A Multiple Subspaces-Based Model: Interpreting Urban Functional Regions with Big Geospatial Data

Jiawei Zhu 1, Chao Tao 1, Xin Lin 2, Jian Peng 1*, Haozhe Huang 1, Li Chen 1 and Qiongjie Wang 3,*

1 School of Geosciences and Info-Physics, Central South University, South Lushan Road, Changsha 410083, China; jw_zhu@csu.edu.cn (J.Z.); kingtaochao@csu.edu.cn (C.T.); PengJ2017@csu.edu.cn (J.P.); hz_huang@csu.edu.cn (H.H.); vchenli@csu.edu.cn (L.C.)
2 Central and Southern China Municipal Engineering Design & Research Institute Co., Ltd., Wuhan 430010, China; linxin@citic.com
3 China Center for Information Industry Development, Beijing 100083, China
* Correspondence: wangqiongjie@ccidthinktank.com; Tel.: +86-010-68209225

Abstract: Analyzing the urban spatial structure of a city is a core topic within urban geographical information science that has the ability to assist urban planning, site selection, location recommendation, etc. Among previous studies, comprehending the functionality of places is a central topic and corresponds to understanding how people use places. With the help of big geospatial data which contain affluent information about human mobility and activity, we propose a novel multiple subspaces-based model to interpret the urban functional regions. This model is based on the assumption that the temporal activity patterns of places lie in a high-dimensional space and can be represented by a union of low-dimensional subspaces. These subspaces are obtained through finding sparse representations using the data science method known as sparse subspace clustering (SSC). The paper details how to use this method in the context of detecting functional regions. With these subspaces, we can detect the functionality of urban regions in a designated study area and further explore the characteristics of functional regions. We conducted experiments using real data in Shanghai. The experimental results and outperformance of our proposed model against the single subspace-based method prove the efficacy and feasibility of our model.

Keywords: urban functional regions; multiple subspaces; big geospatial data; urban spatial structure; sparse subspace clustering

1. Introduction

Analyzing the urban structure of a city is a primary research topic in the urban geographical information sciences. Therein, understanding different functions and their corresponding spatial distribution, i.e., functional regions, has drawn a lot of interest [1–3]. Functional regions describe urban space usage like when and why people visit a place and provide insights for a group of people. First, it helps decision-makers and researchers to calibrate the urban planning, assess the urban spatial structure, and study social and spatial disparities [4–6]. Second, it provides strangers like tourists with a quick understanding of the city and benefits social recommendations [7]. Moreover, it helps local people to expand their knowledge of urban operating regularity and regions with similar functions in the city [8,9]. However, the functionality of a region may change and develop through human mobility and activity [10], which can result in a functional difference from its original or intended use. Therefore, it is necessary to update the knowledge about the urban functional regions in time to capture the urban growth and thus facilitate better urban planning, site selection, location recommendation, etc.

Related works date back to the 20th century when Goddard [11] conducted a case study on functional regions in central London, using taxi flows collected in a traffic survey. However, these traditional methods were based on social surveys [12,13], the access to data...
was costly both in time and labor, and the objectivity of the data as well as the accuracy of study could easily be affected by subjective factors. Fortunately, with the development of information and telecommunication techniques, it is convenient to fetch big geospatial data such as GPS (Global Position System) data, GSM (Global System for Mobile) data, Smart Card Data, Social Media data, etc. This kind of data contains affluent, fine-grained, and large-scale information on human mobility and activity in both time and space dimensions.

With the influx of available information as such, the study of urban spatial structure enters a new paradigm that allows an increasing number of studies to combine data mining techniques with big geospatial data to reveal urban functional regions through human daily mobility and activity. Qi et al. [14] extracted temporal variation of get-on/off amount from taxi traces data and applied agglomerative clustering to characterize the social function of city regions. Pan et al. [15] used an iterative DBSCAN to recognize the social function of a region from the patterns of pick-up/set-down dynamics of taxi trajectories within the area. Yuan et al. [7] discovered the function of regions using both POIs data and taxi trajectory datasets by a topic-based inference model Latent Dirichlet Allocation (LDA). Liu et al. [16] classified Shanghai into six traffic “source-sink” areas that strongly link to six different functionalities by investigating the temporal patterns of taxi trajectory data in a seven-day timespan. Pei et al. [17] validated that land use information can be derived from mobile phone data in a case study of Singapore by implementing the Fuzzy C-means clustering on vectors of which consist of normalized hourly call volume and the total call volume of aggregated mobile phone data. Zhi et al. [18] used social media check-in data to infer urban internal functional regions by the coupling of Low Rank Approximation (LRA) and the K-means algorithm. Gao et al. [19] applied a popularity-based LDA model on POIs and incorporated user check-ins to derive latent thematic topics, functional regions are thus extracted by using clustering methods to group places that are semantically similar. Wang et al. [4] incorporated taxi GPS data with POIs to identify urban functional regions via the non-negative Matrix Factorization (NMF) and spatial semantic mining. Chen et al. [20] proposed a dynamic time warping (DTW) distance-based k-medoids method to delineate functional regions based on building-level social media data. Yu et al. [21] identified functional regions through the clustering method according to the travel characteristics derived from the taxi data and utilized POI data to label functionality.

In summary, in terms of data source, researchers have conducted experiments mainly with taxi data, social media check-in data, mobile phone record data, as well as combinations of data. However, with mobility data like taxi data solely, the characteristics of functional regions cannot be fully captured because of missing the information of what people visit a place for. Due to this reason, after detection, some studies introduced POI data or check-in data to provide travel demands information and label functional regions, the former is implicit and needed inference while the latter can offer explicit information. Therefore, in this study, we use taxi data and check-in data as data sources. Moreover, instead of separately using the mobility and activity information, we incorporate and transform these two into a form that can simultaneously reveal the dynamic of when and why people visit places to facilitate the identification of functional regions.

As for the techniques used to derive functional regions, usually, after preprocessing, the statistics of the data are stored in matrix form, where each vector represents characteristics of one place in terms of human temporal activity. Finding the functional regions, in reality, is equivalent to finding the underlying patterns in the vector space. Some studies directly applied clustering techniques. However, the intrinsic dimension of data is often much smaller. Patterns can be cluttered by redundant information. Therefore, dimensionality reduction is necessary [22], some studies introduced methods such as Singular Value Decomposition (SVD), Low Rank Approximation (LRA), and Non-negative Matrix Factorization (NMF) to extract valid information. These methods basically assume that data points distribute in a single lower-dimensional subspace, which is a subset of its original space. This idea explores the relationships between the data points and the bases of subspace. Then, the relationships can be represented by new lower-dimensional vectors.
In regard to detecting functional regions, the bases represent patterns of human temporal activity and each low-dimensional vector denotes how the patterns are combined into a real place. Therefore, they use shared bases of the low-dimensional space to model observations and further explore functional regions by identifying clustered places with similar patterns. Unfortunately, these methods have a disadvantage. As the functional regions can be different in nature and have intrinsic complexities, diverse bases with distinct dimensionality are needed to depict those regions more concisely and accurately. This challenges methods based on the single space assumption that share the same set of bases.

Therefore, a more general model should consider data lies in a union of low-dimensional subspaces rather than one single low-dimensional subspace [23]. In order to improve our model, we assert the assumption that all observations lie in a vector subspace that can be represented by a union of multiple subspaces, such that they can have diverse sets of bases, and each subspace represents one type of functional region. To discover the subspaces and subordinate points, we introduce a clustering method called Sparse Subspace Clustering (SSC) [24,25]. SSC acknowledges at the same time that the data belongs to a mixture of clusters and each having a different low dimensional structure. It imposes a sparse restriction on the self-expressiveness property of data [26]. That is, data point can only be represented by a linear combination of other points in the same subspace. By this means, data points that share similar patterns are designated to the same subspace. Consequently, the functional regions detected using this method characterize themselves and help us to exhibit the functional structure of the study area. Our contributions are to express how to use SSC to explore functional regions and to propose the metrics termed uniqueness degree and richness degree in order to evaluate the conditions of detected urban functional regions. The metrics are derived from the geometric properties of the subspaces’ representations.

The rest of the paper is organized as follows. Section 2 introduces our proposed model, gives an overview, and explains some preliminaries. Section 3 elaborates on the experiment settings. Section 4 presents the results of our experiment as well as a comparison with a method based on the single space assumption. Finally, the conclusions of the study are in Section 5.

2. Methods and Definitions
2.1. Frequency Matrix

Providing the coverage of the study area, taxi GPS points dataset, and check-in dataset, we can form a frequency matrix \( C \in \mathbb{R}^{M \times N} \). Specifically, first, we partition the study area into \( N \) geographical units and divide the timespan of one day into \( T \) time slots. Then, we remove abnormal GPS points caused by data recording or transfer error and identify taxi trips with passengers to further extract drop-off events. At the same time, we filter the social media check-in dataset to obtain valid check-in events. Next, we match the check-in events with drop-off events based on both coordinates and timestamps. When finished, the demand-tags of check-in events are used to label the purposes of matched drop-off events, and \( D \) is used to denote the number of categories of demands. All the labeled records are then mapped onto the geographical units and time slots to which they belong. After that, for each geographical unit, we can form a column vector \( g_n \in \mathbb{R}^{M \times 1} \), where \( M = T \times D \). Each entry in \( g_n \) denotes the frequency of people visiting this geographical unit in one time slot for a certain demand, such as going home, working, etc. That is, each column vector \( g_n \) presents the dynamic activity patterns of one geographical unit. In the end, \( C \) which represents the dynamic activity patterns of the study area is constructed by stacking \([g_1, g_2, \ldots, g_n, \ldots, g_N]\).

2.2. Sparse Subspace Clustering

Given the frequency matrix \( C \), Sparse subspace clustering (SSC) is then utilized to reveal underlying relationships between geographical units and thus to detect functional regions. The key idea of the SSC algorithm in this scenario is to find the sparsest combi-
nation of geographical units \( g_j \) in the matrix \( C \) to constitute a geographical unit \( g_i \) which belongs to a subspace \( S_k \), i.e., \( g_i = \sum_{j \neq i} Z_{ij} g_j \), where \( Z \) is a coefficient matrix. Finding the sparsest combination restricts to use the geographical units in the same subspace to constitute each other only, which means \( Z_{ij} = 0 \) if \( g_j \notin S_k \), and this characteristic is called self-expression [26]. As finding the sparsest solution is usually NP-hard, \( Z \) is optimized by solving a relaxed convex \( l_1 \) optimization problem to guarantee sparsity.

\[
\text{minimize } ||Z||_{l_1} \text{ s.t. } CZ = C, Z_{ii} = 0 \tag{1}
\]

Notably, the restrain \( Z_{ii} = 0 \) is to avoid trivial solution that each geographical unit is only represented by itself.

Once finished the optimization, we can obtain the matrix \( Z \) and construct a similarity matrix \( W = |Z| + |Z|^T \), where non-zero values denote the relationships between geographical units. Then, we apply spectral clustering techniques to \( W \) to acquire the subspaces segmentation. During this process, the number of subspaces is determined from the normalized Laplacian matrix of \( W \) with the eigengap indication [27]. The Laplacian matrix is defined as \( L = I - D^{-1/2}WD^{-1/2} \), where \( I \) is an identity matrix and \( D = \sum W_{ij} \) is a degree matrix. Shen and Cheng [27] pointed out that when the eigenvalues \( \lambda \) of the normalized Laplacian matrix are in ascending order, and the length of the \( i \)-th eigengap which defined as \( \lambda_{i+1} - \lambda_i \) is the largest, then \( i \) is viewed as the appropriate candidate for the number of subspaces.

Finally, \( C \) can be divided into \( K \) sub-matrices \( [S_1, S_2, \ldots, S_k, \ldots, S_K] \), which standing for \( K \) subspaces according to the clustering results, while each geographical unit \( g_n \) belongs to only one subspace.

\[
C = [g_1, g_2, \ldots, g_n, \ldots, g_N] = [S_1, S_2, \ldots, S_k, \ldots, S_K] \cdot \Gamma \tag{2}
\]

where \( \Gamma \) is a permutation operation that specifies the segmentation of geographical units. Based on above conditions and sparse restrictions, a subspace is formed from geographical units that share the most similar patterns, thus a subspace virtually represents a functional region.

### 2.3. Eigenplace and Significant Eigenplace

Eigenplaces, \( E_k = [e_1, e_2, \ldots, e_r, \ldots, e_R]_k \), represent bases of \( S_k \) which are derived from the Singular Value Decomposition (SVD), and the count \( R \) of bases can be different with respect to subspaces. Eigenplaces reveal which kinds of human temporal activity patterns exist in the functional region \( S_k \). With proper coefficient vectors, any place can be represented by a set of eigenplaces. Subsequently, a significant eigenplace is defined as one of the top \( r \) eigenplaces \( [e_1, e_2, \ldots, e_r]_k \), which determines the nature of a functional region \( S_k \). Notably, the functions of a region are dynamically changing with human mobility and activity. The nature of a functional region refers to the most prominent dynamic characteristic that people would visit the place for.

In summary, the relationships between the subspace \( S_k \) (i.e., functional regions), the bases \( E_k \), and the geographical units \( g_n \) are expressed as follows:

\[
S_k = E_k V_k^T = [e_1, e_2, \ldots, e_r, \ldots, e_R]_k \cdot \begin{bmatrix} V_1^T, V_2^T, \ldots, V_r^T, \ldots, V_R^T \end{bmatrix}_k^T \tag{3}
\]

\[
g_n \in S_k, g_n = E_k \cdot V_r = v_1 e_1 + v_2 e_2 + \ldots + v_R e_R \tag{4}
\]

where each row of \( V_k \) denotes the relationship between one unit and its related set of bases.

### 2.4. Definitions

Some notations and definitions used throughout are elaborated as follows.

1. Affinity between subspaces. The affinity between subspaces is computed from the principal angles between subspaces [25], which depicts the similarity between two
subspaces. It is defined as $\text{aff}(S_k, S_l) = \sqrt{\cos^2 \theta_{kl}^{(1)} + \ldots + \cos^2 \theta_{kl}^{(d_k \cap d_l)}}$, where $S_k$ and $S_l$ are two subspaces of dimensions $d_k$ and $d_l$. $\theta_{kl}^{(i)}$ is the $i$-th largest singular values of $U^{(k)}^T U^{(l)}$, while $U^{(k)}$ and $U^{(l)}$ are orthonbases of $S_k$ and $S_l$, respectively. Note that $\text{aff}(S_k, S_l)$ is high if two compared subspaces are similar.

2. Area Proportion (AP). AP denotes the ratio of the area that one functional region covers in the study area. It is used to complement the assessment of the urban spatial structure.

3. Uniqueness Degree of a Functional Region (UDR). The UDR describes the specialization level of functionality in one functional region and shows how one type of functional region is distinct from the others. If each region has a high UDR, there will be a clear functional division, with less vagueness and overlap between functional regions. The higher the UDR is, the more accurately the detection result will describe the urban structure. As significant eigenplaces characterize a functional region, the UDR is related to significant eigenplaces. It is designed to be inversely proportional to the affinity between subspaces restricted by significant eigenplaces and is defined as $\text{UDR}(S_i) = \sqrt{K - 1 / \sum_{i=1}^{K-1} \text{aff}^2(S_i, S_{-i})}$, where $K$ is the number of subspaces (i.e., regions) and $S_{-i}$ denotes all subspaces except $S_i$.

4. Richness Degree of a Functional Region (RDR). The RDR originates from the reconstruction error of using significant eigenplaces to approximate original functional regions. It is defined as $\text{RDR}(S_i) = \|C(S_i) - \tilde{C}(S_i)\|_F / \|C(S_i)\|_F$, where $C(S_i)$ is the matrix constituted of vectors belonging to subspace $S_i$, and $\tilde{C}(S_i)$ is the reconstructed matrix created by significant eigenplaces of $S_i$. If the reconstruction error is large, more eigenplaces are needed to depict the functional region besides just the dominant ones. Thus, the RDR($S_i$) implies the pluralistic development and function diversity of a region. When considering all regions, the RDA is the summation of each functional region weighted by corresponding AP. Therefore, the calculation is $\text{RDA} = \sum_{i=1}^{K} \text{AP}(S_i) \times \text{RDR}(S_i)$. The RDA examines whether the overall development of the urban spatial structure in the study is balanced.

2.5. Workflow

Figure 1 presents a detailed workflow of our model which is comprised of three major parts: In the first part, geographical units, drop-off events, and check-in events are obtained after preprocessing the coverage of the study area, taxi GPS points dataset, and check-in dataset, respectively. Then, the frequency matrix which indicates the dynamic activity pattern of geographical units is constructed through mapping and matching between drop-off events and check-in events. Later, in the detection part, the frequency matrix is fed into the SSC to detect multiple subspaces, i.e., functional regions. Then Singular Valued Decomposition (SVD) is performed to find the eigenplaces of each functional region, while significant eigenplaces used to label the functionality of detected regions are found by extracting dominant eigenplaces. Finally, based on the results of detection, we can discover the nature of each region and evaluate the condition of urban functional regions by computing the proposed Uniqueness Degree of a Functional Region (UDR) and Richness Degree of a Functional Region (RDR). Besides, we compare area proportion among all regions to complement the assessment of urbanization in the city.
3. Experiment Settings

3.1. Study Area

Shanghai is one of the four municipalities of China. It also plays a role as the commercial and financial hub and has a rapid urbanization growth [28]. The study area we chose covers a large portion of the urban area of Shanghai includes the center area and major transportation hubs, which is shown in Figure 2. Notably, the center area is the core of Shanghai’s development with a dense distribution of residential, business, and commercial facilities.

It is inevitable to partition the study area into geographical units to discover functional regions. However, the choice of research units still lacks standards [29]. A majority of studies divide their study area into regular grids [14,18,30–32], and the grid size varies from $250 \times 250$ to $1000 \times 1000$ m$^2$. In this study, we divided the study area into regular grids with a moderate size of $500 \times 500$ m$^2$. After removing the water zones, the study area is composed of 3166 geographical units.
3.2. Datasets
We used a GPS points dataset generated by 6600 taxicabs in Shanghai from 3 different periods in 2009: 1 June to 5 June, 8 June to 12 June, and 15 June to 19 June. The number of GPS points reaches approximately 714 million at the average sampling rate of 10 s per point within a position accuracy of about ±10 m. In addition, the social media check-in dataset we use has about 15 million social check-in records.

3.3. Data Processing
After removing abnormal GPS points, the trajectories with passengers were extracted. In total, 7.85 million trips were extracted from the taxi GPS points. Based on the trip records, drop-off events are extracted. All the drop-off events are then mapped into geographical units according to their drop-off location. To acquire the dynamic activity pattern of one geographical unit, we select 24 h as the time span where every hour is a single time slot. We derive the categories of demands-tags for each check-in event using the same methods as Zhi et al. [18]. The demands are then divided into six types: Home (H), Transportation (Tr), Work (W), Dining (D), Entertainment (E), and Others (O; refers to going to parks, museums, libraries, and other places). Thus, each grid has a 144-dimensional vector, where every 24 h entry describes the dynamic visiting frequency of a certain demand. In total, the final frequency matrix \( C \) was 144 by 3166.

4. Results
4.1. Results of Detection
As shown in Figure 1, the detection procedure has two steps: using SSC to obtain the functional regions and using SVD to extract eigenplaces. The main results in this section are the spatial distribution of functional regions, the set of significant eigenplaces, and the dynamic characteristics of functional regions.
4.1.1. Functional Region

As computed using SSC, we obtain the relationship between geographical units and generate a similarity matrix. Next, we apply the spectral clustering technique to the similarity matrix in order to acquire clusters, that is, the functional regions in this scenario. By calculating the eigengap mentioned in Section 2.2, the number of functional regions is determined as five. The similarity matrix is visualized in Figure 3, it is permuted to allow geographical units that share similar representations to be indexed sequentially. As the similarity between geographical units indicates that they belong to the same functional region, the number of subspaces can also be inferred as five based on the number of darker blocks.

![Figure 3. Visualization of the similarity matrix. Non-zero entries in the matrix are painted black, while zero entries are not colored. The numbers aside are the indexes of permuted geographical units. Dashed lines are added to highlight the five darker blocks.](image)

Then, we allocated the clustering results to the regular grids, the spatial distribution of the derived functional regions is displayed in Figure 4. As shown, the central zone of the study area is mainly covered by Region 5, while the outer zones contain different functional regions.
4.1.2. Significant Eigenplaces and Characteristics of Each Region

Through SVD, we obtain eigenplaces of each region from the corresponding sub-matrix. As significant eigenplaces reflect the dominant functions of the region, we find the significant ones by evaluating the distribution of eigenvalues. The first five eigenvalues account for more than 90% of the total data in Regions 1, 2, 3, and 4, but less than 90% in Region 5, as shown in Figure 5. Therefore, we choose the top 5 eigenplaces as the significant eigenplaces in each region except for Region 5. As for Region 5, we consider the influence of the top 10 eigenplaces.

Figure 5. Distribution of eigenvalues in corresponding covariance matrix (the vertical axis indicates normalized eigenvalue and the horizontal axis denotes indexes of descending eigenvalues).
Based on the significant eigenplaces, we obtain the patterns of activity and infer the functionality of regions from their dominant activities in corresponding significant eigenplaces. As shown in Figure 6, Home activity is the most active among all significant eigenplaces in Region 1. Dining activity takes second place, followed by Entertainment activity. Therefore, we associate Region 1 with the Residential Area, which is developed with facilities like restaurants and entertainment. Evaluated in the same way, Region 2 is the Transportation Hub due to the high Transportation activity; Region 3 is the Work Space as the major activity is Work; and Region 4 is the Other zones for parks, museums, gas stations, and so on. As for Region 5, we weighed the influence of the top 10 eigenplaces and determined it to be the Business District because Dining and Entertainment hold a large proportion of high activity. Our detected functionality is in line with the functionality of POIs in Figure 4, which verifies our methods.

Combined with the analysis result in Figure 4, we can see the spatial distribution of each functional region. The Business District occupies the center of the study area and sprawls outward, while the outer zones contain other types of functional regions. Furthermore, the separation of Work Space and Residential Area is also implied by Figure 4.

4.2. Udr and Rdr Assessments

We use the proposed UDR and RDR, as well as AP, to explore the condition of the functional regions and urbanization in our study area. The results are shown in Table 1. AP is derived by computing the ratio of the grid number of each detected region to the total number of grids. As shown in Table 1, the Residential Area is slightly larger than the Work Space. The Business District has the highest proportion of land occupancy, which implies the prosperity in the hospitality industry of this section of Shanghai, and the great demand of urban residents in Dining and Entertainment functions. Notably, the
Transportation Hub AP takes up 0.14 of the total area, which means there is a high traveling demand in the area.

In regard to UDR, the Work Space reveals a high uniqueness, which implies a high specialization of workplaces. The lowest UDR belongs to the Residential Area. It can be explained by the Dining and Entertainment demands of residents leaving the residential area to be mixed with other facilities like restaurants and entertainment. This makes it similar to the Business District and therefore less specialized. Still, the overall high UDR shows that the functional regions of the study area are significantly different.

The highest RDR belongs to the Other Zones, meaning it contains the most complicated activity pattern. This is a logical conclusion considering it supplies diverse demands. The low RDR of the Residential Area and Business District can be explained by their relatively simple functionalities which mainly concentrate on Home, Dining, and Entertainment activity. As for the entire area, the value of RDA is 0.538, which is a middle level compared to functional regions, indicating a comprehensive and balanced development of functional regions.

Table 1. Assessment results of functional regions.

| Region          | Residential Area | Transportation Hub | Work Space   | Other Zones   | Business District |
|-----------------|------------------|--------------------|--------------|---------------|-------------------|
| AP              | 0.195            | 0.140              | 0.192        | 0.086         | 0.387             |
| UDR             | 0.898            | 1.080              | 1.151        | 1.047         | 0.946             |
| RDR             | 0.510            | 0.571              | 0.554        | 0.607         | 0.518             |

In summary, urban spatial structure, specifically the function structure, of the study area is predominant in Dining and Entertainment. Home and Work are important components, Transportation is a necessity for the sound growth of this area, and Other Zones adorns the structure.

4.3. Comparison with the Single Subspace-Based Method

In this section, we conduct a comparative experiment to exhibit the differences between our model and single subspace-based methods. The method we used for comparison detects functional regions via a low-rank approximation (LRA) model based on SVD (Zhi et al., 2016). The computations of the LRA method are introduced as

$$\minimize \|C - \hat{C}\|_F, \text{ s.t. rank}(\hat{C}) \leq k \quad (5)$$

$$\hat{C} = \hat{U}\hat{S}\hat{V}^T = \Omega\hat{S}^{\frac{1}{2}}\left(\hat{V}\hat{S}^{\frac{1}{2}}\right)^T \quad (6)$$

where $\hat{C}$ is the best rank-k approximation of a $M \times N$ matrix $C$, and SVD is applied to decompose $\hat{C}$, the resulted $\hat{U}$ is a $M \times M$ unitary matrix, $\hat{V}$ is a $N \times N$ unitary matrix, $\hat{S}$ is a diagonal matrix constraining the singular values $\sigma_i$ of $\hat{C}$. Rows in $\hat{U}\hat{S}^{\frac{1}{2}}$ represent the characteristics distribution of temporal activity pattern in one eigenplace, while rows in $\hat{V}\hat{S}^{\frac{1}{2}}$ denote the characteristics distribution of eigenplaces. Then, the similarity matrix can be constructed as $W = \left|\hat{V}\hat{S}^{\frac{1}{2}}\right| \left|\hat{V}\hat{S}^{\frac{1}{2}}\right|^T$, the number of detected regions is determined as six using the aforementioned eigengap method. The distribution and significant eigenplaces of the functional regions are shown in Figures 7 and 8, respectively.
Figure 7. Geographical distribution of functional regions detected by the low-rank approximation (LRA) Method.

Figure 8. Significant eigenplaces of each region (LRA Method).
Comparing Figure 7 with Figure 4, the distribution of functional regions detected by the LRA method is similar to that discovered by our model, with the exception that the LRA method separates our Region 5 (Business District) into two regions: Region 5 and Region 6. The significant eigenplaces in Figure 8 also show that Region 5 and Region 6 detected by the LRA method have overlapped functionality as Dining and Entertainment are both active in each region. The values of the UDR in Table 2 are generally lower than those in Table 1, which implies that the functions of each region are more overlapped and less distinguished.

| Region | 1    | 2    | 3    | 4    | 5    | 6    |
|--------|------|------|------|------|------|------|
| UDR    | 0.838| 0.918| 0.888| 0.962| 0.836| 0.878|

To further verify whether Region 5 and Region 6 should be integral regions, we use t-SNE to visualize clustering results [33]. t-SNE is a technique commonly used to project high-dimensional data to two-dimensional for visualization. It utilizes joint probabilities to model similarities between data points. To ensure the similarities between data points remain unchanged after dimensioning reduction, t-SNE uses gradient descent to minimize the Kullback–Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. With the frequency matrix as input, t-SNE generates two-dimensional scatter plots, where points represent geographical units. Then, points are rendered with different colors according to cluster labels generated by our model and the LRA method, as shown in Figure 9a,b respectively. Figure 9a shows that the five clusters are well separated, only the blue cluster (Region 3) and the orange cluster (Region 1) are slightly overlapped with the yellow cluster (Region 5). Meanwhile, there shows no clear gap within the yellow cluster, which suggests there is no need for division. In contrast, Figure 9b is more cluttered. A considerable proportion of green points (Region 6) are mixed with yellow ones (Region 5). Additionally, it not only presents a larger mingle among blue (Region 3), orange (Region 1), and yellow (Region 5) clusters, but also a messier distribution of the green cluster (Region 6) and yellow cluster (Region 5) in the area possessed by the pink cluster (Region 2).

Moreover, the #5 eigenplace of Region 5 detected by LRA is dominated by the work activity, while other eigenplaces of the same region are dominated by entertainment and dining. The #3 eigenplace in Region 4 is also inconsistent with others. These inconsistencies within the same region reveal a fuzzy function detection of LRA. In conclusion, our model outperforms LRA in a clearer functional division.
5. Conclusions

In this study, we proposed a model based on the idea that urban spatial structure could be inferred from human temporal activity with the aid of big geospatial data. Specifically, the human temporal activity within a geographical unit tells when and why people visit it, which helps in interpreting the social functionality of the geographical unit. We assumed that features of all geographical units in terms of human temporal activity lie in a high-dimensional space which can be represented by a union of multiple low-dimensional subspaces. Moreover, subspaces are found under the constraint that a geographical unit can be only represented by the combination of other geographical units within the same subspace. By this means, geographical units with the same temporal activity patterns are in the same subspaces. Therefore, recovering subspaces leads to the detection of functional regions.

Our model is divided into three parts: data processing, detection, and assessment. First, we reproduce the human temporal activity of an area with geographical units, drop-off events, and check-in events obtained from big geospatial data, and store it in a matrix form. In the detection procedure, aiming to identify functional regions, we apply Sparse Subspace Clustering to the matrix to discover relationships between geographical units and obtain subspaces (functional regions) using spectral clustering. We adopt a multiple subspaces-based technique under the consideration that different functional regions can be different in nature and have intrinsic complexities. Therefore, diverse bases with distinct dimensionality are needed to depict those regions more concisely and accurately. Therefore, our model is more flexible compared with other single space-based models which adopt the assumption that different functional regions share the same set of bases. After detection, we extract significant eigenplaces which illustrate major activities and their patterns within a region to analyze the characteristics of each detected region and label the functionality of it. Furthermore, we propose calculating UDR and RDR by measuring the affinity of subspaces and reconstruction error to assess the spatial structure of the study area.

This study was performed in an area in Shanghai, China. Experimental results and the comparison with the LRA method prove the outperformance of our model when focusing on distinct functional regions. However, additional works are needed. First, spatial aggregation operations are required in this study, but the scale and shape of geographical units may exert impacts on research results, which is the so-called Modifiable areal unit problem [34]. Though we chose a common scheme under the consideration that larger units may have mixed functions, while the smaller unit may both be computational unfriendly and trivial in generating numerous adjacent parcels that can be merged, quantitative research of MAUP is needed for further exploration. Second, the number of clusters is a parameter needs to be predefined; here, we choose the eigengap as a heuristic and the result is in accordance with the number revealed by the similarity matrix in this study. However, the heuristic may fail in more complicated conditions. As for the main data source of human mobility, we chose the GPS trajectory of taxicabs while daily travel transportation also includes buses, subways, etc. We will improve on our work in a future study by incorporating other sources of big geospatial data and extend our research to more cities.

**Author Contributions:** Conceptualization, Jiawei Zhu; Data curation, Jian Peng, Haozhe Huang and Li Chen; Methodology, Jiawei Zhu and Chao Tao; Software, Jiawei Zhu and Xin Lin; Supervision, Qiongjie Wang; Visualization, Jiawei Zhu; Writing—original draft, Jiawei Zhu and Xin Lin. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by National Natural Science Foundation of China (grant numbers 41571397, 41871364); and the Fundamental Research Funds (No. 2019zzts881) for the Central Universities of Central South University.

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

1. Antikainen, J.; Vartiainen, P. Polycentricity in Finland: from structure to strategy. *Built Environ.* **2005**, *31*, 143–152. [CrossRef]

2. Cranshaw, J.; Schwartz, R.; Hong, J.; Sadeh, N. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, 4–7 June 2012.

3. Yuan, N.J.; Zheng, Y.; Xie, X.; Wang, Y.; Zheng, K.; Xiong, H. Discovering urban functional zones using latent activity trajectories. *IEEE Trans. Knowl. Data Eng.* **2014**, *27*, 712–725. [CrossRef]

4. Wang, Y.; Gu, Y.; Dou, M.; Qiao, M. Using spatial semantics and interactions to identify urban functional regions. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 130. [CrossRef]

5. Ferrão, J.; Mourato, J.M.; Balula, L.; Bina, O. *Functional Regions, Urban-Rural Relations and Post 2013 Cohesion Policy: OB-SEVARA–Observatório de Ambiente e Sociedade, Estudo 29; Instituto de Ciências Sociais: Lisboa, Portugal, 2013.

6. Drobne, S.; Konjar, M.; Liscic, A.; Milanovic, N.P.; Lamovsek, A.Z. *Functional Regions Defined by Urban Centres of (Inter)National Importance—The Case of Slovenia*; Liveable, Healthy, Prosperous: Wien, Austria, 2010.

7. 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, 12–16 August 2012; pp. 186–194.

8. Fan, K.; Zhang, D.; Wang, Y.; Zhao, S. Discovering urban social functional regions using taxi trajectories. In *Proceedings of the 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, Beijing, China, 10–14 August 2015; pp. 356–359.

9. Zhang, D.; Huang, H.; Chen, M.; Liao, X. Empirical study on taxi GPS traces for vehicular ad hoc networks. In *Proceedings of the 2012 IEEE International Conference on Communications (ICC)*, IEEE, Ottawa, ON, Canada, 10–15 June 2012; pp. 581–585.

10. Karlsson, C.; Olsson, M. The identification of functional regions: Theory, methods, and applications. *Ann. Region. Sci.* **2006**, *40*, 1–18. [CrossRef]

11. Goddard, J.B. Functional regions within the city centre: A study by factor analysis of taxi flows in central London. *Trans. Inst. Br. Geogr.* **1970**, *49*, 161–182. [CrossRef]

12. Maria Kockelman, K. Travel behavior as function of accessibility, land use mixing, and land use balance: Evidence from San Francisco Bay Area. *Transp. Res. Rec.* **2004**, *1607*, 116–125. [CrossRef]

13. Steiner, R.L. Residential Density and Travel Patterns: Review of the Literature. *Transp. Res. Rec.* **1994**, *1466*, 37–43.

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 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, IEEE, Seattle, WA, USA, 21–25 March 2011; pp. 384–388.

15. Pan, G.; Qi, G.; Wu, Z.; Zhang, D.; Li, S. Land-use classification using taxi GPS traces. *IEEE Trans. Intel. Transp. Syst.* **2012**, *14*, 113–123. [CrossRef]

16. Liu, Y.; Wang, F.; Xiao, Y.; Gao, S. Urban land uses and traffic ‘source-sink areas’: Evidence from GPS-enabled taxi data in Shanghai. *Landsc. Urban Plan.* **2012**, *106*, 73–87. [CrossRef]

17. Pei, T.; Sobolevsky, S.; Ratti, C.; Shaw, S.L.; Li, T.; Zhou, C. A new insight into land use classification based on aggregated mobile phone data. *Int. J. Geograph. Inform. Sci.* **2014**, *28*, 1988–2007. [CrossRef]

18. Zhi, Y.; Li, H.; Wang, D.; Deng, M.; Wang, S.; Gao, J.; Duan, Z.; Liu, Y. Latent spatio-temporal activity structures: A new approach to inferring intra-urban functional regions via social media check-in data. *Geo-Spatial Inform. Sci.* **2016**, *19*, 94–105. [CrossRef]

19. 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]

20. Chen, Y.; Liu, X.; Li, X.; Liu, X.; Yao, Y.; Hu, G.; Xu, X.; Pei, F. Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method. *Landsc. Urban Plan.* **2017**, *160*, 48–60. [CrossRef]

21. Yu, B.; Wang, Z.; Mu, H.; Sun, L.; Hu, F. Identification of Urban Functional Regions Based on Floating Car Track Data and POI Data. *Sustainability* **2019**, *11*, 6541. [CrossRef]

22. Van Der Maaten, L.; Postma, E.; Van den Herik, H. Dimensionality reduction: a comparative. *J. Mach. Learn Res.* **2009**, *10*, 13.

23. Vidal, R. Subspace clustering. *IEEE Signal Proc. Mag.* **2011**, *28*, 52–68. [CrossRef]

24. Elhamifar, E.; Vidal, R. Sparse subspace clustering. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Miami Beach, FL, USA, 20–25 June 2009; pp. 2790–2797.

25. Soltanolkotabi, M.; Elhamifar, E.; Candès, E.J. Robust subspace clustering. *Ann. Stat.* **2014**, *42*, 669–699. [CrossRef]

26. Elhamifar, E.; Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intel.* **2013**, *35*, 2765–2781. [CrossRef] [PubMed]

27. Shen, H.W.; Cheng, X.Q. Spectral methods for the detection of network community structure: a comparative analysis. *J. Stat. Mech. Theory Exp.* **2010**, *2010*, P10020. [CrossRef]

28. Chu, R.L. Urban sustainability and the urban forms of China’s leading mega cities: Beijing, Shanghai and Guangzhou. *Urban Policy Res.* **2012**, *30*, 359–383. [CrossRef]

29. Liu, Y.; Zhan, Z.; Zhu, D.; Chai, Y.; Ma, X.; Wu, L. Incorporating multi-source big geo-data to sense spatial heterogeneity patterns in an urban space. *Geomat. Inform. Sci. Wuhan Univ.* **2018**, *43*, 327–335.
30. Liu, X.; Gong, L.; Gong, Y.; Liu, Y. Revealing travel patterns and city structure with taxi trip data. *J. Transp. Geograph.* 2015, 43, 78–90. [CrossRef]

31. Reades, J.; Calabrese, F.; Ratti, C. Eigenplaces: Analysing cities using the space—time structure of the mobile phone network. *Environ. Plan. B Plan. Desi.* 2009, 36, 824–836. [CrossRef]

32. Toole, J.L.; Ulm, M.; González, M.C.; Bauer, D. Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD International Workshop on Urban Computing, Beijing, China, 12 August 2012; pp. 1–8.

33. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 2008, 9, 2579–2605.

34. Wong, D. The modifiable areal unit problem (MAUP). *SAGE Handb. Spat. Anal.* 2009, 105, 2.