Research Article

Yingqiao Wang*

Research on territorial spatial planning based on data mining and geographic information visualization

https://doi.org/10.1515/nleng-2025-0090
received August 15, 2024; accepted February 19, 2025

Abstract: At present, big data and mining technology has been more and more applied to urban planning practice. Based on data mining technology and geographic information visualization, this article puts forward a study on the optimization of territorial spatial planning. This article selects the clustering algorithm and decision tree in data mining technology, combines them organically, and puts forward spatial data mining. Cluster analysis is used to conduct spatial cluster analysis of comprehensive spatial data to explore whether there is spatial correlation and large-scale contiguous data in the data; if so, it can realize the purpose of national territorial space division. According to the analysis of the relevant data of land use planning and the characteristics of the business process, the overall design of the land use planning decision support system was carried out. The preliminary effects of land use planning clustering were generated through the usage of the *K*-central factor clustering algorithm, and the consequences with excessive classification accuracy have been chosen as coaching samples mixed with the baseline of time series. Finally, the *K* nearest neighbor classification algorithm primarily based on dynamic time warping is used to classify and perceive the city's practical place again. With the useful resource of point of interest data, the remaining city characteristic identification result is obtained. Finally, the simulation analysis shows that the area under curve (AUC) value of seven types of land use is 0.9376 for cultivated land, 0.85442 for forested land, 0.81747 for grassland, 0.8708 for water area, and 0.86672 for rural residential area. The AUC value of other construction land is 0.80346, and the AUC value of urban construction land is 0.9376. In the urban expansion planning project, multi-source data such as land use status data, population distribution data, and economic development data are integrated, and the spatial agglomeration pattern of land use types is found by using the method proposed in this article, so as to provide a reference for rational planning of urban expansion direction.

Keywords: data mining, clustering algorithm, decision tree, time series, national space planning

1 Introduction

Land is an important basis for human beings to engage in various productive activities. Overall land use planning is a strategic plan that coordinates the allocation of land resources among various departments and the development, utilization, renovation, and protection of land in a relatively long planning period according to the needs of national economic development and the adaptability of land itself [1]. The overall land use plan is the “leader” of the entire land management work. The preparation and implementation of the overall land use plan plays a key role in strictly controlling all types of construction land, effectively protecting cultivated land, and realizing the dynamic balance of the total cultivated land [2]. The overall planning of land use has the function of “connecting the preceding and the following,” involving the fields of social national economy, population development, ecological environment, urban construction, and so on.

Since the 1990s, with the establishment of the geographic information industry and the popularization and application of earth digital products, the development of Geographic Information System (GIS) has entered the user era. During this period, the society’s awareness of GIS has generally improved, and the demand has increased significantly, and GIS has become a necessary working system for many organizations (especially government decision-making departments) [3]. National and even global GIS has become a matter of public concern, and GIS has been included in the “information Superhighway” plan [4],

*** Corresponding author: Yingqiao Wang,** Planning Research and Information Centre, Sanya Municipal Bureau of Natural Resources and Planning, Sanya, Hainan, 572000, China,
e-mail: YingqiaoWang20@outlook.com

which is also an important part of the "digital Earth" strategy proposed by the United States. Dong focuses on ten basic relationship issues [5]. Among them, the relationship between planning and market, spatial planning, and development planning is the basic relationship of territorial spatial planning. Only by dealing with these two relations can we reasonably determine the functional positioning and responsibility boundary of territorial spatial planning and ensure the correct direction of the reform of multi-planning. Lin *et al.* believed that the original intention of establishing the spatial planning system was based on the institutional reform of ecological civilization and improve the supervision system of natural resources, and the implementation of territorial space use control was a connecting point [6].

Traditional spatial planning methods may have limitations in data processing, analysis, and decision support. For example, in the face of massive, multi-temporal, and multi-form land use planning data, it may be difficult for traditional methods to fully explore the hidden laws therein. By introducing data mining technology and improving geographic information visualization, we can make up for these deficits and improve the scientific and rational of territorial spatial planning. This article is based on data mining and geographic information visualization to carry out the research on territorial spatial planning. This article summarizes clustering algorithms and decision tree classification algorithms in spatial data mining, including spatial clustering mining, decision tree classification steps, limitations, and improvement measures. Land planning based on spatial data mining covers the design of the planning decision system, which includes data, organization, mining, and decision four layers, and multi-source data mining urban planning identification, involving data standardization, training sample acquisition, function identification, and so on. Finally, the land planning simulation analysis is carried out, and the influencing factors are analyzed by the heterogeneity model, and the simulation results are given. The design of a land use planning decision support system is proposed by using a cluster algorithm for data mining, and the simulation results show that the simulation accuracy of the model is high.

2 Spatial data mining based on clustering algorithm

2.1 Spatial cluster mining

At present, the research work of spatial data mining has been carried out in China. For example, the Chinese Academy of Sciences has begun the research of the innovative project "spatial data mining and knowledge discovery." The State Key Laboratory of Visual and Auditory Information Processing open project Fund supports research on spatial data online analysis and spatial data mining. This project focuses on the spatial data online analysis and spatial data mining and the underlying spatial data warehouse technology for basic theoretical research and has made good progress in spatial data mining methods. With the increasing spatial characteristics of the amount of stored data, existing data analysis methods are faced with challenges of efficiency and scalability [7]. For example, many data analysis methods in computer science, statistics, machine learning, spatial databases, and other disciplines need to be re-studied.

In order to comprehend spatial data, find spatial relations and non-spatial relations, build a spatial knowledge base, rearrange spatial databases, and optimize queries, spatial data mining technology must integrate data mining technology and spatial database technology [8]. Spatial distribution rules, spatial association rules, spatial clustering rules, and other methods of spatial data mining have been widely used in GIS-based land planning and have achieved certain results.

The basic knowledge types that can be mined from spatial data include spatial clustering rules, general geometry knowledge, spatial distribution laws, *etc.* Among them, spatial cluster analysis refers to clustering spatial targets with similar features into a class, aiming to find the spatial connection, adjacency, and symbiosis of targets [9]. It can be used for spatial generalization and synthesis of GIS, for example, to group spatially similar lands into one group.

The grid-based approach quantizes the object house into a finite wide variety of devices to shape a grid structure. All clustering operations take location on this grid structure. The essential benefit of this technique is that it is very quick, and its processing time is impartial to the quantity of recorded objects and solely associated with the range of devices in every dimension of the quantization space. We recommend a spatial adjacency clustering algorithm with the use of a grid-based algorithm.

In the program, the function MS is the main function, which scans the simulated grid from top to bottom, from left to right, searches for GIS with value 1, and adds it to the dynamic array decision tree. The GIS is zeroed out, and the F function is called. The function of the function detection (FD) is to scan the eight positions adjacent to the grid. The grid coordinates of all adjacent points assigned 1 are pressed into a stack (ST), and the assignment of these adjacent points is changed to 0. Finally, with the stack ST being empty as the loop condition, the loop is repeated until all the grids are scanned.

2.2 Decision tree classification algorithm

Using a choice tree to classify two steps. The first step is to use the coaching set to construct a choice tree and construct a choice tree model. This technique is surely a procedure of obtaining expertise from facts and doing laptop learning. The 2d step is to use the generated choice tree mannequin to classify the unknown facts samples [10]. For an unknown facts sample, the attribute values of the pattern are examined successively from the root node till a leaf node is reached, so as to discover the type in which the statistics pattern resides.

At present, there are a number of algorithms for classification troubles in unique fields, such as statistics, computing device learning, neural networks, and hard-set theory. The choice tree technique derived from laptop getting to know is an extra frequent and deeply studied classification characteristic approximation method.

Entropy is a metric widely used in information theory. The entropy of the relative Boolean classification is defined as

$$\text{Entropy}, S \equiv -P \oplus \log_2 P \oplus -P \log_2 P, \tag{1}$$

where $P\oplus$ represents the proportion of positive examples in S.

If the target property has c different values, then entropy is defined as

$$\text{Entropy}, S \equiv \sum_{i=1}^{c} - P_i \log_2 P_i. \tag{2}$$

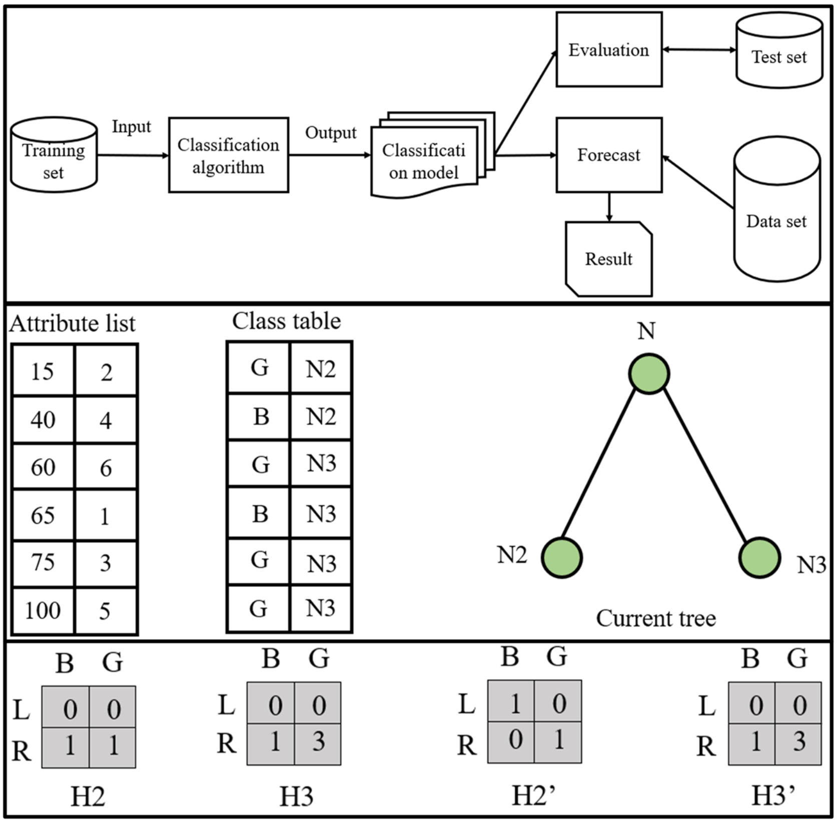
Given a definition of entropy, it is viable to outline a measure of records obtained for the capacity of the attribute classification education data:

$$\text{Gain}(S, A) \equiv \text{Entropy}, S - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}, S_v. \tag{3}$$

Values (A) is the set of all possible values of attribute A, and S_v is the subset of attribute A in S whose value is v.

As shown in Figure 1, when the first item of the attribute table is scanned, and the test condition is $a \leq 15$, the sample distribution that meets the condition in the H2 histogram is recorded in line L, and the sample distribution that does not meet the condition is recorded in line R.

Decision tree algorithm has some limitations in scalability when dealing with large or complex data sets. When constructing the decision tree, the whole data set needs to be scanned several times to calculate the information gain and other indicators to select the best-split attribute. As the



Attribute list

15	2
40	4
60	6
65	1
75	3
100	5

Class table

G	N2
B	N2
G	N3
B	N3
G	N3
G	N3

H2

	B	G
L	0	0
R	1	1

H3

	B	G
L	0	0
R	1	3

H2’

	B	G
L	1	0
R	0	1

H3’

	B	G
L	0	0
R	1	3

Figure 1: Sample decision tree diagram.

size of the data increases, the time to build the decision tree increases significantly.

In terms of computing resources, decision tree algorithms require memory to store data sets, intermediate computation results (such as the information gain value of each node), and to build a good decision tree structure. For large data sets, memory requirements are large. In the process of building decision trees, operations such as calculating information gain require CPU resources, especially when working with high-dimensional data (*i.e.*, data sets with many attributes), a lot of CPU computing power is required to evaluate the extent to which each attribute contributes to the classification.

2.3 Implementation of spatial clustering data mining

K-Means clustering method is a typical spatial clustering method based on partition. The method based on partition is an ancient clustering method and one of the most popular methods at present [11]. The connotation of the method is as follows: the spatial database contains n spatial objects, and for the specified number of clusters k ($k \leq n$), the whole data set is divided into k spatial clusters by adjusting the optimal partition criteria, and each partition is regarded as a classification.

The *K*-Means clustering method has a simple and easy-to-understand algorithm flow and relatively few parameter settings, which makes it easy to implement and operate in practical applications. Because the *K*-Means algorithm adopts iterative optimization and only needs to calculate the distance from the data point to the center of mass and redistribute the clusters in each iteration, its computational efficiency is relatively high. This is of great significance in large-scale data processing and real-time analysis. Density clustering algorithms such as DBSCAN have high requirements for the density distribution of data points, while the *K*-Means algorithm has no strict requirements for the density distribution of data. This makes the *K*-Means algorithm more adaptable when dealing with data sets with uneven density.

In territorial spatial planning, the clustering algorithm and decision tree method are complementary. The clustering algorithm can conduct preliminary grouping and exploratory analysis of the territorial spatial data, find the potential patterns and structures in the data, and provide the basis for the classification of decision trees. After the territorial space is divided into different regional groups by clustering algorithm, the decision tree can further classify and label these groups and determine the functional type of each region. This combination can give full play to the advantages of clustering algorithms in unsupervised data exploration and decision trees in classification and interpretation. Compared with other methods alone, it can analyze and process territorial spatial planning data in a more comprehensive and in-depth way, thus providing more effective decision support for territorial spatial planning.

The core content material of *K*-Means lies in its iterative optimization process. Through non-stop optimization of the rectangular error criterion, K house clusters are sooner or later obtained. The simple manner is as follows (as proven in Figure 2):

(1) k seed points are randomly selected, and each seed point is regarded as the center of mass of k spatial collection clusters.
(2) Calculate the distance between each spatial object and the seed point, and assign the unclustered spatial objects to the nearest seed point according to the principle of the closest distance.
(3) Move the seed point to the center belonging to the "point group" of the point.
(4) The above (2) and (3) steps are repeated until the square error criterion (the sum of squares of the distance from each four objects in the spatial set cluster to the seed point) converges, and the clustering process ends

The above square error criteria are defined as follows:

$$E = \sum_{i=1}^{\mathrm{k}} \sum_{p \in C_i} |p - m_i|^2, \tag{4}$$

where E represents the square error criterion. p stands for spatial object. m_i represents the center of mass of the space set c.

3 Land planning based on spatial data mining

3.1 Design of planning decision system

Since land use planning data come from various sectors of social and economic life, such as land, surveying and mapping, agriculture, and municipal affairs, it has the characteristics of mass, multi-time, and multi-form and is very suitable for spatial data warehouse organization and management [12]. On this basis, using spatial online analytical processing (OLAP), data mining, and other technologies, land use planning data mining and analysis can be carried out according to different themes, and the laws hidden behind the massive historical data of land use planning can be explored. Based on the characteristics of land use planning data and for specific services such as land adaptability analysis, the architecture of the auxiliary decision support system for taxi land use planning with spatial data warehouse as the data management platform, data mining as the knowledge acquisition method, and metadata management as the management means is designed and established and explained [13].

The decision support architecture of land use planning is shown in Figure 3. The four layers are data layer, organization layer, mining layer, and decision layer. From a macro perspective, these four layers are respectively responsible for different stages of land use planning information decision support, including data storage, data processing, data mining, and knowledge expression, forming a complete system [14]. Among them, the data layer is the foundation of the system, the organization layer organizes the data according to different needs, the mining layer completes the data analysis and information mining, which is the core part of the entire system, and the decision layer presents the mined information to the decision-makers.

The data layer includes the data set for land use planning data analysis and mining, which is a land use planning data warehouse composed of various thematic land planning databases, land planning achievement databases, attribute databases, and so on. It includes not only planning historical data and land use status data (including attribute tables, planning maps, *etc.*), but also information sets for different topics established on the basis of land use data warehouse by means of screening, classification, and synthesis [15,16].

Based on SOLAP technology, the organization layer synthesizes all kinds of land use planning data and other related data and establishes multiple multidimensional data cubes for different topics to support data mining and online analysis. In these data cubes, there should be both macroscopic coarse-grained data and fine-grained

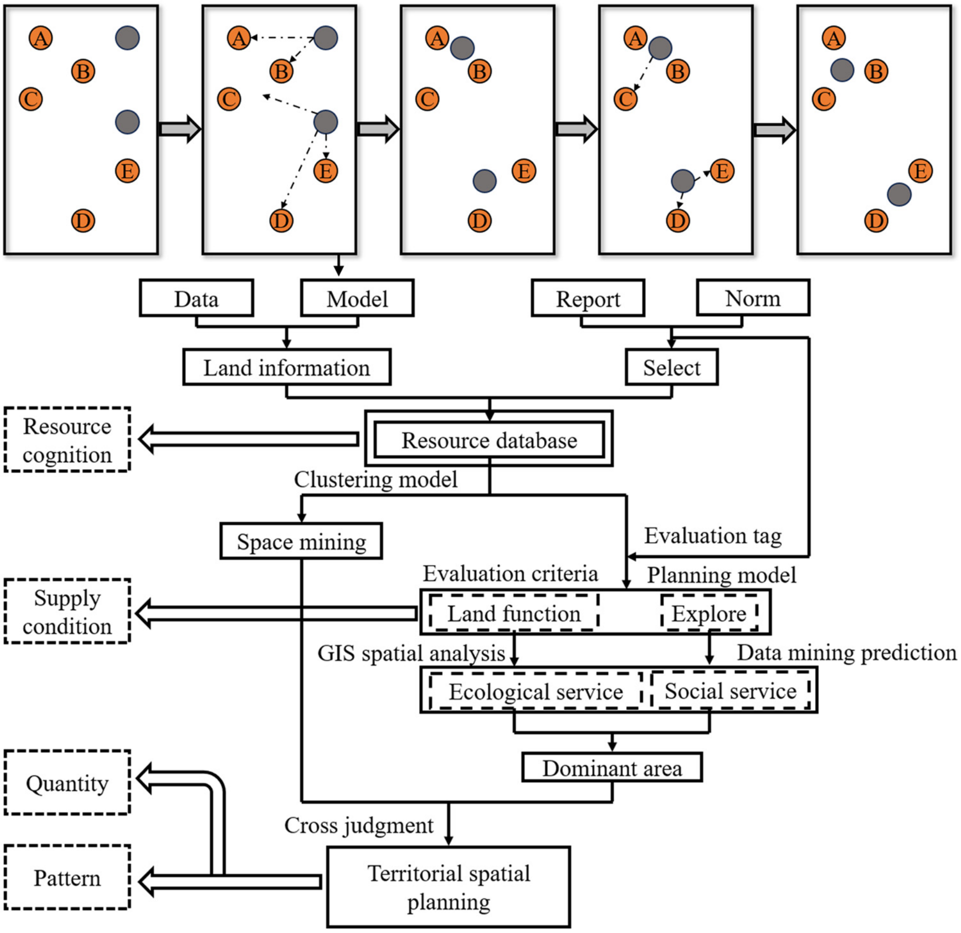
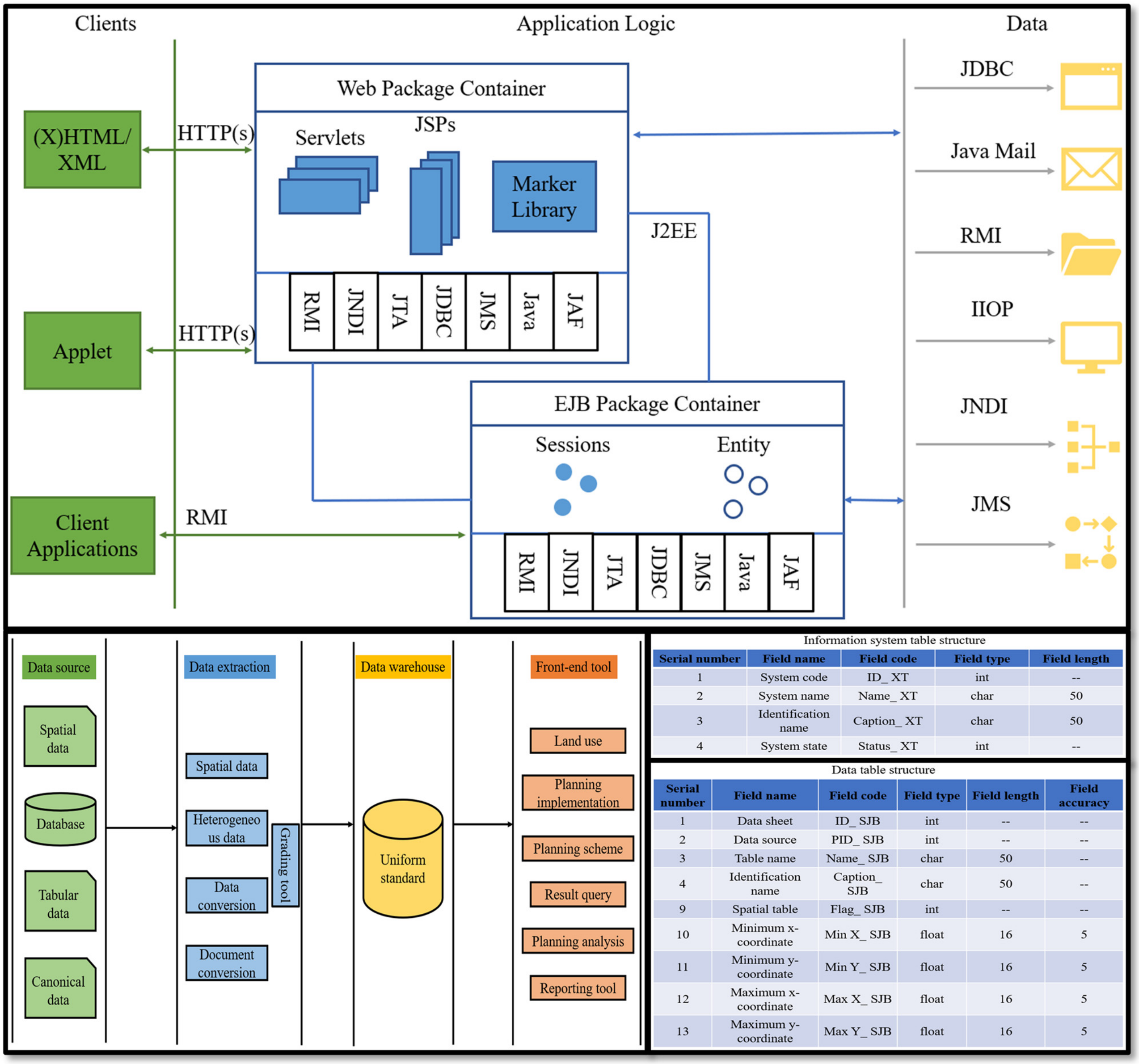


Figure 2: Schematic diagram of the clustering method principle.

data reflecting local details [17]. Organization-level data is managed through metadata.

As the core of the whole architecture, the mining layer mainly includes OLAP and spatial data mining. OLAP generally aims at data preprocessing and advanced application query, while data mining aims at knowledge discovery. In the entire land use planning choice aid system, different applied sciences serve for land use data mining, which can discover hidden and before unknown expertise in a massive range of land use planning facts and associated facts sets so as to be used for choice support.

The most basic metadata is the data of data, which is referred to as the data dictionary in a database system. In contrast, metadata in a data warehouse far exceeds the role of data dictionary. It not only defines the data in the warehouse, describes the data content, its location, and environment, but also specifies the rules of data extraction and conversion [18]. Metadata plays an important role in the whole process of data warehouse construction and operation. It can be divided into metadata about data source, metadata about data model, metadata about data warehouse mapping, and metadata about data warehouse usage.



Information system table structure

Serial number	Field name	Field code	Field type	Field length
1	System code	ID_ XT	int	--
2	System name	Name_ XT	char	50
3	Identification name	Caption_ XT	char	50
4	System state	Status_ XT	int	--

Data table structure

Serial number	Field name	Field code	Field type	Field length	Field accuracy
1	Data sheet	ID_ SJB	int	--	--
2	Data source	PID_ SJB	int	--	--
3	Table name	Name_ SJB	char	50	--
4	Identification name	Caption_ SJB	char	50	--
9	Spatial table	Flag_ SJB	int	--	--
10	Minimum x-coordinate	Min X_ SJB	float	16	5
11	Minimum y-coordinate	Min Y_ SJB	float	16	5
12	Maximum x-coordinate	Max X_ SJB	float	16	5
13	Maximum y-coordinate	Max Y_ SJB	float	16	5

Figure 3: Architecture diagram.

The auxiliary decision support system of land use planning is very complex; it involves a lot of data, algorithms, and applications, so it is necessary to use metadata to organize data effectively. Here, metadata is not only used to describe the data of basic information, but also a method of data organization [19]. Among them, metadata management mainly includes data management, algorithm management, and model management, which provides standardized description and management for the distributed, multilevel, and multi-model data information in the whole system and also sets up a bridge to connect various operations and analysis between the data, so as to systematically and effectively identify, process, and analyze various data to assist decision support.

3.2 Urban planning identification of multisource data mining

In order to achieve the coaching pattern of semi-supervised learning, the pre-processed time sequence facts are clustered with the use of the *K*-Means algorithm. The data is standardized, and the data of different dimensions are converted into values with the same scale. In the land spatial planning data, different types of data may be involved, such as land area (square meters), population number, economic indicators (such as GDP), *etc.*, and these data have different dimensions. Through standardization processing, such as converting the data into a value with a mean of 0 and a standard deviation of 1, different types of data are comparable in the clustering algorithm.

First, the reliability of cluster numbers will be evaluated by using repeated clustering operations. The profile coefficient Silhouette will differ with the range of clusters *K*, as proven in Figure 4. The larger the Silhouette price is, the higher the clustering impact will be. Considering the version of the Silhouette price with K and the dimension of the facts volume, the remaining clustering wide variety is 6. At this stage, urban land is divided into six clusters (C1–C6); however, the precise feature of every cluster is no longer clear, however it can be considered from the discern that the practical distribution sample of the learn about location is essentially round [20].

The *K*-central factor clustering algorithm is a clustering algorithm whose purpose is to divide the objects in a data set into *K* clusters (*K* is a pre-specified number of clusters). Similar to the *K*-Means algorithm, the *K*-central factor clustering algorithm uses the actual points in the data set (called central objects or medoids) as representatives of the cluster, rather than the mean vector of the cluster as *K*-Means does. In dynamic time warping (DTW)-based classification, the DTW distance between the test sample and the time series of the known class (the training sample) is usually calculated first. These distances are then used to determine which category the test sample is most likely to fall into. For example, you can assign a test sample to a category that belongs to the training sample with the smallest distance from its DTW.

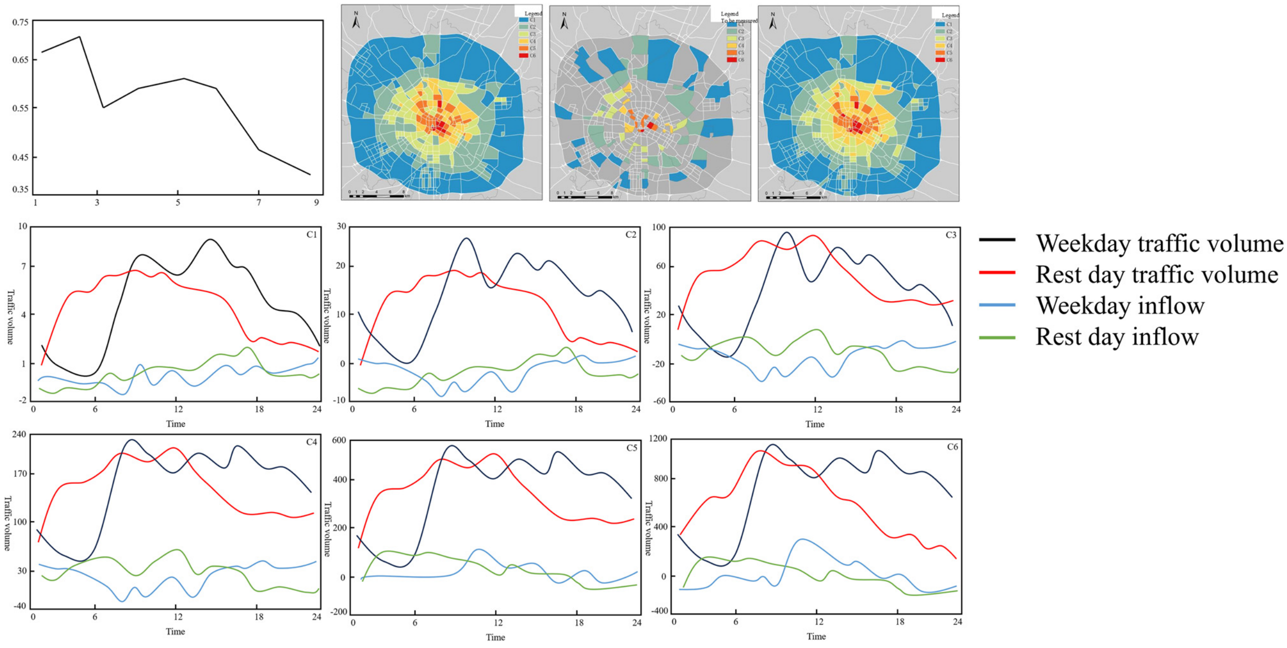


Figure 4: Identification results of spatial data mining.

The classification results of functional areas presented in the figure were obtained. Direct clustering is an unsupervised learning method and there may be inaccurate classification of some regions. In order to make the outcomes greater credible, this article selects 20% of the facts with correct classification impact from the above clustering consequences as coaching samples primarily based on the baseline of every cluster time sequence after classification [21]. Then, *K*-nearest neighbor supervised classification was used to be carried out in accordance to the coaching pattern and the closing end result of practical location classification used to be obtained. At the equal time, the use of the baselines of the time collection of every cluster of classification results, the unit area, visitors quantity, and influx per unit time of every classification in working days and relaxation days can be statistically obtained, so as to locate the journey regulations of residents in a number plots.

When selecting point of interest (POI) data, it is necessary to filter according to the specific needs of the study. For example, if the research focuses on commercial functional areas of the city, then the main selection of business-related POI data, such as shopping malls, supermarkets, commercial streets, *etc.* At the same time, the accuracy and integrity of the data should be considered, and the data sources with high data quality and timely updates should be preferentially selected. For POI data from different data sources, data cleansing and consolidation may be required to remove duplicates and erroneous information.

In order to discover the unique purposeful kinds of exclusive classes, the frequency density and kind ratio of POI information of every cluster (C1–C6) after clustering have been counted in this article (as proven in Table 1 below). The unique features of the neighborhood instructions can be described by means of the residents' tour traits and POI distribution traits after clustering.

C1 is disbursed in the side place of the find-out-about area, and the frequency density of POI is the lowest, so the rule has to be observed from the course of residents' tour characteristics. As shown in the figure, the characteristics of population flow in this area are not obvious on working days [22]. However, on relaxation days, the populace flows in the morning, and there is a greater populace outflow in the afternoon and evening, respectively, which is constant with the regulation of humans enjoying and touring spouse and children in the suburbs on weekends. Therefore, C1 is judged to be a suburban traveler area. It is named 'suburban' because it can be distinguished from major tourist attractions outside the study area.

C2 and C3 are allotted in the transition vicinity between the suburbs and the city center. From the viewpoint of the share of POI types, both C2 and C3 have essential provider services such as shopping, science and education, and clinical care for residents, and C3 has especially extra residential POI. In phrases of journey characteristics, the top of populace outflow and influx in C2 working days is 8:00 and 24:00, respectively, and that in C3 working days is 8:00 and 19:00, respectively. Compared with C3 working days, C3 has greater apparent commuting guidelines in residential areas. C2 belongs to the transition location of C1 and C3, so C3 is judged to be the city residential area, and C2 is the residential/tourist blended area.

C4 and C5 workplace POI (including science and education, medical, company enterprises, authorities agencies, *etc.*) have a greater percentage of kinds in contrast with different regions. In phrases of tour regularity, C4 has a populace outflow in the sunlight hours of working days (7:00–19:00), the boundary factor of populace inflow/outflow happens at 13:00 on relaxation days, the populace outflow height is at 8:00 on relaxation days, and the populace outflow top is at 8:00 on working days. The quantity of site visitors in C5 fluctuates greater regularly for the day

Table 1: Proportion of land planning data types

POI type	C1			C2			C3		
	FD	FDnor	CR	FD	FDnor	CR	FD	FDnor	CR
Catering planning	25.42	0	—	75.84	0.12	9.99%	139.47	0.28	%
Shopping planning	21.78	0	—	54.69	0.25	%	64.27	0.35	%
Leisure planning	20.76	0	—	56.51	0.11	%	109.21	0.31	%
Accommodation planning	3.86	0	—	15.26	0.05	%	26.42	0.14	%
Cultural planning	8.29	0	—	26.41	0.16	%	59.76	0.38	%
Medical planning	9.78	0	—	25.01	0.18	%	56.46	0.48	%
Housing planning	6.41	0	—	20.34	0.08	%	59.91	0.34	%
Corporate planning	27.41	0	—	64.81	0.15	%	98.44	0.21	%
Government organization planning	6.08	0	—	14.17	0.08	%	33.45	0.34	%
Scenic spot	0.74	0	—	1.52	0.09	%	3.24	0.28	%

and produces many extremes, with the influx step by step starting after 8:00 on weekdays and the outflow peaking at 18:00. Compared with C4, C5 has extra apparent commuting time characteristics, so it is judged that C5 is a workplace area. The traits of populace influx on working days in C4 are comparable to those in residential areas, and the fashion of populace influx on relaxation days is identical to that on working days [23]. However, the magnitude of change is small, which is due to the fact people's buying things to do on relaxation days offset the unique fashion to a positive extent, so in summary, C4 is judged to be a blended residential/commercial area.

C6 is solely dispensed in the central region of the city, and from the point of view of tour characteristics, the populace influx peaks at around 10:00 on each weekend and weekdays. However, the outflow begins at 11:00 on relaxation days and at 16:00 on working days. The visitors' quantity on relaxation days and working days peaks at 10:00 and 18:00, respectively, which is extraordinarily comparable to people's purchasing time. Judging from the distribution traits of POI, there are many kinds and portions of POI in this area, and all sorts of services are perfect; this vicinity is judged to be a mature commercial enterprise district.

The method of this article accurately divides different types of land use areas, predicts the changing trend of land use, and provides a basis for the rational allocation of land resources in national spatial planning. This article realizes the visualization of geographic information and presents all kinds of data in territorial space planning in an intuitive graphical way, which is convenient for decision-makers and related personnel to understand the status quo and planning direction of territorial space. This article also has some shortcomings. It fails to visualize complex territorial spatial data in a concise, easy-to-understand, and accurate way and fails to consider the timeliness of visualization, that is, how to update the visual content in time to reflect the dynamic changes of territorial space.

4 Simulation analysis of territorial planning

4.1 Heterogeneity model of spatial data mining

From the standpoint of eye-catching territory space, this study uses two typesof issue detectors and interactive detectors to analyze the influencing elements of territory spatial distribution. Based on the clear evaluation of spatial heterogeneity of geographical phenomena, the thing detector makes use of q fee to quantify the explanatory capability of each independent variable's degree of impact on the dependent variable, and detects the explanatory capability of the component to the established variable by calculating the intra-layer variance and the relationship with the sum of whole variance, which absolutely describes the similarity of spatial modifications amongst geographical data. Its calculation method is as follows:

$$q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2}. \tag{5}$$

The interactive detector is used to quantify the interplay of two explanatory variables on precise goal variables. In this study, the superposition of two influencing factors is used to notice the joint impact of each on the stratification and heterogeneity of territorial space. Determining whether the interaction is enhanced, weakened, or independent can overcome the limitations of traditional regression models, and no longer simply analyze its interaction by the multiplication of two factors. This study mainly conducts comparative analysis under different territorial spatial clustering (or regional division) by detecting the difference in the influence and interaction of various factors.

In order to measure the spatial feature gap on different spatial scales, this study calculated the Gini coefficient and decomposed measure based on the regression model Shapley value, and specifically reflected the contribution degree of different influencing factors to the Gini coefficient inequality. Data acquisition & signal processing extension package is a commonly used tool for inequality analysis, including a variety of inequality indices and their decomposition, and the Gini coefficient is a commonly used index to measure the degree of inequality; the higher the value of the Gini coefficient, the higher the degree of inequality. The Gini coefficient is calculated as follows:

$$\text{Gini} = \frac{-\sum_{i=1}^{N} P_i \ln P_i}{\log_2 N}, \tag{6}$$

$$C = 1 - \text{Gini}. \tag{7}$$

Based on the inequality index, an inequality decomposition method based on a regression equation is produced. The regression equation decomposition method of the Gini coefficient was first proposed by Western economist Oaxaca. On this basis, Chinese scholar Wan Guanghua combined the regression equation with the decomposition principle of Shaply value [24]. The Shaply value decomposition method has a number of advantages, such as there are no restrictions on the use of inequality indicators and the preset regression mode, and the measure of indicators can be analyzed with the help of the estimated regression equation. This method can overcome the limitations of simple regression analysis and conventional exponential decomposition and obtain the contribution degree and ranking of explanatory variables to the explained variables. The calculation result of inequality decomposition is the correlation coefficient between the influencing factor and the level of inequality. The larger the coefficient, the greater the contribution of this factor to inequality. It can be expressed as

$$\ln C_j = \beta X_j + \varepsilon_j. \tag{8}$$

4.2 Simulation analysis

The factor detection results obtained through Geodetector are shown in Table 2. According to the impact factor detection results of national territory planning divided into traditional eastern and western regions, the principal influencing elements in the first column consist of the common elevation of distance from the coastline, the complete populace between 15 and 64 years ancient of GDP per capita, and the percentage of non-agricultural output value. In the 2d column, the object of detection is the land planning with a clustering depth of 65%. The fundamental influencing elements consist of the percentage of nearby residents, the whole populace aged 15–64, the complete population, and per capita GDP. Taking the clustering depth equal to 55% as the clustering standard, the most important elements detected through the elements in the 1/3 column are roughly steady with the effects of 65%. Set the clustering depth to 45%, the fundamental influencing elements bought in the fourth column are roughly regular with the effects of 65 and 55%; in addition, the issue of non-agricultural populace share additionally performs a position of 10%. In general, the four most necessary elements affecting the distribution of land planning are per capita GDP, whole populace aged 15–64, whole population, distance to the coastline, *etc.* These elements essentially consist of the predominant aiding elements of city construction, that is, herbal geographical advantages, prosperous labor sources, and subsequent monetary aid strength. The above consequences are essentially in line with the standard development notion of "time, geography and people," indicating that trendy land planning and building ought to additionally begin from the true scenario of the land, while accomplishing nearby conditions, blessings of development, so that the land planning and herbal landscape, the current texture of higher integration.

Table 2: Cluster analysis results

Influencing factor	Sort			
	Region ($\times 10^{-2}$)	Clustering depth 65% ($\times 10^{-2}$)	Clustering depth 55% ($\times 10^{-2}$)	Clustering depth 45% ($\times 10^{-2}$)
K-POP	9.1	34.1	11.4	15.5
K-POPY	12	42	15.4	20.4
K-PLR	9	46	26.9	32.4
K-GDPPC	15.1	30.8	11.5	15.2
K-PNP	6.5	17.5	7.9	11
K-PNV	10.9	16.4	2.5	4.9
K-DisPC	2.6	1.2	1.5	1.1
K-DisLMC	8.3	11.4	5.3	8.6
K-DisC	44.7	8.5	4.9	2.4
K-DEM	43.8	10.8	7.8	6.1
K-Slope	1.6	2.4	5.5	3.2

Area under curve (AUC) refers to the area under the receiver operating characteristic (ROC) curve (receiver operating characteristic curve). The ROC curve is drawn with the false positive rate as the horizontal coordinate and the true positive rate as the vertical coordinate.

The clustering algorithm and decision tree were used to simulate seven types of land use in a city in 2020. The results showed that the simulation accuracy of the algorithm reached 76%, and the Kappa coefficient was 0.66. Among them, the prediction accuracy of urban land is 76.41%, the prediction accuracy of other construction land is 46.43%, and the prediction accuracy of rural residential areas is 72.23%. The prediction accuracy of grassland was 54.19%, forest land 64.83%, water city 64.71%, and cultivated land 85.07%. It can be seen that the clustering algorithm has strong forecasting ability for land planning in this study, but relatively low forecasting accuracy for other construction land. Figure 5 describes the ROC curve simulated by the clustering algorithm. The AUC value of seven types of land use was obtained by calculating the area under the ROC curve. The AUC fee for cultivated land used to be 0.9376, that of woodland land was once 0.85442, and that of grassland was once 0.81747. The AUC price of the water location is 0.8708, and that of the rural residential vicinity is 0.86672. The AUC fee for different development land is 0.80346, and the AUC price of city development land is 0.9376. The AUC values of cultivated land and city development land are increased to 0.9. The AUC values of different land made use of have been all increased than 0.8.

In land spatial planning, the land use types with high AUC value, such as cultivated land and urban construction land, can rely more on the prediction results of the model in the planning decision. For example, in the development of urban construction land expansion planning, due to the high accuracy of the model's prediction of urban construction land, a reasonable layout can be carried out according to the suitable construction area predicted by the model. For other construction land with low AUC value, it is necessary to be more cautious in planning decisions, and it may be necessary to combine more field research, expert opinions, and other information sources to ensure the rationality of planning.

From the outcomes of the evolution of territorial space, the future city ecological house region will be barely reduced; however, in the context of non-stop populace growth, it is quintessential to enhance its current useful resource and environmental carrying capacity. First of all, in the future, we should rely on the existing ecological resources, integrate with the humanistic resources in the city, and build urban green space systems through various forms such as roof greening, wetland parks, forest parks, *etc.*, to strengthen the service supply capacity of ecological resources. Attention should be paid to the improvement of the ecological environment and the construction of green landscapes in rural areas, the temporal and spatial correlation of ecological areas that open the way for biological migration should be strengthened, and the carrying capacity and risk resistance of ecological networks should be enhanced. Second, in view of the problem of insufficient greening in urban areas, it is suggested to strengthen the construction of urban endogenous green network in various forms such as small parks, streetscape green Spaces and walkway protection green Spaces. According to the existing diversity of ecological resources, differentiated classification and management should be carried out to strengthen the strict protection of ecological resources on one side of the mountain.

In the simulation analysis, the simulation results of the clustering algorithm and decision tree for seven types of land use, such as the simulation accuracy of 76%, the Kappa coefficient of 0.66, and the prediction accuracy and AUC value of different types of land use, indicate that the application of data mining technology in national land spatial planning is effective. This is consistent with the purpose of the paper's research on the optimization of territorial spatial planning based on data mining technology, indicating that data mining technology can play a positive role in territorial spatial planning.

In the actual territorial spatial planning project, attention should be paid to the integration of multi-source data (such as data from land, surveying and mapping, agriculture, municipal, and other departments) as described in the article. These data have characteristics such as massive volume, multiple time frames, and various forms, spatial data warehouse can be used to organize and manage, and then data mining technology (such as clustering algorithm and decision tree, *etc.*) can be used to excavate the hidden rules and provide a basis for planning decisions. In the urban expansion planning project, multi-source data such as land use status data, population distribution data, and economic development data are integrated to find out the spatial aggregation pattern of land use types

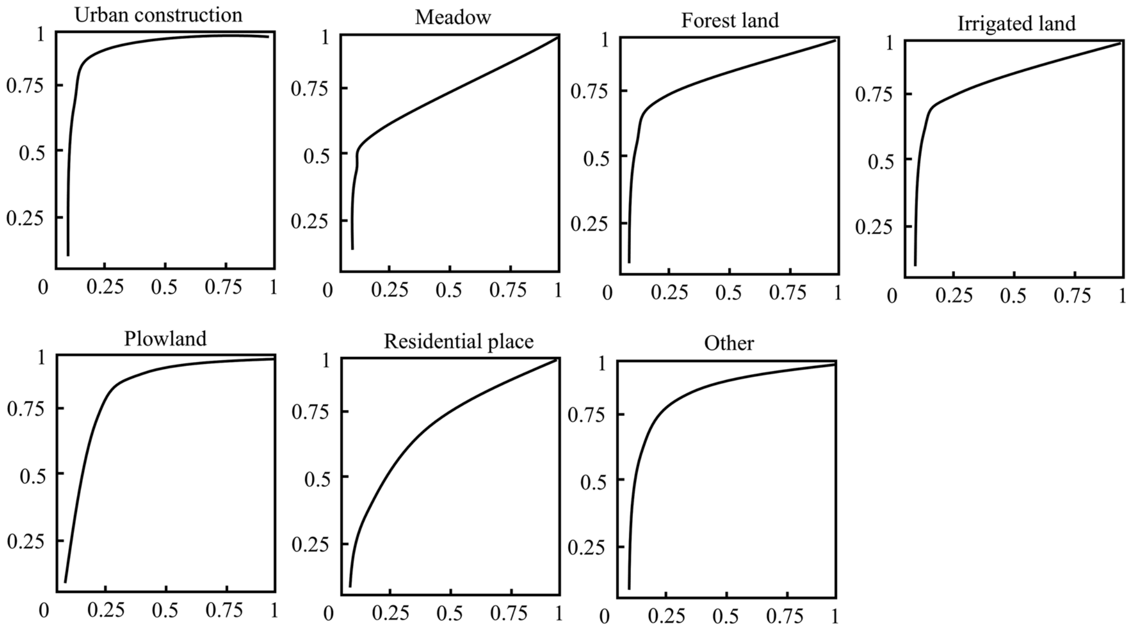


Figure 5: Prediction results of territorial planning based on *K*-Means clustering algorithm.

through cluster analysis, which provides a reference for rational planning of urban expansion direction.

5 Conclusion

Based on the statistics mining science and the visualization of geographic information, this article places ahead the optimization look-up of territorial spatial planning. On the groundwork of in-depth look-up and dialogue of spatial information mining technological know-how and its utility in GIS, this article researches and analyzes the spatial cluster evaluation approach in spatial statistics mining in detail. Combining with the traits of GIS spatial information, in accordance with the spatial cluster evaluation technique primarily based on the grid, the *K*-Means clustering algorithm primarily based on GIS is proposed. The effects are as follows:

1. This article uses a clustering algorithm to carry out data mining and puts forward a system design for a land use planning decision support system. The solution of land use planning data warehouse construction is elaborated in detail, including the overall design of spatial data warehouse, logical modeling, data cube construction, and metadata management. It realizes multi-dimensional organization management and application of massive multi-source heterogeneous data.
2. The simulation results indicate that the simulation accuracy of the algorithm reaches 76%, and the Kappa coefficient is 0.66. Among them, the prediction accuracy of urban land is 76.41%, the prediction accuracy of other construction land is 46.43%, and the prediction accuracy of rural residential areas is 72.23%. The AUC value of cultivated land was 0.9376, that of forest land was 0.85442, that of grassland was 0.81747, that of water area was 0.8708, and that of rural residential area was 0.86672. The predicted results in this article are in good agreement with the real situation, and the result indicates that the simulation precision of the model constructed in article is highly reliable.

Although this article has improved the clustering algorithm, it can study the clustering algorithm that is more suitable for the characteristics of territorial spatial planning data or improve the existing clustering algorithm. For example, consider how to better deal with the impact of noise points and outliers in territorial spatial data on clustering results to improve the accuracy and stability of clustering.

Funding information: Authors state no funding involved.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

[1] López-Acevedo FJ, Escavy JI, Herrero MJ. Application of Spatial Data Mining to national mines inventories for exploration and land planning of high place-value mineral resources. The case of aggregates in Spain. Resour Policy. 2022;79:103096.

[2] Wang S, Yuan H. Spatial data mining: A perspective of big data. Int J Data Warehous Min. 2014;10(4):50–70.

[3] Golmohammadi J, Xie Y, Gupta J, Shekhar S. An introduction to spatial data mining. UCGIS. 2020;4(10):121–3.

[4] Fu W, Liu J. Design and implementation of Land planning Transfer Information Management System based on GIS. Surv Map Spat Geogr Inf. 2012;35(10):82–5.

[5] Dong Z. Ten relationships of territorial spatial planning in the new era. Resour Sci. 2019;11(9):1589–99.

[6] Lin J, Wu Y, Wu J. On the construction of spatial planning system – Analysis of the relationship between spatial planning, territorial space use regulation and natural resources regulation. Urban Plan Rev. 2018;8(5):9–17.

[7] Gajić A, Krunić N, Protić B. Classification of rural areas in Serbia: Framework and implications for spatial planning. Sustainability. 2021;13(4):1596.

[8] Tu X, Fu C, Huang A, Chen H, Ding X. DBSCAN spatial clustering analysis of urban "Production–Living–Ecological" space based on POI data: A case study of central urban Wuhan, China. Int J Environ Res Public Health. 2022;19(9):5153.

[9] Su L, Fu L. Regional land planning based on BPNN and space mining technology. Neural Comput Appl. 2021;33:5241–55.

[10] Tselios V, Stathakis D. Exploring regional and urban clusters and patterns in Europe using satellite observed lighting. Environ Plan B Urban Anal City Sci. 2020;47(4):553–68.

[11] Sang N, Aitkenhead M. Data Mining, Machine Learning and Spatial Data Infrastructures for Scenario Modelling. In: Modelling Nature-Based Solutions: Integrating Computational and Participatory Scenario Modelling for Environmental Management and Planning. Cambridge University Press; 2020. p. 276–304.

[12] Wang M, Yang M, Yang X, Chen J, Yang B. Data mining of national geographical census for decision-making in urban planning: A geo-simulation of urban size in Beijing, China. Sens Mater. 2023;35(2):11–2.

[13] Festa D, Bonano M, Casagli N, Festa D, Bonano M, Casagli N, et al. Nation-wide mapping Nation-wide mapping and classification of ground deformation phenomena through the spatial clustering of P-SBAS InSAR measurements: Italy case study. ISPRS J Photogramm Remote Sens. 2022;189:1–22.

[14] Lee S, Lee S, Lee MJ, Jung HS. Spatial assessment of urban flood susceptibility using data mining and geographic information System (GIS) tools. Sustainability. 2018;10(3):648.

[15] Sufiyan I, Alkali M, Sagir IM. 3d modeling and assessment of flood risk zones using gis and remote sensing in catchment area Terengganu, Malaysia. Malays J Geosci. 2022;6(2):97–100.

[16] Park S, Xu Y, Jiang L, Chen Z, Huang S. Spatial structures of tourism destinations: A trajectory data mining approach leveraging mobile big data. Ann Tour Res. 2020;84:102973.

[17] Zhang Z. Relationship between regional economic development and ecological environment based on spatial data mining. Ekoloji Derg. 2019;9(107):15–9.

[18] Hou J, Zheng M. Online spatial evaluation of residential livability based on POI data mining and LMBP algorithm. Arab J Geosci. 2021;14:1–11.

[19] Kurowska K, Kietlinska E, Kryszk H. Possibilities use to selected methods of spatial data mining in demographic data analytics. Int Sci J Balt Surv. 2018;9:56–62.

[20] Niu H, Silva EA. Crowdsourced data mining for urban activity: Review of data sources, applications, and methods. J Urban Plan Dev. 2020;146(2):04020007.

[21] Bahari Sojahrood Z, Taleai M. Exploring the spatial pattern of urban land uses by utilizing data mining methods. Sci Res Q Geogr Data. 2021;30(119):75–86.

[22] He S, Luo D, Guo K. Analysis of factors affecting the coordinated development of urbanization and the ecological resource environment in southwest China based on data mining. J Urban Plan Dev. 2021;147(3):04021034.

[23] Zhang J, Sui X, He X. Research on the simulation application of data mining in urban spatial structure. J Adv Transp. 2020;2020(6):4–9.

[24] Bohat N, Joshi V. Assessing soil erosion in mandakini river watershed: A sub-watershed scale analysis using rusle model and geospatial tools. Malays J Geosci. 2024;8(1):10–6.