STUDY ON ADAPTIVE PARAMETER DETERMINATION OF CLUSTER ANALYSIS IN URBAN MANAGEMENT CASES

J.Y. Fu[a], C.F. Jing[a, b, *], M.Y. Du[a, b], Y.L. Fu [a], P.P Dai[a, c]

[a] School of Geomatics and Urban Spatial Information of Beijing University of Civil Engineering and Architecture, 100044 Beijing, China-1143896070@qq.com,diecheyanruyu@gmail.com

[b]Key Laboratory for Urban Geomatics of National Administration of Surveying, Mapping and Geoinformation, Beijing, China-jingcf@bucea.edu.cn,dumingyi@bucea.edu.cn

[c]Beijing Digsur Science and Technology co.Ltd, China-854701143@qq.com

Commission VI, WG VI/3

KEY WORDS： DBSCAN algorithm, urban management cases, cluster analysis, data mining

ABSTRACT：

The fine management for cities is the important way to realize the smart city. The data mining which uses spatial clustering analysis for urban management cases can be used in the evaluation of urban public facilities deployment, and support the policy decisions, and also provides technical support for the fine management of the city. Aiming at the problem that DBSCAN algorithm which is based on the density-clustering can not realize parameter adaptive determination, this paper proposed the optimizing method of parameter adaptive determination based on the spatial analysis. Firstly, making analysis of the function Ripley's K for the data set to realize adaptive determination of global parameter MinPts, which means setting the maximum aggregation scale as the range of data clustering. Calculating every point object's highest frequency K value in the range of Eps which uses K-D tree and setting it as the value of clustering density to realize the adaptive determination of global parameter MinPts. Then, the R language was used to optimize the above process to accomplish the precise clustering of typical urban management cases. The experimental results based on the typical case of urban management in XiCheng district of Beijing shows that: The new DBSCAN clustering algorithm this paper presents takes full account of the data's spatial and statistical characteristic which has obvious clustering feature, and has a better applicability and high quality. The results of the study are not only helpful for the formulation of urban management policies and the allocation of urban management supervisors in XiCheng District of Beijing, but also to other cities and related fields.

1. INTRODUCTION

With the rapid growth of large-scale data processing and in-depth analysis of demand in all walks of life, data mining has become a hot area of research for many scholars(Genlin et al.,2014).Refinement, which is an important goal of the urban operation and development, provides the technical support for the delicacy management of city operation(Jing 2014). With the progress of society and science technology, all kinds of issues with respect to urban operation have appeared in succession. According to the city's report on the work of the government, the number of cases about urban management(ChengGuan case in short)also increases year by year, which has influence on the urban appearance and steady running of the city. Therefore, it is of great theoretical and practical value to use the spatial data mining technology to analyze the urban management cases and assist the government decision-making.

* Corresponding author

As a method of data mining, clustering analysis has been widely used. DBSCAN algorithm based on density clustering which has high speed, data adaptability, noise insensitive characteristics was studied by many researchers(Xinyan and Deren,2005). However, the DBSCAN algorithm needs to manually determine the parameters Eps and MinPts, and the values of these two parameters directly affect the quality of data clustering. In view of the problem of how to select the optimal parameters, a large number of literatures have proposed the method of assuming MinPts value and then determining the Eps value. Although avoiding parameter determination artificially, these methods based on the premise of assumption of MinPts are still lack of adaptive parameters Such as Ren Xingping(Xingping et al.,2007)et al take MinPts as 4. According to the forth nearest neighbor distance graph of the data object set, the value of Eps is taken as less than the percentage of noise level; Zhou Dong(Dong and Peng,2009)et al assume that MinPts is 3, and then according to K-dist curve to determine the value of Eps. Some scholars have done some research on the adaptive determination of global parameters Eps and MinPts. Among them, the majority are the research that under the premise of statistical analysis of the data set. For example Xia luning (Luning and Jiwu,2009) et al proposed to k-dist probability curve and statistical model fit peak to Eps, drew the Noise curve and its inflection point MinPts method to achieve the parameters of the adaptive determination, but the whole process is too cumbersome and calculation is large, and the practicability is weak; Li Zonglin(Zonglin and Ke,2016.)et al established a suitable mathematical model to determine the Eps and MinPts values adaptively by using the kernel density estimation theory, but this method is not suitable for data set with large density difference, and the computational complexity of the algorithm is high. There are some scholars to explore the data partition as the premise of the research methods, such as Stefanakis(Stefanakis, 2007), Pandey Abhilash Kumar(Pandey Abhilash Kumar and Dubey Roshni,2014) and others on first divided again clustering of data area. Huang Gang(Gang et al.,2015)et al reduced the number of regional query method to achieve high efficiency clustering algorithm by selecting the seeds on behalf of objects.

In summary, the existing literature on spatial data and spatial statistical characteristics of the research is less. The DBSCAN clustering algorithm based on density still needs to study the data set to explore the statistical characteristics of the data and achieve high quality clustering. This paper uses the Ripley's function and K-D tree to analyze the statistical characteristics of urban management case data, and determine the parameters of DBSCAN algorithm adaptively. The optimized DBSCAN algorithm is used for data mining in typical urban management cases to provide auxiliary decision for urban management policy making, for urban management supervision staff scheduling to provide quantitative analysis support and enhance the city running fine management ability.

2. DATA AND MENTHODS

2.1 The Study Area And Data Source

This paper is based on the Xicheng District of Beijing in 2010. Xicheng District is located in the center of Beijing, which is a set of politics, economic, culture and tourism as one of core development area. It has higher requirements for urban management because of its special geographical position. The specific distribution shown in Figure 1.

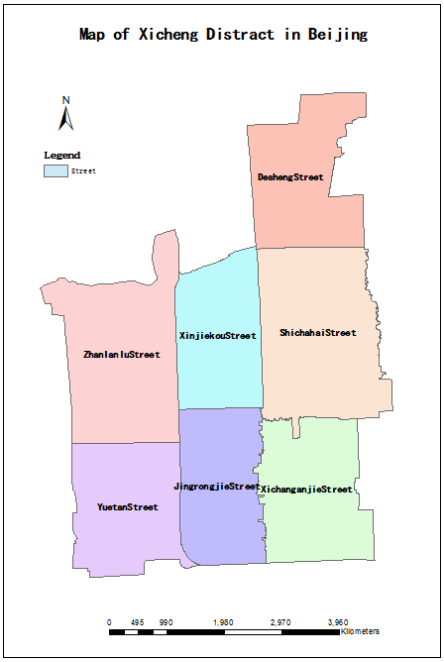


Figure1. The study area

This paper chooses the city grid management case data from 2009 to 2012 in Xicheng District as the research data source (The number of cases is shown in Figure 2). This paper selects the largest proportion of urban environment (39%) and street order (36%) as the typical urban management cases. According

to the national standard definition of "digital city management information system _ second parts: management components and events" (GB/T 30428.2-2013), the case of urban environment refers to the cases that affect the appearance of the city, mainly including the exposed garbage, the dirty green land and the unclean road; Street order cases include unlicensed tour operators, shop management and vagrant begging.

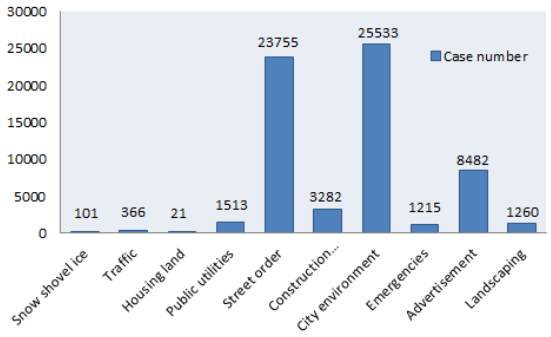


Figure2. Data statistical histogram

2.2 Research Method

This paper makes the global parameters Eps and MinPts determined adaptively based on the classical DBSCAN algorithm, and puts forward an improved DBSCAN algorithm which is applied to urban management data clustering analysis case. Through the correlation analysis between the cluster results and the cases of urban management components, the correlation between clustering results and urban component facilities is verified, which can provide assistant decision for city managers. Among them, for the adaptive determination of parameters Eps and MinPts, this paper proposes that using Ripley 's K method to obtain the best maximum correlation distance aggregation degree as the value of Eps scanning radius. Then, according to the K-D tree principle, each point object in the data set is calculated in the Eps neighborhood of the number of points K, and marking Freq as the frequency of occurrence of K, taking the maximum Freq of K value as MinPts. The specific process of this paper research ideas are shown in Figure 3.

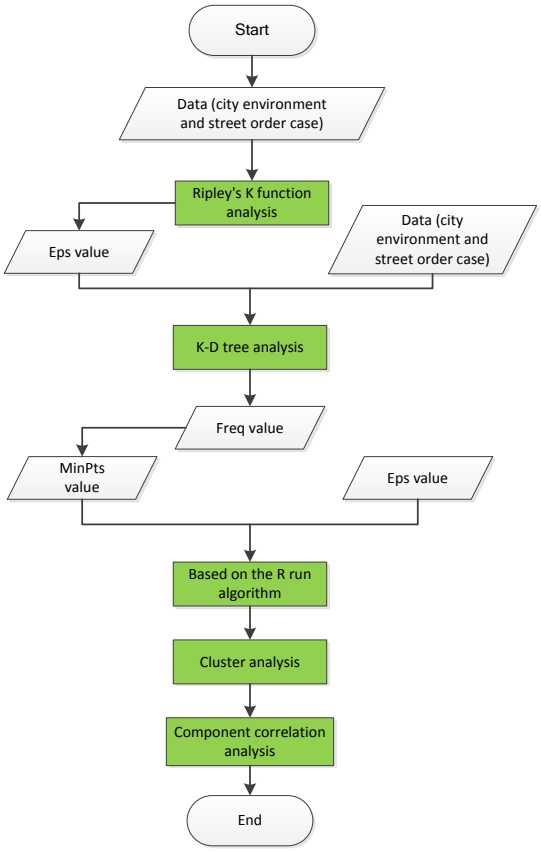


Figure3. Flow chart for the research method

2.2.1 Ripley's K Function

Ripley's K method is a representative spatial point pattern analysis method, which can quantitatively evaluate the spatial distribution characteristic of point patterns(Tang et al.,2015). In this method, Ripley's K function is used to analyze the clustering degree of point datasets at different spatial scales in a certain confidence interval, then the maximum clustering scale of the best clustering effect is quantitatively analyzed according to the expected K value and observed K value. The spatial scale is calculated as follows:

$$L(d) = \sqrt{A\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}K_{i,j}}{\pi n(n-1)}} \qquad (1)$$

In the above formula, d represents the distance, A represents the total area of the area occupied by the feature set, K_{ij} represents the spatial weight.

At a certain specific distance, when the observed K value is

larger than the predicted K value, the clustering degree of the distribution is higher than that of the random distribution of the scale. Therefore, this paper selects the spatial scale with the largest difference between observation K and prediction K as the value of parameter Eps, in order to get a better clustering effect, this paper sets the confidence level to 95% based on the data size of research.

2.2.2 K-D Tree Analysis

K-D tree is similar to the binary tree, it is a data structure with left and right subtrees. The biggest difference between K-D tree and the binary index tree is that K-D tree is stored in the K-dimensional point data. The K-D tree algorithm is composed of two parts, including tree-building and search. K-D tree is divided into left subtree and right subtree according to the max variance of data. In order to make sure the left subtree and the right subtree have the same length, K-D takes the median where array of attribute value as the partition axis. There are two kinds of search methods in K-D tree structure: range search and K-nearest neighbor search. The range search refers to searching the point data within the threshold range for a certain point object in the given searching threshold; K-neighbor search refers to specifying a point object, and then traversing the original data set to find the nearest point of the object K point data.

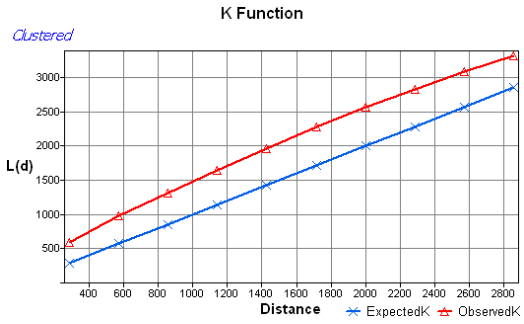
This classical K-D tree algorithm can only be used to search the high efficiency K in the low dimensional case, and the efficiency of searching for high-dimensional data is very low. In view of this problem, some scholars have proposed an improved method(Yi et al.,2016). This paper calculates the point data size K for each point object which is in the Eps threshold, and then takes the max frequency of statistical analysis of the value of K as MinPts value.

2.2.3 Component correlation analysis

Overlapping analysis method has the characteristics of low domain knowledge dependence and clear physical meaning of mining results(Lin et al.,2013)**.**The rationality of the urban infrastructure (such as the urban management component) is the basis for realizing the meticulous management of the city and improving the comprehensive operation ability (Zhonggui,2014). In this paper, the component case in the data has the location attribute of the component, therefore, in order to analyze the correlation between the setting of component facilities and the clustering results more intuitively, the distribution of urban management component cases is visualized by kernel density analysis, then the visualization results of cluster cases are overlaid with the clustering results, to check whether the clustering is consistent with the nuclear density concentrated area of the component, so as to verify the rationality of clustering the results. Moreover it can provide the assistant decision for urban managers to optimize the spatial distribution of the urban components.

3. TEST AND RESULT ANALYSIS

Using the Ripley 's K function to analyze the clustering degree of typical urban management cases. The results are shown in Figure.4.



(a) Street order class

(b) City environment class

Figure4. Ripley's K result

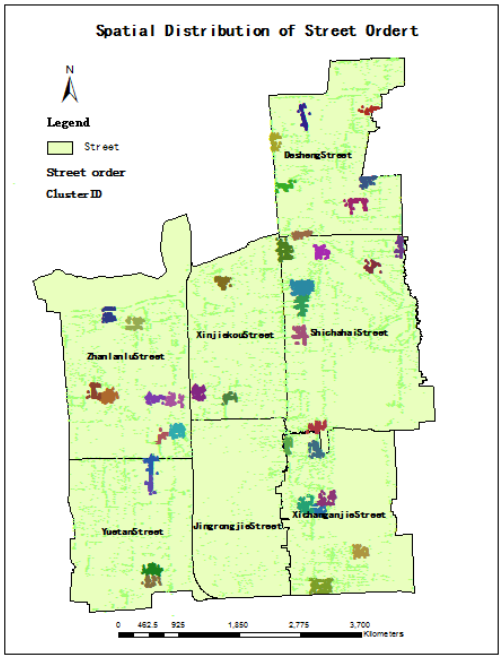
As can be seen from the figures, the observed K values (with △ lines) are both larger than the predicted K values (with × line), which indicates that these two types of cases both have a spatial aggregation mode. After the statistical analysis , the maximum aggregation scale is shown in Table 5.According to the quantitative analysis in the light of the above chart, when aggregation scale of the street order cases is greater than the maximum aggregation scale of 563 meters, the observed K value gradually approaches the prediction K value and they

almost keep parallel; When the aggregation scale of the urban environment cases is larger than the maximum aggregation scale of 792 meters, the observation K value is close to the forecast K value and almost parallel, which shows that as for these two cases, when the spatial scale is larger than the maximum aggregation scale, the degree of aggregation of the data set is decreasing. In view of the fact that the observed K value is higher than the predicted K value at a certain distance, the aggregation degree is higher. Therefore, this paper selects the maximum aggregation scale as the parameter Eps when the aggregation effect is best.

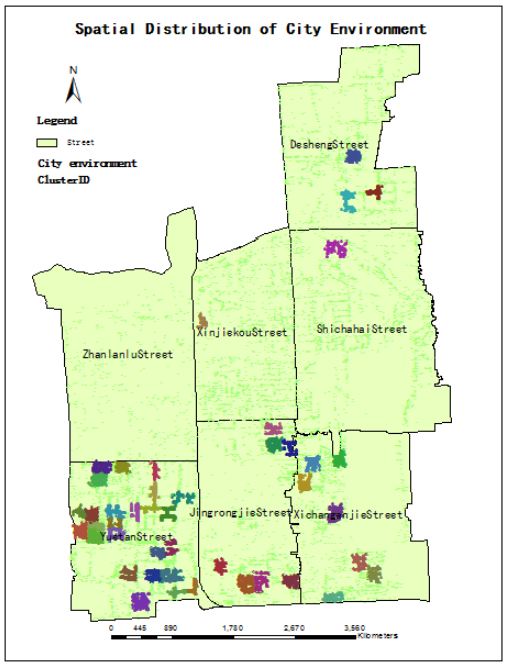
Serial number	Case name	Maximum aggregation scale (m)	Confidence
1	Street order class	563	95%
2	City environment class	792	95%

Table5. The biggest clustering scale

According to the principle of K-D tree, this paper counts the number of data points of each point object of the original data set in the scanning radius Eps is K, and takes the highest K value as MinPts, the DBSCAN algorithm runs based on R, and the clustering results are shown in Figure.5.The similarity between cluster and cluster is the smallest, and the similarity is the largest among clusters, which means that the same clusters are clustered together, the different clusters has a specific distance ,and the scattered points represent noise points that do not constitute a distinct cluster.



(a) Street order class（Eps=560m MinPts=110）



(b) City environment class（Eps=792m MinPts=200）

Figure6. Spatial distribution of cluster results

It can be seen from Figure. 6 that the clustering characteristics of these two cases are obvious and the noise points are almost evenly distributed. The number of cluster types of the city environment was as high as 25 on Yuetan street(The total number of clusters is 44). Combining the definition of density-based clustering: finding the high-density regions isolated by low-density regions, which indicates that the clustering accuracy of the algorithm is high; While the noise points are almost evenly distributed, which shows the advantage that the clustering algorithm can still find the arbitrary shape clusters in the presence of "noisy" sample data(Yupeng,2015).In summary, it can be concluded that the DBSCAN algorithm based on parameter optimization has higher clustering quality.

The clustering results and the parameter values are statistically analyzed, as shown in Table 7. Figure 6 and Table 7 indicate that: (1) In the case of the street order case, there are 38 clusters under the condition of confidence 95%, scanning radius 563m and scanning density 110, which are mainly distributed in the east and west streets of the city center and north of the city. In addition to the less clustered cluster distribution on Yuetan and Xinjiekou streets, apart from Yuetan and Xinjiekou streets with less clusters, the clusters of

West Chang'an Street, Shichahai Street, Desheng Street and Exhibition Road are uniformly distributed. The northern part of the city has larger population mobility, and the majority of the floating population on the convenience, fast shopping needs for unlicensed operators and operators outside the shop to provide a market. And the requirement for convenient and fast shopping of the majority of the floating population provides the unlicensed operators and operators outside the shop with the market. So It is extremely necessary to strengthen the management and deployment of urban managers in the northern part of the city; (2)Under the conditions of confidence 95%, scan radius 792 m and scanning density 200, the cluster of urban environment cases is 44, which are mainly distributed in Yuetan Street, Financial Street and West Chang'an Street in the center of the city, the northern part of the city also has a small amount of distribution, such as Desheng Street, Xinjiekou Street and Shichahai streets, in which the east side of the city, the Yuetan street is especially concentrated, and the number of cluster types is as high as 25, compared with the other streets in Xicheng District, the number of communities (32) and population (150,000) is the largest in Yuetan district. The excessive communities and population make the number of long-term abandoned vehicles on the roadside and garbage piled up on the side of the road larger than other streets. Yuetan streets become a high incidence of urban environment class cases because of lacking of Sanitation facilities and the exposed garbage can not receive timely processing. So we should strengthen city management personnel in Yuetan Street inspections, in order to reduce the number of cases of the occurrence of such cases, improve the current environmental problems.

Serial number	Case name	Clustering results	Scan Radius（m）	Scanning density	Confidence
1	Street order class	38	563	110	95%
2	City environment class	44	792	200	95%

Table7. Statistical analysis of clustering results

In view of the street order case is mainly affected by unlicensed business operators, shop-outside operators and vagrant begging and other issues; City environment cases are mainly affected by the exposed garbage, unclean pavements, accumulation of waste residue and abandoned vehicles and other issues. Lacking of urban public infrastructure is the root cause of a large part of the city environment. Therefore, in order to verify the rationality of the clustering results, this paper choose 14 classes cases about urban public infrastructure and urban environment, such as dustbin, garbage bin, comfort station, storage frame and so on to make correlation analysis according to ” digital city management information system _ second parts: management components and events" (GB/T 30428.2-2013) ”.If there are a lot of duplicate areas between the clusters and the concentrated areas of component cases, which means that urban management cases is unrelated to the component facility. Conversely, If there is only a small amount of overlap between the clusters and the concentrated areas of the component cases, it shows that the urban management case is related to the configuration of the components. The specific correlation analysis results are shown in Figure 8, in which different colors represent different clusters, transition which is from white to black region shows the nuclear density of components from small to large.

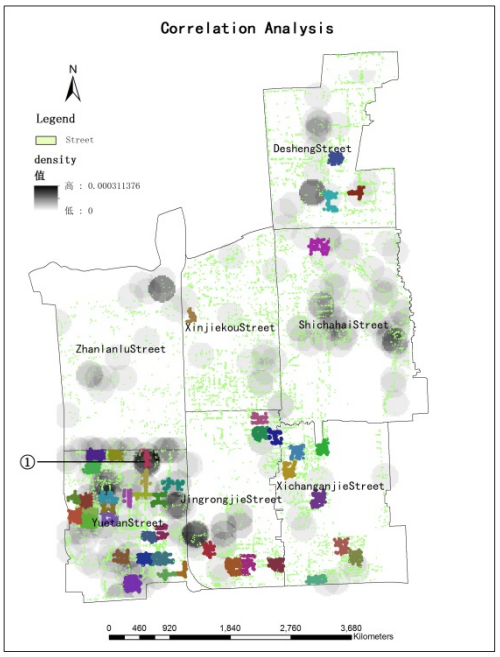


Figure8. Cases management correlation analysis results

As can be clearly seen from the above figure, there are only 1 clusters in the 43 clusters, which overlap with the dense areas of the component cases, this shows that the urban environment

and environment cases have a great relationship with the configuration of urban public infrastructure components, the probability of occurrence of the case in the area where the urban public infrastructure is less (nuclear density is small) is larger than that of the city with more basic infrastructure arrangement (nuclear density), this analysis proves the rationality of the clustering result. Therefore, in the process of dealing with the urban environment cases, we can increase the number of urban public infrastructure such as garbage cans, public toilets and storage stents in order to reduce the probability of the recurrence of cases.

4. DISCUSSION

DBSCAN algorithm which is based on the density is an unsupervised classification method. Without any future knowledge, the values of Eps and MinPts have a great influence on the analysis results. Therefore, the adaptive determination of its parameters is a hotspot of research in recent years. In order to solve the problems of parameter adaptive and clustering quality, this paper proposes a method based on Ripley's K function and K-D tree analysis to optimize the parameter values, this method is applied to the data mining of meticulous management cases, to provide Decision support for reasonable deployment, urban management administrative policy making. The experimental results show that the parameters determined by this method have good adaptability, high clustering quality and obvious effect, and it is of practical significance for fine urban management practice. But because of the limited amount of the accumulated data, the experimental area is relatively small, so collecting more data to verify the algorithm will be the main work of the follow-up. But collecting more data to verify the algorithm will be the main work in the future because the limited amount of the accumulated data and the small experimental area.

REFERENCES

Dong,Z.,Peng,L.,2009. VDBSCAN：varied density based clustering algorithm**.** *Computer Engineering And Applications*, Vol 45 (11):pp.137-141.

Genlin, J., Bin,Z.,2014. A Survey of Spatiotemporal Data Mining for Big Data. *Journal of Nanjing Normal University*, Vol 37(1):pp.1-7.

Gang, H., Chaozhong, W., and Nengchao,L.,2015. A Study of Laser Radar Object Detection Based on Improved DBSCAN Algorithm. *traffic information security*, Vol 33 (3):pp.23-28.

Jing,X.,2014. Conceptual Innovation in Big Data Times and Refined Treatment of Cities. *Shanghai urban management***,** Vol 23(4):pp.23-26.

Luning,X.,Jiwu,J.,2009. SA-DBSCAN :A self-adaptive density-based clustering algorithm. *Journal of Graduate University of Chinese Academy of Sciences*, Vol 26 (4):pp.530-538.

Lin,D., Hong,S., Xiao,N.,2013. Spatial Association Rule Mining Based on Overlay Analysis and Area Calculation. *Journal of Wuhan University, Information Science Edition*, Vol 38 (1): pp.95-99.

Mr. Pandey Abhilash Kumar, Prof. Dubey Roshni.2014. A survey on study of enhanced partition based DBSCAN algorithm. *International Journal of Management, IT and Engineering*, (3):pp.115-123.

Stefanakis,2007. NET-DBSCAN: clustering the nodes of a dynamic linear network. *International Journal of Geographical Information Science***,** (4):pp.427-442.

Tang, Wenwu;Feng, Wenpeng;Jia, Meijuan.2015. Massively parallel spatial point pattern analysis: Ripley's K function accelerated using graphics processing units. *International Journal of Geographical Information Science*,(03):pp.412-439.

Xinyan,L., Deren,L.,2005. DBSCAN spatial clustering algorithm and its application in urban planning. *Journal of Surveying and mapping***,** Vol 30(3):pp.51-53.

Xingping,R., Zhonglong,H.,and Zenghui,M.,2007. The Improved Algorithm of Decide the Parameter Eps in DBSCAN. *Modern electronic technique***,** Vol 30 (11):pp.120-122.

Yi, Z., Gangwu,J., Yanan,D.,and Houpu,L.,2016. An Optimized K-D Tree Search Method for Feature Matching of

UAV Images. *Journal of Surveying and Mapping Science and Technology*, Vol 32 (5): pp.1-5.

Yupeng,X.,2015. Research on spatial clustering Analysis. *Harbin University of Science and Technology*.

Zonglin,L.,Ke,L.,2016. Research on adaptive parameters determination in DBSCAN algorithm. *Computer Engineering and Applications*, Vol 52 (3):pp.70-73.

Zonglin,L.,Ke,L.,2016. Research on adaptive parameters determination in DBSCAN algorithm. *Computer Engineering and Applications*, Vol 52 (3):pp.70-73.