A Study on Crowdsourcing Geospatial Data Mining Based on Spatial Statistics

Jiyuan Geng, Weidong Song, Shangyu Sun
Liaoning Engineering Technology University, surveying and mapping and geographical science college, Liaoning Fuxin123000

Abstract—**By analyzing and mining the source geospatial data, provide a reference for the government macro decision-making and public opinion monitoring and provide the basis for enterprise precision marketing and the individuality service. The experimental results show that, within the scope of the study area tend to check data and its associated attribute value spatial clustering model, clustering of statistically significant hot spots are mainly distributed in the city school, station, district position, city hot spots can be obtained through the sign-in data detection coverage is consistent with the actual urban planning scheme, and has obvious directivity, have very strong application value.**

Keywords-spatial data mining; spatial statistics; exploratory spatial data analysis; the source of geospatial data; plasticity area unit

I. Introduction

The concept of geospatial data is proposed by Ramm (2008), Hudson-Smith (2009), Heipke (2010) and so on. Heipke in his paper described the concept of "the source", in particular, it is difficult to rely on a large number of users to carry out the automation of data collection work

Analysis of the related research is still at the stage of exploring the source of geographic spatial data model, statistical pattern recognition constitutes the theoretical foundation in this field, spatial statistics provides a research method for pattern analysis, GIS provides a practical tool for data modeling, visualization and analysis, constitute a relatively complete system

II. Acquisition and Preprocessing of the Spatial Data of the Public Source

A. Data Acquisition

Public geospatial data, especially social and geographic data generated from the Internet, can not through the traditional aviation, aerospace, ground measurement and digital means to obtain. This article uses the method which obtains the registration data through the micro-blog Sina open platform, this method also applies to other social network platform's public source geographic space data acquisition.

Attendance data acquisition results shown in Figure I

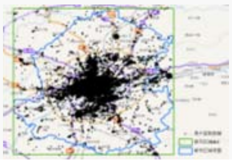

FIGURE I. SIGN IN DATA ACQUISITION RESULTS

B. Data preprocessing

Using data mining method to discover knowledge from massive data requires that the original data must be correct, consistent, complete and reliable, and the data collected during the data acquisition phase is a problem. After treatment, get the sign in place of 30941, a total of 2520128 times the average attendance number, registration number 81.45, the standard deviation is 814.86, the number of 1173882 people in attendance, the average number of 37.94, the standard deviation is 340.16, the difference between the visible and the original data.

III. Basic Theory and Method of Spatial Statistics

A. Spatial Statistics

Spatial statistics is based on the theory of regional variables, using the variation function and Kriging interpolation method as the main tools to study the science of the natural phenomena that are both stochastic and structural in the spatial distribution. Spatial statistics research object, including any features and phenomena with spatial and attribute information. The core statistical methods of spatial statistics are mostly descriptive and exploratory. It is mainly used to describe the central tendency, the degree of dispersion, the direction and the degree of the spatial data.

B. Spatial Sampling

Spatial sampling methods used in spatial statistics mainly include: stratified random sampling, random sampling and random sampling. This paper mainly introduces the spatial sampling method applied to the point pattern analysis. The spatial distribution of commonly used Poisson distribution is used to describe the geographical elements in the grid unit, can be described as a Poisson distribution:

$$p(x) = e^{-\lambda}\lambda^{x}/x! \tag{1}$$

$$\lambda = n/k \tag{2}$$

Among them, n said the number of geographical elements, K indicates the number of grid cells, said the average number of geographic elements of lambda grid unit values in P (x) x probability that geographical elements appear in the grid unit.

When will the Poisson distribution to create a frequency distribution table of frequency distribution table creation obeys CSR distribution and geographical factors when we can preliminary judgment in the study area is geographical elements show some spatial patterns, such as the number of geographical elements contain some grid cells than in the number of CSR distribution were more, while geographical elements may represent a spatial clustering model. Quadrat sampling (quadrat analysis) statistical significance can be measured by K-S or test.

IV. Spatial Sampling and MAUP Effect

A. Spatial Sampling of Registration Data

Sampling is a commonly used quadrat sampling method of space. Through the stack composed of regular shape and the same size grid unit grid in the study area, and calculate the number of falls in each grid unit point entities, will eventually give the corresponding mapping grid unit. The space of data sampling is the sign in place by this method, the study area covers an area of 376.44 square kilometers, the number of sign locations is 30941, the length of sample equal to about 150 meters, in order to avoid the sample size is too large to capture enough detail in the study area, the plot length was reduced to 100 m.

The length of 100 meters of square grid division and superposition operation in the study area, and the number of mapping each grid unit in place to sign the property grid unit value, the result is shown in figure II:

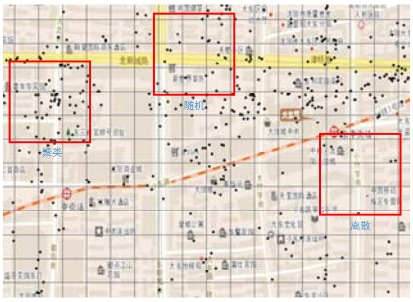

FIGURE II. GRID PARTITION AND SUPERPOSITION RESULTS

By II, the success to the spatial distribution of grid location is not consistent, part of the region is more intensive, part of the region are sparse, some areas almost does not contain any sign in place, this is hidden in the sign in the data space mode and data feature is the paper to research topics.

At the point of entity and space grid connection process, will need to sign in number, registration number and sign location type into account, and the attribute information of attendance data calculated by the user activity and the type of information is assigned to the corresponding grid unit, the basic principle of the method is shown in figure III:

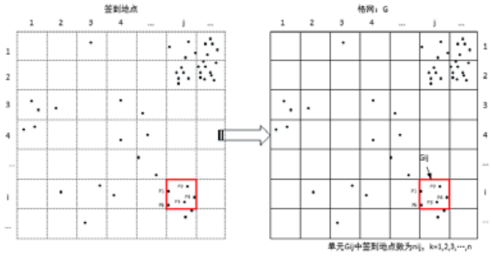

FIGURE III. SIGN IN LOCATION DATA SPACE SAMPLING

Grid user activity and information is calculated as follows:

$$G_{ij} = \sum_k N_{p_k} \times W_{p_k}, W_{p_k} = \frac{U_{p_k}}{U_{G_{ij}}} \tag{3}$$

$$T_{ij} = T_l \cdot \sum_{W_p, T_l} = \max_{1 << l << m} \left\{ \sum_{Wp, T_1}, ..., \sum_{W_p, T_m} \right\} \tag{4}$$

The k=1, 2, 3, N, l=1, 2, 3,..., m, Gij,..., said grid user activity, Pk said the K sign in place, NPl said the number of sign in Pk, WPl said the weight of Pk for UPl, said the number of sign in Pk UGij, said the total number of attendance, Tij said the single grid the type of information, can be accounted for by the same type of maximum weight sign locations and determine.

B. Spatial Autocorrelation of the Plastic Area unit Problem

The modifiable areal unit problem "referred to as MAUP, this concept was introduced by Openshaw and Taylor, can be understood as" changes in spatial patterns due to man-made space unit of geographic phenomena of the continuous problems caused by the". When spatial analysis is performed on the same geographical region with different resolutions or different scales, it is often inconsistent, which is called the scale effect. Spatial analysis of the same region, due to the different regions of the analysis results may not be consistent, which is called the zoning effect. Scale effects and zoning effects are called MAUP effects, because the two are related to the change of the standard of regional unit division.

In order to verify the influence of scale effect on attendance data, the global spatial autocorrelation analysis results, respectively 100m *100m, 500m*500m and 1000m * 1000m three dimension space sampling unit, sampling method using the improved space, attendance data in the study area of spatial sampling, grid partition diagram as shown in figure IV:

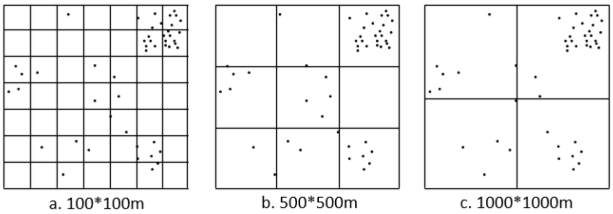
a. 100*100m b. 500*500m c. 1000*1000m

FIGURE IV. THE DIVISION OF REGIONAL GRID

Three respectively on the attendance data grid under incremental spatial autocorrelation analysis, and draw the Z score changes over distance curves, and marked statistically significant peak z scores in the line graph, the experimental results as shown in figure IV-3.6:

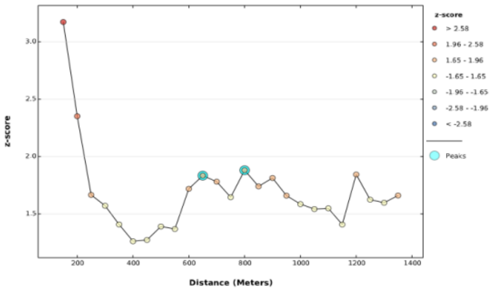

FIGURE V. FAVORED ANALYSIS SCALE (100M*100M PLOTS)

We can see from Figure V, when the spatial sampling unit size is 100m*100m, the most preferential analysis scale of 800m in the study area, a quadrat centroid as the center, extending the distance of 800m. This scale is suitable for the analysis of hot spots in the city (such as schools, stations, shopping district, etc.), which confirms the estimation of the size of the hot spots in the city.

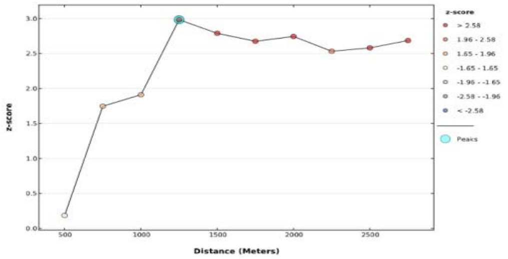

FIGURE VI. FAVORED ANALYSIS SCALE (500M*500M PLOTS)

We can see from figure VI, when the spatial sampling unit size is 500m*500m, the most preferential analysis scale of 1250m in the study area, a quadrat centroid as the center, extending the distance of 1250m. This standard applies to the study of problems in city street area as the basic unit, such as the use of data detection in city ommercial outlets in the most densely populated district.

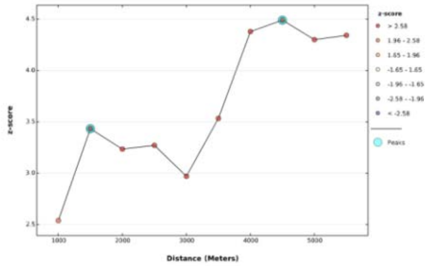

FIGURE VII. FAVORED ANALYSIS SCALE (100M*100M PLOTS)

We can see from figure VII when the spatial sampling unit size is 1000m*1000m, the most preferential analysis scale of 4500m in the study area, a quadrat centroid as the center, extending the distance of 4500m. This scale is applicable to the study of the basic unit in the city area, for example, the spatial correlation between the GDP and the sign data and statistical data.

In summary, the spatial correlation of MAUP effects on the attendance data does have an impact, because of the different ways to sign grid data showed different spatial patterns, detection of attendance data space in a quantitative way need to consider the impact of the distribution scale and zoning effect, especially the application of various spatial statistical methods, otherwise it will cause analysis of the uncertainty of space. Different dimensions of spatial sampling to analyze the scale of the problem for the unit is also different in spatial pattern detection attendance data, we must first determine the size of the problem, and then select the appropriate size of the size of the space space sampling unit in the study area and sampling properties of the grid unit and its associated value of spatial correlation analysis.

V. Pattern Analysis of the Spatial Data of the Public Source

A. *Global Spatial Pattern Detection*

Because the G coefficient is sensitive to the "boundary effect" and is affected by the scale of the cluster region, this paper uses the global s' I Moran index to study the influence degree of the spatial dependence of the sign in data. Distance calculation, as the basis of all spatial analysis, is often used as the most basic variable to describe the relationship between spatial objects. In this paper, we use the Euclidean distance to measure the distance between the sign in place. Spatial weight matrix is used to describe the relationship between spatial objects. The inverse distance is used to model the relationship between the sign in data.

The experimental results show that the property of grid unit and its associated value (user activity) has a strong spatial correlation, and the performance has significant spatial clustering pattern on the probability of this model is created by CSR process is less than 1%. The global spatial autocorrelation analysis of the sign in data is shown in figure VIII:

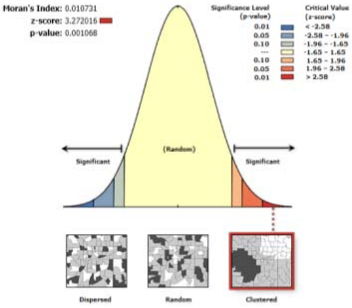

FIGURE VIII. GLOBAL SPATIAL AUTOCORRELATION ANALYSIS OF THE SIGN IN DATA

B. *Local Spatial Pattern Detection*

The local Moran 's I spatial data index can identify statistically significant outliers, and is superior to the Gi* coefficient in the accuracy of detection of clustering center of the region, the local Moran' influence space between s I index sign data dependency

Each grid unit of the study area within the scope of the calculation of the local Moran 's I index, z score, P value, and determine the type of grid unit, are statistically significant in the grid unit 93, the high value of the clustering 68, 19 high value anomaly, low value anomaly 6. There is no statistical significance of the grid unit 13570.

The spatial distribution of clustering and outliers and the visualization of user activity are shown in Figure IX and X:

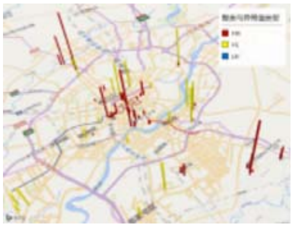

FIGURE IX. SPATIAL DISTRIBUTION OF CLUSTERING AND OUTLIERS

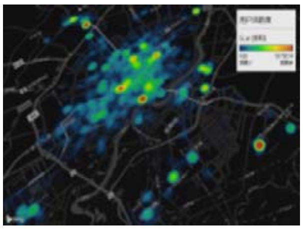

FIGURE X. VISUALIZATION OF USER ACTIVITY

Figure IX and X, can be found in high value area spatial distribution is consistent with the distribution of the local spatial pattern detected by clustering and outliers by user activity space and generate heat map, using local Moran 's I index is detected cluster center of the region, and to identify the abnormal spatial data value.

Taoxian Airport T2, T3 terminal, T2 terminal before the docking station and airport apron were identified as high value clustering, the tendency of users to share its own state information at the airport, and the flight is boarding nearby. The experimental results also show that Starbucks (Taoxian airport shop) user activity was higher than that of other food and beverage brands, which also reflects the actual operating conditions of the catering brand.

VI. CONCLUSION

Sign in data has the characteristics of large amount of data, strong timeliness, frequent updates, and contains a wealth of location, time and semantic information, which provides a new method for the perception of human social activities. This paper takes Sina micro-blog sign data for example, attendance data acquisition platform provided by Sina, micro-blog, and then uses the quadrat sampling method is verified by the experiment results of MAUP effect in the study area in spatial correlation of data, and the use of global and local spatial statistics in spatial pattern detection method to analyses the attendance data. The experimental results show that it is feasible to apply the data in the detection of hot spots in urban areas. But there is also a need to improve the way: the user activity and type of information inference method needs to be further improved and the detection of hot spots in the area of the city's attendance data semantic analysis. Combined with the results of semantic analysis, we can better understand the user's behavior and preferences, and then be able to provide the basis and reference for enterprise management and government management.

REFERENCES

[1] HeipkeC. Crowdsourcing geospatial data [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2010,54 :550-557.

[2] Wald DJ, Quitoriano V, Dengler LA etal. Utilizaiton of the internet for rapid community intensity maps[J]. Seismolo- gical Research Letters, 1999, 70(6): 680-697.

[3] Viglino JM.Handling partner's feedback through the web[R]. Proceedings EuroSDR Workshop, Berne SUI : EuroSDR, 2009.

[4] Havercroft M. Crowdsourcing-some experiences and thoughts[R]. Proceed- -ings EuroSDR Workshop, Berne SUI : EuroSDR, 2009.

[5] Guelet J C. Integration of user generated content into national databases-revision workflow at swisstopo[R]. Proceedings EuroSDR Workshop, Berne SUI: EuroSDR, 2009.

[6] Ming Wang, Qingqvan Li, Qingwu Hu and so on. Quality evaluation method for map spatial data from open source to the public [J]. Journal of Wuhan University (Information Science Edition), 2013, 38 (12): 1490-1494.

[7] Qingwu Hu, Ming Wang, Qingquan Li. Using location data to explore the urban hot spots and business circle [J]. surveying and mapping, 2014, 43 (3): 314-321.