is only 0.0319, which indicates that the ratio for data splitting does affect the prediction results; from the longitudinal perspective, the RMSE and MAPE of the TVP-SSM-1 model are smaller than those of the other benchmark models under different ratios for data splitting.

As shown in Table 9, the null hypothesis of the DM test is rejected at the 10 % significance level, indicating a significant difference in the prediction results of the TVP-SSM-1, TVP-SSM-2, ARIMAX, Seq2Seq, Transformer, MLR, SVR, and NAIVE models under different data-splitting ratios. Additionally, all the DM statistics are negative, which further indicates that the TVP-SSM-1 model has better prediction accuracy than the benchmark models. However, it is worth noting that at 8 and 12 steps, the DM test results of TVP-SSM-1 versus TVP-SSM-2 are not significant, mainly because when the forecasting step is short, the TVP-SSM model fails to adequately learn the features of the existing data. The advantages of the time-varying features in the TVP-SSM become evident only when the forecasting horizon is long.

Case 2: robustness analysis based on different variables

In Case 2, the revenue of the hotel industry in Hangzhou is regarded as the response variable. A forecasting model is then constructed, and it is verified that the TVP-SSM exhibits an excellent predictive effect for different response variables.

Appendix C demonstrates the significant correlation and causality between hotel consumption in Hangzhou and passenger arrivals of Hangzhou Xiaoshan International Airport ($X_{2,t}$) and composite SED variable ($X_{3,t}$), which further validates the reasonableness of using passenger arrivals of Hangzhou Xiaoshan International Airport ($X_{2,t}$) and composite SED variable ($X_{3,t}$) to predict hotel consumption in Hangzhou.

The average values of RMSE and MAPE for the TVP-SSM-1 model are 0.2127 and 0.0685, respectively, and these values are significantly smaller than those of the other benchmark models at different prediction steps, indicating that the TVP-SSM model performs well. Additionally, the DM statistics of Case 2 in Table 10 are all negative, indicating that the prediction results of TSVP-SSM-1, TSVP-SSM-2, ARIMAX, Seq2Seq, Transformer, MLR, SVR, and NAIVE models are significantly different under different prediction steps. The prediction accuracy of the TSVP-SSM-1 model is higher than that of other benchmark models.

Table 8
Comparison of RMSE and MAPE under different ratios for data splitting.

Model	739:20		743:16		747:12		751:8	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
TVP-SSM-1	**2.3069**	**0.1306**	**0.6967**	**0.0339**	**0.7855**	**0.0395**	**0.9007**	**0.0502**
ARIMAX	2.5375	0.1418	0.8636	0.0464	0.8343	0.0447	1.1469	0.0636
Seq2Seq	5.3053	0.3185	5.2193	0.3076	5.2476	0.3137	4.6211	0.2826
Transformer	6.8578	0.4127	6.9414	0.4118	6.9727	0.4213	6.4974	0.3999
MLR	3.2257	0.1827	2.9698	0.1623	2.9677	0.1676	2.4677	0.1448
SVR	2.8996	0.1474	2.8068	0.1463	2.4986	0.1343	2.3279	0.1236
NAIVE	3.7575	0.1735	5.2666	0.2800	3.8532	0.2044	2.0587	0.0730
TVP-SSM-2	4.0900	0.1969	2.9144	0.1634	2.9292	0.1579	4.4576	0.2343
TVP-SSM-3	14.2192	0.8362	13.9015	0.7952	14.5940	0.8623	13.4078	0.7935
TVP-SSM-4	7.9396	0.3277	8.5094	0.3552	9.5390	0.4022	4.4619	0.2147

Note: The best statistical scores are marked in boldface.

Table 9
DM test results under different ratios for data splitting.

TVP-SSM-1 VS.	739:20		743:16		747:12		751:8	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
ARIMAX	−2.8958***	−3.9648***	−2.0800**	−2.4146**	−3.2655***	−3.2147***	−2.6907**	−2.4873**
Seq2Seq	−2.6044***	−3.7653***	−2.5813***	−4.1168***	−2.2139**	−3.1417***	−2.5410**	−2.8445***
Transformer	−3.2268***	−5.0700***	−3.3584***	−5.6343***	−3.1802***	−4.6624***	−5.0445***	−4.3880***
MLR	−3.7276***	−2.5694**	−2.9931***	−2.4052**	−2.8027***	−3.5838***	−3.0104***	−3.0938***
SVR	−4.4316***	−3.3288***	−2.2337**	−2.4246**	−2.6008**	−2.7420**	−3.8739***	−3.8734***
NAIVE	−1.7864*	0.1950	−2.0856***	−2.5881***	−1.7243*	−2.1131**	−2.6505***	−3.8839***
TVP-SSM-2	−1.8970*	−1.6868	−2.5880***	−1.8466*	−0.7561	−1.2223	−1.2076	−1.1410
TVP-SSM-3	−5.0431***	−7.3936***	−4.7905***	−6.0964***	−5.219***	−5.5319***	−4.8028***	−5.9891***
TVP-SSM-4	−3.511***	−3.9736***	−2.8603***	−3.8948***	−3.5505***	−3.9756***	−2.7757***	−2.1737**

Note: ***, **, and * represent 1 %, 5 %, and 10 % significance levels, respectively.

Table 10
Comparison of RMSE and MAPE at different prediction steps.

Case	Model	h = 4		h = 8		h = 12		h = 16	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Case 2	**TVP-SSM-1**	**0.1923**	**0.0611**	**0.2289**	**0.0732**	**0.2172**	**0.0708**	**0.2123**	**0.0687**
	TVP-SSM-2	1.4781	0.4052	1.9013	0.6182	2.2395	0.7203	2.2907	0.8077
	TVP-SSM-3	1.3809	0.4915	1.3808	0.4913	1.3807	0.491	1.3807	0.491
	TVP-SSM-4	0.3205	0.082	0.3203	0.082	0.3206	0.082	0.3204	0.082
	ARIMAX	0.2776	0.0801	0.2910	0.0913	0.2934	0.0893	0.2886	0.0835
	Seq2Seq	0.7765	0.2797	0.7750	0.2791	0.7760	0.2794	0.7751	0.2790
	Transformer	2.3946	0.8938	2.3930	0.8933	2.3906	0.8924	2.3814	0.8890
	MLR	0.3588	0.1141	0.3583	0.1138	0.3588	0.1141	0.3602	0.1147
	SVR	0.3198	0.0956	0.3345	0.1054	0.3324	0.0995	0.3245	0.0912
	NAIVE	0.7664	0.1918	0.5968	0.1564	0.5748	0.1695	0.5070	0.1331
Case 3	**TVP-SSM-1**	**174.5994**	**0.0524**	**162.2061**	**0.0424**	**174.4974**	**0.0504**	**161.4568**	**0.0447**
	TVP-SSM-2	1003.7645	0.3224	2045.5378	0.6278	2432.6180	0.7000	2780.6212	0.8590
	TVP-SSM-3	2573.4767	0.8265	2573.0678	0.8262	2572.3978	0.8257	2572.3978	0.8257
	TVP-SSM-4	1389.6447	0.3207	1418.5824	0.3237	932.4355	0.2521	1408.0734	0.3223
	ARIMAX	184.1614	0.0648	234.4958	0.0560	281.1958	0.0683	271.2229	0.0615
	Seq2Seq	1318.0825	0.4225	1318.7672	0.4228	1319.7890	0.4231	1319.4584	0.4230
	Transformer	1562.1517	0.4355	1564.5224	0.4368	1562.5534	0.4358	1563.4481	0.4363
	MLR	772.7561	0.2453	770.2594	0.2444	772.9813	0.2450	777.3334	0.2468
	SVR	188.1767	0.0821	245.3478	0.0701	282.1134	0.0845	280.4156	0.0713
	NAIVE	809.2093	0.2034	665.9044	0.1707	818.8434	0.1985	741.2329	0.1831

Note: The best statistical scores are marked in boldface.

Case 3: robustness analysis based on Zhoushan data set

Further, in order to verify the fitting effect of the TVP-SSM model on different data sets, robustness analysis was carried out based on the Zhoushan data set in Case 3. Considering the potential heterogeneity in the Zhoushan dataset, the search engine variables, intervention variables, and periodical variables were processed according to Section 4. Subsequently, correlation and causality tests among these variables were conducted, as shown in Appendix D.

As shown by the results of Case 3 in Table 11, for hotel consumption prediction in Zhoushan, the TPP-SSM-1 model achieves an average RMSE of 168.1899 and an average MAPE of 0.0475, both of which are significantly lower than those of the benchmark models across different prediction steps, indicating that the TPP-SSM-1 model demonstrates strong predictive performance. Additionally, the DM test results of Case 3 in Table 11 show that the prediction results of TPP-SSM-1, TPP-SSM-2, ARIMAX, Seq2Seq, Transformer, MLR, SVR, and NAIVE models are significantly different under different prediction steps, and the prediction accuracy of the TVP-SSM model is higher than that of these benchmark models.

Table 11
DM test results under different prediction steps.

	TVP-SSM-1, VS.	h = 4		h = 8		h = 12		h = 16	
		DM (RMSE)	DM (MAPE)	DM (RMSE)	DM (MAPE)	DM (RMSE)	DM (MAPE)	DM (RMSE)	DM (MAPE)
Case	ARIMAX	−2.1911***	−2.9418***	−2.5531**	−2.6532***	−2.7665***	−3.3523***	−3.4509***	−3.4131***
	Seq2Seq	−1.9146*	−2.5677**	−1.7326*	−1.9795**	−1.8761*	−2.4055**	−1.8466*	−2.2898**
	Transformer	−3.0127***	−3.4450***	−3.0062***	−3.4251***	−3.0148***	−3.4219***	−3.0611***	−3.4797***
	MLR	−3.1305***	−3.3946***	−3.1344***	−4.1023***	−3.9578***	−4.1624***	−3.1305***	−3.3946***
	SVR	−2.0839**	−1.8142*	−2.0515**	−1.7793*	−2.0502**	−1.7769*	−2.0468**	−1.7731*
	NAIVE	−1.9130*	−2.4131**	−1.8952*	−2.2027**	−2.7862***	−3.2321***	−3.4246***	−3.7993***
	TVP-SSM-2	−1.0747	−1.7404*	−1.7117*	−2.1586**	−1.8159*	−2.4478**	−2.1328**	−2.7868***
	TVP-SSM-3	−1.7075*	−2.0134**	−1.7066*	−2.0108**	−1.7049*	−2.0036**	−1.7049*	−2.0036**
	TVP-SSM-4	−2.3055**	−2.1904**	−3.3054***	−2.1909**	−3.3058***	−2.1904**	−3.3055***	−3.191***
	ARIMAX	−1.7227*	−1.6761*	−1.7507*	−1.6651*	−2.2379**	−1.8374**	−1.7227*	−1.676*
	Seq2Seq	−3.5978***	−8.6120***	−3.6758***	−8.3124***	−3.6125***	−8.2378***	−3.6442***	−8.2014***
	Transformer	−1.0297	−2.2951**	−1.0379	−2.3884**	−1.0307	−2.2836**	−1.0357	−2.3697**
	MLR	−4.5929***	−4.0273***	−3.7254***	−5.0457***	−3.0304***	−3.6619***	−4.5929***	−4.0273***
	SVR	−4.1116***	−2.8682***	−2.8006***	−2.0014**	−2.3089**	−1.6709*	−4.1116***	−2.8682***
	NAIVE	−3.9792***	−9.1126***	−0.8055	−5.1350***	−0.8508	−4.9516***	−0.7630	−5.6605***
	TVP-SSM-2	−5.1342***	−3.8830***	−1.1579	−2.1761**	−1.5080	−2.0049**	−1.6552*	−2.1625**
	TVP-SSM-3	−2.0729**	−2.5748**	−2.0700**	−2.5702**	−2.0651**	−2.5615**	−2.0651**	−2.5615**
	TVP-SSM-4	−2.5047**	−2.9343***	−2.4847**	−2.9195***	−2.0979**	−2.3456**	−2.4889**	−2.9202***

Note: ***, **, and * represent 1 %, 5 %, and 10 % significance levels, respectively.

Discussion

In this study, we propose a hotel occupancy prediction framework that incorporates multisource data. The study finds that multisource data can effectively improve the accuracy of hotel occupancy predictions, consistent with Pan and Yang (2016). The main reason is that multisource data reflect tourists' behavioral characteristics from multiple dimensions. For example, SED reflects tourists' perceptions of tourist destinations through search keywords, so different search keywords are crucial to the model results. Additionally, the Baidu search index exhibits higher precision in forecasting tourism demand for the Chinese market (Yang et al., 2015). The holiday variable reflects the strong demand for tourism. Generally speaking, holidays are the peak of tourism, and scenic spots and hotels are full (Bi et al., 2021; Liu et al., 2018). Official high-frequency statistics, such as airport passenger flow, reflect tourist behavior (Liu et al., 2022), which is also a significant advantage of this paper compared with other studies. Therefore, integrating multivariate data can effectively enhance the performance of the prediction model.

Second, it has been shown that external intervention events, such as terrorist attacks (Toma et al., 2009) and stock and economic crises (Eugenio-Martin, 2016), have an impact on tourism demand forecasting. In this paper, COVID-19 is regarded as an external intervening variable. The TVP-SSM model is employed to capture the effects of external intervening variables on hotel occupancy. It adjusts model coefficients in real time using Kalman filtering, allowing it to reflect the dynamic impacts of these variables on occupancy across different stages, aligning the shocks with actual conditions.

However, this study exhibits certain limitations that should be taken into account. First, the study only considers the search volume of tourists on the Baidu platform. In fact, many tourists may choose Google, Weibo, Airbnb, and other online platforms for trip planning. Therefore, combining multimodal data may further improve prediction accuracy (Tan et al., 2025). Second, this paper focuses solely on Hangzhou and Zhoushan as case studies. Given the differences in coping strategies adopted by different regions in the face of external shocks, the performance of the model proposed in this study may fluctuate in tourist forecasting scenarios in different regions, and future work can focus on inter-regional heterogeneity. Third, the COVID-19 pandemic is regarded as an external shock and is represented by a dummy variable with only two values (0 and 1). However, in fact there are many types of external shock events that affect tourist flow with diverse degrees of impacts. The application of the model proposed in this study can be discussed separately for these different types of impact factors in the future.

Conclusion and discussion

Conclusion

A TVP-SSM model was constructed in this study, which included the official statistics, SED, intervention variables, and periodical variables. The model was applied to generate short-term predictions of daily hotel occupancy in Hangzhou under the impact of external events, such as the COVID-19 pandemic. The empirical results show that (1) During the comprehensive outbreak stage of the COVID-19 pandemic, the adverse effect of COVID-19 on hotel occupancy increased rapidly at first and then decreased gradually. In the subsequent sporadic outbreak stage, the pandemic situation and restrictions would also decrease hotel occupancy. (2) Adding interventions and SED to the estimation can significantly boost the prediction accuracy. (3) Overall, the TVP-SSM model, incorporating intervention and periodic variables, demonstrated strong performance in forecasting hotel occupancy, effectively capturing the impact of external shocks.

The results of this study have significant implications for hotel operations and management, offering tourism and hospitality industry decision-makers a new approach to emergency scenarios. Specifically, the proposed approach enables the tracking of tourist flows in the short term and ensures that forecasting includes the most recent data without incurring a significant cost increase. Meanwhile, the results can help operators to develop pricing strategies that consider the expected tourism demand. Tourism operators can develop contingency plans to manage fluctuations in tourist flow during specific periods.

Moreover, this paper mainly focuses on analyzing how external intervention variables impact hotel occupancy predictions. To our knowledge, most existing studies on occupancy forecasting primarily utilize time-series data and online metrics, with scant attention paid to intervention factors. However, in reality, sudden disruptions like COVID-19 can have a profound impact on occupancy rates. By systematically examining intervention effects through the integration of multisource data, our work establishes theoretical foundations and offers reference value for future research in this critical direction.

Implications

To evaluate the impact of external events, we propose a new forecasting framework based on multisource data, which holds significant theoretical and practical implications for tourism demand forecasting and management.

First, from a theoretical perspective, external events such as the COVID-19 pandemic make tourism demand forecasting particularly challenging. Hence, based on the TVP-SSM, the forecasting framework constructed in this paper innovatively integrates multiple sources of data, comprehensively reflecting the impact of various factors on tourism demand, which improves the forecasting performance. Additionally, the proposed time-varying model can dynamically simulate the impact of external events on tourism demand, providing solid support for research on tourism demand forecasting under the influence of external events. What is more, the proposed forecasting framework can be extended to analyze the dynamic relations between socio-economic variables, such as the dynamic relations between the relationship between inflation and GDP.

Second, in terms of practical aspects, the results of this study have important implications for hotel business management, providing tourism and hospitality industry decision-makers with a new approach for emergency scenarios. Specifically, the proposed approach allows tourist flows to be tracked in the short run. It can ensure that the forecasting models include the most recent data without a serious cost increase. Meanwhile, the results can help operators to conduct pricing strategies that consider the expected tourism demand. Tourism operators can develop contingency plans to manage fluctuations in tourist flow during specific periods.

CRediT authorship contribution statement

Ji Chen: Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Kang Tong:** Writing – original draft, Visualization, Validation. **Qinglin Yu:** Writing – original draft, Software, Methodology, Investigation. **Sichao Chen:** Software, Methodology, Investigation, Formal analysis. **Tomas Balezentis:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Dalia Streimikiene:** Writing – review & editing, Validation, Investigation, Conceptualization.

Appendix A

To forecast hotel occupancy in Hangzhou, we use daily SED collected from the Baidu Index. First, the basic keywords "Hangzhou food," "Hangzhou hotel," "Hangzhou map," and "Hangzhou attractions" are used, and extended keywords are obtained through Baidu Index's automatic recommendation technology. Then, crawler tools are used to fetch the query volume for each keyword.

To ensure prediction accuracy, we first selected the most relevant keywords from the 56 query terms listed in Table A1, we tested which lag order satisfies the highest correlation with the daily hotel occupancy. Then we used the corresponding time points for further analysis. Second, we lagged the 56 query keywords by orders from zero to seven, turning the query volumes for each keyword into daily data. The Pearson correlation coefficient between the query frequencies (adjusted as discussed above) of each keyword and hotel occupancy was calculated. The threshold was set to 0.6, i.e., keywords with correlation coefficients lower than 0.6 were removed. Third, the lasso regression model was fitted as an additional test. As a result, seven keywords were retained, which are shown in Table A1.

Table A1
Query keywords related to tourism demand in Hangzhou.

Basic Keyword	Extended Keywords	Number of Extended Keywords
Hangzhou food	Hangzhou food, Hangzhou food guide, Hangzhou tourism guide, Hangzhou tourism, Hangzhou travel network, Hangzhou dishes, Introduction of Hangzhou food, Pictures of Hangzhou food, Hangzhou tourist accommodation, Hangzhou travel route, Hangzhou snacks, West Lake vinegar fish, Hangzhou special food, Hangzhou food street, Hangzhou snack street, Tickets for the West Lake, Hangzhou specialty, West Lake Longjing, Song City, Tickets for Hangzhou Song City, West Lake specialty	21
Hangzhou hotel	Hangzhou hotel, Hangzhou hotel reservation, Hangzhou accommodation, Intercontinental Hangzhou hotel, Hangzhou farmhouse, Hangzhou club, Hangzhou bar	7
Hangzhou map	Hangzhou map, Tourist map of Hangzhou, The West Lake, Map of West Lake, West Lake scenery, West Lake picture, Hangzhou bus, West Lake ten views, Map of West Lake in Hangzhou, West Lake in Hangzhou, West Lake attraction, West Lake tourism, Hangzhou subway line map, Boat of West Lake	14
Hangzhou attractions	Hangzhou attractions, Hangzhou Song City, Lingyin Temple, Su Causeway, West Lake broken bridge, West Lake lotus, Hangzhou photography, West Lake tour guide, Yuefei Temple, Lingyin Temple guide, Tickets for Lingyin Temple, West Lake travel service, Bai Causeway, Hangzhou Hefang street	14

Appendix B

To verify the stationarity of hotel occupancy (Y_t), tourist flow in Hangzhou Hubin Road ($X_{1,t}$), passenger arrivals of Hangzhou Xiaoshan International Airport ($X_{2,t}$), and the combined keyword variable ($X_{3,t}$), the ADF test was performed. As shown in Table B1, these four variables satisfy the stationarity condition.

Table B1
Results of the ADF test.

Variable	T-statistic	P-value	Stationarity
$\ln(Y_t)$	−3.9017	0.0021***	Stable
$\ln(X_{1,t})$	−3.5748	0.0065***	Stable
$\ln(X_{2,t})$	−3.0504	0.0309**	Stable
$\ln(X_{3,t})$	−3.0819	0.0284**	Stable

Note: ***, **, and * represent 1 %, 5 %, and 10 % significance levels, respectively.

Appendix C

Table C1
The correlation between hotel consumption in Hangzhou and other variables.

Variables	Correlation coefficient	P-value
$\ln(x_{1,t})$	0.8205	0.0000***
$\ln(x_{2,t})$	0.8106	0.0000***

Note: ***, **, and * represent 1 %, 5 %, and 10 % significance levels, respectively.

Table C2
Granger causality test of hotel consumption in Hangzhou and other variables.

Null hypothesis	F statistic	P-value
$\ln(x_{1,t})$ is not the cause of $\ln(y_t)$	13.4531	0.0000***
$\ln(x_{2,t})$ is not the cause of $\ln(y_t)$	27.8137	0.0000***

Note: ***, **, and * represent 1 %, 5 %, and 10 % significance levels, respectively.

Appendix D

In Case 3, the construction of the SED for Zhoushan and the processing of variables remain the same as in Section 4.2. Finally, the two keywords “Putuo Mountain Tourism Strategy” and “Zhoushan Seafood” were retained, and the correlation with response variables is shown in Table C1, Table C2, Furthermore, the combined keyword variable was obtained by normalizing and summing these two keywords, and its trend is shown in Fig. D1.

Table D1
The correlation between search keywords and hotel consumption in Zhoushan.

Search keywords	Correlation coefficient	Optimal lag order
Putuo Mountain Tourism Strategy	0.8022	1
Zhoushan Seafood	0.7928	1

The settings for the holiday and weekend variables are the same as those in Table 2. However, it is worth noting that the Spring Festival effect is not considered in the forecast of hotel consumption in Zhoushan. The setting of COVID-19 state variables is shown in Table D1, Table D2.

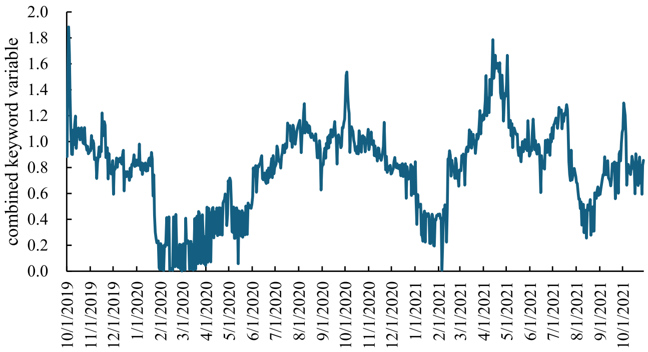

Fig. D1. The trends of the combined keyword variable in Zhoushan.

Table D2
The setting of COVID-19 state variables.

Notation	Influence stage	Value	Periods
$d_{1,t}^{(2)}$	Comprehensive outbreak stage	1	2020.1.22-2020.5.31
		0	Others
	Sporadic outbreak stage	1	2021.1.31-2021.2.4 2021.7.23-2021.8.31
		0	Others

According to Table D3 and Table D4, there is a significant and strong correlation and causality between hotel consumption in Zhoushan, passenger arrivals of Zhoushan Putuoshan Airport, and the combined keyword variable, indicating the rationality of using passenger arrivals of Zhoushan Putuoshan Airport and the combined keyword variable to predict hotel consumption.

Table D3
The correlation between hotel consumption in Zhoushan and various variables.

Variables	Correlation coefficient	P-value
$\ln\left(x_{1,t}^{(2)}\right)$	0.7480	0.0000***
$\ln\left(x_{2,t}^{(2)}\right)$	0.7722	0.0000***

Note: ***, **, and * represent 1 %, 5 %, and 10 % significance levels, respectively.

Table D4
The Granger causality of hotel consumption in Zhoushan and various variables.

Null hypothesis	F statistic	P-value
$\ln\left(x_{1,t}^{(2)}\right)$ is not the cause of $\ln\left(y_t^{(2)}\right)$	16.6747	0.0000***
$\ln\left(x_{2,t}^{(2)}\right)$ is not the cause of $\ln\left(y_t^{(2)}\right)$	6.7238	0.0013***

Note: ***, **, and * represent 1 %, 5 %, and 10 % significance levels, respectively.

References

Aruoba, S. B., Diebold, F. X., Nalewaik, J., Schorfheide, F., & Song, D. (2016). Improving GDP measurement: A measurement-error perspective. *Journal of Econometrics, 191*(2), 384–397. https://doi.org/10.1016/j.jeconom.2015.12.009

Apergis, N., Mervar, A., & Payne, J. E. (2016). Forecasting disaggregated tourist arrivals in Croatia. *Tourism Economics, 23*(1), 78–98. https://doi.org/10.5367/te.2015.0499

Assaf, A. G., Li, G., Song, H., & Tsionas, M. G. (2019). Modeling and forecasting regional tourism demand using the Bayesian Global Vector Autoregressive (BGVAR) model. *Journal of Travel Research, 586*(3), 383–397. https://doi.org/10.1177/004728751875922

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting, 27*(3), 822–844. https://doi.org/10.1016/j.ijforecast.2010.04.009

Bi, J. W., Li, H., & Fan, Z. P. (2021). Tourism demand forecasting with time series imaging: A deep learning model. *Annals of Tourism Research, 90*(3), Article 103255. https://doi.org/10.1016/j.annals.2021.103255

Bi, J. W., Li, C., Xu, H., & Li, H. (2022). Forecasting daily tourism demand for tourist attractions with Big Data: An ensemble deep learning method. *Journal of Travel Research, 61*(8), 1719–1737. https://doi.org/10.1177/00472875211040569

Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association, 70*(349), 70–79. https://doi.org/10.1080/01621459.1975.10480264

Chen, M. H., Jang, S. S., & Kim, W. G. (2007). The impact of the SARS outbreak on Taiwanese hotel stock performance: An even-study approach. *Hospitality Management, 26*(1), 200–212. https://doi.org/10.1016/j.ijhm.2005.11.004

Chen, J., Ying, Z., Zhang, C., & Balezentis, T. (2024). Forecasting tourism demand with search engine data: A hybrid CNN-BiLSTM model based on Boruta feature selection. *Information Processing & Management, 61*(3), Article 103699. https://doi.org/10.1016/j.ipm.2024.103699

Dergiades, T., Mavragani, E., & Pan, B. (2018). Google Trends and tourists' arrivals: Emerging biases and proposed corrections. *Tourism Management, 66*(1), 108–120. https://doi.org/10.1016/j.tourman.2017.10.014

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 20*(1), 134–144. https://doi.org/10.1198/073500102753410444

Dong, Y., Xiao, L., Wang, J., & Wang, J. (2023). A time series attention mechanism based model for tourism demand forecasting. *Information Sciences, 628*(3), 269–290. https://doi.org/10.1016/j.ins.2023.01.095

Emili, S., Figini, P., & Guizzardi, A. (2020). Modelling international monthly tourism demand at the micro destination level with climate indicators and web-traffic data. *Tourism Economics, 26*(7), 1129–1151. https://doi.org/10.1177/135481661986780

Eugenio-Martin, J. L. (2016). Estimating the tourism demand impact of public infrastructure investment: The case of Malaga Airport expansion. *Tourism Economics, 22*(2), 254–268. https://doi.org/10.5367/te.2016.0547

Fan, X., Lu, J., Qiu, M., & Xiao, X. (2023). Changes in travel behaviors and intentions during the COVID-19 pandemic and recovery period: A case study of China. *Journal of Outdoor Recreation and Tourism, 41*(1), Article 100522. https://doi.org/10.1016/j.jort.2022.100522

Gao, J., Peng, P., Lu, F., Claramunt, C., Qiu, P., & Xu, Y. (2024). Mining tourist preferences and decision support via tourism-oriented knowledge graph. *Information Processing & Management, 61*(1), Article 103523. https://doi.org/10.1016/j.ipm.2023.103523

Gunter, U., & Önder, I. (2016). Forecasting city arrivals with Google analytics. *Annals of Tourism Research, 61*, 199–212. https://doi.org/10.1016/j.annals.2016.10

He, K., Ji, L., Wu, C. W. D., & Tso, K. F. G (2021). Using SARIMA–CNN–LSTM approach to forecast daily tourism demand. *Journal of Hospitality and Tourism Management, 49*(3), 25–33. https://doi.org/10.1016/j.jhtm.2021.08.022

Hu, M., Qiu, R. T., Wu, D. C., & Song, H. (2021a). Hierarchical pattern recognition for tourism demand forecasting. *Tourism Management, 84*, Article 104263. https://doi.org/10.1016/j.tourman.2020.104263

Hu, M., & Song, H. (2019). Data source combination for tourism demand forecasting. *Tourism Economics, 26*(7), 1248–1265. https://doi.org/10.1177/1354816619872592

Hu, M., Xiao, M., & Li, H. (2021b). Which search queries are more powerful in tourism demand forecasting: searches via mobile device or PC? *International Journal of Contemporary Hospitality Management, 33*(6), 2022–2043. https://doi.org/10.1108/IJCHM-06-2020-0559

Jiao, E., & Chen, J. (2019). Tourism forecasting: a review of methodological developments over the last decade. *Tourism Economics, 25*(3), 469–492. https://doi.org/10.1177/1354816618812588

Jorge-González, E., González-Dávila, E., Martín-Rivero, R., & Lorenzo-Díaz, D. (2020). Univariate and multivariate forecasting of tourism demand using state-space models. *Tourism Economics, 26*(4), 598–621. https://doi.org/10.1177/1354816619857641

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*(1), 35–45. https://doi.org/10.1115/1.3662552

Lai, S. L., & Lu, W. L. (2005). Impact analysis of September 11 on air travel demand in the USA. *Journal of Air Transport Management, 11*(6), 455–458. https://doi.org/10.1016/j.jairtraman.2005.06.001

Lee, C.-C., Olasehinde-Williams, G. O., & Olanipekun, I. O. (2020). GDP volatility implication of tourism volatility in South Africa: A time-varying approach. *Tourism Economics, 28*(2), 435–450. https://doi.org/10.1177/1354816620970001

Li, H., Hu, M., & Li, G. (2020). Forecasting tourism demand with multisource big data. *Annals of Tourism Research, 83*, Article 102912. https://doi.org/10.1016/j.annals.2020.102912

Li, X., & Law, R. (2019). Forecasting tourism demand with decomposed search cycles. *Journal of Travel Research, 59*(1), 52–68. https://doi.org/10.1177/0047287518824158

Li, Y., Yang, D., Guo, J. E., Sun, S., & Wang, S. (2023). Daily tourism demand forecasting before and during COVID-19: data predictivity and an improved decomposition-ensemble framework. *Current Issues in Tourism, 27*(8), 1–21. https://doi.org/10.1080/13683500.2023.2202308

Liu, H., Liu, W., & Wang, Y. (2021). A study on the influencing factors of tourism demand from Mainland China To Hong Kong. *Journal of Hospitality & Tourism Research, 45*(1), 171–191. https://doi.org/10.1177/1096348020944435

Liu, P., Zhang, H., Zhang, J., Sun, Y., & Qiu, M. (2019). Spatial-temporal response patterns of tourist flow under impulse pre-trip information search: From online to arrival. *Tourism Management, 73*, 105–114. https://doi.org/10.1016/j.tourman.2019.01.021

Liu, Y., Feng, G., Chin, K. S., Sun, S., & Wang, S. (2022). Daily tourism demand forecasting: the impact of complex seasonal patterns and holiday effects. *Current Issues in Tourism, 26*(10), 1573–1592. https://doi.org/10.1080/13683500.2022.2060067

Liu, Y.-Y., Tseng, F.-M., & Tseng, Y.-H. (2018). Big Data analytics for forecasting tourism destination arrivals with the applied Vector Autoregression model. *Technological Forecasting and Social Change, 130*, 123–134. https://doi.org/10.1016/j.techfore.2018.01.018

Lu, R., Bai, R., Li, R., Zhu, L., Sun, M., Xiao, F., Wang, D., Wu, H., & Ding, Y. (2024). A novel Sequence-to-Sequence-based Deep Learning Model for multistep load forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15. https://doi.org/10.1109/tnnls.2023.3329466

Ma, M., Ren, P., Chen, Z., Ren, Z., Liang, H., Ma, J., & De Rijke, M. (2023). Improving Transformer-based sequential recommenders through preference editing. *ACM Transactions on Information Systems, 41*(3), 1–24. https://doi.org/10.1145/3564282

Monache, D. D., Petrella, I., & Venditti, F. (2020). Price dividend ratio and long-run stock returns: A score driven State Space Model. *Journal of Business & Economic Statistics, 39*(4), 1–31. https://doi.org/10.1080/07350015.2020.1763805

Nicholas, A. (2021). Forecasting US overseas travelling with univariate and multivariate models. *Journal of Forecasting, 40*(6), 963–976. https://doi.org/10.1002/for.2760

Ozdemir, O., Kizildag, M., Dogru, T., & Madanoglu, M. (2022). Measuring the effect of infectious disease-induced uncertainty on hotel room demand: a longitudinal analysis of US hotel industry. *International Journal of Hospitality Management, 103*, Article 103189. https://doi.org/10.1016/j.ijhm.2022.103189

Ozdogan, A., & Ozdogan, Z. (2023). A vector-autoregression-intervention analysis of PKK terrorism and Turkey's counterterrorism. *Journal of Policing, Intelligence and Counter Terrorism, 18*(22), 189–212. https://doi.org/10.1080/18335330.2022.2117568

Pan, B., Wu, C., & Song, H. (2012). Forecasting hotel room demand using Search Engine Data. *Journal of Hospitality and Tourism Technology, 3*(3), 196–210. https://doi.org/10.1108/17579881211264486

Pan, B., & Yang, Y. (2016). Forecasting destination weekly hotel occupancy with big data. *Journal of Travel Research, 56*(7), 957–970. https://doi.org/10.1177/0047287516669050

Pezenka, I., & Weismayer, C. (2020). Which factors influence locals' and visitors' overall restaurant evaluations? *International Journal of Contemporary Hospitality Management, 2*(9), 2793–2812. https://doi.org/10.1108/IJCHM-09-2019-0796

Prilistya, S. K., Permanasari, A. E. and Fauziati, S. (2021). The effect of the COVID-19 pandemic and Google trends on the forecasting of international tourist arrivals in Indonesia. *2021 IEEE Region 10 Symposium (TENSYMP)*, 1-8. doi: 10.1109/TENSYMP52854.2021.9550838.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/AOS/1176344136

Semenoglou, A. A., Spiliotis, E., & Assimakopoulos, V. (2023). Data augmentation for univariate time series forecasting with neural networks. *Pattern Recognition, 134*, Article 109132. https://doi.org/10.1016/j.patcog.2022.109132

Seong, B., & Lee, K. (2021). Intervention Analysis Based on Exponential Smoothing Methods: Applications to 9/11 and COVID-19 Effects. *Economic Modelling, 98*, 290–301. https://doi.org/10.1016/j.econmod.2020.11.014

Shafiqah, A., Dharini, P., & Aerambamoorthy, T. (2022). Forecasting the volatility of Cryptocurrencies in the presence of COVID-19 with the State Space Model and Kalman Filter. *Mathematics, 10*(17), 3190. https://doi.org/10.3390/MATH10173190. –3190.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting: A review of recent research. *Tourism Management, 29*(2), 203–220. https://doi.org/10.1016/j.tourman.2007.07.016

Song, H., Li, G., Witt, S. F., & Athanasopoulos, G. (2011). Forecasting tourist arrivals using time-varying parameter structural time series models. *International Journal of Forecasting, 27*(3), 855–869. https://doi.org/10.1016/j.ijforecast.2010.06.001

Song, H., & Wong, K. K. F. (2003). Tourism Demand Modeling: A time-varying parameter approach. *Journal of Travel Research, 42*(1), 57–64. https://doi.org/10.1177/0047287503253908

Srdelić, L., & Dávila-Fernández, M. J. (2024). International trade and economic growth in Croatia. *Structural Change and Economic Dynamics, 68*, 240–258. https://doi.org/10.1016/j.strueco.2023.10.018

Sun, S., Hu, M., Wang, S., & Zhang, C. (2022). How to capture tourists' search behavior in tourism forecasts? A two-stage feature selection approach. *Expert Systems with Applications, 213*(1-3), Article 118895. https://doi.org/10.1016/j.eswa.2022.118895

Sun, S., Wei, Y., Tsui, K. L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management, 70*(2), 1–10. https://doi.org/10.1016/j.tourman.2018.07.010

Sun, S., Sun, H., Xu, H., Li, H., & Wang, S. (2025). Unlocking the power of multimodal online reviews: A multisensory perspective. *Tourism Management, 111*, Article 105206. https://doi.org/10.1016/j.tourman.2025.105206

Sun, Y., Zhang, J., Li, X., & Wang, S. (2023). Forecasting tourism demand with a new time-varying forecast averaging approach. *Journal of Travel Research, 62*(2), 305–323. https://doi.org/10.1177/00472875211061206

Tan, J., Cheng, M., Chen, J., Zhu, J., Yu, Q., & Chen, S. (2025). Multimodal destination image and user engagement: A sequential research design. *Tourism Management, 111*, Article 105209. https://doi.org/10.1016/j.tourman.2025.105209

Tian, F., Yang, Y., Mao, Z., & Tang, W. (2021). Forecasting daily attraction demand using big data from search engines and social media. *International Journal of Contemporary Hospitality Management, 33*(6), 1950–1976. https://doi.org/10.1108/IJCHM-06-2020-0631

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, 58*(1), 267–288. www.jstor:stable/2346178.

Toma, M., McGrath, R., & Payne, J. E. (2009). Hotel tax receipts and the "Midnight in the Garden of Good and Evil": a time series intervention seasonal ARIMA model with time-varying variance. *Applied Economics Letters, 16*(7), 653–656. https://doi.org/10.1080/13504850701221808

Uniejewski, B., Marcjasz, G., & Weron, R. (2019). Understanding intraday electricity markets: Variable selection and very short-term price forecasting using LASSO. *International Journal of Forecasting, 35*(4), 1533–1547. https://doi.org/10.1016/j.ijforecast.2019.02.001

Waris, I., & Mohd Suki, N. (2025). Predicting sustainable tourism: examining customers' intention to visit green hotels using an extended norm activation model. *Management of Environmental Quality: An International Journal, 36*(1), 228–248.

Wen, L., Liu, C., & Song, H. (2019). Forecasting tourism demand using search query data: A hybrid modelling approach. *Tourism Economics, 25*(3), 309–329. https://doi.org/10.1177/1354816618768317

Wickramasinghe, K., & Ratnasiri, S. (2021). The role of disaggregated search data in improving tourism forecasts: Evidence from Sri Lanka. *Current Issues in Tourism, 24*(19), 2740–2754. https://doi.org/10.1080/13683500.2020.1849049

Wu, D., Li, G., & Song, H. (2012). Economic analysis of tourism consumption dynamics: A time-varying parameter demand system approach. *Annals of Tourism Research, 39*(2), 667–685. https://doi.org/10.1016/j.annals.2011.09.003

Wu, D., Song, H., & Shen, S. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management, 29*(1), 507–529. https://doi.org/10.1108/IJCHM-05-2015-0249

Wu, D. C., Zhong, S., Wu, J., & Song, H. (2024). Tourism and hospitality forecasting with Big Data: A systematic review of the literature. *Journal of Hospitality & Tourism Research.* , Article 267315866. https://doi.org/10.1177/10963480231223151

Wu, E. H. C., Hu, J., & Chen, R. (2022). Monitoring and forecasting COVID-19 impacts on hotel occupancy rates with daily visitor arrivals and search queries. *Current Issues in Tourism, 25*(3), 490–507. https://doi.org/10.1080/13683500.2021.1989385

Xiang, Z., & Pan, B. (2011). Travel queries on cities in the United States: Implications for search engine marketing for tourist destinations. *Tourism Management, 32*(1), 88–97. https://doi.org/10.1016/j.tourman.2009.12.004

Xie, G., Qian, Y., & Wang, S. (2021). Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach. *Tourism Management, 82*, Article 104208. https://doi.org/10.1016/j.tourman.2020.104208

Xue, G., Liu, S., Ren, L., & Gong, D. (2023). Forecasting hourly attraction tourist volume with search engine and social media data for decision support. *Information Processing & Management, 60*(4), Article 103399. https://doi.org/10.1016/j.ipm.2023.103399

Xiong, Z., Liu, J., Sriboonchitta, S., & Ramos, V. (2018). Forecasting China's inbound tourist arrivals using a state space model. *Journal of Physics: Conference Series, 1053*(1), Article 1012134. https://doi.org/10.1088/1742-6596/1053/1/012134

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46*, 386–397. https://doi.org/10.1016/j.tourman.2014.07.019

Yang, Y., Fan, Y., Jiang, L., & Liu, X. (2022). Search query and tourism forecasting during the pandemic: When and where can digital footprints be helpful as predictors? *Annals of Tourism Research, 93*, Article 103365. https://doi.org/10.1016/j.annals.2022.103365

Yang, Y., Pan, B., & Song, H. (2013). Predicting hotel demand using destination marketing organization's web traffic data. *Journal of Travel Research, 53*(4), 433–447. https://doi.org/10.1177/0047287513500391

Zhan, L., Cheng, M., & Zhu, J. (2024). Progress on image analytics: Implications for tourism and hospitality research. *Tourism Management, 100*, Article 104798. https://doi.org/10.1016/j.tourman.2023.104798

Zhang, B., Li, N., Law, R., & Liu, H. (2021a). A hybrid MIDAS approach for forecasting hotel demand using large panels of search data. *Tourism Economics, 28*(7), 1823–1847. https://doi.org/10.1177/13548166211015515

Zhang, B., Li, N., Shi, F., & Law, R. (2020). A deep learning approach for daily tourist flow forecasting with consumer search data. *Asia Pacific Journal of Tourism Research, 25*(3), 323–339. https://doi.org/10.1080/10941665.2019.1709876

Zhang, C., & Tian, Y. (2022). Forecast daily tourist volumes during the pandemic period using COVID-19 data, search engine data and weather data. *Expert Systems with Applications, 210*(30), Article 118505. https://doi.org/10.1016/j.eswa.2022.118505

Zhang, C., Luo, L., Liao, H., Mardani, A., Streimikiene, D., & Al-Barakati, A. (2019). A priority-based intuitionistic multiplicative UTASTAR method and its application in low-carbon tourism destination selection. *Applied Soft Computing, 88*, Article 106026. https://doi.org/10.1016/j.asoc.2019.106026

Zhang, H., & Lu, J. (2022). Forecasting hotel room demand amid COVID-19. *Tourism Economics, 28*(1), 200–221. https://doi.org/10.1177/13548166211035569

Zhang, H., Jiang, Z., Gao, W., & Yang, C. (2022b). Time-varying impact of economic policy uncertainty and geopolitical risk on tourist arrivals: Evidence from a developing country. *Tourism Management Perspectives, 41*, Article 100928. https://doi.org/10.1016/j.tmp.2021.100928

Zhang, H., Song, H., Wen, L., & Liu, C. (2021b). Forecasting tourism recovery amid COVID-19. *Annals of Tourism Research, 87*(4), Article 103149. https://doi.org/10.1016/j.annals.2021.103149

Zhang, T., Zhang, Z., & Xue, G. (2023a). Mitigating the disturbances of events on tourism demand forecasting. *Annals of Operations Research.* , Article 264385362. https://doi.org/10.1007/s10479-023-05626-6

Zhang, Y. (2023b). Circular economy model for elderly tourism operation based on multi-source heterogeneous data integration. *Applied Artificial Intelligence, 37*(1), Article e2205228. https://doi.org/10.1080/08839514.2023.2205228

Zhao, E., Du, P., Azaglo, E. Y., Wang, S., & Sun, S. (2022). Forecasting daily tourism volume: A hybrid approach with CEMMDAN and multi-kernel adaptive ensemble. *Current Issue in Tourism, 26*(7), 1112–1131. https://doi.org/10.1080/13683500.2022.2048806

Zhu, J., Cheng, M., & Wang, Y. (2024). Viewer in-consumption engagement in pro-environmental tourism videos: A video analytics approach. *Journal of Travel Research.* , Article 266947035. https://doi.org/10.1177/00472875231219634