Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Extraction and Analysis of Crowd Activity Vergence Model in Space-Time Vector Field

To cite this article: Haiyan Liu *et al* 2022 *J. Phys.: Conf. Ser.* **2294** 012031

Extraction and Analysis of Crowd Activity Vergence Model in Space-Time Vector Field

Haiyan Liu[1], Jing Li[1,*], Qiang Guo[2], Youwei Zhang[3], Chuanwei Lu[1], Fang Hu[2], Hongjian Wu[4]

[1] PLA Strategic Support Force Information Engineering University, Zhengzhou, 450052, China;

[2] Henan Big Data Center, Zhengzhou, 450052, China;

[3] Zhengzhou Xinda Institute of Advanced Technology, Zhengzhou, 450052, China;

[4] Zheng Shu Network Technology Co., Ltd, Zhengzhou, 450052, China;

* Correspondence: 351800040@qq.com

Abstract. The vergence model of crowd activity is one of the core contents of human mobility research. Traditional methods do not consider the mobility of crowd activities in terms of extracting vergence models. In this paper, the model extraction problem is transformed into a time series clustering problem, and the mobility of crowd activities is dynamically modeled by introducing vector field theory. Then, the vergence of the crowd is calculated by the divergence. Finally, a time series composed of the crowd vergence is constructed to obtain the main vergence model of crowd activity through clustering. The method proposed in this paper is experimentally verified on the Didi Chuxing data in Haikou City, and four main vergence models of the crowd activity are extracted, which proves that the method proposed in this paper is effective and provides research ideas and method support for exploring human mobility.

1. Introduction

The different functions of the spatial structure in the city make the crowd activities have a certain tendency and purpose [1]. Paying attention to the vergence models of inter-regional crowd activities can help identify core traffic corridors and hot spots, which has important research significance in traffic planning, emergency avoidance, and urban construction [2]. Studying the vergence characteristics of crowds in cities is one of the core research contents to explore the laws of crowd activities, but crowd activities are complex and difficult to be directly observed. In recent years, with the rapid development of big data technology and network sensing technology, the spatial and temporal process of crowd movement is positively or passively recorded by sensing devices, resulting in large-scale sensing data (such as taxi trajectory data, Didi travel data, social media check-in data, etc.). These perceptual data contain a large amount of explicit and implicit geographic information, which provides research materials and data support for exploring the relationship between crowd activities and urban structure.

Common crowd activity aggregation feature analysis methods include density-based methods and flow-based methods. For example, kernel density analysis is used to identify areas with high crowd density [3], or the degree of aggregation can be judged by the inflow and outflow of crowds within a certain period of time [4]. In addition, lacking consideration of the dynamic flow of crowds between regions, the methods mentioned above are all relatively static point-based methods, which cannot reflect the future trend of crowd activities.

Therefore, in this research, the vector field[5] that can be used to describe the fluidity of things is introduced to model the fluidity of crowd activities. The vergence degree of crowd activities is quantitatively calculated by the divergence operator, and finally a time series composed of crowd vergence is constructed to obtain the main crowd activity vergence models through time series clustering. Based on the data of Didi Chuxing in Haikou City, the four main crowd activity models are extracted in the experiment, and combined with the distribution of POI types. The four models are explained semantically, providing research ideas and method support for exploring human mobility.

2. Overall Research Framework

The overall research framework of this method is shown in Figure 1. Specific steps are as follows:

- Divide the original data into time units and geographic units to construct spatial and temporal units;
- The main direction of crowd activities in each spatial and temporal unit is calculated by the wind direction mean method under the constraint of angle space, and the spatial and temporal vector field of crowd activities is constructed by treating the projection value of each direction on the main direction as the main intensity;
- The vector field divergence is introduced to calculate the vergence of each spatial and temporal unit to construct the spatial and temporal sequence of crowd activities;
- Dynamic Time Warping (DTW) is used to calculate the similarity between time series, and K-means is applied to clustering the time series according to the similarity. The clustering result is the main crowd activity vergence pattern.

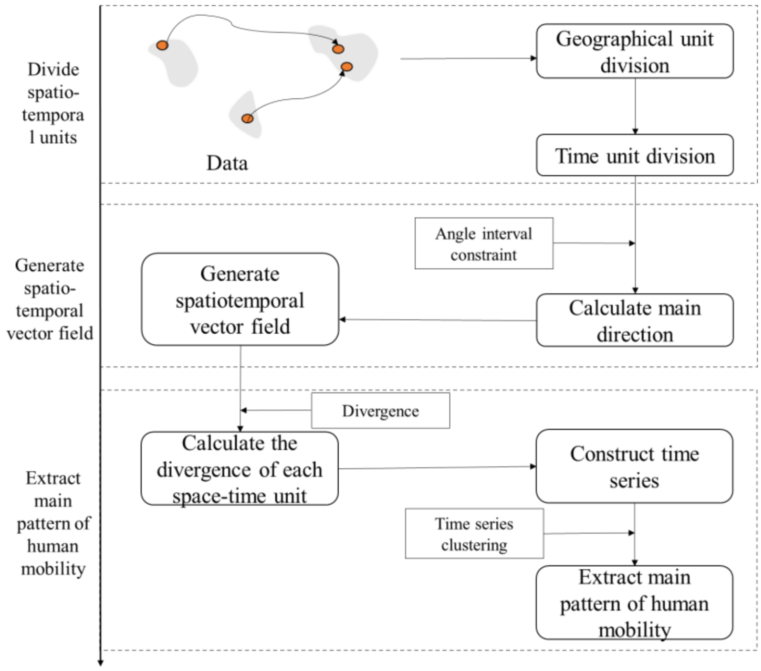


Figure 1. Overall research framework

2.1. Calculation of Main Body Direction

In this paper, the calculation method of the angle means proposed by Mardia et al. [6] is used, and the formula is as follows:

$$\beta = \arctan\left(\frac{\sum \sin \beta_i v_i}{\sum \cos \beta_i v_i}\right) \tag{1}$$

Journal of Physics: Conference Series **2294** (2022) 012031 doi:10.1088/1742-6596/2294/1/012031

Where, β_i refers to the direction angle in the angle interval *M*, v_i represents the number of people in the direction β_i. Since the direction angle corresponds to an individual in the study, it is set to be 1 for easy calculation.

2.2. Body Strength Calculation

After finding the main direction of the area, it is necessary to calculate the intensity of the crowd activity in the main direction. The crowd activity intensity is calculated by the sum of the projection values of the crowd numbers in all directions in the subject direction, as shown in the figure. The specific formula is as follows:

$$v_{main} = \sum v_i \cos \alpha_i^{main} \tag{2}$$

Among them, α_i^{main} is the angle between the direction α_i and the main direction α_{main}, and $v_i \cos \alpha_i^{main}$ refers to the projection value of the number of people in the direction α_i on the main direction γ_{main}.

2.3. The Calculation of Time Series Similarity Based on Dynamic Time Warping

The existence distance matrix for two unequal time series *a* and *b* is $D_{m\times n} = (d_{ij})_{m\times n}$, as shown in Figure 2. Among them, d_{ij} represents the distance between x_i and y_j, which is generally the square of the Euclidean distance, namely $d_{ij} = (x_i - y_j)^2$. It is assumed that the dynamic time-curved paths of *a* and *b* are $W = \{w_1, w_2, \cdots, w_k\}$, where $w_k = (d_{ij})_k$ is the k th element of the path, and *W* needs to satisfy the following conditions:

- $\max\{m, n\} \le k \le m + n - 1$;
- If the conditions $w_1 = d_{11}$ and $w_k = d_{mn}$ are satisfied, the two sequences must match at the beginning and the end, and the other parts are matched in sequence.
- If $w_k = d_{ij}$, $w_{k+1} = a_{kl}$, there will be $0 \le k - i \le 1$ and $0 \le l - i \le 1$, indicating that the dynamic time-curved path increases monotonically with time.

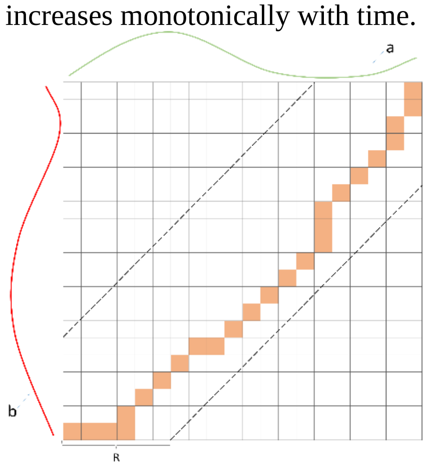


Figure 2. Distance matrix of time series

Since the dynamic time curved path that satisfies the above three conditions is not unique, the path with the smallest $\sqrt{\sum_{i=1}^{k} w_i}$ value among all the dynamic time curved paths can be selected as the optimal dynamic time curved path, and the corresponding distance is the dynamic time curved distance. The dynamic time curved distance can be expressed as:

$$DTWdistance(a,b) = \min\left(\sqrt{\sum_{i=1}^{k} w_i}\right) \tag{3}$$

3. Experiment and Analysis

3.1. Experimental Data Source and Processing

1. OD Data Set

Since the taxi is one of the main means of transportation for urban people to travel, the trajectory of the taxi can reflect the travel law of people living in the urban spatial structure [7]. The experiment uses the Haikou Didi Chuxing data set from the Gaia Open Data Project[8] for analysis and verification. This data set is a typical class of origin-destination (OD) data, which collects the information on crowd activities using Didi software every day from May 1, 2017, to October 31, 2017, in Haikou City, including the starting and ending latitude and longitude of the order, the order attribute data of order type, travel category, and the number of passengers. In the experiment, the study area is divided into 968 geographic units with 22 rows x 44 columns. Each geographic unit is a regular grid of 1000 meters x 1000 meters. A day is divided into 24 time units at 1 hour intervals.

2. POI Data Set

To study the relationship between crowd activity models and urban functional structure, the research results in this paper are explored and interpreted based on the Point of Interest (POI) in the study area. 20395 POI data of the research area in the Baidu map in 2017 is selected in the experiment. The data contains ten fields, from which eight main fields are selected for experimental study. The main fields and examples of POI data are shown in Table 1, and the geographic distribution of POI points is shown in Figure 3. Baidu map service divides POI points into first-level categories, including food, hotel, shopping, tourist attractions, etc.[9]. The research mainly explains the analysis results of the experiment according to the category of POI points.

Table 1. Main fields and examples of POI data

Field type	Example
province	Hainan Province
city	Haikou City
district	Meilan District
POI name	Crossing the Sea Memorial Plaza
POI address	200 meters east of the intersection of Haishun Road and Bihai Avenue
POI type	tourist attraction
POI longitude	110.330528
POI latitude	20.077276

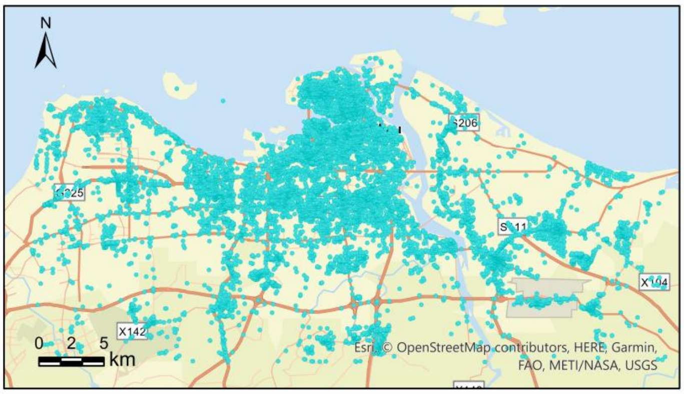


Figure 3. Geographical distribution of POI points

3.2. Extraction and Analysis of Crowd Spatial and Temporal Activity Vergence Models

The study analyzes the models of crowd vergence on non-working days. By calculating the aggregation degree of each geographic unit under the geographical flow field of crowd activities on non-working days, and normalizing and classifying the aggregation degree, the time series of crowd vergence grades are constructed. Moreover, when using DTW to calculate time series, to improve the operation efficiency, a search window is set to control the search range, and the window size of DTW is set to 3 in the experiment. The number of clusters *k* affects the validity of the clustering results. In the experiment, the DBI index is calculated under different *k* values, and it is found that when k is 5, the DBI value will be the lowest, so the study sets the number of clustering categories to 5.

According to the method proposed in this paper, the spatial and temporal models of crowd activities are extracted into five categories. Due to the interference of blank geographic units without data on the clustering results, there is a category of clustering results that has no obvious fluctuation in time and is near the value of 0. Therefore, after removing this category, the vergence spatial and temporal models of the four main crowds are finally extracted. The study visualizes the mean of all-time series in the four models as the cluster centre. On the whole, there are obvious differences in the aggregation and diffusion of the four models during the day, but no obvious aggregation and diffusion from 2:00 a.m. to 6:00 a.m., since people are mainly resting during this period, and the intensity of crowd activities is weak. The vergence of the four main models are as follows:

- C1 model: Overall, there is a "diffusion model for most of the time period." For example, from 7:00 to 22:00, the diffusion is continuously present; the degree of diffusion is the highest, from 17:00 to 18:00 in the afternoon.
- C2 model: As a whole, it presents a "daytime gathering while nighttime diffusion model." The model continues to gather from 6:00 a.m. to 9:00 a.m. and from 12:00 a.m. to 19:00 a.m. In particular, the degree of aggregation is the highest, from 8:00 to 9:00 and from 16:00 to 17:00, which diffuses from 20:00 to 23:00.
- C3 model: As a whole, there is a " diffusing in the morning and aggregating from afternoon to evening model." This model shows diffusion from 6:00 to 10:00, and the highest diffusion level shows from 8:00 to 9:00. After 14:00 in the afternoon, the overall aggregation trend is shown, and the time periods with the highest aggregation levels are from 17:00 to 18:00 and from 22:00 to 23:00.
- C4 model: On the whole, there is a model of "aggregating in the morning, diffusing in the afternoon, and aggregating in the evening." The model shows a clustering trend from 6 to 12, a diffuse trend from 13 to 19, and a clustered model from 20 to 23. Moreover, the degree of aggregation and diffusion in this model is low. There are no periods of high aggregation and high diffusion, indicating that the mobility of the population in this model is weak.

To further understand the spatial distribution of the four models and the underlying crowd activity rules, the study combines the 17 POI types defined by the Baidu Map of Haikou City to explain the four models. In the experiment, the POI points in the coverage area of each model are counted by category, and then the statistical results of each category are divided by the total number of POIs of each category in the entire study area to indicate the significance of POI types in each model. The formula is as follows:

$$\gamma^a = \frac{N_C^a}{N^a} \tag{4}$$

In the formula, γ^a refers to the significance of POI type *a* in model C, N_{Ci}^a is the number of POI type *a* in model C, and N^a is the number of POI type *a* in all models.

The study calculates the significance of 17 POI types in the four models, as shown in Table 2. It can be seen from Table 2 that:

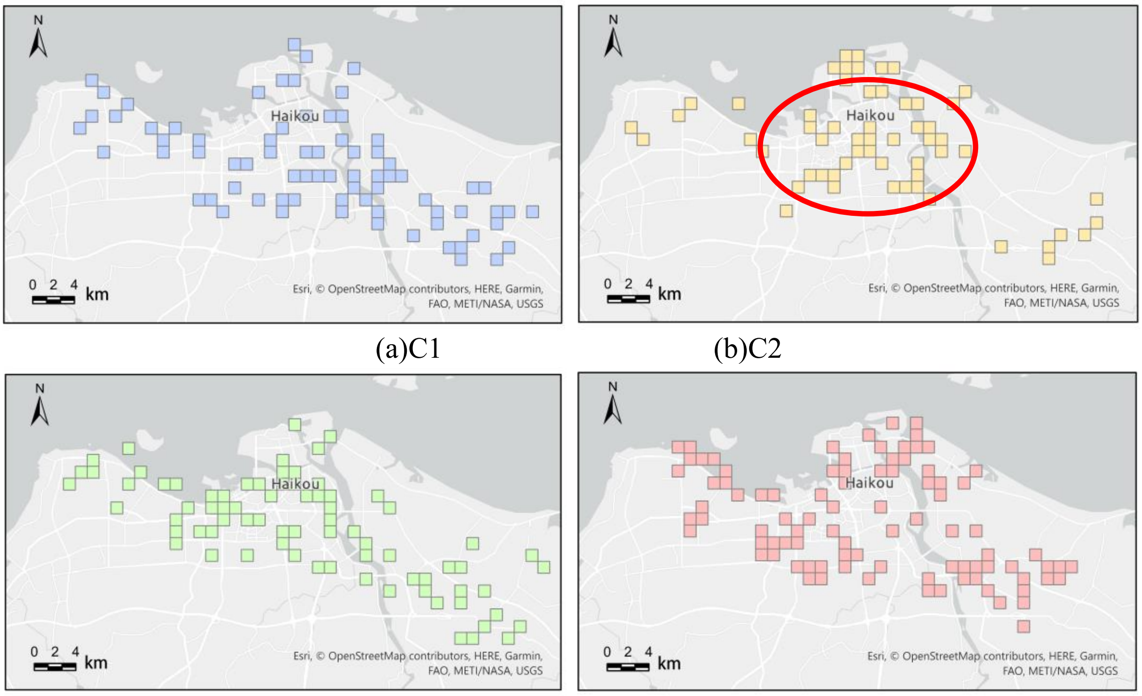
- Crowd activities show a diffusion pattern in most of the time periods in the C1 model, but it is obviously unreasonable for a geographic unit to diffuse continually. By combining POI types, it is found that the most significant POI type in the C1 model is transportation facilities. Traffic facilities include large intersections and parking lots, and the data used in the experiment is taxi data. People tend to ride at intersections, parking lot exits, and other areas. Therefore, continuous diffusion is prone to occur in areas with many traffic facilities.

- In the C2 model, most of the daytime is in a trend of aggregation, and at night it shows a trend of diffusion. By combining the POI types, it is found that companies, finance, cultural media, and government agencies in the geographic unit of the C2 model belong to the dominant POI types, showing that the C2 model is mainly located in geographical units with many enterprises and companies, so the crowd will show an aggregation pattern during the day, especially during the rush hour (from 8:00 a.m. to 9:00 a.m).
- The C3 mode mainly presents the diffusion in the morning and aggregation from the afternoon to the evening. The dominant POI types in this model include residential, hotel, etc. Since people in residential areas usually choose to go out in the morning, especially it is the rush hour from 8:00 a.m. to 9:00 a.m., its level of diffusion is the highest. On the one hand, in the afternoon, especially from 17:00 to 18:00, there is a high aggregation pattern, which is in line with the life pattern of people returning home from getting off work. On the other hand, people usually choose to stay in hotels in the afternoon and evening and check out in the morning, so it also conforms to the C3 vergence model.
- The C4 pattern presents a pattern of aggregating in the morning and spreading in the afternoon, and aggregating in the evening. Combined with the significant degree of POI type, it is found that this mode is located in the geographical unit where shopping and food types are more significant. In addition, the crowd tends to shop in supermarkets and vegetable markets and dine during non-working hours of the morning and evening, conforming to the living habits of the crowd.

Table 2. Significance of POI types in four models

C1		C2		C3		C4	
POI type	significance	POI type	significance	PO type	significance	POI type	significance
Transporta-tion Facilities	**0.51**	**Enterprises**	**0.48**	**Housing**	**0.42**	**Shopping**	**0.40**
car services	0.35	Finance	0.48	hotel	0.37	delicacy	0.32
Domestic services	0.25	Culture media	0.35	Leisure and entertain-ment	0.36	Education and training	0.32
Medical treatment	0.21	government organization	0.32	tourist attraction	0.36	medical treatment	0.32
enterprises	0.20	tourist attraction	0.29	finance	0.30	Domestic services	0.31

The study visualizes the geographic distribution of the four models, as shown in Figure 4. On the whole, there are differences in the geographical distribution of the four models, among which C1, C3, and C4 are relatively scattered in distribution, with no obvious regularity. As for the C2 model, it is mainly distributed in places with many enterprises and companies, which are mainly located in the commercial and economic center of the study area, as shown in the red circle in Figure 4(b).



(a)C1 (b)C2

(c)C3 (d)C4

Figure 4. Geographical distribution of the four types of temporal and spatial clustering models of crowd activities

4. Conclusion

In this paper, the model extraction problem is transformed into a time series clustering problem, and the mobility of crowd activities is dynamically modeled by introducing vector field theory. Then, the vergence of the crowd is calculated by the divergence, and finally, a time series composed of the crowd vergence is constructed to obtain the main crowd activity vergence model through the clustering of the time series. The experiments in this paper are carried out on the Didi Chuxing dataset in Haikou City, and four main crowd activity vergence models are extracted. Combined with the distribution of POI types, the four models are semantically explained.

The method proposed in this paper can extract the main crowd activity vergence models and provide certain method support for human mobility research. However, the experiment in this paper adopts a single data source, which has certain limitations. In the future, the fusion of multi-modal data will be considered to more comprehensively reflect the law of crowd activities.

References

[1] Yang X P, Fang Z X, Zhao Z Y, et al. 2016. Exploring urban human spatio-temporal convergence-dispersion models: a case study of Shenzhen city. Journal of Geo-information Science, 18(4):486-492.

[2] Xiao S Y, Fang Z X, Chen B Y, et al. Spatio-temporal GIS analysis of urban crowd activities [M], Beijing: Science Press, 2017.11.

[3] Yang X P, Fang Z X, Zhao Z Y, et al. Analyzing Space-Time Variation of Urban Human Stay Using Kernel Density Estimation by Considering Spatial Distribution of Mobile Phone Towers[J]. Geomatics & Information Science of Wuhan University, 2017, 42(01):49-55.

[4] Yang X, Fang Z, Xu Y, et al. Understanding Spatiotemporal models of Human Convergence and Divergence Using Mobile Phone Location Data[J]. ISPRS International Journal of Geo-Information, 2016, 5(10):177-177.

[5] Mazzoli M, Molas A, Bassolas A, et al. Field theory for recurrent mobility[J]. Nature Communications. 2019.
[6] Mardia, K. V.; Jupp, P. E. Directional statistics, 1st ed.[M]; John Wiley & Sons, Inc.: New York, USA, 2000; pp. 15-17.
[7] Zhang Y, Li Y, Zheng X, Liu H. Research on trajectory network and flow model based on Taxi Data[J]. Geospatial Information,2021,19(07):1-6.
[8] GAIA Open Dataset[EB/OL]. https://gaia.didichuxing.com.
[9] https://lbsyun.baidu.com/index.php?title=lbscloud/poitags