Research on the Crime of Burglary Based on Cegra Clustering Algorithm

Tuo Shi[1],*, Xiaolan Liu[2], Gao He[1]

[1] Beijing Police College, Beijing, China
[2] Hengshui College, Hengshui, China

Abstract: The crime of burglary has not only been one of the factors affecting the stability of public security, but the tough case that the public security organs face. Due to the difference of spatial environment factors, the criminal law and the characteristics in the space area always show different trends, which means the process of gradual differentiation and relative concentration of the criminal space under the control of this difference. And this kind of differentiation and concentration is not only represented in spatial locations, but the space area caused by the criminal environment. And on the basis of relevant theories, this paper adopts CEGRA clustering ensemble algorithm to divide the criminal spatial differentiation pattern, in order to better explore the law and characteristics of burglary crime under the criminal space differentiation pattern. And the related experiment result has fully validated the good application effect of this method in the criminal spatial differentiation, especially the obtained conclusion can fully reflect the impact of the criminal environment on the crime of burglary, which can have a strong guiding significance to the specific practices.

Keywords: Cegra; Criminal spatial differentiation; Clustering ensemble; Environmental factor

1. Introduction

The crime of burglary has not only been one of the factors affecting the stability of public security, but the tough case that the public security organs face. And this kind of non-criminal activity would constantly change the way and means of committing crimes along with the progress of society and technological innovation, in order to avoid being captured, which has also become a tough battle with the characteristics of the higher occurrence frequency and the lower possibility to the solution for the public security organs. In essence, this research can effectively provide more guidance for the further practices and better improve the ability to prevent the burglary by exploring its underlying causes, conducting the corresponding predicting and warning based on the case characteristics, and what's more, it can even put forward the reasonable prevention countermeasures for the public security organs to promote the efficient allocation of police resources.

Founded by the representative Shaw and Mcka from Chicago school, social disorganization theory is the earliest and most classic theory of criminal geography, which is originally from the research on the juvenile crimes in Chicago. Specifically, on this basis, it has been found that the place with the highest crime rate is located in the degraded area close to the central business district, and the gradual closer to the city center can further help prove the close relationship with the expansion pattern of concentric cities. And since then, there have been a large number of the relevant empirical researches, thus leading to the further development of those theories. In 1989, Sampson and Groves started the exploration at the medium micro level through validating this theory from the perspective of the community [1]. And in the later period, many scholars have tried to verify this theory from many different perspectives of population, income, family, social voluntary, neighborhood relationship and urban process, and among them, the relatively influential research result is the famous assertion of "social environment creating crimes, but criminals only becoming a tool for the crimes " proposed by Ketterer [2]. And there are countless researches on the relationship between spatial environment factors and crimes in China, and the scholar Zheng Wensheng studies the law of the total amount of urban crimes in China along with time [3]. Xu Xiao verified the existence of time and space clustering in burglary cases on the basis of analyzing the regional distribution of burglary cases, as well as the space and time clustering mode in Jiang'an District of Wuhan City [4]. Lu Juan analyzed the spatial distribution characteristics of crime rate of burglary, theft of motorcycles and non-motor vehicles in Beijing through the Moran index, which can provide an important means for quickly understanding the spatial distribution pattern of crimes [5]. Through the analysis on the crime rate, crime density, time and spatial distribution law of burglary in the eight districts of Hangzhou, Chen Man found that such kind of crimes was mainly affected by traffic flow, housing type, anti-theft system and many other factors [6]. Many scholars have studied the law of crimes from the perspectives of space and the integrated environmental factor, in order to explain the mechanism of crimes, and to further predict the trends. In view of that, this paper mainly carries out the spatial differentiation research of burglary from the perspective of social environmental factor, and further adopts the clustering ensemble technology to study and predict the crimes of burglary under the spatial differentiation pattern, in order to better provide support and reference for the prevention and control of the burglary.

2. Criminal Spatial Differentiation Based on Environmental Factors

In the previous studies, it can be found that most scholars qualitatively discuss the crimes from the perspective of functional spatial division when discussing the differentiation of criminal space. For example, Wang Fazeng divided the urban space into a downtown area, a living area, an industrial area, an administrative area, a cultural area, an external transportation area, a storage area, a recreation area, a suburban area, as well as the urban road system and green land system crossing or infiltrating in each area, and he believed that the difference, centripetal and functional characteristics in spatial environment differentiation indeed have an important impact on urban crimes [7]. And based on the related theories of criminology and psychology, Liu Dan also studied the defense and safety of the functional composite space in Changchun. Specifically, she summarized the functional composite space of Changchun as four types according to the spatial form formed by the relationship between architecture and road layout, mainly including linear functional composite space, mesh functional composite space, decentralized functional composite space and mixed functional composite space, which can effectively provide new ideas for the defense and safety design of urban space [8].And taking the geographical location attribute of criminal space as the research object, some other scholars carry out the research on spatial differentiation and concentration from the perspective of geographic information processing of spatial correlation.

The criminal spatial differentiation refers to the criminal law in the spatial area due to the differences of spatial environmental factors, making the characteristics of the crimes show different trends, which means the gradual differentiation and relative concentration process of the crime space under the control of this difference. And this kind of differentiation and concentration is not only represented in spatial locations, but the space area caused by the criminal environment. In general, the overall framework and the trend of spatial shape and environmental differences are relatively stable in a certain historical period though the differentiation pattern of spatial environment has been in the process of continuous changing and development. In other words, the impact of environmental factors on the differentiation of criminal space is relatively stable and continuous, meaning the constant existence of its differentiation and concentration.

Based on the above analysis, it can be found that different environmental factors have created different criminal spaces, while different criminal spaces also have different criminal characteristics and laws. And in this way, it may inevitably lead to some criminal spaces becoming the high-risk area with different types of crimes, thus forming the “sources” of crimes. And those areas with many crimes, which are also called hot spots, often have many commonalities and similarities with the spatial environmental factors.

3. Cegra Algorithm

Based on the CEGRA algorithm that is the clustering ensemble algorithm of gray absolute correlation degree, this paper divides the criminal spatial differentiation pattern, and then calculates the gray absolute correlation degree between data objects. After obtaining the calculation result, the paper further transforms the matrix and the gray link matrix into a weight relationship matrix between the data objects by utilizing the gray link matrix mathematically to describe and process the relationship between the data and the cluster, in order to use the method of multiple consensus function to acquire the final clustering results [9]. And the overall process is clearly shown in Figure 1.

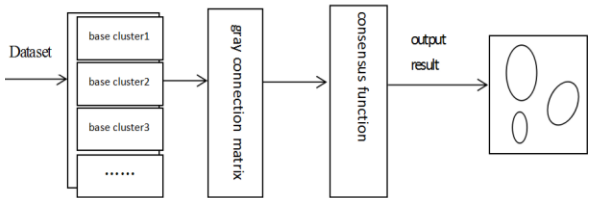

Figure 1. Clustering ensemble process

Through applying this algorithm to conduct the cluster analysis of space based on environmental factor data, the paper can obtain the convergence and differentiation results of crime spaces. Specifically, on the basis of the differentiation pattern, the paper carefully analyzes the criminal law and characteristics of each different zone group, and then makes the explanation to the crime phenomenon and the induced cause in the crime space by combining with the realities.

4. Experiment and Result Analysis

Taking the burglary data of a city in China as an empirical research case, this paper selects 49 indicators that reflect social environmental factors (the specific contents are shown in TABLE I). And these environmental factors describe five-dimensional characteristics of space economy, population, transportation, social supporting resources and geographic attributes, which can fully and comprehensively reflect the social environment in the space.

Considering that the data indicators being selected should timely conduct the clustering, the K-means++ algorithm should be adopted to further determine the number of clustering clusters due to the lack of supervision for this data set.In this paper, the determination of the number of clusters mainly includes two aspects: the first aspect is to the set number of clusters between 5-9 according to the experts experience, while the second step is to set different parameters ranging from 3 to 18 based on the traversing method in this paper. And in this way, the final number of clusters can be determined by investigating the silhouette coefficients of the clustering results of the CEGRA algorithm.

Table 1. Criminal spatial clustering indicator

number of legal entities _ wholesale and retail industry
number of legal entities _ transportation, warehousing and postal services industry
number of legal entities_real estate industry
sales expense
business tax and surcharges
total energy
expenditure on staff
labor costs
number of subway stations
liquefied petroleum gas
number of legal entities_resident service, repairs and other services
number of bus stops
number of bus lines
public utility expenditure
total liabilities
number of legal entities_medium sized
number of legal entities_small sized
total profit
number of floating population
number of resident population
number of registered population
number of legal entities_construction industry
number of legal entities_education
number of legal entities_public management, social security and social organization
number of legal entities_domestic investment
number of legal entities_Hong Kong, Macao and Taiwan investment
number of legal entities_central authority
number of legal entities_local authority
the final number of people engaged_manufacturing industry
the final number of people engaged_electricity, gas and water production and supply industry
the final number of people engaged_wholesale and retail industry
the final number of people engaged_ transportation, warehousing and postal services industry
the final number of people engaged_accommodation and catering industry
the final number of people engaged_real estate industry
the final number of people engaged_education
the final number of people engaged_health and social work
the final number of people engaged_public management, social security and social organization
the final number of people engaged_domestic funded
the final number of people engaged_Hong Kong, Macao and Taiwan investment
the final number of people engaged_foreign investment
the final number of people engaged_large sized
operating profit
total welfare amount
total electricity consumption
total coal consumption
total gas consumption
total petrol consumption
total kerosene consumption

Silhouette coefficient is not only an effective method to evaluate the clustering effect, but a dimensionless measurement [10], which means it can use two indicators of cohesion and resolution to measure the clustering effect.And the silhouette coefficient is often used to evaluate the effects of different algorithms or their different operating modes on the clustering results on the basis of the same raw data. The specific calculation step for the silhouette coefficients is as follows:

- Calculating the average distance from the data sample i to all other samples in the same cluster. The smaller value a_i indicates that the sample should be divided into this cluster, and in this way, a_i should be referred to as the intra-cluster dissimilarity of the sample, and the a_i mean of all samples in the cluster is defined as the dissimilarity of the cluster C .
- Calculating the average distance b_{ij} from a sample i to all samples of another cluster C_j , and defining as the degree of dissimilarity between the sample i and the cluster C_j ,which is called intra-cluster dissimilarity of the sample i . For the $b_i = \min(b_{i1,} b_{i2},...,b_{ik})$,the larger value of b_i indicates that the sample i does not belong to other clusters.
- The silhouette coefficient of sample i can be defined as the Equation 5-19 according to the intra-cluster dissimilarity a_i and inter-cluster dissimilarity b_i of the sample i :

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

$s(i)$ that is closer to 1 indicates the more reasonable characteristics for the sample i being divided into the cluster, while $s(i)$ that is closer to -1 indicates that the sample i should be classified into another cluster. And $s(i)$ approximated to 0 shows the sample i is on the boundary of the two clusters and subordinate to any cluster.

- $Sihouette\_Score$ is the mean of all samples $s(i)$, and the calculation formula is shown in 5-20:

$$Sihouette\_Score = \sum_{i=1}^{n} s(i) \tag{2}$$

And $Sihouette\_Score$ is the silhouette coefficient of the results of clustering, which is an effective measurement to evaluate the efficient and reasonable characteristics of clustering algorithm, meaning the larger value of the silhouette coefficient can bring the better clustering effect. And the silhouette coefficient can be used in the supervised learning and unsupervised learning situations.

In the experiment of this chapter, it is necessary to set two parameters based on the criminal spatial clustering division of environmental factors, including the number of clusters k of clustering and the number of base cluster

M . And in order to obtain the best parameter setting effect, this paper adopts the grid search method for parameter tuning, which can effectively optimize the model performance by traversing the given parameter combination, and the evaluation criteria of parameter optimization mainly refers to the silhouette coefficient, setting separately as $k = [3,4,5,6,...,18]$ and $M = [10,20,30,40,50]$. On the basis of finding the highest parameter settings with the best silhouette coefficient after 50 experiments under each combination of parameters, the optimal parameter k=6, and the best silhouette coefficient value M=30 can be determined and obtained, thus getting the best Sihouette_Score=0.96 And the results of the division are shown in Table 2 below:

Table 2. Division results of each label

Category of label	Number of sample
0	82
1	149
2	28
3	26
4	10
5	20

The criminal space can be divided into six categories through experiments according to the level of spatial environmental factors, and the specific classification is as follows:

Category 1: Such kind of criminal space involves the lowest level of space environment traffic factor, which is reflected in the smallest number of subway stations, the bus stations and routes, as well as the poorer public transportation level, and such areas generally belong to those underdeveloped areas of public transportation. The lowest economic factor level is reflected in the the lowest level of the regional total profit, total personnel expenditure, total regional sales expenses, business tax and surcharges that are far lower than other categories, including the lowest level of the regional income level, consumption level and taxation ability. The population factor also has the lower level, and the number of permanent residents and the number of registered households are higher than that of floating population in the space, indicating that the space in such areas is less attractive to foreigners, and further leading to the less population mobility. What's more, the level of the social resource factor is also the lowest, which is not only reflected in the lower employment level and the living facilities for residents, but the relatively scarcity of social resources in this kind of criminal space, in this way, the number of such space crimes is also at the lowest level.

Category 2: Such kind of criminal space transportation factors are underdeveloped and the public transportation resources are at a low level. The economic factor level is also at a lower level, which is reflected in the regional operating profit, total profit, total regional sales expenses and personnel expenditures, and especially the slightly higher debt can fully reflect the lower level of consumption in the regional space, the tax amount payable and the economic level. For the the population factor, the number of permanent residents and the number of registered households are higher than that of floating population, reflecting that the relatively fixed population and the lower population mobility. The overall level of social resource factors is at the middle lower level, specifically reflecting in the employment resources, living facilities, and energy consumption. Therefore, the number of such cases in the area is at a lower level.

Category 3: Such kind of criminal space transportation factors are relatively developed, and the level of public transportation is at the moderately lower level. The economic factor is also at a general level, reflecting in the moderately lower level of the regional operating profit and personnel expenditure, as well as the and the relatively general economic level. However, the floating population in such areas is much larger than the resident population and the registered population, but the number of permanent residents and floating population is always at the middle lower level. What's more, the level of social resource factors, including the employment and living facilities, is generally lower, and leading to the general level of the overall resources. Therefore, it should belong to the area to be developed, making the lower level of the number of cases.

Category 4: Such kind of criminal space transportation factors are relatively developed, and public transportation resources are relatively abundant, and such areas generally belong to those more developed areas with higher level of public transportation. The economic factors is also at a relatively developed level, which is reflected in the higher levels of regional operating income, total personnel expenditure and regional sales expenses. The population factor is relatively at the higher level, especially the number of permanent residents is the highest, while the number of registered population and floating population is relatively fixed and stable. The overall level of social resource factors is relatively high, making the abundant employment resources and the high level of social support. Therefore, this category belongs to the cross region of production and residence or the urban areas at the development and construction level, making a high level of the number of cases.

Category 5: Such kind of criminal space transportation factors are relatively developed, and public transportation resources are relatively abundant, and such areas generally belong to those relatively developed areas with higher level of public transportation. The economic factors is also at a relatively developed level, which is reflected in the higher levels of regional operating income, total personnel expenditure and regional sales expenses with the third ranking. The population factor is always at the relatively higher level, in which the number of permanent residents is at a middle level and the number of floating population is usually large. What's more, the overall level of social resource factors is higher, which is reflected in the larger number of employment enterprises and the higher level of living resources, including the catering, services and energy supply. Therefore, this category belongs to the urban-rural integration area, making the relatively high level of the number of cases.

Category 6: Such kind of criminal space transportation factors is the most developed, and the public transportation resources are abundant with the highest values, therefore, such areas generally are developed areas with the higher level of public transportation. The economic factors is also at the most developed level, which is reflected in the regional operating income, total personnel expenditure, total regional sales expenses and total tax payment with the first ranking. The population factor is relatively at the high level, in which the resident population is at a higher level, the number of the household registration population is the largest that is slightly higher than that of the floating population, while the number of the floating population is in the second level among all spatial types, indicating the higher personnel mobility. The social resource factors is generally at the developed level focusing on energy consumption, especially the highest level of the real estate departments, liquefied petroleum, accommodation, catering, wholesale and retailing, mainly involving the upper-middle level of resident consumption. In view of that the criminal space has the small area and this category also should belong to the developed commercial area, the number of cases is also at the highest level.

In addition, in order to compare the performance advantages of the CEGRA algorithm adopted in this paper on the differentiation pattern division of criminal space, the paper mainly uses silhouette coefficient to be the evaluation standard of the unsupervised learning, and the specific experimental results are shown in Figure 2.

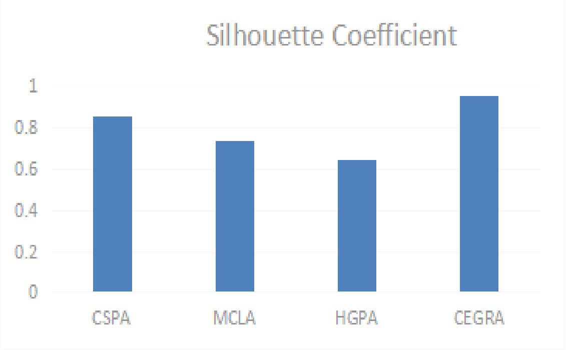

Figure 2. Comparison of cegra algorithm result in the data set of this paper with that of other algorithms

Based on the number of clusters $k = 6, M = 30$ in the above figure, the CEGRA algorithm used in this paper has the best clustering effect on the crime data set of burglary, which means the highest clustering silhouette coefficient can show the excellent clustering performance.

5. Conclusion

Based on the crime theory of burglary induced by the space environment factor, this paper conducts the research on the spatial differentiation of crimes by taking the relevant environmental factors as indicators to classify the crime space, and further adopts the clustering method to gather and differentiate criminal space, in order to break the overall analysis model of all the data in the past. In addition, the paper also carries out the separate crime analysis for each type of criminal space after differentiation, which not only can provide support for revealing the law of crimes, but make the crime space analysis become more targeted rather than simply analyzing the overall space. What’s more, in order to complete the criminal spatial division more reasonably, the paper adopts CEGRA clustering ensemble algorithm, which can form good clustering effect, better robustness and flexibility after the empirical verification, thus it can be widely used in the research of criminal spatial differentiation, and the analysis results also have strong practical guiding significance.

Acknowledgment

This work was supported in part by a grant from SJ2019 Technology Application 03 Project, and Beijing Police College Key Scientific Project 2019KZD42.

References

[1] Robert J. Sampson, W. Byron Groves. Community Structure and Crime: Testing Social-Disorganization Theory. American Journal of Sociology, 1989, 94(4): 774-802.

[2] Quetelet A. Adolphe Quetelet's Research on the propensity for crime at different ages. Anderson Pub. Co. 1994.

[3] Zheng Wensheng, Zhuo Rongrong, Luo Jing, et al. Research On the "Robbery, Grabbing and Theft" Crime Distribution Environment in Wuhan City Based On Spatial Syntax. Acta Geographica Sinica, 2016, 71(10): 1710-1720.

[4] Xu Xiao, Zhu Jingwei, Wu Ling, et al. Research on the Analysis of the Time and Space Pattern of Urban Burglary. Geomatics & Spatial Information Technology, (8): 15-18.

[5] Lu Juan. Analysis of the Global Distribution Model of the Crime Rate of Burglary Based on Moran Index. Police Technology, (1): 45-46.

[6] Chen Man, Wu Yiwen, Yan Jiaqi. Analysis of Temporal and Spatial Distribution of Burglary Crimes in Hangzhou from 2014 to 2015. Geospatial Information, 2017, 15(12).

[7] Wang Fazeng. Research on Comprehensive Management of Urban Criminal Space Blind Zone. Geographical Research, 2010, 29(1): 57-67.

[8] Liu Dan. Research on Space Functional Defense Safety Design of the Urban Defense. Jilin Jianzhu University, 2016.

[9] Shi T, Jiang W, Luo P. A Method of Clustering Ensemble Based on Grey Relation Analysis. Wireless Personal Communications, 2018.

[10] Kaufman L, Rousseeuw P J. Finding Groups in Data: An Introduction to Cluster Analysis. DBLP, 1990.