A CLUSTERING ALGORITHM OF LAND GRADES BASED ON CLOUD HISTOGRAM

HU Shiyuan[*a, b], LI Deren[c], LIU Yaolin[a,b], LI Deyi[d]

[a]School of resource and environment science, Wuhan University, 129 Luoyu Road, Wuhan 430079,China
[b]Key Laboratory of Geographic Information System , Ministry of Education , Wuhan University, 129 Luoyu Road, Wuhan 430079,China
[c]State Key Laboratory of Information Engineering in Surveying,Mapping and Remote Sensing,, Wuhan University, 129 Luoyu Road, Wuhan 430079,China
[d]Institue of Electronic System Engineering, 6 Wanshou Road, Beijing 100039, China,

KEY WORDS: Land Grade; Cloud Model; Histogram; Clustering Algorithm

ABSTRACT:

This paper analyzes the drawbacks of traditional land grade clustering method, introduces cloud model into the frequency histogram drawing process, and advances the land grade clustering method based on cloud model and histogram. Turning the traditional clear interval data to cloud, this method draws the cloud histogram according to the cloud fuzzy frequency of each appraisal unit calculated from its total score, and promptly and accurately detects the peak points and valley points of the histogram through the automatic detecting algorithm to determine the land grade boundary. The method successfully takes into consider the fuzzy of data intervals, diminishes the abnormity in traditional histogram, and makes the land grade clustering results more reasonable and practical.

1. INTRODUCTION

Data clustering is to divide data into different groups according to their "similarities", widening differences among groups while keeping the similarities of the objects within each group.

There are two kinds of land grades clustering, including the graph-based method and model-based methods. Graph-based method includes score axis determination method, score frequency curve method, score frequency histogram method. Although these methods have their merits such as intuitiveness and comprehensibly, the drawbacks cannot be ignored: (1) all these methods depends on human judgment, which adds to the chances of accidental errors; (2) as the distribution of the total score value of appraisal unit varies differently, in actual practice, there exists either very few obvious "critical points" or irregular "faultages", which brings lots of difficulties on human recognition. The model-based methods includes iterative clustering (K-Means Cluster), Division-based method (K－medoid algorithm), Hierarchy-based method (BIRCH and CURE algorithm), Density-based method (DBSCAN algorithm), Grid-based method (STING algorithm, CLIQUE algorithm and Wave-cluster algorithm), and neural-network-based method, etc. These methods quantify land grade compartmentalization process with mathematical models and use the advanced computer recognition technology to overcome the drawbacks of graph-based methods, but still have some shortages as far as time-efficiency is concerned. In concurrent land evaluation, regular grids are used to divide the appraisal zones into several land evaluation units. As the size of the grids ranges from 25mX25m, 50mX50m to 100mX100m, based on the scale of the city, the number of appraisal units in a city varies from tens of millions to hundreds of millions. As a result, it will be huge time-consuming process for the model-based methods to cluster such large number of data, while the graph-based methods such as the frequency histogram could overcome the time-inefficiency of model-based methods. This paper analyzes the traditional score frequency histogram, introduces cloud model to improve it, and develops the clustering algorithm of land grades based on cloud histogram.

2. ANALYSIS OF CLUSTERING BASED ON SCORE FREQUENCY HISTOGRAM

Score frequency histogram method is a process that makes statistical analysis of the total scores of land evaluation units, divides the score range into several tiny ones, draws frequency histogram through frequency statistics towards the total score of every land evaluation unit in each percentile range, and finally, delimits the land level boundary according to total score frequency distribution.

Through histograms, we can sketchily understands the distribution of the total score value of each appraisal unit. However, in practical work, multi-peaks or other abnormal shapes exists in histograms, as shown in Figure 1. According to our research, these abnormities are not necessarily the actual distribution of the total score, but may be the result of the hardware division of the total score, which divides the range of the total score into several clearly-defined tiny intervals so that the total score of each appraisal unit only belongs to one interval. This hardware division may cause impractical results. For example, the scores 69.99 and 70.01 have only a slight difference of 0.02, which is not significant at all, but will be abruptly assigned into two intervals as one is less than 70 while the other is more than 70. On the contrary, the scores 70.01 and 79.99 will be assigned into the same interval since both of them are less than 70 and more than 80. As a result, there will be a jumping of scores. In the next, we will analyze the problems that may occur when applying traditional the score frequency histogram method to cluster data and choose critical points for land grades.

Figure 2 shows a segment of score frequency histogram for Wuhan City's commercial land grades, which belongs to 35 to 43 score interval. In the segmental histogram, there are 2000 unit belonging to the (38, 39] interval, and 2400 units belonging to the (39, 40] interval. However, there are also 1650 unit belonging to

the (38.5, 39] interval, 2100 units belonging to the (39, 39.5] interval and 300 units belonging to the (39.5, 40] interval, which means that 3750 appraisal units belong to the (38.5, 39.5] interval while only 300 unit belong to the (39.5, 40] interval. Therefore, it is much more appropriate to choose the score 39 as the critical point for land grades, compared with the score 38 chosen in the traditional score frequency histogram method.

Besides, in practical work, the data collected are always limited because of various reasons, and the limited data can't accurately describe the distribution of the parent population. To avoid this, we should find a new method. In the next section, we will adopt the cloud theory to divide the total scores of land appraisal units and draw cloud histogram. Making clustering analysis through cloud histogram can solve the problem said above.

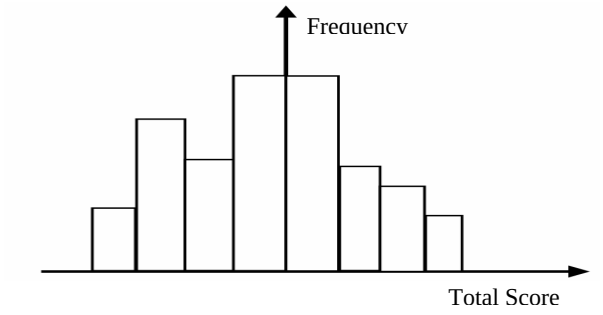

Figure 1. Traditional Histogram

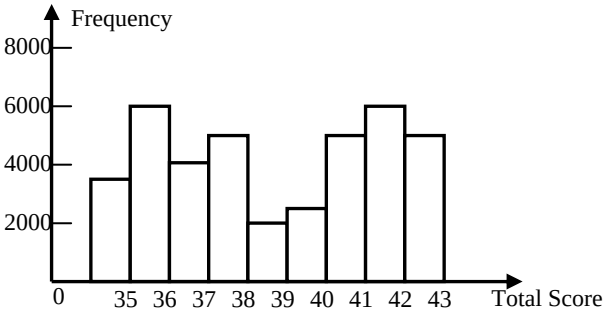

Figure 2. Histogram for Commercial Land Grades

3. CLOUD HISTOGRAM DRAWING

The key point of drawing cloud histogram is to turn the traditional clearly-defined interval data into cloud, according to cloud model theory. The range of the total score of each appraisal unit is [0,100], and the intervals of the scores are clearly-defined as shown in Table 1.

Name	1	2	…	i	…	100
Interval	(0 , 1]	(1 , 2]	…	(i-1 , i]	…	(99 , 100]

Table1. Intervals of the Scores

Consider the total score f as a fuzzy number. When fi < f <=fi + 1, f simultaneously belongs and only belongs to that two interval fi and fi + 1.The certainty degrees that f belongs to fi is $\mu_{fi}(f)$ and to fi + 1 $\mu_{fi+1}(f)$. The digital features of the cloud model's each interval are designated according to experts' knowledge, as shown in Table 2. For the specific total score of each evaluation unit, drive the X-condition cloud generator, make its value as the input of the X-condition cloud generator, and separately calculate its certainty degrees towards its two neighboring qualitative concepts, μ_1 and μ_2, which are used as the frequency of the total score corresponding to the two qualitative concepts to join in the frequency statistics of the two concepts respectively.

Name	The digital features of interval cloud model
About 1	$V_1 = V(1,1/3,0.05)$
About 2	$V_2 = V(2,1/3,0.05)$
…	…
About i	$V_i = V(i,1/3,0.05)$
…	…
About 100	$V_{100} = V(100,1/3,0.05)$ $f \in (99,100]$

Table2. Interval Cloud

The method to draw cloud histogram according to cloud fuzzy frequency is as follows. Make the x-coordinate represent the total scores of appraisal units and Y-coordinate cloud fuzzy frequency. Sign the endpoint of each clear-defined interval on the x-coordinate. Then draw rectangles with the distance d of each interval as the bottom and the cloud fuzzy frequency as the height, and we finally get the cloud histogram. Figure 3 shows an example of cloud histogram using the total score data from the score frequency histogram for commercial land grades in Figure 2.

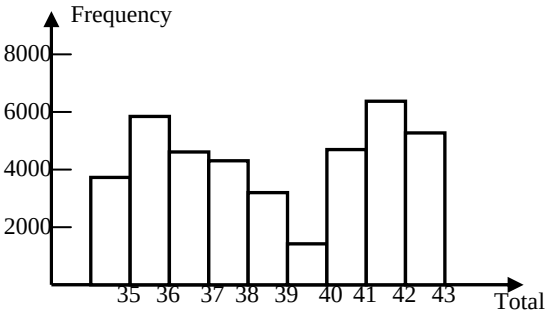

Figure3. Cloud Histogram for Commercial Land Grades

As shown in Figure 3, the cloud histogram diminishes the abnormities in Figure 2, such as multi-peaks and the abnormal critical points between 38 and 39. It indicates that the abnormities in Figure 2 is not caused by the total scores of appraisal units but by the hardware division of total scores. The cloud histogram takes into consider the fuzziness of the intervals and divides the total scores into two adjacent intervals, which is more in line with the reality and makes the clustering results more reasonable. Of course, this method is still a primary

delimitation of land grades. Before we can finally define the land grades boundary, the result of land evaluation must be tested based on either the differentiated land revenue or the market land price.

4. AUTOMATICALLY THRESHOLDS (CRITICAL POINTS FOR LAND GRADES)-SELECTING ALGORITHM

The score frequency histogram method reflects the statistical features of the distribution of the total score and is the conventional method for land grades delimitation. Generally speaking, the valley points of the histogram can be chosen as the thresholds. The most obvious feather of valley points is that the frequencies of the points are locally smallest, while the feather of peak points is just opposite. In practical work, however, there always exist a few random interference points that meet the feature of valley points and peak points due to complexity of land price field or noise, which makes it difficult for the computer to recognize the valley points and peak points automatically. Certain measures must be taken to eliminate these disturbances before detecting the valley points and peak points automatically. As a result, how to promptly and accurately recognize the inference points and detect the valley points and peaks points has always been a common interest of researchers. In this section, we adopt a simple algorithm to detect the valley points and peak points automatically, based on the cloud histogram, and achieve a satisfactory result.

The number of urban land grades depends on the city's scale and complexity as well as the type of land grades. According to the urban land grades regulation, the number of land grades is determined as indicated in Table 3.

Consider Pi as the corresponding score frequency for the total score fi. Compare the frequency Pi with frequencies Pi-1 and Pi+1, which are the corresponding score frequency for the two neighboring total scores of fi, f i-1 and f i+1, respectively. The results of the comparison are as follows:

If P i < P i- 1 and P i < P i+ 1, the point fi is a valley point;

if P i > P i- 1 and P i > P i+ 1, the point fi is a peak point.

The recognition process seems to be cursorily in practical work, as the selected valley points or peak points may not be reasonable ones and may include pseudo valley points or peak points. In the following, we designed the automatic detecting algorithm of valley points and peak points of the histogram, based on the given number of land grades determined by the regulation and urban scale.

(1) Determine the number of land grades K, based on the urban land grades regulation and the scale of the city.
(2) Scan the cloud histogram from left to right to obtain the set of primary valley points f{f1,f1,……,fn}, based on the definition of valley points; the beginning and ending scores are also included in the set f, with f_1 = beginning score and f_n =ending score.
(3) If the number of primary valley points n equals to K+1, all the primary valley points in the set are considered as critical points for land grades, and the score intervals of land grades are defined as: 1st-Class [f_1, f_2], 2nd-Class （ f_2, f_3],……kth-Class （ f_{n-1}, f_n];
(4) If the number of primary valley points n exceeds K+1, calculate the absolute distance d_i between each valley points in the subset { f_2, f_3 ……, f_{n-1} } of the set of primary valley points { f_1, f_2, ……, f_n } ($d_i = f_i - f_{i-1}$). Find out the minimum distance $d_{\min}$ in all the distance calculated, as well as its two corresponding valley points f_i and f_{i-1}. Keep one point of the two whose frequency is smaller, and delete the other point from the primary valley points set.
(5) Turn to step (4), until the number of primary valley points n equals to K+1.
(6) All the primary valley points in the set are considered as critical points for land grades.

5. CONCLUSION

This paper analyzes the drawbacks of the score frequency histogram method for land grades clustering, tries to integrate the cloud model with the frequency histogram, and develops the land grade clustering method based on cloud model and histogram. Based on the cloud model theory, this method turns the traditional clear interval data into cloud, draws the cloud histogram according to the cloud fuzzy frequency of each appraisal unit, and diminishes multi-peaks and abnormity in critical points in traditional histogram. The cloud histogram takes into consider the fuzzy of data intervals, divides the total scores into two adjacent intervals, which is more in line with the reality and makes the clustering results more reasonable and makes the land grades clustering results more reasonable and practical. In the meantime, based on the drawing of the cloud histogram, this paper also advances the automatic critical points-selecting algorithm, which promptly and accurately recognizes pseudo valley points and peak points, detects the valley points and peaks points in the histogram, and achieves satisfactory results.

Urban Scale / Types of Land Grades	Large City	Medium-sized City	Small City
Integrated Land Use	5~10 Grades	4~7 Grades	3~5 Grades
Commercial Land Use	6~12 Grades	5~9 Grades	4~7 Grades
Residential Land Use	5~10 Grades	4~7 Grades	3~5 Grades
Industrial Land Use	4~8 Grades	3~5 Grades	2~4 Grades

Table3. The number of land grades

ACKNOWLEDGMENT

Financial supports from the National 863 Program of China(No.2007AA12Z225) and the "11th Five-Year Plan" of National Scientific and Technological Supporting Project(No.2006BAJ05A02) are highly appreciated.

REFERENCES

Li Deyi. Artificial Intelligence with Uncertainty[M]. Beijing:National Defence Industry Press,2005

Li Deren, Wang Shuliang and Li Deyi. Spatial Data Minning Theories and Applications, Science Press, Beijing,2006.

Yan Xing and Lin Zengjie.Real Estate Appraisal[M].China Renmin University Press, Beijing,1999.

Wang Shuliang. Data Field and Cloud Model Based Spatial Data Mining and Knowledge Discovery[D].Wuhan: Wuhan University,2002

Di Kaichang. Spatial Data Mining and Knowledge Discovery[M]. WuHan:WuHan University Press,2003.

Li Xingsheng. Study on Classification and Clustering Mining Based on Cloud Model and Data Field[D]. Beijing: PLA University of Science and Technology,2003.

Wang Xinzhou,Zou Shuangchao,Zou Jingui and Hua Xianghong. Fuzzy Histogram and Its Appl ication to Surveying Data Processing[J]. Geomatics and Information Science of Wuhan University,2004, 29(5): 385-388

Lu Rong,Sheng Yi. Image threshold segmentation method based on an improved 2-D histogram[J].Systems Engineering and Electronics, 2004,26(110): 1487-1490

About the first author: HU Shiyuan, associate professor, PH.D, majors in land evaluation and spatial data mining.
E_mail: shiyuanhu@ sina.com