1. Introduction

Understanding the distribution of different functional regions (e.g. residential, business, and transportation areas, etc.) in a city is an important theme in urban studies (Antikainen 2005; Cranshaw et al. 2012; Yuan, Zheng, and Xie 2012). For many years, the exploration of the functional regions in urban areas has mainly relied on socio-demographics data and aggregate areas with high economic interaction (Karlsson and Olsson 2006). However, the process of updating such data is laborious and time-consuming (Wu et al. 2009), so the results are often stagnant and cannot reflect the dynamic property of local urban areas.

With the advent of the era of big data, two kinds of geospatial big data, movement-based data and activity-based survey data, have been widely used to understand our socioeconomic environments (Liu et al. 2015). To uncover the association between the spatio-temporal patterns of human movements and the functional regions, a branch of research attempted to utilize the movement-based data, including mobile phone data (Reades, Calabrese, and Ratti 2009; Toole et al. 2012), taxicab data (Qi et al. 2011; Liu et al. 2012), smart card data (Liu et al. 2009; Pelletier, Trépanier, and Morency 2011), and Wi-Fi data (Calabrese, Reades, and Ratti 2010). Unlike the movement-based data, the activity-based survey data are adopted to explore spatio-temporal activity patterns and then illustrate the functional regions of a city (Steiner 1994; Kockelman 1997). Both of the two kinds of data have their own limitations. The activity-based survey data require long-term observation, high time cost, and high financial cost (Wu et al. 2009; Toole et al. 2012). Moreover, the outcomes usually do not scale and only uncover a partial depiction of characteristics of functional regions (Cranshaw et al. 2012). For the movement-based data, they do not contain the travel demand information and cannot be used to depict the detailed characteristics of regions (e.g. the temporal variation of travel demands). Therefore, the cluster type is inferred by the empirical analysis and it is hard to distinguish the non-home/work activities (Jiang, Ferreira, and Gonzalez 2012). To overcome the limitations mentioned above, Yuan, Zheng, and Xie (2012) considered both point of interests (POIs) data of
a region and human mobility data to identify functional regions and inferred users’ travel demands by linking the movement with POI data. However, this method may fail because it is difficult to precisely match human movements with POIs. For instance, one individual goes to a shopping mall and then to a campus. If the shopping mall is not included in the POI data set, his/her movement will be linked to the education purpose instead of shopping.

Fortunately, with the proliferation of social media, such as Facebook, Twitter, Foursquare, and Flickr, millions of registered members are recording their surroundings and sharing their movement routes with friends via check-in (Noulas et al. 2011). Unlike cell phone and car trajectories data derived from GPS trackers, check-in data not only contain location information, but also record users’ travel demands. Although false check-in exists (e.g. one user who is not actually at airport pretends to create his/her check-in location at airport), Cheng et al. (2011) have announced a series of rules to filter out the false check-in records. Wu et al. (2014) also proposed five criteria to eliminate the fake check-ins and trips. As a consequence, check-in data have advantages to depict the intra-urban functional regions than the previous three kinds of data. Cranshaw et al. (2012) discovered the suburban areas, named Livehoods, from check-in data. Silva et al. (2012) utilized check-in data to measure the dynamics of eight cities in a large scale. However, they did not systematically examine the interdependence between the functional regions and human activity via check-in data.

In this article, we proposed a novel model via a low rank approximation (LRA) method to infer intra-urban functional regions from a data-set containing about 15 million check-in records during a year-long period in Shanghai, China. We found that a series of latent structures, entitled Latent Spatial-Temporal Activity Structure (LSTAS), can well represent the underlying associations between the functional regions and human activities. Additionally, we showed that the LSTAS had clear geographical meanings, such that LSTAS could serve as a feature indicator to urban activity structures. With the indicator function, LSTAS was then used to identify the territory of functional regions without any predefined functional region classification. The results show that our model well infers the functional regions of a city.

2. Materials and methods

2.1. Data and study area

In this study, we investigated about 15 million social media check-in records during a year-long period from September 2011 to September 2012 in Shanghai, the city with the largest population in China (Chan 2007). These records have been applied to model the intra-urban human mobility patterns (Wu et al. 2014) and are also used as a partial set to investigate inter-urban trips and spatial interactions (Liu et al. 2014). Considering both the computational efficiency and heterogeneous distribution of check-in data points, we chose the central part of the city (50 km × 35 km) as the study area and divided it into square grids (1 km × 1 km). As shown in Figure 1(a), the red lattices represent the study area with two airports (Pudong Airport and Hongqiao Airport) and two railway stations (Shanghai Railway Station and Shanghai South Railway Station). Moreover, we grouped the travel demands into six types: Home (H), Transportation (Tr), Work (W), Dining (D), Entertainment (E), and Others (O), since some check-in demand-tags signified the similar demand. As shown in Figure 1(b), one check-in record is geo-referenced as one point according to its location, where different colors of the points denote different activities.

2.2. Model framework

In this article, the proposed model was constructed mainly according to the following four steps (Figure 2).

1. Construct a region and temporally dependent travel demands matrix.
2. Adopt a LRA method to find the best estimation of the original matrix.
96

Y. ZHI eT AL.

(3) Extract the LSTAS from the LRA matrix.

(4) Adopt the K-means clustering algorithm to aggregate regions into several significant types according to several top LSTAS.

2.3. Matrix of region and temporally dependent travel demands

This article focuses on the underlying relationship between functional regions and human activity. We have three basic variables.

(1) \( D = \{d_1, d_2, \ldots, d_j\} \) denotes the domain of travel demands and \( j \) means total types of demands. For instance, in this study there are six types of demands: \( H, Tr, W, D, E, \) and \( O. \)

(2) \( T = \{t_1, t_2, \ldots, t_k\} \) is the collection of time intervals set and \( k \) means total temporal intervals in one day.

(3) \( G = \{g_1, g_2, \ldots, g_m\} \) represents the domain of the urban area set and \( m \) means the number of square lattices in the previous section. Each square lattice is called as subregion and can be marked with a certain number from left to right and from bottom to top.

The human activity demands are temporally relevant in nature. For instance, lunch and dinner are not the same demand from the view of semantics and will also cause different impacts on corresponding travel activities. Therefore, we used the Cartesian product of travel demands and time interval set to denote the temporally dependent travel demands (TTD) which can be expressed as:

\[
\text{TTD} = D \times T = \{td_1, td_2, \ldots, td_n\}, \quad n = j \times k
\]  

The subregion of \( G \) over TTD forms a region and TTD relation matrix is denoted by \( R\text{-TTD} \) or \( M \) for short.

\[
M = \begin{bmatrix} u_{g,td} \end{bmatrix}_{m \times n}
\]

where \( u_{g,td} \) is the intensity of the TTD in the subregion, the occurrence frequency of the travel demand \( d \) at time \( t \) conditioned within the subregion \( g. \)

2.4. Exploring the low-dimensional representation via LRA

In order to explore the lower dimensional representation, analyze the eigen-structures, and interpret the interdependence between functional regions and human activities, we adopted a LRA method. Because this method can explore the latent structure between two associated factors in the high-dimensional matrix, LRA has been widely applied in the fields of information retrieval (Deerwester et al. 1990), face recognition (Ma et al. 2012), and salient object detection (Peng et al. 2013). The principal component analysis (PCA) could also be applied to explore the eigen-structure. For example, Eagle et al. suggested that human movement patterns could be represented as a repeating structure, termed eigen-behaviors (Eagle and Pentland 2009). Researchers have connected eigen-behaviors with functional regions and used the term, eigen-place, to identify the recurring patterns of urban dynamics (Reades, Calabrese, and Ratti 2009; Calabrese, Reades, and Ratti 2010). However, PCA has some limitations. With the covariance matrix, the PCA method has to analyze the spatial and the temporal characteristics as two separate sets of features. By contrast, the LRA method could project the spatial or the temporal characteristics simultaneously into the same subspace directly which could show the connections between the functional regions and human activities.

The details of the LRA method are introduced as follows.

(1) Any matrix can be decomposed into two parts as (Candès et al. 2011):

\[
M = \hat{M} + N
\]

where \( \hat{M} \) is a LRA matrix of \( M \) and \( N \) is a perturbation matrix which indicates noise. The best rank-\( k \) estimate of \( \hat{M} \) can be denoted as:

\[
\text{minimize} \|M - \hat{M}\|_F
\]

s.t. rank (\( \hat{M} \)) \( \leq k \)

(4)
The matrix $\hat{M}$ can be decomposed into three matrices according to the Singular Value Decomposition (SVD) method:

$$\hat{M} = \hat{U} \hat{S} \hat{V}^T$$  \hspace{1cm} (5)

where $\hat{U}$ is a $m \times m$ unitary matrix (i.e. $\hat{U}^T \hat{U} = I_{m \times m}$); $\hat{V}$ is a $n \times n$ unitary matrix (i.e. $\hat{V}^T \hat{V} = I_{n \times n}$); $\hat{S}$ is a diagonal matrix constraining the singular values $\sigma$ of $\hat{M}$. The $\hat{M}$ can be rewritten as the sum of rank-1 matrices:

$$M \approx \hat{M} = \sum_{i=1}^{r} \hat{\sigma}_i \hat{u}_i \hat{v}_i^T$$  \hspace{1cm} (6)

where $\hat{\sigma}_i$ is larger than $\hat{\sigma}_{i+1}$, $i = 1, \ldots, r$; $r$ equals to the number of non-zero singular values (it also equals to the rank of $\hat{M}$).

To construct $\hat{M}$, the value of $r$ should be determined. It is suggested to use the Frobenius norm to evaluate the similarity between the $\hat{M}$ and the original matrix $M$ as:

$$E = \frac{\|M - \hat{M}\|_F}{\|M\|_F} \times 100\%$$  \hspace{1cm} (7)

where,

$$\|M\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} = \sqrt{\text{trace}(A^T A)} = \sqrt{\sum_{i=1}^{\min(m,n)} \sigma_i^2}$$

2.5. Uncovering the latent urban spatio-temporal activity structure

Based on Equations (6) and (7), the LRA method could be viewed as a transformation process for projecting a high-dimension matrix to a series of low-dimensional subspaces. These subspaces can be regarded as multiple intrinsic feature space embedded in the original high-dimension data space. Hence, we treated this kind of subspaces as the LSTAS. One LSTAS is viewed as the combination of the columns in the matrix $\hat{U}$ and the columns in the matrix $\hat{V}$, correspondingly. For example, the combination of $\hat{U}$’s first column and $\hat{V}$’s first column can be viewed as the first LSTAS. Considering that each column in matrix $\hat{U}$ is orthogonal to the other, one column in $\hat{V}$ denotes one unique feature among regions, called LSTAS for the spatial characteristics (LSTAS-SC). Similarly, each column in matrix $\hat{V}$ is also orthogonal to the other and one column in $\hat{V}$ can be viewed as one unique feature among TTD, called LSTAS for TTD characteristics (LSTAS-TTDC). As a result, one LSTAS can express the corresponding relationship between the spatial distribution pattern and TTD.

2.6. Identifying the territory of functional regions

This step aggregates similar formal regions in terms of LSTAS by performing the K-means clustering. In particular, each row of $\hat{U}$ indicates the characteristics distribution of original subregion in LSTAS and each row of $\hat{V}$ denotes the characteristics distribution of original TTD in LSTAS. As a consequence, this result allows us to simultaneously project the subregions and TTD into the same subspace for identifying the territory of functional regions, indicating that the $i, j$ cell of the $\hat{M}$ can be obtained by the dot product of the $i$ and $j$ row of matrix $\hat{V} \hat{S}$ as:

$$\hat{M} = \hat{U} \hat{S}^{1/2} (\hat{V} \hat{S}^{1/2})^T$$  \hspace{1cm} (8)

where $\hat{U} \hat{S}^{1/2}$ is the new coordinate for the original subregions and $\hat{V} \hat{S}^{1/2}$ is the new coordinate for the original TTD.

Then we used the K-means clustering algorithm to aggregate similar subregions and TTD. Therefore, one cluster indicates one functional region which has two kinds of characteristics: the spatial distribution and the function characteristics (the set of TTD in a cluster). K-means clustering algorithm is widely used and has a number of derived methods (Jain 2010), such as K-medoids, K-SVD (Aharon, Elad, and Bruckstein 2006). However, the K-means algorithm has two problems: the determination of the distance and the selection of the optimal number of clusters.

The study is based on trait representation by vectors in the feature space. Therefore, we used the cosine distance to measure the dissimilarity of those relationships which is a common method to measure the similarity between two vectors. Moreover, we should estimate the optimal number of clusters for the K-means algorithm. That is to say, it is necessary to find the optimal number of clusters for the inherent partition of the data. The most common approaches used to validate the clustering results include the following three aspects. First, in the method of external criteria, previous knowledge about the data was used as external reference. Second, the method of internal criteria is based on the quantity and intrinsic features of the data and no prior knowledge about the data is introduced. Third, in the method of relative criteria, the best clustering scheme is selected according to a pre-specified criterion without any statistical test (Brun et al. 2007). Therefore, we used three typical internal validation criteria: Dunn’s index (Dunn 1974), Silhouette index (Rousseeuw 1987), and Davies–Bouldin index (DBI) (Davies and Bouldin 1979) to determine the optimal number of clusters. Higher Dunn’s index or Silhouette index indicates a better clustering number, while the opposite holds for DBI.

3. Results and analysis

In this work, we set the number of demand types as $|D| = 6$ and the number of time intervals as $|T| = 24$ since 1-h intervals were adopted as the temporal unit.
to the temporal characteristics, some TTDs are highly correlated with others. For example, the frequencies for two demand types of Home and Entertainment are very high after 19:00, while the frequencies for Traffic and workplace are relatively high in the daytime. Such characteristics indicate that the matrix \( M \) contains redundant structures and a lower dimensional representation exists.

3.2. Lower dimensional representation of region-TTD matrix

To explore a lower dimensional representation \( \hat{M} \) and compare it with original matrix \( M \), according to Equation (7), it is required to determine an optimal \( r \). We determined the value of \( r \) considering the following three aspects: the distribution of singular values, the reconstruction accuracy compared with \( M \), and the reconstruction accuracy compared with the actual temporal variation of travel demands.

3.2.1. Distribution of singular values

We set normalized singular values (ratio to the maximum) as the vertical axis and index of singular with...
order from high value to low one (Figure 4). In Equation (6), the $\hat{M}$ can be written as the sum of $r$ rank-1 matrices. From the view of significance, most features can be represented by the first few principal structures. Less required singular values generally indicate more notable features. As shown in Figure 4, the first 30 singular values account for about 90% of energy in $M$. Such a distribution also tells us that the original matrix $M$ includes a large number of redundant information, which covers the most valuable and important information about the relationship between the functional region and human activity in the urban area.

### 3.2.2. Reconstruction accuracy compared with $M$

Using Equation (7), we can calculate the dissimilarity between the reconstruct matrix $\hat{M}$ and original matrix $M$ with different values of $r$, as shown in Figure 5. The horizontal axis represents the number of $r$ and the vertical axis represents the reconstruction error. When $r$ equals to 30, the reconstruction error is only 0.06.

### 3.2.3. Reconstruction accuracy compared with the real temporal variation of travel demands

The above two methods indicate that it is appropriate to set $r$ to be 30. Figure 6 plots the difference between the reconstruct TTD and original TTD when $r = 30$. The horizontal axis represents the time intervals in one day, and the vertical axis represents the frequency of travel demands. We used different colors to distinguish the travel demands and adopted different lines to discriminate the reconstructed temporal variation of demands from the original ones (dash line for the reconstructed demand and solid line for the original demand). Figure 6 illustrates that the original distribution of TTD can be well approximated when $r = 30$.

### 3.3. Interpreting the embedded region-activity subspace

In this research, we set $r = 30$, and there are 30 LSTAS to represent the relationship between the functional region and human activity. Each LSTAS consists of two kinds of structures: the spatial structure (one of the columns in $\hat{U}$) and the TTD structure (one of the columns of $\hat{V}$). To have a better understanding of LSTAS, we take the top six LSTAS as examples. Figure 7 illustrates the TTD characteristics of the top six LSTAS. Correspondingly, Figure 8 shows the spatial characteristics of the top six LSTAS.

In Figure 7, the vertical axis represents the travel demand and the horizontal axis implies the time from 0:00 to 23:00 during a day. The first three LSTAS present the significant compound temporal activity patterns, while the latter three LSTAS mainly show some specific single activity patterns. As shown in Figure 8, the green points are the 10 municipal commercial circles from a Collection of Policies for Development in Shanghai Commercial Sector 2010 (http://images.mofcom.gov.cn/accessory/201006/1277190604257.pdf). The green circles indicate the 10 municipal commercial circles. The first four LSTAS are mainly accumulated in the central part while the latter two LSTAS are in the discrete space out of the central part involving the transportations stations and the resident areas.

Represented by the first column in Figure 7, the first LSTAS-TTDC denotes the pattern of remarkable compound characteristics of two activities, including the activity for entertainment from noon to 22:00 and the activity for dining at noon and in the evening. Meanwhile, from the Figure 8(a), we can get spatial distribution insights that the first LSTAS-SC is mainly accumulated in the most of municipal commercial circles, especially in the Nanjing East Road and Huaihai Middle Road that are well known for the pedestrianized tourist street. Compared to the first LSTAS, the second one has the lower correlation to the entertainment and the higher correlation to three kinds of temporal activities: the activity for traffic in the morning and evening rush hours, the work in the daytime, and the activity for dining from noon to 22:00. Although the spatial distribution of the second LSTAS is also near to the commercial circles (as shown in Figure 8(b)), it mainly represents workspaces of the commercial circles rather than the entertainment parts. The latter three LSTAS just...
used the K-means clustering algorithm to aggregate similar subregions and TTD. Without any prior knowledge or pre-specified clustering numbers, we used three typical internal validation criteria: Dunn’s index, Silhouette index as shown in Figure 9 (a), and DBI as shown in Figure 9(b), to determine the optimal number of clusters. By analyzing the results of the three indexes, we suggest the optimal number of clusters is 5 in this study.

As a consequence, we find that five types of clusters have significant spatio-temporal characteristics in travel demands, as shown in Figures 10 and 11(a). Figure 10 shows the similarities between each TTD and the center of each cluster and Figure 11(a) shows the result of region aggregation. Cluster #1 presents a high degree of the entertainment and dining activities at noon and in the evening (Figure 10(a)), and covers all municipal

3.4. Inferring the functional regions

In order to demonstrate the importance of LSTAS, we applied the LSTAS to infer the functional regions. With the top 30 LSTAS as the reference, we simultaneously projected both the spatial and TTD structures in the same coordinate system. According to Equation (7), we used the K-means clustering algorithm to aggregate similar subregions and TTD. Without any prior knowledge or pre-specified clustering numbers, we used three typical internal validation criteria: Dunn’s index, Silhouette index as shown in Figure 9 (a), and DBI as shown in Figure 9(b), to determine the optimal number of clusters. By analyzing the results of the three indexes, we suggest the optimal number of clusters is 5 in this study.

As a consequence, we find that five types of clusters have significant spatio-temporal characteristics in travel demands, as shown in Figures 10 and 11(a). Figure 10 shows the similarities between each TTD and the center of each cluster and Figure 11(a) shows the result of region aggregation. Cluster #1 presents a high degree of the entertainment and dining activities at noon and in the evening (Figure 10(a)), and covers all municipal
commercial circles. The spatial distribution of Cluster #1 is shown in red in Figure 11(a). As a result, we suggest that Cluster #1 is commercial-dominant (CD). Unlike the CD, Cluster #2 is also highly correlated to the demand for transportation activities (Figure 10(b)), and contains some important transportation stations including airports and railway stations in Shanghai. Cluster #2, the blue area in Figure 11(a), is therefore viewed as the transportation-dominant (TD). Both Cluster #3 and Cluster #4 have a strong association with the demand for home activities. However, they are dissimilar in the features for other demands and the spatial distribution. As shown in Figures 10(c) and (d), all values of other demands in Cluster #3 are positive, indicating that the activities for the other demands in Cluster #3 are also active. By contrast, the values of other demands in Cluster #4 are negative, indicating that other demands are not associated with Cluster #4. From the view of spatial distribution, Cluster #3 is closer to the CD than Cluster #4, as shown in Figure 11(a). Therefore, we suggest that Cluster #3 is the developed residential-dominant (developed-RD) and that Cluster #4 is the developing residential-dominant (developing-RD). Since Cluster #5 shows a low degree of all the demands, as shown in Figure 10(e), we suggest that Cluster #5 is other-dominant (OD).

To verify the results, we first adopted the standard deviational ellipse method to determine the directional extent of the check-in data-set following the study by Liu et al. (2012). It is obvious that the urban development of Shanghai is confined by the coastal line as well as the Yangtze River. We then draw nine ellipses with the same center and the major axis increases from 1 to 17 km by an increment of 2 km, as shown in Figure 11(a). In each zone, the proportion of every cluster area is computed.
Figure 11(b) shows the distribution of various clusters within each zone. From inner zone to outer zone, the ratio of CD area and developed-RD area dramatically decreases while the developing-RD and OD significantly increase with the increase in the radius. Unlike the other three areas, the TD areas first increase and then decrease after the radius is longer than 9 km. It indicates a gradually declining activity intensity distribution from center to outer zones, and the corresponding change in region types from mostly CD and developed-RD areas to TD area, then outwards more developing-RD and OD areas, suggesting a concentric form of Shanghai.

The result is not only consistent with the official design, a Collection of Policies for Development in Shanghai Commercial Sector 2010, but also conforms to the findings of previous results obtained with the census and survey data (Li, Wu, and Gao 2007) or the taxi data (Liu et al. 2012). Moreover, this result proves that the LSTAS can be utilized to infer the functional regions without any predefined knowledge and provide a clear picture of the correlation between the groups of regions and different travel demands at different time of a day in the city.

Figure 9. Cluster validity indices of the K-means algorithm via (a) Dunn’s and Silhouette index, and (b) Davies-Bouldin index.

Figure 10. Similarities among each TTD and centers of the 5 Clusters. (a) Cluster #1; (b) Cluster #2; (c) Cluster #3; (d) Cluster #4. (e) Cluster #5.
systems. In the future, we will investigate more cities and further improve the applicability of the proposed model.

Funding
This work was supported by the Open Research Fund Program of Shenzhen Key Laboratory of Spatial Smart Sensing and Services (Shenzhen University), and sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry [grant number 50-20150618]; National Natural Science Foundation of China [grant numbers 41001220, 51378512, 41571397, and 41501442]. This work was also supported by the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (The Second Phase).

Notes on contributors
Ye Zhi is an assistant researcher of Road Traffic Safety Research Center of the Ministry of Public Security. His research interests include trajectory data mining, urban computing and geographic information engineering and applications.

Haifeng Li is currently an associate professor with the School of Geosciences and Info-Physics, Central South University. His current research interests include geo big data analysis, remote sensing, geographic information services, sparse representation, and machine learning.

Dashan Wang is an associate fellow and deputy director of the informatization of traffic management office at Road Traffic Safety Research Center of the Ministry of Public Security. His research interests include urban planning, Intelligent Transportation System (ITS) and traffic engineering and applications.

Min Deng is currently a professor with the School of Geosciences and Info-Physics, Central South University. His current research interests include geo big data analysis, temporal–spatial data mining.

Shaowen Wang is Professor of Geography and Geographic Information Science (Primary), Computer Science, Library and Information Science, and Urban and Regional Planning at the University of Illinois at Urbana-Champaign (UIUC), where he is named a Centennial Scholar. He is also Associate Director of the National Center for Supercomputing Applications (NCSA) for CyberGIS and Lead of NCSA’s Earth and Environment Theme, and Founding Director of the CyberGIS Center for Advanced Digital and Spatial Studies and the CyberInfrastructure and Geospatial Information Laboratory at UIUC. His current research interests include advanced cyberinfrastructure and cyberGIS, high-performance parallel and distributed computing, and spatial analysis and modeling.

Jing Gao was a postdoctoral researcher at the CyberGIS Center for Advance Digital and Spatial Studies, University of Illinois at Urbana-Champaign. Her current research interests include big geo-data analysis and machine learning.

Zhengyu Duan is currently an associate professor with the Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University. His current research interests include advanced cyberinfrastructure and cyberGIS, high-performance parallel and distributed computing, and spatial analysis and modeling.
research interests include intelligent transportation system, big data transportation.

Yu Liu is a professor at the Institute of Remote Sensing and Geographical Information Systems, Peking University. His research interests include GIScience and big geo-data mining.

ORCID
Yu Liu  http://orcid.org/0000-0002-0016-2902

References
Aharon, M., M. Elad, and A. Bruckstein. 2006. “K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation.” *IEEE Transactions on Signal Processing* 54 (11): 4311–4322.

Antikainen, J. 2005. “The Concept of Functional Urban Area.” *Informationen Zur Raumentwicklung* 7: 447–456.

Brun, M., C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. R. Dougherty. 2007. “Model-based Evaluation of Clustering Validation Measures.” *Pattern Recognition* 40 (3): 807–824.

Calabrese, F., J. Reades, and C. Ratti. 2010. “Eigenplaces: Segmenting Space through Digital Signatures.” *IEEE Pervasive Computing* 9 (1): 78–84.

Candès, E. J., X. Li, Y. Ma, and J. Wright. 2011. “Robust Principal Component Analysis?” *Journal of the ACM* 58 (3): 1–37.

Chan, K. W. 2007. “Misconceptions and Complexities in the Study of China's Cities: Definitions, Statistics, and Implications.” *Eurasian Geography and Economics* 48 (4): 383–412.

Cheng, Z., J. Caverlee, K. Lee, and D. Z. Sui. 2011. “Exploring Millions of Footprints in Location Sharing Services.” Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM), Barcelona, Catalonia, Spain, 81–88.

Cranshaw, J., R. Schwartz, J. I. Hong, and N. M. Sadeh. 2012. “The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City.” Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM), Dublin, Ireland, 58–65.

Davies, D. L., and D. W. Bouldin. 1979. “A Cluster Separation Measure.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1 (2): 224–227.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman. 1990. “Indexing by Latent Semantic Analysis.” *Journal of the American Society for Information Science* 41 (6): 391–407.

Dunn, J. C. 1974. “Well-separated Clusters and Optimal Fuzzy Partitions.” *Journal of Cybernetics* 4 (1): 95–104.

Eagle, N., and A. S. Pentland. 2009. “Eigenbehaviors: Identifying Structure in Routine.” *Behavioral Ecology and Sociobiology* 63 (7): 1057–1066.

Jain, A. K. 2010. “Data Clustering: 50 Years beyond K-Means.” *Pattern Recognition Letters* 31 (8): 651–666.

Jiang, S., J. Ferreira, Jr., and M. C. Gonzalez. 2012. “Discovering Urban Spatial-temporal Structure from Human Activity Patterns.” Proceedings of the ACM SIGKDD International Workshop on Urban Computing, China, Beijing, 95–102.

Karlsson, C., and M. Olsson. 2006. “The Identification of Functional Regions: Theory, Methods, and Applications.” *The Annals of Regional Science* 40 (1): 1–18.

Kockelman, K. M. 1997. “Travel Behavior as Function of Accessibility, Land Use Mixing, and Land Use Balance: Evidence from San Francisco Bay Area.” *Transportation Research Record: Journal of the Transportation Research Board* 1607 (1): 116–125.

Li, Z., F. Wu, and X. Gao. 2007. “Polarization of the Global City and Sociospatial Differentiation in Shanghai.” *Scientia Geographica Sinica* 27 (3): 304–311.

Liu, L., A. Hou, A. Biderman, C. Ratti, and J. Chen. 2009. “Understanding Individual and Collective Mobility Patterns from Smart Card Records: A Case Study in Shenzhen.” Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis Missouri, United States, 1–6.

Liu, Y., X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi, and L. Shi. 2015. “Social Sensing: A New Approach to Understanding Our Socio-economic Environments.” *Annals of the Association of American Geographers* 105 (3): 1–19.

Liu, Y., Z. Sui, C. Kang, and Y. Gao. 2014. “Uncovering Patterns of Inter-urban Trip and Spatial Interaction from Social Media Check-in Data.” *PLoS ONE* 9 (1): e86026.

Liu, Y., F. Wang, Y. Xiao, and S. Gao. 2012. “Urban Land Uses and Traffic 'Source-sink Areas': Evidence from GPS-enabled Taxi Data in Shanghai.” *Landscape and Urban Planning* 106 (1): 73–87.

Ma, X., H. Quang Luong, W. Philips, H. Song, and H. Cui. 2012. “Sparse Representation and Position Prior Based Face Hallucination upon Classified Over-complete Dictionaries.” *Signal Processing* 92 (9): 2066–2074.

Noulas, A., S. Scellato, C. Mascolo, and M. Pontil. 2011. “An Empirical Study of Geographic User Activity Patterns in Foursquare.” Proceedings of the 9th International Conference on Weblogs and Social Media (ICWSM), Barcelona, Catalonia, Spain, 11: 570–573.

Pelletier, M., M. Trépanier, and C. Morency. 2011. “Smart Card Data Use in Public Transit: A Literature Review.” *Transportation Research Part C: Emerging Technologies* 19 (4): 557–568.

Peng, H. W., B. Li, R. Ji, and W. Hu. 2013. “Salient Object Detection via Low-rank and Structured Sparse Matrix Decomposition.” Proceedings of Twenty-Seventh AAAI Conference on Artificial Intelligence, Washington, USA, pp 796–802.

Qi, G., X. Li, S. Li, G. Pan, Z. Wang, and D. Zhang. 2011. “Measuring Social Functions of City Regions from Large-scale Taxi Behaviors.” Proceedings of the 9th Annual IEEE International Conference on Pervasive Computing and Communications Workshops, Seattle, WA, 384–388.

Reades, J., F. Calabrese, and C. Ratti. 2009. “Eigenplaces: Analysing Cities Using the Space – Time Structure of the Mobile Phone Network.” *Environment and Planning B: Planning and Design* 36 (5): 824–836.

Rousseeuw, P. J. 1987. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” *Journal of Computational and Applied Mathematics* 20: 53–65.

Silva, T. H., P. O. Melo, J. M. Almeida, J. Salles, and A. A. Loureiro. 2012. “Visualizing the Invisible Image of Cities.” Proceedings of IEEE International Conference on Green Computing and Communications Workshops, Seattle, WA, 384–388.

Steiner, R. L. 1994. “Residential Density and Travel Patterns: Review of the Literature.” *Transportation Research Record* 1466: 37–43.

Toole, J. L., M. Ulm, M. C. González, and D. Bauer. 2012. “Inferring Land Use from Mobile Phone Activity.”
Proceedings of the ACM SIGKDD International Workshop on Urban Computing, Beijing, China, 1–8.
Wu, L., Y. Zhi, Z. Sui, and Y. Liu. 2014. “Intra-urban Human Mobility and Activity Transition: Evidence from Social Media Check-in Data.” *PLoS ONE* 9 (5): e97010.
Wu, S., X. Qiu, E. L. Usery, and L. Wang. 2009. “Using Geometrical, Textural, and Contextual Information of Land Parcels for Classification of Detailed Urban Land Use.” *Annals of the Association of American Geographers* 99 (1): 76–98.
Yuan, J., Y. Zheng, and X. Xie. 2012. “Discovering Regions of Different Functions in a City Using Human Mobility and POIs.” Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 186–194.