

## Research Article

# Study of Flight Departure Delay and Causal Factor Using Spatial Analysis

Shaowu Cheng,<sup>1</sup> Yaping Zhang ,<sup>1</sup> Siqi Hao ,<sup>1</sup> Ruiwei Liu,<sup>2</sup> Xiao Luo,<sup>3</sup> and Qian Luo<sup>3</sup>

<sup>1</sup>School of Transportation Science and Technology, Harbin Institute of Technology, Harbin 150001, China

<sup>2</sup>State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin 150001, China

<sup>3</sup>The Second Research Institute of Civil Aviation Administration of China, Chengdu 610041, China

Correspondence should be addressed to Siqi Hao; [siqihao47@163.com](mailto:siqihao47@163.com)

Received 15 February 2019; Revised 5 May 2019; Accepted 23 May 2019; Published 4 June 2019

Academic Editor: Eneko Osaba

Copyright © 2019 Shaowu Cheng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Analysis of flight delay and causal factors is crucial in maintaining airspace efficiency and safety. However, delay samples are not independent since they always show a certain aggregation pattern. Therefore, this study develops a novel spatial analysis approach to explore the delay and causal factors which is able to take dependence and the possible problem involved including error correlation and variable lag effect of causal factors on delay into account. The study first explores the delay aggregation pattern by measuring and quantifying the spatial dependence of delay. The spatial error model (SEM) and spatial lag model (SLM) are then established to solve the error correlation and the variable lag effect, respectively. Results show that the SEM and SLM achieve better fit than ordinary least square (OLS) regression, which indicates the effectiveness of considering dependence by employing spatial analysis. Moreover, the outcomes suggest that, aside from the well-known weather and flow control factors, delay-reduction strategies also need to pay more attention to reducing the impact of delay at the previous airport.

## 1. Introduction

With the rapid development of the civil aviation industry, airspace has become increasingly crowded. This crowdedness causes increasingly frequent delays in most major airports worldwide. This situation seriously affects airports, airlines, and passengers. From 2007 to 2017, the annual flights in China consistently increased from 3.65 million to 10.83 million, with an average increasing rate of approximately 12.2% in the past five years. Meanwhile, the rate of flights arriving on time decreased from 83.19% in 2007 to 71.67% in 2017. The annual cost of flight delays in China was estimated to be more than \$7.4 billion. Such high economic costs of delay necessitate delay causal factor analysis and delay-reduction strategies.

Several approaches have been taken to analyze the factors that affect flight arrival and departure delay. Allan et al. [1] studied several determining causes of flight delay at the Newark International Airport (EWR) using a comprehensive approach. The results show that adverse weather conditions, low ceilings, and low visibility conditions strongly influence flight delays. Similarly, Asfe et al. [2] investigated the major

causal factors of flight delays by ranking different factors using the analytical hierarchical process. They found technical failure and delayed entries as two of the most influential factors. Based on the identification of causal factors, further researches explored the quantitative effect of each factor on flight delay. By analyzing the characteristics of flight departure and arrival delays by constructing probability density functions, Mueller et al. [3] explored several causal factors of delays, such as traffic volume, aircraft type, aircraft maintenance, airline operations, weather conditions, change of procedures en route, capacity constraints, customer service issues, and late aircraft or crew arrival. The results show that weather contributed to 69% of the delays. Different results can be achieved by different method and variables; research results of Kwan and Hansen [4] show that airport congestion contributed to approximately 32% of the average delays, in which a series of econometric models was established to identify the key causal factors of flight delays, including airport congestion, total traffic, and en route weather. In addition to identifying the causal factors and their quantitative effect on flight delay, more studies focus on the development of models

to determine the probability of aircraft delay. Wesonga et al. [5] proposed and evaluated a multiple parametric approach, which includes the apparently significant meteorological and aviation parameters, to predict the probability of aircraft delay. Recent research and development effort in delay probability prediction are seeking to develop asymmetric Bayesian logit model to take the asymmetric distribution pattern of the dependent variable into consideration (see Perez-Rodriguez et al. [6]). By using data from BTS and IATA, this article corroborates the necessity and superiority of the proposed asymmetric Bayesian logit model, as well as identifying new significant factors affecting the probability of arrival delay.

In addition to traditional statistical methods, machine learning algorithms were used by several studies. Bayesian network was a commonly used approach to establish delay model to explore the delay propagation mode and estimate delay [7, 8]. Artificial neural network was also utilized to examine the relationship between departure delay and different causal factors comparing to linear and nonlinear regressions [9]. Deep learning models have also been investigated for air traffic delay prediction tasks [10]. Moreover, a number of studies attempted to determine the major causal factors of flight delays by detecting the time series data trend. Abdel-Aty et al. [11] applied the “two-stage approach” to detect periods of regularly repeating patterns in their data and to identify the factors correlated with them. Tu et al. [12] employed a smoothing spline model to identify the relationship between seasonal trends, random effects, and daily delay propagation pattern. Delay propagation has also been deeply investigated by many researches to help to understand the air congestion [13–15] and alleviate flight delay [16, 17]. The effects of day and time were assumed to be additive, and the residuals were assumed to be identically and independently distributed in the study.

However, delays show a certain aggregation pattern in the temporal dimension; high delays are normally clustered; and low delays tend to be surrounded by low delays. In other words, the delay value of samples with shorter distance between them is normally similar compared to the delay value of delays with longer distance between them. The correlation between two delay values depends on their spatial attribute such as spatial location and spatial distance. Without doubt, there is high degree of spatial dependence among delays in a space organized by hour and by day. Given that most of the aforementioned methods were based on certain assumptions which either ignore or simplify the correlation of samples in the dataset, Diana [18] initially introduced the approach of spatial analysis for delay prediction, which is able to take the spatial dependence in every direction into account. In the study, delay was considered as a spatially distributed variable in a space coordinated by day and time. A spatial error model (SEM) was built to consider spatial dependence in error.

Actually, flight departure delay is a complex problem with substantial direct causal factors and many concealed indirect causal factors. Flight departure delay is caused by the abovementioned factors, as well as by the flight delays that occur earlier [12], as the operation resources required by the current flight, such as the crew, aircraft, and passenger gates, might have been utilized by previously delayed flights. This

resources correlation may lead to delay daily propagation. The spatial dependence exists in every direction since the aggregation is observed in both day of week and hour of day, which probably lead to error correlation and variable lag effect of causal factors on delay [19].

Motivated by the exploration of the main causal factors of flight departure delays in consideration of correlation between delay samples, our study analyzes departure delay as a geographic problem instead of a statistical problem by assuming delay as a spatially distributed variable organized by hour and by day. Causal factor analysis using spatial analysis enables the existence of spatial dependence in variables, which solves the problem of sample correlations among hours and days simultaneously. Specifically, spatial regression models were built to absorb the delay spatial dependence by adding a spatial independent variable. The spatial lag model (SLM) and spatial error model (SEM) are established in our study to solve the variable lagged effect and the error correlation, respectively. Comparisons between the SLM, the SEM, and the OLS estimation are also conducted.

This paper is structured as follows. Section 2 introduces the spatial analysis methodology. Section 3 describes the data sources, defines the variables, and describes the data-processing methodology. Section 4.1 shows the exploration analysis of flight departure delay with a distribution map and trend analysis. Section 4.2 demonstrates identification of delay pattern. Section 4.3 maps the semivariogram to quantify the spatial dependence of flight departure delay. Based on the results of Section 4.3, Section 4.4 illustrates spatial prediction considering spatial autocorrelation by employing ordinary kriging method. Section 4.5 discusses the establishment of the classical regression model, SEM, and SLM, as well as the comparative analysis of the three models to explore the main causal factors of flight departure delays. Finally, this paper is concluded with a summary.

## 2. Methods

This study employs the spatial analysis method to explore the delay distribution pattern and causal factors of flight departure delays while considering delay spatial dependence. Delay is assumed to be a spatially distributed variable. Spatial analysis is a quantifying technique used in the study of spatial variables [20].

### 2.1. Delay Pattern Analysis

(1) *Exploring Delay Distribution.* The first step to analyze delay pattern is to explore the distribution. By defining a space with  $x$  coordinate of day of week and  $y$  coordinate of hour, the delay is added to each hour unit as an attribute. The delay distribution can be plotted with different colors as the delay minutes. The time when an intense delay occurred is recognized in the distribution map. 3D trend analysis can be used to visualize the departure delay distribution and trend in the temporal dimension.

(2) *Identifying the Pattern of Delay.* The pattern of delay is then identified by calculating Moran's  $I$  and general  $G$

to measure the degree of delay spatial dependence among observations. Positive autocorrelation suggests that the values of the one hour unit and its neighbors are similar. Negative autocorrelation suggests that the values of the one hour unit and its neighbors are different. No autocorrelation suggests that the values are randomly distributed over the space.

Moran's  $I$  is calculated as

$$I = \frac{\sum_i^n \sum_j^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{S^2 \sum_i^n \sum_j^n w_{ij}} \quad (1)$$

$$S^2 = \frac{1}{n} \sum_i^n (x_i - \bar{x})^2 \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2)$$

$$Z = \frac{1 - E(I)}{\sqrt{\text{VAR}(I)}} \quad (3)$$

where  $I$  is the value of Moran's  $I$ ,  $x_i$  is the total minutes of departure delay during the hour unit  $i$ , and  $w_{ij}$  is the spatial weight matrix.  $Z$  value is generally used to test Moran's  $I$  value. A test result against the null hypothesis indicates that no spatial autocorrelation exists.

Most of the spatial weight matrices are built based on spatial connectivity and spatial distance. The weight matrix in this study is generated based on distance measured by the inverse Euclidean distance between two hour units. The value of Moran's  $I$  ranges from  $-1$  to  $1$ . Moran's  $I$  identifies the similarity between units with delay and the spatial distribution pattern. However, it cannot distinguish high-from low-value clusters. General  $G$  identifies the two different patterns of spatial cluster; it is computed as

$$G(d) = \frac{\sum_i^n \sum_j^n w_{ij}(d) x_i x_j}{\sum_i^n \sum_j^n x_i x_j} \quad (4)$$

$$Z = \frac{G - E(G)}{\sqrt{\text{VAR}(G)}}$$

When the  $Z$  value is significant, a general  $G$  value that is greater than its average indicates a high-value cluster, as opposed to a general  $G$  value that is less than its average. A general  $G$  value that is equal to its average indicates no autocorrelation.

The cluster type in the flight departure delay is then identified, and the hot and cold spots of flight departure delay are explored.

(3) *Quantifying Delay Spatial Dependence.* After measuring the degree of spatial dependence, the semivariogram is modeled to quantify the departure delay spatial dependence and to analyze its random and structural properties. Departure delay is considered as a regionalized variable because it is correlated with the hour and the day. Structural property indicates the existence of the autocorrelation between the departure delay at location  $x$  and at location  $x + h$  ( $h$  is the distance from  $x$ ). Semivariance calculates the average difference on departure

delays between pairs of hour units at a given interval [9]; it is computed as

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2, \quad (5)$$

where  $Z(x_i)$  is the total minutes of departure delay of location  $x_i$ ;  $Z(x_i + h)$  is the total minutes of departure delay of the locations with  $h$  distance from  $x_i$ ; and  $N(h)$  is the number of locations with  $h$  distance from  $x_i$ .

(4) *Delay Prediction.* After the spatial dependence structure of a variable is determined, the measured data can be used to estimate the variable at unmeasured locations. This interpolation method is known as kriging interpolation. Based on the unbiased estimation and the minimum variance principle, the kriging interpolation method can quantify the spatial dependence between the known sample and the estimated point according to the statistical characteristics and spatial variation of the sample.

2.2. *Causal Factor Analysis.* After the identification of delay dependence, causal factor analysis is performed using spatial analysis, which enables the existence of spatial dependence in variables. To explore the causal factors of flight departure delay, spatial econometric models were built to absorb the delay spatial dependence by adding a spatial independent variable, and the outcomes of the SLM, SEM, and classical regression model are compared.

(1) *Classical Regression Model.* A classical regression model can be written as

$$Y = \beta X + \varepsilon, \quad (6)$$

where  $Y$  represents the total minutes of departure delay at the target airport and  $X$  represents the factor variables.  $\beta$  represents the effect of the independent variables on the dependent variable, and  $\varepsilon$  is the random error term vector subjected to normal distribution.

(2) *SEM.* SEM is able to consider the spatial dependence in error terms by adding spatial error term as an explanatory variable. The SEM takes the following form:

$$Y = \beta X + \varepsilon$$

$$\varepsilon = \lambda W \varepsilon + \mu, \quad (7)$$

where  $Y$  is the total minutes of departure delay and  $X$  is the factor variables.  $\beta$  represents the effect of the independent variables on the dependent variable,  $\varepsilon$  is the random error term vector,  $\lambda$  is the spatial error coefficient,  $W$  is the spatial weight matrix of error term generated based on distance measured by the inverse Euclidean distance between two hour units, and  $\mu$  is the random error term vector subjected to normal distribution.

TABLE 1: Frequency and effect of each factor of flight departure delay.

Factors of departure delay	Frequency	Rank of frequency	Average delay minutes	Rank of average delay minutes
T	21	7	206.5	4
TL	34	5	235.8	2
W	30	6	238.6	1
WL	161	2	154.8	6
WD	9	9	176.1	5
WR	13	8	212.2	3
CF	650	1	99.2	11
CR	71	4	126.9	8
A	115	3	134.8	7
F	3	11	110.7	10
P	2	12	98.0	12
D	7	10	118.0	9

(3) *SLM*. SLM is able to consider the spatial autocorrelation in delay variable by adding a spatial lag variable as an explanatory variable. The SLM takes the following form:

$$Y = \rho WY + \beta X + \varepsilon, \quad (8)$$

where  $Y$  is the total minutes of departure delay and  $X$  is the factor variables.  $\beta$  represents the effect of the independent variables on the dependent variable;  $W$  is a spatial weight matrix of the dependent variables generated based on distance measured by the inverse Euclidean distance between two hour units;  $\rho$  is the spatial regression coefficient, which reflects the effects of the delay in the neighbor hours  $WY$  on the delay in one hour  $Y$ ; and  $\varepsilon$  is the random error term vector subjected to normal distribution.

### 3. Data Collection

*3.1. Data Sample.* The data in this study are obtained from the database of an international hub airport in China in June 2016. To maintain the privacy of the institution, the name of the airport is not revealed. In June 2016, 8788 flights departed from the target airport, among which 18 flights returned, 51 flights were canceled, and 5357 flights (60.96%) were delayed for more than 15 minutes; 3180 flights (36.19%) were delayed for more than half an hour; 1528 flights (17.39%) were delayed for more than one hour; and 489 flights (5.56%) were delayed for more than two hours. The most severe delay lasted for 888 minutes. Approximately 70% of the delays were within 60 minutes. The data are organized by day of week and hour of day. To demonstrate the spatial dependence of delay distribution intuitively, the study assumed delay as a spatially distributed variable. The space is defined with day of week as the  $x$  coordinate and hour of day as the  $y$  coordinate. Compared with the total number of flights (8788), there were few flights (72) from 0:00 to 7:00, and hour units with less than five flights are not considered since it could bias the average. The study area covers 7:00 to 24:00, including a total of 510 hour units with departure delays.

*3.2. Definitions of Variables.* First step of variable construction is to find out factors affecting flight departure delay.

The flight delay determinants considered in previous studies include weather, delay propagation, flight schedule, airplane shortage, air route, airplane type, flight order, air traffic flow, hub airport, ability of the airline to pay debt, ability of the airline to profit, load factors of the airline, load rate of the airline, and other factors [21, 22]. Chinese aviation determined the following factors of flight delay. Technical failure includes technical failure at the target airport (T) and technical failure at the previous airport (TL). Weather refers to weather conditions at the target airport (W), at the previous airport (WL), at the destination airport (WD), and en route (WR). Control factors include flow control (CF) and route restriction (CR). Other factors include the airline (A), airport facility (F), passenger (P), and capacity allocation (D).

Then, nominal factors are selected by calculating the frequency and the effect of each factor in our dataset. Effect of each factor of flight delay in Table 1 is measured by average delay minutes caused by each factor. As shown in Table 1, the frequency of the flow control factor is significantly higher than the others, but the average delay minutes caused by the flow control factor is lower. Conversely, the frequency of the weather condition at the target airport and the technical failure at the previous airport are significantly lower, with high average delay minutes.

All factors are classified into three categories: high frequency and low effect, low frequency and high effect, and low frequency and low effect. Flow control, airline factor, route restriction, and weather condition at the previous airport caused most of the departure delays; however, these factors can be usually controlled well, and the delay can be eliminated in a short time. The effects of weather conditions at the target airport and en route and the technical failure at the target, previous, and destination airports, although they did not happen often, have dramatic impacts with long departure delays. Airport facility, passenger, and capacity allocation are the minor reasons for flight departure delay, and we will not focus on these factors in the following discussion.

In addition, delay can be related to time period (morning, afternoon, night, and weekday or weekend). The total traffic and passengers are also important factors. Aviation industry experts are interviewed about the limitations of the data



TABLE 2: Descriptive statistics and definitions of the variables used in the model.

Variables	Definition	Mean	s.d.	Min	Max
<i>Delay</i>	Total minutes of departure delay	587.824	428.433	0	2341
<i>T</i>	Technical failure at the target airport (equals 1 if technical failure occurs)	0.039	0.194	0	1
<i>TL</i>	Technical failure at the previous airport (equals 1 if technical failure occurs)	0.065	0.246	0	1
<i>W</i>	Weather condition at the target airport (equals 1 if weather is adverse)	0.018	0.132	0	1
<i>WL</i>	Weather condition at the previous airport (equals 1 if weather is adverse)	0.251	0.434	0	1
<i>WR</i>	Weather condition en route (equals 1 if weather is adverse)	0.016	0.124	0	1
<i>CF</i>	Flow control (equals 1 if conduct flow control)	0.571	0.496	0	1
<i>CR</i>	Route restriction (equals 1 if conduct route restriction)	0.122	0.327	0	1
<i>NF</i>	Scheduled departure traffic	16.943	4.414	4	27

collection, and the final list included 15 factors that affected flight departure delays.

We then conducted a stepwise-backwards regression in variable construction and determined a significant level of introduced independent variable as  $\alpha_{in} = 5\%$  and a significant level of excluded independent variable as  $\alpha_{out} = 10\%$ . Seven factors are excluded, and the remaining 8 explanatory variables comprise the regression model (Table 2). All variables are calculated in an hour unit.

**3.3. Data Processing.** The data are processed with various software. The exploration analysis module in ArcGIS 10.2 is used for the distribution mapping and the 3D trend analysis. The geostatistic module in ArcGIS 10.2 software is adopted to generate the theoretical and empirical semivariograms, as well as the kriging interpolation. The Geoda software is used to develop the spatial econometric models.

## 4. Results and Discussion

### 4.1. Exploring Delay Distribution

*(1) Distribution Map.* The distribution map is a commonly used spatial data visualization method. Each grid defined by day and hour is colored according to the departure delay in minutes that occurred in an hour unit. Red represents high departure delay, whereas dark blue indicates low departure delay. The distribution map highlights delay intensity, the day on which the delay occurred, and the duration of the delay. As demonstrated in Figure 1, the delay levels between neighborhoods are usually similar, which indicates obvious spatial cluster characteristics. The distribution map also shows that intense delay occurred mostly at 16:00, 18:00, and 21:00, especially from June 18 to June 22 in our dataset.

*(2) 3D Trend Analysis.* The trend analysis generated a 3D trend map of the departure delay. In Figure 2, the  $x$ -axis and  $y$ -axis represent the day and the hour of delay, respectively, and the  $z$ -axis represents the total minutes of departure delay. The green line in the  $x$ - $z$  plane and the blue line in the  $y$ - $z$  plane indicate the trend of the delay. The figure shows that flight departure delay is intensively late in the month and exhibits a parabola with the peak at 18:00 in our dataset.

**4.2. Identifying the Pattern of Delay.** As mentioned in Introduction, there exists a high degree of spatial dependence among delays. Moran's  $I$  and general  $G$  are calculated to measure the degree of spatial autocorrelation and to identify the pattern of delay. The Moran's  $I$  and general  $G$  values of all variables are calculated. Values with a significant autocorrelation are listed in Table 3.

Table 3 shows significant positive spatial autocorrelation in variables such as the total minutes of departure delay, weather conditions at the target airport, weather conditions at the previous airport, weather conditions en route, flow control, total departure traffic, and number of passengers. General  $G$  test showed that all of the abovementioned variables are high-value clusters, indicating that the hours of intense delay are clustered.

The hot and cold spots of delay are explored after the degree of autocorrelation is measured and the hour units of high-value clusters in flight departure delay are identified. A high degree of delay from June 18 to 22 that lasted for 8 hours from 14:00 to 22:00 is noted, as shown by the red area in Figure 3. Among all the factors responsible for flight departure delay, this large-scale cluster is probably caused by an exogenous variable such as sudden adverse weather conditions. This conclusion corresponded to the actual weather report record of the target airport in June 2016. Between

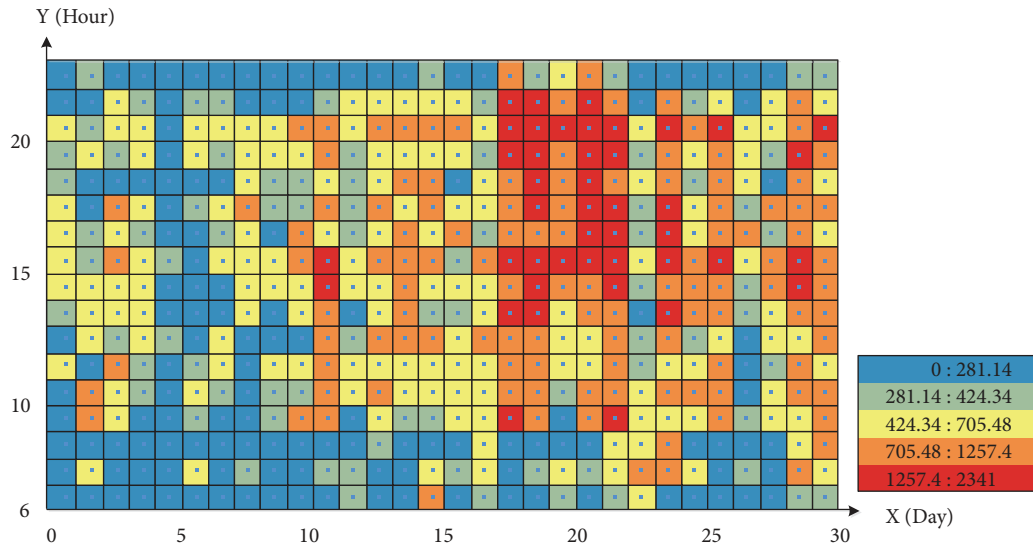


FIGURE 1: Distribution map of departure delay.

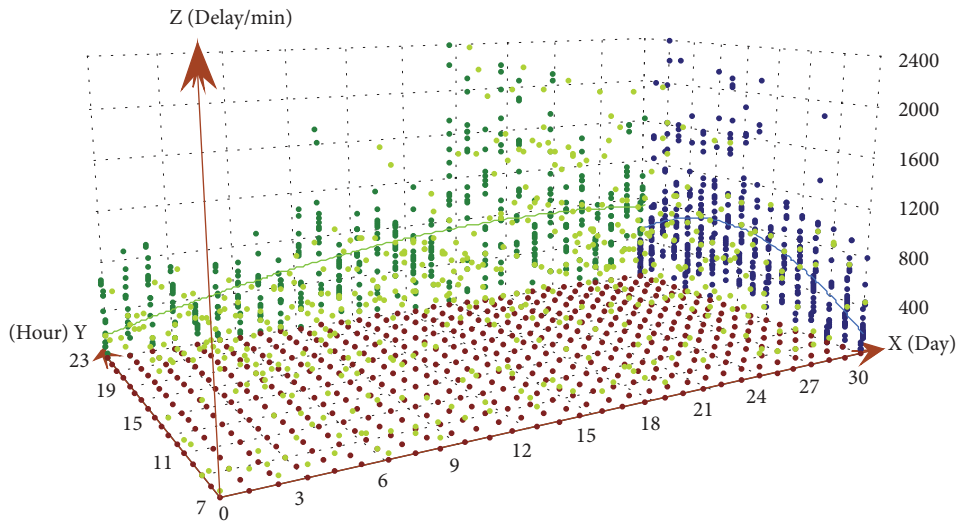


FIGURE 2: 3D trend map of departure delay.

TABLE 3: Autocorrelation statistics for selected variables.

Variables	Coefficient	Observed	Expected	Stddev	Z	P
Delay	Moran's <i>I</i>	0.580509	-0.001965	0.001016	18.275923	0.000000
	General <i>G</i>	0.010119	0.007496	0.000000	19.158604	0.000000
W	Moran's <i>I</i>	0.397390	-0.001965	0.000915	13.198725	0.000000
	General <i>G</i>	0.194444	0.007496	0.000201	13.177839	0.000000
WL	Moran's <i>I</i>	0.161547	-0.001965	0.001021	5.116695	0.000000
	General <i>G</i>	0.011565	0.007496	0.000001	5.576605	0.000000
WR	Moran's <i>I</i>	0.248777	-0.001965	0.000901	8.352514	0.000000
	General <i>G</i>	0.142857	0.007496	0.000260	8.398764	0.000000
CF	Moran's <i>I</i>	0.341630	-0.001965	0.001024	10.738782	0.000000
	General <i>G</i>	0.009717	0.007496	0.000000	11.624289	0.000000
NF	Moran's <i>I</i>	0.532762	-0.001965	0.001018	16.761879	0.000000
	General <i>G</i>	0.007859	0.007496	0.000000	14.762860	0.000000
NP	Moran's <i>I</i>	0.489845	-0.001965	0.001019	15.405064	0.000000
	General <i>G</i>	0.007888	0.007496	0.000000	12.831494	0.000000

TABLE 4: Theoretical model fit comparison of isotropic semivariogram.

Model	Nugget	Sill	Range	RSS	R <sup>2</sup>
Exponential	66979.4200	219070.4200	17.6000	5.48E+08	0.977
Spherical	90023.7000	206937.5788	14.0439	1.12E+09	0.955
Gaussian	105523.0000	203931.7816	11.4025	1.35E+09	0.947

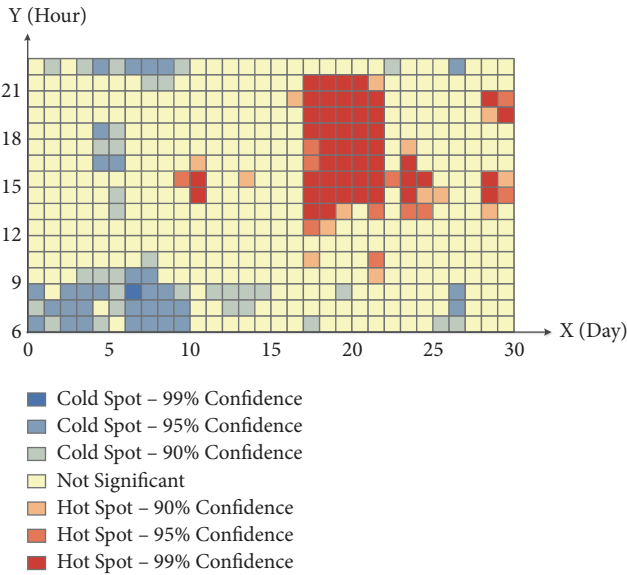


FIGURE 3: Hot and cold spots of departure delay.

June 18 and 22, there were cloudy skies for 5 days and a thunderstorm for 3 days.

4.3. *Quantifying Delay Spatial Dependence.* After measuring the degree of delay spatial dependence between observations, the variogram is utilized to quantify the spatial dependence based on the theory of regionalized variables. The experimental semivariogram is mapped to quantify the spatial dependence of delays and to provide the spatial structure for the subsequent kriging interpolation.

The following are the key parameters in the semivariogram:

(1) Nugget effect  $C_0$  is estimated from the empirical variogram at  $h = 0$ . This represents the measurement error or random property of the departure delay. The  $C_0$  value reflects the variation caused by the stochastic factor.

(2) Range  $A_0$  is the distance where the variogram reaches plateau. This represents the largest distance of autocorrelation. Data can be considered uncorrelated if their distance exceeds the range.

(3) Sill  $C+C_0$  is the plateau at which the variogram reaches the range. This represents the total variance of regionalized variables, which is equal to the sum of the autocorrelation and stochastic variances.

The following are the other two parameters that can be calculated from the three parameters mentioned above:

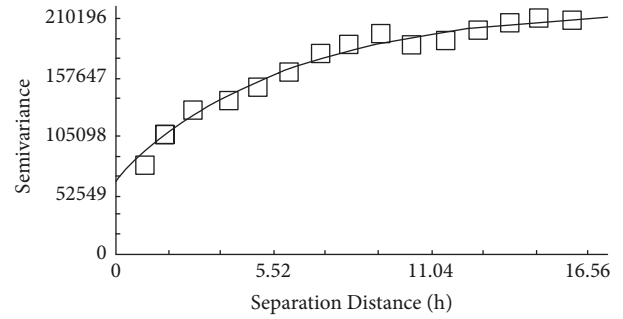


FIGURE 4: Isotropic semivariogram of departure delay.

(4) Structural variance  $C$  represents the structural property of delay. The value of  $C$  reflects the variance caused by the autocorrelation.

(5) Nugget-sill ratio  $C_0/(C + C_0)$  represents the percentage of variance caused by randomness. A low nugget-sill ratio reflects that the variation is mainly affected by the autocorrelation factors.

Theoretical semivariogram is necessary to obtain the spatial structure of delay. The experimental semivariogram generated from limited samples is used to estimate the correlation in the whole area by fitting a theoretical semivariogram to an empirical semivariogram. Different theoretical models are compared in Table 4. The comparison shows that, when fitting the isotropic semivariogram, the exponential model is more effective than others, such as the spherical and Gaussian models.

According to results of semivariogram, the low nugget-sill ratio (30.6%) suggests that the variation of delay is mainly caused by autocorrelation (69.4%). In Figure 4, the delays separated by short intervals are strongly correlated to one another. The correlation decreases as the intervals increase to a distance of 17.60.

4.4. *Delay Prediction.* Spatial interpolation allows us to further comprehend the overall situation of the entire study area from a limited number of spatial sample points. We randomly select 10% of the sample dataset as the test set and the remaining 90% as the training set. Spatial autocorrelation undermines the accuracy and effectiveness of some commonly used interpolation methods such as trend surface method or inverse distance weighting (IDW) method. We use the ordinary kriging method to interpolate delays since it can take spatial dependence into account by considering spatial structure obtained by semivariogram. Similar to the IDW method, the ordinary kriging method predicts the value on unmeasured position by generating weights of the surrounding points. IDW generates weights according to

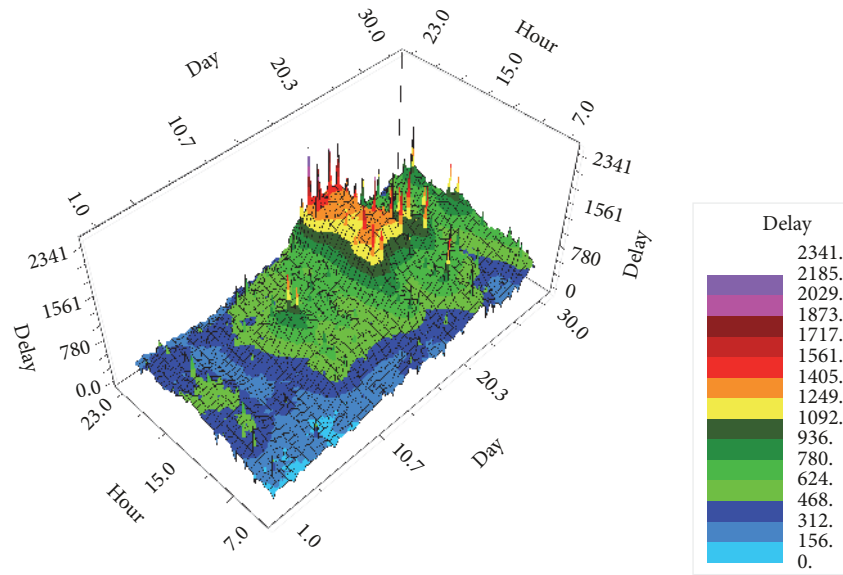


FIGURE 5: Prediction surface of delay.

TABLE 5: Results of the comparison of cross-validation between different interpolation methods.

Interpolation method	M	RMS	A_Std	Std_M	Std_RMS	
IDW	8.0616	298.4739				
Ordinary kriging	Exponential	1.7808	312.6256	0.0040	0.9665	
	Spherical	3.0301	307.9240	331.4682	0.0077	0.9312
	Gaussian	2.9745	316.4641	336.4019	0.0075	0.9426

the distance between unmeasured position and surrounding points. Different from IDW, kriging method generates weight from the semivariogram, which is developed by considering spatial properties and spatial structure of the data. The interpolation results of the prediction surface are shown in Figure 5, in which the  $x$ -axis represents the day of week,  $y$ -axis represents the hour of day, and the  $z$ -axis represents the value of delay. In this way the value of unmeasured locations can be interfered according to the value of measure locations and the spatial relation between them.

After the generation of prediction surface, it is important to evaluate the interpolation precision, which is conducted by cross-validation. Cross-validation leaves one point out and uses the rest to predict a value at that location. The point is then changed to another in turn, and finally this process is performed for all samples in the dataset. Similar to another typical interpolation method, prediction performance can be evaluated by Mean Error (M) and Root Mean Square Error (RMS). The smaller the RMS, the better. Besides, ordinary kriging has other indicators to evaluate prediction performance, including Average Standard Error (A\_Std), which measures the average of the prediction standard errors; Mean Standardized Error (Std\_M), whose value should be close to 0; Root Mean Square Standardized Error (Std\_RMS), which should be close to 1. A Std\_RMS greater than 1 indicates underestimating the variability in the predictions. A Std\_RMS less than 1 indicates overestimating the variability in the predictions.

Cross-validation can also be an effective selection approach between different interpolation methods. Comparing the cross-validation results, exponential semivariogram shows the minimum RMS and Std\_RMS closest to 1. Therefore, the best result is obtained in this study using the exponential fitting semivariogram for ordinary kriging interpolation (Table 5).

**4.5. Spatial Econometric Analysis.** The regression model is commonly used to analyze the factors of departure delay. First, we perform an ordinary least square (OLS) estimation based on the classical regression model (as in (6)). The estimation results of each variable are demonstrated as  $\hat{\beta}$  in Table 6. After the model parameters are estimated, it is necessary to perform the statistical test of the model, which includes the goodness of fit test, the significance test of the equation, and the significance test of the variables.

The goodness of fit test can be reflected by  $R^2$ .  $R^2$  value is the ratio of the sum of the squares of the regression and the sum of the squares of the total deviations, and it indicates the degree of interpretation of all the explanatory variables to the variation of the dependent variables. The value is between 0 and 1; the closer to 1, the better the estimated regression model fits.

The F test is a joint significance test for multiple coefficients to infer whether the linear relationship between the dependent variable and explanatory variables is significant.



TABLE 6: Estimation results of delay and causal factors for each model.

Variable	OLS estimation		Spatial lag model		Spatial error model	
	$\hat{\beta}$	$p$	$\hat{\beta}$	$p$	$\hat{\beta}$	$p$
Constant	-189.2450***	0.0008	-275.3832***	0.0000	81.4065	0.2570
T	192.5201***	0.0066	163.5734***	0.0029	197.2891***	0.0003
TL	179.6183***	0.0013	150.6237***	0.0005	140.6904***	0.0006
W	1059.5640***	0.0000	699.4476***	0.0000	769.7684***	0.0000
WL	247.8793***	0.0000	181.4880***	0.0000	161.2410***	0.0000
WR	275.8793**	0.0187	90.3014	0.3228	147.4748	0.1265
CF	292.8883***	0.0000	139.7609***	0.0000	128.7156***	0.0000
CR	205.2254***	0.0000	141.3503***	0.0002	120.0013***	0.0002
NF	28.3640***	0.0000	19.8473	0.0000	28.7060***	0.0000
$\rho$			0.5806***	0.0000		
$\lambda$					0.6926***	0.0000
R <sup>2</sup>	0.5977		0.7928		0.7884	
F	62.0591***					
Log likelihood	-3638.21		-3539.13		-3554.8640	
SC	7332.53		7140.61		7165.84	
AIC	7294.43		7098.26		7127.73	
Likelihood Ratio			198.1630***		166.6973***	

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

TABLE 7: Test results of OLS residuals' spatial dependence.

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.3640	11.5242	0.0000
Lagrange Multiplier (lag)	1	195.3046	0.0000
Robust LM (lag)	1	68.0750	0.0000
Lagrange Multiplier (error)	1	127.4078	0.0000
Robust LM (error)	1	10.1783	0.0000

The null hypothesis ( $H_0$ ) of the F test is that all the parameters to be estimated are simultaneously zero. The larger the F value, the less likely the null hypothesis.

The  $p$ -value measures the probability of correctly rejecting the null hypothesis when testing the significance of a single variable. A larger  $p$ -value indicates greater probability of erroneously rejecting the null hypothesis.

In Table 6, the OLS estimation shows an F value of 62.0591 at a 1% level of significance and a goodness of fit  $R^2$  value of 0.5977, which indicates that the explanation variables and the dependent variable have relatively significant linear correlation, and the dependent variable can be effectively predicted by the explanation variables.  $p$ -value indicates significant variables such as the technical fault at the target airport (T), the technical fault at the previous airport (TL), the weather condition at the previous airport (WL), the weather condition at the airport of departure (W), the flow control (CF), the route restriction (CR), and the number of scheduled departure flights (NF).

However, the Moran's  $I$  test shows a significant spatial autocorrelation in the residuals of the OLS estimators. The spatial dependence test results of the error and lag are positive, as shown in Table 7.

Classical regression model fails to reflect the spatial dependence between hour units and the influence of their interactions on the total minutes of departure delay. Therefore, spatial factors are introduced into the regression model, and spatial econometric analysis is necessary. SEM and SLM are built to measure the spatial dependence in error terms and the spatial dependence of delay between the hour units, respectively [23].

The spatial lag variable and spatial error terms are considered as the explanatory variables because of the spatial effects. The use of the OLS results in a biased and irregular estimation. Therefore, the maximum likelihood estimation method is used in this study. The model selection is based on the value of Log likelihood (Log L), the Akaike information criterion (AIC), and the Schwartz criterion (SC), which are fit statistic measures of the accuracy of the model, as well as the test for goodness of fit ( $R^2$ ). A greater Log likelihood and goodness of fit value and smaller Akaike information and Schwartz criteria indicate a better model fit.

Comparing the estimation results between the OLS estimation, spatial lag model, and spatial error model in Table 6, the goodness of fit  $R^2$  is 0.7918 for the SLM, which is greater

than 0.5978 in the OLS estimation and 0.7884 in the SEM. The AIC (7098.26) and SC values (7140.61) of the SLM are both less than the values of the OLS estimation (AIC 7294.43, SC 7332.53) and the SEM (AIC 7127.73, SC 7165.84). Moreover, the SLM and OLS estimations are nested, as with the SEM. Increasing the model parameters must result in high likelihood scores. Therefore, judging the fit of the model based on the log likelihood value is inaccurate. We conduct the likelihood ratio test for both models.

The likelihood ratio test uses a likelihood function to evaluate a simple model and a complex model with parameter constraints. The likelihood ratio is defined as the ratio of the maximum value of the likelihood function under constrained conditions to that under unconstrained conditions. A statistic that obeys the chi-square distribution can be constructed based on the likelihood ratio. The null hypothesis  $H_0$  is that there is no significant difference in the goodness of fit between model A and model B. The rejection or acceptance of the null hypothesis can be judged based on the constructed chi-square statistic value or  $p$ -value. In this way, we can judge whether the difference between models is significant. The results of the likelihood ratio test of SLM-OLS and SEM-OLS show that the likelihood ratio values are greater than the chi-squared distribution with the degree of freedom of 1 at 1% significance level, and the null hypotheses are rejected, which means that the SLM and SEM provide significantly better fit compared with the OLS estimation, and the explanatory capacity is enhanced by adding the spatial effect to the model.

Among the explanatory variables, the effects of the weather condition at the previous airport, the weather condition at the airport of departure, flow control, and the number of scheduled departure flights are the most significant. Moreover, the technical failure at the target airport and the previous airport and the route restriction also significantly affect departure delay. Adverse weather is the primary cause of flight departure delays with harsher influence than flow control. Delay reduction primarily focuses on weather forecasts and dynamically adjusts to weather changes.

Besides, the comparison results in Table 6 show that the WR variable, which is significant in the OLS estimation, is insignificant in the SLM and SEM. Delay-reduction strategies may focus on the weather prediction on route according to the OLS estimation but would not achieve a significant effect in delay reduction according to SLM and SEM. The SLM shows that the spatial lag variable is at the 1% significance level, which indicates a strong spillover effect of the departure delay in the temporal dimension.

Comparing with the results of causal factors obtained from the previous study, this study also indicates that the effect of weather condition at the target airport on flight delay is much greater than that of other factors. However, this study exhibits an interesting finding that the technical failure and weather condition at the previous airport have a larger effect on departure delays than flow control, which is one of the two most significant factors that affect delays aside from weather condition. This finding suggests that dealing with technical failure and weather prediction at the previous airport is crucial in delay reduction.

## 5. Conclusions

This study studied the flight departure delay and its causal factors by developing a novel spatial analysis method, which enables the correlation in data samples. The main conclusion can be presented as below.

First, spatial analysis is confirmed as a useful method in the delay and causal factor analysis in this study. Exploration analysis can intuitively demonstrate the distribution pattern of flight departure delay in the temporal dimension, semi-variogram can quantify the spatial structure of the delay, and kriging interpolation allows delay estimation at unmeasured locations.

Besides, the results of the spatial econometrics models achieve better fit performance by taking the spatial dependence into consideration, since the fit of SLM and SEM is better than that of OLS estimation. Results achieved by this study reconfirm the significant effect of the weather condition and technical failure on flight departure delay.

This study also indicates that the weather condition and technical failure at the previous airport significantly affect departure delay. These effects are more significant than the flow control factor, which is regarded as one of the two most important factors that affect delay. This result suggests that delay-reduction strategies must also focus on reducing the impact of delay at the previous airport.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

The authors would like to acknowledge the financial support from the Research and Development Project of Scientific and Technological Cooperation between Sichuan Provincial Colleges and Universities (Grant No. 2019YFSY0024), Key Research and Development Projects of Sichuan Science and Technology Plan Project (Grant No. 2019YFG0050), and National Natural Science Foundation of China (Grant No. U1533203) (Grant No. 61179069).

## References

- [1] S. S. Allan, J. A. Beesley, J. E. Evans et al., "Analysis of delay causality at Newark International Airport," in *Proceedings of the 4th USA/Europe Air Traffic Management R&D Seminar*, 2001.
- [2] M. Kazemi Asfe, M. Jangi Zehi, M. N. Shahiki Tash, and N. M. Yaghoubi, "Ranking different factors influencing flight delay," *Management Science Letters*, vol. 4, no. 7, pp. 1397–1400, 2014.
- [3] E. R. Mueller and G. B. Chatterji, "Analysis of aircraft arrival and departure delay characteristics," in *Proceedings of the AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*, pp. 1–14, Los Angeles, Calif, USA, October 2002.

- [4] I. Kwan and M. Hansen, "US flight delay in the 2000s: an econometric analysis," in *Proceedings of the Transportation Research Board 90th Annual Meeting*, vol. 11-4283, 2011.
- [5] R. Wesonga, F. Nabugoomu, and P. Jehopio, "Parameterized framework for the analysis of probabilities of aircraft delay at an airport," *Journal of Air Transport Management*, vol. 23, pp. 1–4, 2012.
- [6] J. Pérez-Rodríguez, J. Pérez-Sánchez, and E. Gómez-Déniz, "Modelling the asymmetric probabilistic delay of aircraft arrival," *Journal of Air Transport Management*, vol. 62, pp. 90–98, 2017.
- [7] Y.-J. Liu and S. Ma, "Flight delay and delay propagation analysis based on Bayesian network," in *Proceedings of the 2008 International Symposium on Knowledge Acquisition and Modeling, KAM 2008*, pp. 318–322, China, December 2008.
- [8] W. Cao and X. Fang, "Airport Flight departure delay model on improved BN structure learning," *Physics Procedia*, vol. 33, pp. 597–603, 2012.
- [9] D. M. Dai and J. S. Liou, "Delay prediction models for departure flights," *Transportation Research Record*, 2006.
- [10] Y. J. Kim, S. Choi, S. Briceno, and D. Mavris, "A deep learning approach to flight delay prediction," in *Proceedings of the 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pp. 1–6, Sacramento, Calif, USA, September 2016.
- [11] M. Abdel-Aty, C. Lee, Y. Bai, X. Li, and M. Michalak, "Detecting periodic patterns of arrival delay," *Journal of Air Transport Management*, vol. 13, no. 6, pp. 355–361, 2007.
- [12] Y. Tu, M. O. Ball, and W. S. Jank, "Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern," *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 112–125, 2008.
- [13] X. Dai, M. Hu, W. Tian, and H. Liu, "Modeling congestion propagation in multistage schedule within an Airport Network," *Journal of Advanced Transportation*, vol. 2018, Article ID 6814348, 11 pages, 2018.
- [14] N. Kafle and B. Zou, "Modeling flight delay propagation: A new analytical-econometric approach," *Transportation Research Part B: Methodological*, vol. 93, pp. 520–542, 2016.
- [15] Q. Wu, M. Hu, X. Ma, Y. Wang, W. Cong, and D. Delahaye, "Modeling flight delay propagation in airport and Airspace Network," in *Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3556–3561, Maui, HI, USA, November 2018.
- [16] W. Wu, C.-L. Wu, T. Feng, H. Zhang, and S. Qiu, "Comparative analysis on propagation effects of flight delays: a case study of china airlines," *Journal of Advanced Transportation*, vol. 2018, Article ID 5236798, 10 pages, 2018.
- [17] N. Ivanov, F. Netjasov, R. Jovanović, S. Starita, and A. Strauss, "Air Traffic Flow Management slot allocation to minimize propagated delay and improve airport slot adherence," *Transportation Research Part A: Policy and Practice*, vol. 95, pp. 183–197, 2017.
- [18] T. Diana, "Predicting arrival delays: An application of spatial analysis," *Journal of Aircraft*, vol. 48, no. 2, pp. 462–467, 2011.
- [19] A. Reynolds-Feighan, "Competing networks, spatial and industrial concentration in the US airline industry," *Spatial Economic Analysis*, vol. 2, no. 3, pp. 237–257, 2007.
- [20] R. Haining, *Spatial Data Analysis: Theory and Practice*, Cambridge University Press, 2003.
- [21] J.-T. Wong and S.-C. Tsai, "A survival model for flight delay propagation," *Journal of Air Transport Management*, vol. 23, pp. 5–11, 2012.
- [22] X. Yang, J. Wang, and J. He, "The identification of flight delay's determinants and their impact degree based on the analysis of dynamic queues model," *Statistic Information Forum*, vol. 4, no. 29, pp. 88–95, 2014.
- [23] L. Anselin, *Spatial Econometrics: Methods and Models*, vol. 4, Springer Science & Business Media, 2013.





**Hindawi**

Submit your manuscripts at  
[www.hindawi.com](http://www.hindawi.com)

