J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 5435--5451

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

A Study on the Impact of Urban Functional Mixing on Neighborhoods and Social Inclusion Based on Big Data Analysis

Wenshu Li[1], Yichen Yang[1], Chunsong Li[1] ✉

[1] *School of Urban Construction, Beijing City University, Beijing, 101309, China*

ABSTRACT
Increasing the degree of mixed use of urban land and building diverse and multifunctional urban spaces are important ways to shape urban vitality and promote healthy development of neighborhoods and social inclusion. Taking the urban area of City A as the research object, the article screens and classifies the collected POI data, and realizes the division and identification of functional areas in the core urban area of City A by calculating the degree of chaotic urban land use in parcels based on entropy under the fine-grained grid scale of the road network. Subsequently, the calculation methods of spatial weights and bandwidths of the model based on ordinary least squares and the Moran's index eliciting the GWR model are introduced. Finally, eight factors that have an impact on neighborhood and social inclusion were selected as explanatory variables, and an empirical study of the spatial distribution of neighborhood and social inclusion and the influencing factors was carried out using the geographically weighted regression model. The study found that the functional mixing degree in the main urban area of City A generally shows the spatial distribution of high-mixing degree plots of land with "center clustering and multi-point scattering", and locally shows the characteristics of piecewise clustering in the central area, linear clustering along the main roads, and pointwise clustering around the subway stations. The four influencing factors of common habits, psychosocial distance, social contact behavior and external behavioral interference are positively correlated with the changes of neighborhood relationship and social inclusion.

Keywords: POI, Moran index, GWR model, least squares method, urban functional mixing degree

1. Introduction

In the process of continuous development of a city, urban functional zoning and urban spatial structure planning become more and more important. If the city is regarded as a complex organic whole, the urban functional space is a number of individual parts that constitute this organic whole [1]. The design of urban functional zoning can learn from the historically proven traditional urban architectural

✉ Corresponding author.
E-mail address: 18686823702@163.com (C. Li).
Received 05 March 2024; Revised 20 April 2024; Accepted 05 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-304

forms, and focus on the nature of daily interactions and neighborhood culture, to create a shared and diversified neighborhood spatial relationship in contemporary and future urban living spaces, thus shaping the future neighborhood scenario with a socially inclusive humanistic atmosphere [2-6]. In this context, strategies for urban functional zoning and urban spatial structure planning are proposed in order to reduce the generation of negative spaces in urban settlements [7-9]. At present, China is in the process of rapid urbanization, in order to achieve a precise grasp of the urban spatial structure and the current situation of urban functions, researchers need to make effective scientific decisions on the functional zoning of the city, so as to realize the identification and evaluation of the degree of mixing of urban functional areas [10-14].

In the traditional urban spatial functional identification, it mainly relies on the land use status map, on-site research, questionnaire survey, expert judgment and other ways to carry out the functional identification, which is obviously subjective, and the accuracy of the identification results is affected to some extent [15-17]. In recent years, with the advent of the big data era, there are new ideas for planning research and application related to urban functional areas. Relying on the characteristics of big data technology, such as large data volume, simple access, and wide coverage, quantitative analysis of relevant data is carried out to quantitatively calculate the urban functional zones, which is used to identify and evaluate the degree of urban functional mixing [18-21]. The results of the study will help urban planners to better grasp the urban spatial structure and provide ideas and reference for urban planning research.

With the development of remote sensing technology, scholars began to use the spectra in remote sensing images to identify urban spatial functions. Tu, W. et al. used remote sensing and remote sensing data fusion algorithms to identify the functional zoning of a city as a bridge connecting human activities and urban planning, and to provide scientific support for the formulation of sustainable development policies for the city [22]. Chen, Y. et al. gave full play to the advantages of machine learning and deep learning in feature extraction to analyze the image data collected by high-resolution satellites to realize the recognition of feature information on a fine scale, which is of great significance for to urban planning and sustainable development [23]. Zhao, W. et al. proposed a context-aware segmentation network for the difficult problem of large-scale mapping of datasets caused by the division of urban functional areas by land boundaries or road networks, which introduces building type division criteria on the original basis and improves the accuracy of spatial unit and type identification of urban functional areas [24]. Zhou, W. et al. introduced the application of super-objective-convolutional neural network on urban functional zone delineation in ultra-high resolution remote sensing images, which can produce functional zone classification results at higher accuracy and is important in the study of small-scale functional spatial structure [25]. However, the features of remote sensing images alone cannot accurately obtain the specific functional information of intra-city spaces.

With the rapid development of geographic information technology and big data, more and more types of data are available to researchers, and some scholars analyze, integrate, and calculate these data to identify urban spatial functions. Cao, R. et al. showed that the introduction of social perception on the basis of remote sensing image recognition will effectively improve the recognition performance of high-density urban functional areas, and proposed deep learning strategies such as one-dimensional convolutional neural networks to fuse multi-source asynchronous remote sensing images and social perception data [26]. Xie, L. et al. pointed out that social perception data can provide additional information for urban functional area identification, and adopted a multimodal deep learning method incorporating the attention mechanism, which effectively fused the differences in the data forms of remotely sensed data and social perception data to provide data support for urban planning and management [27]. Qian, Z. et al. developed an urban functional zone identification model based on satellite images and cab GPS trajectories, which calculates the spatio-temporal information entropy of cab trajectories within the urban spatial unit and combines with numerical algorithms to classify them, which significantly improves the effectiveness of the functional zoning of the city [28]. Chen, W. et al. established a framework for mapping urban green space with remote sensing data and social perception data as data sources, and the fusion of multi-source data through algorithms can improve the accuracy of identifying social function types in urban areas, which can help governmental departments to promote urban planning decisions [29]. Chang, S. et al. pointed out that the function of an urban area is determined by the activities occurring in that spatial unit, and the introduction of cell phone dynamic signaling data, which has more identification accuracy, to reflect people's activity characteristics, makes the functional zoning of the city more practical value [30]. However, the high cost and difficulty coefficient of these data acquisition makes it difficult to be used as a universal identification path. Therefore, the development of a big data technology with the characteristics of easy access, easy processing, and large sample size for effective identification of urban functional mixing degree is a prerequisite for exploring the link between urban functional zoning and neighborhood relations and social inclusion.

Starting from the research perspective of POI big data in City A, this paper elaborates on the research area and data sources, innovatively applies the method of POI multi-influence factor weight assignment for entropy value calculation, and identifies the functional mixing degree of the land use in the main urban area to summarize the spatial distribution characteristics. Subsequently, the theoretical sources and related mathematics of the feature value model based on the least squares method and the Moran index are elaborated, and the Gaussian kernel function calculation formula of the geographically weighted regression model and the AIC bandwidth selection method are studied. Finally, eight factors that have an impact on neighborhood relations and social inclusiveness are selected as explanatory variables to study the spatial heterogeneity characteristics of the factors that shape neighborhood relations and social inclusiveness using the GWR model.

2. Subjects and methodology of the study

2.1 Study area and data

2.1.1. Overview of the study area. City A is located at the intersection of China's coastal open zone and the Yangtze River basin development zone, and is the core city of the world-class city cluster in the Yangtze River Delta. City A is the capital of Jiangsu Province, (118°22'-119°14'E,31°14'-32°37'N), and is a national ecological garden city. This paper chooses the core urban area of A city as the research object.

2.1.2. Data sources:

1) POI data. This study utilizes POI data from the 2023 City A Sky Map Update Project to obtain a total of 14,000 points of interest (POI) data for the core urban area. Each POI data contains several attribute values, which are expressed as field names, including serial number, latitude, longitude, code, area, importance, name, address, large class, medium class, small class, and so on. The POI data in this paper are categorized into 3 levels of large, medium and small classes according to the linear classification method and according to the subordination relationship. The major categories include residential land, public administration and public service land, commercial land, industrial, mining and warehousing land, transportation land, scenic land, green space and square, etc. There are 7 categories. The medium category includes 16 medium categories such as neighborhoods, science, education and culture, healthcare, business facilities, factories and parks. The subcategories include 42 subcategories such as kindergartens, elementary schools, middle schools, hospitals, shopping malls, and so on.

2) Basic Data.The basic data utilized in this study are the road network, water system, buildings and points of interest from the 1:500 topographic map project of City A in 2023, of which the road network data has a high degree of accuracy and completeness, which is suitable to be applied in this study of urban functional land use classification. For the acquired road network data, the road network data of

the main and secondary road fields are retained, and the redundant and messy road network is processed with appropriate tolerance to get the road network data needed for the study, and the road network is used to divide the urban area of City A into basic research units.The existing urban land use data of City A is mainly referred to the DLTB (land use plot) layer in the Third National Land Survey Project of City A, which is mainly related to the land use of Class 05 (commercial land), Class 06 (industrial, mining and storage land), Class 07 (industrial, mining and storage land), and Class 08 (industrial, mining and storage land). (industrial, mining and warehousing land), 07 (residential land), 08 (public administration and public service land), 10 (transportation land) and other land class patches, which can be used as the basic data to compare and analyze with the data of urban functional areas obtained from the experimental results of this paper.

2.2 Research methodology

2.2.1. Data pre-processing. Firstly, the POI data in the 2023 City A sky map update project were data cleaned by land classification, and non-classified data and redundant data were deleted. Secondly, the administrative divisions, road network data and POI data are loaded, and the main roads are used to divide the administrative divisions to form divided parcels. Then the POI spatial area weights and influence weights are harmonized to get the final POI weight table. Finally, the entropy value is calculated according to the formula, and the mixing degree distribution of each parcel can be obtained.

2.2.2. Normalization of POI weights. Since POIs are only point elements and different functional areas within the city have different influence and relevance, the categorized POIs are assigned with categorical weights [31]. In this paper, the final POI weights are obtained by reconciling the influence weights with the spatial area weights.
1) Influence weight. Different types of POIs have large variability in the degree of significance in terms of urban influence and public awareness, and this paper introduces a new evaluation index of influence weight to describe this significance. For POI data with multi-level classification, the normalized influence weights are obtained by combining the methods of fuzzy index evaluation and AHP hierarchical analysis model, using total sorting and hierarchical single sorting.
2) Spatial area weights. According to the current business classification standard in China, based on the construction area of the POI category in the classification standard, the construction area of each category of POI is calculated by analogy, and the POI area weights of 42 sub-categories are assigned through the Z-Score standardization method. Z-Score normalization refers to the standardization of data based on the mean and standard deviation of area weights.

$$y_i = \frac{x_i - \overline{x}}{s} \tag{1}$$

Among them:

$$\begin{gathered} \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{gathered} \tag{2}$$

where $x_1, x_2, \cdots, x_n$ is the sequence of area weights. $y_1, y_2, \cdots, y_n$ is the new sequence after Z-Score normalization.
3) Reconciliation weights. The weight values of the first two items are normalized in a way of very small value normalization to get two new columns of weights, and the formula is shown in equation (3).

$$\begin{cases} M_{ij}^{*} = \dfrac{M_{ij}}{\min(M_{ij})} \\ M_{j}^{0} = M_{1j}^{*} + M_{2j}^{*} \end{cases} \tag{3}$$

where M_{ij} is the weight of item j of category i. M_{ij}^{*} is the new value after the normalization of the minimal values. M_{j}^{0} is the final weight value of each subcategory after reconciliation. $i = 1,2; j = 1,2,\cdots,42$.

2.2.3. Entropy Calculation of Functional Mixing Degree. In this paper, we take the cleaned POI as the data source, and do the refinement of road network data administrative division to get the partitioned parcel, and determine the functional mixing degree of each type of POI by calculating the diversity of POI types within a single parcel; the larger the entropy value is, the larger the uncertainty is, which indicates that there are more types of POI functions within the parcel: the smaller the entropy value is, the smaller the uncertainty is [32]. The calculation formula is shown in equation (4):

$$H(X) = -\sum_{i=1}^{n} P_i \log P_i, \text{ among } P_i = \frac{M_j^0 \times R_k}{M_j^0 \times \sum_{k=1}^{m} R_k} \tag{4}$$

where $H(X)$ is the functional mix of POIs for a given parcel. n is the total number of POI types. p_i is the weighted average of the number of Type i POIs within the parcel over the number of all POIs. R_k is the number of Type k POIs within the parcel. M_j^0 is the weighted value of the reconciled POIs of that type. m is the total number of POIs within the parcel.

3. POI-based urban functional area mixing degree identification study

3.1 Analysis of Results and Optimization of Measurement Methods

3.1.1. Functional Mixedness Results Judgement. According to the general rule of spatial pattern development in city A, it is known that the functions of city center are complicated and diversified, and the high-grade commercial and business functions gradually replace the low-grade functions to promote the comprehensive development of city center, and the spatial distribution results of the functional mixing degree of the land use in the main city area of city A are also consistent with this rule. Dividing the old and new city areas by the Ming City Wall, the old city area has a high degree of functional mixing, while the areas outside the old city have a low degree of functional mixing.

3.1.2. Optimization of measurement methods. Through the status quo research visits, it is found that in multiple plots of similar size, whether it is an area with diverse types of functions or an area with rich numbers of functions, the crowd satisfaction of both is basically the same, and both can satisfy the diversified needs of the residents' daily life, i.e., the study considers that the impact of the types of functions and the number of functions on the functional mix of the land use has the same weight. A value closer to 1 means that the functional mix is higher and the functional types are more diverse, while a value closer to 0 means that the functional mix is lower and the functional types are more homogeneous.

3.1.3. Comparison of the two measures. The statistics of the number of parcels and area ratio before and after the optimization of the measurement method are shown in Table 1. It can be found that compared with the measurement method that only considers whether the functional types are balanced or not, the optimized results need to consider both the “quality” and “quantity” of the mix, so the number and area of plots with a low degree of mixing have increased significantly. For example, the number of optimized sites with low mixing degree increased by 11.24%. The mixing level of most sites has decreased, while the mixing level of a small number of sites has increased, and the characteristics of center clustering have become more obvious. Most of the sites with decreasing mixing grades are those with a low number of POIs. The sites with increased mixing grade are mostly commercial complex sites. The optimized results clearly distinguish the two differences and improve the accuracy of the measurement results.

Table 1. The quantity and area ratio of the plots are calculated by the measure method

Measure result	The number of plots is the case/%		The area of the plot is the ratio/%	
	Preoptimize	After optimization	Preoptimize	After optimization
Low mixing	41.7	52.94	45.02	60.66
Medium low mixing	19.51	19.26	30.89	25.62
Medium mixing	15.69	14.33	15.41	9.15
Medium high mixing	13.94	9.53	6.61	3.34
High mixing degree	9.34	4.07	2.24	0.59

3.2 Characterization of the functional mix of the city

3.2.1. General distribution characteristics. Generally speaking, the main urban area of City A shows a high-mixed spatial distribution pattern of "center clustering and multi-point scattering".

1) Center agglomeration. The attenuation curves of different land use function mixing degree and the distance from the center of the old city are shown in Figure 1. Generally speaking, the old city is more mature and has a high degree of land use mixing, while the overall functional mixing degree of the areas outside the old city is low, which is still far from a comprehensive and multi-functional urban environment. The city of City A mainly shows the outward expansion centered on the old city, and the spatial distribution of functional mixing degree corresponds to it, which is significantly higher than that of the peripheral areas of the old city. In addition to the differences between the old and new cities, the construction sequence within each area also affects the degree of mixing. For example, in Hexi area, bounded by Jiqingmen Street, the development of the northern part of Hexi is 10-20 years earlier than the development of the southern part of Hexi, and the degree of functional mixing is 130.25% higher than that of the southern part.

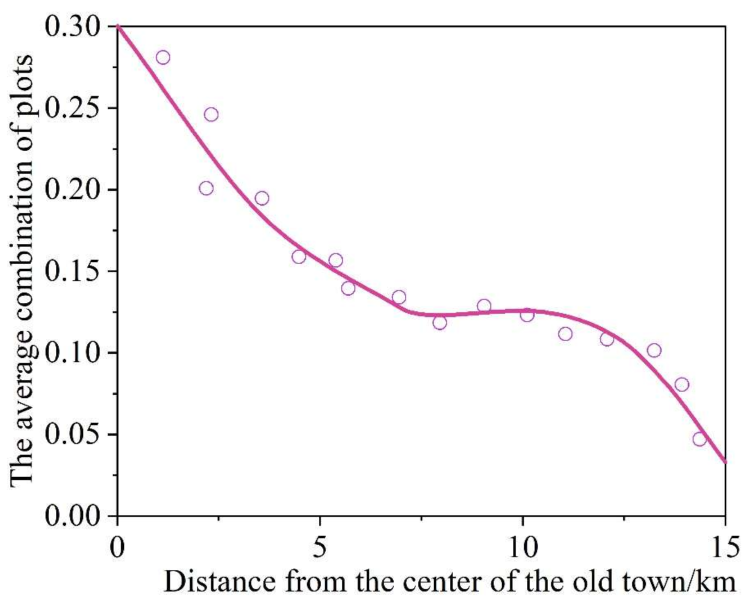


Fig. 1. The function of different land functions and the distance attenuation curve

2) Multi-point distribution. The multi-point scattering characteristics of the main urban area are shown in Figure 2. Except for the slice-shaped agglomeration area centered on Xinjiekou in the old city, the rest of the areas with high functional mixing degree have scattered distribution characteristics, among which there are still 8 agglomerations with medium-high and above mixing degree in the old city, and 26 agglomerations outside the old city, with the average distance of each agglomeration point being 1.9km. According to the average closest-neighbor analysis to test the statistical significance of spatial autocorrelation, the final standard deviation multiplier Z value score is greater than 0, indicating that the agglomerations as a whole are in the state of decentralized arrangement. Superimposing the boundaries of commercial center system, subway stations and mixed-degree agglomeration points, it is found that they are highly correlated spatially.

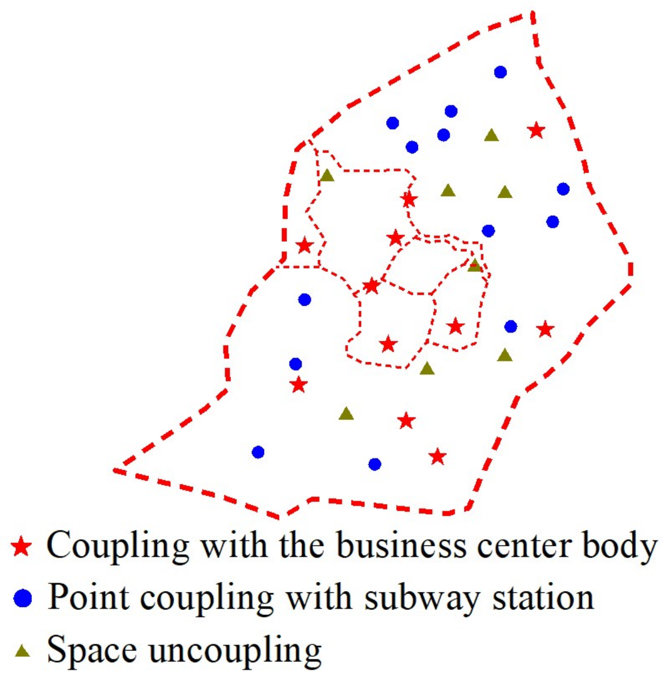


Fig. 2. The main urban area is dotted with the characteristics of multi-point

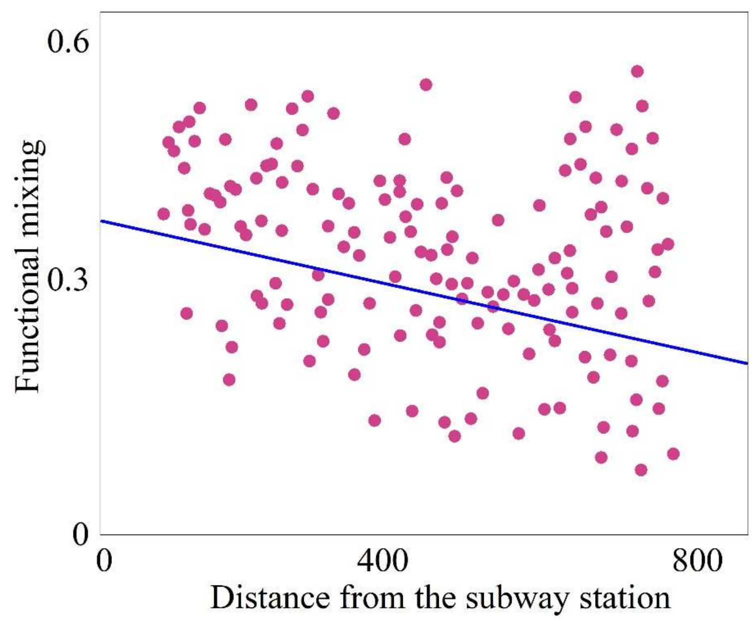
3.2.2. Localized aggregation characteristics:

1) Patchy Aggregation. Large continuous areas of medium-high and higher mixing degree are mainly concentrated in the south-central part of the old city. Xinjiekou, as the largest and most fully developed traditional business district in City A, has a large degree of influence on the mixing degree of land use and a wide range of influence. In Xinjiekou, 21.55% of the total number of sites in the main urban area have medium-high or above mixing degree, and the average functional mixing degree is as high as 0.41 in the 400m core circle, then drops to 0.34 in the 800m circle, and gradually decreases in the more peripheral circles.

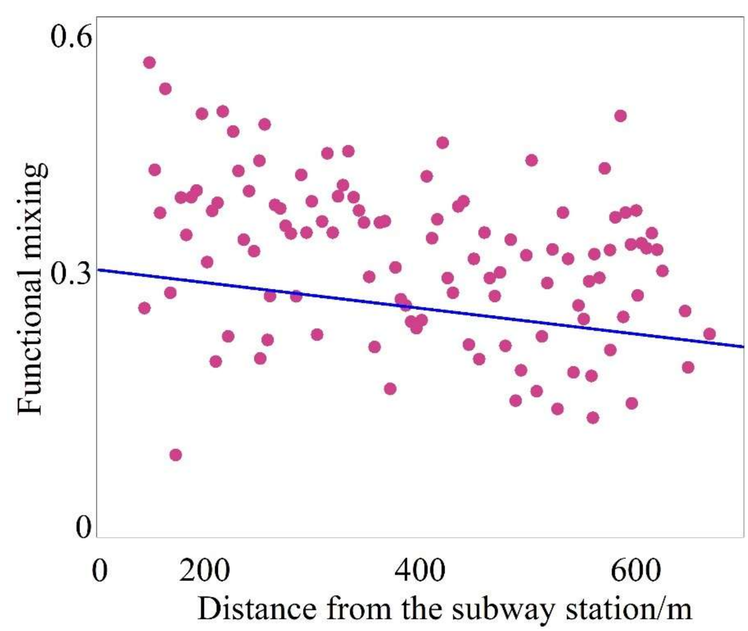
2) Linear aggregation. Some of the areas with medium-high or higher mixing degrees are concentrated on both sides of the roads, forming a linear aggregation pattern based on the main roads. For example, there is a high proportion of medium-high mixing areas within 200 meters along the main roads of the city such as Jiangdong North Road and Jiangdong Middle Road, which are highly correlated with the layout of commercial offices and residential land.

3) Point Agglomeration. The distribution of the functional mixing degree of Fuzimiao, Floating Bridge and Xinglong Street subway stations is shown in Figure 3 (Figures a~c are Fuzimiao subway station, Floating Bridge subway station and Xinglong Street subway station, respectively). In the main urban area, there are also a small number of areas with high values of functional mixing degree and point-like aggregation distribution characteristics, mainly relying on the development of rail transit stations

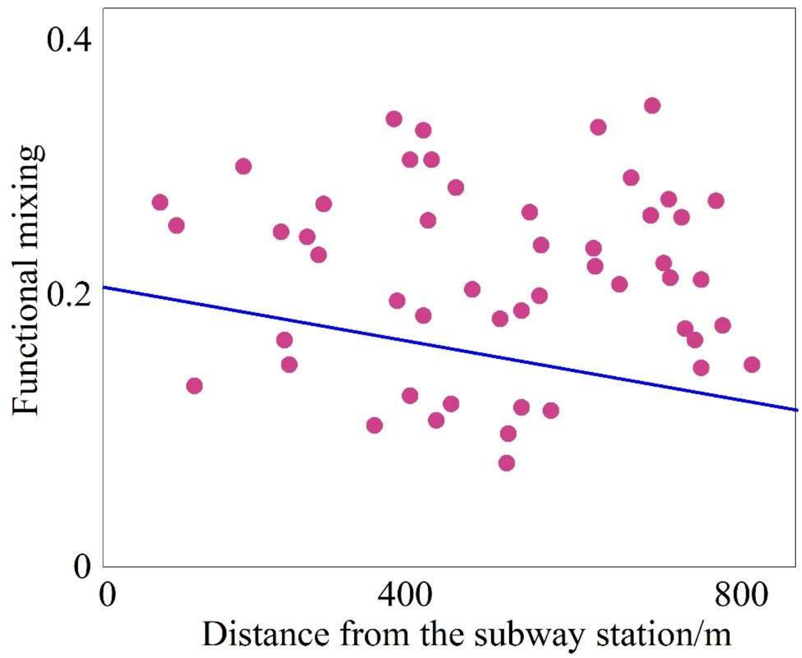
or low-grade commercial centers, mainly commercial or public service functions, with obvious spontaneity. This study takes the subway station as the center of the circle to delimit the sites within 800m, uses the Euclidean distance tool to calculate the distance between each site and the nearest subway station, and then carries out a linear regression analysis between the results of the calculation and the degree of functional mixing. The study selects three subway station sites with obvious distribution of medium-high and above mixing degree clusters, namely, Fuzimiao, Floating Bridge and Xinglong Street, for analysis. The following pattern can be drawn: the degree of functional mixing is negatively correlated with the distance of the site from the subway station, i.e. the closer the site is to the rail transit station, the higher the degree of functional mixing is.



(a)Fu zitan subway station



(b) Floating bridge subway station



(c) Boom street station

Fig. 3. Functional mixing distribution

4. Introduction to the GWR model

4.1 OLS modeling

One-way linear regression refers to the use of one influence to explain the dependent variable. Let the dependent variable be $y, x_1, x_2, \cdots, x_p$ be the independent variable and when the relationship between the independent variable and the dependent variable is linear, the multiple linear regression model is (OLS) is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{5}$$

Here, $\beta_0, \beta_1, \ldots, \beta_p$ is the $p+1$ unknown parameter and ε is the random error that satisfies the following assumptions:

$$\begin{cases} E(\varepsilon) = 0 \\ Var(\varepsilon) = \sigma^2 \end{cases} \tag{6}$$

For a practical problem, if n set of observations $(x_{i1}, x_{i1}, \cdots, x_{ip}; y_i)(i = 1, 2, \cdots, n)$ is obtained, the linear regression model can be expressed as:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ \cdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n \end{cases} \tag{7}$$

is written in matrix form as:

$$y = X\beta + \varepsilon \tag{8}$$

Among them:

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}, \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{9}$$

In order to facilitate parameter estimation and hypothesis testing of the model, the following assumptions are made about the regression equation:

1) The explanatory variable $x_1, x_2, \cdots, x_p$ is a deterministic variable, not a random variable, and $rank(X) = p+1 < n$ is required, which means that X is a full rank matrix and the columns of the independent variables are not correlated with each other, and the number of sample sizes should be greater than the number of explanatory variables.

2) The random error term satisfies a normal distribution with mean 0 and independent equal variances, i.e:

$$\begin{cases} \varepsilon_i \sim N(0,\sigma^2), & i=1,2,\cdots,n \\ \operatorname{cov}(\varepsilon_i,\varepsilon_j)=0, & i \neq j \end{cases} \tag{10}$$

The mean of the random error term is 0, which assumes that there is no systematic error in the observations, and the covariance of the random error term is 0, which indicates that the random error term is uncorrelated across sample points, is not serially correlated, and has the same precision.

Each regression parameter $\beta_0, \beta_1, \ldots, \beta_p$ can be regarded as a "slope" between an independent variable and a dependent variable, the regression solution is the solution of β, in the regression equation, it is to find the estimated value $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$ of parameter $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$, so that the sum of squares of the dispersion reaches a very small level, and the solution methods include least squares method, gradient descent method, etc., but the least squares method is generally used:

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)^T = (X^T X)^{-1} X^T Y \tag{11}$$

4.2 The Moran'I index

In multiple linear regression analysis, the slopes in the model are assumed to be generalized within the study area, and the model ignores the spatial properties among the variables because the homogeneity of the relationships among the variables is assumed prior to the analysis [33]. This spatial non-stationarity should be taken into account when analyzing spatial data, leading to new modeling strategies. The principle of Moran's index is the final representation of spatial correlation through the product of attributes and spatial relationships, which is calculated by the formula:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{12}$$

where z_i is the deviation of the attribute of element i from its mean value, i.e., $z_i = (x_i - \hat{x})$, $w_{i,j}$ is the spatial weight between element i and element j, n is the total number of elements, and S_0 is the aggregation of all spatial weights:

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} \tag{13}$$

4.3 Geographically weighted regression model (GWR)

The geographically weighted regression model follows the first law of geography: all things are related, and the closer they are to each other, the more closely they are related. Information about the spatial location of the data is also included in the regression parameters when performing the regression, and point-by-point parameter estimation is performed using locally weighted least squares. The non-stationarity of the spatial relationships can be detected visually by observing how the parameter estimates at each geospatial location vary with geospatial location, and it is also possible to see the role of each variable in different geographic locations. The GWR coefficients are compared to the global regression coefficients as a function of spatial location variations in a model that allows for the existence of different relationships at different points in space [34]. It is modeled as follows:

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{ik} + \varepsilon_i \tag{14}$$

The spatial weight matrix is a diagonal matrix with each element on the diagonal being a function of the observed position. Let the weighting matrix at position i be $W(u_i, v_i)$. Then the parameter vector at position i is estimated as:

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y \tag{15}$$

Among them:

$$W(u_i, v_i) = diag(w_{i1}, w_{i2}, \cdots, w_{in}) \tag{16}$$

When estimating the parameters for the estimated regression point i, more weight should be given to the points near i. From a mathematical point of view, the relationship between w_{ij} and d_{ij} can be expressed by choosing a continuously monotonically decreasing function, thus achieving the purpose of the more distant the less weight. Although there are many functions that satisfy the continuous monotonic decreasing function, the most widely used is the Gaussian function, which has the following functional form:

$$w_{ij} = \exp[-(d_{ij} / b)^2] \tag{17}$$

Where d_{ij} is the Euclidean distance from point i to point j, the closer the two points are, the greater the weight between them. As d_{ij} becomes larger, this weight decreases as d_{ij} increases, gradually approaching zero. Parameter b is the bandwidth, which is a non-negative decay parameter that controls the rate at which the weights decrease and describes the functional relationship between weights and distance. The larger the bandwidth, the slower the weights decay with distance, and the smaller the bandwidth, the faster the weights decay with distance. In the two most extreme cases: when $b = 0$, only regression point i has a weight of 1, and the weights of the other observations tend to 0, when there is only $\hat{y}_i = y_i$, i.e., the estimation process is just a re-representation of the data. And when b tends to infinity, the weights of all observations tend to 1, at which point the model becomes a traditional least squares. For a given bandwidth, when $d_{ij} = 0$, $w_{ij} = 1$, the weights reach the maximum, and as the distance of the data points from the regression point increases, w_{ij} it gradually decreases, and when j the points are far enough away from the i points, d_{ij} it approaches 0, i.e., these points have almost no effect on the parameter estimation of the regression point.

In the practical use of geographically weighted regression, the results of the regression analysis show a more sensitive response to changes in bandwidth than the choice of kernel function. If the bandwidth is chosen too large, the bias of the regression parameter estimates will be too large. However, if the bandwidth is too small, the variance of the regression parameter estimates will be too large. In geographically weighted regression analysis, the least squares method cannot be used for bandwidth selection, because for $\sum_{i=1}^{n}[y_i - \hat{y}_i(b)]^2 = \min$, the smaller the bandwidth selected, the smaller the weight of the data points participating in the regression analysis, so that $\hat{y}_i(b)$ will be closer to y_i, and therefore the value of $\sum_{i=1}^{n}[y_i - \hat{y}_i(b)]^2$ will be smaller, which means that the optimal bandwidth selected by the least squares method is the narrow area that contains only one sample point.

The most common choices for bandwidth are the cross-validation method and the minimum information criterion (AIC), which was used in the article to do the geographically weighted regression.The AIC criterion, modified from the great likelihood, takes into account the number of independent parameters and improves the quality of the fit by means of increasing the number of free parameters, thus avoiding overfitting. With this approach, a model is found that best explains the data but contains the fewest number of free parameters. For the general case, let the likelihood function of the model be $\ln(\theta, X)$, θ have dimension p and x be a random sample, then the AIC is defined as:

$$AIC = -2\ln(\hat{\theta}_L, x) + 2q \tag{18}$$

Where $\hat{\theta}_L$ is the maximum likelihood estimate of θ and q is the number of unknown parameters, the size of the AIC depends on the number of independent parameters and the value of the model's maximum likelihood function, which is a weighted function of the accuracy of the fit and the number of unknown parameters.The smaller the value of the AIC, the larger the value of the maximum likelihood function, which indicates that the model is accurate, and the smaller the value of the AIC.The AIC not only improves the degree of fit of the model, but it also introduces a penalty term to keep the model parameters as small as possible, which helps reduce the likelihood of overfitting.

In geographically weighted regression analysis the weight function bandwidth is chosen with the formula:

$$AIC = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left[\frac{n + tr(S)}{n - 2 - tr(S)}\right] \tag{19}$$

$tr(S)$ is a function of bandwidth b, and $\hat{\sigma}$ is a great likelihood estimate of the variance of the random error term, i.e., the smaller the values of $\hat{\sigma} = \left|\frac{RSS}{n - tr(S)}\right|$, $\ln(\hat{\sigma})$, the better the fitted model. For data from the same sample cohort, the bandwidth corresponding to the geographically weighted regression weight function that minimizes the AIC value should be chosen.

5. Analysis of the role of spatial heterogeneity of influencing factors

5.1 Model-based analysis of influencing factors

In this section, a linear form of the function is used and the estimation method utilizes the Ordinary Least Squares (OLS) method for model construction and computation. The Ordinary Least Squares (OLS) tool of ArcGIS 10.2 software was utilized for the construction and calculation of the function, and the results of the OLS model parameter estimation are shown in Table 2. In the table, as shown by the robust probability, the eight indicators of influencing factors, namely, common habits of life, neighborhood communication and interaction, social network establishment, psychosocial distance, participation in community activities, community care and support, social contact behaviors, and external behavioral interference, are significantly correlated at a significance level of 5%. Social network establishment and community activity participation influencing factors have a positive correlation effect on the influence of neighborhood relations and social inclusiveness, as can be seen by the indicator values of VIF in the table, all the VIF values are small, which indicates that there is no covariance problem between the independent variables and the independent variables, and the results are consistent with the results of the analysis of the covariance of the influencing factors.

Table 2. OLS model parameter estimation results

Variable	Coefficient	Standard deviation	T statistic	probability	Robust probability	VIF
Intercept	432.61	21.63	12.623	0.000000	0.000000	
Common habits of life	-91.395	22.843	-4.017	0.000000	0.000000	1.475
Neighborhood exchanges	-216.192	67.201	-3.22	0.00134	0.02825	1.373
Social networking	0.106	0.1	2.285	0.01942	0.01581	2.661
Social psychological distance	-0.016	0.108	-0.361	0.97549	0.97646	1.332
Community activity	0.201	0.134	1.579	0.10982	0.03269	1.379
Community care support	-0.319	0.137	-2.644	0.00964	0.00359	1.473
Social contact behavior	-0.409	0.106	-5.355	0.000000	0.000000	2.406
External behavior interference	-0.058	0.738	-0.079	0.98069	0.97967	1.098

5.2 Spatial Heterogeneity Analysis of Influential Factors Based on GWR

The geographically weighted regression model can well explain the changes of neighborhood and social inclusion in different spatial locations and the characteristics of influencing factors on their spatial heterogeneity, therefore, this paper applies the geographically weighted regression model to further simulate the spatial evolution patterns of influencing factors of neighborhood and social inclusion.

The regression coefficient of the estimated parameters in the OLS model is a global stable value, which cannot well explain the phenomenon of spatial heterogeneity of neighborhood and social inclusion caused by the influencing factors, whereas in the GWR model, the estimated regression coefficient is a dynamic value affected by the localization of adjacent observations. In this paper, the sample data has spatial non-homogeneity, and at the same time, combined with the characteristics of the sample data itself, the Gaussian weight function method is selected to determine the spatial weight function, and the bandwidth is determined in the geographically weighted regression using the AICc criterion method, which is a common method for optimal bandwidth selection. The model influencing factors finally selected eight factors are common habits of life, neighborhood communication and interaction, social network establishment, social psychological distance, community activity participation, community care support, social contact behavior and external behavioral interference.

In this paper, the geographically weighted regression modeling tool of Arc GIS10.2 software was used for regression analysis, and the comparison of the fitting effects of the OLS model and the GWR model is shown in Table 3. The model fit R^2 increased from 0.55 to 0.62 in OLS, the corrected R^2 increased from 0.5 to 0.57 in OLS, and the AICc indicator decreased from 12922.32 to 12634.62 in OLS. The results of this model fully demonstrate that local regression based on geographically weighted regression model effectively improves the simulation accuracy, which is superior to that of the General Linear Regression (OLS) model calculation. Results.

Table 3. The OLS model is compared with the GWR model

Model	R^2	Correct R^2	AICc
OLS	0.55	0.5	12922.32
GWR	0.62	0.57	12634.62

Geographically weighted regression analysis with 670 valid sample data of streets in the main urban area of city A, to derive the regression coefficients corresponding to the degree of mixing of urban functions and each influencing factor, this paper selects the statistic of descriptive degree of contribution, and the descriptive statistics of the regression coefficients of the dependent variable of GWR are shown in Table 4. As seen in the table, the four types of influencing factors, namely, common habits of life, psychosocial distance, social contact behavior and external behavioral interference, are positively correlated with the changes in neighborhood relations and social inclusiveness in the streets of the main urban area of A. The standard deviation of these four influencing factors is 393.0%. Their standard deviations are 393.299, 0.191, 0.217 and 0.288 respectively.

Table 4. The GWR factor regression coefficient is descriptive

Variable	Minimum value	Lower quadrate	Median	Last quarter	Maximum value	Average	Standard deviation
Common habits of life	568.417	1543.212	1822.005	2004.668	2319.793	1737.978	393.299
Neighborhood exchanges	-184.838	-110.646	-73.514	-64.002	11.501	-88.054	34.942
Social networking	-352.696	-246.103	-233.322	-198.962	-62.94	-217.968	43.71
Social psychological distance	-0.236	0.22	0.28	0.413	0.548	0.292	0.191
Community activity	-0.345	-0.012	0.093	0.387	0.673	0.129	0.265
Community care support	-0.787	-0.555	-0.475	-0.406	-0.056	-0.466	0.159
Social contact behavior	-0.892	-0.782	-0.548	-0.401	-0.082	-0.635	0.217
External behavior interference	1.724	2.464	2.461	2.641	3.193	2.519	0.288

In order to test the overall fitting effect of the model, this paper further analyzes the residuals of the results obtained from geographically weighted regression (GWR), and the ideal result should be that the residuals are randomly distributed in space. In this paper, the residuals are analyzed for spatial autocorrelation, and the spatial distribution of the standard residual values of the GWR model is shown in Figure 4. From the figure, we can see that Moran's I index is 0.014125, and the Z score value is 1.516637, the results show that the residual values are completely randomly distributed in space, and the overall fitting results of the GWR model are good.

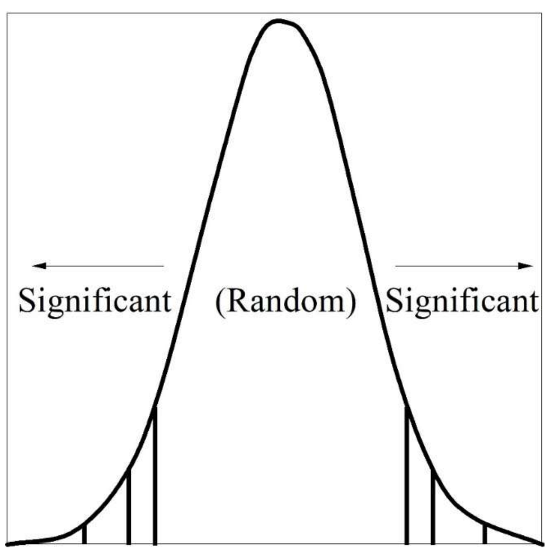


Fig. 4. GWR model standard residual space distribution

6. Conclusion

This paper takes the streets in the main urban area of City A as the study area, and conducts GIS spatial analysis and geographically weighted regression modeling for collecting various types of data to explore the impact of urban functional mixing on neighborhood relations and social inclusion from both global and local perspectives.

In the experiments comparing the fitting effect of the OLS model and the GWR model, the model fit R^2 was improved by 0.07, the correction R^2 was improved by 0.02, and the AICc index was reduced from 12922.32 to 12634.62 in the OLS, and the localized regression based on the geographically weighted regression model effectively improves the simulation precision, which is superior to that of the results calculated by the general linear regression (OLS) model.

Through the descriptive statistics of the regression coefficients of the dependent variable with GWR, the four categories of influencing factors, namely, common habits of life, psychosocial distance, social contact behavior and external behavioral interference, play a positive role in the changes of neighborhood relations and social inclusion in the streets of A main city.

References

[1] Liu, B., Deng, Y., Li, M., Yang, J., & Liu, T. (2021). Classification schemes and identification methods for urban functional zone: A Review of Recent Papers. Applied Sciences, 11(21), 9968.

[2] Pozoukidou, G., & Chatziyiannaki, Z. (2021). 15-Minute City: Decomposing the new urban planning eutopia. Sustainability, 13(2), 928.

[3] Mouratidis, K. (2018). Built environment and social well-being: How does urban form affect social life and personal relationships?. Cities, 74, 7-20.

[4] Jennings, V., & Bamkole, O. (2019). The relationship between social cohesion and urban green space: An avenue for health promotion. International journal of environmental research and public health, 16(3), 452.

[5] Mouratidis, K. (2019). Compact city, urban sprawl, and subjective well-being. Cities, 92, 261-272.

[6] Xia, C., Yeh, A. G. O., & Zhang, A. (2020). Analyzing spatial relationships between urban land use intensity and urban vitality at street block level: A case study of five Chinese megacities. Landscape and Urban Planning, 193, 103669.

[7] Hao, F., Zhang, J., Duan, Z., Zhao, L., Guo, L., & Park, D. S. (2020). Urban area function zoning based on user relationships in location-based social networks. IEEE Access, 8, 23487-23495.

[8] Schmidt, A. (2023). Health and the City: How Cities Impact on Health, Happiness, and Well-Being. In The Palgrave Encyclopedia of Urban and Regional Futures (pp. 812-819). Cham: Springer International Publishing.

[9] Ryckewaert, M., Zaman, J., & De Boeck, S. (2021). Variable arrangements between residential and productive activities: conceiving mixed-use for urban development in brussels. Urban Planning, 6(3), 334-349.

[10] Ye, C., Zhang, F., Mu, L., Gao, Y., & Liu, Y. (2021). Urban function recognition by integrating social media and street-level imagery. Environment and Planning B: Urban Analytics and City Science, 48(6), 1430-1444.

[11] Chen, C., Yan, J., Wang, L., Liang, D., & Zhang, W. (2020). Classification of urban functional areas from remote sensing images and time-series user behavior data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 1207-1221.

[12] Du, S., Du, S., Liu, B., Zhang, X., & Zheng, Z. (2020). Large-scale urban functional zone mapping by

integrating remote sensing images and open social data. GIScience & Remote Sensing, 57(3), 411-430.

[13] Zhang, Y., Li, Q., Tu, W., Mai, K., Yao, Y., & Chen, Y. (2019). Functional urban land use recognition integrating multi-source geospatial data and cross-correlations. Computers, Environment and Urban Systems, 78, 101374.

[14] Zhou, Y., Tu, M., Wang, S., & Liu, W. (2018). A novel approach for identifying urban built-up area boundaries using high-resolution remote-sensing data based on the scale effect. ISPRS International Journal of Geo-Information, 7(4), 135.

[15] Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., & Hong, Y. (2017). Classifying urban land use by integrating remote sensing and social media data. International Journal of Geographical Information Science, 31(8), 1675-1696.

[16] Lin, A., Huang, B., Wu, H., & Luo, W. (2024). An MIU-based deep embedded clustering model for urban functional zoning from remote sensing images and VGI data. International Journal of Applied Earth Observation and Geoinformation, 128, 103689.

[17] Zhang, Y., Qin, K., Bi, Q., Cui, W., & Li, G. (2020). Landscape patterns and building functions for urban land-use classification from remote sensing images at the block level: a case study of Wuchang District, Wuhan, China. Remote Sensing, 12(11), 1831.

[18] Lin, A., Sun, X., Wu, H., Luo, W., Wang, D., Zhong, D., ... & Zhu, J. (2021). Identifying urban building function by integrating remote sensing imagery and POI data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 8864-8875.

[19] Xu, S., Qing, L., Han, L., Liu, M., Peng, Y., & Shen, L. (2020). A new remote sensing images and point-of-interest fused (RPF) model for sensing urban functional regions. Remote Sensing, 12(6), 1032.

[20] Mao, L., Zheng, Z., Meng, X., Zhou, Y., Zhao, P., Yang, Z., & Long, Y. (2022). Large-scale automatic identification of urban vacant land using semantic segmentation of high-resolution remote sensing images. Landscape and Urban Planning, 222, 104384.

[21] Bao, H., Ming, D., Guo, Y., Zhang, K., Zhou, K., & Du, S. (2020). DFCNN-based semantic recognition of urban functional zones by integrating remote sensing data and POI data. Remote Sensing, 12(7), 1088.

[22] Tu, W., Hu, Z., Li, L., Cao, J., Jiang, J., Li, Q., & Li, Q. (2018). Portraying urban functional zones by coupling remote sensing imagery and human sensing data. Remote sensing, 10(1), 141.

[23] Chen, Y., He, C., Guo, W., Zheng, S., & Wu, B. (2023). Mapping urban functional areas using multi-source remote sensing images and open big data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

[24] Zhao, W., Li, M., Wu, C., Zhou, W., & Chu, G. (2022). Identifying urban functional regions from high-resolution satellite images using a context-aware segmentation network. Remote sensing, 14(16), 3996.

[25] Zhou, W., Ming, D., Lv, X., Zhou, K., Bao, H., & Hong, Z. (2020). SO–CNN based urban functional zone fine division with VHR remote sensing image. Remote Sensing of Environment, 236, 111458.

[26] Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., ... & Qiu, G. (2020). Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS Journal of Photogrammetry and Remote Sensing, 163, 82-97.

[27] Xie, L., Feng, X., Zhang, C., Dong, Y., Huang, J., & Liu, K. (2022). Identification of urban functional areas based on the multimodal deep learning fusion of high-resolution remote sensing images and Social Perception Data. Buildings, 12(5), 556.

[28] Qian, Z., Liu, X., Tao, F., & Zhou, T. (2020). Identification of urban functional areas by coupling satellite images and taxi GPS trajectories. Remote Sensing, 12(15), 2449.

[29] Chen, W., Huang, H., Dong, J., Zhang, Y., Tian, Y., & Yang, Z. (2018). Social functional mapping of urban green space using remote sensing and social sensing data. ISPRS Journal of Photogrammetry and Remote Sensing, 146, 436-452.

[30] Chang, S., Wang, Z., Mao, D., Liu, F., Lai, L., & Yu, H. (2021). Identifying urban functional areas in China's Changchun City from Sentinel-2 Images and social sensing data. Remote Sensing, 13(22), 4512.

[31] Tiange Zhang,Xia Zhu,Yuanping Liu,Cui Jia & Huimin Bai. (2024). Research on the Spatial Structure of the Beijing–Tianjin–Hebei Urban Agglomeration Based on POI and Impervious Surface Coverage. Buildings(6),1793-1793.

[32] Na Li & Miao Zhang. (2024). Distribution characteristics and influencing factors of tourist attractions in Tianjin based on POI big data. Environment, Development and Sustainability(prepublish),1-27.

[33] Wondwossen Taddesse Gedamu,Uwe Plank Wiedenbeck & Bikila Teklu Wodajo. (2024). A spatial autocorrelation analysis of road traffic crash by severity using Moran's I spatial statistics: A comparative study of Addis Ababa and Berlin cities.. Accident; analysis and prevention107535-107535.

[34] Thabiso Moeti,Tholang Mokhele & Solomon Tesfamichael. (2024). Associating socioeconomic factors with access to public healthcare facilities using geographically weighted regression in the city of Tshwane, South Africa.. Geospatial health(2).