Study on Urban Spatial Function Mixture and Individual Activity Space From the Perspectives of Resident Activity

LINA LIU\(^1\), HONG CHEN\(^1\), AND TAO LIU\(^2\)

\(^1\)College of Transportation Engineering, Chang’ an University, Xi’an 710064, China
\(^2\)College of Cyber Science, Nankai University, Tianjin 300350, China

Corresponding author: Hong Chen (glch@chd.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 52002030, in part by the Humanities and Social Sciences Foundation of the Ministry of Education under Grant 20XJCZH011, and in part by the Fundamental Research Funds for the Central Universities under Grant 300102219301.

ABSTRACT The research on the relationship between residents’ daily activities and urban spatial structure is of considerable significance to urban planning engineering and the organization of urban functions. However, little research considers the perspective of micro-spatial scale or resident perception. The increasing user-generated activity check-in data in social networks provides a database for this research. In this study, we first divided the urban space into nine functions that satisfy the residents’ activities, then used the small-scale grid to divide the city blocks and used information entropy to evaluate the mixed degree of land use functions. We then introduced the latent Dirichlet allocation (LDA) topic model to identify 15 mixed patterns of land use functions and each spatial unit’s topic distribution. Moreover, the JS divergence index was employed to measure spatial units’ similarity, fit the distance-activity intensity decay curve, and studied the influence of the individual spatial function distribution choice. We demonstrate that in urban space, residents’ daily activities mold the blending of urban area functions and shift single-function urban planning to mixed-use, consisting of single-function dominant and multi-function mixed. Besides, the functional complementarity between the activity units weakens the distance attenuation effect of the activity-space interaction intensity to some extent. The research on the interaction between active space and spatial activities expect to support the combination of urban land use types, the layout of facilities, and the guidance of residents’ activities.

INDEX TERMS Mixed urban land use, spatial interaction, text analysis, topic modeling.

I. INTRODUCTION

With the development of the transportation industry and information communication technologies (ICTs), human movement is becoming more powerful, resulting in activity compression of time and space [1]. Space distance is no longer a barrier to activities, making people’s activities more fragmented, personalized, and sophisticated in time and space [2], [3]. Mobile phones and Global Positioning System (GPS) have recorded many urban population activity data, providing opportunities for exploring the urban spatial structure and functional zoning [4]. The “human” factor has received more attention in urban space research, emphasizing the relationship between human behaviors and space from human subjectivity perspectives. Human activities can utilize the function of land with fewer space-time limits, and the structure of the function of land becomes complex, which changes the formation of the urban multi-functional activity area and then affects the traditional urban functional zoning.

The relationships among residents, places, and activities are considerably complicated, and the traditional space theory based on the law of distance decay is challenging. Therefore, the study of the interactions between urban space and residents’ behavior should be based not only on the conditions of space and land but also on personal behavior. The interaction between the urban spatial functional organization and residents’ behavior has become a research hotspot. More and more scholars have paid their attention to optimizing urban living space’s functional structure and balancing urban infrastructure resources to people’s perspective. Material space and activity space explain urban space from static and dynamic angles [5]. Urban spatial structure is various urban functional regions’ geographical position and its combination. For urban material space, the urban areas are represented by kinds of functional regions and their combinations, while for urban activity space, the urban area is various types of residential...
The rapid development of ICT and LBS technology provides a new opportunity for data collection of spatial behavior and behavioral space research. Large-scale real-time activities tracking, scene interaction and exploratory analysis produced many spatial behavior-based mining algorithms. Besides, the people affected by ICTs often have frequent network activities and high space mobility. The submission and sharing of geo-tagged information supported by social media sites have facilitated crowdsourced data. Since much of the information is user-generated and location-specific, Goodchild coined the term voluntary geographic information (VGI) for it [7]. More and more platforms offer location-based service (LBS), making it easy for users to record and share the time, location, content, and emotion. Objectively, it has retained the traces of its space-time activities to the maximum extent. The people’s activities play an essential role in reshaping urban space’s functional structure [8]. Therefore, analyzing the spatial distribution of resident check-in data is essential to study the urban spatial structure and function. Although a person’s activities are entirely random, as the number of activities increases, it is likely to find potential patterns. On the one hand, at the physical space level, these data enable us to study the relationship between spatial activity distribution in a region and spatial functions. On the other hand, at the residents’ activities, this paper studies the relationship between spatial function distribution and people’s activities distribution.

In this study, the functional zoning of urban space and land use break through the traditional space theory of the more substantial limits of the scope of activities, the use of ICT development brought about by the data resources, considering the impact of human subjective activities. We used social media check-in data to examine how urban residents’ activities and urban spaces mold each other. On the one hand, the mixed-use of urban land functions is discussed based on residents’ urban space activities’ distribution and intensity. On the other hand, the relationship between activity intensity and activity distance discusses urban space functions’ mixed background.

The remainder of this paper is organized as follows. Section 2 presents a brief overview of current studies related to spatial behavior and behavioral space, region function, and space-behavior interactions. Section 3 introduces the study area and data. Section 4 describes the methodology of modeling the function mixed pattern of urban units and activity-space interaction under spatial similarity. Section 5 focuses on the analysis and results. Section 6 summarizes the main findings of this study and points out the directions for future studies.

II. RELATED WORK

In this section, the relevant research works are highlighted in three categories, namely, spatial behavior and behavioral space, urban region extraction function, and space-behavior interactions.

A. SPATIAL BEHAVIOR AND BEHAVIORAL SPACE

Spatial behavior is the choice result of a resident under the space restriction. Behavioral space is how individuals meet their needs and carry out their daily activities [9].
analysis tools, and software platforms [20]. ICT has led to activity feature reconstruction and weakens spatial and temporal human activities [21]. At present, the impact of ICT on spatial behavior studies mainly focuses on the macro-scale, including the urban spatial structure, influencing factors [22], and the urban system in network space [23].

It is an important research direction to use space-behavior interaction theory to do a case study, model analysis, and methodology construction. Research content that focused on spatial behavior expands the use of ICT and spatial constraints. Big data from social networks combine with individual behavior, consider databases with different geographic scales, different accuracy, and rich attributes. In terms of research methods, the city is regarded as a collection of individual activities, behaviors, and reactions and described by “what happened” rather than the quantitative characteristics of land use types. Integrating the decision-making activities and behavior interactions can help the urban behavior space evolution from micro-individual action. The study of non-work behavior such as shopping and leisure provides essential theoretical and practical significance for optimizing commercial space, public service space, and transportation [24].

B. EXTRACTING URBAN REGION FUNCTION FROM LOCATION BASED NETWORK

Human activity usually occurs in different types of (Point of Interest) POIs, and the region function depends on the activity type [25]. The same type POI can locate in different land using types that support different functions. It helps to make up the semantic difference between land use classification and urban function. Integrated Circuit Card (I.C. card) data, GPS data, mobile phone data, and social media data have fine-grained spatial analysis capabilities, helping to understand urban space utilization levels for individuals and groups. Long et al. [26] identify the urban functional regions in Beijing using I.C. card records and POIs data. Yuan et al. [27] adopt the topic model using the GPS trajectory data and POIs data to find the different urban functional regions. As crowdsourced data content providers, social media contains a wealth of information about residents’ spatial interactions and the location semantics, which has widely used to understand the urban functional structure and how residential activities mold urban systems [28]. At present, many scholars use social media data [29] to study urban dynamics. Using Twitter messages, Han et al. [30] determined users’ Spatio-temporal distribution, analyzed people’s co-occurrence density change and revealed the relationship between urban regions. Long et al. [31] used land functions reflected by POIs data and Sina Weibo check-in data to overlay Beijing’s current and planned land use. After the grid processed the research area, use the mixed land use index method proposed by Frankie et al. [32] to calculate each grid’s degree. It evaluated a fine-scale land-use function mixing degree.

Under the joint influence of land use and economic activities, similar activities show a high degree of spatial aggregation and form functional zones. In order to meet residents’ needs, different functional units have gradually formed in the city. The spatial distribution and combination pattern of functional regions reflects the spatial distribution and combination of the city material elements, an integral part of urban spatial structure. The change of resident activity amount in urban space partly reflects the region’s function; that means, similar functional regions have similar resident activity patterns. Therefore, the resident movement’s spatial characteristics can apply to the study of urban functional regions [33]. The functional regions divide from the perspective of resident perception. Liu et al. [34] studied the taxi activity intensity change at different city locations, then used cluster analysis to find six functional regions. Combined with POI data, Yuan et al. [27] used taxi data to quantify the pedestrian flow between regions and built a D-Model to find the different functional regions. With the help of Weibo’s data, Wang Bo et al. [35] analyzed the dynamic change of residents’ movement space from time, space, and activity, and divided urban space into regional functions.

The past research mainly evaluated the mixed land use function according to different land-use types at the regional scale, and the research scale was coarse. Based on big crowdsourced data, research mainly focuses on identifying or evaluating the whole functional region and lacks the in-depth quantitative analysis of the refined region units’ mixed function. Besides, due to the lack of human-land interaction, region units are rarely classified from residents’ perception. Few studies employed POI information and human activities in LBS networks simultaneously to get the urban functional regions [36].

C. SPACE-BEHAVIOR INTERACTIONS

There are two paths to construct the coupling relationship between population heterogeneity and geospatial environmental heterogeneity in urban research. One uses the similarity of individual behavior characteristics [37] or the strength of social connection [38] to classify the population, deal with the heterogeneity, and determine the spatial distribution of different population categories. The other is dividing the urban space, quantifying the behavior characteristics in different spatial units, and classifying the spatial units. Due to the difficulty of obtaining individual granularity data and the lack of attribute dimension data [39], the current research is mainly carried out from the second route. (1) Geographical data reflect the spatial distribution of the corresponding activity events. Assuming that in different locations performing certain activities has a similar proportion over time, the extracted point distribution can be used as an estimated [40] population distribution model. (2) Text and photo with spatiotemporal labels can analyze the semantic information that reflects people’s cognition about the association of different geographical units [41]. (3) Geographic big data can quantify the spatial interaction intensity between geographic units [42]. Second-order features expressed by interaction patterns between geographic units reflect their
role in the urban spatial flow [43]. Spatial interaction between geographical entities helps us understand the region’s spatial structure and plan effective spatial form. The interaction intensity measure in terms of pedestrian flows [44], trade flows, currency flows, and even the co-occurrence of place names [45]. Due to the complexity of spatial interactions involving multiple spatial node pairs, various research has done to visualize spatial interactions effectively and delineate meaningful sub-regions [46], [47].

Most spatial interactions affected by distance decay [48], the gravity model and the power function fit the relationship between the distance and the intensity geographical entities intersection [49], [50]. Human behavior is affected by distance and the functional characteristics of the destination, which is especially important in daily activities. As the main body of behavior, the residents form specific cognition and preference to space and constrained by space [51].

At present, many studies use the data with spatiotemporal labels to illustrate the interaction patterns among geographical units. Less research focuses on the interaction between residential activities and geographical units. Looking for the anchor point of human activity from human activity space, we can use it as a virtual geographical entity and expand the interactive mode between humans and space. Besides, from individual residents’ perspectives, we can consider the relationship and rationality of different behavior spaces projected to the city’s physical space.

III. STUDY AREA AND DATA
A. STUDY AREA
New York City is about 1214.4 square kilometers, divided into five administrative districts, has a complex land type and abundant functions. Therefore, New York is chosen as the research city in this paper. Manhattan is the core of New York City and has the highest population density. Moreover, Manhattan was the first district to develop from an urban perspective, followed by the Bronx and Brooklyn, and Queens and Staten Island were the last. Because of the population density on Staten Island is very spare, only the other four districts are selected as the study area.

B. DATA SETS
POIs can introduce the necessary information of each functional unit in a city and represent city’s functional regions’ distribution to some extent, and it is also a spatial anchor point when other types of data are fused. To start the research, we need a series of check-in data. This study used a 227428 Foursquare check-in data set from New York City. Using ArcGIS software to select the data, we got 180074 data in the study area. Each check-in data includes user ID, venue ID, venue category, latitude, longitude, and time stamp. The POI data in this research is from the OpenStreetMap open data platform. There are 29,418 POIs in the study area that matches the check-in location. Each data includes the attributes name, latitude, longitude, category, and address. We formalize the resident check-in point as a set of points P = {p1, p2, ..., pn}. For each pi ∈ P, it includes latitude (pi.Lat), longitude (pi.Lon), and position category label (pi.Category).

IV. METHOD
The study was based on check-in data from social media and POI data. We discuss the mixed-use of urban land function at the fine-scale, divide the mixed degree and mode of land-use function according to activity intensity’s characteristics, and establish a functional area identification model. The interactive characteristics between residents and land functions are illustrated based on the mixture of regional functions and the distribution of individual residents’ activities.

First, the area studied was divided into grids. The land use function is then integrated according to the activity type and regarded as a POI label. Moreover, information entropy was used to describe the quantity ratio mixed degree of land-use. Based on the activities proportion in each spatial unit, a series of mixed land-use patterns were generated using the topic modeling method-Latent Dirichlet Allocation Analysis (LDA), it realizes the mixed pattern of identifying the function of the spatial unit and the representation of subject ratio by using the function distribution of each specific spatial unit. Finally, the Kullback-Leibler divergence (KL divergence) is employed to evaluate the similarity of function distribution among units, and the effect of unit function on activity intensity-distance was discussed at the individual level.

A. THE MIXED USE OF URBAN SPACE UNIT
The impacts of behavior on space are mainly reflected in the continuous adjustment of activity space for people to obtain pleasurable activities. Choosing the daily activity space, adjusting the long-term living space, and reconstructing the urban space (including the material space and the social space) in the choosing process all affect the urban space organization structure. From the perspectives of people’s activities, urban functional regions are represented by various types of residential space. Our study is based on the function division of urban activity space, the activity content, and intensity in the region, to study the mixed-function.

1) ACTIVITY CATEGORY CLASSIFICATION
The advantage of check-in data from social media is the ability to identify the activity purpose. Each check-in data reports the functional category information of the place visited. Therefore, the functional categories in the check-in data are very detailed, which are over 100 categories. When discussing land functions at a sufficient scale, such a detailed functional classification would make the feature distribution pattern hard to extract. Therefore, we classify the types of check-in location functions into nine categories in Table.1 according to the activities’ purpose [52].

The activity function cannot be determined entirely by the check-in location category. For instance, a place marked as entertainment could be the workplace for a resident, and a place marked as a home may not be the resident’s home but...
TABLE 1. Location function category classification.

| Function   | Type of location category                  |
|------------|---------------------------------|
| Home       | Home (private), Residential Building, Apartment / Condo |
| Work       | Design Studio, Office             |
| Eating     | Café, Food Joint, Food Place, Food Shop, Restaurant, Bakery, Deli/Bodega, Diner, Tea Room, Winery |
| Education  | School, College & University, Museum, College Academic Building, Elementary School, High School, Law School, Library, Medical School, Middle School, Music School, Nursery School, Planetarium, Professional & Other Places, Student Center, Trade School |
| Entertainment | Arcade, Arts & Entertainment, Bar, Beer Garden, Casino, College Stadium, College Theater, Comedy Club, Concert Hall, Event Space, Gastropub, General Entertainment, Movie Theater, Music Venue, Nightlife Spot, Other Nightlife, Performing Arts Venue, Photography Lab, Stadium, Theater |
| Recreation | Aquarium, Athletic & Sport, Beach, Bowling Alley, Campground, Garden, Gym / Fitness Center, Historic Site, Laundry Service, Moving Target, Other Great Outdoors, Outdoors & Recreation, Park, Playground, Pool, Public Art, Racetrack, Rest Area, Scenic Lookout, Sculpture Garden, Ski Area, Spa / Massage, Tanning Salon, Zoo |
| Shopping   | Antique Shop, Arts & Crafts Store, Automotive Shop, Bike Shop, Board Shop, Bookstore, Bridal Shop, Camera Store, Candy Store, Clothing Store, Convenience Store, Cosmetics Shop, Department Store, Donut Shop, Drugstore / Pharmacy, Electronics Store, Flea Market, Flower Shop, Furniture / Home Store, Gift Shop, Hardware Store, Hobby Shop, Jewelry Store, Mall, Market, Miscellaneous Shop, Mobile Phone Shop, Motorcycle Shop, Music Store, Newstand, Paper / Office Supplies Store, Pet Store, Plaza, Record Shop, Shop & Service, Smoke Shop, Sporting Goods Shop, Storage Facility, Thrift / Vintage Store, Toy / Game Store, Video Game Store, Video Store |
| Social service | Animal Shelter, Art Gallery, Bank, Brewery, Car Dealership, Car Wash, Castle, Cemetery, Church, City, Convention Center, Distillery, Embassy / Consulate, Factory, Fair, Financial or Legal Service, Fraternity House, Funeral Home, Government Building, Harbor / Marina, Hotel, Housing Development, Medical Center, Military Base, Mosque, Nail Salon, Neighborhood, Pet Service, Post Office, Recycling Facility, Salon / Barbershop, Shrine, Sorority House, Spiritual Center, Synagogue, Tattoo Parlor, Temple |
| Transportation | Airport, Bike Rental / Bike Share, Bridge, Bus Station, Ferry, Gas Station / Garage, General Travel, Light Rail, Parking, Rental Car Location, River, Road, Subway, Taxi, Train Station, Travel & Transport, Travel Lounge |

could be his/her friend’s home. In this case, we should be cautious in discussing the residents’ activities distribution. For other location categories, it is reliable to classify activity function by check-in.

2) BUILD A GRID LAND SEGMENTATION

Some studies of spatial-temporal data preprocessing divide the study area into several regular grids with the same distance. When studying the population spatial distribution, the grid size is set as 200m, 500m, or 1000m [53]. Then, use the human mobility characteristics reflected by big geographical data to illustrate the spatial distribution and interaction among grid units.

To estimate the spatial change of urban land-use patterns, we delineate the urban functional units at a fine scale. Divide the research area into 4006 regular square grids of 300m × 300m. The grids are represented by $G = \{A_{0,0}, A_{0,1}, \ldots, A_{i,j}\}$, in which $A_{i,j}$ is a unique unit grid, and $i, j \in \{0, 1, 2, \ldots, n\}$ are the serial number in the horizontal and vertical directions. In this way, map residents’ check-in records to the grids according to activity purposes. Then, we obtain the intensity spatial distribution map of various activities shown in Fig.1.

Then, use the human mobility characteristics reflected by big geographical data to illustrate the spatial distribution and interaction among grid units.

To estimate the spatial change of urban land-use patterns, we delineate the urban functional units at a fine scale. Divide the research area into 4006 regular square grids of 300m × 300m. The grids are represented by $G = \{A_{0,0}, A_{0,1}, \ldots, A_{i,j}\}$, in which $A_{i,j}$ is a unique unit grid, and $i, j \in \{0, 1, 2, \ldots, n\}$ are the serial number in the horizontal and vertical directions. In this way, map residents’ check-in records to the grids according to activity purposes. Then, we obtain the intensity spatial distribution map of various activities shown in Fig.1.

To estimate the spatial change of urban land-use patterns, we delineate the urban functional units at a fine scale. Divide the research area into 4006 regular square grids of 300m × 300m. The grids are represented by $G = \{A_{0,0}, A_{0,1}, \ldots, A_{i,j}\}$, in which $A_{i,j}$ is a unique unit grid, and $i, j \in \{0, 1, 2, \ldots, n\}$ are the serial number in the horizontal and vertical directions. In this way, map residents’ check-in records to the grids according to activity purposes. Then, we obtain the intensity spatial distribution map of various activities shown in Fig.1.

To estimate the spatial change of urban land-use patterns, we delineate the urban functional units at a fine scale. Divide the research area into 4006 regular square grids of 300m × 300m. The grids are represented by $G = \{A_{0,0}, A_{0,1}, \ldots, A_{i,j}\}$, in which $A_{i,j}$ is a unique unit grid, and $i, j \in \{0, 1, 2, \ldots, n\}$ are the serial number in the horizontal and vertical directions. In this way, map residents’ check-in records to the grids according to activity purposes. Then, we obtain the intensity spatial distribution map of various activities shown in Fig.1.
units are marked as single functional areas, accounting for 41.06%. There are a lot of concentrated mixed functional regions in Manhattan.

FIGURE 2. Single functional area and the mixed functional area.

Information entropy is a physical concept to measure the complexity and equilibrium of a system. The city is also a system in which the structure and formation is the mapping of the system function to space. The level of information entropy can reflect the equilibrium degree of land-use. The higher the entropy value is, the more the land functions. While the smaller the area difference between functions type is, and the more balanced land distribution [55]. Quantifying the mixed land-use can reflect the equilibrium degree of land-use. The higher the proportion of each spatial unit activity throughout the city to generate a series of topics for mixed land-use patterns and discuss the probability of meeting these topics for a particular spatial unit.

The probabilistic topic model is widely used in recent years. Among them, the latent Dirichlet allocation (LDA) is the most popular one, and it is an unsupervised generation probability model. Blei et al. [56] put forward the LDA to speculate the distribution of article topics in 2003. It generates the topic probability distribution for each document in the document set so that analyzing some documents and extracting their topic distribution can do topic clustering or text classification according to the subject distribution. If every article in the corpus has multiple topics, and each word in the article supports these topics when setting all words in all articles as the observed value, the model is able to discover the hidden topics. This idea can be adopted to discover the region’s function. Regarding a spatial unit as a document, the actual check-in data in the spatial unit as the word in the document, and the spatial unit’s function as the document topic, we can obtain each spatial unit’s functional distribution using the topic model. The generation process in the topic model does not assume any particular word order, and it only cares about the word quantity. This assumption also applies to the proposed mixed land-use model.

1) LDA MODEL

Document corpus $d_m$ represents the word set and the word frequency in the document. $d_m = \{(w_1, f_1), (w_2, f_2) \cdots, (w_n, f_n)\}, m \in [1, M], n \in [1, N_m]$, where $w_i$ represents the i-word in document $d_m$, and $f_i$ represents the word frequency.

Let us assume that document $d_m$ includes $k$ topics. These topics form a mixed polynomial distribution $z \sim Multi(\theta_m), \theta_m$ is a random variable subject to Dirichlet’s prior distribution, which has a prior parameter $\alpha$, and $\theta \sim Dirichlet(\alpha)$. As for the relationship between topic and word, the words in a topic are subject to a polynomial distribution $w \sim Multi(w|\phi_k). \phi_k$ represents the lexical distribution of topic $k$, which is also a random variable subject to Dirichlet’s prior distribution. The prior parameter is $\beta$, and $\beta \sim Dirichlet(\beta)$.

The graphical representation for latent Dirichlet allocation showed in Fig.3.

Overall, LDA topic generation process can be represented as a joint distribution, as shown in the following formula:

$$p \left( \tilde{z}_{in}, \tilde{\alpha}_m, \tilde{\theta}_m, \emptyset | \alpha, \beta \right) = \prod_{m=1}^{M} p(w_{m,n}|\phi_m) p(\tilde{\theta}_m | \tilde{\alpha}) p(\emptyset | \beta)$$ (3)

The parameter estimation is to determine the topic-word parameter matrix $\phi$ and the document-topic matrix $\theta$. This manuscript uses Gibbs sampling method to estimate the parameters.

B. ACTIVITY DENSITY-BASED TOPIC MODEL

To take the city as a whole subject and make it convenient for the manager to classify land use. We put forward the topic ratio to describe the mixed-use of land. The topic ratio uses
LDA process:
Step 1: Randomly generate topics.
Traversing all words of the all corpus’ documents, randomly assign a topic for each word, \( z_{m,n} = k \sim \text{Multi}(1/k) \).
Step 2: Variable initialization.
Suppose \( n_m^{(k)} \) is the quantity of the topic \( k \) appears in document \( d_m \). \( n_m \) is the total topics in document \( d_m \). \( n_k \) is the word \( t \) appears in topic \( k \). \( n_k \) are the total words in topic \( k \).
Step 3: Iteration.
Iterate over all the words in the Corpus, calculate the topic probability of each word, and sample the new topic according to the topic probability. If the word \( t \) in the current document \( d_m \) corresponds to topic \( k \), update the four variables to \((n_m^{(k) - 1}; n_m - 1; n_k^{(t)} - 1; n_k - 1)\). Sample the new topic according to the topic probability distribution, and update the variable to \((n_m^{(k) + 1}; n_m + 1; n_k^{(t)} + 1; n_k + 1)\).
Step 4: Repeat iteration.
Repeat step 3 until the hyperparameters \( \alpha \) and \( \beta \) converges.
Step 5: Output hyperparameters.
Output topic-word parameter matrix \( \varphi \) and document-topic matrix \( \theta \).

2) REGIONAL FUNCTION HYBRID PATTERN RECOGNITION BASED ON LDA

Within a region, a specific type of POI is dominant in quantity, but it did not always represent the region’s function. For example, the area where many universities congregate is likely to be an education center, but there are also many restaurants. It is because of the generally higher restaurant quantity in the city. Therefore, we need to consider the POIs in which residents’ activity happens. This question is similar to the topic finding, where we should consider the word frequency and the contribution. Similarly, we use the POI quantity and the check-in frequency to build the topic model.

The application of LDA in urban area function identification firstly appeared in the work of Microsoft Research Asia (MSRA) by Yuan et al. [57], [58], and then this method has been widely used [59]. POI and function label are always used as the auxiliary data for classification, but in this paper, they are analogous to the document’s words and directly treated by the LDA model. Besides, function labels in the third-party data set classified according to the activity type reduce the data complexity and consider the residents’ activities.

3) DETERMINE TOPICS SIZE

In the LDA model, it is essential to find a decent number of potential topics. The general solution is to use perplexity [60], which is a standard performance measure of probabilistic models in machine learning, and theory like this:

\[
\text{perplexity} (D_{\text{test}}) = \exp \left( \frac{\sum_{d=1}^{M} \log(p(w_d))}{\sum_{d=1}^{M} N_d} \right) 
\]

\[
p(w_d) = \sum_z p(z)p(w|z, \gamma) 
\]

in which \( M \) is the number of documents in the testing set, \( N_d \) is the number of words in each document. \( k \) is the topic size, \( w \) stands for a specific word, and \( z \) is a particular topic generated from \( \theta \), \( \gamma \) is document-topic distribution in the test set. Perplexity is the logarithm of words in the test set to the negative likelihood estimate of the generated document, so the more powerful the model, the smaller the value of perplexity.

Repeating the following steps and averaging the results at last using different \( K \); 1) dividing documents into a training set and testing set; 2) training topic model in the training set; 3) computing the corresponding perplexity in the testing set. The relationship between perplexity and topic size, i.e., \( K \), could be plotted. The smaller the perplexity values, the better the model’s effects are so that the optimal topic size could be determined.

C. ACTIVITY-SPACE INTERACTION UNDER THE INFLUENCE OF SPATIAL SIMILARITY

When mixed-use patterns of urban space are obtained, they can be used to discuss further whether individual residents’ activities are affected. In order to better discuss the interactive features of individual residents’ activities and spaces, firstly, the similarity of grid spaces is judged, and similar grids can provide similar functions for residents. Distance and functional demand for travel destinations are two competitive factors, so it is necessary to consider these two factors when analyzing spatial attraction from the individual perspective.

1) SPATIAL UNIT SIMILARITY MEASURE

After running the LDA topic model, the spatial unit represents a polynomial distribution of \( k \) dimensional mixed topics, as a probability vector \( \alpha_i = [p_{i1}, p_{i2}, \ldots, p_{im}, \ldots, p_{ik}] \), where the sum of all the probability values is 1. Therefore, we can apply a variety of probabilistic distance or similarity measures to quantify the distribution similarity of spatial units’ function such as the Hellinger Distance [61], cosine similarity [62], and Jensen-Shannon divergence (JSD) [63].
Specifically, the cosine distance is a measure of similarity between two eigenvectors, and Hellinger distance is a triangle inequality distance.

KL divergence is a measure of the asymmetry of the difference between two probability distributions $A_i$ and $A_j$. JS divergence measures the similarity of two probability distributions. It is a variant of KL divergence and solves the problem of KL Divergence Asymmetry. If the two distributions do not intersect at all, then JS divergence is a maximum of 1; if the two distributions are identical, then JS divergence is a minimum of 0.

The entropy of the probability distribution is:

$$H(p) = -\sum_{m=1}^{k} p(m) \log p(m)$$  \hspace{1cm} (6)

The KL divergence of $A_i$ and $A_j$, is:

$$D_{KL}(A_i \mid A_j) = \sum_{m=1}^{k} p(m) \log \frac{A_i(m)}{A_j(m)}$$  \hspace{1cm} (7)

Based on KL divergence, the JS divergence of $A_i$ and $A_j$, is:

$$D_{JS}(A_i \mid A_j) = D_{JS}(A_j \mid A_i)$$

$$= \frac{1}{2} D_{KL}(A_i \mid \frac{A_i + A_j}{2}) + \frac{1}{2} D_{KL}(A_j \mid \frac{A_i + A_j}{2})$$  \hspace{1cm} (8)

To measure the similarity between spatial units more intuitively, define the JS divergence similarity matrix:

$$M_{DJS} = 1 - D_{JS}(A_i \mid A_j)$$  \hspace{1cm} (9)

2) INTERACTIVE CHARACTERISTICS BETWEEN THE INDIVIDUAL ACTIVITIES AND SPATIAL UNITS

If regarding the activity space unit as the trip node, the individual can produce a trip between nodes. However, it is impossible to infer the travel distribution among nodes and the relationship among spatial units only by activities spatial distribution. Because the individual activities gather unevenly in a specific spatial scope, we can cluster the residents according to their check-in positions, and then the activities distribute around a core point. Take this core point as the focal point, and this paper discusses the spatial distribution features of residents’ activities, e.g., the interactive features between the core point and the spatial units.

In the travel activities, the farther away from the activity distribution center, the lower the activity probability is. In the interactions between an individual and a spatial unit, as the distance increases, the strength of the interaction between the core point and the spatial unit decreases, and this phenomenon is called distance decay [64]. On the macro-level, the activity intensity between residents and spatial units negatively correlates with the distance under the condition that other variables are relatively stable. On the micro-level, the probability of individuals moving to different spatial units negatively correlates with the distance.

$$G_{ij} = CP_{i}P_{j} f(d_{ij})$$ is usually used to quantify the effect of distance on spatial interaction. $C$ is a constant coefficient, $G_{ij}$ represents the interaction strength between the core point $i$ and the spatial unit $j$. $P_{i}$ and $P_{j}$ reflect the size of the core point and the spatial unit. The distance decay function $f(d_{ij})$ has an independent variable distance $d_{ij}$, which is used to describe the influence of distance factor quantitatively. The most commonly used distance decay functions are exponential [65], power [66], and gauss function [67].

Exponential distance decay function:

$$f (d) = Ce^{-\alpha d} + d_{0} (\alpha > 0)$$  \hspace{1cm} (10)

Power distance decay function:

$$f (d) = Cd^{-\beta} + d_{0} (\beta > 0)$$  \hspace{1cm} (11)

Gauss distance decay function:

$$f (d) = Cd^{-\alpha d^{2}} + d_{0} (\alpha > 0)$$  \hspace{1cm} (12)

For individual residents, the choice of a spatial unit is to choose the land function. Residents’ initiative activities are not only affected by distance, but also by other subjective factors. Residents care about whether multiple functional requirements can fill within a smaller footprint and whether the functionality provided by one location can be replaced by other frequently visited or familiar locations. It means that spatial units with similar functions in different locations will compete, while the spatial units with different functions in similar locations are complementary [62].

Based on spatial units’ topic distribution similarity and the distribution of individual activities, we can distinguish the interaction between the daily activities and spatial unit functions. It is possible to determine whether functional complementarity or competition among spatial units affects the distance attenuation of individual activity-spatial unit interactions, and the impact.

V. RESULT

A. LDA TOPIC MODEL RESULT

1) TOPICS QUANTITY DETERMINATION

We first divide the check-in functions and the quantity in each spatial unit to represent the document-word Matrix (the spatial unit-function occurrence Matrix). For the LDA input matrix, each row represents check-in activities quantity for each spatial unit, and each column represents the activity type.

To obtain the proper number of topics for the region-functional hybrid pattern, test the dataset’s perplexity. We randomly select 70% as the training set, take the remaining as the test set, calculate the perplexity, repeat this operation five times in each topic size, and calculate the corresponding perplexity of the topic size $k$. The K ranges from 0 to 25 is used to show the relationship between the degree of perplexity and the topic size in Fig.4. The Lower degree perplexity indicates that the better model performance. The results showed that the perplexity decreased gradually with the size of the topic in the five tests. When the topic size surpasses 15, the perplexity decreased slowly, so we chose 15 as topic size.
2) FUNCTION HYBRID PATTERN RECOGNITION

After running the LDA model, the distribution of 9 functions in 15 topics is obtained. A specific function has a significant proportion in some topics, and the other functions’ contribution is tiny. There are two or three functions that account for a high proportion, and function mixing is prominent in other topics.

The 15 topics in Fig.5 are divided into two categories according to the degree of functions mixing.

The first category is single functional-dominant: social service-dominant (topic #1), eating-dominant (topic #2), entertainment-dominant (topic #3), recreation-dominant (topic #7), transportation-dominant (topic #6), recreation-dominant (topic #7), shopping-dominant (topic #10) and home-dominant (topic #13). Each of the above-mentioned single function types has its corresponding dominant topic. An important reason for the lack of a work-dominant topic may be residents’ general reluctance to share their work status.

The second category is the multi-functional: entertainment, social service, transportation mixed (topic #4); entertainment, recreation, transportation mixed (topic #8); shopping, eating, transportation mixed (topic #9); transportation, recreation, entertainment, eating, shopping mixed (topic #12); transportation, social service, shopping, shopping, recreation, eating, home mixed (topic #14); social, eating mixed (topic #15).

The probability of transportation type is always high in the topic with multi-functions, so transportation is an essential reason for the mixed degree of land use functions. The impact of land use on transportation demand is primarily generated by the impact of activities. Besides, activities related to non-commuting like leisure, entertainment, shopping, and eating always appear together. The education and home functions are less intermingled with other functions, possibly because these two types of land use functions typically occupy a larger area than a single spatial unit.

3) TOPIC DISTRIBUTION FOR EACH SPATIAL UNIT

The LDA model obtains the topic distribution of each spatial unit. Each spatial unit is denoted as a polynomial distribution of a k-dimensional mixed topic. Ten spatial units are randomly selected, and their topic distributions are displayed in Table.3, according to which we can find that most spatial units have a topic, and its probability is much higher than other topics.

According to all spatial units’ topic distribution, we find that only 746 grid units’ second-largest topic probability is over 0.05, which means 81% of the topic with the highest probability can represent spatial units. Moreover, only 22 grids units’ third-largest topic probability is over 0.05, which means that 99.4% of the spatial units can be represented by the topics with the most considerable two probabilities.

For New York City, the urban space can be divided into two categories with different functional mixing degrees, and further divided into 15 meaningful mixing patterns. For each
spatial unit, one or two of these 15 patterns can be represented. It provides a suitable classification method for space management from a micro perspective.

FIGURE 6. Visualization for similarity matrix.

B. ACTIVITY-SPACE INTERACTION RESULT

1) SPATIAL UNIT SIMILARITY MEASURE

According to the proposed spatial unit’s similarity metric \( D_{JS} \), we can analyze the pairwise similarity between spatial units. Fig. 6 shows the \( D_{JS} \) similarity matrix for all the spatial units in New York City, in which each unit represents a 15-dimensional vector of the topics. The similarity value ranges from 0 to 1. The higher the value, the more similar the distribution of topics between the two units. The values on the diagonal are all 1. Visualize the similarity matrix, use the grid number as the abscissa and ordinate. It is easy to identify some blue bars with low similarity values in Fig. 6. It indicates that the distribution of the functions of these spatial units is significantly different from others. Further investigation reveals that the units are mainly located in developed areas, such as Wall Street, Broadway. The frequent simultaneous occurrence of different types of POIs in these spatial units within the region led to a low similarity to other locations. Therefore, given any spatial unit, we can find the most or least similar spatial unit according to this similarity matrix. The similarity results in this study distributed in the range of \((0.560337, 1)\). The degree of similarity is divided into four grades, namely, grade 1 \((0.9, 1)\), grade 2 \((0.8, 0.9)\), grade 3 \((0.7, 0.8)\) and grade 4 \((0.560337, 0.7)\).

2) SPATIAL UNIT FUNCTIONAL INTERACTION

To study the interaction between the resident and spatial unit, we use the grid number as a distance unit denoted by \( d_{gn} \), recognize the spatial unit \( A_{ij} \) as a node, and distribute each resident’s activity in the grid graph. The grid coordinates of the core points \( C_p \) are obtained by averaging the grid numbers \( G_p = \{ A_{ij}, \text{User ID} = p \} \) of each resident \( p \)’s activity.

\[
(C_p, Lat) = \text{ave}\{i, G_p\}
\]  

(13)

Calculate the distance between the check-in activities and the core point use

\[
d_{p-s} = \sqrt{((C_p, Lat) - i)^2 + ((C_p, Lon) - j)^2}
\]  

(15)

as well as statist the relative frequency of check-in activity, and then we obtain the distance-frequency graph. Use exponential, power, and gauss distance decay functions to fit the distance and check-in activity frequency curve, as shown in Fig. 7. The R-square values of the three functions are 0.91389, 0.79364 and 0.93926. The gauss distance decay function is the best, so we employ it in this paper.

The 180074 data in the check-in dataset correspond to 1,085 residents. However, our check-in data sets are uneven for individuals because each resident has a different propensity to initiate share their location in the social network. The more check-in data a resident has, the more distinctive the location will be. To better discuss the effect of spatial interaction distance decay at the individual resident level, we first ruled out residents with fewer than 200 check-ins. The remaining residents are divided into four equal parts according to the activity space’s scope. A resident number is

\[
(C_p, Lon) = \text{ave}\{j, G_p\}
\]  

(14)
randomly selected from four data sets. We selected four residents with numbers 354,293,315,84 for Quantitative analysis and recorded them as A, B, C, and D.

Firstly, by distributing their activities in grid coordinates, it can be found that the four residents’ spatial range of activities is quite different. Resident A activity distributes over 275 grids in the range ((2, 72), (11, 125)). Resident B activity distributes in 58 grids within ((9, 62), (16, 108)). Resident C activity distributes in 121 grids ((10, 70), (20, 124)), and Resident D activity distributes in 80 grids ((8, 55), (36, 117)). We can see that although residents’ activities can distribute in spatial units far-off, the access rate to accessible spatial units is not high and does not exceed 5 percent. Therefore, activities concentrate on a small number of space units, and space unit-individual interactive intensity is not entirely following the distance decay. Therefore, discuss the effects of individual selection in the following.

**FIGURE 8.** The curve of distance and check-in activity frequency for resident A.

For Resident A, as shown in Fig.8, the distance-interaction frequency curve is significantly different from the gauss curve. The primary activity radius is 50 $d_{gn}$ equaling 15 km, close to the city radius. There is no apparent decay between distance 0 to 50 $d_{gn}$, and there are four peaks at the distance of 10, 18, 27, 42, and the last peak is the highest. By observing the activity distribution, we find the spatial clustering phenomena at the four distances where the interaction intensity peaks occur. Select the spatial units with high interactive intensity in the aggregated areas and capture the similarity data in the similarity matrix. We can see that the fourth peak’s main reason is that the similarity of the function distribution between the fourth peak and the other peaks lies in the third and fourth grades, and their function distribution forms a complementary relationship. The demand for resident A activity for location function mostly overcomes the attenuation resistance of distance.

**FIGURE 9.** The curve of distance and check-in activity frequency for resident B.

For Resident B, as shown in Fig.9, the interaction intensity decreases with distance, and the trends are steeper than the gauss curve. The primary activity radius is 20 $d_{gn}$, which is equivalent to 6 km. There are also clusters of active regions at several distances with high interaction intensity.

**FIGURE 10.** The curve of distance and check-in activity frequency for resident C.

For Resident C, as shown in Fig.10, activities are basically in the range of 3 $d_{gn}$, converted to 0.9 km, a very short activity distance. By observing the activity distribution, we can find that the spatial units near the core point have a high functional mixing degree. The functional similarity degree between the core point and some spatial units far away is in the first grade. Residents can meet the needs of daily activities at a minimal distance, so the distance factor largely determines the intensity of activity-space interaction.

**FIGURE 11.** The curve of distance and check-in activity frequency for resident D.

For resident D, as shown in Fig.11, most activities are concentrated in 15 $d_{gn}$, converted to 4.5 km, and 40 grids distance also covers some activities. By observing the activity distribution, we can find that there are three activity clusters, and the activity intensity of the cluster at 5 $d_{gn}$ is much higher than that at 40 $d_{gn}$. The similarity of the three activity regions is at the first level, and the functional distribution forms a complementary relationship.
By combining the above four residents’ intensity-distance distribution, we can find that the distance decay curve’s effect is not the same for individuals. The range of activity varies significantly from person to person. Distance as a limitation to activity frequency has a few influences for the person with an extensive range. As for the complementary functions provided by the different urban spaces, all residents show the demand and access. People show strong access desire to the locations far away but can meet their activity function needs. The subjective demand for activities is an essential factor in overcoming the distance limit.

VI. CONCLUSION
The utilization of urban space function is an essential task in the field of urban planning and design. This paper provides a new perspective from residential activities’ footprint for exploring the land use patterns and the urban activity space patterns. This work uses the vast amount of check-in data from social media to explore how spatial functions and activities model each other from urban residents’ activities perspective. This paper quantitatively analyzed the degree of mixed-use of urban functions and studied the mixed mode of functions in small-scale spatial units. On the premise of applying information entropy to evaluate the mix of land-use functions, the LDA topic model is introduced to identify the mix patterns. Use the intensity of residential activities as the input of the model, we get the 15 mixed land use pattern topics and obtain the topic distribution of each spatial unit. Furthermore, on that basis, we discussed the effects of spatial function on individual activity based on the distance feature in activity-space interaction. Using the JS divergence of the spatial unit’s similarity measure makes it easy to obtain the similarity matrix. On the premise of fitting the gauss distance decay function, we discuss the variation of the interaction intensity between the individual and space. It proved that the spatial function distribution similarity leads to the complementation and competition in the residents’ subjective decision-making. The influence of subjective on activity-space interaction is to satisfy the functional requirements of the activity. This study provides a new idea for studying urban land use patterns and urban spatial function from a multi-disciplinary perspective. The way of bringing active location information from Social Network check-in data into urban space, the methods of GIS, information entropy, pattern recognition, text mining, spatial similarity are integrated and cross-utilized. We hope that this approach can be applied to other areas of future urban development, such as benefiting urban planning, public services, and location-based recommendations.

We only used a predefined grid system (300 m * 300 m) as urban land use spatial analysis unit. The size was adopted by following previous research works and the size of check-in data. This spatial scale is in line with the requirements of micro-scale urban functional area quantitative identification, and the number of check-in data in each grid is reasonable for the topic model training. However, we may need to examine our methods’ sensitivity using different sizes of grids and different sizes of check-in data sets. Although our study has successfully used check-in data in New York City to identify 15 functional urban distribution topics, LDA topic modeling is an unsupervised method with some limitations in finding proper urban functions. We plan to study additional considerations to use a supervised topic model to discover functional distribution performance in the future. Although the method used in this study applies to different cities, the thematic model results are different, representing the city’s spatial characteristics. Due to the lack of check-in data in other cities, it is impossible to compare different cities’ spatial characteristics through contrast experiments. The unitary social network data can only represent the location where the activity occurs, ignoring the time sequence and transportation, and covering up some activities characteristics. In the following research, multi-source data should be used to mine further the characteristics of individual residents’ whole-day trip-activity chain. The interaction between residents and spatial units is mainly based on qualitative descriptions and a lack of quantitative surveys based on residents’ travel choice tendencies.

REFERENCES
[1] M. Moberg, “The condition of postmodernity: An enquiry into the origins of cultural change,” DAVID HARVEY, “Amer. Ethnologist, vol. 21, no. 4, pp. 915–916, Nov. 1994.
[2] B. Alexander, D. Ettema, and M. Dijst, “Fragmentation of work activity as a multi-dimensional construct and its association with ICT, employment and sociodemographic characteristics,” J. Transp. Geography, vol. 18, no. 1, pp. 55–64, Jan. 2010.
[3] C. Hubers, T. I. M. Schwanden, and M. Dijst, “ICT and temporal fragmentation of activities: An analytical framework and initial empirical findings,” Tijdschrift Voor Economische en Sociale Geografie, vol. 99, no. 5, pp. 528–546, 2008.
[4] T. Schwanden and M.-P. Kwan, “The Internet, mobile phone and space-time constraints,” GeoForum, vol. 39, no. 3, pp. 1362–1377, May 2008.
[5] W. Tu, J. Cao, Y. Yue, S.-L. Shaw, M. Zhou, Z. Wang, X. Chang, Y. Xu, and Q. Li, “Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns,” Int. J. Geographical Inf. Sci., vol. 31, no. 12, pp. 2331–2358, Dec. 2017.
[6] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, “Urban computing: Concepts, methodologies, and applications,” ACM Trans. Intell. Syst. Technol., vol. 5, no. 3, pp. 1–55, 2014.
[7] M. F. Goodchild, “Citizens as sensors: The world of volunteered geography,” GeoJournal, vol. 69, no. 4, pp. 211–221, Nov. 2007.
[8] M. Dijst, “ICTs and accessibility: An action space perspective on the impact of new information and communication technologies,” in Advances in Spatial Science, 2004, pp. 27–46.
[9] R. G. Golledge and R. J. Stimson, Spatial Behavior: A Geographic Perspective. New York, NY, USA: Guilford Press, 1997.
[10] F. Horton and D. Reynolds, “An investigation of individual action spaces: A progress report,” in Proc. Assoc. Amer. Geographers, vol. 1, 1969, pp. 70–75.
[11] B. J. Lorch and M. J. Smith, “Pedestrian movement and the downtown enclosed shopping center,” J. Amer. Planning Assoc., vol. 59, no. 1, pp. 75–86, Mar. 1993.
[12] R. M. Downs and Y.-F. Tuan, “Space and place: The perspective of experience,” Geographical Rev., vol. 68, no. 3, p. 375, Jul. 1978.
[13] D. I. Walmsley, Human Geography: Behavioural Approaches. Hoboken, NJ, USA: Wiley, 1984.
[14] H. Forsberg, “Institutions, consumer habits and retail change in Sweden,” J. Retailing Consum. Services, vol. 5, no. 3, pp. 185–193, Jul. 1998.
[15] R. Palm, “The constitution of society: Outline of the theory of structuration,” Political Geography Quarte, vol. 20, nos. 2–3, pp. 200–203, 2003.
[16] B. Lemtora, “Time-geography-at the end of its beginning,” GeoJournal, vol. 48, no. 3, pp. 155–158, 1999.
L. Liu et al.: Study on Urban Spatial Function Mixture and Individual Activity Space From the Perspectives of Resident Activity

[66] P. Haggett, “Locational analysis in human geography,” Edward Arnold, London, U.K., Tech. Rep. 1, 1997.

[67] F. Wang, “Measurement, optimization, and impact of health care accessibility: A methodological review,” Ann. Assoc. Amer. Geographers, vol. 102, no. 5, pp. 1104–1112, Sep. 2012.

LINA LIU received the bachelor’s degree in transportation engineering from Chang’an University, Xi’an, China, in 2018, where she is currently pursuing the master’s degree in transportation planning and management. Her research interests include location-based social networks and pedestrian trip behavior.

HONG CHEN received the M.S. and Ph.D. degrees from Chang’an University, Xi’an, USA, in 1999 and 2006, respectively. She is currently a Professor with the Department of Transportation, Chang’an University. Her research interests include transportation management, environment, and spatial-temporal trajectory. She has been an Executive Member of the council—Traffic Engineering Branch, China Highway and Transportation Society, and the council—Highway Environmental Sustainability Chapter, China Highway and Transportation Society, since 2004.

TAO LIU is currently pursuing the bachelor’s degree with the Department of Computer Science and the Department of Law, Nankai University, Tianjin, China. His research interest includes information security.

* * *