

Article

# Recognition of Urban Functions and Mixed Use Based on Residents' Movement and Topic Generation Model: The Case of Wuhan, China

Haifu Cui <sup>1</sup>, Liang Wu <sup>1,2,\*</sup> , Sheng Hu <sup>1</sup>, Rujuan Lu <sup>1</sup> and Shanlin Wang <sup>1</sup>

<sup>1</sup> School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China; cuihaifu@cug.edu.cn (H.C.); husheng@cug.edu.cn (S.H.); lurujuan.cug@cug.edu.cn (R.L.); cugwsl@cug.edu.cn (S.W.)

<sup>2</sup> National Engineering Research Center of Geographic Information System, Wuhan 430074, China

\* Correspondence: wulian@cug.edu.cn

Received: 28 July 2020; Accepted: 4 September 2020; Published: 6 September 2020



**Abstract:** The rapid evolution of cities has brought new challenges to urban planning and management. The accurate evaluation of urban functional structure and mixed use is critical, especially at a fine scale such as by blocks. The composition and mixing of urban spatial functions calculated by remote sensing and statistics are non-quantitative and undetailed. The text topic models are often applied to process text data, but are rarely used to mine semantic information in quantitative data. Therefore, this paper attempts to carry out research on the recognition of urban functions and mixed use using a text topic generation model based on resident mobile data. First, the area within Wuhan Third Ring Road was divided into 2451 units at a grid size of 500 m × 500 m. The histogram-latent Dirichlet allocation (H-LDA) and information entropy were applied to assign different grid units to correct the functional topics and topic information entropy (TIE). Second, the functional categories of different analysis units were calculated using the point of interest (POI), frequency density (FD) and category proportion (CP) indicators, while the functional information entropy (FIE) based on the POI was calculated. Then, the urban functions and mixtures identified by the two kinds of data were compared and analyzed. Finally, referring to the geographic information in streetscape map and applying correlation analysis, the function and mixing results obtained from the experiment were verified. Studies have shown that the H-LDA model can identify bridges, which the POI data have shown is challenging to identify without attributes such as length. The function recognition accuracy of the H-LDA model is 89.3%, which is higher than K-means algorithm and Word2vec models. The correlation coefficient between FIE and TIE is 0.587, indicating that both are highly correlated. These explain the accuracy and rationality of identifying city functions and mixtures based on the H-LDA model. The H-LDA model can be applied to functional computing and fine-scale urban mixed function planning.

**Keywords:** LDA; information entropy; GPS trajectories; POI; function and mixed use

## 1. Introduction and Related Work

The diversification of urban functions is the basis of urban development, and different functions increase convenience for people by facilitating living, working, recreation and transportation [1]. To scientifically and rationally plan the urban spatial structure, many references have conducted various functional zoning studies.

Traditionally, urban functional areas are divided mainly based on satellite imagery, population size, statistical surveys, etc., and various types of index systems are used to reflect the characteristics of functional zoning [2–8]. These classification methods have the disadvantages of poor timeliness and

high cost. They also ignore the interaction between human behavior and the space in which people are located and rarely use human activity data to identify urban areas with common characteristics.

There is a close relationship between human mobility and urban spatial structure [9–13]. The behavioral data of micro objects can be used to better perceive the complex spatial structure of cities. Big data obtained through multiple sensors (e.g., Global Positioning System (GPS), Global System for Mobile Communication (GSM), smart card data (SCD), social media data (SMD)) provide new opportunities for describing and understanding the urban spatial structure. These mobile positioning data have low acquisition costs, wide coverage, and temporal characteristics. Therefore, the use of large-scale, high-quality individual spatiotemporal data to calculate people's daily behavior and reflect the different functional divisions of the city is a new method and approach for urban research.

References [14–16] have used GPS data to build various models through data mining tools, analyzed the relationship between passenger boarding and alighting characteristics and social functions in different regions of the city, and determined the urban spatial structure. References [17–19] proposed dividing the urban spatial structure using the information generated by the mobile phone base station network. Joh and Hwang [20] used 10 million SCD for Seoul to analyze the travel trajectory of bus card holders and the characteristics of urban land use. Wang et al. [21] used Sina Weibo data to spatially, temporally and behaviorally analyze the dynamic changes in urban activity and then divided urban areas into five functions, confirming that urban functional areas are characterized by hybridization and blurred borders. Some references have tried to use point of interest (POI) data to analyze urban spatial structure. POI data are very easy to capture, and include a comprehensive set of types, making them widely used. Wang et al. [22] and Yao et al. [23] used POI data to complete the identification of urban spatial functions through clustering and word vector models. Yuan et al. [24] and Chen et al. [25] used taxi trajectory data and POI data to identify urban functional areas with distinct characteristics at different scales. Han et al. [26] built a city functional area identification model based on bus credit card data and POI data to quickly and effectively identify each functional area in Beijing. Yue et al. [27] used POI data to measure the urban function mixing degree. Long and Liu [28] used the check-in data of Sina Weibo and POI data to calculate different types of land-use function and cited the mixed land use index proposed by Frank et al. [29] to realize the evaluation of mixed land-use functions.

The above researchers used a variety of big data to identify the functional structure of a city through different data analysis methods. The problems and deficiencies in previous research also deserve further consideration. (1) Previous research mainly focused on the overall identification and evaluation of urban functional areas and simply divided the city into single functional types, such as commercial functional areas, residential functional areas, and industrial functional areas. Although some references have studied the distribution of mixed urban functional areas (such as mixed commercial and residential functional areas), there are deficiencies in the refinement of research units and the quantitative analysis of functional mixing. (2) Research on the mixing of urban functions has been based on analyses of POI data, and the potential of other city data to recognize functional mixing has not been fully explored. There are also limitations in the research on functional recognition using POI data. For example, there is no relatively uniform standard for data classification, different categories and subordinate categories overlap, and the area of the entity object is ignored so that its boundary, scale, usage status and other information cannot be reflected, which can lead to inaccurate classification of mixed functional areas. (3) In the past, the identification of urban functional structure was mostly based on the number of incidents, such as active areas and hotspots, and the semantic information contained in the data has been rarely analyzed. Semantic information analysis is a new method for city function recognition after statistics and clustering algorithms, and it is playing an increasingly important role.

The latent Dirichlet allocation (LDA) model is widely used in semantic information mining. It is a document topic generation model, and also a three-layer Bayesian probability model that includes words, topics, and documents [30]. LDA is an unsupervised machine-learning technique for modeling discrete data sets. By modeling a corpus, it can be used to identify potential topic information in

a large-scale document collection or corpus. Wei et al. [31] applied the LDA model to information retrieval and proved that it can obtain more reasonable efficiency than using cluster-based models. Janowicz et al. [32] used the LDA model to identify a category for each document, and then used the expectation maximization (EM) algorithm to estimate the maximum likelihood. However, this model can only identify a single category of documents. Gao et al. [33] used the check-in data and the LDA model to calculate the relationship between different scenic spots in the city, excavated areas with similar thematic features, and achieved the purpose of perceiving the urban spatial structure from human activities. Tong et al. [34] applied the LDA and max-pooling to propose a point set multi-level aggregation feature extraction and fusion method, and proved its effectiveness for point cloud classification.

There is also some literature on improving the LDA model [35,36], but its application in the analysis of urban functional structure has the following problems: the research on the functional information mining is mainly based on POI semantics; the function type defined in the research scale is single; and the scale is not fine enough to quantify the functional composition, especially for the mixed function structure. The study of urban functional zoning and mixed use at a fine scale can facilitate rational and healthy urban planning. Appropriate mixed land use is conducive to improving land utilization, increasing government revenue, and reducing citizens' traffic flow between different functional areas. Considering the relationship between urban traffic behavior and urban functions and the rise of semantic topic analysis methods, this paper intends to conduct research on the identification of urban spatial functional structure and mixed use using the big data of urban floating cars and the latent semantic models. Therefore, this paper proposes to mine the implicit semantic information of urban functions from people's movement trajectory data, and analyze urban functions and mixed use from a quantitative and refined perspective.

This study attempts to use the LDA model which can quickly mine the latent semantics in massive texts, and use floating vehicle trajectory data to mine the hidden functions and mixed use in different areas of the city. First, we used one week of taxi GPS data for Wuhan to find the average O/D (origin/destination) data distribution at different times of day and calculated the workdays and weekends separately to characterize the actual use of various functions by people in the city. Second, using the histogram-latent Dirichlet allocation (H-LDA) model, the potential semantic information is extracted from the floating vehicle trajectory data, and the temporal and spatial patterns of human activities are analyzed to draw a probability histogram of functional topics in the research units. The topic information entropy (TIE) of the grid is calculated by the information entropy method. Third, the POI data are divided into 7 categories. The functions of different analysis units are calculated using the frequency density/category proportion (FD/CP) index, and the quantitative analysis of the mixed use of urban functions (functional information entropy, FIE) is completed. Finally, the urban function recognition results calculated by GPS data are compared with those based on POI data, and the two information entropy values are also compared. The K-means, Word2vec and correlation analysis methods are used to prove that the functional topic model (H-LDA) can effectively identify the different types of urban function and mixed use. Combining the analysis of static data (POI) and dynamic data (GPS) improves the real-time and accurate identification and evaluation of urban functions and mixed use and can provide a foundation and support for city block planning, policy formulation, and resource allocation.

## 2. Methodology

This study is based on the LDA topic model and information entropy method to calculate the urban functions and mixed use. Therefore, the following is a detailed introduction to the document topic probabilistic model (LDA), information entropy, and the calculation framework of this article, including a description of the improved functional topic model (H-LDA).

### 2.1. Latent Dirichlet Allocation (LDA)

This part focuses on the principle and calculation process of the LDA probability model, as well as the connection with urban function calculation.

The LDA [30] is a topic model proposed by Blei based on the Bayesian algorithm that uses the prior distribution to estimate the likelihood of the data and ultimately obtains the posterior distribution of the data. The LDA topic model is widely used in the field of natural language processing [35–38], and it can mine well the semantic information hidden in text.

LDA is an unsupervised machine learning model that does not require manually labeling of the training set, it only needs the number of documents and topics to be set. It uses the bag of words method and treats each document as a word frequency vector, thereby transforming text information into digital information that is easy to model. Each document represents a probability distribution composed of multiple topics, and each topic represents a probability distribution composed of many words. The calculation of the process of document topic probability based on LDA model is shown in Figure 1.

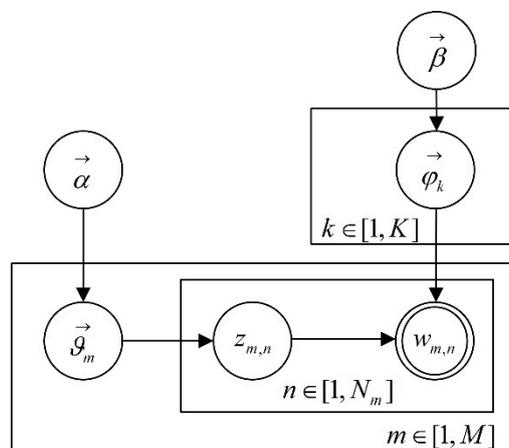


Figure 1. Latent Dirichlet allocation (LDA) model representation.

In Figure 1,  $K$  is the number of topics,  $M$  is the total number of documents, and  $N_m$  is the total number of words in the  $m$ -th document.  $\vec{\beta}$  is the Dirichlet prior parameter of the polynomial distribution of the word under each topic, and  $\vec{\alpha}$  is the Dirichlet prior parameter of the polynomial distribution of the topic under each document.  $z_{m,n}$  is the topic of the  $n$ -th word in the  $m$ -th document, and  $w_{m,n}$  is the  $n$ -th word in the  $m$ -th document. The two hidden variables  $\vec{\vartheta}_m$  and  $\vec{\varphi}_k$  represent the distribution of topics under the  $m$ -th document and the distribution of words under the  $k$ -th topic, respectively.

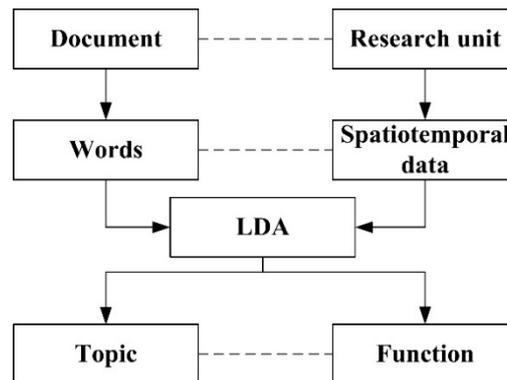
The process of generating the LDA topic model generally includes the following 4 steps:

1. Suppose there are  $M$  documents covering a total of  $K$  topics.
2. Each document has its own topic distribution, which is a polynomial distribution. The parameters ( $\vec{\alpha}$ ) of the polynomial distribution obey the Dirichlet distribution.
3. Each topic has its own word distribution, which is a polynomial distribution. The parameters ( $\vec{\beta}$ ) of the polynomial distribution obey the Dirichlet distribution.
4. For the  $n$ -th word in a document, first sample a topic from the topic distribution of the document, and then sample a word from the word distribution corresponding to the topic. Repeat this random generation process for all  $M$  documents.

The model has two parameters that need to be inferred: one is the “document-topic ( $\vec{\alpha}$ )” distribution, and the other is the “topic-word ( $\vec{\beta}$ )” distribution. By learning these two parameters, the

topic of the document and the proportion of topics covered by each document can be known. The inference methods mainly include the variational-EM algorithm and the Gibbs sampling method used in this paper.

Urban functional zoning research and text semantic research are similar. As shown in Figure 2, a research area is regarded as a document, the spatiotemporal data of human activity in the area are treated as words in the document, and the functional distribution of the area corresponds to the topic distribution of the document. Therefore, the concept of the LDA model can also be applied to regional function recognition to mine the functional structure of cities based on spatiotemporal data.



**Figure 2.** Relationship between document-topic and area-function.

## 2.2. Information Entropy

The amount of information measures the information delivered by the occurrence of a specific event. Considering all possible values of the random variable, entropy is the expectation of the amount of information for all possible events. Entropy represents a measure of the uncertainty in random variables in information theory. The Information Entropy formula [39] is:

$$H(X) = -\sum p(x_i) \log(p(x_i)) \quad (i = 1, 2, \dots, n) \quad (1)$$

where  $X$  is a random variable, the value of the random variable is  $(x_1, x_2, \dots, x_n)$ ,  $p(x_i)$  represents the probability of occurrence of event  $x_i$ , and  $\sum p(x_i) = 1$ . The unit of Information Entropy is a bit.

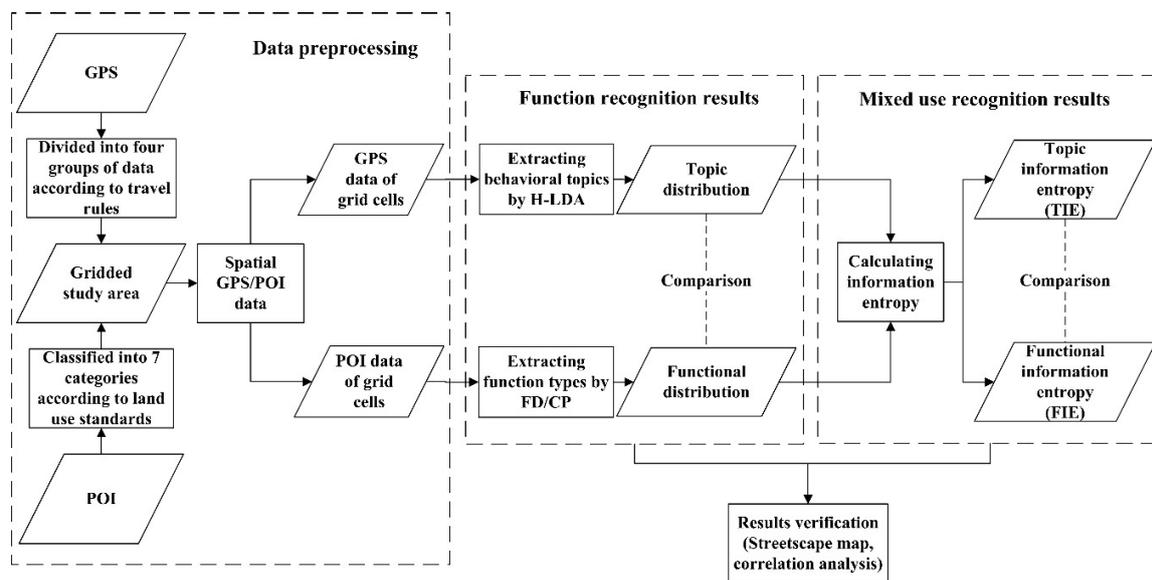
According to this formula,  $H(X)$  is the average information amount of the random variable  $X$ , that is, the expectation. The less probable it is that something happens, the greater the amount of information generated. Therefore, the information amount of a specific event should decrease as its probability of occurrence increases and cannot be negative.

Information Entropy is a physical concept used to measure the complexity and equilibrium of a system. However, the definition of Information Entropy makes it clear that its measurement is unrelated to the content of the event and does not change with the specific expression of the information, which is a statistical abstraction. Therefore, information entropy can also be applied to the analysis of urban functional structure.

Cities can also be regarded as a system, in which the structure and form of the city is a spatial mapping of the system, and their function is core contents of the study of urban structure and form. The research challenge lies in how to quantitatively describe and mathematically analyze the different functions produced by urban structure and form. The level of information entropy can reflect the mixed function distribution of different research units in the city. When the distribution of function types in a certain research unit in a city is more balanced, a larger the entropy value reflects a higher degree of functional mixing in the unit [40]. The information entropy model can provide an effective method for a quantitative analysis of the mix of urban functions.

### 2.3. Framework

This paper uses the human activity information contained in floating vehicle trajectory data and the urban function semantic information described by the POI data to achieve an in-depth analysis of urban spatial function and mixed use. Urban function mixing refers to the mixing of two or more urban functions within a certain space and time range. The information entropy of urban functional structure can reflect the distribution of mixed urban functions well. To this end, based on the LDA model and histogram method, the H-LDA model is proposed, which can generate the human activity topic probability information, the Information Entropy concept is applied to analyze the degree of urban function mixture, and the POI data are used to define the functional category of the research units. Finally, the urban functional area and the mixed use are identified. The technical flow of this paper is shown in Figure 3.



**Figure 3.** Process for identifying and analyzing urban function and mixed use.

1. The Wuhan Third Ring Road is taken as the research area and divided into the same grid units. To explore the urban functions and the mixed use in detail, the study area is divided into 2451 grid units sized  $500\text{ m} \times 500\text{ m}$ , which will provide results that can improve the utilization rate of urban space and serve as a reference for planning at the block scale [41]. Each grid is the basic unit of the spatial function analysis, and the discrete trajectory data and POI data are mapped to the corresponding grid unit according to their spatial coordinates.
2. The floating car trajectory data are collected approximately every 5 s. The basic information recorded includes the taxi license plate number, time, latitude and longitude, speed, direction and passenger status. This article uses a week of GPS data from 9 May 2015 (Saturday) to 15 May 2015 (Friday). Through the processing of trajectory data (in Section 3.1), the O/D data are extracted from the massive floating vehicle trajectory and divided into four groups of experimental data according to weekday and weekend. At the same time, the grid unit is used as an object to extract the behavior patterns in human travel at different times (hours), and the movement mode and topic probability distribution of each grid is calculated based on the H-LDA model.

The traditional LDA model determines the topic with single probability, ignoring other probabilities, and thus it cannot calculate the degree of topic mixing. Therefore, it is fused with the histogram method to describe the topic distribution of each grid, which makes it possible to calculate the mixing of functional topics. Through the probability distribution histogram of different topics for grids, it can also clearly describe the mixing of topics for each grid. Figure 4 shows the

calculation process of the H-LDA model. Each grid in the research unit is equivalent to a document in the LDA model; the frequency distribution of O/D point per hour in a day in a grid is equivalent to the words in the LDA model; the urban land function is equivalent to the document topic in the LDA model.

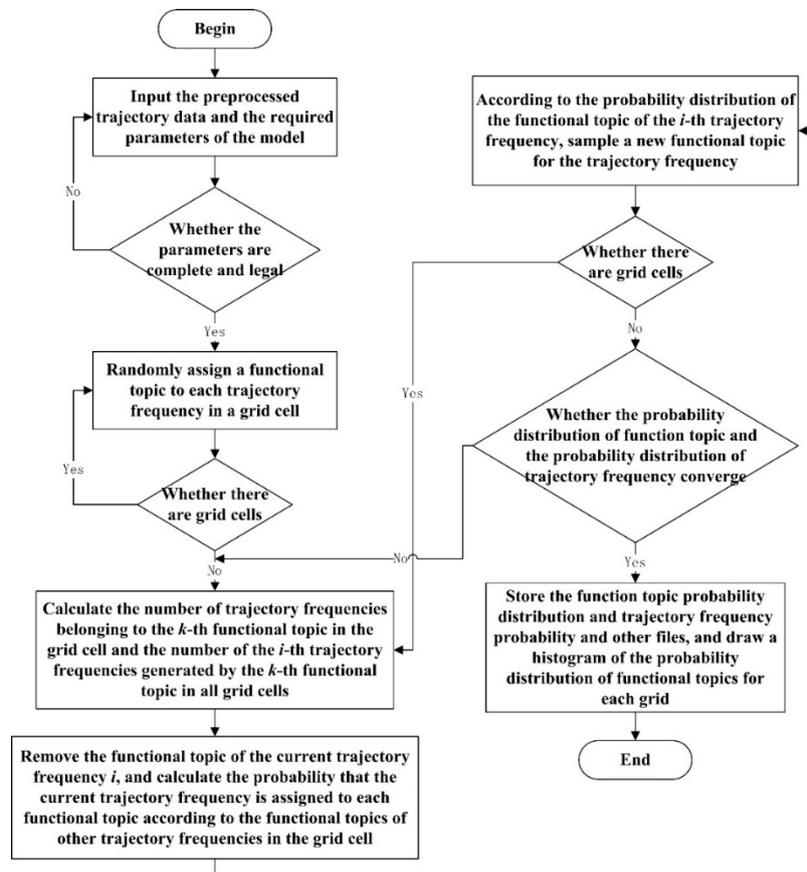


Figure 4. Histogram-latent Dirichlet allocation (H-LDA) model calculation process.

There are 2451 documents in this experiment, and each document has 24 words (each word is the same or different). The number of the urban function topics is set to 8, which refers to the number of POI categories (7 categories) divided in the Table 1 and the case of no data. By inputting these parameters and the iterative calculation of the distribution of function topics and trajectory frequencies, as described in Figure 4, the probability distribution of the urban functional topics in each grid can be obtained.

The probability assigned to each of the different topics of each grid is calculated by applying the H-LDA model, and the functional topic of each grid is determined according to the maximum probability value. The H-LDA model not only facilitates viewing the distribution of urban functional topics in different grids but also makes it possible to calculate the topic information entropy (TIE).

According to the principle of information entropy, the probability of the distribution of the 8 topics of each grid in the histogram is used to calculate the topic mixing degree of that grid. In other words, the probability of events in the information entropy formula is the probability of different topic distributions in a grid (the proportion of topic category  $i$  in a grid unit is  $p(x_i)$  in the information entropy formula), and the sum of the topic distribution probabilities for each grid is 1. Therefore, the distribution of the TIE for each grid can be obtained.

**Table 1.** Point of interest (POI) categories.

| ID | Category Name   | Quantity |
|----|---|----------|
| 1  | Office (government agency, company, transportation facility)          | 19,419   |
| 2  | Catering (restaurant, casual dining)                                  | 14,440   |
| 3  | Residence (real estate community, guesthouse, hotel)                  | 9491     |
| 4  | Medical and health (medical institution, social security institution) | 3088     |
| 5  | Culture and education (college, cultural media)                       | 7176     |
| 6  | Leisure and sports (sports venue, scenic spot)                        | 6582     |
| 7  | Entertainment and shopping (financial institution, shopping mall)     | 70,879   |

3. Different categories of POI data and the FD/CP index are applied to identify the function of the grid and calculate the functional information entropy (FIE) of the POI.

POI data have the advantages of a large sample size and detailed coverage information. Through the processing and analysis of POI data, the quantitative division of urban single and mixed functional areas can provide a reference for understanding the urban spatial structure. Therefore, this study uses POI data as supplementary information to help identify the functional attributes of cities.

After using the H-LDA model to calculate grid units with the same functional topic, they are identified according to the actual function of the area, that is, they are labeled according to the POI recognition results. The absolute quantity of POI may mask the actual attribute information in the area. To avoid the influence of differences in the number of POI in the grid unit on the recognition results, the POI frequency density (FD) and category proportion (CP) indicators are introduced to identify the functional properties. The FD represents the proportion of a certain category of POI in the total number of POIs of this category. The CP represents the proportion of the FD of a certain category of POI to the FD of all categories of POI in the grid unit. The function of each grid is determined according to the calculated CP value, that is, the category with the largest CP value in the grid unit is the function of the grid. When the grid unit does not contain POI data, the CP is null, and the unit belongs to the no-data area. According to the category definition of the POI, the functional distribution in the Wuhan Third Ring Road is calculated.

By applying the principle of the information entropy model, the FIE based on the POI data is constructed. This paper processes the grid for the urban space to calculate the functional mixed distribution. The proportion of different functions of each grid is calculated according to the FD/CP index, and the information entropy formula is applied to calculate the degree of functional mixed use for each grid (the CP index of function category  $i$  in a grid unit is  $p(x_i)$  in the information entropy formula). Therefore, the distribution of the FIE for each grid based on the POI data can be obtained.

4. Based on the distribution of urban function topics and the information entropy calculated by the O/D and POI data, the functions and the mixed use are calculated, compared and analyzed.
5. Streetscape map is used to verify the accuracy of the functional recognition results based on the H-LDA, K-means, and Word2vec methods, and correlation analysis is applied to verify the effectiveness of the H-LDA model for dividing the mixed use.

Considering the clustering algorithms and text information mining models commonly used in urban function calculations, the classic K-means and Word2vec methods is selected for comparative analysis of the results in the same study area. Simultaneously, the geographic information of the streetscape map is used to verify the accuracy of urban function recognition based on the three methods. In this paper, from the 2451 grid segments, statistical calculations are performed on the sample based on a 95% confidence level; 150 grids are randomly selected as the test area. By comparing the function of each selected grid with the streetscape map, the number of grids consistent with the actual category can be found.

The application of the H-LDA model to people's travel data to calculate functional mixing is a supplement to the traditional calculation of mixed use based on POI data. Its effectiveness can be

shown by the correlation analysis, specifically the analysis of the spatial correlation between the FIE based on POI data and the TIE based on taxi data. The FIE and TIE values are obtained through grid sampling, and their correlations are calculated accordingly. If the FIE and TIE pass the significance test, the two are related. This correlation can explain the reliability of calculating the urban function mixing degree based on the H-LDA model.

### 3. Experiment and Results

Based on the method and framework proposed in this article, the next part introduces the area and data used in the experiments. At the same time, it shows the urban function structure and mixing degree calculated based on O/D data and POI data, and compares the results.

#### 3.1. Data and Zones

GPS data record the activity information of people, cars and other mobile objects in the city. According to the 2016 Wuhan Statistical Yearbook, taxi passenger volume is nearly 1.1 million passengers per day. Therefore, using large-scale taxi datasets can reflect people's activities in Wuhan well and tap the functional structure of the city.

In this paper, the Wuhan Third Ring Road is used as the research area, as shown in Figure 5. Although the area of the seven main districts (Jiangnan, Qingshan, Jianghan, Qiaokou, Wuchang, Hanyang, Hongshan) in the Third Ring Road represents only 10% of Wuhan City, its population accounts for more than 60%. Therefore, population movement in this area can reflect its functions and mixed use. The study area is divided into 2451 grid units, which will serve as a reference for planning at the block scale. The basic data include GPS data from more than 10,000 taxis and open source big data include more than 130,000 POIs obtained through a Baidu Maps application programming interface (API); these are both large-scale multisource heterogeneous data sets.

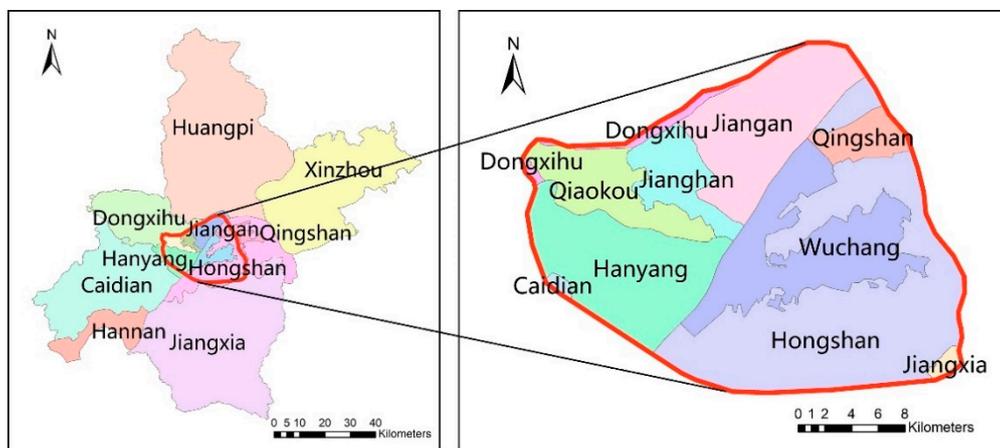


Figure 5. The case study area: Third Ring Road of Wuhan, China.

For subsequent experiments, the GPS data were preprocessed as follows: (1) delete non-conforming data according to the area and time period of the study; (2) extract the trajectory of passengers based on whether the vehicle is carrying people. The traffic behavior of vehicles with passengers reflects people's purposeful activities and commuting in the city, while the trajectory for no passengers is often to find the next passenger and cannot represent the travel characteristics of urban traffic; (3) generally, the purpose of travel is different between working days and rest days, and the origins and destinations have different types of functions, which reflect people's different behavior patterns. Therefore, the origin point (O), destination point (D), and corresponding time information of each traffic activity are extracted separately into working days and rest days based on the driving trajectory data of the person; (4) map each GPS point to the map, find the research unit where the coordinates are located, and

count the number of origin points and destination points of each research unit in hours. The purpose of dividing by hours is to try to exclude any effects from an excessive concentration of vehicles in a specific time period (such as morning and evening peaks). In this way, the extraction and spatialization of the origin point (O), destination point (D), and their respective times (T) based on the floating car data are realized.

The POI data mainly refer to certain geographical entities that are closely related to people's lives, such as schools, banks, and supermarkets. It is the basic information that defines the functions of each unit of the city [42]. Therefore, the POI data can to a certain extent represent the spatial distribution of urban functions, and they are also an important reference when fusing other types of city data. The POI data obtained in this study describe the location and attribute information of geographic entities, such as the entity's name, address, and coordinates. Combined with the "Urban Land Classification and Planning and Construction Land Standard" issued by the Ministry of Housing and Urban-Rural Development of the People's Republic of China [43] and the POI data classified by Baidu Maps [44], taking into account the correlation between urban functions and people's activities, the POI data are divided into 7 representative categories of office, catering, residence, medical and health, culture and education, leisure and sports, entertainment and shopping, as shown in Table 1.

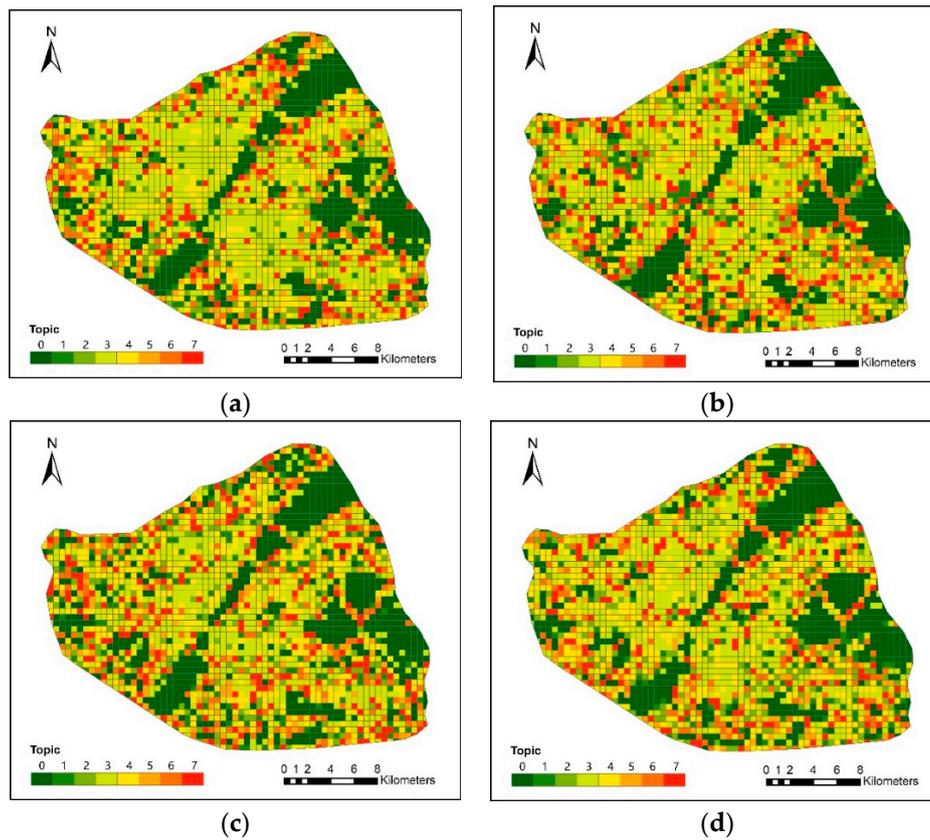
### 3.2. Origin/Destination (O/D)-Based Functions and Mixed Use

In the study area, the functional topic calculation is performed using the floating car data to obtain the distribution of the functional topics of the O/D points on weekdays and weekends. The O/D data mapped in each grid unit are averaged for weekdays and weekends, that is, the average O/D number per hour for the five weekdays from 11 to 15 November, and the average O/D number per hour for weekend on the 9 and 10 November, ultimately forming four groups of experimental data.

In this study, the LDA model is integrated with a histogram to calculate and display the distribution of urban functional topics, as this not only facilitates viewing the distribution of topics in different grids but also makes it possible to calculate the degree of functional mixing. The probability assigned to each of the different topics of each grid is calculated by applying the H-LDA model, and the functional topic of each grid is determined according to the maximum probability value. The 2451 grids of the study area are divided into 8 functional topics (7 POI categories and the case of no data). The experimental results are shown in Figure 6 and Table 2.

Through the combination of Figure 6 and Table 2, it can be concluded that in the four groups of data function topic distributions, the grids for topics 0 and 3 are the most common, while there are fewer grids for topics 1, 5, and 6. In general, the four groups of data have similar distribution rules for the functional topics. Therefore, a set of data (D point of weekday) is used as an example for comparison with the POI data to analyze the feasibility of applying the H-LDA model to identify urban functions and mixed use.

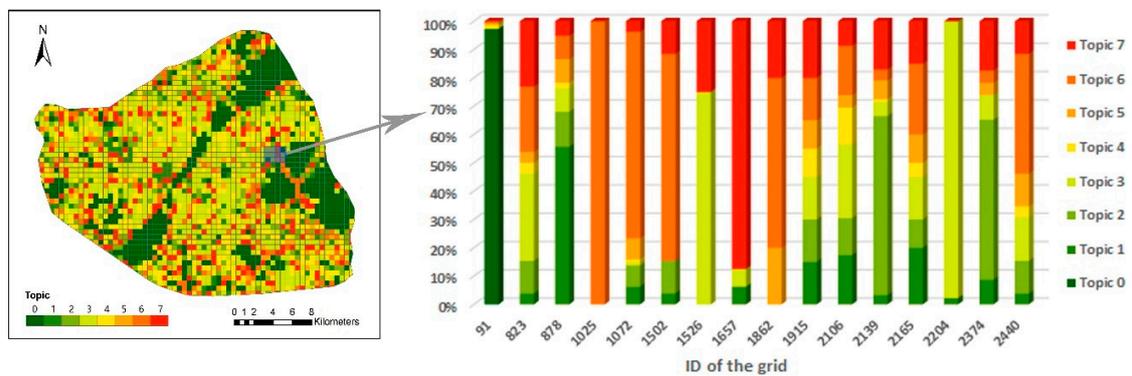
At the same time, the H-LDA model can describe the function probability distribution of each grid. Figure 7 shows the probability distribution of different topics for some grids in Figure 6b, from which we can clearly see the mixing of topics for each grid.



**Figure 6.** Topic distribution under four groups of data: (a) O point of weekday; (b) D point of weekday; (c) O point of weekend; and (d) D point of weekend.

**Table 2.** Topic distribution under four groups of data.

| Data/Topic | 0   | 1   | 2   | 3   | 4   | 5   | 6   | 7   |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|
| Weekday(O) | 531 | 132 | 161 | 770 | 421 | 84  | 138 | 214 |
| Weekday(D) | 505 | 129 | 160 | 763 | 416 | 112 | 129 | 237 |
| Weekend(O) | 613 | 155 | 150 | 543 | 434 | 136 | 171 | 249 |
| Weekend(D) | 588 | 128 | 174 | 609 | 456 | 122 | 169 | 205 |

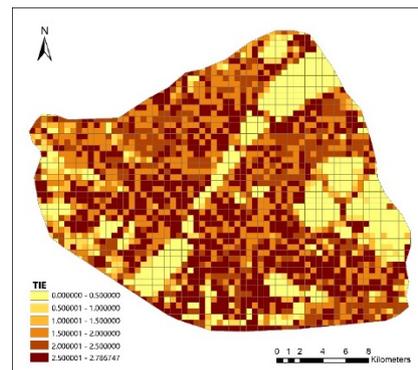


**Figure 7.** The proportion of different topics by the H-LDA model.

According to the function probability distribution calculated by the H-LDA model, the TIE for each grid based on the big data of taxis can be obtained. It can be found from the literature that the number of GPS data in different areas can characterize people’s different behavioral characteristics and rules [45]. The greater the number of GPS data in an area and the more evenly distributed they are

over different time periods, the more complex the function types and the higher the degree of mixed use; this can supplement the FIE calculation based on the POI data.

Figure 8 shows the distribution of the entropy of the functional topics in the Wuhan Third Ring Road calculated based on the H-LDA model. It is divided into 6 levels (0 to 0.5, greater than 0.5 to 1, greater than 1 to 1.5, greater than 1.5 to 2, greater than 2 to 2.5, greater than 2.5); the darker the color is, the higher the degree of mixed use. According to the spatial distribution in the figure, the value of topic mixed use is high. The number of grids in which the TIE value exceeds 2.5 (6th level) is the largest, and the number in the 1–1.5 (3rd level) level is small. Next, the topic mixed use is compared with the functional mixed use calculated by POI data.

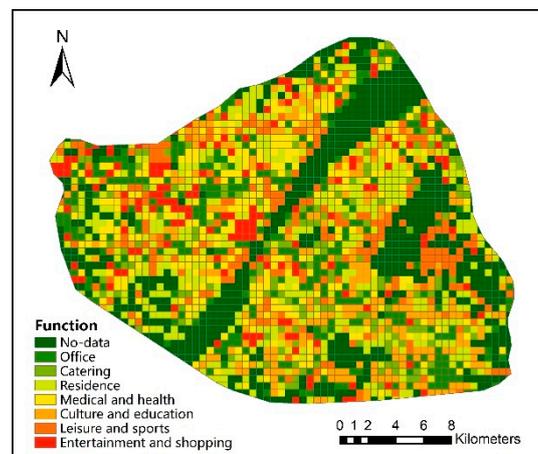


**Figure 8.** Topic information entropy (TIE) based on the H-LDA model.

### 3.3. Point of Interest (POI)-Based Functions and Mixed Use

The use of POI data can quantitatively divide the urban's single and mixed functional areas and provide a reference for understanding the urban spatial structure. Therefore, this study uses POI data as supplementary information to help identify the functional attributes of cities.

At the same time, in order to avoid the influence of differences in the number of POI in the grid unit on the recognition results, the POI FD and CP indicators [46] are used to identify the functional properties. The category with the largest CP value in the grid unit is the function of the grid. When the CP is null, the unit belongs to the no-data area. According to the category definition of the POI and the case of no data, the function of the Wuhan Third Ring Road is divided into 8 categories, as shown in Figure 9.



**Figure 9.** Functional distribution based on POI data

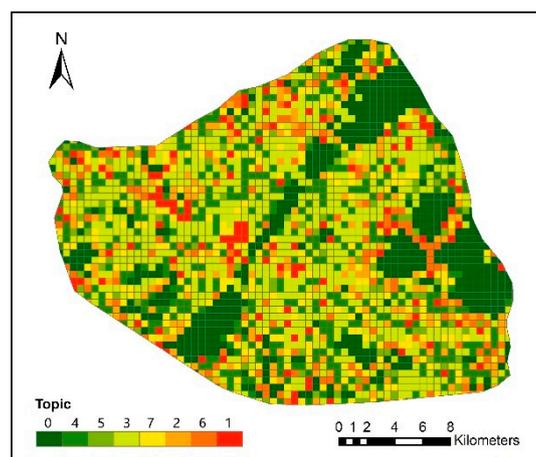
It can be seen in Figure 9 that the no-data and residence categories have the most grids, while the catering and entertainment and shopping categories are fewer. Although there are many POIs in the

catering and entertainment and shopping categories, the topic distribution is different. This shows that the application of the FD/CP index analysis function distribution is not affected by the number of POIs, and the analysis results are reliable. The grid topics calculated by the H-LDA model are compared with the grid functions calculated by the POI data, and the function category with the most grids in the same topic is defined as the actual function category of the grid.

Based on a comparison between the functional distribution calculated with the POI data and the functional topic distribution calculated by the H-LDA model, the proportion of different functions included in each topic is obtained. As seen in Table 3, 64.8% of the grids with topic 0 calculated by the H-LDA model belong to the no-data function category, 3.8% of the grids belong to the office function category, and so on. The data distribution of the no-data function is ignored when identifying other function categories, and the function category with the largest proportion is the corresponding function of the topic. The final function recognition results are shown in the bold font, that is, topic 0 calculated by the H-LDA model corresponds to the no-data function category calculated with the POI data, topic 1 corresponds to the entertainment and shopping function category, topic 2 corresponds to the culture and education function category, topic 3 corresponds to the residence function category, topic 4 corresponds to the office function category, topic 5 corresponds to the catering function category, topic 6 corresponds to the leisure and sports function category, and topic 7 corresponds to the medical and health function category. The topic distribution calculated based on the H-LDA model is matched to the functional distribution calculated based on the POI data, and the results are shown in Figure 10.

**Table 3.** Proportion of different functional categories in each topic.

| Topic | No-Data      | Office       | Catering     | Residence    | Medical and Health | Culture and Education | Leisure and Sports | Entertainment and Shopping |
|-------|--------------|--------------|--------------|--------------|--------------------|-----------------------|--------------------|----------------------------|
| 0     | <b>0.648</b> | 0.038        | 0.026        | 0.113        | 0.024              | 0.057                 | 0.067              | 0.028                      |
| 1     | 0.202        | 0.070        | 0.031        | 0.163        | 0.085              | 0.078                 | 0.132              | <b>0.240</b>               |
| 2     | 0.213        | 0.113        | 0.056        | 0.163        | 0.113              | <b>0.225</b>          | 0.056              | 0.063                      |
| 3     | 0.127        | 0.127        | 0.072        | <b>0.218</b> | 0.166              | 0.126                 | 0.098              | 0.066                      |
| 4     | 0.142        | <b>0.219</b> | 0.063        | 0.204        | 0.108              | 0.101                 | 0.101              | 0.063                      |
| 5     | 0.232        | 0.089        | <b>0.205</b> | 0.188        | 0.080              | 0.080                 | 0.063              | 0.063                      |
| 6     | 0.256        | 0.147        | 0.054        | 0.116        | 0.054              | 0.116                 | <b>0.202</b>       | 0.054                      |
| 7     | 0.211        | 0.139        | 0.063        | 0.177        | <b>0.177</b>       | 0.089                 | 0.084              | 0.059                      |



**Figure 10.** Topic distribution based on the H-LDA model.

By comparing Figures 9 and 10, with verification by streetscape map, it can be seen that the H-LDA model can identify no-data categories such as rivers and lakes. Because bridges over the waters also have human movement data, the H-LDA model can identify bridges, which is difficult to do with POI data, which lacks attributes such as length. Because different bridges have different levels of importance, there are differences in the number of times people use a bridge, resulting in different

types of topic function. The experimental results show that bridges are mainly divided into topic 3 and topic 4.

Comparing the function distribution calculated based on the H-LDA model and the POI data, it can be seen that the null value (no-data) with the topic of 0 includes rivers and lakes such as the Yangtze River and East Lake. However, the number of grids with no-data categories has decreased, indicating that some areas in the city that cannot be identified using POI data can be identified using human movement trajectories.

Topic 1 (entertainment and shopping) includes Wuhan International Plaza on Jiefang Avenue, Hanshang Group of Wangjiawan, Wanda plaza of Chuhe Han Street, Guanggu Pedestrian Street, etc., covering the main shopping and entertainment outlets in the Hankou, Hanyang and Wuchang districts.

Topic 2 (culture and education) includes major cultural and educational sites such as Wuhan University, Huazhong University of Science and Technology, China University of Geosciences, Hubei TV Station, and the Cultural Palace.

Topic 3 (residence) is distributed around various functional categories; therefore, many grids in other topics are also divided into residential functions. The residence category is thus the main functional category after the no-data category.

Topic 4 (office) includes major transportation and office sites such as Wangjiadun central business district (CBD), Optics Valley Software Park, Government Hall, station, airport, and the Yangtze River Bridge.

Topic 5 (catering) includes Hubu Lane, Wansongyuan Food Street, various small restaurants located near residential land and other places for eating.

Topic 6 (leisure and sports) includes Hankou River Beach, Zhongshan Park and other recreational areas. It should be noted that the waters of East Lake are classified in the no-data category, but its surroundings are within the leisure function category.

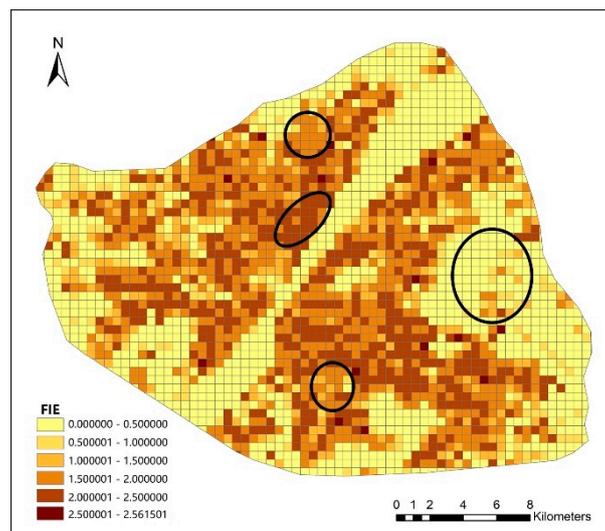
Topic 7 (medical and health) includes provincial and municipal hospitals, and community medical and health institutions. Because medical institutions are mostly distributed around residential areas, the proportion of residential functional categories in this topic is also large.

The comparison and identification of some POIs show that the H-LDA model can accurately identify the different functions of the city.

The level of information entropy can reflect the degree of mixed use of urban functions. The higher the entropy value is, the more the function categories exist in that grid, and the smaller the difference in the number of function categories. The proportion of different functions of each grid is calculated according to the FD/CP index, and the information entropy formula is applied to calculate the degree of functional mixed use for each grid. The FIE is divided into 6 levels, and the results are shown in Figure 11.

It can be seen from Figure 11 that most of the FIE in Wuhan Third Ring Road is higher than 1.5, and the degree of functional mixed use is high overall. Comparing the TIE calculated based on the H-LDA model (Figure 8) with the FIE calculated from the POI data, it can be found that the TIE is larger, reaching up to 2.79. There are many grids belonging to the 6th level of the TIE but few for the same level of the FIE. This shows that human travel data can provide fine-grained, high-sensitivity recognition of city functions, so that using the H-LDA model to calculate the functional mixed use will obtain a higher value. A comparison of Figures 8 and 11 shows the following.

The Hankou River Beach area has high FIE and TIE and a high degree of functional mixed use. The public facilities and functions in this area are relatively mature, and it concentrates living, working and entertainment. The consumption level is also high, so the frequency of taxi travel is high.



**Figure 11.** Functional information entropy (FIE) based on POI data.

The Houhu District, in which the FIE is higher than the TIE, has the main function of residence. The population is dense, and service facilities and entertainment facilities are relatively abundant. The residents here are mostly office workers, and their daily travel mainly depends on buses and subways.

The Nanhu District, in which the FIE is lower than the TIE, has the main function of residence and office. The density of people is high, and this area also has a higher per capita income for Wuhan. However, there is only one subway line here, and there is insufficient connectivity with other areas, so the frequency of taxi trips is relatively high.

The Donghu District has relatively low FIE and TIE, indicating that this area is a single functional area mainly for leisure. Facilities such as buses and subways are relatively mature and frequent, so both entropy values are low.

Table 4 compares the number of grids between the FIE and the TIE. The number for the TIE is higher than that for the FIE, which indicates that the H-LDA model calculates a higher functional mixed use for the Wuhan Third Ring Road: nearly half of the grids have a TIE value higher than 2; but only one-fifth have an FIE value higher than 2 calculated by POI. This shows that the recognition of city functions based on human movement patterns is more refined, while calculating urban functional mixed use based on the purpose of travel can supplement calculations of urban functional mixed use based on POI data.

**Table 4.** Number of grids with different levels of FIE and TIE.

| Level      | 1   | 2   | 3   | 4   | 5   | 6   |
|------------|-----|-----|-----|-----|-----|-----|
| <b>FIE</b> | 918 | 189 | 224 | 590 | 514 | 16  |
| <b>TIE</b> | 525 | 118 | 95  | 563 | 378 | 772 |

The above experiments demonstrate that the H-LDA model can be used to mine semantic information from the human activity pattern implicit in the floating car data and use it to identify the city function and mixed use; this can supplement and expand the quantitative identification of urban functions and mixed use based on POI data.

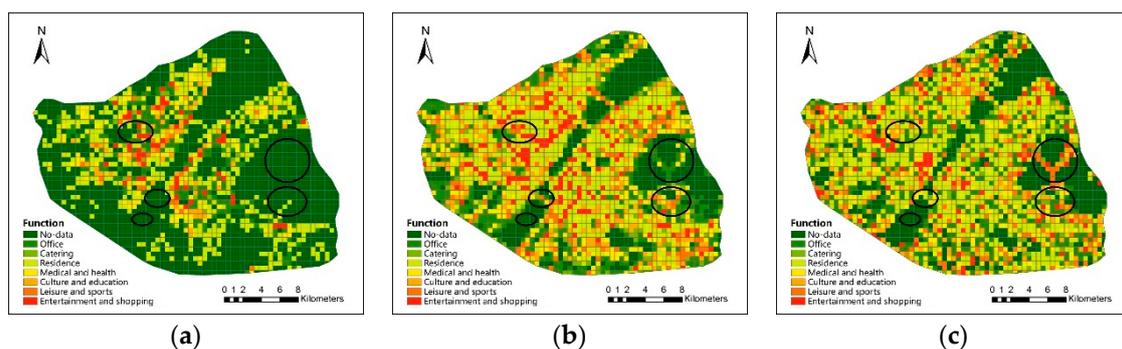
## 4. Discussion

### 4.1. Verification of Functional Recognition Results

The geographic information of Google Maps was used to verify the accuracy of urban function recognition based on the H-LDA model. The classic K-means [47] and Word2vec [48] methods were

selected for comparative analysis of the results in the same study area. K-means is a classic clustering method, which has the characteristics of less parameter input, unsupervised, and high computational efficiency. Therefore, it is often used to mine data sets with similar characteristics in spatiotemporal big data. Word2vec is a more important model in natural language processing. It converts words into computable and structured vectors. Through word embedding, the similarity between texts can be measured and the classification of text information can be realized. Clustering and semantic information mining algorithms are often used in urban calculations. Therefore, the K-means and Word2vec methods were selected to compare with the H-DLA model to analyze the effectiveness of the method proposed in this paper in urban function recognition.

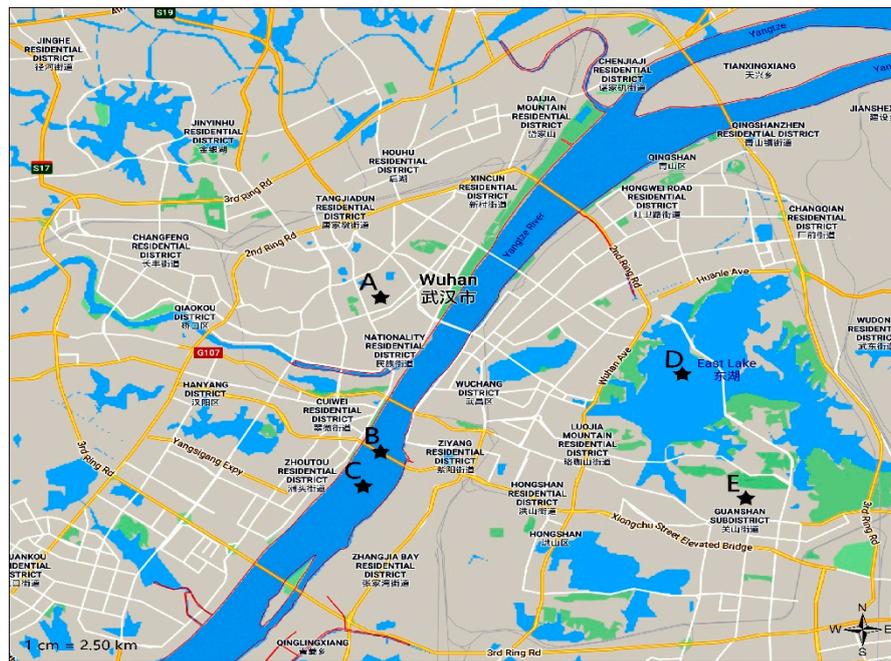
For comparability between the methods, we selected the same experimental data (D point of weekday) and the same function category parameters (8 functional categories) to calculate the city function. Finally, POI data is used to identify the function category of the corresponding classification result. Figure 12a was the result of K-means clustering on the total frequency of trajectories in a day in the grid, where the number of clusters was set to 8. Figure 12b used Word2vec to map different word vectors according to the hourly trajectory frequency in the grid, and then used K-means to calculate similar categories, and the number of clusters was also set to 8. Figure 12c was the function recognition result calculated by the H-LDA method proposed in this paper, which is consistent with Figure 10. According to the functions of the POI data division mentioned above, the functional categories were defined based on the experimental results of different algorithms, as shown in Figure 12.



**Figure 12.** Function recognition results of different methods. (a) K-means algorithm (b) Word2vec model (c) H-LDA model

In this paper, from the 2451 grid segments, statistical calculations were performed on the sample based on a 95% confidence level. Ultimately, 150 grids were randomly selected as the test area. By comparing the function of each selected grid with the streetscape map, the number of grids consistent with the actual category was found: 97 (in the K-means algorithm), 118 (in the Word2vec model), and 134 (in the H-LDA model), for an accuracy of 64.7%, 78.7%, and 89.3%, respectively. The comparison results for several typical regions are shown in Table 5.

The compared areas shown in Table 5 include the no-data areas of rivers and lakes (Place C), the office areas of roads and bridges (Place B), the leisure and sports areas of parks and scenic spots (Places A and D), and the culture and education areas of colleges and universities (Place E). As seen in Figure 12, the types of function calculated by the three methods are generally similar, but there are still differences in some details. As shown in Figure 12a, the function recognition of area A is insufficient, and the function information of areas B, D, and E is lost; although the recognized function is more detailed in Figure 12b, it is greatly affected by the adjacent area, and neighboring areas such as areas A and D are classified into the same category; the functional categories shown in Figure 12c are in line with reality, especially for areas A, B, and D.

**Table 5.** Comparison of the functions identified in this study and streetscape map.

A. Zhongshan Park



B. Parrot Island Yangtze River Bridge



C. Yangtze River



D. East Lake



E. Huazhong University of Science and Technology

The K-means algorithm uses Euclidean distance when calculating data similarity, so smaller differences in the functional information contained in the data will be ignored, and the classification results are affected by the order of the experimental data input, making the results unstable. The Word2vec model calculates the similarity by mapping data to high-dimensional vectors, so the calculation result is better than that of the K-means algorithm, but the computation is intensive and influenced by the data context, which can lead to incorrect functional categorization. Compared with the above two methods, the H-LDA model applied in this paper not only reduces the dimensionality of multidimensional data but also continuously optimizes the parameters based on a priori knowledge. The H-LDA model can divide the functional topic types in a probabilistic manner with higher overall accuracy.

For inaccurately divided areas, there may be mixed functions. Functional mixing is manifested not only in the coexistence of functions on different spatial scales but also in the functional transformation of the same space at different time periods. For example, roads and bridges are necessary features for people to access work during the day, and in the evening, they can become places for dining, leisure and sports. The actual degree of functional mixing in cities is much higher than that calculated from the POI data.

The H-LDA model can use human movement patterns to identify urban functions and comprehensively considers the complexity of residents' travel and the highly mixed context of urban land. The test results show that the H-LDA model can more accurately identify the functions of the city.

#### 4.2. Verification of Functional Mixing Results

Correlation analysis is the process of describing the closeness of the relationship between objective things and expressing it with appropriate statistical indicators. It is classical to use POI data when calculating functional mixing. Therefore, the correlation analysis method is used to calculate the correlation degree between the FIE based on POI data and the TIE based on taxi data, and then the reliability of calculating the urban function mixed use based on the H-LDA model is explained.

The FIE and TIE are spatially correlated, and two corresponding entropy values are obtained through grid sampling. Similar to Section 4.1, 150 grids were randomly selected from the 2451 grids segments. According to the values of the FIE and TIE in 150 grids, the correlation between them is calculated. The FIE and TIE passed the significance test (significance level at 0.01), and the Pearson correlation coefficient was 0.587, so the two were highly correlated.

Generally, the higher the FIE, the higher the TIE. This relationship not only explains the similarity between the two entropy values in the spatial distribution of functional mixing but also proves the reliability of calculating the function mixing degree based on the H-LDA model.

The H-LDA model focuses on the volume of data (numerical data and text data) on urban computations. For volunteered geographic information data such as social media data, it only needs to extract the check-in data from different locations to perform calculations without the user's personal information. Therefore, research on urban spatial structure can be carried out while protecting user privacy.

### 5. Conclusions

This paper used big data capturing people's daily movement to study the spatial structure of urban functions at a fine scale, which alleviated previous problems of difficult and slow data acquisition, large analytical scale, single functional classification of research units, and some subjectivity in the recognition of urban functions and mixed use. Instead of the traditional statistical and clustering methods, text classification was applied to study the city functions and mixed use, which expands the methods for city research.

First, GPS floating car data were divided into four groups according to people's different travel times and locations. The H-LDA model was used to divide the Wuhan Third Ring Road into eight topic categories. Experiments showed that four groups of data had similar topic distributions, topics 0 and 3 had the largest number of grids, and topics 1, 5, and 6 had a smaller number of grids. Therefore, one set of data (D point of weekday) was selected for mixed-use analysis. In addition, based on the information entropy method, the TIE of each grid unit was calculated and divided into 6 levels for display. These analyses showed that the Wuhan Third Ring Road had a high degree of functional topic mixing, and the largest grids had information entropy values exceeding 2.5 (level 6th).

Second, the POI data were combined with the FD/CP index and the information entropy method to calculate the functional category and mixing degree of the study area. By comparing the function distribution calculated by POI data with the topic distribution calculated by the H-LDA model, the actual function category corresponding to each topic was identified. It was found that the H-LDA model can identify bridges, which the POI data found challenging to identify without attributes such as length. Based on a comparison with typical POIs, it can be concluded that the H-LDA model can identify city functions accurately.

Comparing the FIE with the TIE showed that the TIE value was larger, up to 2.79. Nearly half of the grids had a TIE value higher than 2; while only one-fifth had an FIE value higher than 2. This shows that basing the recognition of urban function mixed use on human movement patterns produces more refined results. This method can distinguish the functions within cities based on the purpose of travel, which can serve as a supplement to calculating the mixed use of urban functions based on the POI data. The combination of the two types of data can be used to analyze the urban spatial structure in more real-time and multiple-dimensions.

Finally, the K-means and Word2vec methods were used for function recognition, and the results were compared with those from the H-LDA model. Through verification with the streetscape map, it was found that the H-LDA model delivered the highest accuracy. Using the correlation method to analyze the relationship between FIE and TIE, the Pearson correlation coefficient reached 0.587 (significance level at 0.01), which indicated that the two entropies were highly correlated. This also explains the accuracy and rationality of identifying city functions and mixtures based on the H-LDA model.

The above analysis showed that the experimental conclusions were in line with the actual context of Wuhan. Furthermore, it demonstrated that mining the semantic information of the human activity pattern implied by the floating car data (dynamic data) using the H-LDA model delivered both accurate and rational quantitative results in identifying city functions and mixed use. This approach can supplement and extend the identification of city functions and mixed use based on POI data (static data), furthermore, it will provide a new method to the urban mixed function planning on the block scale.

Selecting only taxi GPS data to represent people's daily travel behavior is one limitation of this article, so further research is planned that includes bus and subway data. In addition, the POI data used for functional verification do not consider the possible impact of different levels for the same category on the results. Furthermore, there are some shortcomings in the use of geospatial big data to identify function and mixed use. For example, the frequency of activities is closely related to population density, and the characteristics of frequency may be routine or noise caused by special events. Therefore, in future research, the analysis of urban functional structure will be based on multisource data such as demographic data.

**Author Contributions:** Conceptualization, H.C.; methodology, H.C., L.W.; validation, H.C., L.W., S.H.; formal analysis, H.C., L.W.; investigation, H.C.; data curation, H.C., S.H., R.L., S.W.; writing—original draft preparation, H.C.; writing—review and editing, H.C., L.W.; visualization, H.C., S.H.; supervision, L.W.; project administration, L.W.; funding acquisition, L.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was financially supported by the National Natural Science Foundation of China (41871311), the National Key Research and Development Program (Grant No: 2017YFB0503600), the Teaching Innovation Funds of Central Universities (2019G56).

**Acknowledgments:** The authors are thankful to the anonymous referees for their comments and suggestions that improved the quality of this paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Li, D. *Principles of Urban Planning*; China Architecture & Building Press: Beijing, China, 2001.
2. Li, J. Detecting Urban Land-Use and Land-Cover Changes in Mississauga Using Landsat TM Images. *J. Environ. Inform.* **2003**, *2*, 38–47. [[CrossRef](#)]
3. Wang, H. Rise of New Special Development Zones and Polarization of Socio-economic Space in Xian. *Acta Geogr. Sin.* **2006**, *61*, 1011–1024. [[CrossRef](#)]
4. Zhong, Y.; Zhu, Q.; Zhang, L. Scene Classification Based on the Multifeature Fusion Probabilistic Topic Model for High Spatial Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6207–6222. [[CrossRef](#)]
5. Hu, T.; Yang, J.; Li, X.; Gong, P. Mapping Urban Land Use by Using Landsat Images and Open Social Data. *Remote Sens.* **2016**, *8*, 151. [[CrossRef](#)]
6. Li, Y.; Yu, H.; Wang, H.; Deng, Y. The Response and Countermeasures of Urban Planning Research in the Era of Big Data. *Urban Dev. Stud.* **2017**, *24*, 1. [[CrossRef](#)]
7. Yao, Y.; Zhang, J.; Hong, Y.; Liang, H.; He, J. Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. *Trans. GIS* **2018**, *22*, 561–581. [[CrossRef](#)]
8. He, Z.; Deng, M.; Cai, J.; Xie, Z.; Guan, Q.; Yang, C. Mining spatiotemporal association patterns from complex geographic phenomena. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 1162–1187. [[CrossRef](#)]

9. Goodchild, M.F.; Klinkenberg, B.; Janelle, D.G. A Factorial Model of Aggregate Patio-Temporal Behavior: Application to the Diurnal Cycle. *Geogr. Anal.* **1993**, *25*, 277–294. [[CrossRef](#)]
10. Brockmann, D.; Hufnagel, L.; Geisel, T. The scaling laws of human travel. *Nature* **2006**, *439*, 462–465. [[CrossRef](#)]
11. Jiang, B.; Yin, J.; Zhao, S. Characterizing Human Mobility Patterns in a Large Street Network. *Phys. Rev. E* **2009**, *80*, 1136–1146. [[CrossRef](#)]
12. Batty, M. *Cities as Complex Systems: Scaling, Interactions, Networks, Dynamics and Urban Morphologies*; Springer: Berlin, Germany, 2010. [[CrossRef](#)]
13. Chen, T.; Hui, E.C.M.; Wu, J.; Lang, W.; Li, X. Identifying urban spatial structure and urban vibrancy in highly dense cities using georeferenced social media data. *Habitat Int.* **2019**, *89*, 102005. [[CrossRef](#)]
14. Qi, G.; Li, X.; Li, S.; Pan, G.; Wang, Z.; Zhang, D. Measuring Social Functions of City Regions from Large-Scale Taxi Behaviors. In Proceedings of the International Conference on Pervasive Computing and Communications Workshops, Seattle, WA, USA, 21–25 March 2011; pp. 384–388. [[CrossRef](#)]
15. Ladle, A.; Galpern, P.; Doyle-Baker, P. Measuring the use of green space with urban resource selection functions: An application using smartphone GPS locations. *Landsc. Urban Plan.* **2018**, *179*, 107–115. [[CrossRef](#)]
16. Du, Z.; Zhang, X.; Li, W.; Zhang, F.; Liu, R. A multi-modal transportation data-driven approach to identify urban functional zones: An exploration based on Hangzhou City, China. *Trans. GIS* **2020**, *24*, 123–141. [[CrossRef](#)]
17. Soto, V.; Friasmartínez, E. Automated land use identification using cell-phone records. In Proceedings of the 3rd ACM International Workshop on MobiArch, Bethesda, MD, USA, 28 June 2011; pp. 17–22. [[CrossRef](#)]
18. Toole, J.L.; Ulm, M.; Bauer, D.; González, M.C. Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD International Workshop on Urban Computing, Beijing, China, 23–27 August 2012; pp. 1–8.
19. Niu, X.; Ding, L.; Song, X. Understanding Urban Spatial Structure of Shanghai Central City Based on Mobile Phone Data. *Urban Plan. Forum* **2014**, *6*, 61–67. [[CrossRef](#)]
20. Joh, C.; Hwang, C. A Time-Geographic Analysis of Trip Trajectories and Land Use Characteristics in Seoul Metropolitan Area by Using Multidimensional Sequence Alignment and Spatial Analysis. In Proceedings of the AAG Annual Meeting, Washington, DC, USA, 14–17 April 2010.
21. Wang, B.; Zhen, F.; Zhang, H. The Dynamic Changes of Urban Space-time Activity and Activity Zoning Based on Check-in Data in Sina Web. *Sci. Geogr. Sin.* **2015**, *35*, 151–160. [[CrossRef](#)]
22. Wang, F.; Gao, X.; Xu, Z. Identification and classification of urban commercial districts at block scale. *Geogr. Res.* **2015**, *34*, 1125–1134. [[CrossRef](#)]
23. Yao, Y.; Li, X.; Liu, X.; Liu, P.; Liang, Z.; Zhang, J.; Mai, K. Sensing Spatial Distribution of Urban Land Use by Integrating Points-of-Interest and Google Word2Vec Model. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 825–848. [[CrossRef](#)]
24. Yuan, J.; Zheng, Y.; Xie, X. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 23–27 August 2012; pp. 181–189.
25. Chen, S.; Tao, H.; Li, X.; Zhuo, L. Discovering urban functional regions using latent semantic information: Spatiotemporal data mining of floating cars GPS data of Guangzhou. *Acta Geogr. Sin.* **2016**, *71*, 471–483. [[CrossRef](#)]
26. Han, H.; Yu, X.; Long, Y. Identifying Urban Functional Zones Using Bus Smart Card Data and Points of Interest in Beijing. *City Plan. Rev.* **2016**, *40*, 52–60.
27. Yue, Y.; Zhuang, Y.; Yeh, A.G.O.; Xie, J.; Ma, C.; Li, Q. Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 658–675. [[CrossRef](#)]
28. Long, Y.; Liu, X. How mixed is Beijing, China? A visual exploration of mixed land use. *Environ. Plan. A* **2013**, *45*, 2797–2798. [[CrossRef](#)]
29. Frank, L.D.; Andresen, M.A.; Schmid, T.L. Obesity Relationships with Community Design, Physical Activity, and Time Spent in Cars. *Am. J. Prev. Med.* **2004**, *27*, 87–96. [[CrossRef](#)] [[PubMed](#)]
30. Blei, D.; Ng, A.; Jordan, M. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.

31. Wei, X.; Croft, W.B. LDA-based document models for ad-hoc retrieval. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, USA, 6–11 August 2006; pp. 178–185. [CrossRef]
32. Janowicz, K.; Scheider, S.; Pehle, T.; Hart, G. Geospatial semantics and linked spatiotemporal data—Past, present, and future. *Semant. Web* **2012**, *3*, 321–332. [CrossRef]
33. Gao, S.; Janowicz, K.; Couclelis, H. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Trans. GIS* **2017**, *21*, 446–467. [CrossRef]
34. Tong, G.; Li, Y.; Zhang, W.; Chen, D.; Zhang, Z.; Yang, J.; Zhang, J. Point Set Multi-Level Aggregation Feature Extraction Based on Multi-Scale Max Pooling and LDA for Point Cloud Classification. *Remote Sens.* **2019**, *11*, 2846. [CrossRef]
35. Pang, X.; Wan, B.; Wang, P. Micro-blog’s Text Classification Based on MRT-LDA. *Comput. Sci.* **2017**, *44*, 236–241. [CrossRef]
36. Pavlinek, M.; Podgorelec, V. Text classification method based on self-training and LDA topic models. *Expert Syst. Appl.* **2017**, *80*, 83–93. [CrossRef]
37. Li, W.; Sun, L.; Zhang, D. Text classification based on labeled-LDA model. *Chin. J. Comput.* **2008**, *31*, 620–627. [CrossRef]
38. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L. Social sensing: A new approach to understanding our socioeconomic environments. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 512–530. [CrossRef]
39. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 623–656. [CrossRef]
40. Wang, X.; Li, X.; Feng, Z. Research on Urban Extension Based on Shannon Entropy. *China Popul. Resour. Environ.* **2010**, *20*, 89–92.
41. Schwanke; Phillips, D.; Spink, P.; Lockwood, F.; Versel, C.; Fader, D.; Holst, S.; Jerschow, L.; Myerson, O.; Deborah. *Mixed-Use Development Handbook*; Urban Land Institute: Washington, DC, USA, 2003.
42. Krosche, J.; Boll, S. The xPOI Concept. In Proceedings of the First International Workshop on Location and Context Awareness, Oberpfaffenhofen, Germany, 12–13 May 2005.
43. The Ministry of Housing and Urban-Rural Development of the People’s Republic of China. Urban Land Classification and Planning and Construction Land Standard. 2011. Available online: <http://www.mohurd.gov.cn/> (accessed on 19 June 2020).
44. Baidu. Baidu Map Open Platform. 2020. Available online: <http://lbsyun.baidu.com/index.php?title=lbscloud/poitags> (accessed on 19 June 2020).
45. Li, M.; Ma, Y.; Sun, X.; Wang, J.; Dang, A. Application of Spatial and Temporal Entropy Based on Multi-Source Data for Measuring the Mix Degree of Urban Functions. *City Plan. Rev.* **2018**, *42*, 97–103. [CrossRef]
46. Yu, L.; He, X.; Liu, J. Discovering urban functional regions based on semantic mining from spatiotemporal data. *J. Sichuan Univ.* **2019**, *56*, 64–70. [CrossRef]
47. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Symposium on Mathematics Statistics and Probability, Oakland, CA, USA, 21 June–18 July 1967; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297.
48. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. In Proceedings of the International Conference on Learning Representations, Scottsdale, AZ, USA, 2–4 May 2013; pp. 1–12.

