Research Article

A Big-Data-Driven Framework for Parking Demand Estimation in Urban Central Districts

Yunlin Guan,[1] Yun Wang [ID],[1] Xuedong Yan,[1] Haonan Guo,[1] and Yu Zhou[2]

[1]*MOT Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*
[2]*Department of Civil and Environmental Engineering, National University of Singapore, 117576, Singapore*

Correspondence should be addressed to Yun Wang; wang.yun@bjtu.edu.cn

Parking planning is a key issue in the process of urban transportation planning. To formulate a high-quality planning scheme, an accurate estimate of the parking demand is critical. Most previous published studies were based primarily on parking survey data, which is both costly and inaccurate. Owing to limited data sources and simplified models, most of the previous research estimates the parking demand without consideration for the relationship between parking demand, land use, and traffic attributes, thereby causing a lack of accuracy. Thus, this study proposes a big-data-driven framework for parking demand estimation. The framework contains two steps. The first step is the parking zone division method, which is based on the statistical information grid and multidensity clustering algorithms. The second step is parking demand estimation, which is extracted by support vector machines posed in the form of a machine learning regression problem. The framework is evaluated using a case in the city center in Cangzhou, China.

1. Introduction

As the growth of motor vehicle ownership throughout the world continues to lead to various traffic problems, solutions to mitigate issues of traffic safety, congestion, noise, air pollution, and parking, are becoming increasingly urgent [1, 2]. The majority of previous research studies had focused on the reduction of traffic accidents that alleviated environmental pollution and the relieve of traffic jams, while limited attention had been paid to parking problems [3, 4]. However, parking problems are becoming more and more serious as the parking demand has increased rapidly owing to the explosive growth of privately owned vehicles. In the UK, households currently possessing one vehicle account for approximately 45% of all registered vehicles, while households possessing two or more vehicles account for 30% of the total number of vehicles [5]. In view of these considerations, the parking problem during peak hours becomes the main burden of road traffic. Research has indicated that the average volume of traffic related to parking during peak hours

can reach 30–50% of the total traffic [6]. Therefore, the formulation of a reasonable parking planning is of great importance to both ease the burden of large road traffic volumes and guarantee an increased level of parking service during peak parking hours [7].

Traditional parking planning has focused on providing an improved parking supply to meet the demand. However, the ever-increasing parking demand can no longer be met owing to the limited number of parking lots and land resources. To mitigate this problem, parking planning has started to shift to demand management. Essentially, parking demand should be taken into consideration. For large and medium-sized cities, a reasonable demand management should be conducted based on accurate parking demand estimation. Otherwise, unrealistic demand management can hardly contribute to handling the parking problems.

This study focuses on the parking demand estimation problem. An accurate estimation of the parking demand will be useful for the determination of the gap between the supply and demand and for the provision of essential

information for parking planning schemes. Without reliable parking demand estimation results, the allocation of parking resources in parking planning will be difficult to match the reality. As a result, it will become occasionally difficult to park owing to the lack of available parking spaces. At the same time, some parking lots throughout the city will be wasted as a result of the unbalanced parking supply and demand. Moreover, relevant demand management policies can be applied to control the growing demand for parking lots. Accurate estimation of the parking demand is a primary goal of parking demand management, as it will help achieve a sustainable transportation system.

Traditional parking demand estimation methods based on parking survey data for the entire study area cannot be accurate, as manual errors are recorded and the impact of the increase in private vehicle ownership on the parking demand is unaccounted for. Meanwhile, the cost of parking surveys that cover the entire study area is high and execution of these surveys is time consuming. Fortunately, with the rapid development of traffic information technology, a huge amount of diverse traffic data has accumulated, including ride hailing, floating car, and various detector data [8]. Owing to the increasing maturity and popularization of data storage and processing technologies, big data can be used efficiently and appropriately to deal with various transportation problems [9]. Regarding the parking demand estimation problem, traditional methods can hardly obtain or handle the origin-destination (OD) big data from urban road networks. However, OD data can reflect the travel characteristics, such as the departure and arrival spatial distributions, that directly affect the estimation accuracy of parking demand. To our knowledge, no prior research has used OD big data resources in parking demand estimation problems. Thus, this study proposes a generalized big-data-driven framework for parking demand estimation.

During the estimation of the parking demand, the division of parking zones is an essential step that must be performed prior to the estimation of the parking demand. Therefore, this study proposes a parking demand estimation framework that consists of two steps, namely, parking-zone division and parking demand estimation. First, an integrated parking zone division method is designed based on the combination of statistical information grid (STING) clustering and a multidensity clustering algorithm. The proposed zone division method can significantly improve the computational efficiency of big OD data without sacrificing the traffic characteristics of each grid. Second, the parking demand estimation is performed with support vector regression (SVR). The model can accurately estimate the parking demand based on partial parking survey data, OD data with peak parking hour attributes, and land use information contained in each parking zone. Compared with traditional parking survey data, the partial parking survey data for this study only covers part of the study area that required a lower survey cost. In addition, the OD data and land use information can be obtained directly through modern traffic information technology and results in a lower survey cost and higher accuracy. The proposed parking demand estimation framework is based on the relationship between the travel characteristics obtained from modern transportation systems and parking characteristics and can accurately estimate the parking demand via OD big-data resources.

The remainder of this paper is organized as follows. Section 2 reviews relevant studies in the literature. Section 3 develops a framework for two step parking demand estimation, including data description, statistical information grid, and the two-step method. In Section 4, a case study based Cangzhou parking planning is performed to complete parking demand estimation. Finally, Section 5 illustrates the accuracy comparison between the method in this study and the general regression estimation method.

2. Literature Review

Early parking demand estimation methods mainly performed calculations in accordance with the principles outlined in applicable engineering manuals, which contain specific requirements for different cities or countries. The Institute of Transportation Engineers (ITE) published a parking generation rate report in 1978, which included the range of parking generation rates associated with 64 land types. Following development of the city, ITE has continued to update their findings pertaining to parking research studies. The addition of land use types in the fifth edition of the parking generation rate report published in 2019 serves as an important reference for many traditional estimation methods of parking demand. Their report proved that land usage has a significant effect on parking demand [10]. It is also necessary to confirm the land use characteristics of each parking zone.

The division of parking zones is an important preliminary step for estimating the parking demand. The principle of division for traditional parking zones focuses on the preservation of the integrity of the geographical environment and administrative functions [11]. As parking surveys are often divided by administrative area and parking survey data often limits the flexibility and rationality of the division of parking zones to some extent. Several problems may exist in this case, including (a) different travel density levels mixed in one parking zone or (b) a high-travel-density area is split into multiple adjacent parking zones in accordance with the traditional method, which precludes achievement of the goal of delineating the parking zones with traffic attributes as the leading factor. Considering that traffic characteristics are an essential part of the estimation of parking demand, insufficient correlation between the parking zone and traffic attributes reduces the rationality of additional parking demand estimation. Based on the correlation between the traffic characteristics and parking features, this study establishes the dividing method of the parking zones as the first step in the proposed parking demand estimation framework.

Recently, parking demands for more targeted estimation methods that take into consideration the actual characteristics of the researched city have increased. Currently, the most extensively used parking demand estimation models around the world include the parking generation rate,

multiple regression, and trip generation/attraction models. The parking generation rate model is a classic estimation model. The core of the parking generation rate model is based on the relationship between different land use types and the parking demand. In addition, it sets the number of parking lots per unit area for different land types. However, the parking generation rate is neither general nor universal. This means that various countries must still obtain local land use classifications and parking generation rate data via costly and time-consuming traffic surveys. Khaled and Jamil [12] referred to the international parking generation rate calculation model proposed by ITE and the land surveys of certain cities in Palestine, and they proposed the parking generation rates for 73 types of land usage. Ultimately, they established a reference standard for local parking demand estimation. In terms of model application, Xie Ying et al. [13] established a time-segment land structure estimation model and refined the application range of the parking demand estimation model based on the temporal dimension with consideration for parking land type and parking time. Although the parking generation rate model is simple and can be easily calculated, it requires a comprehensive and accurate parking survey to ensure accuracy of the parking demand estimation. A multivariate regression model can estimate the parking demand by defining the influencing factors as independent variables and the parking demand as the dependent variable. Dai et al. [14] improved the multivariate regression model based on the traffic volume and service level to guarantee the accuracy of the estimate. Selection of the influencing factors in the multiple regression model will directly affect the accuracy of the estimated results. Accordingly, comprehensive consideration of the influencing factors will cause difficulties in parameter calibration and the high cost of surveys. Based on consideration of the land use, the trip generation/attraction model establishes a relationship between the parking demand and regional travel attraction and estimates the parking demand with the OD data between parking zones. Li [15] established a predictive model based on preset parking zones and vehicular OD data. Incorporation of several OD data in the model allows for calculation of the parking demand in each parking zone. Correspondingly, the estimation result is affected by preset parking zones and the assumption that the vehicle travel OD is restricted by the amount of parking space and travel distance. In this way, it is unable to comprehensively and accurately estimate the parking demand within an urban central district.

The above parking demand estimation models require a considerable amount of time and manual survey costs to comprehensively record the main factors of parking demand to ensure the accuracy of the estimated results. Fortunately, with the rapid development of traffic information technology, large amounts of big data, such as mobile phone [16–18], floating car [19, 20], transit smart card [21], pass-recording [22], and loop detector and remote sensor data [23, 24], are generated during routine operation of transportation systems. In the area of big data, the acquisition of massive parking operation data within the study area may provide better opportunities to refine the estimation models

of the parking demand. For example, Lim et al. [25] studied the impact of adjacent shared parking and parking costs on the parking demand based on the central business district (CBD) of Knoxville, which includes 11,276 off-street and 1,024 on-street parking spaces. Massive parking and travel data have become the new direction of parking demand research. Combined with the considerations listed above, this study proposes a parking demand estimation model driven by the big-data method.

Based on an extensive literature review, we know that there are several estimation research studies about parking demand on a worldwide scale. However, no relevant studies have focused on the travel intensities caused by construction and the location function of the city, which constitutes an important factor that affects the spatial and temporal distribution of parking. This is attributed to the limitations of the model and the difficulty of acquiring and processing traffic big data. As such, this study proposes a big-data-driven framework for parking demand estimation based on the relationship between the travel characteristics obtained from traffic big data sources, land use types, and parking characteristics and dividing parking zones based on various travel densities that can accurately estimate the parking demand for set parking zones.

3. Methods

3.1. Data Description and Processing

3.1.1. Data Description. This study applied OD data, parking survey data for some of the subregions in the study area and land use data. The vehicular OD data was the main data foundation for the two-step parking demand estimation. In the OD data, "O" denotes the trip origin and "D" denotes the trip destination. The OD data can illustrate the travel density characteristics and attractiveness of different subareas of the study area, which directly affects the scale of the parking demand. In this study, adjacent subareas with the same level of travel density were designated as the parking zones. To obtain the parking zones and their travel characteristics, the vehicular OD data had to include the origin and destination coordinate information and departure and arrival times (as shown in Table 1). More importantly, the OD data had to span a large data scale and completely cover the study area.

In addition to OD data, the data involved in this study also included the parking survey data and land use information. The parking survey contained the parking facilities on public land within the subregions of the study area and included urban roadside, independent public, and other pertinent parking facilities. The parking survey data content included the coordinates, peak parking times, and numbers of the parking facilities. Land use type was an important piece information that was used to describe the urban functional characteristics that affect parking demand. This study divided the types of land used into five categories: residential, administrative, education, commercial, and other types. Aggregates of the OD data for each type of land were then used to complete the data precalculation based on the parking zones. The parking zones were used as the basis

Table 1: Origin-destination data description.

Notation	Field	Data type
ID	Trip number	Varchar
O_{lng}	Origin longitude	Double
O_{lat}	Origin latitude	Double
D_{lng}	Destination longitude	Double
D_{lat}	Destination latitude	Double
T_{dept}	Departure time	Datetime
T_{arr}	Arrival time	Datetime

of the parking demand estimation. The parking survey data, OD data, and land use information contained in each parking zone comprised the data required to establish the parking demand estimation model.

3.1.2. Statistical Information Grid (STING).

Parking zones are difficult to divide based on spatially distributed OD data as massive OD data are randomly scattered. Nonetheless, processed data in a disaggregated form is also associated with problem related to excessive calculation. To solve these problems, this study proposes an integrated parking zone division method that combines the statistical information grids clustering algorithm with a multidensity clustering algorithm. The STING clustering algorithm is a grid-based multiresolution clustering method that divides a spatial region into rectangular units [26]. In reference to the data aggregation form of the STING algorithm, this study aggregates traffic data based on the STING subject to a single-resolution condition. As shown in Figure 1, the STING can divide the studied area into grids according to the equidistant side length. It then loads the OD, land use, and parking data into each grid separately.

Each grid is used as the basic information unit of traffic element. Therefore, the grid length should be selected in such a way as to avoid (a) the integration of multiple travel characteristics based on an excessively large grid-side length or (b) too many parking lots divided into different adjacent parking zones owing to the small length of the grid side. The grid can not only preserve the complete traffic characteristics of the research area but also ensure the accuracy and efficiency of subsequent parking zone division calculations. According to the relevant literature, approximately half of the parking lot service radius is 100 m, and approximately 3/4 of the parking lot service radius is less than 200 m [27]. Considering the service scope of the parking lot, a relatively small grid length is recommended. It is worth noting that the aggregated parking zone usually contains several grids that cover the service range of a parking lot, which is several times larger than the grid. Meanwhile, a relatively small side length is beneficial because it divides the parking zones to have unified travel intensities.

3.2. Illustration of the Two-Step Method.

In this section, we demonstrate the two-step parking demand estimation method proposed in this study. Before we illustrate the concept, the relationships between the traffic attributes,

parking demand, and land use should be verified based on the OD data, parking data, and land use data.

This study analyzed the spatial relationship between the OD distribution and allocation of parking lot resources. As shown in Figure 2(a), the red color denotes high traffic density and the green color denotes low traffic density. To ensure the continued operation of urban traffic, subareas with a high-travel occurrence and attraction density (i.e., a dense OD distribution) also require more parking spaces to meet the parking demands. The correspondence between the distribution of the travel density and the spatial allocation of parking resources supports this research. In addition, we analyzed the obvious correlation between the travel density and the land use attributes. Figures 2(a) and 7 show that areas with high-travel occurrences and density, such as the delineation shown in the figure, often correspond to commercial and administration land sites.

The generalized framework used to solve the parking demand estimation problem contains two steps, namely, (a) parking zone division and (b) parking demand estimation. The specific methodology framework is shown in Figure 3. First, based on the STING, the studied area was divided into grids that contained vehicular OD data. Various parking zones with the same travel density levels were then identified using a multidensity clustering algorithm. The parking data, land use information, and OD data were aggregated separately for each parking zone in accordance with the relevant data within the statistical information grids. Considering the missing data and manual recording errors in the parking survey data of some of the parking zones, the parking zones with high OD density levels, and low or no peak parking numbers, were defined as the parking zones for estimating, while the zones that remained after screening were defined as the effective parking zones needed for the establishment of the parking demand estimation model.

Second, the peak parking times for the parking zone estimation were calculated using the Bayes classifier, which was based on the complete parking survey data in the effective parking zones. Thus, the parking data, land use information, and OD data associated with the effective parking zones allowed the establishment of the parking demand estimation model with the SVM and calculation of the parking demand of the parking zones for estimating.

3.2.1. Step 1: Parking Zone Division Based on Statistical Information Grids.

Evidently, parking zones with the same travel density levels exhibit an irregular boundary given the influence of the land use type and construction level. This feature matches the computational performance of the multidensity clustering algorithm. The multidensity clustering algorithm based on the density-based spatial clustering of applications with noise (DBSCAN) was selected as the guiding principle for dividing parking zones. DBSCAN is one of the most representative algorithms in density clustering. The core idea of this algorithm is based on the definition of the search neighborhood and density threshold followed by the hierarchical density point aggregation in clusters [28, 29]. This method does not preset the core region

FIGURE 1: Concept diagram of statistical information grid and data aggregation.

or the cluster boundary and can identify clusters of different shapes and sizes in noisy data.

The improved multidensity clustering algorithm based on STING data forms can accurately identify parking zones in the entire area of study and has been associated with multiple travel density levels. The following is a list of the definitions of variables involved in the multidensity clustering algorithm:

(a) ε neighborhood: the area within the search radius, ε, for a given object is called the ε neighborhood of the point.

(b) MinPts: the threshold for multidensity clustering is the initial condition for a given object to become a core object within the ε neighborhood. This threshold gradually decreases according to the density threshold iteration formula.

(c) Object, O: clustered data unit. In this case, it is the information grid loaded with OD data.

(d) Core object, p: if the density of object O within the ε neighborhood is not less than MinPts, then the object is the core object, p, and the core objects that are adjacent and satisfy the same density threshold condition are aggregated into one cluster.

(e) Boundary point, q: if the density of the ε neighborhood of object O is less thanMinPts and the object is adjacent to a core object, p, it is included in the cluster and is defined as a boundary point.

(f) Noise object: a data object that does not belong to any class and whose direct density is reachable is called a noise object.

The multidensity clustering algorithm can identify parking zones with the same travel intensities driven by massive OD data and can, therefore, highlight the characteristic relationship between the travel density and parking demand. The specific description of the algorithm is as follows:

Step 1: the studied area is divided into STINGs with identical side lengths. This is followed by the precalculation and storage of the OD data for each grid as object O.

(a)

Legend:
- Residential land
- Administration land
- Education land
- Commercial land
- Other lands

(b)

FIGURE 2: (a) Origin-destination kernel density distribution coupled with parking resource allocation. (b) Distribution of land types.
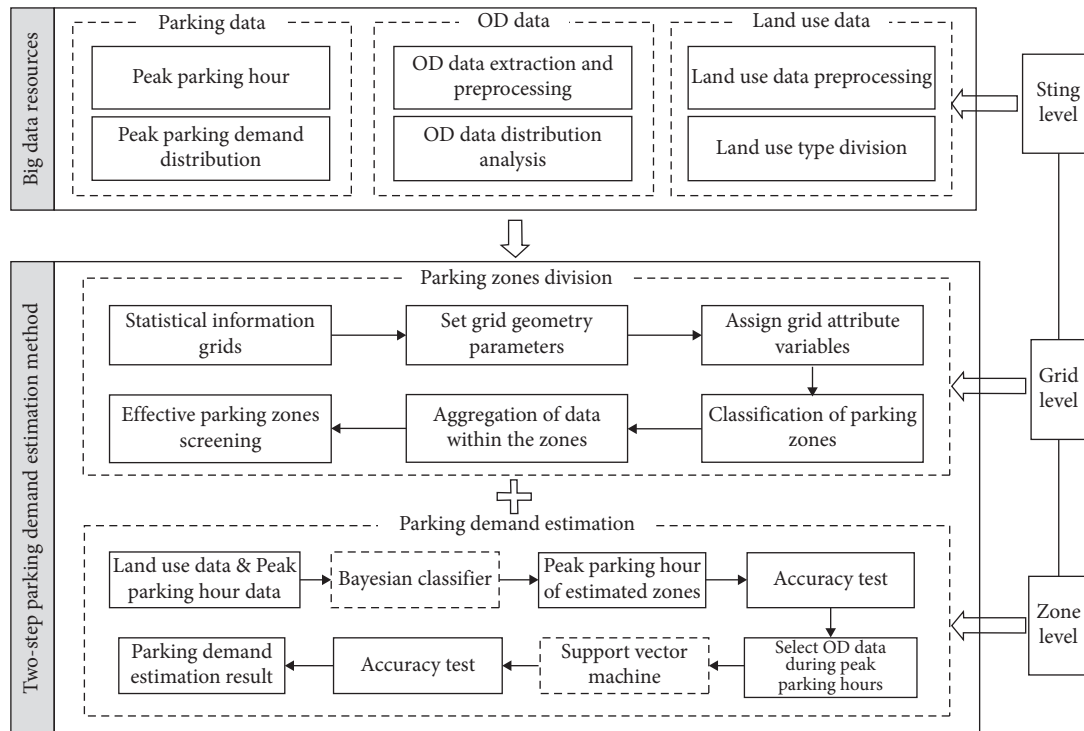

FIGURE 3: Methodology framework adopted to resolve the parking demand estimation problem.

Step 2: the object set Set_i^O is obtained by scanning the entire study area to count the density of each object, O, in the neighborhood of ε, and a descending sort is executed for the object set Set_i^O based on the density.

Step 3: set $i = 0$ and perform a sequential clustering operation according to the order of Set_i^O. Set_i^O is filtered according to the initial density threshold MinPts_0. The object with a number of OD data above the threshold is then obtained to build the core object set, p. The object is selected from p, and the objects within the neighborhood of ε are aggregated in the same cluster if the density is higher than the threshold condition. The aggregation clusters are the parking zones. The objects in the clusters are then moved out of the object set Set_i^O.

Step 4: set $i = i + 1$ and update object set Set_i^O to Set_{i+1}^O. Meanwhile, the threshold condition is changed to MinPts_i based on the density threshold iteration formula:

$$\text{MinPts}_i = \text{MinPts}_0 - \alpha i, \qquad (1)$$

where α is the rate of density threshold change and i is the number of iterations.

Return to step 2 and repeat the steps until $\text{MinPts}_i = 0$. By setting the density threshold formula, the threshold is gradually reduced to zero according to a certain gradient. The parking zones with higher travel intensities are identified first, and the parking zones with the same travel density level are then divided among the overall study area.

3.2.2. Step 2: Parking Demand Estimation Based on Parking Zones.
There is a correlation between the OD data during the day and the parking demand, but this is not obvious given that the parking demand is the sum of the peak parking numbers in the parking zone. If the OD data can be selected according to the time characteristics reflected by the parking demand of each parking zone separately, this will ensure the estimation accuracy of the improved correlation. The temporal characteristics of adjacent parking lots in the same parking area are strongly correlated, especially in the peak parking time. When the vehicle travels to a parking lot and cannot park owing to space unavailability, it often goes to the adjacent parking lot to park. This phenomenon occurs more frequently during the peak-parking periods. Owing to the small scope of the parking zones, all the parking lots on public land within the zone include the adjacent parking lots that can complement each other to serve the parking demand during peak-parking periods. Therefore, the peak parking time of a zone is obtained by calculating the average value of the peak time of each parking lot in the parking zone. The precalculation of the OD data before the estimation of the parking demand is the selection of the OD data during the peak parking hour of each parking zone.

It is necessary to obtain the peak parking time for each estimated parking zone to screen out the OD data of the corresponding time period for parking demand estimation purposes. However, owing to the lack of complete and accurate parking survey data in the estimated parking zones, it is difficult to determine the peak parking hour of the zones. Estimating the peak parking hour of the estimated parking zones according to the parking data of the effective parking zones is necessary. The peak parking hour of different parking zones depend on the land use characteristics and the function of the zone in the city. Traffic generated in different lands are usually associated with different travel purposes. Meanwhile, the parking demand for various travel purposes also exhibit significant differences in the parking time distribution. Therefore, according to the area ratio of five different types of land use in the parking zones, the peak parking time is estimated by the Bayes classifier.

Based on the prior probability, the classification principle of the Bayes classifier applies the Bayes formula to calculate and select the class with the largest posterior probability as the class to which the object belongs to. The Bayes classifier is a method of supervised learning [30]. This study established the classifier through parking survey and land use data associated with the effective parking zones to calculate the peak parking time of the estimated parking zones with similar land-type combination characteristics. Each variable is represented as a node (referred to as the child) and the class variable is represented as a label (referred to as the parent) for all the other variables. The Bayes classifier is defined as

$$v_{\text{MAP}} = \text{argmax}_{v_j} \in v^{P(v_j)} \prod_{i=1}^{n} P(x_1, x_2, \ldots, x_n | v_j). \qquad (2)$$

The most probable target value is denoted by v_{MAP}, while v is a finite set building on every target value, v_j, that shows the probability of a specific peak parking time. Let X_i ($i = 1, 2, \ldots, n$, for n variables) assume that each variable has the respective values of $x_1, x_2, \ldots, x_n$, that describe the area ratios of each land type within the parking zones. One set of training instances with a specific class is given to assign the most probable target value of the test instance using the Bayes approach for classification. The established classifier can estimate the class distribution of an instance with its class unknown.

The naive Bayes (NB) classifier makes the independence assumption that the input variables are conditionally independent of each other. This is suitable to the features that the area ratio correlation between various land types in the parking zones is low. The combination of various land types determines the function and development direction of this subarea. This can directly affect the travel characteristics, but the area ratios of different land types do not affect each other considerably. The NB classifier is best suited when the independence assumption is valid. At the same time, the NB classifier can not only maintain good computational efficiency but also update the probability values of different classes of classifiers in real time according to the new training data. As the city develops, each updated training dataset may cause changes in the classification probability values based on the relationship among the traffic density, land use type, and changed peak parking time. The updateable classifier ensures that the estimation of the peak

parking time can be adjusted according to urban development. Therefore, this can maintain increased reliability pertaining to the estimated results. The NB classifier can be defined as follows:

$$v_{\mathrm{NB}} = \mathrm{argmax}_{v_j} \in v^{P(v_j)} \prod_{i=1}^{n} P(x_j | v_j). \tag{3}$$

Each variable node in NB has a class node as its parent but does not have any parent from variable nodes. Based on the effective parking zones, the NB classifier is established to estimate the peak parking hours of the estimated parking zones. Subsequently, the OD data during the peak parking time of each estimated parking zone is selected for the parking demand estimation calculation.

The data of the parking zones used for parking demand estimation include the peak parking number, area ratios of different land types, the numbers of OD data on various types of land usages, and the entire parking zone during the peak parking times and have increased dimensionality. Considering the characteristics of the dataset, this study uses the SVM model to establish the estimation model for parking demand. SVM is a machine learning method based on statistical learning theory, VC dimensional theory, and structural risk minimization principle. Accordingly, it possesses obvious advantages in solving nonlinear and high-dimensional pattern recognition problems subject to small sample size conditions. SVR models are based on SVM to fit

curves and conduct regression analyses. The generalized regression problem for training SVR is based on the linear classification hypothesis [31]. Accordingly, $f(x) = <\omega, x_i> + b$ can be modelled for $\omega \in R^n$ and the offset $b \in R$. The objective function in the SVM is used to maximize the classification interval based on the point-to-line distance formulas to derive the objective function that is written as a convex optimization problem.

$$\mathrm{minimize} \frac{1}{2} \|\omega\|_2^2,$$

$$\mathrm{subject\ to} \begin{cases} y_i - <\omega, x_i> - b \leq \varepsilon, \\ \\ <\omega, x_i> + b - y_i \leq \varepsilon. \end{cases} \tag{4}$$

The constraint of the optimization problem was based on the setting of the estimated error of $f(x)$ to a value less than ε, where (x_i, y_i) is the spatial eigenvector, ω is the N-dimensional weight vector, b is a constant called deviation, C is the cost parameter, and ξ_i and ξ_i^* denote the slack variables.

The convex optimization problem is only realized based on the hypothesis that the training set data is linearly separable but cannot easily deal with high-dimensional OD and parking data stored in the parking zones. When the training set is linearly inseparable, the optimization problem needs to be calculated by introducing slack variables.

$$\mathrm{minimize} \frac{1}{2} \|\omega\|_2^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*),$$

$$\mathrm{subject\ to} \begin{cases} y_i - <\omega, x_i> - b \leq \varepsilon + \xi_i, & \forall i, \\ <\omega, x_i> + b - y_i \leq \varepsilon + \xi_i^*, & \forall i, \\ \xi_i, \xi_i^* \geq 0, & \forall i, \end{cases} \tag{5}$$

where C is the weighting parameter of the error cost and ξ_i and ξ_i^* are the slack variables. By introducing the Lagrange function to deal with the constraint conditions, the dual problem of the optimization problem is derived according to

$$f(x) = \sum_{i=1}^{n} (a_i + a_i^*)\langle x_i, x \rangle + b, \tag{6}$$

where α_i and α_i^* are the Lagrange multipliers. When the training set cannot be divided in the original low-dimensional space, the support vector machine can achieve the goal of mapping the low-dimensional data in the high-dimensional space by defining the kernel function $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ instead of the inner product $\langle x_i, x \rangle$ used in (6). To construct an estimation model for the parking demand with a good performance, the choice of the kernel function is the most critical step. Most research studies commonly use the linear kernel, polynomial kernel functions, and the radial basis

function kernel (RBF). The RBF kernel can be readily applied to datasets that comprise large sample sizes and different data dimensions and can be used to map a sample to a higher dimensional space. The RBF kernel requires fewer parameters than polynomial kernel functions, and the lower functional complexity makes the model have a better computational efficiency. This study estimated the parking demand with the SVR model based on RBF kernel function. The model has two important parameters, C and σ, where C is the penalty factor, which is the tolerance for the error, and σ is a parameter defined after the selection of the RBF function as the kernel that implicitly determines the distribution of data after mapping to a new feature space.

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right), \tag{7}$$

where $\sigma > 0$ is the control parameter of the RBF kernel function.

Based on the *K*-fold cross-validation used to calculate the cost parameter *C* and the kernel function parameters, this study established the parking demand estimation model with the SVM. Meanwhile, several other estimation models were proposed for comparison, such as the naive SVM, linear regression, and quadratic polynomial regression models. The naive SVM model does not consider land use information. The linear regression and the quadratic polynomial regression models are established through the strong correlation between the OD data during the peak parking time and the parking demand of the parking zones. Evaluating the accuracies of the various estimation models for parking demand (as listed above) can be accomplished based on relative error indicators.

4. Results and Discussion

4.1. Data Description. A rectangular area defined by latitude and longitude (Lng ∈ [116.806, 116.891], Lat ∈ [38.291, 38.315]) in the city center of Cangzhou was selected as the research area. There are two municipal districts in Cangzhou (Xinhua District and Yunhe District) with a total area of 218 square kilometers. Besides, the research area is located in the central of the two districts, covering 72.5 square kilometers, accounting for 33% of the total urban area. This area can provide sufficient traffic survey data because travel is more frequent and the parking demand is larger than in the suburban areas. Actually, the total of 207,997 OD data of online car-hailing in the surveyed area in February 2017 is considered. Besides, all the parking facilities on the public land within the research area were surveyed.

The parking surveys were conducted over a period of 6 days (from 8 am to 8 pm) during the entire week (working days and weekend). Parking survey data include the coordinates, the peak parking time, and the numbers of the parking facilities. Considering that the parking facilities do not change during short time periods, the study selected OD data within the parking survey period (8 am to 8 pm) given that the OD data can reflect the travel characteristics more appropriately based on the development of the online car-hailing platform. The numerical abnormalities or formatting errors in the dataset were screened based on the total of 207,997 OD data of online car-hailing in the research area. This study also considered land use information. Many traditional parking demand estimation models considered these factors, including the parking generation rate model related to the land use type. The parking generation rate parameter refers to the parking number per unit area for each land type and land use. The land types are classified based on the actual land use characteristics, as shown in Figure 4.

4.2. Parking Zone Division Results. The foundation of parking zone division is the statistical information grid. Considering the service scope of the parking lot and the unity of travel density, this study set up STINGs with side lengths of 100 m and 300 m for comparisons, as shown in Figure 5. The figure illustrates the spatial distribution (center position) and peak parking number (circle radius) of all the public parking facilities and the different travel density distributions distinguished by color in the destination data kernel density map. The parking zones identified by the 200 m side length grids cover most of the service range of the parking lot. Meanwhile, the grids with a side length of 100 m have travel density levels that are close to the adjacent grids and can, therefore, be used to divide the information grids with similar travel levels into the same parking zones. In summary, this study selected the STING with a side length of 100 m as the foundation for parking zone division and the calculation time is 11.3 min based on Intel Core i5-6500 3.20 GHz CPU and 8 GB RAM. Effective parking zones with complete parking data and estimated parking zones with insufficient or missing parking data were thus determined based the parking survey data in different parking zones.

4.3. Parking Demand Estimation Results. There are 65 effective parking zones in the studied area of which 45 were used for model training referred to as training parking zones. Of these, 20 were used for accuracy analyses of the estimation model commonly referred to as the test parking zones. Considering the OD data of online car-hailing and area ratio of various types of land use in each parking zone, the parking demand estimation model was established by SVR, since the number of parking zones divided in this study is limited, and there are high-dimensional attribute data, which is suitable for SVR calculation characteristics. Clearly, the instance is solved in a workstation equipped with Intel Core i5-6500 3.20 GHz CPU and 8 GB RAM and the time to calculate SVR is 1.2 s. Accordingly, it is also feasible to use the parking demand estimation framework proposed in this study to calculate the parking demand of a larger area including suburbs. Meanwhile, several other parking demand estimation models were used for comparative analyses. The linear regression model and the quadratic polynomial regression model were established based on OD data and the parking demands of the training parking zones. In addition, we used the *R*-square value in the monadic regression model and its adjustment in the multivariate regression model as the correlation evaluation indices. Table 2 listings show that there is a high correlation between the independent variable and the peak parking demand, and the correlation can be further improved by considering the land attribute data in the independent variables. The table also illustrates that the estimation accuracy of the SVR algorithm is considerably higher than the traditional regression model and considering the land use type factor can further improve the reliability of the estimated model.

y_i is the parking demand in the parking zone i, x_{di} is the number of trip arrival data in parking zone i, and x_{oi} is the number of trip departure data in parking zone i.

Figure 6 illustrates the estimated relative error results of the three parking demand estimation models with the best estimation accuracy. The estimated method proposed in this study yields higher estimation accuracies in most parking zones. Based on this analysis, the specific relative error in the
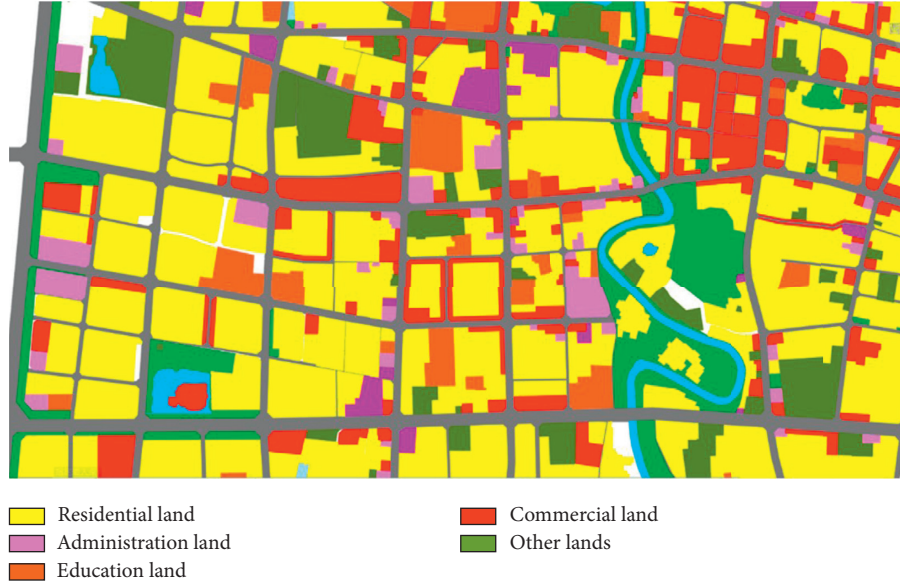
Figure 4: Schematic of land use types in the central area of Cangzhou.
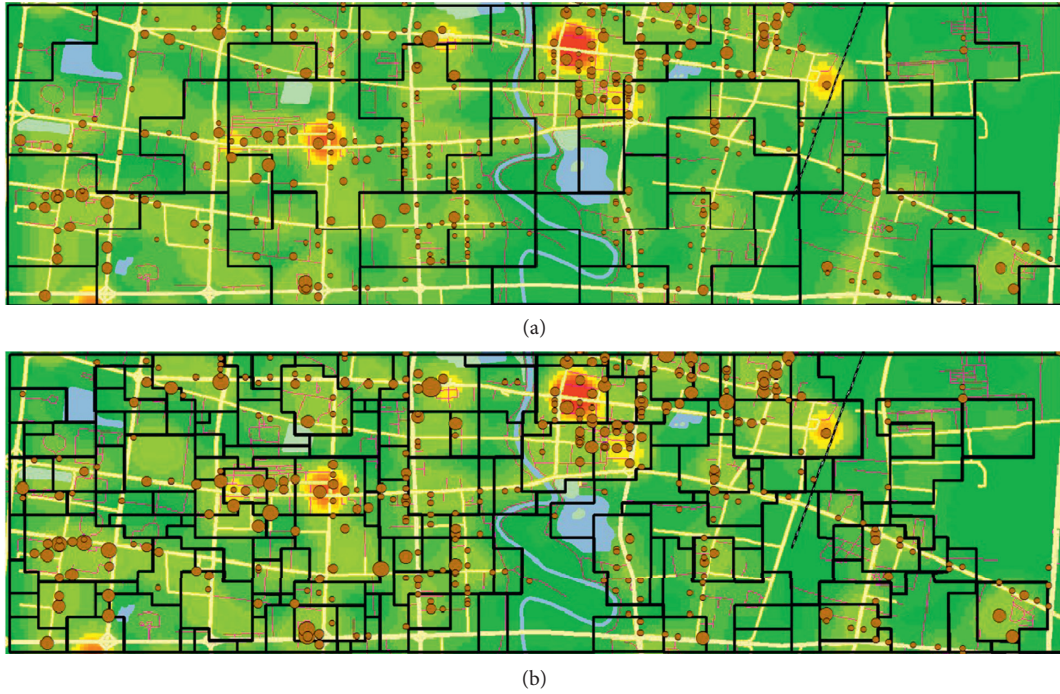

(a)

(b)

Figure 5: Parking zone division outcomes at different side-length conditions. (a) Statistical information grids with a side length of 300 meters and (b) Statistical information grids with a side length of 100 meters.

Table 2: Verification results of different regression models.

| Regression model	Independent	R^2	Regression equation	Difference (%)
Support vector regression (SVR)	OD data and area ratio of land types	0.86	——	18.06
SVR (without land attributes)	Travel OD data	0.80	——	30.78
Linear regreesion	Origin data	0.80	$y_i = 0.43x_{\mathrm{oi}} - 38.17$	50.65
	Destination data	0.73	$y_i = 0.25x_{\mathrm{di}} + 34.11$	37.83
Quadratic polynomial	Origin data	0.82	$y_i = 5.95 \times 10^{-5}x_{\mathrm{oi}}^2 + 0.27x_{\mathrm{oi}} + 5.42$	36.37
	Destination data	0.73	$y_i = 5.16 \times 10^{-7}x_{\mathrm{di}}^2 + 0.25x_{\mathrm{di}} + 34.55$	37.65
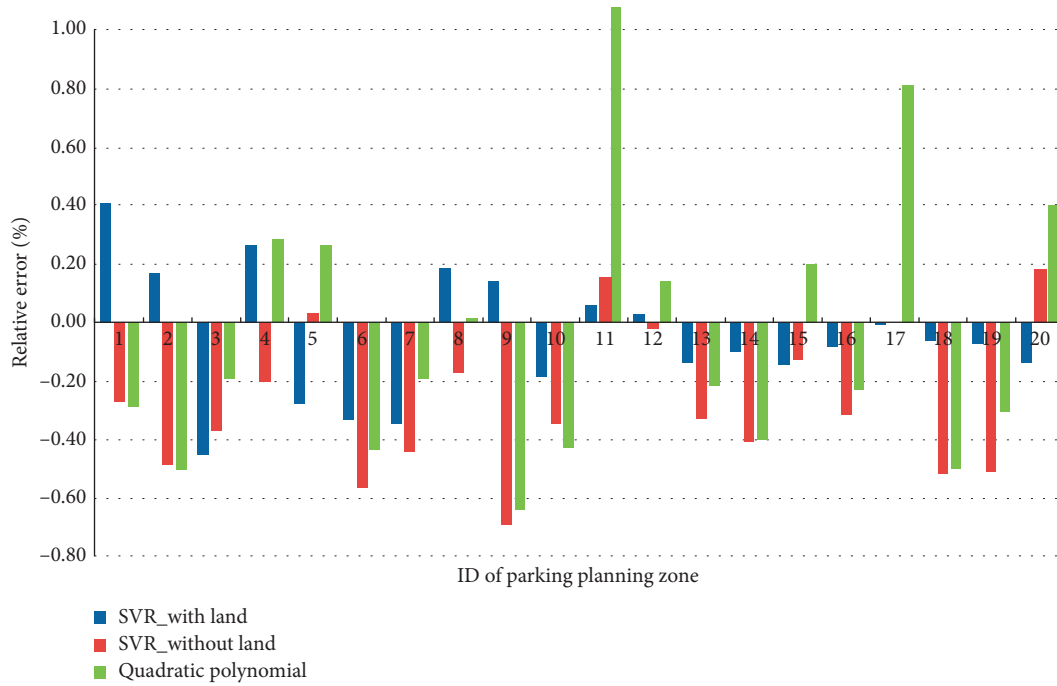
FIGURE 6: Bar graph of the estimation error distribution in each tested parking zone.
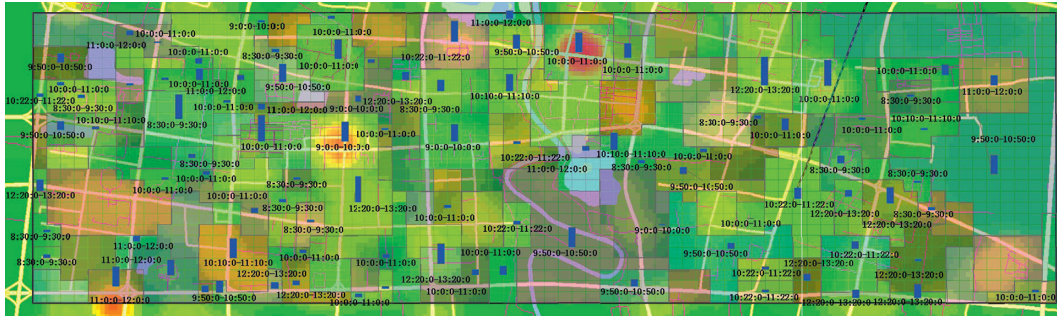

FIGURE 7: Peak parking hours and parking demand column for estimated parking zones.

20 tested parking zones can also indicate that there are more cases where the predicted value is smaller than the actual observed value. This indicates that the independent variable data is insufficient to some extent when the parking demand is characterized, but it is acceptable to analyze these shortcomings from the perspective of accuracy compared with the traditional parking demand estimation methods.

The parking demand estimation model proposed in this study can yield the most accurate estimation results with a maximum estimated accuracy of 82%. The parameters of the SVR model with RBF kernel function were obtained by the K-fold cross-validation ($C = 2$, $\sigma = 0.05$). The zones without the estimated results in Figure 7 are the effective parking zones used for model training and testing. The peak parking hours and parking demand of each estimated parking zone can be calculated based on the big-data-driven framework proposed in this study. The height of the blue rectangle represents the relative scale of the parking demand. Considering the kernel density map of the travel destination data, the estimated parking zone with high-travel density or larger area has a higher parking demand that is consistent with the law of the actual density of the demand for parking resources.

5. Conclusions

This study proposed a big-data-driven framework for the estimation of the parking demand. The method can precalculate and store big data with multiple source isomerisms, such as travel big data, parking survey data, and land use information based on STINGs and divided parking zones with the same travel density level through multidensity clustering algorithm. Furthermore, the method adopts the machine learning method to combine various types of data resources to estimate the parking demand, which not only reduces the traditional parking survey cost but also ensures the efficiency and accuracy of big-data calculation.

The city center of Cangzhou was used for experiments to test the performance of the proposed methodology. A detailed analysis was then conducted based on the comparison

with several other parking demand estimation models in which the relative errors of all estimations were calculated. The results showed that the estimated accuracy of this methodology was as high as 82% in most parking zones. Additionally, the framework proposed in this study can be used to estimate a larger range of parking demand based on related data among cities and suburbs.

This study proved that by mining the relationship between travel and parking survey data, it is possible to accurately estimate the parking demand based on the OD data generated in modern traffic systems. This broadens the research opportunities for parking demand estimation based on the application of traffic big data. Further research work is recommended to obtain parking data that is automatically recorded every minute in intelligent parking lots for the construction of a historical real-time database that can promote real-time, parking-demand estimation research. Research efforts will also allow the development of more efficient methods to solve the parking demand estimation problem based on the optimization of the structure and the algorithm of the framework.

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 they have no conflicts of interest.

Acknowledgments

References

[1] R. Rastogi, "Preference of vehicle parker's in Kota City-a stated preference study," *Indian Highways*, vol. 34, no. 3, pp. 45–59, 2006.

[2] P. Barter, "Off-street parking policy surprises in Asian cities," *Cities*, vol. 29, no. 1, pp. 23–31, 2012.

[3] Z. Zhu, Y. Lu, J. Zeng, and H. Chen, "Highway traffic accident influence area based on vehicle bypass decision-making index," *Journal of Advanced Transportation*, vol. 2017, no. 1-8, Article ID 5270951, 2017.

[4] O. Alvear, N. R. Zema, E. Natalizio, and C. T. Calafate, "Using UAV-based systems to monitor air pollution in areas with poor accessibility," *Journal of Advanced Transportation*, vol. 2017, no. 1-14, Article ID 8204353, 2017.

[5] S. D. Clark and S. Rey, "Temporal dynamics in local vehicle ownership for Great Britain," *Journal of Transport Geography*, vol. 62, pp. 30–37, 2017.

[6] D. C. Shoup, "Cruising for parking," *Transport Policy*, vol. 13, no. 6, pp. 479–486, 2006.

[7] M. D. Sanjaykumar, G. J. Joshi, K. Ravinder, and N. Gore, "Data monitoring for the assessment of on-street parking demand in CBD areas of developing countries," *Transportation Research Part A: Policy and Practice*, vol. 126, pp. 152–171, 2019.

[8] E. Ben-Elia and D. Ettema, "Rewarding rush-hour avoidance: a study of commuters travel behavior," *Transportation Research Part A Policy & Practice*, vol. 45, no. 7, pp. 567–582, 2011.

[9] J. Li, C. H. Liu, S. Y. Hu, and F. Wang, "Travel time prediction model based on traffic data fusion technology," *Journal of Hunan University Natural Sciences*, vol. 41, no. 1, pp. 33–38, 2014.

[10] ITE, *Parking Generation (Report No. IR-034C)*, ITE, Washington, DC, USA, 2019.

[11] Y. Shan, *Research on Parking Demand Forecast of Tonghua City*, Jilin University, Changchun, China, 2017.

[12] A. S. Khaled and H. Jamil, "Establishing parking generation rates/models of selected land uses for Palestinian cities," *Transportation Research Part A: Policy and Practice*, vol. 91, pp. 213–222, 2016.

[13] Y. Xie, B. Cheng, and K. Dong, "Parking demand estimated model based on analysis of land," *Transportation Standardization*, vol. 19, no. 13-18, 2014.

[14] Y. Dai, L. Wu, X. Zhang et al., "Parking demand estimated based on traffic flows and service level," *Technology & Economy in Areas of Communications*, vol. 15, no. 6, pp. 35–39, 2013.

[15] X. L. Li, "Study on estimated model of parking demand based on OD matrix," *Experimental Technology & Management*, vol. 31, no. 10, pp. 83–85, 2014.

[16] J. Steenbruggen, E. Tranos, and P. Rietveld, "Traffic incidents in motorways: an empirical proposal for incident detection using data from mobile phone operators," *Journal of Transport Geography*, vol. 54, pp. 81–90, 2016.

[17] T. Bantis and J. Haworth, "Who you are is how you travel: a framework for transportation mode detection using individual and environmental characteristics," *Transportation Research Part C: Emerging Technologies*, vol. 80, pp. 286–309, 2017.

[18] C. Gariazzo, M. Stafoggia, S. Bruzzone, A. Pelliccioni, and F. Forastiere, "Association between mobile phone traffic volume and road crash fatalities: a population-based case-crossover study," *Accident Analysis & Prevention*, vol. 115, pp. 25–33, 2018.

[19] Y. Huang, D. Sun, and L.-H. Zhang, "Effects of congestion on drivers' speed choice: assessing the mediating role of state aggressiveness based on taxi floating car data," *Accident Analysis & Prevention*, vol. 117, p. 318, 2018.

[20] R. Günther, T. Wenzel, M. Wegner, and R. Rettig, "Big data driven dynamic driving cycle development for busses in urban public transportation," *Transportation Research Part D: Transport and Environment*, vol. 51, pp. 276–289, 2017.

[21] N. Nassir, M. Hickman, and Z.-L. Ma, "A strategy-based recursive path choice model for public transit smart card data," *Transportation Research Part B: Methodological*, vol. 126, pp. 528–548, 2019.

[22] W. Yuan, P. Deng, T. Taleb, J. Wan, and C. Bi, "An unlicensed taxi identification model based on big data analysis," *IEEE Transactions Intelligent Transportation Systems*, vol. 17, no. 6, pp. 1703–1713, 2016.

[23] J. Tang, F. Liu, Y. Zou, W. Zhang, and Y. Wang, "An improved fuzzy neural network for traffic speed prediction considering periodic characteristic," *IEEE Transactions Intelligent Transportation System*, vol. 99, no. 1-11, 2017.

[24] Y. Zou, H. Yang, Y. Zhang, J. Tang, and W. Zhang, "Mixture modeling of freeway speed and headway data using multivariate skew-t distributions," *Transportmetrica A: Transport Science*, vol. 13, no. 7, pp. 657–678, 2017.

[25] H. Lim, G. T. Williams, D. Abdelqader et al., "Alternative approach for forecasting parking volumes," *Transportation Research Procedia*, vol. 25, pp. 4171–4184, 2017.

[26] W. Wang, "STING: a statistical information grid approach to spatial data mining," in *Proceedings of the 23rd Very Large Database Conference*, Athens, Greece, August 1997.

[27] H. Z. Guan and X. M. Li, *Parking Planning Design and Management*, China Communications Press, Beijing, China, 2003.

[28] C. F. Tsai and C. Y. Sung, "MIDBSCAN: an efficient density-based clustering algorithm," in *Proceedings of the Sixth International Symposium on Neural Networks (ISNN 2009)*, May 2009.

[29] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the International Conference on Knowledge Discovery & Data Mining*, Portland, OR, USA, August 1996.

[30] E. Manino, L. Tran-Thanh, and N. R. Jennings, "On the efficiency of data collection for multiple Naïve Bayes classifiers," *Artificial Intelligence*, vol. 275, pp. 356–378, 2019.

[31] I. Zoppis, G. Mauri, and R. Dondi, "Kernel methods: support vector machines," *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1, pp. 502–510, 2019.