

Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model

Yao Yao, Xia Li, Xiaoping Liu, Penghua Liu, Zhaotang Liang, Jinbao Zhang & Ke Mai

To cite this article: Yao Yao, Xia Li, Xiaoping Liu, Penghua Liu, Zhaotang Liang, Jinbao Zhang & Ke Mai (2016): Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model, International Journal of Geographical Information Science, DOI: [10.1080/13658816.2016.1244608](https://doi.org/10.1080/13658816.2016.1244608)

To link to this article: <http://dx.doi.org/10.1080/13658816.2016.1244608>



Published online: 23 Oct 2016.



Submit your article to this journal [↗](#)









View related articles [↗](#)



View Crossmark data [↗](#)



Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model

Yao Yao ^a, Xia Li ^a, Xiaoping Liu^a, Penghua Liu ^b, Zhaotang Liang ^b,
Jinbao Zhang ^b and Ke Mai ^b

^aSchool of Geography and Planning, Guangdong Key Laboratory for Urbanization and Geo-simulation, Sun Yat-sen University, Guangzhou, China; ^bSchool of Geography and Planning, Sun Yat-sen University, Guangzhou, China

ABSTRACT

Urban land use information plays an essential role in a wide variety of urban planning and environmental monitoring processes. During the past few decades, with the rapid technological development of remote sensing (RS), geographic information systems (GIS) and geospatial big data, numerous methods have been developed to identify urban land use at a fine scale. Points-of-interest (POIs) have been widely used to extract information pertaining to urban land use types and functional zones. However, it is difficult to quantify the relationship between spatial distributions of POIs and regional land use types due to a lack of reliable models. Previous methods may ignore abundant spatial features that can be extracted from POIs. In this study, we establish an innovative framework that detects urban land use distributions at the scale of traffic analysis zones (TAZs) by integrating Baidu POIs and a Word2Vec model. This framework was implemented using a Google open-source model of a deep-learning language in 2013. First, data for the Pearl River Delta (PRD) are transformed into a TAZ-POI corpus using a greedy algorithm by considering the spatial distributions of TAZs and inner POIs. Then, high-dimensional characteristic vectors of POIs and TAZs are extracted using the Word2Vec model. Finally, to validate the reliability of the POI/TAZ vectors, we implement a K-Means-based clustering model to analyze correlations between the POI/TAZ vectors and deploy TAZ vectors to identify urban land use types using a random forest algorithm (RFA) model. Compared with some state-of-the-art probabilistic topic models (PTMs), the proposed method can efficiently obtain the highest accuracy (OA = 0.8728, kappa = 0.8399). Moreover, the results can be used to help urban planners to monitor dynamic urban land use and evaluate the impact of urban planning schemes.

ARTICLE HISTORY

Received 21 March 2016




Accepted 28 September 2016

KEYWORDS

Land use; Word2Vec; point-of-interest; deep learning; topic model

1. Introduction

Land use and land cover (LULC) are extremely important geospatial features (Ellis 2007, Arsanjani *et al.* 2013) and play important roles in many fields such as environmental monitoring, urban planning and government management (Williamson *et al.* 2010, Yin

CONTACT Xia Li  lixia@mail.sysu.edu.cn; Xiaoping Liu  liuxp3@mail.sysu.edu.cn  School of Geography and Planning, Guangdong Key Laboratory for Urbanization and Geo-simulation, Sun Yat-sen University, Guangzhou 510275, Guangdong province, China

© 2016 Informa UK Limited, trading as Taylor & Francis Group

et al. 2011, Hayashi and Roy 2013, La Rosa and Privitera 2013, Liu *et al.* 2014, Regan *et al.* 2015). In recent years, rapid urbanization and modern civilizations have generated diverse and sophisticated urban land use types, such as residential areas, education facilities and business districts, at different scales in China. Moreover, urban or regional land use patterns are not only determined by urban layouts specified by governments but also affected by people's lifestyles, which cannot be stereotyped and are continuously changing with further urban development (Yuan *et al.* 2012). Hence, sensing the spatial structures of urban land use quickly and identifying urban function structures accurately are of great significance in formulating effective policies and regulations for urban planning.

With the rapid development of remote sensing (RS) and computing technologies, RS images with high spatial resolution (HSR) have been widely used to extract and analyze LULC. Object-oriented classification (OOC) is one of the most popular methods for LULC analysis. There are many studies in which OOC is used to extract urban land use patterns from HSR images through the physical features of ground objects such as spectral, shape and texture features (Blaschke 2010, Dupuy *et al.* 2012, Hu and Wang 2013, Blaschke *et al.* 2014). However, without considering spatial relationships among ground objects, OOC methods can only recognize land cover information with low-level semantic features. To narrow the 'Semantic Gap' for land use classification, Bratasanu (2011) first introduced the concept of 'Scene Classification' applied to HSR RS images by building virtual words from spectral features (Bratasanu *et al.* 2011). Through building a Bag-of-Virtual-Words (BoVW) model, multifeature information, such as spectral, texture and SIFT, can be fused together, thus improving the LULC classification performance. Recent studies have mainly focused on how to integrate BoVW and probabilistic topic models (PTMs) to identify land use types with high-level semantic information, such as airports, residential districts and schools (Yang and Newsam 2010, Sun *et al.* 2012, Chen *et al.* 2013, Zhao *et al.* 2013, Tokarczyk *et al.* 2015, Zhang and Du 2015, Zhong *et al.* 2015, Wen *et al.* 2016). However, methods derived for pure RS images can only reflect the natural properties of ground objects. Land use types in a region often have strong relationships with inner social-economic activities, which is difficult to detect from pure RS imagery.

To address the above problems of using RS for urban applications, the concepts of 'social sensing' and 'urban computing' have been developed to monitor land use dynamics and further enable citywide computing to better serve residents and their cities (Zheng *et al.* 2014, Liu *et al.* 2015). Multisource geospatial big data, such as POIs, mobile phone signals, trajectories of floating cars and social media data, have been deployed for urban computing (Zheng *et al.* 2014, Liu *et al.* 2015). For example, POIs can effectively present regional functions as a result of their high accessibility from the Internet (Yuan *et al.* 2012, Zheng *et al.* 2014). In recent years, numerous in-depth discussions have been conducted to classify urban land use via POIs (Tian and Shen 2011, Yuan *et al.* 2012, Jiang *et al.* 2015). Through extracting various indicators based on frequencies of inner POI categories, regional land use types can be estimated using regression models or empirical models (e.g., the LUTE model) (Jiang *et al.* 2015, Long and Liu 2013, Rodrigues *et al.* 2012). Moreover, due to the complexity of urban land use, it is clearly unsatisfactory to analyze land use patterns via POI frequencies alone. Therefore, Yuan *et al.* (2012) proposed a POI-based semantic analysis model to Discover Regions of different Functions (DRoF) for the first time in 2012 (Yuan *et al.*

2012). By regarding regions as documents, functions of zones as topics, categories of POIs as metadata and human mobility patterns as words, DRoF constructs a Latent Dirichlet Allocation (LDA) model to mine regional senior semantic information and urban land use types, therein successfully improving the accuracy compared with pure image-based methods (Yuan *et al.* 2012, Zhang and Du 2015).

The abovementioned studies only take frequencies of POIs as the judgment of a region's land use types without considering inner spatial correlations, which may lead to most of the spatial information of POIs being wasted. If we regard regions as documents, urban land use types as topics and inside POIs as basic words, the spatial distributions of POIs in a region can then be considered as word sequences in a document. Thus, the relationships between POI-based sequences and land use types can be quantified through a continuous space language model (Schwenk 2007). By exploiting the potential of context relationships, information inside POIs can be better mined. Based on these assumptions and research results, an open-source deep-learning language model named Google Word2Vec is introduced in this study. The Word2Vec model can project words to high-dimensional vector spaces based on context relationships in documents (Mikolov *et al.* 2013a, 2013b, 2013c). Recent studies have proved that Word2Vec has a superior ability to analyze correlations between word pairs and recognize the sentiments of texts (Yu and Dredze 2014, Lilleberg *et al.* 2015, Mikolov *et al.* 2013c, Zhang *et al.* 2015).

In this study, we attempt to sense spatial distributions of urban land use at a fine local scales by integrating a Word2Vec model with POIs. We first construct a corpus based on traffic analysis zones (TAZs) and POIs and then quantify POI categories into characteristic vectors via a CBOW-based Word2Vec model. TAZ vectors can be estimated through the averaged summation of inside POI vectors. To validate the reliability of the produced characteristic vectors, we design several experiments using K-Means-based region aggregation and a random forest algorithm (RFA)-based land use classification based on TAZ vectors. A deep-learning language model is also incorporated to mine spatial distribution information of ground objects effectively and to identify urban land use at the local scale using POIs alone. Finally, this method is used to identify land use types in the Pearl River Delta (PRD), one of the most developed and largest urban agglomerations in Southern China. A comparison is further conducted between the proposed method and some state-of-the-art topic models.

2. Study area and data description

The study area includes the whole PRD, with a total area of 54,002km², which is one of the most important economic zones in China. As the political, cultural and economic center of South China, this region is characterized by the highest population density in the Guangdong Province. Urban structures in the PRD are of high complexity, therein containing a mix of a huge number of land use types, such as residential communities, shopping malls, clinical facilities and educational buildings. According to the governmental land use data from the *Bureau of Land and Resources of Guangdong Province*, the study area has 99,065 land parcels at the unit of the TAZ and 14 urban land use types. The spatial distribution and quantitative proportion of each land use type in the study area are illustrated in Figure 1 and Table 1, respectively.

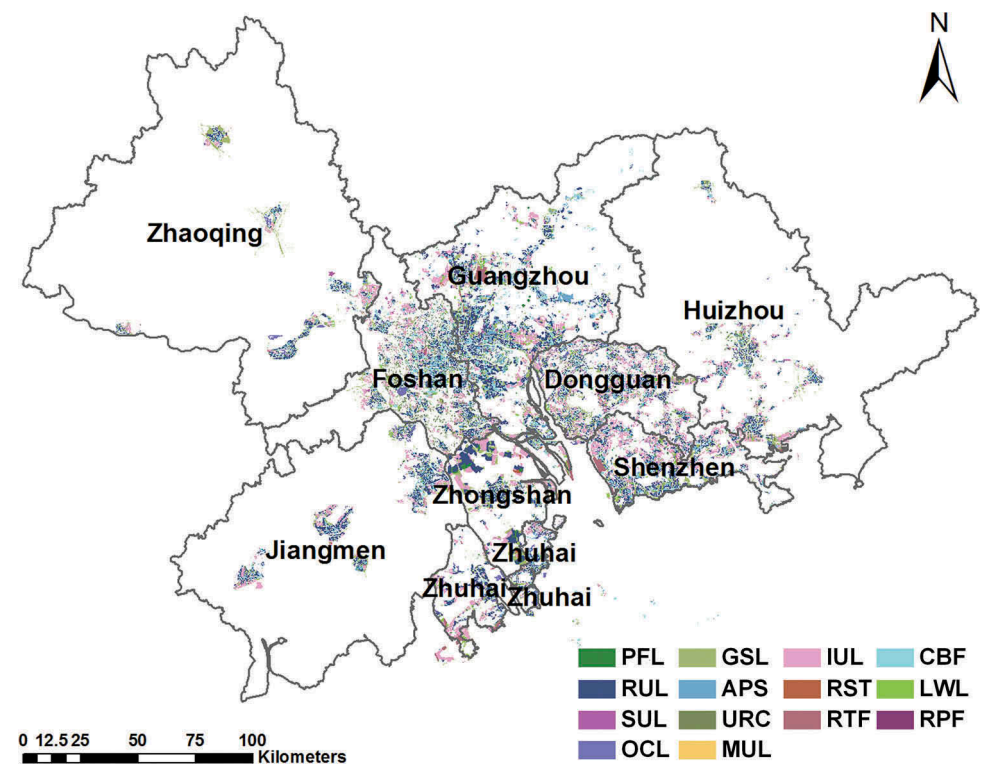


Figure 1. The Pearl River Delta in Guangdong Province, including Guangzhou, Shenzhen, Dongguan, Huizhou, Foshan, Zhongshan, Zhuhai, Zhaoqing and Jiangmen. The background data are urban land use data in 2010 within the unit of TAZ level (Data source: Bureau of Land and Resources of Guangdong Province). Land use types: PFL: Public facilities, GSL: Green space and square, IUL: Industrial land, CBF: Commercial and business facilities, RUL: Residential land, APS: Administration and public services, RST: Road, street and transportation, LWL: Logistics and warehouse, SUL: Special use land, URC: Urban and rural construction land, RTF: Regional traffic facilities, RPF: Regional public facilities, OCL: Other construction land, MUL: Mining use land.

Table 1. Quantitative proportion of each land use type in study area.

ID	Short code	Land use type	Proportions
1	PFL	Public facilities	3.97%
2	GSL	Green space and square	37.31%
3	IUL	Industrial land	8.19%
4	CBF	Commercial and business facilities	13.23%
5	RUL	Residential land	20.34%
6	APS	Administration and public services	7.52%
7	RST	Road, street and transportation	1.68%
8	LWL	Logistics and warehouse	1.03%
9	SUL	Special use land	0.40%
10	URC	Urban and rural construction land	5.29%
11	RTF	Regional traffic facilities	0.59%
12	RPF	Regional public facilities	0.02%
13	OCL	Other construction land	0.38%
14	MUL	Mining use land	0.05%

In this study, the POI dataset is fetched via application programming interfaces (APIs) provided by Baidu Map Services (<http://map.baidu.com>), which is the most widely used search engine and map service provider in China. We fetched 1,403,453 records of Baidu

POIs with multilevel categories. Along with the category-level upgrades, the descriptions of POIs are provided in greater detail, and the information is given in the form of independent Chinese phrase, which does not need word segmentation in advance. For example, the top-level and second-level categories of POIs whose final level is 'middle school' would be 'education' and secondary education, respectively. The description of 'secondary education' from the second-level categories is more specific than 'education' from the top-level category. In our POI dataset, there are 20 labels in the top-level category and more than 400 labels in the final level. Specifically, these 20 labels of POIs at the top level are corporations (COR), shopping (SHP), catering (CAT), life service (LIF), residential community (RSC), government (GOV), clinic facilities (CLF), roads (ROD), traffic facilities (TRA), automobile services (AMS), financial industry (FII), administrative landmark (ADL), education (EDU), hotel (HOT), entertainment (ENT), location annotation (LOC), business building (BUB), natural mountain (MOU), scenic spots (SCE) and green space (GRE). A division based on TAZs is set up as the same as the unit of governmental land use data (2010), which are offered by the Bureau of Land and Resources of Guangdong Province.

3. Methodology

The flowchart of the proposed method is illustrated in [Figure 2](#). The method attempts to transform Baidu POIs into high-dimensional vectors using the Google Word2Vec model and then apply the POI vectors to sensing urban land use and functional structures at the unit of the TAZ. The procedure includes three parts: (1) Baidu POIs and TAZs are used to build a TAZ-POI corpus. (2) Based on the TAZ-POI corpus, a characteristic vector for all POI categories is obtained using the Word2Vec model and a TAZ vector is computed by averaging the sum of POI vectors inside. (3) Urban land use types of TAZs are extracted via TAZ vectors computed from POI vectors. To demonstrate the effectiveness and potential of POI vectors, we will analyze the correlation among POIs and then evaluate the classification accuracy. Moreover, to test the reliability of the proposed method, comparisons with state-of-the-art topic models are presented, the details of which are given in the fourth section.

3.1. Building TAZ-POI corpus

A corpus often refers to a large and organized collection of well-sampled and processed texts in the field of natural language processing (NLP) (Ng *et al.* 1997). Specifically, a corpus consists of a number of documents, and each document contains several words. In a corpus, the sequential order of documents and words represents the context relationships, similar to natural language. Based on these concepts, our study area can be seen as a corpus, and each TAZ contained within can be regarded as a document, where POIs can be seen as words. To obtain a sufficient number of words, we select the descriptions from level-4 categories of Baidu POIs to compose documents.

Through the organized composition of a corpus, the context relationships of words are able to reveal spatial distribution attributes and positional relationships of POIs to some extent. To associate the words in an organized form and assign each document to realistic meanings, we propose a 'shortest path' method based on the idea of a 'Greedy

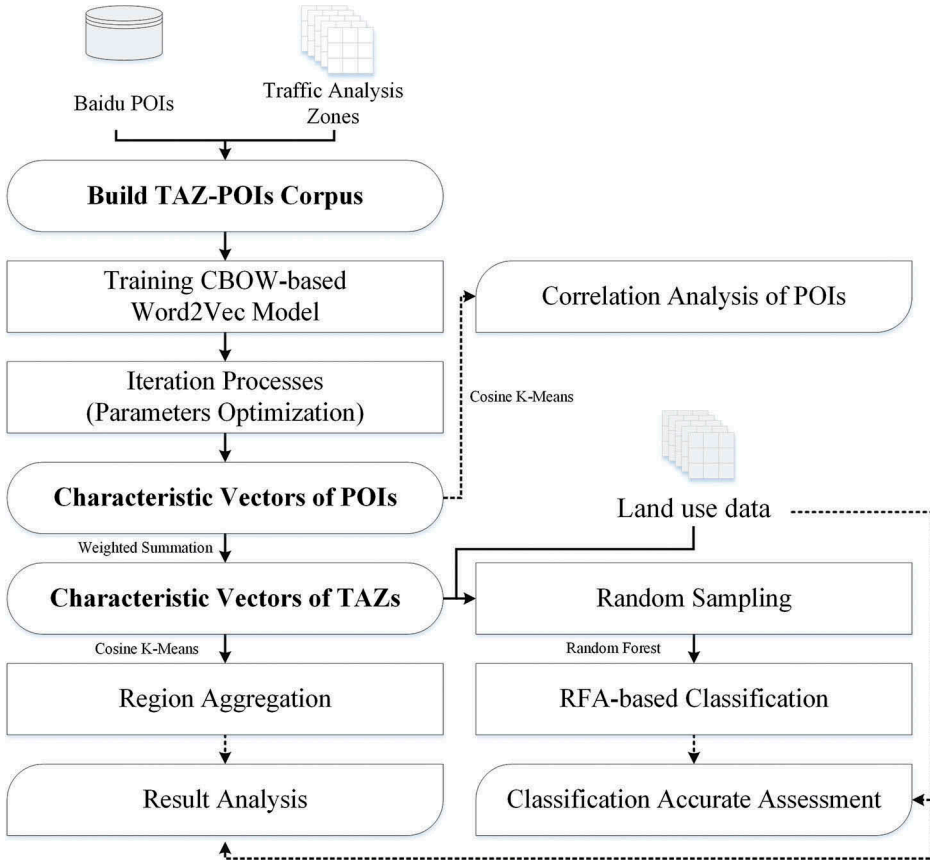


Figure 2. Workflow of urban land use identification using Google Word2Vec model.

Algorithm' to construct the TAZ-based documents. Using this method, POIs are connected to each other by their spatial relations. First, we find the shortest path in each TAZ that passes all the POIs and record the POIs in sequential order. In the next step, TAZ-based documents are constructed using the words according to the sequence of POIs. Suppose that there exists N POIs in a given TAZ. Denoting POIs as $P_1(x_{p1}, y_{p1}), P_2(x_{p2}, y_{p2}), \dots, P_N(x_{pn}, y_{pn})$ with their location coordinates (x, y) and using the subscript index i to represent the i -th POI $(i \in \{1, 2, 3, \dots, n\})$, the sequential order of POIs can be obtained as follows:

- (1) First, compute the Euclidean distance S of every POI pair $\langle P, P \rangle$ and select the furthest POIs (denoted as P_s and P_e) as the path's endpoints. Thus, after the operation is completed, the POIs' sequential order of the shortest path is $L = \{P_s, P_e\}$, and the wait-to-insert set of POIs is $A = \{P_x | x \in I \cap P_x \notin L\}$. Obviously, the path length l is the distance summation of each adjacent POI pair in L . Let t be the indicator of the current moment; the path length would now be $l(t) = S_{s,e}$.
- (2) The task of this step is to insert and update the shortest path. In this step, we exploit the idea of a greedy algorithm to ensure that each insertion of POIs is locally optimal in an attempt to find a global optimum for a shortest path.

Specifically, our objective is to select P_i from a set A and insert it into the correct location of the POI sequential order L. The ‘correct location’ criterion is that the path length is the shortest. An iteration loop was built for location selection. Suppose that, at the moment $t + 1$, P_i is inserted between P_m and P_n , which belong to order L; therefore, the new path length can be mathematically computed as $l_{(t+1)} = l_{(t)} - S_{m,n} + S_{m,i} + S_{n,i}$. Repeatedly inserting P_i into different locations in L and comparing the path length $l_{(t+1)}$, the iteration loop will end after all the location traversing is completed. Assuming that the insertion of P_i between P_m and P_n makes the path the shortest, the point P_i will then be fixed at this location, which means that the POI sequential order L would be updated to $\{..., P_m, P_i, P_n, ...\}$, and the wait-to-insert set A would become $A\{P_x | x \in I \cap P_x \notin L\}$.

- (3) The previous step is repeated until all the points in the wait-to-insert set A are placed into the sequential order L. Finally, the sequential order of POIs in TAZ-based documents can be successfully obtained.

Similarly, the arrangement of documents in the corpus is also based on the shortest path algorithm. Presuming that there exists M TAZs in the study areas, for legibility, we denote the TAZs as $T_1(x_{T1}, y_{T1}), T_2(x_{T2}, y_{T2}), \dots, T_m(x_{Tm}, y_{Tm})$, where (x_{Tj}, y_{Tj}) denote the centroid coordinates of the j – th TAZ. Applying the above greedy algorithm to the TAZ centroid coordinates, the sequential order of TAZ-based documents in the corpus can be obtained. Because the acquisition of the sequential order of POIs in TAZ-based documents and the sequential order of TAZ-based documents in the corpus are both achieved, our Word2Vec training corpus, namely, TAZ-POIs, is thus successfully constructed.

3.2. Computing characteristic vector of each POI category

Word2Vec, a model made open source by Google in 2013 (<https://code.google.com/p/word2vec/>), is a deep-learning tool for transforming words into high-dimensional spatial vectors (Mikolov *et al.* 2013a). By building a Neural Network Language Model (NNLM) using an input training corpus, the Word2Vec model can map each word to characteristic real-valued vectors via its contextual content (Mikolov *et al.* 2013a, 2013b, 2013c). In this study, we assume that the size of the TAZ-POI corpus is T, the final-level categories of the t – th POI in the corpus are w_t , and the sampling window size of the context centered on the t – th POI is denoted as c ; therefore, the maximum likelihood function of the NNLM can be estimated using Equation (1) (Mikolov *et al.* 2013a, 2013b, 2013c, Yu and Dredze 2014):

$$l(\theta) = \log L(\theta) = \frac{1}{T} \sum_{t=1}^T \log p(w_t | w_{t-c}^{t+c}). \quad (1)$$

In Equation (1), w_{t-c}^{t+c} denotes a set of words at the center of w_t whose context sampling window size is c , where the center w_t is excluded from the sample set. The Word2Vec model provides two mathematical models for solving Equation (1) such as Continuous Bag-of-Words (CBOW) and Skip-Gram. Compared with the stochastic sampling of word pairs in the Skip-Gram model, the continuous input and training process of CBOW can better reflect the context relationships characterizing words (Yu and Dredze 2014). Therefore, in this paper, a CBOW-based

Word2Vec model is adopted to extract POI vectors. CBOW defines $p(w_t|w_{t-c}^{t+c})$ as follows (Mikolov *et al.* 2013c):

$$p(w_t|w_{t-c}^{t+c}) = \frac{\exp(-E(w_t, w_{t-c}^{t+c}))}{\sum_{i=1}^T \exp(-E(w_i, w_{t-c}^{t+c}))}. \quad (2)$$

E is an energy function, where $E(w_i, w_j) = -(w_i \cdot w_j)$. Equation (2) shows the occurrence probability of the t – th POI when the current context is c . During the iterative training process of the CBOW-based Word2Vec model, optimal POI vectors can be estimated by integrating Hoffman trees with a stochastic gradient descent (SGD) algorithm.

Theoretically, characteristic vectors of similar POI categories have approximately the same angles and orientations in high-dimensional spaces. This means that the similarity of POI pairs has an inverse correlation with the angle between related POI vectors. To demonstrate the effectiveness and reliability of POI vectors obtained by Google Word2Vec, Cosine-distance-based K-Means is adopted for POI vector clustering. The cosine distance between the i – th and j – th POIs can be determined mathematically by the following equation:

$$D(P_i, P_j) = 1 - \cos(\theta) = 1 - \frac{\sum_{k=1}^K v_{ik} \cdot v_{jk}}{\sqrt{\sum_{k=1}^K v_{ik}^2} \cdot \sqrt{\sum_{k=1}^K v_{jk}^2}}. \quad (3)$$

In Equation (3), $\cos(\theta)$ is the cosine distance from the i – th POI to the j – th POI in the range of $[-1, 1]$, and v_i and v_j are the i – th and j – th POI vectors with dimensions of K . After clustering the POI vectors, we estimate the reliability of the POI vectors by comparing the clustering results of low-level POI categories with those of high-level categories. It should be noted that the performance of K-Means depends on the initial cluster centers; therefore, we introduce the modified iterated anomalous pattern (AP) method to obtain initial cluster centers (Rutkowski 2007). The average silhouette value can be used as a criterion to determine the number of POI clusters K (Rousseeuw 1987, Yuan *et al.* 2012). The silhouette value of the i – th POI in the dataset is denoted by $s(i)$ in the range of -1 to 1 . A value of $s(i)$ approaching 1 indicates that the POI is clustered appropriately and far from other clusters, $s(i)$ close to 0 indicates that the POI is on the border of two natural clusters, and $s(i)$ close to -1 indicates that the POI would be more appropriate if it was clustered in its neighboring cluster. Therefore, the average silhouette value \bar{s} of all the POIs can reflect the reliability of the clustering results. Consequently, the optimal cluster number K can be estimated by the average silhouette \bar{s} near 1 , and the best clustering results will be applied to analyze urban structures in the following sections.

3.3. Extracting urban land use type for each TAZ

3.3.1. Region aggregation via POI vectors

Supported by the CBOW-based Word2Vec model, we can obtain characteristic vectors of all POI categories. Recent studies on text semantic orientation analysis have shown that

document vectors computed using the weighted average of all words in texts can effectively portray a document (Xue *et al.* 2014, Zhang *et al.* 2015). Therefore, in our study, we obtained the TAZ vectors in adopting this method by averaging inside POI vectors with weightings. Suppose that there are N POIs $(P_{i,1}, P_{i,2}, \dots, P_{i,n})$ in the i -th TAZ. The TAZ vectors can then be specified mathematically by Equation (4):

$$\text{vectors_of_TAZ}_i = \frac{\sum_{k=1}^N \text{vectors_of_type}(P_{i,k})}{N}. \quad (4)$$

In Equation (4), $\text{vectors_of_type}(P_{i,k})$ denotes the characteristic vectors of the k -th POI type in the i -th TAZ, and TAZ vectors are estimated by a weighted average of inside POI vectors. To estimate the relationship between TAZ vectors and real land use, the proposed Cosine-distance-based K-Means method is adopted to cluster TAZs, and the clustering results are compared with governmental land use data in terms of the functions of the TAZs.

3.3.2. RFA-based supervised classification

Furthermore, a supervised classification method is adopted to predict the land use type of all TAZs, which is used for comparison with the proposed method. Previous studies have indicated that support vector machines (SVMs) can perform well in solving high-dimensional and nonlinear classification problems (Mountrakis *et al.* 2011, Huang and Zhang 2013, Mordelet and Vert 2014). However, SVMs are highly sensitive to initial parameters, therein causing uncertainty during the training and predicting processes (Liu *et al.* 2003). RFAs are a state-of-the-art nonlinear and nonparametric classification model that can address the problem of the correlation among variables and overfitting in the field of high-dimensional and nonlinear classification (Breiman 2001, Biau 2012, Palczewska *et al.* 2014).

Assume that $X_{ij}(i \in [1, M], j \in [1, N])$ and $Y_i(i \in [1, K])$ are the characteristic vectors and land use types of a TAZ, where M is the total number of TAZs in the training dataset, N denotes the dimensions of the TAZ vectors and K is the total count of each type of urban land use. Using the Bagging method, the RFA randomly takes $m \times n$ -dimensional ($m \ll M, n \ll N$) samples depending on the dimensions of the training dataset. C trees are trained on these selected sample data without pruning operations. In the RFA method, variables are not totally taken to split nodes; instead, only parts of the variables are randomly selected to make decisions. Using this approach, the correlation of each decision tree can be reduced, thus enhancing each decision tree's classification accuracy. Furthermore, the generalization error of the RFA can be calculated by averaging the errors of the decision trees via out-of-bag (OOB) estimation after the training process. In previous studies, it has been indicated that the fitting model using an RFA overcomes the multiple correlative problems among spatial variables, especially in higher dimensional fitting situations (Fakhraei *et al.* 2014). Finally, the land use types of TAZs can be identified as the maximum voting type through the use of the decision trees of a random forest.

In this study, by randomly selecting part of the land use data and related TAZ vectors as the training dataset, an RFA-based model is applied to land use type sensing at the unit of the TAZ in the study area. To guarantee the reliability of the classification results,

the classification experiments are repeated 100 times, and the average of the overall accuracies (OAs) and kappa coefficients are used as the accuracy criteria. Moreover, several state-of-the-art topic models (TF-IDF, LDA and pLSA), which can properly extract topic features and train RFA classifiers, will be adopted to form comparisons with our proposed method.

4. Results and discussion

Our research team built a software application and realized all models proposed in Section 3 using C++ on Windows 8.1(x64). Several open-source C/C++ libraries, such as Google Word2Vec (<https://code.google.com/p/word2vec/>), GDAL (<http://www.cgal.org/>) and Shark (<http://image.diku.dk/shark/>) libraries, were applied to this project for extracting POI vectors and identifying urban land use types at the unit of the TAZ. The source codes of the LDA-based topic model are provided by Princeton University (<http://www.cs.princeton.edu/~blei/topicmodeling.html>) (Blei *et al.* 2003). The related application and results (POI vector data) can be downloaded from our GeoSOS website (<http://www.geosimulation.cn/>).

4.1. Implementation and results

4.1.1. POI vector extraction and correlation analysis

In the study area, 944,698 Baidu POIs are evenly distributed among 37,584 TAZs, and TAZs that contain no POIs are excluded from the analysis. Final-level categories of POIs, totaling 419, are adopted to build the TAZ-POI corpus based on the greedy algorithm. During the process of building the POI-based CBOW model, we set the dimension of the output word vectors to 200 and the sample window size to 5; the number of iterations is set to 20 and the other parameters are set to the recommended values. By transforming the spatial distributions of POIs into the context relationships of visual words, 419 characteristic vectors of final-level POI categories can be estimated by the Google Word2Vec model.

The cosine distance is used to indicate the correlation between different categories of POIs. The K-Means-based clustering result of POI vectors can effectively quantify the relationship between different POI categories. The results of the POI vectors and the Cosine-distance matrix are freely available on the Internet (<http://pan.baidu.com/s/1gene5IN>) for download. For example, the result suggests that POIs with the category of 'real estate' have strong correlations with POIs of 'sales office', 'residential community' and 'parking area' and the 'ATM' POI is highly related to bank POIs. As illustrated in Figure 3, we find that the average silhouette value is maximized when $K = 2, 4$ and 10 . To analyze the correlations between POI clusters and POI categories, we calculate the proportion of different types of POIs in each cluster using top-level POI categories, as demonstrated in Table 2.

When $K = 2$, cluster **C1** mainly contains virtual POIs, which are used as place name tags only, such as road labels (ROD), location annotation (LOC) and natural mountain (MOU); the components of cluster **C2** represent a vast majority of actual ground objects that cover most urban land uses. When $K = 4$, we can observe that **C1** represents virtual tag POIs, **C2** refers to government (GOV) and residential community (RSC) and **C3** is a

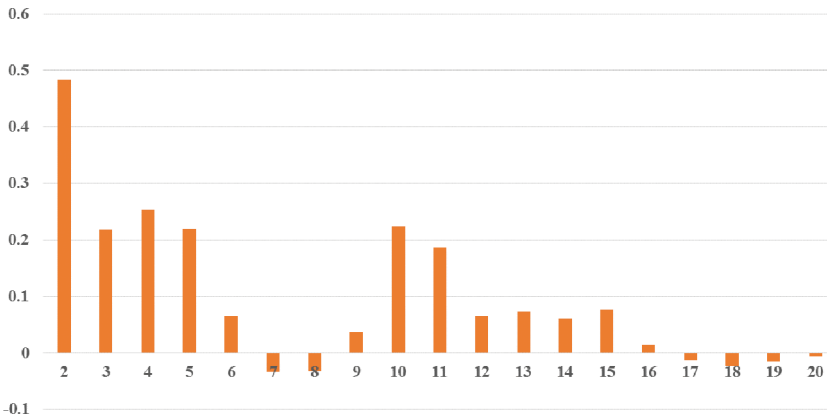


Figure 3. Change of average Silhouette value (y-axis) of clustering results (POI vectors) with increases of K value (x-axis).

mix of shopping (SHP) and life service (LIF). **C4** mainly covers commercial and business facilities. Apparently, with increasing cluster number K, POIs with similar functions become concentrated in one cluster, and the heterogeneities among homogenous POIs gradually become distinctive. When K reaches 10, it is obvious that the function of each cluster has been identified by one or two dominating POI categories such as government (**C2**), shopping (**C3**), clinic (**C5**), business/corporation (**C6**) and entertainment (**C10**). The clustering results of the POI vectors indicate that urban micro entities with homogeneous functions have similar spatial distribution features, and certain levels of spatial correlations also exist between heterogeneous entities. Therefore, we have reason to believe that POI vectors are effectively able to quantify spatial semantic features of POIs, and the results of multi-scale clustering indicate that POI vectors can be used to reveal the correlations between actual regional land use types and the spatial structures of inside POIs.

4.1.2. Identifying urban land use

(1) K means-based region aggregation

Based on Equation (3), TAZ vectors can be obtained using a weighted average of inside POI vectors. As illustrated in Figure 4, average silhouette values fluctuate strongly with increasing cluster number K. This indicates that the cluster number K can be used to measure the functional heterogeneity between different regions at the scale of a TAZ. In other words, by specifying the proper value of the cluster number K, two TAZs can be separated into different clusters in terms of the similarity degree of urban land use patterns. Figure 4 shows that the top-3 average silhouette values of the TAZ vector clustering results are 2, 3 and 4, which means the TAZ vectors are clustered appropriately when taking these values. Therefore, we use these values of K to analyze land use types. Based on governmental land use data and HSR RS imagery, Figure 5 demonstrates the proportions of different land use regions in each cluster with varying K. Figure 6 maps the clustering results for Guangzhou, Shenzhen, Zhongshan and Zhuhai. The region cluster results with different values of K are annotated as follows:

K = 2:

Table 2. Clustering result of POI vectors by using Cosine-based K means. The values which are greater than 50% are shown in boldface.

K	Class	COR	SHR	CAT	LIF	RSC	GOV	CLF	ROD	TRA	AMS	FII	ADL	EDU	HOT	ENT	LOC	BUB	MOU	SCE	GRE
2	C1	8.1%	33.3%	20.3%	33.3%	0.0%	14.3%	4.3%	100.0%	50.0%	62.5%	5.4%	16.7%	25.0%	0.0%	15.8%	100.0%	33.3%	100.0%	0.0%	50.0%
	C2	91.9%	66.7%	79.7%	66.7%	100.0%	85.7%	95.7%	0.0%	50.0%	37.5%	94.6%	83.3%	75.0%	100.0%	84.2%	0.0%	66.7%	0.0%	100.0%	50.0%
4	C1	28.6%	6.7%	6.8%	0.0%	0.0%	7.1%	0.0%	87.5%	58.8%	50.0%	2.7%	33.3%	0.0%	0.0%	13.2%	100.0%	33.3%	100.0%	0.0%	50.0%
	C2	21.4%	0.0%	0.0%	0.0%	75.0%	81.0%	43.5%	0.0%	2.9%	0.0%	16.2%	50.0%	50.0%	7.1%	18.4%	0.0%	0.0%	0.0%	0.0%	0.0%
	C3	19.0%	66.7%	25.4%	63.0%	0.0%	0.0%	8.7%	0.0%	0.0%	12.5%	5.4%	0.0%	0.0%	7.1%	5.3%	0.0%	0.0%	0.0%	0.0%	0.0%
	C4	31.0%	26.7%	67.8%	37.0%	25.0%	11.9%	47.8%	12.5%	38.2%	37.5%	75.7%	16.7%	50.0%	85.7%	63.2%	0.0%	66.7%	0.0%	75.0%	50.0%
10	C1	11.9%	13.3%	15.3%	14.8%	0.0%	4.8%	0.0%	37.5%	2.9%	0.0%	2.7%	16.7%	0.0%	0.0%	2.6%	33.3%	0.0%	100.0%	0.0%	0.0%
	C2	9.5%	0.0%	0.0%	0.0%	0.0%	66.7%	8.7%	0.0%	2.9%	0.0%	0.0%	50.0%	18.8%	0.0%	13.2%	0.0%	0.0%	0.0%	0.0%	0.0%
	C3	14.3%	60.0%	10.2%	55.6%	0.0%	0.0%	4.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.6%	0.0%	0.0%	0.0%	0.0%	0.0%
	C4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	8.3%	0.0%
	C5	2.4%	0.0%	3.4%	0.0%	0.0%	0.0%	82.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	8.3%	0.0%
10	C6	42.9%	0.0%	3.4%	3.7%	25.0%	14.3%	4.3%	0.0%	0.0%	12.5%	70.3%	0.0%	0.0%	0.0%	5.3%	0.0%	80.3%	0.0%	16.7%	0.0%
	C7	7.1%	0.0%	3.4%	0.0%	0.0%	4.8%	0.0%	25.0%	20.6%	62.5%	2.7%	0.0%	0.0%	7.1%	2.6%	33.3%	0.0%	0.0%	0.0%	0.0%
	C8	9.5%	10.0%	5.1%	0.0%	25.0%	0.0%	0.0%	37.5%	50.0%	0.0%	2.7%	16.7%	6.3%	14.3%	23.7%	33.3%	19.7%	0.0%	25.0%	50.0%
	C9	2.4%	0.0%	0.0%	14.8%	50.0%	7.1%	0.0%	0.0%	8.8%	25.0%	8.1%	16.7%	68.8%	0.0%	21.1%	0.0%	0.0%	0.0%	16.7%	0.0%
	C10	0.0%	16.7%	59.3%	11.1%	0.0%	2.4%	0.0%	0.0%	14.7%	0.0%	13.5%	0.0%	6.3%	78.6%	28.9%	0.0%	0.0%	0.0%	25.0%	50.0%

Types of POIs: COR = Corporation, SHP = Shopping, CAT = Catering, LIF = Life Service, RSC = Residential community, GOV = Government, CLF = Clinic facility, ROD = Road, TRA = Traffic facility, AMS = Automobile service, FII = Financial industry, ADL = Administrative landmark, EDU = Education, HOT = Hotel, ENT = Entertainment, LOC = Location annotation, BUB = Business building, MOU = Natural Mountain, SCE = Scenic spot, GRE = Green space

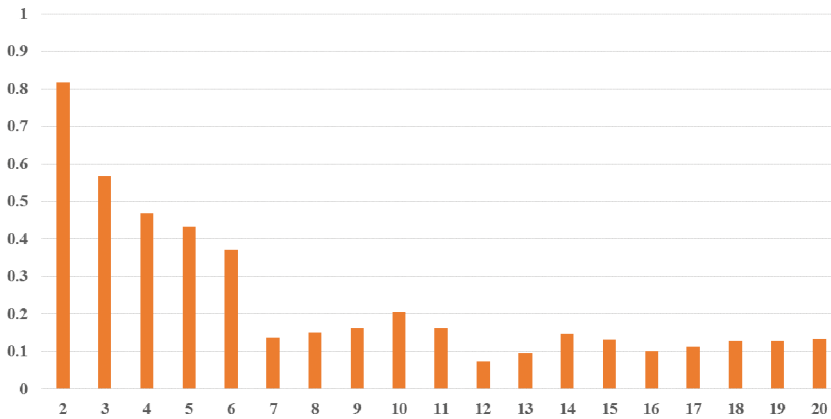


Figure 4. Change of average Silhouette value (y-axis) of clustering results (TAZ vectors) with increases of K value (x-axis).

Residential and shopping areas [K = 2, C1]: In this cluster region, more than 50% of the land use types are related to residents' livelihoods, such as residential areas, business and commercial districts, administrative areas and public services. We can observe from Figure 6 that this cluster region is mainly located in residential areas and old city centers, which are heavily mixed areas of residential blocks, business districts, roads and public service facilities.

Working and construction areas [K = 2, C2]: In this cluster region, major categories of land use type are regional public/traffic facilities, industrial and construction lands. In addition, Figure 6 demonstrates that this cluster region largely covers the emerging business center, new urban districts and factory yards, indicating that the corresponding demographic organization includes fewer permanent residents and a greater working and floating population.

K = 3 and K = 4:

Developed residential areas [K = 3, C1] and [K = 4, C1]: Compared with the residential proportion obtained using K = 2, this cluster eliminates some regions whose dominating functions are business and entertainment. For example, west of Ersha Island is a luxury residential district, and the eastern area is densely populated with entertainment and public facilities, which can be easily recognized when K is greater than 2.

Public facilities and construction areas [K = 3, C2] and [K = 4, C2]: Compared with the situation when K = 2, we have reason to believe that this cluster is stripped from 'working and construction areas' [K = 2, C2]. In this cluster region, the highest proportions of land use types are denoted as 'regional public facility' and 'urban and rural construction land', which originally belonged to 'working and construction areas' when K = 2. In Figure 6, east of Ersha Island (Central Pearl River), the center square of the Guangzhou Higher Education Mega Center (southeast of Guangzhou) and construction sites around the cities can be clearly seen in this region, and their corresponding city functions fall within the scope of public facilities and construction areas.



Figure 5. Proportion of land use types in different clusters ($K = 2, 3, 4$) Land use types: Public facilities (PFL), Green space and square (GSL), Industrial land (IUL), Commercial and business facilities (CBF), Residential land (RUL), Administration and public services (APS), Road street and transportation (RST), Logistics and warehouse (LWL), Special use land (SUL), Urban and rural construction land (URC), Regional traffic facilities (RTF), Regional public facilities (RPF), Other construction land (OCL), Mining use land (MUL).

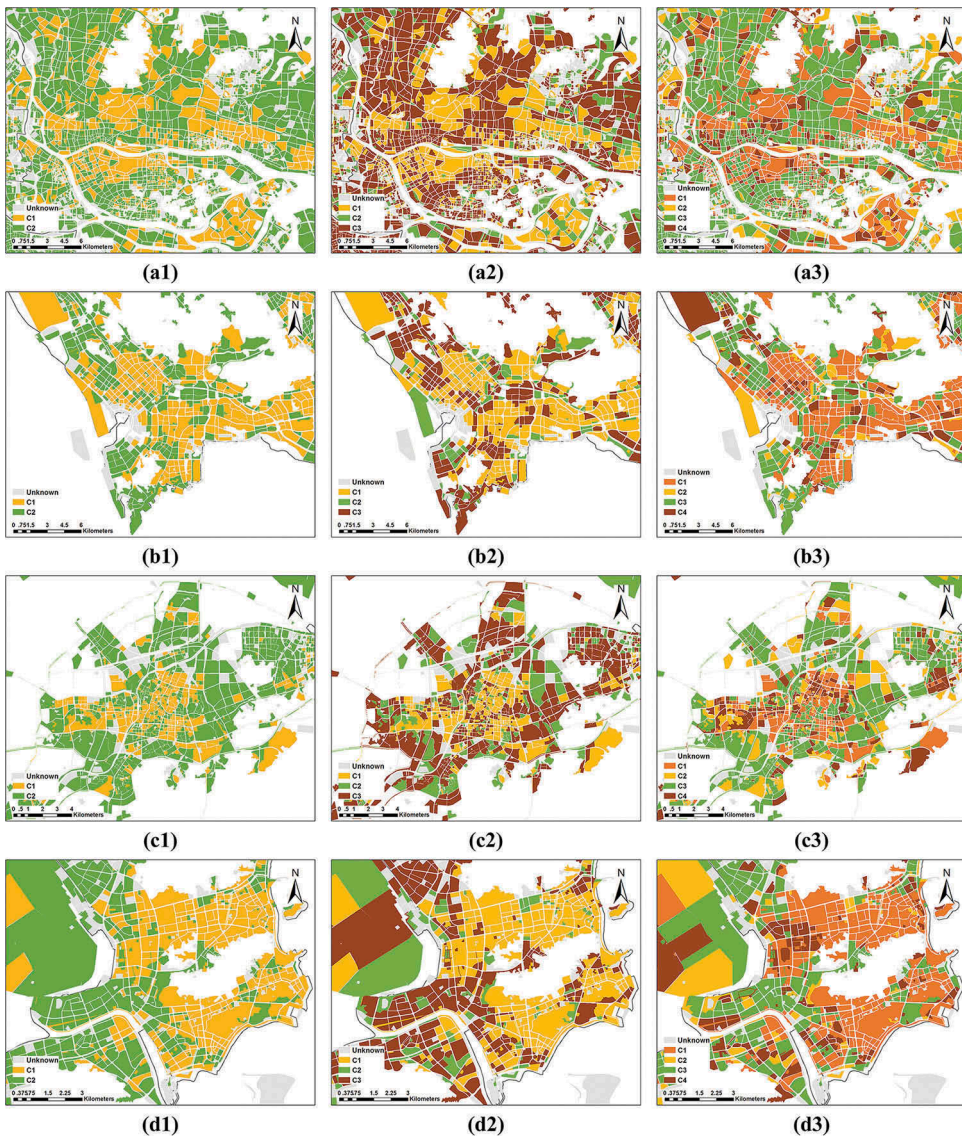


Figure 6. K-Means-based clustering results of TAZ vectors via TAZ vectors in study area (Same color settings of clusters as Figure 5). (a1)~(a3) Downtown areas of Guangzhou ($K = 2,3,4$), (b1)~(b3) Nanshan districts of Shenzhen ($K = 2,3,4$), (c1)~(c3) Downtown areas of Zhongshan ($K = 2,3,4$), (d1)~(d3) Downtown areas of Zhuhai ($K = 2,3,4$).

Working areas [$K = 3, C3$] and [$K = 4, C3$]: Compared with [$K = 2, C2$], this cluster region eliminates parts of public traffic areas, all the public facilities and all the construction areas while including most industrial areas, business centers and logistics and warehouses. As depicted in Figure 6, the business centers, Central Business Districts and emerging commercial districts, such as Beijing road (a famous shopping center), Pearl River Metro (CBD) of Guangzhou and Gongbei (a shopping center) of Zhuhai, are located in this cluster region.

Emerging residential areas [K = 4, C4]: This cluster region exhibits a weak correlation with construction and regional facilities because it provides a balanced land use configuration for residents' livelihood, therein including, for example, green spaces, living services and residential areas. Scrutinizing this cluster, we can find that this cluster is extracted from [K = 3, C2]. By carefully comparing it with HSR RS imagery, we finally classify this region cluster as emerging residential areas.

(2) RFA-based classification

Through the above K-Means-based region aggregation, it has been shown that TAZ vectors have strong correlations with regional land use types. In this section, we employ some RFA-based methods to evaluate the performance of the proposed method. Some governmental land use data were selected as the training samples. Several state-of-the-art semantic topic models, such as term frequency-inverse document frequency (TF-IDF) (Aizawa 2003, Yuan *et al.* 2012), probabilistic latent semantic analysis (pLSA) (Bosch *et al.* 2006) and LDA (Blei *et al.* 2003, Li *et al.* 2010), were also used for the classification for comparison with our proposed method. In this study, the settings of the semantic language models and parameters were configured as follows:

- **Proposed Method:** 50% of land use data and related TAZ vectors were randomly chosen to train an RFA-based classifier.
- **TF-IDF:** First, the TF-IDF values of the TAZs were calculated using the distribution frequencies of inside POIs. Then, 50% of the TAZs and their TF-IDF features were randomly selected to build a RFA-based classifier.
- **pLSA:** The model parameters of vector dimensions and number of iterations were set to 200 and 100, respectively. Similar to TF-IDF, 50% of the TAZs and their pLSA-based semantic vectors were randomly selected to build an RFA-based classifier.
- **LDA:** The number of topics and alpha were separately set as 200 and 0.025, and the maximum number of iterations and minimum errors between iterations were tuned to 100 and 0.0001, respectively. A total of 50% of the TAZs and their Dirichlet topic distribution probabilities were randomly selected to build an RFA-based classifier.

The above RFA-based classifiers, which were used to predict land use types of all TAZs in the study area, were implemented using the recommended number (100) of decision trees provided by the Shark v3.0 library (<http://image.diku.dk/shark/>). For cross-validation, the percentage of the training dataset and out-of-bag dataset was set as 0.5 and 0.5, respectively. To guarantee the reliability of the classification results, the experiments, including random sampling and land use classification, were repeated 100 times for each language model. Table 3 shows the results of the accuracy assessment for land use classification. Figure 7 illustrates the confusion matrixes of these methods for the study area and Figure 8 displays the land use classification maps of four selected cities.

Table 3. Accurate assessment of land use classification results by using different semantic models.

Methods	Training process		Predicting process		Avg. comp. time
	OOB avg. error	OOB RMSE	OA	Kappa	Unit: seconds
Proposed method	0.1028 ± 0.0009	0.2301 ± 0.0013	0.8728 ± 0.0012	0.8399 ± 0.0007	161.0040
TF-IDF	0.1527 ± 0.0028	0.3384 ± 0.0043	0.5526 ± 0.0051	0.4162 ± 0.0023	146.3670
pLSA	0.1081 ± 0.0012	0.2361 ± 0.0021	0.7431 ± 0.0018	0.6719 ± 0.0010	3673.2530
LDA	0.1038 ± 0.0008	0.2375 ± 0.0016	0.6763 ± 0.0016	0.5841 ± 0.0026	1221.5500

W2V	PFL	GSL	IUL	CBF	RUL	APS	RST	LWL	SUL	URC	RTF	RPF	OCL	MUL	TFIDF	PFL	GSL	IUL	CBF	RUL	APS	RST	LWL	SUL	URC	RTF	RPF	OCL	MUL
PFL	0.81	0.09	0.01	0.01	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	PFL	0.51	0.03	0.06	0.37	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GSL	0.00	0.89	0.01	0.01	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	GSL	0.00	0.69	0.02	0.04	0.23	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IUL	0.00	0.06	0.86	0.01	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	IUL	0.00	0.32	0.41	0.02	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CBF	0.00	0.06	0.01	0.82	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	CBF	0.00	0.28	0.03	0.39	0.28	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RUL	0.00	0.05	0.01	0.01	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	RUL	0.00	0.20	0.02	0.03	0.74	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
APS	0.00	0.06	0.01	0.01	0.09	0.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	APS	0.00	0.27	0.02	0.04	0.29	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RST	0.00	0.10	0.01	0.02	0.07	0.01	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	RST	0.00	0.39	0.03	0.04	0.24	0.02	0.27	0.00	0.00	0.00	0.01	0.00	0.00	0.00
LWL	0.00	0.07	0.05	0.01	0.07	0.00	0.00	0.80	0.00	0.00	0.00	0.00	0.00	0.00	LWL	0.01	0.30	0.08	0.03	0.30	0.00	0.27	0.00	0.01	0.01	0.00	0.00	0.00	0.00
SUL	0.00	0.04	0.01	0.01	0.10	0.01	0.00	0.00	0.84	0.00	0.00	0.00	0.00	0.00	SUL	0.01	0.21	0.04	0.04	0.37	0.01	0.00	0.32	0.00	0.00	0.00	0.00	0.00	0.00
URC	0.00	0.08	0.02	0.01	0.09	0.00	0.00	0.00	0.00	0.80	0.00	0.00	0.00	0.00	URC	0.00	0.39	0.04	0.03	0.26	0.00	0.00	0.00	0.27	0.00	0.00	0.00	0.00	0.00
RTF	0.00	0.06	0.02	0.01	0.09	0.01	0.00	0.00	0.00	0.00	0.82	0.00	0.00	0.00	RTF	0.00	0.33	0.04	0.06	0.31	0.01	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00
RPF	0.00	0.29	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.57	0.00	0.00	RPF	0.00	0.43	0.14	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00
OCL	0.00	0.07	0.01	0.01	0.11	0.01	0.00	0.00	0.00	0.00	0.00	0.79	0.00	0.00	OCL	0.00	0.32	0.04	0.03	0.29	0.01	0.00	0.00	0.00	0.01	0.00	0.30	0.00	0.00
MUL	0.00	0.08	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.83	0.00	MUL	0.00	0.42	0.00	0.00	0.42	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00

(a) Word2Vec

(b) TF-IDF

pLSA	PFL	GSL	IUL	CBF	RUL	APS	RST	LWL	SUL	URC	RTF	RPF	OCL	MUL	LDA	PFL	GSL	IUL	CBF	RUL	APS	RST	LWL	SUL	URC	RTF	RPF	OCL	MUL
PFL	0.59	0.16	0.03	0.03	0.19	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	PFL	0.43	0.32	0.03	0.03	0.17	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GSL	0.00	0.79	0.03	0.02	0.14	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	GSL	0.00	0.78	0.03	0.02	0.16	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IUL	0.00	0.13	0.70	0.01	0.16	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	IUL	0.00	0.23	0.58	0.02	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CBF	0.00	0.13	0.02	0.64	0.20	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	CBF	0.00	0.21	0.03	0.58	0.17	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RUL	0.00	0.10	0.02	0.02	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	RUL	0.00	0.15	0.02	0.02	0.80	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
APS	0.00	0.12	0.02	0.03	0.20	0.63	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	APS	0.00	0.21	0.03	0.02	0.18	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RST	0.00	0.17	0.01	0.04	0.12	0.01	0.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	RST	0.00	0.30	0.03	0.02	0.18	0.03	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LWL	0.01	0.14	0.07	0.01	0.18	0.01	0.00	0.59	0.00	0.00	0.01	0.00	0.00	0.00	LWL	0.00	0.23	0.08	0.02	0.20	0.00	0.00	0.46	0.00	0.00	0.00	0.00	0.00	0.00
SUL	0.00	0.13	0.03	0.03	0.24	0.01	0.00	0.00	0.56	0.00	0.00	0.00	0.00	0.00	SUL	0.00	0.17	0.04	0.03	0.24	0.01	0.00	0.52	0.00	0.00	0.00	0.00	0.00	0.00
URC	0.00	0.16	0.05	0.01	0.17	0.01	0.00	0.00	0.00	0.60	0.00	0.00	0.00	0.00	URC	0.00	0.33	0.03	0.02	0.18	0.01	0.00	0.00	0.00	0.44	0.00	0.00	0.00	0.00
RTF	0.00	0.13	0.06	0.02	0.20	0.01	0.00	0.00	0.00	0.00	0.58	0.00	0.00	0.00	RTF	0.00	0.26	0.04	0.04	0.24	0.01	0.00	0.02	0.00	0.01	0.39	0.00	0.00	0.00
RPF	0.00	0.57	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00	RPF	0.00	0.71	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00
OCL	0.00	0.16	0.01	0.03	0.23	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.56	0.00	OCL	0.00	0.16	0.04	0.02	0.28	0.01	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00
MUL	0.00	0.17	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.58	MUL	0.00	0.67	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17

(c) pLSA

(d) LDA

Figure 7. Confusion matrixes of classification results via features extracted from (a) Word2Vec, (b) TF-IDF, (c) pLSA and (d) LDA.

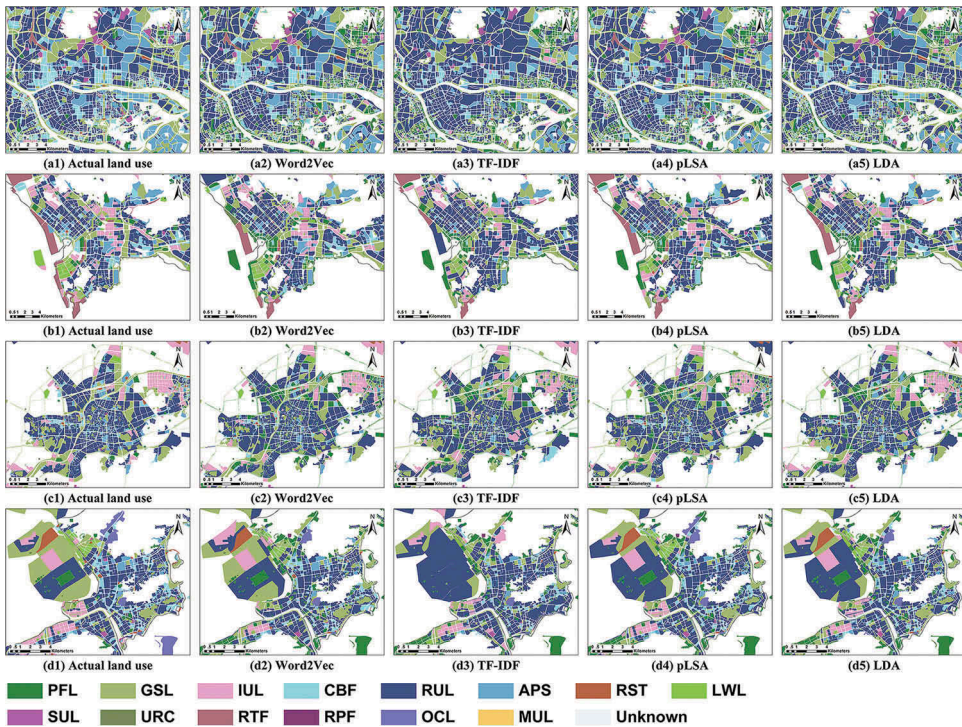


Figure 8. RFA-based land use classification results using several language models (Word2Vec/TF-IDF/pLSA/LDA) in study area. (a1)~(a5) Downtown areas of Guangzhou, (b1)~(b5) Nanshan districts of Shenzhen, (c1)~(c5) Downtown areas of Zhongshan, (d1)~(d5) Downtown areas of Zhuhai. In this figure: Land use types are Public facilities (PFL), Green space and square (GSL), Industrial land (IUL), Commercial and business facilities (CBF), Residential land (RUL), Administration and public services (APS), Road street and transportation (RST), Logistics and warehouse (LWL), Special use land (SUL), Urban and rural construction land (URC), Regional traffic facilities (RTF), Regional public facilities (RPF), Other construction land (OCL) and Mining use land (MUL).

As expected, the TF-IDF method performed poorly in the land use classification. In this method, TF-IDF values are only acquired using the distribution frequencies of POIs, thus overlooking the relationship between similar POIs and obtaining poor performances in extracting potential semantic information of documents (TAZs) (Blei *et al.* 2003, Li *et al.* 2010). PTMs, such as pLSA and LDA, are able to mine potential semantic features and model the interactions between features and different categories of scenes (Bosch *et al.* 2006, Zhang and Du 2015). PTMs have been widely applied in the scene classification of HSR RS imagery (Zhang and Du 2015, Zhong *et al.* 2015); however, they obtained unsatisfactory results in this study. To our knowledge, urban land use types relate to not only categories and quantities of inner micro entities but also their spatial distribution patterns (Jiang and Yao 2010). Conventional PTMs only consider the proportions of POIs and ignore their spatial correlations and context relationships, which may result in low land use classification precision. Moreover, the performance of the LDA-based model is quite sensitive to the setting of the initial hyper-parameter, and the optimal parameter settings vary according to how the LDA-based model is used to address specific tasks (Lu *et al.* 2011).

The above analysis has shown that our proposed method can obtain the highest classification accuracy with the lowest computation time compared with RFA-based methods. By considering both the quantitative features and spatial distribution features of POIs, the proposed method maps POIs into a high-dimensional feature space and obtains a better result, therein demanding fewer parameters compared with the RFA. Therefore, the proposed method can map POIs into a high-dimensional feature space using fewer parameters. This method should be useful for quantifying the correlation relationships between urban land use types and the spatial distribution patterns of POIs.

4.2 Discussion

In this section, we will first discuss the influence of POI vector dimensions and the sampling window size setting of the POIs. So far, there have been no reports on the parameter sensitivities of Word2Vec. Instead, the computational cost of the Word2Vec model has been discussed in some articles. For example, Mikolov's study suggested that increasing the vector dimension would double the computational complexity (Mikolov *et al.* 2013a). Therefore, we designed two experiments to analyze the parameter sensitivity of the proposed model. Figures 9 and 10 present the fluctuations of the classification accuracy under different experiments. We note that the classification accuracy based on POI vectors first increases and then achieves steady state with increases along the x-axis, which represents the vector dimensions and sampling window size in Figures 9 and 10, respectively. It can be stated that, when the sizes of the vectors and sample windows are set too small in the training process, the input information becomes insufficient, and the rapid over-convergence problem results, further generating inaccurate word (POI) vectors and resulting in poor classification accuracy. Conversely, when these two parameters are set to a sufficient scale, the OA of land use classification can be increased to approximately 0.85 and become stable.

Our overall land use classification accuracy is slightly lower than those of previous studies that conducted natural text sentiment classification using the Word2Vec model (Xue *et al.* 2014, Lilleberg *et al.* 2015, Zhang *et al.* 2015). This can be mainly attributed to

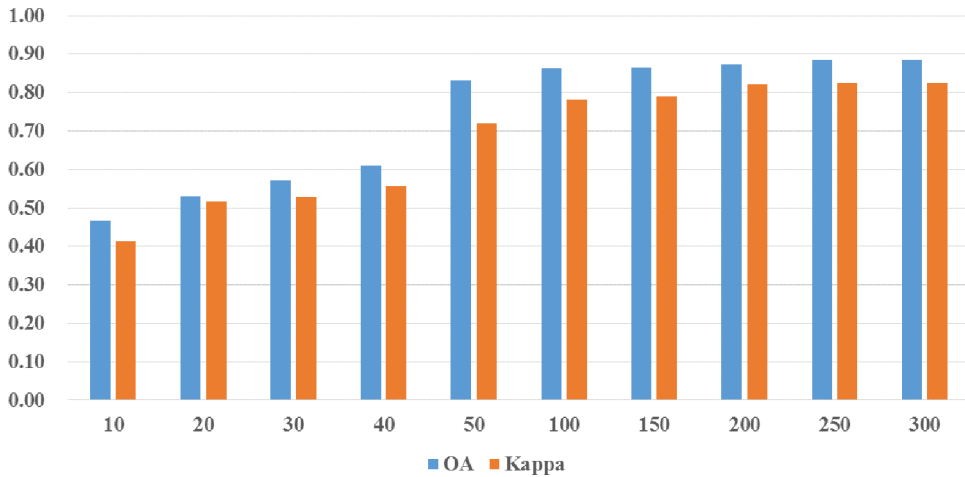


Figure 9. Changes of land use classification accuracy (Overall accuracy and Kappa, y-axis) by different POI vector dimensions (x-axis) while sampling window size is set to 5.

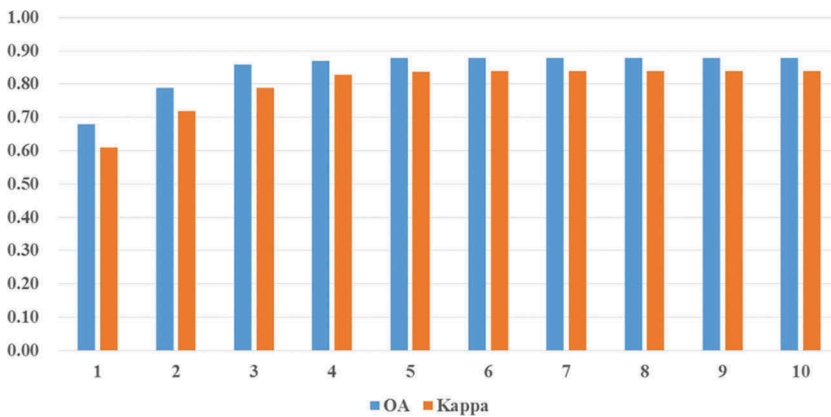


Figure 10. Changes of land use classification accuracy (Overall accuracy and Kappa, y-axis) by different sampling window size (x-axis) while POI vectors dimension is set to 200.

two causes: on the one hand, compared with natural text analysis, the complex relationship between different land use types causes difficulties in determining the actual functions of regions; on the other hand, land use layouts, which are proposed by planning departments, may be disagreement with the actual situation due to China's rapid development (Tian and Shen 2011, Long *et al.* 2012). Moreover, in this study, our validation data exhibit significant differences compared with Baidu POI data (2015) in terms of timescale. This will inevitably increase land use classification errors.

In addition, there are a lot of mixed land use types in the interior of the downtown, and the land use types are the hard classification results of artificial interpretation. The main purpose of this study is to investigate the potential and effectiveness of POI vector, so when selecting the classification samples, we did not taking into account the issue across the pure patches and mixed patches, which left errors of omission in classification

to some extent. In future research, we will further develop the RFA model by taking into account the statistical sampling of patches with pure land use type in order to get the land use ratio of each patch.

Since POI data obtained from Baidu API lacked attributes of land area and population, POIs within TAZ were all specified to the same weight during the construction of TAZ vector. Though lacking the detailed attributes we desired, the proposed method still obtained fine clustering and classification result. It can be explained that in a city, common ground objects (convenience stores, residential blocks, etc.) and uncommon ground objects (universities, hospitals, etc.) can be seemed as common words and uncommon words in natural language, which are interpreted via context relationships. In Google Word2vec model, weights of uncommon words would be lifted and common words would be assigned to lower importance according to their occurrence in the context, which assures good performance in text classification of POIs without thick-enough attribute information (Lilleberg *et al.* 2015). In future studies, we will involve more multi-source spatial data (such as HSR images and official statistical data) in the model and enter into detailed discussion on the weighting issue of different types of POIs.

Despite some of the abovementioned insufficiencies, the proposed method is able to incorporate spatial factors into the corpus-building process. This method is the first to map POIs to a high-dimension spatial vector by considering spatial distribution features of POIs. The analysis indicates that POI vectors have the potential to identify urban function structures and land use types. For example, according to the cosine distance among POI vectors, we can learn that those vectors with the highest spatial correlations with the 'Pharmacy' POI are 'Snapshot Printing', 'Supermarket' and 'Convenience store'. Furthermore, through region aggregation via TAZ vectors, we can understand multi-scale urban spatial structures in depth by specifying different clustering numbers. To our knowledge, due to a lack of effective methods, previous studies on urban land use classification only used the frequency of POI categories as judgment and did not consider the underlying features of the spatial distributions of POIs (Yuan *et al.* 2012, Jiang *et al.* 2015). In this regard, our study provides a new method for mining the spatial distribution features of POIs.

Moreover, this study can help urban planners to monitor dynamic urban land use changes and evaluate the impacts of urban planning schemes. In this study, POI vectors are used to detect urban land use types, and the POIs used are continually updated on the Internet. Therefore, this indicates that urban land use changes can be monitored by keeping pace with the POI updated speed. In future studies, the proposed method can be used to map finer urban land use distributions by integrating HSR RS images. Regions with the same functions not only have semblable social economic attributes (e.g., spatial distributions of POIs) but also present similar spatial patterns. Therefore, in our forthcoming work, how to combine HSR RS images and geospatial big data to obtain more accurate and sophisticated land use patterns is the first question to be resolved.

5. Conclusions

In this study, we established a framework for sensing spatial structures of urban land use at the unit of the TAZ by integrating POIs with a deep-learning language model (Google Word2Vec). The objective is to classify urban land use at a fine patch scale by considering

the context relationships of POIs. First, a greedy algorithm-based shortest-path model was used to build a TAZ-POI corpus, in which TAZs and POI categories are considered as documents and basic words, respectively. In the next step, POIs and TAZs were mapped to a high-dimensional spatial vector by building a CBOW-based Word2Vec model, and the results demonstrated that the spatial correlations of POIs have been reasonably quantified by spatial vectors. Moreover, the results of K-Means-based region aggregation effectively revealed complex urban structures at multiple scales. By comparing with some state-of-the-art topic models (TF-IDF, pLSA and LDA), the proposed method was applied to the classification of urban land use and obtained the highest accuracy and efficiency (OA = 0.8728, kappa = 0.8399, average computation time = 161.0040 s). In our future study, the integration of the proposed method with HSR RS images will be conducted to satisfy a variety of applications in urban planning and environmental management.

Acknowledgment

We thank the anonymous reviewers for their useful comments and suggestions. This study was supported by the Key National Natural Science Foundation of China (Grant No. 41531176), the National Natural Science Foundation of China (Grant No. 41371376) and the National Science Fund for Excellent Young Scholars of China (Grant No. 41322009).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was supported by the Key National Natural Science Foundation of China [Grant No. 41531176] and the National Natural Science Foundation of China [Grant No. 41371376] and the National Science Fund for Excellent Young Scholars of China [Grant No. 41322009].

ORCID

Yao Yao  <http://orcid.org/0000-0002-2830-0377>
 Xia Li  <http://orcid.org/0000-0003-3050-8529>
 Penghua Liu  <http://orcid.org/0000-0002-8574-891X>
 Zhaotang Liang  <http://orcid.org/0000-0001-9261-5261>
 Jinbao Zhang  <http://orcid.org/0000-0001-8510-149X>
 Ke Mai  <http://orcid.org/0000-0002-3532-9872>

References

- Aizawa, A., 2003. An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39 (1), 45–65. doi:10.1016/S0306-4573(02)00021-3
- Arsanjani, J.J., et al., 2013. Toward mapping land-use patterns from volunteered geographic information. *International Journal of Geographical Information Science*, 27 (12), 2264–2278. doi:10.1080/13658816.2013.800871
- Biau, G.E.R., 2012. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13 (1), 1063–1095.

- Blaschke, T., 2010. Object based image analysis for remote sensing. *Isprs Journal of Photogrammetry And Remote Sensing*, 65 (1), 2–16. doi:10.1016/j.isprsjprs.2009.06.004
- Blaschke, T., et al., 2014. Geographic object-based image analysis – Towards a new paradigm. *Isprs Journal Of Photogrammetry And Remote Sensing*, 87, 180–191. doi:10.1016/j.isprsjprs.2013.09.014
- Blei, D.M., et al., 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.
- Bosch, A., Zisserman, A., and Oz, X. 2006. Scene classification via pLSA. In: *Computer Vision - ECCV 2006, European Conference on Computer Vision*, 7–13 May 2006, Graz, vol. 3954, 517–530.
- Bratananu, D., Nedelcu, I., and Datcu, M., 2011. Bridging the semantic gap for satellite image annotation and automatic mapping applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4 (1), 193–204. doi:10.1109/JSTARS.2010.2081349
- Breiman, L., 2001. Random forests. *Machine Learning*, 45 (1), 5–32. doi:10.1023/A:1010933404324
- Chen, K., et al., 2013. Semantic annotation of high-resolution remote sensing images via gaussian process multi-instance multilabel learning. *IEEE Geoscience and Remote Sensing Letters*, 10 (6), 1285–1289. doi:10.1109/LGRS.2012.2237502
- Dupuy, S., et al., 2012. Land-cover dynamics in Southeast Asia: contribution of object-oriented techniques for change detection. In: R. Queiroz Feitosa, ed. *4th international conference on GEOgraphic Object-Based Image Analysis (GEOBIA)*, Rio de Janeiro, Brazil, 217–222.
- Ellis, E. 2007. Land-use and land-cover change. In: R. Pontius and C.J. Cleveland, eds. *Encyclopedia of earth*. Available from: http://www.eoearth.org/article/Land-use_and_land-cover_change
- Fakhraei, S., Soltanian-Zadeh, H., and Fotouhi, F., 2014. Bias and stability of single variable classifiers for feature ranking and selection. *Expert Systems with Applications*, 41 (15), 6945–6958. doi:10.1016/j.eswa.2014.05.007
- Hayashi, Y., and Roy, J. 2013. *Transport, land-use and the environment*. Springer Science & Business Media.
- Hu, S. and Wang, L., 2013. Automated urban land-use classification with remote sensing. *International Journal Of Remote Sensing*, 34 (3), 790–803. doi:10.1080/01431161.2012.714510
- Huang, X. and Zhang, L., 2013. An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 51 (1), 257–272. doi:10.1109/TGRS.2012.2202912
- Jiang, B., and Yao, X. 2010. *Geospatial analysis and modelling of urban structure and dynamics*, vol. 99. Springer Science & Business Media.
- Jiang, S., et al., 2015. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and Urban Systems*, 53, 36–46. doi:10.1016/j.compenvurbsys.2014.12.001
- La Rosa, D. and Privitera, R., 2013. Characterization of non-urbanized areas for land-use planning of agricultural and green infrastructure in urban contexts. *Landscape and Urban Planning*, 109 (1), 94–106. doi:10.1016/j.landurbplan.2012.05.012
- Li, M., Maitre, H., and Datcu, M., 2010. Semantic annotation of satellite images using latent dirichlet allocation. *IEEE Geoscience and Remote Sensing Letters*, 7 (1), 28–32. doi:10.1109/LGRS.2009.2023536
- Lilleberg, J., Zhu, Y., and Zhang, Y. 2015. Support vector machines and Word2vec for text classification with semantic features. In: *Cognitive Informatics & Cognitive Computing (ICCI* CC)*, 2015 IEEE 14th International Conference on, July. IEEE, 136–140.
- Liu, B., et al., 2003. Building text classifiers using positive and unlabeled examples. In: *IEEE Third International Conference on Data Mining, 2003 (ICDM 2003)*. IEEE, 179–186.
- Liu, X., et al., 2014. Simulating urban growth by integrating landscape expansion index (lei) and cellular automata. *International Journal of Geographical Information Science*, 28 (1), 148–163.
- Liu, Y., et al., 2015. Social sensing: a new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers* (ahead-of-print), 105(3), 512–530.
- Long, Y., Gu, Y., and Han, H., 2012. Spatiotemporal heterogeneity of urban planning implementation effectiveness: evidence from five urban master plans of Beijing. *Landscape and Urban Planning*, 108 (2–4), 103–111. doi:10.1016/j.landurbplan.2012.08.005

- Long, Y., and Liu, X., 2013. Automated identification and characterization of parcels (AICP) with OpenStreetMap and Points of Interest. *arXiv preprint arXiv:1311.6165*.
- Lu, Y., Mei, Q., and Zhai, C., 2011. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, 14 (2), 178–203. doi:10.1007/s10791-010-9141-9
- Mikolov, T., et al., 2013a. Efficient estimation of word representations in vector space. *arXiv Preprint Arxiv:1301.3781*.
- Mikolov, T., et al., 2013b. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. 3111–3119.
- Mikolov, T., Yih, W. T., and Zweig, G., 2013c. Linguistic regularities in continuous space word representations. In: *HLT-NAACL*, June, vol. 13, 746–751.
- Mordelet, F. and Vert, J.-P., 2014. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37, 201–209. doi:10.1016/j.patrec.2013.06.010
- Mountrakis, G., Im, J., and Ogole, C., 2011. Support vector machines in remote sensing: a review. *Isprs Journal of Photogrammetry and Remote Sensing*, 66 (3), 247–259. doi:10.1016/j.isprsjprs.2010.11.001
- Ng, H.T., et al., 1997. Corpus-based approaches to semantic interpretation in NLP. *Ai Magazine*, 18 (4), 45.
- Palczewska, A., et al., 2014. Interpreting random forest classification models using a feature contribution method. In: *Integration of reusable systems*. Springer, 193–218.
- Regan, C.M., et al., 2015. Real options analysis for land use management: methods, application, and implications for policy. *Journal of Environmental Management*, 161, 144–152. doi:10.1016/j.jenvman.2015.07.004
- Rodrigues, F., et al., 2012. Automatic classification of points-of-interest for land-use analysis. In: *Proceedings of the fourth international conference on advanced geographic information systems, applications, and services (GEOProcessing)*, January, 41–49.
- 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. doi:10.1016/0377-0427(87)90125-7
- Rutkowski, L., 2007. Clustering for data mining: a data recovery approach. *Psychometrika*, 72 (1), 109–110. doi:10.1007/s11336-005-1358-y
- Schwenk, H., 2007. Continuous space language models. *Computer Speech & Language*, 21 (3), 492–518. doi:10.1016/j.csl.2006.09.003
- Sun, H., et al., 2012. Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. *IEEE Geoscience and Remote Sensing Letters*, 9 (1), 109–113. doi:10.1109/LGRS.2011.2161569
- Tian, L. and Shen, T., 2011. Evaluation of plan implementation in the transitional China: a case of Guangzhou city master plan. *Cities*, 28 (1), 11–27. doi:10.1016/j.cities.2010.07.002
- Tokarczyk, P., et al., 2015. Features, color spaces, and boosting: new insights on semantic classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 53 (1), 280–295. doi:10.1109/TGRS.2014.2321423
- Wen, D., et al., 2016. A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multi-index scene representation. *IEEE Transactions on Geoscience And Remote Sensing*, 54 (1), 609–625. doi:10.1109/TGRS.2015.2463075
- Williamson, I.P., et al., 2010. *Land administration for sustainable development*. Redlands, CA: ESRI Press Academic.
- Xue, B., Fu, C., and Shaobin, Z., 2014. A study on sentiment computing and classification of sina weibo with word2vec. In: *2014 IEEE International Congress on Big Data*, June. IEEE, 358–363.
- Yang, Y., and Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. In: *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, November. ACM, 270–279.
- Yin, J., et al., 2011. Monitoring urban expansion and land use/land cover changes of Shanghai metropolitan area during the transitional economy (1979–2009) in China. *Environmental Monitoring and Assessment*, 177 (1–4), 609–621. doi:10.1007/s10661-010-1660-8

- Yu, M., and Dredze, M., 2014. Improving lexical embeddings with semantic knowledge. In: *Association for Computational Linguistics (ACL) (2)*, June, 545–550.
- Yuan, J., Zheng, Y., and Xie, X., 2012. *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*, August. ACM, 186–194.
- Zhang, D., et al., 2015. Chinese comments sentiment classification based on word2vec and SVM perf. *Expert Systems with Applications*, 42 (4), 1857–1863. doi:[10.1016/j.eswa.2014.09.011](https://doi.org/10.1016/j.eswa.2014.09.011)
- Zhang, X. and Du, S., 2015. A linear dirichlet mixture model for decomposing scenes: application to analyzing urban functional zonings. *Remote Sensing of Environment*, 169, 37–49. doi:[10.1016/j.rse.2015.07.017](https://doi.org/10.1016/j.rse.2015.07.017)
- Zhao, B., Zhong, Y., and Zhang, L., 2013. *Hybrid generative/discriminative scene classification strategy based on latent Dirichlet allocation for high spatial resolution remote sensing imagery*. In: *2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS*. IEEE, 196–199.
- Zheng, Y., et al., 2014. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5 (3), 38.
- Zhong, Y., Zhu, Q., and Zhang, L., 2015. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 53 (11), 6207–6222. doi:[10.1109/TGRS.2015.2435801](https://doi.org/10.1109/TGRS.2015.2435801)