IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 25, 2018, accepted August 27, 2018, date of publication September 10, 2018, date of current version October 25, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2869434

Design and Application of an Attractiveness Index for Urban Hotspots Based on GPS Trajectory Data

LI CAI[1,2], FANG JIANG[2], WEI ZHOU[2], AND KEQIN LI[3], (Fellow, IEEE)

[1]School of Computer Science, Fudan University, Shanghai 200433, China
[2]School of Software, Yunnan University, Kunming 650091, China
[3]Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

Corresponding author: Wei Zhou (zwei@ynu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grants 61663047 and 61762089.

ABSTRACT Urban hotspots refer to regions where flourishing shopping centers are located, the travel volume is very large, and there is high traffic. The formation of hotspots is strongly correlated with many features, i.e., time, space, and the distribution of points of interest; however, most studies have used qualitative analyses to describe the relevance of these features, and there is a lack of quantitative analyses. Therefore, we propose the concept and a model of the attractiveness index of a hotspot that is used to quantify the spatio-temporal distribution of the hotspots and determine the degree of attractiveness to the residents. In addition, a novel algorithm of hotspot similarity is designed and implemented to improve the efficiency of summarizing the hotspot data, which is usually performed manually. We mine the data and determine different hotspots using a one-week GPS trajectory dataset collected from 6599 taxis in Kunming. Furthermore, we calculate the attractiveness index of hotspots and visualize their characteristics. The research results provide a scientific basis and reference for public infrastructure planning, land-value evaluation, store location planning, consumption recommendations, and other applications.

INDEX TERMS Attractiveness index, clustering mining, hotspots, similarity calculation, trajectory data.

I. INTRODUCTION

Urban hotspots are areas of intense urban economic development and are an important part of urban comprehensive competitiveness [1]. A region is designated a hotspot because it contains many types of infrastructure that are necessary for the daily life of residents, such as restaurants, entertainment, schools, and hospitals. The use of data mining to determine hotspots of interest to residents can provide a scientific basis and a reference for location-based services. Taxis are a public transportation type characterized by all-weather operating and real-time data. Therefore, taxi trajectories are a good representation of the residents' travel patterns and commuting behavior and have become an important source of data mining for the detection of urban hotspots [2], [3].

Liu *et al.* [4] proposed a variant of the density peaks clustering (DPC) approach for discovering demand of hot spots from a low-frequency, low-quality taxi fleet operational dataset. Their approach could be of use to both taxi fleet operator and traffic planners in guiding drivers and setting up taxi stands. Pan *et al.* [5] used an iterative density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm to find urban hotspots, which were then used to create a classification of urban land functions. Chen *et al.* [6] planned a city's night bus route based on the detected hotspots. Li *et al.* [7] analyzed the pick-up/drop-off behaviors of taxi drivers derived from the GPS trajectories of taxicabs and recommended the best locations where the drivers could find passengers. In addition to using the GPS trajectory data of vehicles as a data source, mobile location data and check-in data have also become a new data source for data mining of urban hotspots. Klessig *et al.* [8] used a mobile location dataset of a 3G mobile network in a northern European city and automatically detected and tracked the traffic hotspots using image processing techniques. Louail *et al.* [9] determined the number of hotspots (locations where mobile phone users congregate) as a function of the overall city population and observed the spatial and temporal stationarity of these hotspots, which represented the hearts of cities. Hu and Zhang [10] use the Weibo check-in data in Wuhan which is between 2011 and 2015 to mine commercial hotspots.

After urban hotspots are discovered, most literatures mainly focus on their spatio-temporal patterns, but seldom analyze the factors affecting the formation of hotspots.

For example, large office buildings and railway station are both hotspots. Large office buildings are characterized by high travel volume, but the time forming a hotspot is limited to two time periods of commuting. However, except late at night, the railway station has a large traffic volume all day. In addition, even the same type of hotspots, such as shopping malls, because of their location, scale and surrounding environment, the degree of attraction to citizens is completely different. Due to the lack of quantitative analysis, the differences between these hotspots cannot be compared. Therefore, the objectives of this paper is how to establish an effective quantitative model to describe the attractiveness of each hotspot.

II. RELATED WORK

The traditional analysis for urban hotspots has used spatio-economic data such as land use changes, population, and density to reveal the attraction of tourist scenic spots, the spatio-temporal changes of prosperous districts and the spatial distribution of passenger flow by the questionnaires. At present, it is the most commonly method to study the urban hotspots through GPS trajectories collected from vehicles and individuals. The extraction techniques of hotspots include spatio-temporal clustering, mesh generation and spatial statistics.

Wang *et al.* [11] investigated the real-time demand-supply level for taxis and made an adaptive tradeoff between the utilities of drivers and passengers for different hotspots. Then, they constructed a recommendation system that achieve a remarkable improvement on the global utility and make equilibrium between the utilities of drivers and passengers, simultaneously. Zheng *et al.* [12] mined interesting locations and classical travel sequences within a given geospatial region using the GPS trajectories generated by multiple users and provided the travel recommendation for mobile tourists. Kisilevich *et al.* [13] presented P-DBSCAN, a new density-based clustering algorithm based on DBSCAN for analysis of hot places and events using a collection of geotagged photos. Gui *et al.* [14] proposed a MapReduce-based two-steps distributed parallel algorithm for extracting traffic hotspot areas from the taxi track. Chen *et al.* [15] referred to image-based clustering algorithm applied on matching map of the vehicle networking data and ultimately obtained the traffic hot spot map of the urban roads by the clustering analysis.

For discovering efficient and inefficient passenger-finding strategies from a large-scale taxi GPS dataset, Li *et al.* [7] partitioned Hangzhou metropolitan area into 40×20 grids with equal intervals, then they counted all the pick-up and drop-off events during the selected 15 days in each region and selected the top 99 busiest regions as the hotspots area. Rong *et al.* [16] studied the influential factors in passenger seeking strategies and found algorithms to guide taxi drivers to passenger hotspots with the right timing. Their proposed strategies may predict the taxi rides at different times per day, and increase the taxi drivers income levels by controlling appropriate mileage per trip and following the route across more urban hot spots.

Li *et al.* [17] used kernel density estimation to identify the urban' hotspots. The region with high kernel density corresponds to the area with larger distribution of trajectory points. Qi [18] combined the information entropy theory with the spatial statistical analysis theory, and analyzed the spatial distribution characteristics of multi-day taxi's pick-up points in Shenzhen for determining the high gathering places.

Cao *et al.* [19] performed the random walks model that exploit massive amounts of GPS records collected from multiple users for identifying top-k significant semantic locations. Moreira-Matias *et al.* [20] proposed a prediction model based on probability model, which can be well applied to a recommend system for helping taxi drivers to choose the best taxi stand. Scholz and Lu [21] proposed a method to detect the dynamics of space-time development of urban activity patterns that are embedded in large volume trajectory data, at the same time, they used Poisson distribution in the process of detecting hotspots.

Several thorough and detailed studies have been conducted on the use of data mining for the detection of urban hotspots. However, most of the existing studies only used qualitative methods to analyze the spatio-temporal pattern of hotspots and few quantitative analyses have been conducted. For example, Zhao [22] divided the hotspots generated by the taxi trajectories in Wuhan into persistent hotspots and transient hotspots but he did not compare the spatio-temporal differences in these regions in detail. Chen [23] performed a quantitative analysis of urban hotspots but she only considered the travel volume and travel frequency in the hotspots and these two indicators were relatively simple. In order to improve on these studies, we redefine the concept of an attractiveness index for a hotspot and use multiple quantitative indicators to describe the attractiveness degree of hotspots to residents. In addition, we develop an algorithm to compute the attractiveness index of the hotspots. The calculation of the attractiveness index requires the comparison and analysis of the relationships among multiple hotspots to identify similar regions and merge them. However, the commonly used manual summarization of the hotspots is often inefficient. Therefore, a hotspot similarity algorithm is proposed and implemented in this study to improve the efficiency of summarizing the hotspots.

III. DEFINITIONS AND MODELS OF THE ATTRACTIVENESS INDEX FOR URBAN HOTSPOTS

The formation of hotspots is strongly correlated with some features and there are also various differences between hotspots. Therefore, the concept of the hotspot attractiveness index is proposed to quantitatively analyze the relationship between the hotspots and the related factors. The basic components and the models of the attractiveness index are given below.

A. DEFINITIONS

Definition 1 [Hotspot(HS)]: An HS refers to an region where the travel volume is high and which is visited frequently.

In this study, the HS represents the results of the clustering mining.

Definition 2 [Hotspots Attractiveness Index(HSAI)]: This assessment index describes the hotspot's attractiveness to residents and it is comprised of four indicators including travel volume, time distribution, travel distance, and the density of the points of interest (POI).

Definition 3 [Residents Travel Volume Index (RTVI)]: This index denotes the residents' travel volume from and to the hotspots. In this study, the number of pick-up/drop-off points in a certain region is used to represent the travel volume.

Definition 4 [Time Distribution Index (TDI)]: This index represents the frequency of occurrence of hot spots under given time period. The time period can be hour or day.

Definition 5 [Residents Travel Distance Index (RTDSI)]: The travel distance index includes the distance that the residents travel when entering and leaving the hotspots.

Definition 6 [Hotspot POIs Index (HSPOI)]: A POI is a venue such as a shopping mall or school with a name, address, coordinates, category, and other attributes. The hotspot's POI index is used to evaluate the POI density for each category in a particular region, that is, the infrastructure configuration in this region.

B. MODELS OF THE ATTRACTIVENESS INDEX

The *HSAI* is the sum of the four indicators (*RTVI*, *TDI*, *RTDSI*, *HSPOI*) after normalization, the range of the indicators is (0,1). In this paper, we believe that these four indicators are equally important, so we do not consider their weights, in other words, their weights are all 0.25. The equation for the *HSAI* is:

$$HSAI = RTVI + TDI + RTDSI + HSPOI \tag{1}$$

The *RTVI* is defined by Eq. (2) and Eq. (3):

$$HSTD_{t1} = N_{up}^{t1} + N_{down}^{t1} \tag{2}$$

$$RTVI = \frac{2}{1 + e^{-\alpha \times HSTD}} - 1 \tag{3}$$

In Eq. (2), $HSTD_{t1}$ represents the travel volume of the hotspot in time period $t1$, N_{up}^{t1} and N_{down}^{t1} represent the number of pick-up and drop-off points in $t1$ respectively. The travel volume in different time periods is added to obtain the total travel volume in this area for one day. In Eq. (3), α is a parameter to be determined and the calculation of α will be described in the experimental section. The range of *RTVI* is (0,1). The closer it is to 1, the higher the traffic volume of the hotspot is.

The *TDI* is defined by Eqs. (4) and (5):

$$TDI = \frac{\sum_{i=1}^{m} \sqrt{\rho_i} - 1}{\sqrt{m} - 1} \tag{4}$$

$$\rho_i = \frac{HSTD_i}{\sum_{i=1}^{m} HSTD_i} \tag{5}$$

In Eq. (4), a day is divided into 13 time periods, m is equal to 13, and the expression $\sum_{i=1}^{m} \sqrt{\rho_i}$ represents the sparseness of the time period. The range of *TDI* is also (0,1). The closer the value is to 1, the more even the distribution of the time periods, which means that a hotspot can be formed in multiple time periods of a day; the closer the value is to 0, the sparsity of time distribution is, and a hotspot is appeared in only 1 or 2 time periods.

TABLE 1. POI category taxonomy.

Code	Category	Code	Category
1	entertainment,sports and theaters	9	banking and insurance service
2	shopping mall	10	public utilities
3	restaurant	11	science and education
4	car service/car sales/car repair	12	corporate business
5	hotel	13	transportation facilities
6	living service	14	residence
7	hospital	15	governmental agencies and public organizations
8	scenic spot		

The *RTDSI* is defined by Eqs. (6) and (7):

$$HSDS_{t1} = \bar{DS}_{arr}^{t1} + \bar{DS}_{dep}^{t1} \tag{6}$$

$$RTDSI = \frac{2}{1 + e^{-\beta \times HSDS}} - 1 \tag{7}$$

In Eq. (6), $HSDS_{t1}$ represents the travel distance to and from the hotspot in time period $t1$, $\bar{DS}_{arr}^{t1}$ is the average travel distance when entering the area in $t1$, and $\bar{DS}_{dep}^{t1}$ is the average travel distance when departing from this area in $t1$. The travel distance in the different time periods is added to obtain the total travel distance in this area for one day. In Eq. (7), β is also a parameter to be determined. The range of *RTDSI* is (0,1). The closer the value is to 1, the more the hotspot attracts visitors who live far away; otherwise, it indicates that this hotspot only attracts nearby visitors.

In order to obtain the POI distribution in all hotspots, we divide the POI into 15 categories, as shown in Table 1.

Then, we compute an average POI density in a hotspot, where the density d_j of the jth POI category in hotspot r_i is calculated using Eq. (8) and the *HSPOI* is calculated using Eq. (9):

$$d_i = \frac{Number\ of\ POIs\ of\ the\ jth\ POI\ category}{Area\ of\ hotspots\ r_i} \tag{8}$$

$$HSPOI = \frac{\sum_{j=1}^{15} d_j}{15} \tag{9}$$

The value range of *HSPOI* is (0,1). A value closer to 1 indicates better infrastructure in the hotspot and more urban functions for the residents; a value closer to 0 indicates that the hotspot provides only a few urban functions, such as airport or bus terminals or that there is less infrastructure.

IV. SIMILARITY CALCULATION OF URBAN HOTSPOTS

After using the clustering algorithm to mine the GPS trajectory data, we obtain multiple clusters but each cluster does not represent one hotspot because several clusters may exist in close proximity and they only occur during different

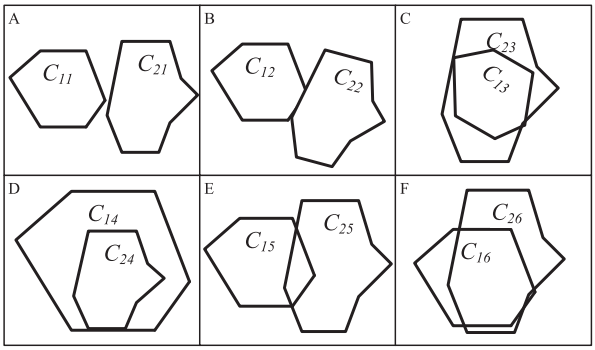

FIGURE 1. The spatial relationship between two polygons.

time periods. Therefore, we need to merge the clusters with similar locations into one hotspot and those dissimilarity clusters are identified as the new hotspot.

A. ANALYSIS OF HOTSPOTS

On the map, the hotspot is depicted as a polygon that encompasses multiple pick-up/drop-off points. The spatial relationship between any two polygons can be described as separation, containing, intersecting, or touching [24]. Figure 1 shows these four positional relationships. In Figure 1, a 2×3 grid is shown and each grid represents the positional relationship of a polygon pair; therefore, there are six cases in total. C_{ij} represents the j clustering result of the polygon in i time period. Grid A indicates that C_{11} and C_{21} are separated and they form two hotspots. Grid B indicates that C_{12} and C_{22} are touching and they are still treated as two hotspots. Grid C shows that C_{23} contains C_{13}; therefore, the larger portion C_{23} is used to represents the hotspot. Grid D is exactly the reverse of the C grid but the larger result C_{14} is still used to represent the hotspot. Grid E shows that the two polygons C_{15} and C_{25} intersect but their intersection is less than a given threshold θ_s; therefore, they are treated as two hotspots. Grid F indicates that C_{16} and C_{26} intersect and the intersection is larger than the threshold θ_s; therefore, their union is treated as a hotspot.

In this study, when determining if two polygons are similar, the main point of interest is whether they overlap. The hotspot consists of multiple point sets; therefore, is inconvenient to determine the overlapping relationship manually. In order to solve this problem, we use a bounding box to replace the points. A bounding box is a relatively simple closed space that contains all the points, objects, or a group of objects. A bounding box is commonly used to expedite specific testing processes [25]. There are five types of bounding boxes, i.e., a surrounding sphere (SS), an axis-aligned bounding box (AABB), an oriented bounding box (OBB), a fixed-direction hull (FDH), and a convex hull (CH) [26].

In Figure 2, the AABB refers to a box whose axis is parallel to the coordinate axis. It is the rectangle formed by selecting the maximum and minimum horizontal and vertical coordinates in each vertex of the two-dimensional shape and

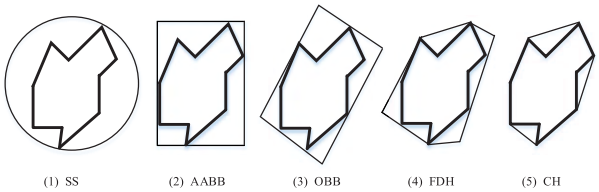

FIGURE 2. The types of bounding box.

is one of the most commonly used bounding box types. The axis of the OOB can have a random direction and it is an improvement of the AABB. The FDH is a fixed-direction bounding box. The CH tightly surrounds the object and it has the smallest area of the bounding boxes. In order to efficiently describe the polygonal shape, the CH is used to represent the hotspot. At present, there are many algorithms for computing CH including incremental method, Graham scan method, gift-wrap-ping method and divide-and-conquer method. Among them, the Graham algorithm is more commonly used.

After all the CHs and their areas are calculated, we use the overlapping relationship to reflect the similarity of the CHs and the area threshold θ_s is introduced here. Assuming that the areas of two CHs P and Q are SP and SQ respectively and their intersection area is SI, then θ_s can be specified as two-thirds of the maximum value of the two CH areas. Due to space limitations, we don't discuss the computation process of the parameter θ_s, and its value came from the experimental results. If $SI \geq \theta_s$, then P and Q are similar and the union of P and Q forms a new hotspot. Otherwise, P and Q are dissimilar and P and Q each form one hotspot. In addition to the separation and intersection relationship, two CHs can touch or one can contain the other. The external touching relationship is treated as a special separation relationship in this study and they are judged as dissimilar. If all the vertices of one CH are inside of another CH and their edges are not intersected, they belong to the containing relationship and are similar.

B. ALGORITHM OF HOTSPOT SIMILARITY

Using the above-mentioned definitions, a hotspot similarity algorithm (HSSA) is proposed. Its basic process is as follows:

(1) The Graham algorithm [28] is used to calculate the CHs resulting from the clusters.

(2) The minimum area of each CH is calculated using the algorithm described in [27].

(3) The spatial relationship between any two CHs is evaluated to determine their similarity. The simple method of calculating the distance between the centroids of two CHs can be used to determine the positional relationship. If this distance is greater than or equal to the distance threshold θ_d, then the CHs are separated and the similarity is 0. If the distance is less than θ_d, the CHs may have touching, containing, or intersecting relationships. First, for the containing relationship, if one CH contains another, the similarity is 1.

TABLE 2. Notations for the algorithm of hotspots similarity.

Notation	Description
C_1,C_2	Two polygons generated from the clustering results, each cluster consists of many points
CP_1,CP_2	Two CHs formed by C_1 and C_2
$S_{CP_1}^{MARB}$, $S_{CP_2}^{MARB}$	The minimum areas of these two CHs
$d(CP_1, CP_2)$	The centroid distance of C_1 and C_2
I, SI	The intersection and their area of C_1 and C_2
θ_d, θ_s	Distance threshold, area threshold
$flag$	The comparison results of two CHs, 0 for separation, 1 for inclusion, 2 for intersection and similarity,3 for intersection and not similarity

Algorithm 1 Algorithm of Hotspots Similarity

Require: C_1 and C_2
Ensure: Similarity comparison result of C_1 and C_2

Compute CP_1, CP_2
Compute $S_{CP_1}^{MABR}$ and $S_{CP_2}^{MABR}$
Compute $d(CP_1, CP_2)$
if $d(CP_1, CP_2) \geq \theta_d$
  $flag = 0$; return $flag$
else
  if $CP_1 \subseteq CP_2$ *or* $CP_2 \supseteq CP_1$
    $flag = 1$; return $flag$
  else
    Compute I
    Compute SI
    θ_s=2/3 $\times$ $max(S_{CP1}^{MABR}, S_{CP2}^{MABR})$
    if $SI \geq \theta_s$
      $flag = 2$; return $flag$
    else
      $flag = 3$; return $flag$
    endif
  endif
endif

Touching is treated as a special case of an intersection. Next, for the intersection, the intersection area and area threshold θ_s of the intersecting CHs are calculated. If the intersection area is greater than or equal to θ_s, the two CHs are similar and the return value is 1; otherwise, the two CHs are dissimilar and the return value is 0. Two methods are used to determine the distance threshold θ_d.

Method 1: the value θ_d is manually determined based on experience; for example, 10 km.

Method 2: first, the centroid distance between any two hotspots is calculated; second, we choose the minimum distance and maximum distance from all distances and calculate the mean value according these two values, then θ_d is set to the mean value.

We list the notations for describing the HSSA in Table 2.

The pseudocode for the HSSA is is given in Algorithm 1.

The Graham algorithm is used to calculate the CHs of the points. According to the relevant proofs [26], the time complexity of the Graham algorithm is O($n$log$n$) for polygons with n vertices. Assuming that the number of edges of the polygon is n and the number of CH edges is m (where $m \leq n$), the time complexity of the minimum area of the CH is O(m^2) [29]. A new algorithm based on plane scan lines and intersection subdivision is used to determine whether two CHs intersect [30]. Assuming that the total number of edges of all CHs that are compared is m and the total number of intersecting edges of all CHs is k, then the time for this algorithm to complete the entire traversal is O($(m + k)$log$(m + k)$). Since $k \leq m^2$, the total time consumption for determining whether several CHs intersect is O$(m + k)$log(m). In summary, the time complexity of the HSSA is O($n$log$n$)+O(m^2)+O$(m+k)$log(m), where n, m and k represent the vertices of the polygons to be calculated, the number of edges of the CHs created by the polygons, and the total number of intersecting edges of all CHs respectively, and their relationship is $n \gg m$, $m^2 \geq k$. Therefore, the most time-consuming calculation of the HSSA lies in the calculation of the CHs. Finally, the time complexity of the HSSA is O($n$log$n$).

TABLE 3. Notations for the algorithm of hotspots attractiveness index.

Notation	Description
D_{ij}	Results of Clustering, i represents day and j represents time period
HS	The set of hotspot
$Index$	The set of attractiveness index for hotspots
$HSAI$	The set of indices for hotspots
FD	Travel volume of the hotspot
TD	Travel time period of the hotspot
DIS	Travel distance of the hotspot
$flag1$, $flag2$	The results of intersection and similarity comparison of CHs

V. THE ALGORITHM OF ATTRACTIVENESS INDEX FOR HOTSPOTS

This section introduces the algorithm of attractiveness index for hotspots (HSAIA). The principle of the HSAIA is as follows: first, we analyze and compare each generated cluster for the different time periods. Second, we merge similar clusters into a hotspots or identify independent clusters as new hotspots. Third, we calculate the four evaluation indicators of each hotspot. Finally, the *HSAI* for each hotspot and the combined hotspots are obtained. We list the notations used in this algorithm in Table 3.

The pseudocode of the HSAIA algorithm is given in Algorithm 2.

VI. EXPERIMENT AND ANALYSIS

In order to validate the effectiveness of the attractiveness index, the GPS trajectory data provided by the Kunming Taxi Vehicle Administration Office is used as a validation dataset. This dataset contains 55,851,192 GPS records generated by 6,599 taxis within one week in August 2012 (Aug. 13-19) [11]. After performing data preprocessing (such as data cleaning, map matching) on the original GPS trajectory dataset, the DBSCAN algorithm [31] is used to mine the

Algorithm 2 Algorithm of Hotspots Attractiveness Index

Require: $D = \{D_{11}, D_{12}, \cdots, D_{ij}\}$
Ensure: HS, $Index$

initial $m = 1, HS = \{\}, HSAI = \{\}, Index = \{\}, FD = 0, TD = 0, DIS = 0$
for $i = 1$ to 7 //days
  for $j = 1$ to 13 //time period
    for $k = 1$ to count(D_{ij})
      if $HS = \phi$ then
        $FD =$ compute_HSTD(D_{ijk}), $TD = 1$
        $DIS =$ compute_HSDIS(D_{ijk})
        $HSAI = HSAI + (FD, TD, DIS)$
        add(D_{ijk}) to HS
      else
        $A = D_{ijk}$
        $FD_A =$ compute_HSTD(HS_A^j), $TD_A = 1$
        $DIS_A =$ compute_HSDIS(HS_A^j)
        for $n = 1$ to count(HS)
          $B = HS_n$
          $flag1$=Inter(A, B) //intersection?
          if $flag1$=1 then // Yes
            $flag2$ =Sim(A, B) // Similarity?
            if $flag2 = 1$ then //Yes
              $HSAI_n.FD = HSAI_n.FD + FD_A$
              $HSAI_n.TD = HSAI_n.TD + TD_A$
              $HSAI_n.DIS = HSAI_n.DIS + DIS_A$
            end // end flag2
          end // end flag1
          add(A) to HS
          $TD_A = 1$
          add(FD_A, TD_A, DIS_A) to $HSAI$
        end
      end
    end
  end
end // end i
for $i = 1$ to count($HSAI$)
  $RTVI_i =$ compute_RTVI(HS_i)
  $TDI_i =$ compute_TDI(HS_i)
  $RTDSI_i =$ compute_RTDSI(HS_i)
  $HSPOI_i =$ ccompute_HSPOI(HS_i)
  $Index_i = RTVI_i + TDI_i + RTDSI_i + HSPOI_i$
  $Index = Index + Index_i$
end
return HS, $Index$

data for hotspots. DBSCAN algorithm is a classical clustering algorithm, many researchers use it or its improved algorithm to discover urban hot regions. After many experiments, it is concluded that the parameters suitable for this algorithm are $Eps = 360$ m and $Minpts = 150$. The number of clusters formed by passengers that are picked up/dropped off within a week is 905 and includes 3,128,289,000 GPS points. Although the number of clusters formed in a week is 905, the number of hotspots in a city may be actually less because many areas are similar and can be merged into the same hotspot.

FIGURE 3. Clustering results of hotspot (ID = 16).

A. SIMILARITY RESULTS OF HOTSPOTS

The similarity calculation of the 905 clustering results is performed in accordance with the HSSA, that is, the clustering results based on the pick-up/drop-off data for one week are calculated and 37 urban hotspots are finally obtained. The hotspots' similarity results are compared with the results of manual summarization, and their numbers and locations are basically the same. These hotspots vary widely in terms of time period and travel volume. The hotspots may form in several time periods in one week in some areas, whereas some areas only occasionally become hotspots. In order to display the similarity results of the hotspots, the clustering results of the pick-up points in four different time periods on August 19 are shown in Figure 3 for visual analysis and different colors are used. It is evident that the hotspots for the four clustering results are spatially very similar; therefore, they should belong to the same hotspot. Therefore, the results of the algorithm are consistent with our visual interpretation and the results are accurate and credible.

B. ATTRACTIVENESS INDEX OF THE HOTSPOTS

After the 37 hotspots are obtained, the HSAIA is used to calculate the four evaluation indicators for each hotspot. The parameter α is important in the calculation of the *RTVI*. The parameter α represents the mean value of the *RTVI* in the range of (0,1), that is, the greater the travel volume of the hotspot, the closer the value of α is to 1; otherwise, the value of α is closer to 0. According to Eq. (3) we can deduce Eq. (9) for calculating α, which is as follows:

$$\alpha = -\frac{\log\left(\frac{2}{RTVI+1} - 1\right)}{HSTD} \tag{10}$$

It can be seen from Eq. (10) that the parameter α is jointly determined by the *HSTD* and the *RTVI*. First, the *HSTD* of each hotspot is calculated and then the *HSTD* values are ranked from largest to smallest to determine the median and mean values of the travel volume. At this point, the corresponding *RTVI* value should be 0.5. Then, the parameters $\alpha 1$

TABLE 4. Attractiveness indices of 10 hotspots.

ID	*HSTD*	*RTVI*	*Freq*	*TDI*	*RTDSI*	*HSPOI*	*INDEX*
1	18433	0.8393	46	0.3237	0.8688	0.3029	2.3347
2	67216	0.9997	140	0.9999	0.9851	0.0534	3.0382
3	31061	0.9676	61	0.4317	0.7949	0.2720	2.4662
4	32252	0.9723	92	0.6475	0.8396	0.2122	2.6716
5	17112	0.8115	86	0.6115	0.9930	0.0083	2.4243
6	6132	0.3846	30	0.2086	0.7356	0.0047	1.3335
7	5254	0.3341	30	0.2086	0.6566	0.0012	1.2005
8	58054	0.9991	76	0.5396	0.9042	0.3449	2.7878
9	6298	0.3939	29	0.2014	0.5268	0.2017	1.3239
10	13115	0.6999	37	0.2590	0.5096	0.3960	1.8645

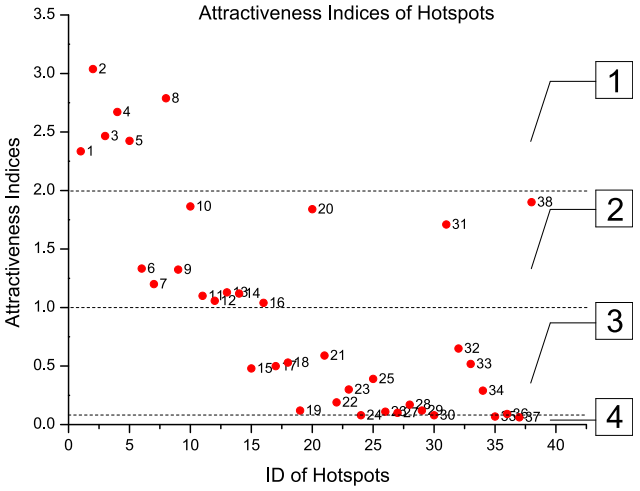

FIGURE 4. Scatter plot of attractiveness indices for hotspots.

and $\alpha 2$ are calculated and obtained based on the median and mean values respectively. Finally, $\alpha 1$ and $\alpha 2$ are used in Eq. (3) and the *RTVI* of the hotspots is calculated. According to the experimental results, $\alpha 1$ has a better effect than $\alpha 2$ because the travel volume index is evenly distributed between 0 and 1. Therefore, the value of $\alpha 1$ is set to 1.3565e-04.

For the *RTDSI*, we use a similar method to calculate the parameter β and its value is 0.0073.

After determining the values of α and β, the *RTVI* and *RTDSI* for all hotspots can be calculated. Then, their *TDI* and *HSPOI* are obtained using Eqs. (4),(5),(8), and (9). The attractiveness indices of 10 hotspots are shown in Table 4. We find that the difference between the hotspots is very large. The hotspots with a high travel volume have a high visiting frequency and large travel distance and the time periods of the hotspot formation exhibit high regularity. The hotspots with a low travel volume have a low visiting frequency and small travel distance and no regularity in the time periods of the hotspot formation is observed. However, the *HSPOIs* of these hotspots are not high. After performing the quantitative analysis of all hotspots, we use a scatter plot to depict them, as shown in Figure 4.

In Figure 4, the X-axis represents the ID of hotspots and the Y-axis represents the attractiveness index. Three dotted lines are plotted on the 0.1, 1, and 2 intervals on the Y-axis

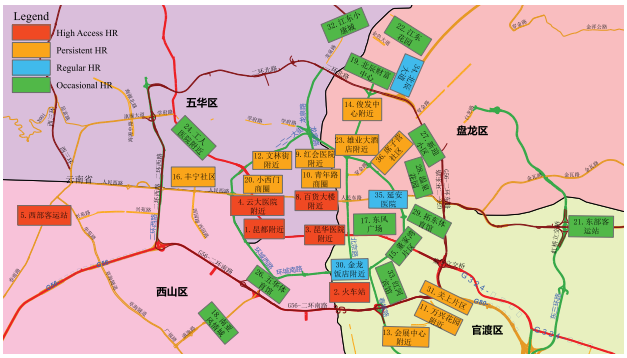

FIGURE 5. Visualization of traveling hotspots in kunming.

and the attractiveness index is divided into four categories. The first category contains 6 hotspots, which are called high-access hotspots. Their attractiveness index is between 2 and 3 and they have the highest travel volume, the highest travel frequency, the longest travel distance, and the most POI. The second category contains 11 hotspots, which are called persistent hotspots. Their attractiveness index is between 1 and 2, and their travel volume, travel frequency, travel distance, and POI are lower than those in the high-access hotspots. The third category contains 16 hotspots and we call them regular hotspots; they have attractiveness indices between 0.1 and 1. Their travel volume and travel frequency are low and the travel distance is short. Moreover, there are few POI and most hotspots belong to this category. The fourth category contains only 4 hotspots and they are called the occasional hotspots. Their attractiveness index is less than 0.1. These hotspots form temporarily, they only appear 1 or 2 times a week, and they do not follow any spatio-temporal rules.

C. VISUAL ANALYSIS AND EVALUATION OF THE HSAI

1) VISUAL ANALYSIS OF HOTSPOTS

This subsection describes the visual evaluations of the hotspots. Kunming is divided into four administrative districts (Wuhua District, Panlong District, Xishan District, and Guangdu District). Four colors are used to show the hotspot categories. Red represents the high-access hotspots, yellow for the persistent hotspots, green for the regular hotspots, and blue for the occasional hotspots. Since there are many hotspots and their locations are very dispersed, some common hotspots outside the Third Ring Road, such as the airport and the two bus terminals (North bus terminal and South bus terminal) are not shown on the map for clarity. The remaining hotspots are shown in Figure 5.

Several travel rules of the urban residents can be deduced from Figure 5:

(1). The number of hotspots in the four categories is 6, 11, 16, and 4, indicating that the number of regular hotspots is largest. Regular hotspots are often located around commercial centers and mostly appear after the rush hour. This shows that residents prefer to choose commercial zone near their

TABLE 5. Error evaluation result.

Methods	RMSE	MAE
RTVI+TDI	2.860	1.808
RTVI+TDI+ RTDSI	2.296	1.455
RTVI+TDI+ HSPOI	1.414	0.909
RTVI+TDI+ RTDSI+ HSPOI	1.348	0.727

workplace or residential areas when considering shopping and entertainment after work.

(2). More than 2/3 of the high-access hotspots and nearly half of the persistent hotspots are on First Ring Road, which is located in the Wuhua District and near the traditional shopping malls, supermarkets, universities, and parks. These areas represent the preferred destination of residents' travel. Therefore, these locations are prone to traffic jams and traffic dispersion is needed.

(3). More than half of the regular hotspots are on First Ring Road and between First Ring Road and Second Ring Road. These are mainly residential areas and areas where schools and hospitals are located.

(4). There are a few regular hotspots between Second Ring Road and Third Ring Road. They are mainly commercial centers that are surrounded by densely populated residential areas.

(5). The formation of the hotspots is significantly associated with the scale, abundance, and functionality of the infrastructure. Therefore, they are basically distributed within Third Ring Road.

2) EVALUATION OF ATTRACTIVENESS INDEX

In order to verify the result of attractiveness index, 11 hotspots located on Second Ring Road were selected for the evaluation, including 5 business circles, 4 hospitals, and 2 large residential areas. Our evaluation indicators are compared with 2 evaluation indicators from [12]. We perform a field study in the 11 hotspots in Kunming and analyze the actual traffic patterns via video capture and proof our results by [32]–[34]. We then rank these hotspots based on the actual trips and the results of our method and measure the difference in the two ranks using Eqs. (11) and (12):

$$RMSE = \sqrt{\frac{\Sigma_i (y_i - \widehat{y_i})^2}{n}} \tag{11}$$

$$MAE = \frac{\Sigma_i |y_i - \widehat{y_i}|}{n} \tag{12}$$

The results are shown in Table 5. The performance is measured by two metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), where $\widehat{y_i}$ is an inference and y_i is the ground truth; n is the number of instances. Table 5 shows that the ranking result using *RTVI* and *TDI* only has the largest error, the ranking result of the three indices has the larger error, but the ranking error of the four indices is the smallest. Our method is superior to the other methods.

Next, we use 5 business circles as an example to illustrate their differences in the attractiveness index. They are located

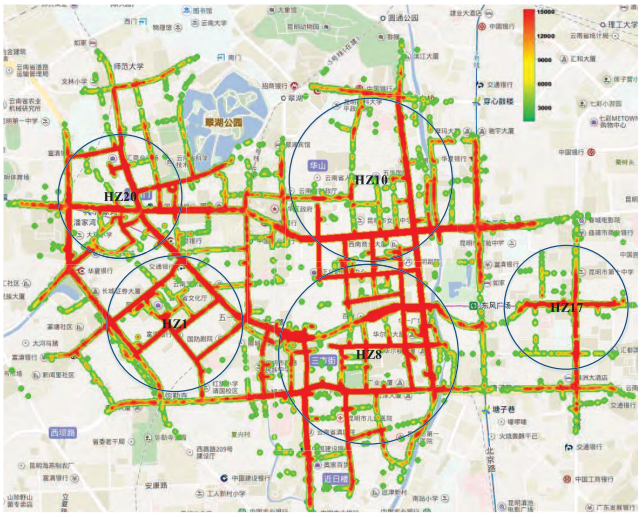

FIGURE 6. Heat map of 5 business circles (clustering results) on weekdays in kunming.

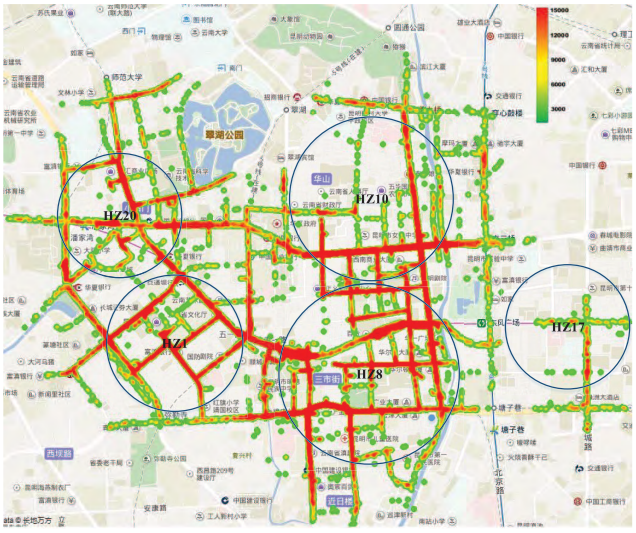

FIGURE 7. Heat map of 5 business circles (clustering results) on weekends in kunming.

TABLE 6. Attractiveness indices of 5 business circles.

ID	*RTVI*	*TDI*	*RTDSI*	*HSPOI*	*INDEX*	District
HS1	0.8393	0.3237	0.8688	0.3029	2.3347	Wuhua
HS8	0.9991	0.5396	0.9042	0.3449	2.7878	Wuhua
HS10	0.6999	0.2590	0.5096	0.3960	1.8645	Wuhua
HS17	0.1543	0.0863	0.2316	0.0649	0.5370	Panlong
HS20	0.1853	0.0360	0.5998	1	1.8212	Wuhua

on First Ring Road; 4 are located in the Wuhua district and 1 is located in the Panlong area. All of them are core commercial centers. Their heat maps on weekdays and weekends are shown in Figures 6-7.

The results show that the 5 business circles attract many urban residents and the travel volume is large regardless of the day of the week. However, there are some differences between them. The indices are higher for HS1 and

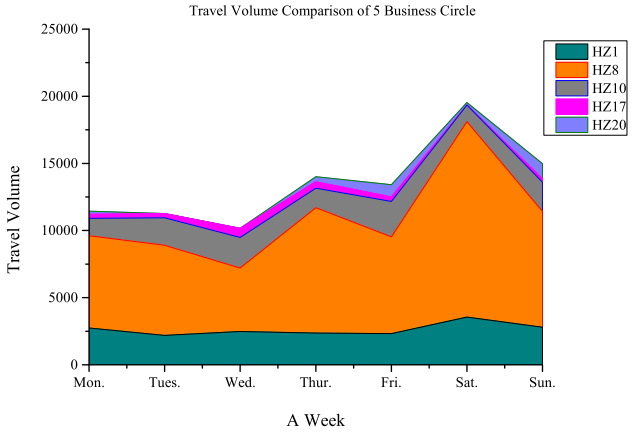

FIGURE 8. Travel volume comparison of 5 business circle.

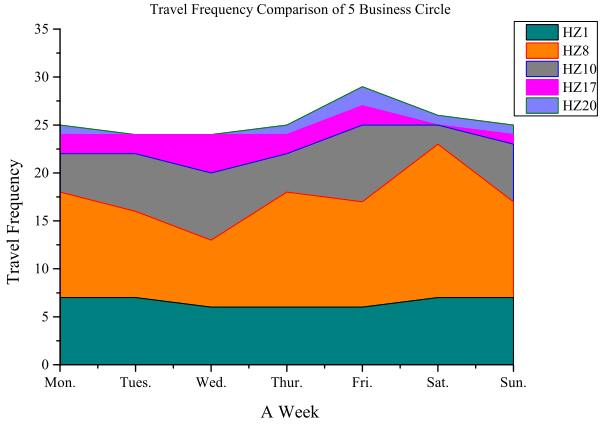

FIGURE 9. Travel frequency comparison of 5 business circle.

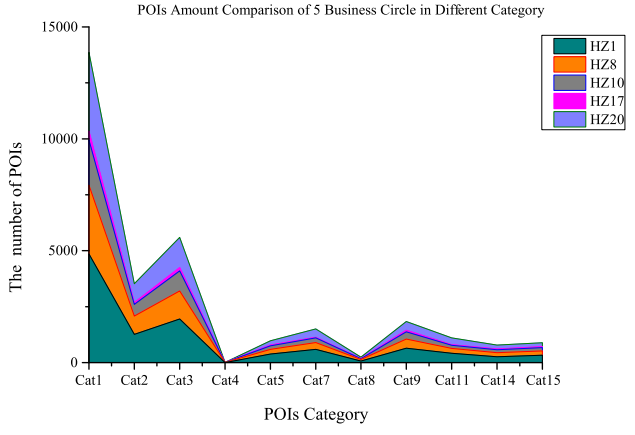

FIGURE 10. POIs amount comparison of 5 business circle in different category.

HS8 than for the other 3 hotspots and HS17 has the lowest indices. A comparison of the evaluation indices is shown in Table 6 and Figures 8-11.

If the attractiveness index is not used, no differences are observed between the areas. However, when the travel volume, travel frequency, average travel distance in one week, and the number of POI are taken into consideration, there are

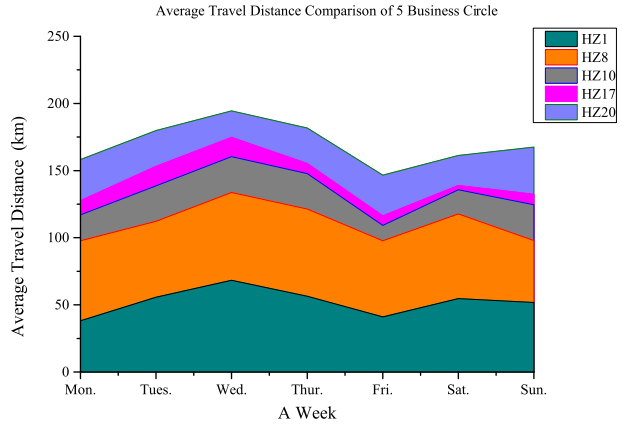

FIGURE 11. Average travel distance comparison of 5 business circle.

apparent differences, which is in agreement with the visual analysis. HS1 and HS8 are high-access hotspots, HS10 and HS20 are persistent hotspots, and HS17 is a regular hotspot. Therefore, these 5 business circles have different attractiveness values and different business values.

VII. CONCLUSION

In the intelligent transportation field, determining urban hotspots is an important research topic. However, there are few quantitative analysis methods for evaluating them. Therefore, a novel evaluation indicator, the attractiveness index, is proposed to describe the degree of correlation between the hotspots and the time period, travel frequency, travel volume, travel distance, and POI. In this study, the concept, mathematical models, parameter generation, and the algorithms of the attractiveness index are described in detail. Actual GPS trajectories are used to validate the effectiveness of the attractiveness index. The urban hotspots are categorized into high-access hotspots, persistent hotspots, regular hotspots, and occasional hotspots based on the attractiveness index values and they are visualized on a map. The proposed attractiveness index model can also be used for the creation of hotspots using bus, subway, or mobile location data. Therefore, it has a wide range of applications. In this study, the attractiveness index was limited to data acquired in one week; in future studies, a larger dataset will be used to validate the effectiveness of this method.

REFERENCES

[1] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proc. Knowl. Discovery Data Mining*, 2012, pp. 186–194.
[2] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.
[3] W. Vandenberghe, E. Vanhauwaert, S. Verbrugge, I. Moerman, and P. Demeester, "Feasibility of expanding traffic monitoring systems with floating car data technology," *IET Intell. Transp. Syst.*, vol. 6, no. 4, pp. 347–354, Dec. 2012.
[4] D. Liu, S. Cheng, and Y. Yang, "Density peaks clustering approach for discovering demand hot spots in city-scale taxi fleet dataset," in *Proc. Int. Conf. Intell. Transp. Syst.*, Las Palmas, Spain, Sep. 2015, pp. 1831–1836.
[5] G. Pan, G. Qi, Z. Wu, D. Zhang, and S. Li, "Land-use classification using taxi GPS traces," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 113–123, Mar. 2013.

[6] C. Chen, D. Zhang, N. Li, and Z.-H. Zhou, "B-planner: Planning bidirectional night bus routes using large-scale taxi GPS traces," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 4, pp. 1451–1465, Aug. 2014.

[7] B. Li *et al.*, "Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset," in *Proc. Int. Conf. Pervasive Comput. Commun.*, 2011, pp. 63–68.

[8] H. Klessig, V. Suryaprakash, O. Blume, A. Fehske, and G. Fettweis, "A framework enabling spatial analysis of mobile traffic hot spots," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 537–540, Oct. 2014.

[9] T. Louail *et al.*, "From mobile phone data to the spatial structure of cities," *Sci. Rep.*, vol. 4, no. 1, p. 5276, 2015.

[10] Q. Hu and Y. Zhang, "An effective selecting approach for social media big data analysis—Taking commercial hotspot exploration with Weibo check-in data as an example," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 537–540, Mar. 2014.

[11] X. Wang, H. Zhang, L. Wang, and Z. Ning, "A demand-supply oriented taxi recommendation system for vehicular social networks," *IEEE Access*, vol. 6, pp. 863–868, 2018.

[12] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proc. Int. Conf. World Wide Web*, Madrid, Spain, Apr. 2009, pp. 791–800.

[13] S. Kisilevich, F. Mansmann, and D. Keim, "P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos," in *Proc. Int. Conf. Exhib. Comput. Geospatial Res. Appl.*, Washington, DC, USA, Jun. 2010, p. 38.

[14] Z. Gui, Y. Xiang, and Y. Li, "Parallel discovering of city hot spot based on taxi trajectories," *J. Huazhong Univ. Sci. Technol.*, vol. 40, no. Sup. I, pp. 187–190, 2012.

[15] Z. Chen, Z. Xie, and X. Feng, "Image-based clustering analysis of massive vehicle networking data," in *Proc. Int. Conf. Wireless Commun. Signal Process.*, Yangzhou, China, Oct. 2016, pp. 1–5.

[16] H. Rong *et al.*, "Mining efficient taxi operation strategies from large scale geo-location data," *IEEE Access*, vol. 5, pp. 25623–25634, 2017.

[17] Q. Li *et al.*, *Traffic Geographic Information System Technology and Forward Development*, 1st ed. Beijing, China: Beijing Press, 2012.

[18] W. Qi, "Analyzing spatial characteristics of taxi pick-up with GPS data," M.S. thesis, Dept. Automot. Eng., Jilin Univ., Changchun, China, 2013.

[19] X. Cao, G. Cong, and C. S. Jensen, "Mining significant semantic locations from GPS data," in *Proc. Int. Conf. Very Large Data Bases*, 2010, pp. 1009–1020.

[20] L. Moreira-Matias, R. Fernandes, J. Gama, M. Ferreira, J. M. Moreira, and L. Damas, "On recommending urban hotspots to find our next passenger," in *Proc. Int. Workshop Ubiquitous Data Mining Articial Intell.*, Beijing, China, Aug. 2013, pp. 17–23.

[21] R. W. Scholz and Y. Lu, "Detection of dynamic activity patterns at a collective level from large-volume trajectory data," *Int. J. Geograph. Inf. Sci.*, vol. 28, no. 5, pp. 946–963, 2014.

[22] P. Zhao, "Research on the method of extracting and analyzing urban-hotspots based on trajectory clustering," M.S. thesis, Dept. Urban Planning, Wuhan Univ., Hubei, China, 2015.

[23] H. Chen, "Mining of urban hot region and attractiveness analysis based on GPS trajectory data," M.S. thesis, Dept. Softw. Eng., Yunnan Univ., Kunming, China, 2016.

[24] F. Martinez, A. Rueda, and F. R. Feito, "A new algorithm for computing boolean operations on polygons," *Comput. Geosci.*, vol. 35, no. 6, pp. 1177–1185, 2009.

[25] Y. He, *Geometric Calculation*, 1nd ed. Beijing, China: Higher Education Press, 2013.

[26] K. M. Babu and M. V. Raghunadh, "Vehicle number plate detection and recognition using bounding box method," in *Proc. Int. Conf. Adv. Commun. Control Comput. Technol.*, Tamil Nadu, India, May 2016, pp. 106–110.

[27] P. Cheng, H. Yan, and H. Zhen, "An algorithm for computing the minimum area bounding rectangle of an arbitrary polygon," *J. Eng. Graph.*, vol. 4, no. 1, pp. 122–126, 2008.

[28] R. Liu, *Algorithm Art and Informatics Olympiad: Algorithm Competition Classic*, 1st ed. Beijing, China: Tsinghua Univ. Press, 2009.

[29] S. M. F. Preparata, *Computational Geometry: An Introduction*, 1st ed. New York, NY, USA: Springer, 1985.

[30] J. Huang, S.-J. Yan, X.-L. Zhu, and J.-W. Li, "Algorithm for intersection and union between convex polygons," *J. Eng. Graph.*, vol. 4, no. 1, pp. 122–126, 2008.

[31] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Knowl. Discovery Data Mining*, vol. 96, no. 34, pp. 226–231, 1996.

[32] T. Dui, "The rise of the business circle of Renmin road in Kunming, rebirth of urban pattern," *China Real Estate News*, pp. 1–2, 2013.

[33] Z. Cha, *Where Will the City Business Circle go in Kunming?* Yunnan Daily, 2016, p. A1.

[34] *2010 Research Report of Business in Downtown Kunming*. Accessed: Dec. 1, 2017. [Online]. Available: http://doc.mbalib.com/view/ad11a8a37de7cc076e9fc16494700986.html/

LI CAI was born in Kunming, China, in 1975. She received the M.S. degree in computer application from Yunnan University, China, in 2007. She is currently pursuing the Ph.D. degree with the School of Computer Science, Fudan University, China.

From 1997 to 2002, she was a Research Assistant with Network Center. Since 2010, she has been an Associate Professor with the School of Software, Yunnan University. Her research interests include intelligent transportation, machine learning, visualization, and data quality.

FANG JIANG was born in Jiujiang, China, in 1993. She received the B.S. degree in digital media technology from GanNan Normal University, Gannan, China, in 2017. She is currently pursuing the master's degree with Yunnan University. Her current research interest focuses on intelligent transportation, cloud computing, and mobile computing.

WEI ZHOU was born in 1974. He received the M.S. degree in computer science from the Kunming University of Science and Technology, Kunming, China, in 2002, and the Ph.D. degree in computer science from the University of Chinese Academy of Sciences in 2008. Since 2009, he has been with the Software School, Yunnan University, where he became a Full Professor in 2016. He has authored one book and over 50 articles.

His current main research interests are about distributed computing, cloud computing, and bioinformatics.

KEQIN LI (F'15) is currently a SUNY Distinguished Professor of computer science with the State University of New York. He has authored over 590 journal articles, book chapters, and refereed conference papers. His current research interests include parallel computing and high-performance computing, distributed computing, energy-efficient computing and communication, heterogeneous computing systems, cloud computing, big data computing, CPU–GPU hybrid and cooperative computing, multicore computing, storage and file systems, wireless communication networks, sensor networks, peer-to-peer file sharing systems, mobile computing, service computing, Internet of Things, and cyber-physical systems. He received several best paper awards. He is currently serving or has served on the editorial boards of the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON CLOUD COMPUTING, the IEEE TRANSACTIONS ON SERVICES COMPUTING, and the IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING.