Spatial Analysis of Villages in Jilin Province Based on Space Syntax and Machine Learning

Deli Liu[✉] and Keqi Wang[✉]

School of Architecture and Planning, Jilin Jianzhu University, Changchun, China
`1661654037@qq.com, wkq0431@126.com`

Abstract. The development of machine learning technology gives architects and urban planners a new tool that can be used for research and design. The topic of this paper is to analyze the rural space of Jilin Province with the machine learning algorithms and space syntax theory, and to obtain the inherent formation and development laws of rural spatial forms, which can be used as a reference and evaluation system for subsequent rural development, and also can emphasize the locality and continuity of rural development. First, based on geographic information data, researching the connection between the distribution of villages and geographic data at a macro level and to classify them. Then, from each category, selecting one township and use all villages in its area as samples for the more specific study. Spatial features of individual village are extracted based on space syntax theory, and representative spatial features which can as feature values for cluster analysis are selected through comparative analysis. Then classify villages from high-dimensional data and explore their type characteristics. Finally, we hope the result of this study can help provide useful theoretical references for rural construction and nature conservation in the future.

Keywords: Spatial characteristics · Village · Space syntax · Machine learning

1 Introduction

In China, most cities have reached the level of stock development, while the countryside is still lagging behind. More and more people are concentrating their attention on rural construction in the context of rural rejuvenation. However, the countryside, unlike the city, does not develop from the top down, but more from the bottom up, which is a very natural and slow development process. This means that it is difficult to have a very clear development pattern like the city. The characteristics of the space in village are very implicit and highly influenced by local environmental and cultural factors. As a result, a method distinct from urban design is required in dealing with rural concerns. Otherwise, there would be a slew of issues, including the adoption of a single paradigm in the face of significantly disparate villages, a failure to respect history and natural conditions and so on.

In the process of building the countryside, it is important to follow the objective rules and propose development strategies with a targeted approach. So as a first step we need to have a more accurate understanding of the countryside and be able to grasp its objective laws. However, as mentioned in the previous part of this article, the development of the countryside is a natural process. So, it is difficult to find obvious patterns of the individual countryside, and we need to draw conclusions from a large amount of data in comparison and analysis. When dealing with such a large and complex amount of data, machines are able to be more efficient and can extract information from high-dimensional data in a way that the human brain cannot. In the process of feature extraction of individual villages, space syntax theory is introduced, which can be used to describe the spatial form, and complete the spatial analysis of individual village.
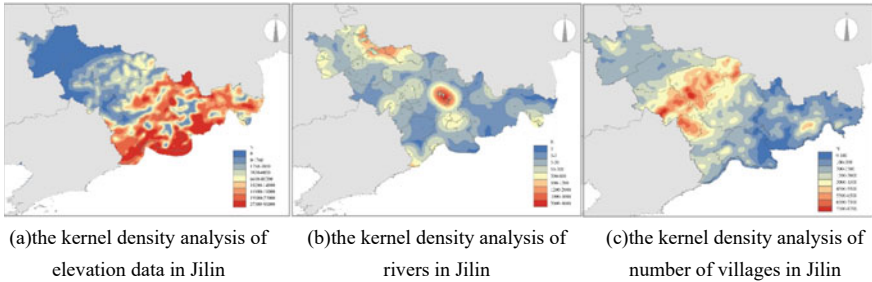
2   Related Work

The theory of space syntax was proposed by Bill Hillier, Julienne Hanson and colleagues at The Bartlett, University College London in the late 1970s to early 1980s. It's a theory about architecture and urban based on the theory of society and space in the earlier the Social Logic of Space [2]. Space syntax theory is mostly used in studies related to urban design [4–6], and its initial object of research is also the city, in order to reveal the'deep structure' in buildings and cities. It illustrates how urban spatial patterns are constituted through spatial laws that link the emergence of particular spatial patterns of the city to other factors [3]. While studying cities and buildings, human behavioral factors are often included in the study. Such as analyzing the connection between spatial structure and human behavior factors. Alabi [1] tried to use the space syntax as a research theory to find out how the urban form and socio-economic factors that influence human behavior patterns, and designers can use the findings to design successful transit and pedestrian-oriented developments. Montello [7] tried to combine space syntax and environmental psychology theories. Space syntax theory can provide a rich and diverse set of quantitative indices for characterizing places in many ways that are potentially relevant to a variety of psychological and behavioral responses.

In China, some scholars have also tried to do research with the Chinese villages as the object based on space syntax theory. Niu et al. [8] use four typical villages in Xiahuayuan District of Zhangjiakou city as samples to obtain the spatial structure characteristics based on space syntax theory, which could provide theoretical basis for the current village plan. Xu et al. [10], taking Jiangshan Hejia-Wujia village in Guchen township, Gaochun county, Nanjing as an example, in terms of Space Syntax, analyzes the spatial characteristics of traditional villages, and they hope to provide new thoughts for systematically protecting and healthily developing of the traditional villages.

However, the sample sizes of these studies are very small and can only reflect the spatial characteristics of individuals, and no common patterns can be derived. Therefore, in this study, we expand the sample size and expect to obtain some implicit common laws through further analysis of results of individual village analysis with machine learning algorithms, which can be used to provide some theoretical references for future rural development.

3   The Geographical Distribution Characteristics of Villages in Jilin Province
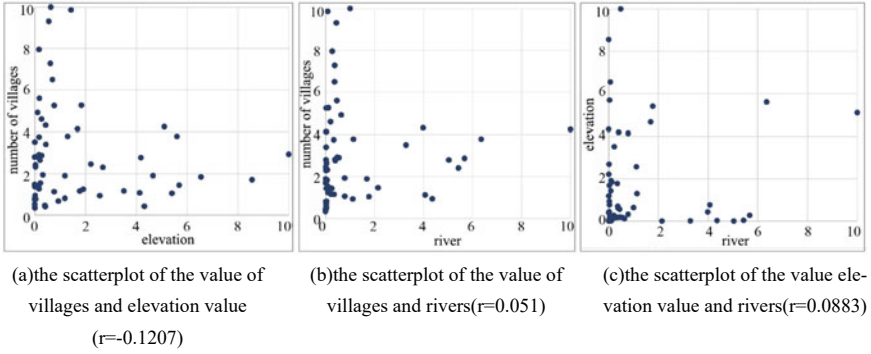
The geomorphic feature of Jilin Province varies greatly. We can obtain the contour data of Jilin Province from China National Catalogue Service For Geographic Information, and the elevation values were used as the population value for kernel density analysis. The analysis result shows that the topography of Jilin Province slopes from southeast to northwest, showing a clear feature of high southeast and low northwest (Fig. 1a). Bound by the central Big Black Mountain, it can be divided into two major landscapes: the southeastern mountains and the central and western plains. The water resources data of Jilin Province were also obtained from China National Catalogue Service For Geographic Information, and the area value of rivers was used as the population value for kernel density analysis (Fig. 1b), which shows that it is mostly concentrated in the central and northwestern regions.

(a)the kernel density analysis of elevation data in Jilin

(b)the kernel density analysis of rivers in Jilin

(c)the kernel density analysis of number of villages in Jilin

Fig. 1. The kernel density analysis of elevation value, rivers and villages in Jilin

According to the kernel density analysis of the village POI data obtained from Amap (Fig. 1c), it can be seen that villages in Jilin Province are mostly concentrated in the central region, and its distribution is less in the southeast, which is strongly related to the topography of the mountain range. The raster data from the kernel density analysis of villages, elevation and water resources were summed up at the county level area of Jilin Province, and the results were mapped uniformly into the [0,10] value domain to draw a scatter plot (Fig. 2). Then the correlation coefficient between the elements is calculated according to Eq. (1). From Fig. 2 and correlation coefficient values, it is clear that the topography of the mountain ranges is more related to the distribution of villages, and the overall trend is that as the number of mountains increases, the distribution of villages decreases.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{1}$$

(a)the scatterplot of the value of villages and elevation value (r=-0.1207)

(b)the scatterplot of the value of villages and rivers(r=0.051)

(c)the scatterplot of the value elevation value and rivers(r=0.0883)

Fig. 2. The scatterplot of the value of rivers, number of villages and elevation data in Jilin

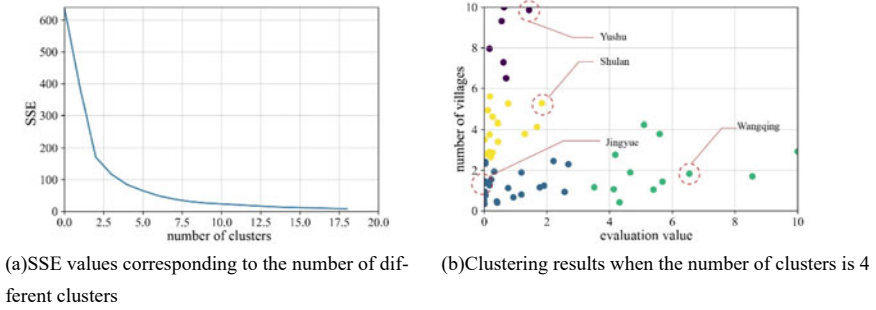## 4 Village Spatial Feature Extraction Based on Space Syntax

4.1 Sample Selection for the Research

From the above correlation analysis, we know that the geographical distribution of villages is closely related to the elevation value, so the elevation value and the number of villages are used as influencing factors to complete the classification analysis, and then one township from each category is selected as the research object. The space syntax is used as the theory to extract the spatial characteristic values of individual village and use them for subsequent analysis. First, the number of villages and elevation data are used as feature values to cluster the county-level regions of Jilin Province, and the K-means algorithm of scikit-learn algorithm library is chosen for the clustering analysis, which can organize the data with similar feature values for classification. When selecting the number of clusters, the SSE value is used as a criterion to evaluate the good or bad clustering results, and different numbers of clusters and their corresponding SSE values are calculated according to Eq. (2) and plotted as a line graph.

$$\text{SSE} = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{2}$$

It can be found that the SSE value has a plummeting inflection point when the number of clusters is 2 and 3 (Fig. 3a), so it will be better to choose 2 or 3 for the number of clusters. However, since a representative from each cluster will be selected for specific analysis at a later stage, in order to increase the sample size, so 4 was chosen as the number of clusters for cluster analysis, and the clustering results are shown in Fig. 3b. One county-level region was randomly selected in each of the four clusters obtained, and one township-wide villages in each region was selected as the study sample. The four county-level regions selected were Jingyue District, Shulan City, Wangqing County, and Yushu City. The townships selected from these four county-level areas are Xinhu Township in Jingyue District, Xihe Township in Shulan City, Baichaogou Township in Wangqing County and Yumin Township in Yushu City. From their satellite maps

(Fig. 4), it can be seen that Baicaogou Townships is located in the mountainous area, Xihe Township is located in the hilly area, while Xinhu Township and Yumin Township are located in the plain.

(a)SSE values corresponding to the number of different clusters

(b)Clustering results when the number of clusters is 4

Fig. 3. Cluster analysis based on elevation value and the number of villages

(a)Xinhu Township   (b)Xihe Township   (c)Baicaogou Township   (d)Yumin Township
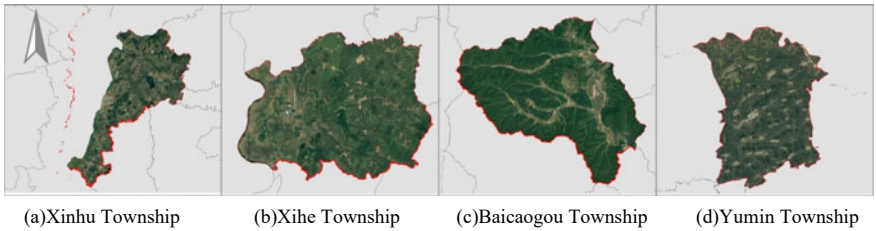
Fig. 4. The satellite map of four Townships

4.2 Spatial Analysis of the Individual Village

The analysis of axial maps in cities can reflect the external spatial structure and there are many studies to confirm this [9]. The axial maps were also used in the study about Chinese villages. Selecting connectivity, integration and intelligibility values as the values of the variables describing the spatial characteristics of the villages. The connectivity value can indicate the number of spaces that are directly connected to the space where that axis is located. The axis with higher integration means that the average topological depth value of that axis reaching all axes of the whole system is lower, indicating its higher accessibility. The intelligibility value can be represented by the R-squared value of the linear fit of the connectivity value of the axes and the integration value. The higher the R-squared value means that the better the connectivity value and the integration value are fitted, which means it is easier for people to infer the overall spatial structure from the local perception, and the intelligibility of the space is higher.

The R-squared value represents the linear correlation between the connectivity value and the integration value of all axes in a village, which is also a value that can be directly used as a feature value for the subsequent village clustering analysis. However, In the

case of integration and connectivity analysis, each axis has a value attached to it, so it is difficult to describe the overall status of individual axial maps. In this case, mean, median, mode, standard deviation and range values are used as the statistical indicators of the list data. And among these representative values, only the mean and standard deviation give a better description of the overall condition of the data.

Four villages were randomly selected and then DepthmapX software was used to analyze them. From the result (Fig. 5), it can be seen that the mean value of connectivity of village-2 is 3.45, which is greater than the mean value of village-4, so the degree of spatial interconnection represented by its axes is higher than that of village-4. The standard deviation value reflects the dispersion of the data, although the average of the connectivity values of village-2 and village-3 are relatively close, the standard deviation value of village-3 is larger than that of village-2, indicating that its connectivity values are not evenly distributed. It is also obvious from the figure that the connectivity values of village-3 appear extreme, with a very obvious main axis, and its range values are also significantly larger than those of village-2. The integration can reflect the accessibility of the spaces where the axes of a village are located. The higher the value of integration, the easier it is to reach other spaces in the village. The standard deviation of the integration of all axes can reflect the difference of how easily to access to different space in the village. In terms of spatial intelligibility, we can find that villages whose R-squared values are closer to 1 have relatively simpler spatial structures. In the case of the four villages, they are sorted by their R-squared values close to 1, and their order is Village 4, Village 2, Village 3 and Village 1. From the axial maps, we can also clearly see that Village 4 has the simplest spatial structure, with a very obvious main axis. It is easier for people to perceive the whole external space in it. While village 1, which has the farthest R-squared difference value from 1, we can see that its space is more diffuse and the whole village shape is more organic.

From the above analysis, we can learn that when using the connectivity and integration values to describe the spatial characteristics, the standard deviation and the mean values are chosen to be more comprehensive because they can reflect all the data very well. The R-squared value, on the other hand, is already the result of the comparison of all the connectivity and integration values, and can also complete the description of the spatial characteristics very well.

5   Clustering of Villages Based on Spatial Features

There are four townships in this study, and axial maps of all villages are depicted by hand. There are 165 effective axial maps, including 27 axial maps in Baicaogou Township, 42 in Xihe Township, 54 in Xinhu Township, and 42 in Yumin Township. The selected spatial eigenvalues are the R-squared values, the mean and standard deviation of the connectivity values and the integration value. The clustering algorithm be selected is also the K-means algorithm, which is able to find classifications with similar values of each eigenvalue. Since all villages come from four different townships and it does not have a clear inflection point for the SSE analysis, the cluster number 4 was chosen as the cluster number, and the clustering results are shown in Fig. 6.
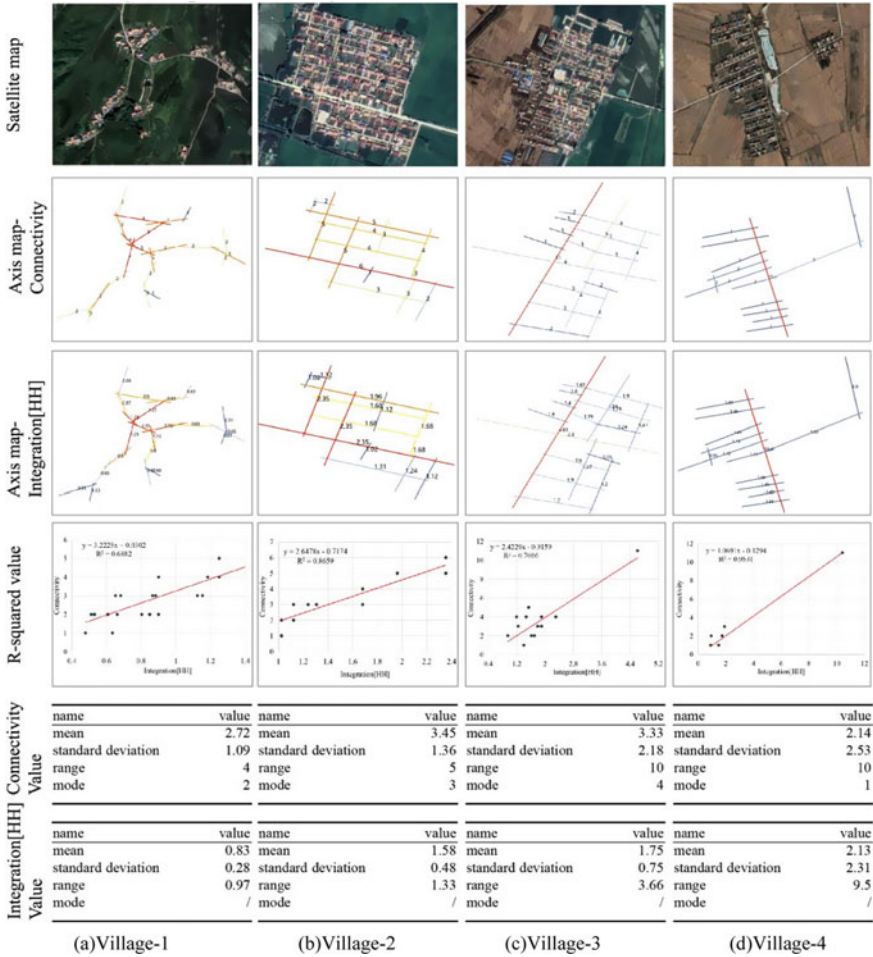
Fig. 5. The four villages and their spatial features

The figure compares four villages across the following rows: Satellite map, Axis map-Connectivity, Axis map-Integration[HH], R-squared value, Connectivity Value, and Integration[HH] Value.

R-squared values:

- Village-1: $y = 3.2221x - 0.5102$, $R^2 = 0.6482$
- Village-2: $y = 2.6476x - 0.7174$, $R^2 = 0.3659$
- Village-3: $y = 2.422x - 0.9359$, $R^2 = 0.7666$
- Village-4: $y = 1.069/x - 0.5294$, $R^2 = 0.9631$

Connectivity Value:

name	(a)Village-1	(b)Village-2	(c)Village-3	(d)Village-4
mean	2.72	3.45	3.33	2.14
standard deviation	1.09	1.36	2.18	2.53
range	4	5	10	10
mode	2	3	4	1

Integration[HH] Value:

name	(a)Village-1	(b)Village-2	(c)Village-3	(d)Village-4
mean	0.83	1.58	1.75	2.13
standard deviation	0.28	0.48	0.75	2.31
range	0.97	1.33	3.66	9.5
mode	/	/	/	/

5.1 Spatial Characteristics of the Villages in Each Cluster

Table 1 shows the countryside feature parameters in each different set of clusters, and it can be seen that among the four clusters, the mean intelligibility in cluster-1 (red) and cluster-4 (blue) is higher than that in Cluster-2 (purple) and Cluster-3 (green). So, it can be known that the spatial structure of villages in cluster-1 (red) and cluster-4 (blue) is simpler than the other two, and it is easier for people to perceive the whole rural space in this cluster. In the eigenvalues of connectivity, there is no great variability in the mean values among the four clusters. In the eigenvalues of integration, Cluster-1 (red) has a higher average value of integration than the other three, indicating that the topological depth between its individual axes is relatively low and there are more connections between the individual spaces. The lowest mean value of integration in
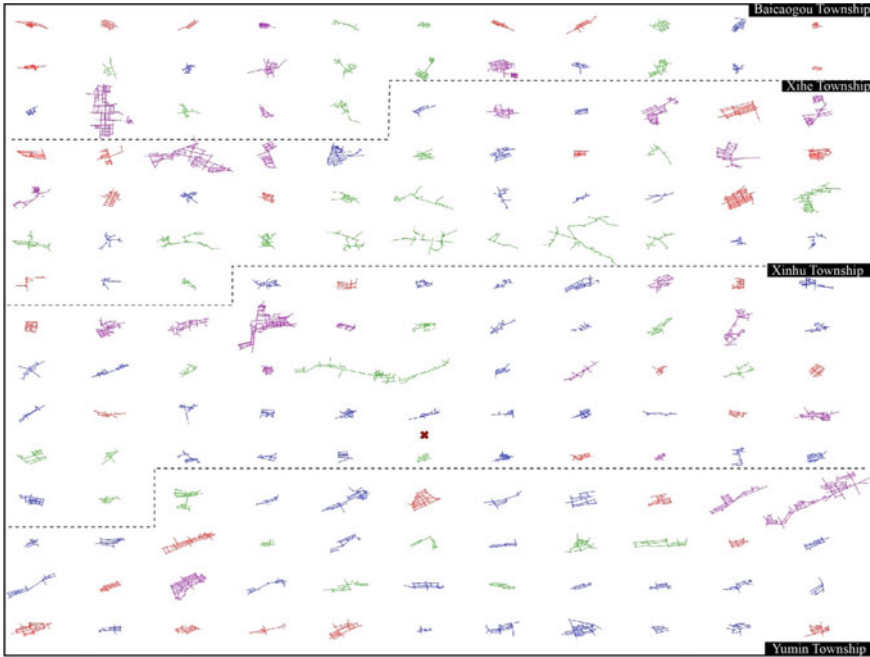
Fig. 6. The results of village clustering. This figure shows the villages sorted by belonging to the same township, with different colors representing the different clusters they belong to

Cluster-3(green) indicates that the spaces represented by each axis are less connected and more diffuse overall.

In summary, these four clustering results have the following characteristics. In Cluster-1 (red), they have high integration and connectivity values. So, the space where the axes are located in the village are more closely connected to each other, and all the spaces in the spatial system where the axes are located reach each other more easily. They also have high intelligibility value, which means the spatial structure of these villages is simpler, and it is easier for people to understand the structure of all the space system. In Cluster-2 (purple), the spatial intelligibility value is low, which means that the axial spatial structure is more complex. We can find that the central settlements in all four townships are in this classification, and their scales are larger compared with other villages. They have a higher integration values, it is means the space where the axes are also located in the village are more closely connected to each other.

In Cluster-3(green), they have the lowest intelligibility, integration and connectivity values. Therefore, the spatial structure of this type of countryside is the most complex and difficult to understand, as can be seen from Fig. 6, which shows that the axial maps of this type of countryside are more natural and disorderly compared to the other three classifications. The spatial intelligibility is higher in Cluster-4 (blue) and its spatial structure is relatively simple compared to Cluster-2 (purple) and Cluster-3 (green). Its connectivity value is not high, indicating that the direct connection between each axis is also weak, and it can be found that the number of axes in each axial map in this cluster

Table 1. The spatial feature values of the four clustering results

		Cluster-1 (red)	Cluster-2 (purple)	Cluster-3 (green)	Cluster-4 (blue)
Intelligibility	Range	[0.691,0.980]	[0.239, 0.632]	[0.116, 0.582]	[0.569, 0.931]
	Mean	0.836	0.459	0.398	0.764
Connectivity_ mean value	Range	[2, 4.385]	[2.750, 3.887]	[1.867, 3.415]	[1.867, 3.533]
	Mean	3.191	3.252	2.577	2.717
Integration[HH]_mean value	Range	[1.419, 2.1259]	[0.974, 1.671]	[0.45, 1.079]	[0.785, 1.442]
	Mean	1.67894	1.211644	0.824427	1.163224
Connectivity_ standard deviation	Range	[1.099, 2.941]	[1.187, 3.304]	[0.667, 1.762]	[0.786, 1.795]
	Mean	1.772424	1.879552	1.153918	1.300722
Integration[HH]_ standard deviation	Range	[0.371, 2.312]	[0.240, 0.431]	[0.117, 0.372]	[0.247, 0.571]
	Mean	0.716854	0.334469	0.222923	0.394921

is low. If we arrange these four clusters according to the degree of simpler or regular spatial structure and similar morphology of the villages in the cluster. Then the order should be Cluster-1 (red), Cluster-4 (blue), Cluster-2 (purple) and Cluster-3 (green).

5.2 Conclusion Based on Comparative Analysis of the Spatial Characteristics of Villages in the Four Townships

In these four townships, the ratio of these Cluster-1 (red),Cluster-2 (purple), Cluster-3 (green) and Cluster-4 (blue), four types of villages in Baicaogou Township is 0.297: 0.185: 0.333: 0.185, the ratio in Xihe Township is 0.214: 0.167: 0.333: 0.286, the ratio in Xinhu Township is 0.148: 0.185: 0.185: 0.482, and the ratio in Yumin Township is 0.238: 0.071: 0.167: 0.524. In all villages, the ratio of this four type villages is: 0.212: 0.152: 0.242: 0.394. In Baicaogou Township and Xihe Township, the proportion of villages of cluster-3 (green) type is higher, while the largest number of all villages in these four townships is the cluster-4 (blue) type. We know that the spatial feature of villages in cluster-3 (green) have the minimum mean value of Intelligibility and Integration[HH]_mean value, which means that these villages have a complex spatial structure, and they have a variety of spatial forms with natural patterns and fewer regularities.

In Xinhu Township and Yumin Township, the number of villages belong to cluster-4 (blue) is larger than the other types of villages, which means the plane of villages in this two cluster is more likely controllable and less affected by geography. In the previous part of this article we said Baicaogou Township is in the mountainous area, Xihe Township is located in the hilly area, while Xinhu Township and Yumin Township are located in the

plain area. In fact, it is not difficult to imagine that in the mountainous and hilly areas, the plan of villages is more natural and the spatial structure is more disorderly due to the geographical topography. Villages located in the plain area, on the other hand, are relatively more regular in their planar texture, less constrained by geography and more controlled by human factors.

6 Conclusion

In this paper, we uses space syntax and machine learning algorithms to complete the analysis of the spatial characteristics of villages in Jilin Province, to find patterns and draw conclusions from the high-dimensional data, so that we can better understand the characteristics of rural spatial characteristics and apply them to the subsequent work on villages. It is worth noting that this analysis method does not yield unique results. For example, when changing the number of clusters in the final cluster or replacing the feature values, or weighting the feature values, the clustering results will become different, and different conclusions can be drawn based on the analysis of the clustering results. This is the characteristic of using machine learning algorithms to complete the analysis, it does not filter for us subjectively, while it just analyzes and outputs the results without any preference. When we change the input parameters, the output results will also be diverse. Algorithms and theories just help us to have more and deeper understanding of the object we are analyzing when we use them, and finally it is up to us to interpret and summarize the diverse results to get the information we need. This corresponds to the statement in the abstract that machine learning algorithms only provide a new analysis or design tool. They do not make decisions for us, but provide us with new references when we making decisions. Finally, this is also an attempt to apply space syntax theory and machine learning algorithms to the study of the countryside in question. Although there are still some shortcomings, however, we still hope it can give new reference and inspiration to researchers in the same field.

References

1. Alabi MO (2021) Space syntax: evaluating the influence of urban form and socio-economy on walking behaviour in neighbourhoods of Akure, Nigeria. Urban, Plan Transport Res 9(1):579–597
2. Hillier B, Hanson J (1989) The social logic of space. Cambridge University Press
3. Hillier B (2007) Space is the machine: a configurational theory of architecture. Space Syntax
4. Hillier B (1999) The hidden geometry of deformed grids: or, why space syntax works, when it looks as though it shouldn't. Environ Plann B Plann Des 26(2):169–191
5. Jiang B, Claramunt C, Klarqvist B (2000) Integration of space syntax into GIS for modelling urban spaces. Int J Appl Earth Obs Geoinf 2(3–4):161–171
6. Karimi K (2012) A configurational approach to analytical urban design:'Space syntax'methodology. Urban Design Int 17(4):297–318

7. Montello DR (2007) The contribution of space syntax to a comprehensive theory of environmental psychology. In: Proceedings of the 6th international space syntax symposium, pp 012.1–012.14
8. Niu MJ, Xu WG, Li YX (2020) Quantitative research and analysis of the rural settlement structure of Xianhuayuan district based on space syntax. Contemp Architect 2020(06):138–140
9. Penn A (2003) Space syntax and spatial cognition: or why the axial line? Environ Behav 35(1):30–65
10. Xu H, Zhao HS, Liu F (2016) Preliminary study in syntactical of the spatial form of traditional villages: a case study of Jiangshan Hejia-Wujia village in Guchen Town, Gaochun County, Nanjing. Modern Urban Res 2016(01):24–29