

Article

Integrating the Eigendecomposition Approach and k-Means Clustering for Inferring Building Functions with Location-Based Social Media Data

Feng Gao ^{1,2,3,†} , Guanping Huang ^{1,4,†}, Shaoying Li ^{1,*} , Ziwei Huang ¹ and Lei Chai ¹

¹ School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China; 2111801048@e.gzhu.edu.cn (F.G.); 2111801062@e.gzhu.edu.cn (G.H.); 2111901046@e.gzhu.edu.cn (Z.H.); 2112101059@e.gzhu.edu.cn (L.C.)

² Guangzhou Urban Planning & Design Survey Research Institute, Guangzhou 510060, China

³ Guangdong Enterprise Key Laboratory for Urban Sensing, Monitoring and Early Warning, Guangzhou 510060, China

⁴ Guangdong Provincial Institute of Land Surveying & Planning, Guangzhou 510062, China

* Correspondence: lsy@gzhu.edu.cn

† These authors contributed equally to this work.



Citation: Gao, F.; Huang, G.; Li, S.; Huang, Z.; Chai, L. Integrating the Eigendecomposition Approach and k-Means Clustering for Inferring Building Functions with Location-Based Social Media Data. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 834. <https://doi.org/10.3390/ijgi10120834>

Academic Editor: Wolfgang Kainz

Received: 2 November 2021

Accepted: 10 December 2021

Published: 13 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Understanding the relationship between human activity patterns and urban spatial structure planning is one of the core research topics in urban planning. Since a building is the basic spatial unit of the urban spatial structure, identifying building function types, according to human activities, is essential but challenging. This study presented a novel approach that integrated the eigendecomposition method and k-means clustering for inferring building function types according to location-based social media data, Tencent User Density (TUD) data. The eigendecomposition approach was used to extract the effective principal components (PCs) to characterize the temporal patterns of human activities at building level. This was combined with k-means clustering for building function identification. The proposed method was applied to the study area of Tianhe district, Guangzhou, one of the largest cities in China. The building inference results were verified through the random sampling of AOI data and street views in Baidu Maps. The accuracy for all building clusters exceeded 83.00%. The results indicated that the eigendecomposition approach is effective for revealing the temporal structure inherent in human activities, and the proposed eigendecomposition-k-means clustering approach is reliable for building function identification based on social media data.

Keywords: social media data; building function; eigendecomposition; k-means clustering; Guangzhou

1. Introduction

The relationship between human activity patterns and urban spatial structure has been a key research topic in urban geography and urban planning [1,2]. As the fundamental structural elements of urban physical space [3], buildings are the basic spatial unit for urban spatial structure and urban form studies [4,5]. Buildings are also the important carries of human activities (e.g., living, working, and entertainment) in the urban socioeconomic space [5], which can serve as the basic unit for analyzing human socioeconomic activities and urban functional areas. Identifying building functions is significant for understanding urban spatial structure and urban functional areas, which can assist urban management and future smart city planning.

Traditional methods for identifying functional urban areas rely on land use maps and questionnaire data, which are not time-effective and tend to be subjective in classification [6–8]. Recently, remote sensing data, such as high-resolution images and light detection and ranging (LiDAR) data were widely used to identify building types based on the physical information such as the outline, spectrum, and texture of the buildings [9–13].

However, such methods lack sufficient socioeconomic attributes. The socioeconomic functions of buildings are often closely related to human activities [3]. As network information technology has continued to develop in recent years, massive data about individuals' real-time mobile trajectory are generated through location-based services (LBS) such as mobility trajectory data [14–17] and social media data [1,2,18–21]. As these LBS data are usually collected from individuals, they have shown advantages in and potential for reflecting human activities at fine spatiotemporal resolutions [19,22–27]. Recently, attempts have been made to identify building functions by using these LBS datasets. For example, Chen et al. (2017) assumed that social media activities in buildings of similar functions have similar temporal patterns and applied a dynamic time warping distance based k-medoids method to delineate urban function areas from building level social media data [1]. Niu et al. (2017) integrated multisource big data (taxi trajectory data, social media data, and point of interest (POI) data) and proposed a density-based method to infer urban building functions [19]. Zhuo et al. (2019) identified building functions based on the population density and interactions of people among buildings [3].

In conclusion, on the one hand, these studies have revealed the relationship between the temporal patterns of human activities and building types and proved that LBS datasets, which reflect individual behaviors, are effective in building function inference. On the other hand, these existing studies applied various types of clustering methods to delineate building functions based on massive volume and high dimensional data ignoring the data dimensionality reduction. However, reducing high dimensions is a necessary step of data preprocessing in geospatial big data analysis due to the following reasons. Firstly, facing the massive volume and high dimensional data, dimensionality reduction, through the eigendecomposition approach, is helpful to avoid data redundancy; then, one can use fewer variables to explain most of the information in the original data and transform many highly correlated variables into independent and irrelevant variables. Finally, modeling or clustering results would be more stable by reducing the data dimension. Second, only the original temporal patterns of the raw data (e.g., hourly population activity of a building) were presented and explored in previous studies.

Little is known, however, about the hidden characteristics of the temporal patterns of population activity at building scale from the literature. The eigendecomposition method is capable to construct new features, that is, uncover the hidden structures of population activity temporal patterns by disclosing how they resemble or deviate from the base mode (average pattern) of the study area. How dynamic population activity changes over time within different buildings and how these temporal patterns vary among buildings and the base mode of the study area still remains unknown.

Therefore, to fill the research gap of previous studies, we proposed a novel method, an integrated method of the eigendecomposition approach and k-means clustering, to infer building functions based on location-based social media data, Tencent user density (TUD) data. The eigendecomposition approach has been employed to uncover human activity patterns [28–30]. For example, Eagle and Pentland (2009) used the eigendecomposition method to identify the structure inherent in human daily behavior, and they demonstrated that this dimensionality reduction technique can be used to represent behavioral structures in related research [28]. This method was employed by Gong et al. (2017) to capture the common patterns of passengers' variation over time from a metro smart card dataset to explore the spatiotemporal structure of dynamic urban space [29]. In Xu et al. (2019)'s recent study, the eigendecomposition method was proposed to unravel the landscape and pulses of cycling activities from a dockless bike-sharing system [30]. In our study, the eigendecomposition method was employed to capture the hidden structures of the temporal patterns of human activities using TUD data at building level. By using the eigendecomposition method, we can extract the principal components (PCs) from a 48 h TUD index on weekdays and weekends and characterize the temporal patterns of human activities for each building with low dimensional structures to remove data redundancy. The k-means clustering method was then used to classify the buildings according to the

extracted PCs. The proposed method was illustrated with a case study of Tianhe district, Guangzhou. This study aims to examine the integration of eigendecomposition approach and k-means clustering in inferring building function types. The remainder of this paper is organized as follows. Section 2 presents the study area and dataset. Section 3 introduces the methods used in this study, and Section 4 presents the results. The discussion and conclusions are summarized in Section 5.

2. Study Area and Data

2.1. Study Area

Tianhe district, which is located in the downtown area of Guangzhou, China, was selected as our study area (Figure 1). Tianhe covers an area of 96.33 km² and had a residential population of 1,545,700 in 2015 (Bureau of Statistics of Guangzhou 2015, <http://www.gzstats.gov.cn/tjgb/qstjgb/>, accessed on 15 December 2015). The area has highly concentrated and diverse resources in housing, commerce, transportation, and education. In particular, more than 20 colleges and universities are located in this area. After 40 years of reform and opening up, China has seen rapid growth in urbanization and modernization. Tianhe District has transformed from a group of traditional villages to the new central business district (CBD) of Guangzhou City. In recent years, its economy has been at the forefront of China's CBDs. Compared with other districts and counties, Tianhe District has a wider variety of buildings. The classification of building functions in the district, especially the analysis of the functions of mixed-use buildings, can provide scientific evidence and a valuable reference for urban planning and decision making in government.

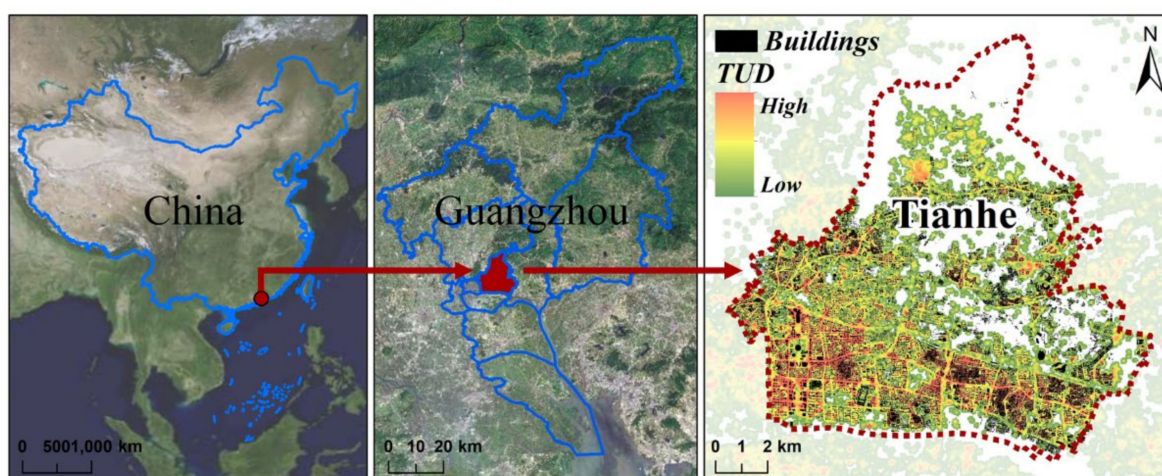


Figure 1. Study area.

2.2. Data

Four different datasets, including building footprint data, TUD dataset, POIs data, and 0.55 m high-resolution remote sensing images in Tianhe, were used in this study.

The TUD dataset includes data about Tencent EasyGo (EasyGo, <http://ur.tencent.com/articles/100>, accessed on 15 June 2015) obtained via the Internet for a continuous week (15–21 June 2015) with temporal and spatial resolutions of 1 h and 25 m, respectively. The dataset records location information of users of smart terminal devices for Tencent's LBS apps such as Tencent QQ, WeChat, and Tencent Maps. Tencent is one of the largest platforms for instant messaging and social networking in China. Guangzhou is one of China's first-tier cities, and more than 93% of its population uses Tencent products (<http://bigdata.qq.com>, accessed on 15 January 2016). Moreover, TUD is a type of aggregated dataset, which can avoid the risk of user privacy. Without any private information on users, TUD records the coordinates, the number of active users, and the recording time stamp of the sample points, which can be regarded as dynamic population density grid data.

Given the huge user coverage of TUD in China and its high spatial-temporal resolution, the data quality of TUD is enough to help perform geospatial analysis at fine scale, and it can well represent realtime and dynamic information of human activities in the city [20,21]. Therefore, this study aims to infer building functions based on the daily rhythms of human activity intensity within buildings at a fine spatial and temporal scale. Firstly, we assumed that the human activity intensity of different functional buildings varies greatly at different times. For example, office buildings in CBD usually have the highest people flow during working hours (daytime) and the least at night, with the characteristics of morning peak and night decline. On the contrary, the flow of people in residential buildings usually starts to decrease in the morning and increase in the evening, and it is relatively stable at night.

Building footprint data were obtained from Amap images in 2017 (<http://ditu.amap.com/>, accessed on 15 January 2017) through web crawler technology. In total, 24,064 building outlines were captured, including the information on height. The POIs dataset was obtained via Baidu Map API (<https://lbsyun.baidu.com/>, accessed on 15 January 2015) in 2015, including more than 100,000 items including information on name, address, and category. Considering the purpose of this research, the POIs data were grouped into 20 categories such as colleges and universities, residential areas, office buildings, and shopping malls. The 0.55 m remote sensing images from April 2015 were obtained from the official website of bigemap (<http://www.bigemap.com/>, accessed on 15 January 2015).

3. Methodology

Previous study has shown that buildings with different functions have different TUD temporal patterns, and human activities in buildings of similar functions have similar spatiotemporal patterns [1]. As the TUD data exhibited a strong periodic regularity over one week, we aggregated the TUD dataset at building level per hour and averaged the TUD data for weekdays and weekends. Hence, we obtained the average TUD data of 48 h for each building. To reveal the hidden structures of the temporal patterns of building level TUD data, an eigendecomposition was performed to extract the effective PCs to remove data redundancy from a large set of 48 indicators. The PCs were used to characterize the underlying temporal patterns of human activities for each building. Furthermore, the k-means clustering, an unsupervised iterative clustering algorithm that is simple, efficient, and easy to implement, was employed to perform a cluster analysis of buildings based on the extracted PCs. Then, the POI density index was employed to help interpret the building function types of different clusters. The overall workflow of the study is presented in Figure 2. More details about each method are presented in the following sections.

3.1. The Eigendecomposition Method

In this study, an eigendecomposition method was used to obtain the effective PCs from a total of 48 variables of TUD data, which were most consistent with the inherent variation of the TUD dataset. The application of the eigendecomposition method in this study had the following advantages. First, the resulting PCs, ranked according to the scores of the eigenvalues or the explained variance, represent the basic structure of the TUD dataset. That is to say, the extracted PCs can indicate the underlying temporal patterns of human activities of buildings. Second, there may be correlation between the obtained 48 TUD variables. The eigendecomposition method can transform a set of correlated variables into several uncorrelated orthogonal variables [31]. We can represent the human activity temporal patterns with the extracted PCs, the lower dimensional, and uncorrelated variables. Third, the eigendecomposition method is a widely used method for removing data redundancy due to its simplicity and straightforward interpretation.

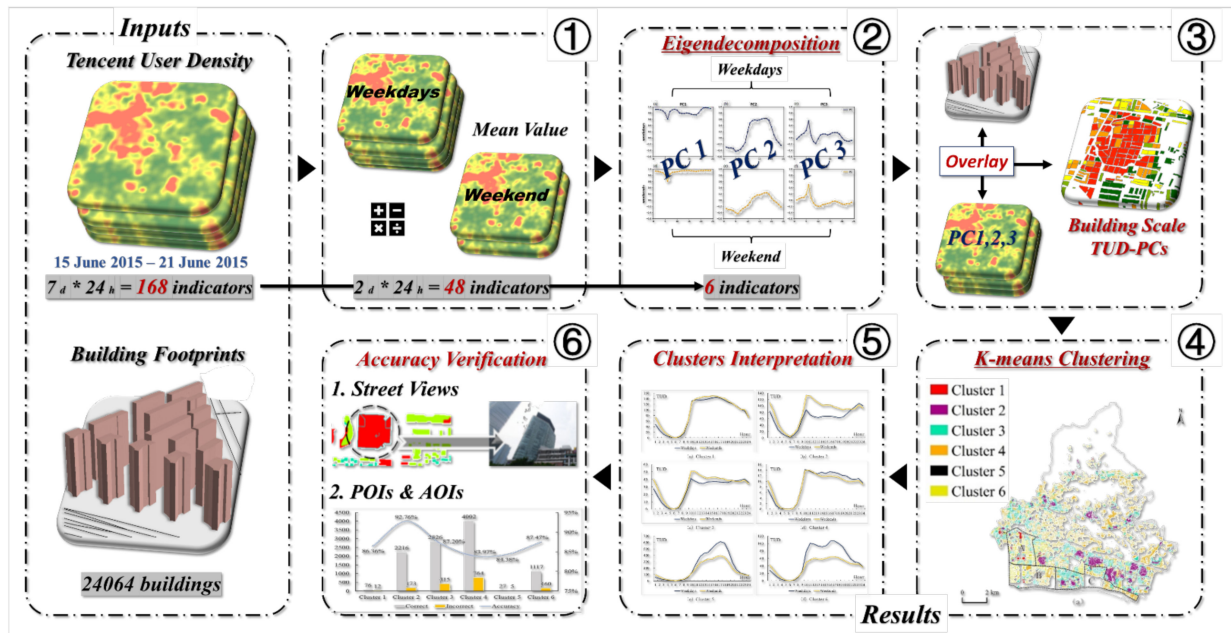


Figure 2. Workflow of the study.

Eigendecomposition is a method to decompose a matrix into its eigenvalues and eigenvectors. Let X be a vector of n random variables and C be the covariance or correlations matrix of X . In this study, X is a vector of 48 variables obtained from hourly TUD data. The symmetric matrix C can be expressed as follow,

$$C = \begin{bmatrix} r_{1,1} & \cdots & r_{1,48} \\ \vdots & \ddots & \vdots \\ r_{48,1} & \cdots & r_{48,48} \end{bmatrix} \quad (1)$$

In matrix C , $r_{i,j}(i, j = 1, \dots, 48)$ refers to the correlation coefficient of the original variables x_i and x_j . The matrix C can be decomposed into eigenvalues and eigenvectors according to the following equation,

$$ACA^T = V \quad (2)$$

where V is the diagonal matrix of eigenvalues, $A = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$ is the matrix of eigenvectors, and T is the transposition function. The eigenvalues of the matrix C are $\lambda_1, \lambda_2, \dots, \lambda_n$, where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots \geq \lambda_n$.

As the eigenvector of C is a transformation of X into the corresponding PC, we can obtain the PCs Z from the original dataset X . The transformation is based on the following equation [29]:

$$Z = \{X\alpha_1, X\alpha_2, \dots, X\alpha_n\} = XA \quad (3)$$

A can also represent the loadings of X with its PCs, and its columns are the eigenvectors. The larger the eigenvalue is, the larger the variance of the matrix on the corresponding eigenvector is, and the more information there is. Thus, we can evaluate the importance of each PC according to the eigenvalue that reflects the variance of the original variables. Generally, the first PC with the maximum eigenvalue, Z_1 obtains the maximum variance of the n original variables, indicating that Z_1 contains the maximum information of the original variables. In the same way, Z_i contains the i th maximum information of the original variables. Based on the empirical rules, we can extract the first few PCs whose variances reach 85% to 95% of the original dataset, to represent the original dataset. Moreover, we can use the extracted effective PCs to unravel the underlying temporal patterns of

human activities of buildings by analyzing the coefficients of these PCs. In summary, the function of the eigendecomposition method in this study is to first remove redundancy in the original dataset of 48 TUD indicators, then to transform correlated 48-variables into several uncorrelated orthogonal components to uncover the hidden human activity temporal patterns with low dimension structures in a unified manner.

3.2. *k*-Means Clustering

The *k*-means and K-medoids algorithm have been widely applied in the identification of urban function [32–34]. Previous studies on urban problems show that *k*-means clustering has a good performance in clustering index variables after dimensionality reduction [35–37]. In this study, *k*-means clustering was employed to perform a cluster analysis of buildings according to the extracted PCs, which reflected the temporal patterns of human activities [33].

The *k*-means algorithm, which is an unsupervised clustering algorithm, is usually used to classify a sample set of objects into *k* different clusters according to the similarity of their attributes [34]. The sample set is divided into *i* clusters according to the distance between the samples [32]. The purpose of clustering is to minimize the distance of samples within the same cluster and maximize the distance between clusters. The cluster is assumed to be divided into (C_1, C_2, \dots, C_k) , then, the ultimate goal is to minimize the sum of squared error (SSE), as calculated below [32],

$$E = \sum_{i=1}^k \sum_{x \in C_i} |x - \mu_i|^2 \quad (4)$$

where *x* represents the value of the sample object, and μ_i is the mean value of the cluster C_i calculated as follows:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \quad (5)$$

3.3. Identification of Building Function

The types of building functions were determined by observing and analyzing the characteristics of time series of human activities on weekdays and weekends in various clusters based on changes in frequencies and patterns of TUD in different types of buildings on weekdays and weekends. Then, a POI dataset was introduced to help interpret the clustering results. The density index of POI, $F_{i,l}$ was adopted to indicate the richness of land-use types [37], which was represented as follows,

$$F_{i,l} = \frac{n_{l,i}/n_i}{N_l/N} \quad (6)$$

where $F_{i,l}$ denotes the degree of richness of Type *l* POIs in *i* number of buildings, $n_{l,i}$ represents the number of Type *l* POIs in the buffer zone of the *i*-th building, n_i is the total number of all POIs in the buffer zone of *i* number of buildings, N_l is the total number of Type *l* POIs in the buffer zones of all buildings, and *N* denotes the total number of POIs in the entire study area (with a 10 m buffer zone). The greater the density index, the larger the number of types in the cluster, and the more concentrated the distribution.

4. Results

4.1. The Temporal Structures of Human Activities

We first performed the eigendecomposition to extract the effective PCs to remove data redundancy from a large set of 48 TUD indicators, including 24 h on weekdays and weekends. Table 1 presents the percentage of variance explained by the top few PCs. According to the results, PC1 accounted for 77.92% of the total variance, PC2 accounted for 11.51% of the total variance, and PC3 accounted for 2.47%. The first three PCs with eigenvalues

greater than one were extracted, which explained about 91.90% of the total variance, in combination, indicating a good approximation of the original 48 TUD indicators. Hence, the first three PCs were taken as the effective comprehensive indicators to characterize the human activities patterns of buildings in this study.

Table 1. Variance explained by the principal components.

Principal Component (PC)	Eigenvalue	Variance Explained by (%)	Accumulative Variance Explained by (%)
1	37.40	77.92	77.92
2	5.52	11.51	89.43
3	1.19	2.47	91.90

We further explored the hidden temporal structures of human activities at building level from the eigendecomposition (Figure 3). PC1, which explained the largest percentage of the original dataset (77.92%), had very large positive values during early morning (1:00–4:00), morning peak (7:00–9:00), and evening after work (22:00–24:00) on weekdays (Figure 3a). This indicated that the amount of human activity during these periods differed among different buildings. For weekends, PC1 had large positive values during 8:00–24:00 (Figure 3b), indicating a large spatial variation of human activity at building level for this period. PC2 explained a lower but still significant amount of the total variance (11.51%) and exhibited notable peaks during working hours on weekdays (Figure 3a). This means it was more distinguishable among different buildings during this period. Regarding weekends, the PC2 showed a small peak during 14:00 to 19:00, which is leisure time (Figure 3b). PC3, which explained the lowest amount of the total variance (2.47%) among the three extracted PCs, showed a peak value at 6:00 (the starting point for the increase in Tencent users in one day) on both weekdays and weekends (Figure 3). This indicated the significant spatial variation of human activity among buildings at this time.

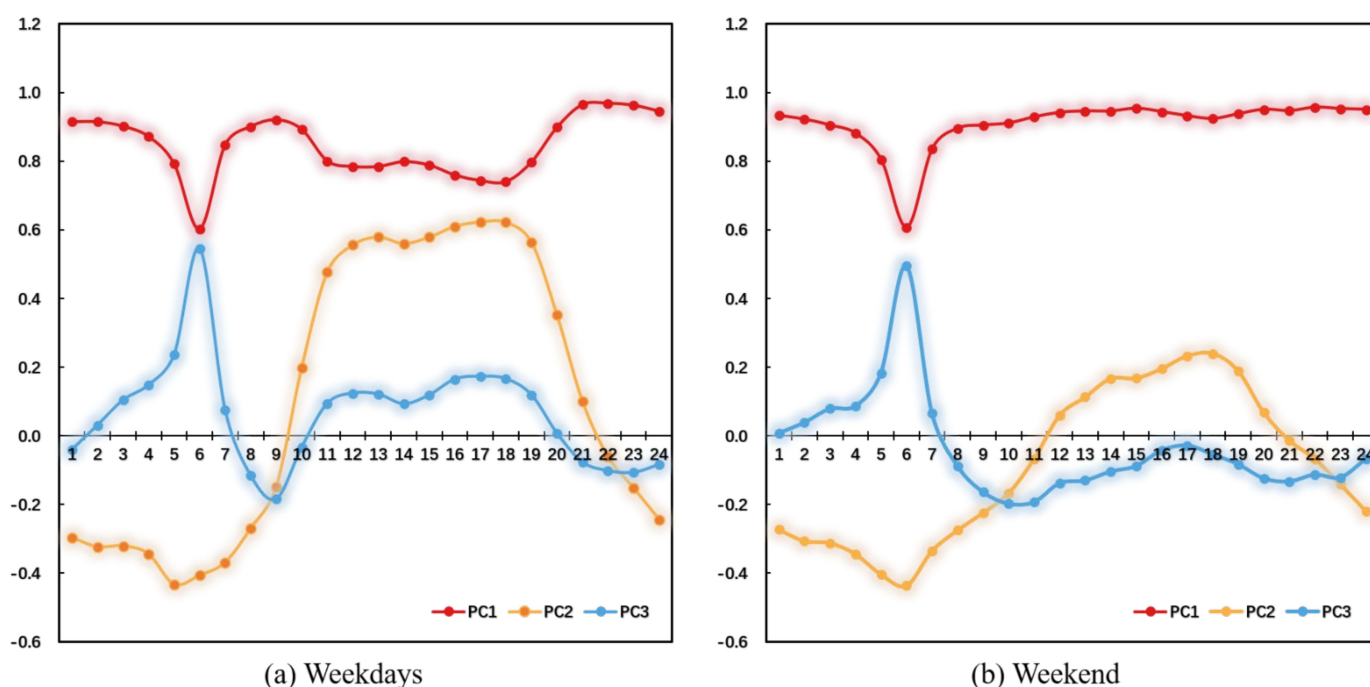


Figure 3. Results of the eigendecomposition. (a) PCs on weekdays, (b) PCs on weekend.

The resultant PCs, ranked by the fraction of the variance explained, indicated the hidden structure of the TUD dataset. The corresponding coefficients associated with a particular entity (e.g., temporal population activity pattern in a building) demonstrate its

deviation from the norm (i.e., average pattern of the study area). Hence, the eigendecomposition method can be adopted to answer our research question—how do the temporal patterns of population activity within different buildings resemble or deviate from the base mode of the study area?

4.2. Results of Building Function Inference and the Spatial Patterns

As the top three PCs approximated the original TUD data, the temporal patterns of human activities at building level can be characterized by the linear combination of the original variables (X) and the corresponding coefficients (A) based on Equation (3). Based on the three extracted PCs, the k-means clustering was employed to cluster the buildings into K types, and the POI density index was used to help identify the building functions.

The optimal number of clusters was determined by the sum of squared errors (SSE, Equation (4)). The smaller the SSE, the better the cluster effect. As Figure 4 shows, the SSE decreased rapidly as the number of clusters increased, and its value appeared to stabilize when $k \geq 8$. To ensure the reliability of results, clustering experiments were conducted with $k = 6$ –15. It was found that clustering with $k = 8$ can represent cases with $k > 8$. Therefore, $k = 8$ was selected for the cluster analysis in this study.

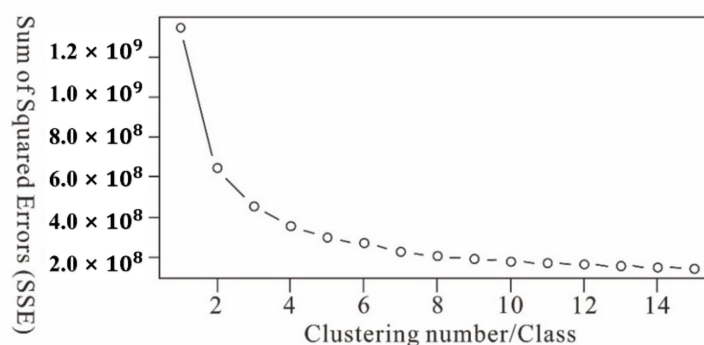


Figure 4. The relationship between Clustering number and the SSE.

It was found that the TUD curves of some building categories were relatively close, from the results of $k = 8$. Hence, the similar categories were merged, and six categories of buildings were obtained. Figure 5 shows the average hourly TUD on the weekdays and weekends for the six clusters. Then, the POI density index was employed to help interpret the building function types of different clusters. We calculated the density index of POI, $F_{i,l}$, (Equation (6)) for all buildings and calculated the average of each building cluster (Table 2) to help interpret the building functions. Currently, no unified standards exist for the classification of building functions. This study classified the functions of buildings as residential buildings; commercial buildings; and buildings for wholesale and retail, work, and science and education. The spatial distributions of the building clusters are shown in Figure 6.

Cluster 1 (commercial/wholesale and retail): This cluster was characterized by high POI density values for the CP (commercial plaza) and the WR (wholesale and retail) type (Table 2). In terms of the TUD curve (Figure 5a), the morning peak was between 9:00 and 10:00 and continued to fall after the evening peak. The TUD during the daytime on weekends was slightly higher than that on weekdays, suggesting that more people came then for entertainment and shopping. The overall temporal profiles matched the opening times (9:00–22:00) of CP and WR. This type of building was concentrated on one side of the street. This cluster included large commercial plazas such as Grandview Plaza and the Tianma Clothing Wholesale Market on Shahe Street (Figure 6a,b).

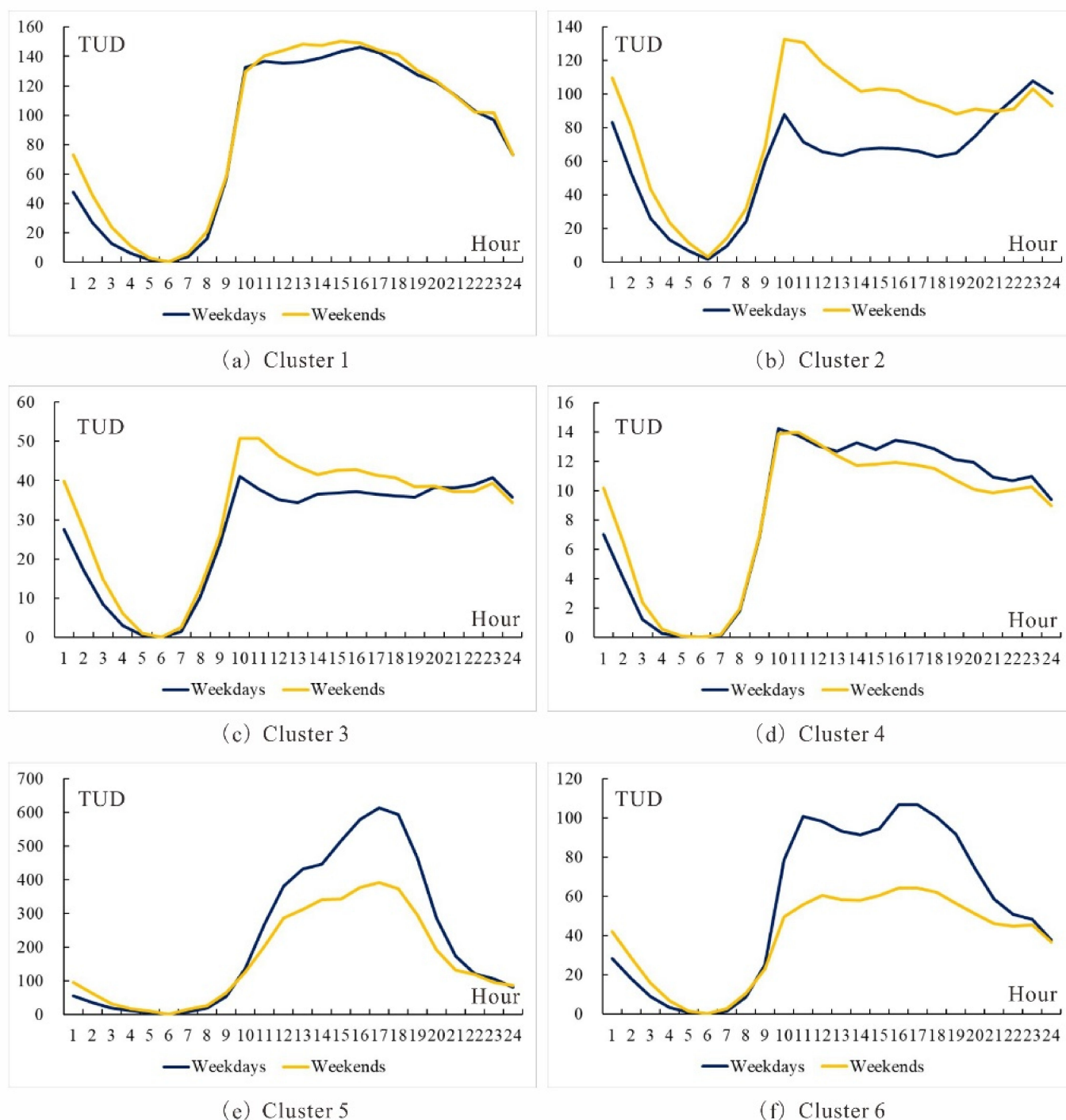


Figure 5. The average hourly TUD for the six clusters on weekdays and weekends.

Cluster 2 (urban village/residential district): Regarding the TUD curve (Figure 5b), the TUD value increased significantly between 6:00 and 9:00 as well as 19:00 to 23:00, and it was low on weekdays between 10:00 and 19:00. On weekends, the TUD values were much higher than weekdays. This is a typical temporal pattern of a residential building, which refers to the population's outflow during working hours on weekdays and staying home during the daytime on weekends. In terms of the POI density index (Table 2), SHS (scenic and historic spots), CS (catering services), HA (hotels and apartments), and LS (living services) had high values. Note that most SHS in this category were ancestral halls in urban villages. This cluster consisted of many urban villages in Tianhe District (e.g., Tangxia, Shipai, and Chebei) and other nearby residential districts. This cluster had obvious

features of residential functions (Figure 6a,c). Urban villages usually have low rent, so they attract a large number of migrants and have many living facilities. These results indicate that the TUD time series can also well capture the specific pattern of human activities in urban villages.

Table 2. Density index values of POIs in different categories grouped by clusters.

POIs Density	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
CS	0.98	1.42	1.03	0.56	0.58	0.87
IP	0.07	0.07	0.83	3.89	0.00	0.99
SHS	0.41	2.57	0.75	1.16	0.00	0.25
HEI	0.03	0.02	0.52	5.05	0.00	0.87
EP	0.49	0.33	0.93	2.47	0.00	1.18
CP	3.13	0.57	0.57	0.57	2.80	1.18
TF	0.88	0.28	1.01	1.52	0.50	1.42
FIS	0.57	0.12	0.44	0.69	0.47	2.87
HA	0.74	1.27	0.74	0.63	0.70	1.44
SRI	0.16	0.06	0.39	5.05	0.00	0.96
WR	1.43	0.97	1.01	0.95	1.72	0.81
OB	0.55	0.29	0.69	1.15	0.72	2.14
DB	0.40	0.25	0.79	0.45	0.69	2.45
LS	0.71	1.21	1.04	0.82	0.86	0.98
SR	0.89	0.62	0.90	1.29	0.61	1.38
MS	0.77	1.16	1.05	0.92	0.00	1.00
GA	0.33	0.58	1.20	1.81	0.25	0.99
VTs	0.42	0.10	0.84	3.77	0.00	0.88
PSS	0.00	0.14	1.43	3.78	0.00	0.20
RC	0.48	0.74	1.36	1.55	0.27	0.71

Note: CS = catering services; IP = industrial parks; SHS = scenic and historic spots; HEI = higher education institutions; EP = enterprises; CP = commercial plaza; TF = transportation facilities; FIS = financial and insurance services; HA = hotels and apartments; SRI = scientific research institutions; WR = wholesale and retail; OB = office buildings; DB = dual-purpose buildings (commercial and residential); LS = living services; SR = sports and recreation; MS = medical services; GA = government agency; VTs = vocational and technical schools; PSS = primary and secondary schools; and RC = residential community.

Cluster 3 (residential district/urban village): This cluster was characterized by high POI density values for the RC (residential community) and LS (living services) (Table 2). This cluster had comprehensive community facilities and was a mature residential area. House prices in this cluster were generally higher than in Cluster 2, but it had a lower density. The overall characteristics of the temporal profiles were similar to that of Cluster 2 (Figure 5b,c). This cluster spatially surrounded Cluster 2 and consisted of residential districts on both sides of the streets and a small number of buildings in urban villages (Figure 6a,d). The biggest difference between Cluster 3 and Cluster 2 was the different ratio of urban villages to residential districts in the cluster.

Cluster 4 (science and education/work): This type of building was the most widely distributed and the most numerous in the study area. Relatively high values on the POI density index were found for HEI (higher education institutions), SRI (scientific research institutions), IP (industrial parks), VTs (vocational and technical schools), and EP (enterprises) (Table 2). The TUD value increased from 6:00 and peaked at 10:00. Its values on weekdays between 13:00 and 24:00 were higher than on weekends (Figure 5d). This indicates that more people flow into this cluster on weekdays than weekends. Workplaces usually have fewer people on weekends, and students in colleges and universities have more freedom in their activities and tend to go out for entertainment. In terms of spatial distribution, this cluster was closely connected with Cluster 3. It consisted of a large number of scientific and educational institutions as well as industrial parks including South China University of Technology, Guangdong Academy of Agricultural Sciences, National Software Industry Bases, and Guangxin Creative Industry Park (Figure 6a,e).

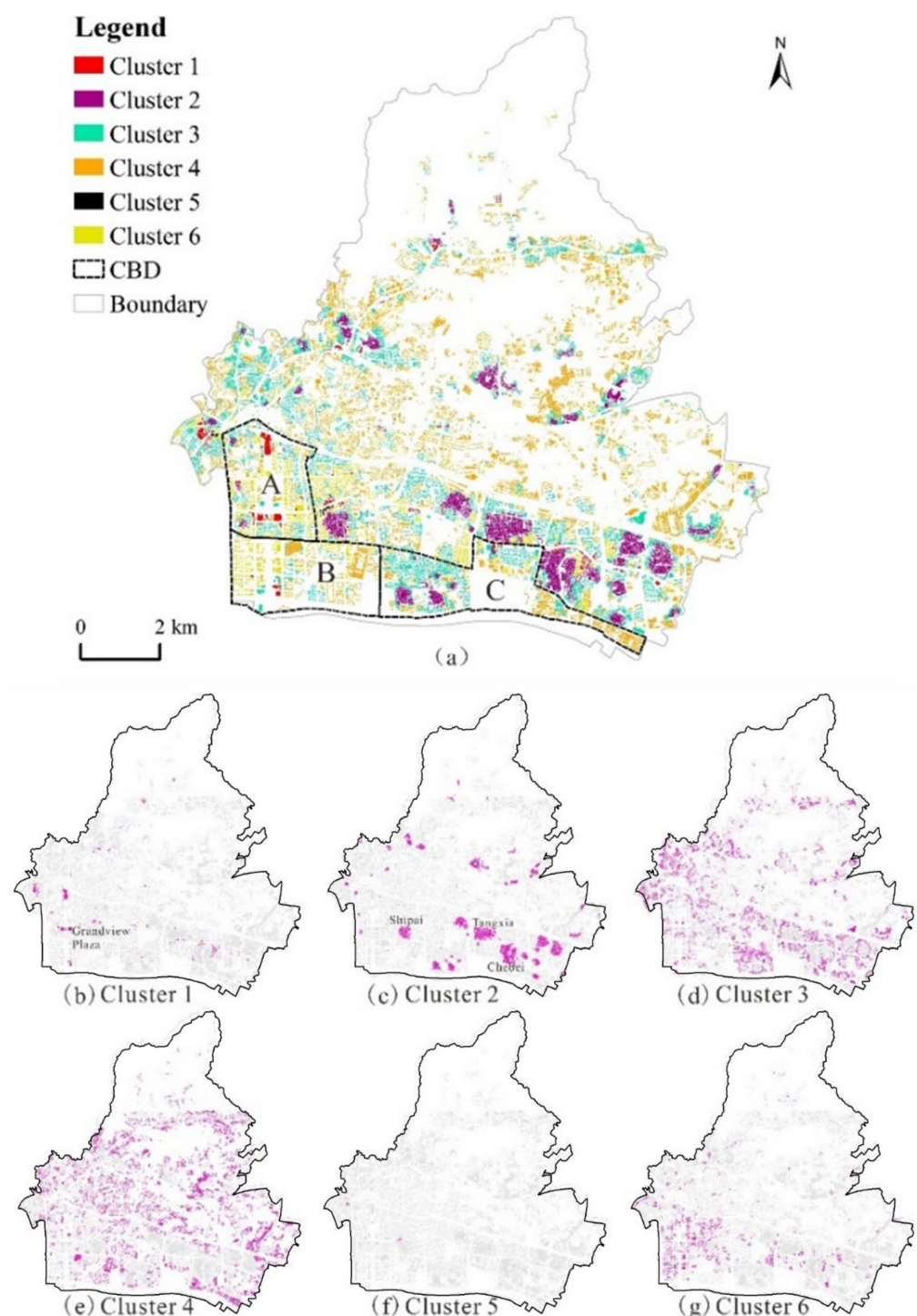


Figure 6. The spatial distribution of the building clusters.

Cluster 5 (wholesale and retail): In the ranking of the POI density index values, WR had high ranks in Cluster 5 (Table 2). Regarding the TUD temporal profiles, the population was concentrated on weekdays between 10:00 and 22:00 and peaked at around 17:00. This is usually the peak time for buyers to purchase goods and for logistics companies to receive and ship items, which is also close to dinner time; in addition, similar patterns of activities were also observed on weekends. Despite the above, the TUD value was relatively low in general. As one of the first-tier cities in China, Guangzhou has a strong ability to collect and distribute a wide range of goods and services. Low-level cities or regions often purchase goods on weekdays, while weekends are the best times for sub-merchants to return to the

place of sale to sell goods. This cluster mainly consisted of buildings for electronic WR (e.g., Pacific Digital Plaza) and buildings for clothing WR on Shahe Street (Figure 6a,f).

Cluster 6 (work/residential district): This type of building was concentrated in the southern part of Tianhe District (Figure 6a,g). The cluster consisted of many office buildings as well as dual-purpose buildings with both residential and commercial components. These parts of the buildings, with Clusters 3 and 4, constitute the CBD in Tianhe District. In terms of the POI density index (Table 2), high values were observed for buildings for FIS (financial and insurance services), DB (dual-purpose buildings (commercial and residential)), OB (office building), HA (hotels and apartments), and RC (residential community). On weekdays, the TUD value was very high with obvious peaks in the morning and evening. The value dropped slightly at lunch time and was higher on weekdays between 10:00 and 20:00 than on weekends (Figure 5f). These temporal patterns of human activities are the opposite of the patterns observed in Clusters 2 and 3. During the weekend, the distribution of commercial and residential buildings and residential districts show that this type still has a certain level of TUD value. This may be due to working overtime on weekends or the residential population resting at home.

4.3. Accuracy Verification

To verify the clustering results, 11,800 samples were randomly selected from 24,600 buildings in the study area. The selected samples were very representative in number and spatial distribution (Figure 7a). Figure 6b shows that Cluster 2 had the highest accuracy (92.76%) and Cluster 1 the lowest (83.36%). Clusters 3, 4, 5, and 6 had accuracy rates of 86.97%, 87.20%, 84.38%, and 87.47%, respectively. The classification accuracy for all clusters exceeded 83.00%. Typical buildings were selected from each cluster. The results were compared with the attribute information in the AOIs data and the spatial location of building units. In addition, the results were verified according to the street views in Baidu Maps (Table 3). It was confirmed that the extracted typical buildings were all aligned with the clustering results. Thus, it can be concluded that the clustering results of this study demonstrate good reliability.

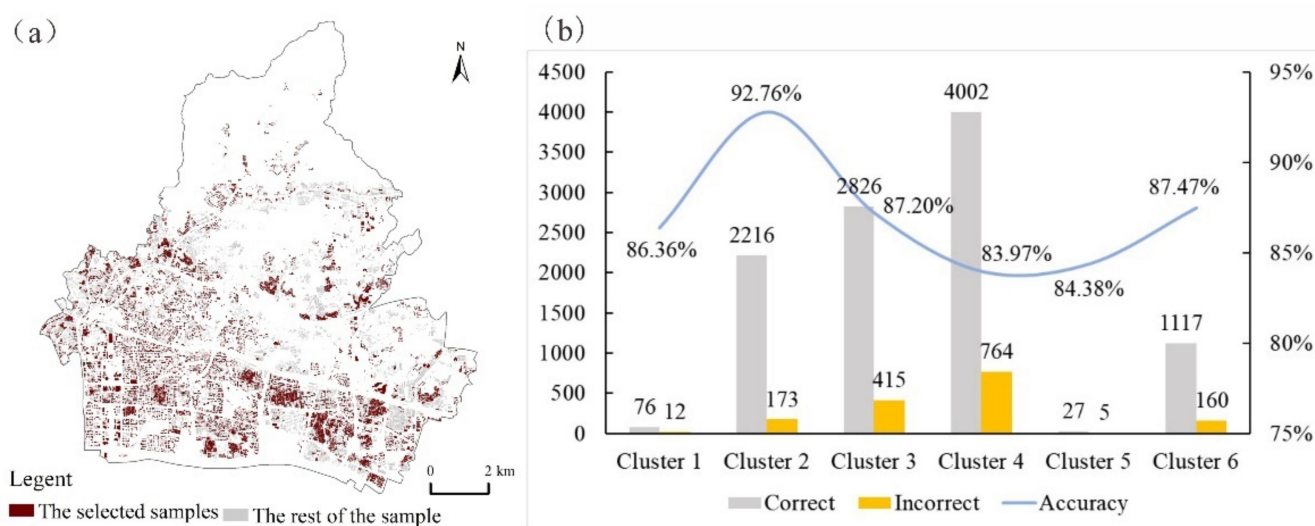
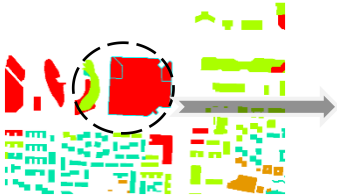

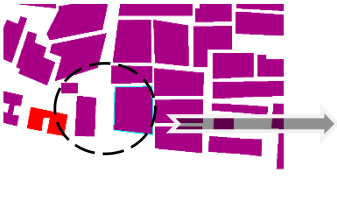



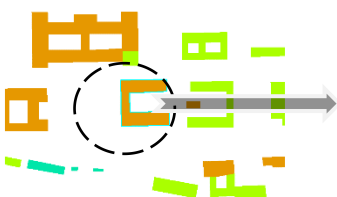



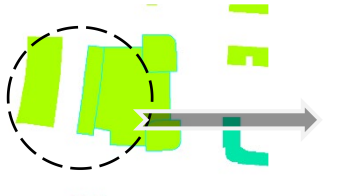



Figure 7. The Spatial distribution of validation samples and the accuracy of the clustering results. (a) Sample buildings (b) Accuracy of the results.

Table 3. Comparison analysis of clustering results and current situation of Baidu street view map.

Cluster	Classification Results	Spatial Position	Baidu Street View Maps	Building Selection Fit
Cluster 1	commercial/ wholesale and retail			Fit: Zhengjia Square is a famous shopping mall.
Cluster 2	urban village/ residential district			Fit: The building in Tangxia Urban Village is one of the places where migrants gather in Guangzhou.
Cluster 3	residential district/urban village			Fit: Poly heart language garden residence is a typical real estate residential area.
Cluster 4	science and education/work			Fit: South China Normal University is located in Wushan Street, where higher learning centers are clustered.
Cluster 5	wholesale and retail			Fit: Pacific Digital Plaza is an electronic products wholesale and retail distribution center.
Cluster 6	work/residential district			Fit: Goldland Plaza is a typical business office building in Zhujiang New Town.

5. Conclusions

This study presented a comprehensive method that integrated the eigendecomposition approach and k-means clustering for inferring building function types based on location-based social media data, Tencent user density (TUD) data. We employed the eigendecomposition method to extract the effective principal components (PCs) and characterized the temporal patterns of human activities for each building with low dimensional comprehensive variables. The k-means clustering method was used to classify the buildings, which was combined with a POI density index to interpret the building function types of different clusters. The building functions in the study area, Tianhe district, Guangzhou,

were classified into six clusters, and the results were verified through the random sampling of AOI data and street views in Baidu Maps. The accuracy of the classification for all clusters exceeded 83.00%. The accuracy assessment demonstrated that the proposed method is reliable to identify building function types based on location-based social media data.

The results of this study might shed light on urban planning and management. First, this study proved that the eigendecomposition approach can effectively characterize the temporal patterns of building-level human activities reflected by social media data. This can maximize the usage of social media data in urban studies. Second, the proposed eigendecomposition–k-means clustering method can address the difficulty of identifying buildings with mixed functions. Some mixed function buildings, such as commercial and residential dual-purpose buildings were identified in this study. Last, the proposed method, which is an easy-to-use method with social media data, has great potential for various applications. The building function inference results reflect the inherent heterogeneity of urban functional areas, which can help understanding the urban spatial structure at fine scale and assist urban planning and management. In particular, the urban village buildings, which are rarely identified in other studies, were identified in our study. The spatial distributions of urban villages can provide decision support for urban planning and management, such as urban renewal, infrastructure planning, urban village population management, etc.

This study also points to future study directions that need further research. First, by using the eigendecomposition method, the coefficients of the effective PCs for different types of buildings can be further used to explore the relationship between human activities and the urban spatial structures. Second, other big data that can reflect the interaction between different buildings, such as taxi data [3,19] and mobile phone data [38,39], can be further combined with TUD data, to further improve the building function inference method. In addition, the bias and limits of social media data should also be recognized. First, though the user coverage of TUD is high in China, it still has gaps such as the elderly and young children who rarely use smartphones [40]. In the future, other data sources on urban population and mobility should be incorporated to supplement the gap in social media big data.

Author Contributions: Conceptualization, Shaoying Li; methodology, Guanping Huang and Feng Gao; software, Guanping Huang and Feng Gao; validation, Guanping Huang and Feng Gao; formal analysis, Guanping Huang and Feng Gao; investigation, Guanping Huang, Feng Gao, Ziwei Huang, Lei Chai; resources, Shaoying Li; data curation, Shaoying Li; writing—original draft preparation, Feng Gao and Guanping Huang; writing—review and editing, Feng Gao, Shaoying Li; visualization, Guanping Huang and Feng Gao; supervision, Shaoying Li; project administration, Shaoying Li; funding acquisition, Shaoying Li and Feng Gao. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China [grant number 41871290, 41401432], Guangdong Enterprise Key Laboratory for Urban Sensing, Monitoring and Early Warning (No. 2020B121202019), The Science and Technology Foundation of Guangzhou Urban Planning & Design Survey Research Institute (RDI2210205064).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. 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](#)]
2. Zhi, Y.; Li, H.F.; Wang, D.S.; Deng, M.; Wang, S.W.; Gao, J.; Duan, Z.Y.; Liu, Y. Latent spatio-temporal activity structures: A new approach to inferring intra-urban functional regions via social media check-in data. *Geo-Spat. Inf. Sci.* **2016**, *19*, 94–105. [[CrossRef](#)]

3. Zhuo, L.; Shi, Q.; Zhang, C.; Li, Q.; Tao, H. Identifying Building Functions from the Spatiotemporal Population Density and the Interactions of People among Buildings. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 247. [\[CrossRef\]](#)
4. Hecht, R.; Meinel, G.; Buchroithner, M. Automatic identification of building types based on topographic databases—A comparison of different data sources. *Int. J. Cartogr.* **2015**, *1*, 18–31. [\[CrossRef\]](#)
5. Chen, W.; Zhou, Y.Y.; Wu, Q.S.; Chen, G.; Huang, X.; Yu, B.L. Urban Building Type Mapping Using Geospatial Data: A Case Study of Beijing, China. *Remote Sens.* **2020**, *12*, 2805. [\[CrossRef\]](#)
6. Heiden, U.; Heldens, W.; Roessner, S.; Segl, K.; Esch, T.; Mueller, A.J.L.; Planning, U. Urban structure type characterization using hyperspectral remote sensing and height information. *Landsc. Urban Plan.* **2012**, *98*, 361–375. [\[CrossRef\]](#)
7. Tian, G.J.; Wu, J.G.; Yang, Z.F. Spatial pattern of urban functions in the Beijing metropolitan region. *Habitat Int.* **2010**, *34*, 249–255. [\[CrossRef\]](#)
8. Van de Voorde, T.; Jacquet, W.; Canters, F. Mapping form and function in urban areas: An approach based on urban metrics and continuous impervious surface data. *Landsc. Urban Plan.* **2011**, *102*, 143–155. [\[CrossRef\]](#)
9. Belgiu, M.; Tomljenovic, I.; Lampoltshammer, T.J.; Blaschke, T.; Höfle, B. Ontology-Based Classification of Building Types Detected from Airborne Laser Scanning Data. *Remote Sens.* **2014**, *6*, 1347–1366. [\[CrossRef\]](#)
10. Lu, Z.Y.; Im, J.; Rhee, J.; Hodgson, M. Building type classification using spatial and landscape attributes derived from LiDAR remote sensing data. *Landsc. Urban Plan.* **2014**, *130*, 134–148. [\[CrossRef\]](#)
11. Huang, Y.H.; Zhuo, L.; Tao, H.Y.; Shi, Q.L.; Liu, K. A Novel Building Type Classification Scheme Based on Integrated LiDAR and High-Resolution Images. *Remote Sens.* **2017**, *9*, 679. [\[CrossRef\]](#)
12. Li, M.; Stein, A.; Bijker, W.; Zhan, Q. Urban land use extraction from Very High Resolution remote sensing imagery using a Bayesian network. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 192–205. [\[CrossRef\]](#)
13. Gilani, S.A.N.; Awrangjeb, M.; Lu, G.J. An Automatic Building Extraction and Regularisation Technique Using LiDAR Point Cloud Data and Orthoimage. *Remote Sens.* **2016**, *8*, 258. [\[CrossRef\]](#)
14. Liu, Y.; Seah, H.S. Points of interest recommendation from GPS trajectories. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 953–979. [\[CrossRef\]](#)
15. Gong, L.; Liu, X.; Wu, L.; Liu, Y. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartogr. Geogr. Inf. Sci.* **2016**, *43*, 103–114. [\[CrossRef\]](#)
16. Gao, F.; Li, S.; Tan, Z.; Zhang, X.; Lai, Z.; Tan, Z. How Is Urban Greenness Spatially Associated with Dockless Bike Sharing Usage on Weekdays, Weekends, and Holidays? *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 238. [\[CrossRef\]](#)
17. Gao, F.; Li, S.; Tan, Z.; Wu, Z.; Zhang, X.; Huang, G.; Huang, Z. Understanding the modifiable areal unit problem in dockless bike sharing usage and exploring the interactive effects of built environment factors. *Int. J. Geogr. Inf. Sci.* **2021**, *35*, 1–21. [\[CrossRef\]](#)
18. Crooks, A.; Pfoser, D.; Jenkins, A.; Croitoru, A.; Stefanidis, A.; Smith, D.; Karagiorgou, S.; Efentakis, A.; Lamprianidis, G. Crowdsourcing urban form and function. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 720–741. [\[CrossRef\]](#)
19. Niu, N.; Liu, X.P.; Jin, H.; Ye, X.Y.; Liu, Y.; Li, X.; Chen, Y.M.; Li, S.Y. Integrating multi-source big data to infer building functions. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1871–1890. [\[CrossRef\]](#)
20. Li, S.; Lyu, D.; Huang, G.; Zhang, X.; Gao, F.; Chen, Y.; Liu, X. Spatially varying impacts of built environment factors on rail transit ridership at station level: A case study in Guangzhou, China. *J. Transp. Geogr.* **2020**, *82*, 102631. [\[CrossRef\]](#)
21. Li, S.; Lyu, D.; Liu, X.; Tan, Z.; Gao, F.; Huang, G.; Wu, Z. The varying patterns of rail transit ridership and their relationships with fine-scale built environment factors: Big data analytics from Guangzhou. *Cities* **2020**, *99*, 102580. [\[CrossRef\]](#)
22. Huang, W.; Li, S.; Liu, X.; Ban, Y. Predicting human mobility with activity changes. *Int. J. Geogr. Inf. Sci.* **2015**, *29*, 1569–1587. [\[CrossRef\]](#)
23. Song, C.M.; Qu, Z.H.; Blumm, N.; Barabasi, A.L. Limits of Predictability in Human Mobility. *Science* **2010**, *327*, 1018–1021. [\[CrossRef\]](#)
24. Shen, Y.; Karimi, K. Urban function connectivity: Characterisation of functional urban streets with social media check-in data. *Cities* **2016**, *55*, 9–21. [\[CrossRef\]](#)
25. Chen, W.; Huang, H.; Dong, J.; Zhang, Y.; Tian, Y.; Yang, Z. Social functional mapping of urban green space using remote sensing and social sensing data. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 436–452. [\[CrossRef\]](#)
26. Tu, W.; Cao, J.Z.; Yue, Y.; Shaw, S.L.; Zhou, M.; Wang, Z.S.; Chang, X.M.; Xu, Y.; Li, Q.Q. Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 2331–2358. [\[CrossRef\]](#)
27. Song, Y.M.; Huang, B.; He, Q.Q.; Chen, B.; Wei, J.; Mahmood, R. Dynamic assessment of PM2.5 exposure and health risk using remote sensing and geo-spatial big data. *Environ. Pollut.* **2019**, *253*, 288–296. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Eagle, N.; Pentland, A.S. Eigenbehaviors: Identifying structure in routine. *Behav. Ecol. Sociobiol.* **2009**, *63*, 1057–1066. [\[CrossRef\]](#)
29. Gong, Y.; Lin, Y.; Duan, Z. Exploring the spatiotemporal structure of dynamic urban space using metro smart card records. *Comput. Environ. Urban Syst.* **2017**, *64*, 169–183. [\[CrossRef\]](#)
30. Xu, Y.; Chen, D.C.; Zhang, X.H.; Tu, W.; Chen, Y.Y.; Shen, Y.; Ratti, C. Unravel the landscape and pulses of cycling activities from a dockless bike-sharing system. *Comput. Environ. Urban Syst.* **2019**, *75*, 184–203. [\[CrossRef\]](#)
31. Xia, L.; Yeh, G.O. Integration of principal components analysis and cellular automata for spatial decisionmaking and urban simulation. *Sci. China* **2002**, *45*, 521–529. [\[CrossRef\]](#)
32. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [\[CrossRef\]](#)

-
33. Rahman, M.A.; Islam, M.Z. A hybrid clustering technique combining a novel genetic algorithm with K-means. *Knowl. Based Syst.* **2014**, *71*, 345–365. [[CrossRef](#)]
 34. Chang, D.-X.; Zhang, X.-D.; Zheng, C.-W. A genetic algorithm with gene rearrangement for K-means clustering. *Pattern Recognit.* **2009**, *42*, 1210–1222. [[CrossRef](#)]
 35. Wu, P.; Zhang, S.; Li, H.; Dale, P.; Ding, X.; Lu, Y. Urban parcel grouping method based on urban form and functional connectivity characterisation. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 282. [[CrossRef](#)]
 36. Gutiérrez, A.; Domènech, A. Identifying the Socio-Spatial Logics of Foreclosed Housing Accumulated by Large Private Landlords in Post-Crisis Catalan Cities. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 313. [[CrossRef](#)]
 37. Verburg, P.H.; Nijs, T.; Eck, J.; Visser, H.; Jong, K.D. A method to analyse neighbourhood characteristics of land use patterns. *Environ. Urban Syst.* **2004**, *28*, 667–690. [[CrossRef](#)]
 38. Zhang, X.; Gao, F.; Liao, S.; Zhou, F.; Cai, G.; Li, S. Portraying Citizens' Occupations and Assessing Urban Occupation Mixture with Mobile Phone Data: A Novel Spatiotemporal Analytical Framework. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 392. [[CrossRef](#)]
 39. Deng, X.; Liu, Y.; Gao, F.; Liao, S.; Zhou, F.; Cai, G. Spatial Distribution and Mechanism of Urban Occupation Mixture in Guangzhou: An Optimized GeoDetector-Based Index to Compare Individual and Interactive Effects. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 659. [[CrossRef](#)]
 40. Li, S.; Huang, Z.; Gao, F.; Wang, F.; Lin, J.; Tan, Z. Evaluating the performance of LBSM data to estimate the gross domestic product of China at multiple scales: A comparison with NPP-VIIRS nighttime light data. *J. Clean. Prod.* **2021**, *328*, 129558. [[CrossRef](#)]