# Extracting Urban Functional Regions from Points of Interest and Human Activities on Location-based Social Networks

Song Gao, Krzysztof Janowicz, Helen Couclelis

Department of Geography,

University of California, Santa Barbara, CA, USA

**Abstract:** Data about points of interest (POI) have been widely used in studying urban land use types and for sensing human behaviors. However, it is difficult to quantify the right mix or the spatial relations among different POI types indicative of specific urban functions. In this research, we develop a statistical framework to help discover semantically meaningful topics and functional regions based on the co-occurrence patterns of POI types. The framework applies the latent Dirichlet allocation (LDA) topic modeling technique and incorporates user check-in activities on location-based social networks. Using a large corpus of about 100,000 Foursquare venues and user check-in behaviors in the ten most populated urban areas of the United States, we demonstrate the effectiveness of our proposed methodology by identifying distinctive types of latent topics and further, by extracting urban functional regions using K-means clustering and Delaunay triangulation spatial constraints clustering. We show that a region can support multiple functions but with different probabilities, while the same type of functional region can span multiple geographically non-adjacent locations. Since each region can be modeled as a vector consisting of multinomial topic distributions, similar regions with regard to their thematic topic signatures can be identified. Compared to remote sensing images which mainly uncover the physical landscape of urban environments, our popularity-based POI topic modeling approach can be seen as a complementary social sensing view on urban space based on human activities.

**Keywords:** urban functions, topic modeling, latent Dirichlet allocation, LBSN, place type.

# 1 Introduction

Cities support a variety of functions that relate to land use types, including residential, commercial, industrial, transportation, and business regions and infrastructure, while affording different types of human activities, such as living, working, commuting, shopping, eating, and recreation. Rapid urbanization and new construction have caused land use changes and urban expansions in many areas. Remote sensing images together with spatial metrics have been widely used to classify urban land use and monitor change at different spatial scales (Barnsley and Barr, 1996; Herold et al., 2005; Banzhaf and Netzband, 2012). However, human activities usually take place in different types of points of interest (POIs). Remote sensing techniques perform well in extracting physical characteristics, such as land surface reflectivity and texture of urban space but are not good in identifying functional interaction patterns or in helping understand socioeconomic environments (Pei et al., 2014; Liu et al., 2015). Compared to other datasets and methods in remote sensing and field mapping, using POI data, social media, and their associated methods can lead to a better understanding of individual-level and group-level utilization of urban space at a finegrained spatial and temporal resolution. Rich *social sensing* techniques can help bridge the semantic gap between land use classification and urban functional regions. The function of a place is determined by what type of activities can occur there (Janowicz, 2012; Zhong et al., 2014). The same types of POIs can be located in different land use types and may also support different functions. For example, *restaurants* are found in residential areas and in commercial areas, as well as in industrial areas. The main function of *universities* is education, but they also support sports activities, music shows, and so on. Previous studies have demonstrated that different POI types have distinctive semantic signatures (Janowicz, 2012) (i.e., spatial, temporal, and thematic distributions) based on crowd-sourced location-based social media data analysis, in analogy to spectral bands in remote sensing (McKenzie et al., 2015). There is a growing trend of using location-awareness sensing data (e.g., trajectories from mobile phones), POI data, and social media feeds to study the spatial and social structure of urban environments (Pei et al., 2014; Hu et al., 2015; Jiang et al., 2015; Liu et al., 2015; McKenzie et al., 2015; Steiger et al., 2016; Yao et al., 2017). However, few studies have investigated the latent relationships among different types of POIs and how they spatially interact with each other to support urban functions, such as education, business, and shopping. In this research, we aim to develop a data-driven framework to discover urban functional regions from POIs and associated human activities on location-based social networks (LBSN).

We argue that geographic knowledge and measures of spatial distribution over POI types (categories) can be employed to derive latent classification features for these said types, which will then enable the detection and the abstraction of higher-level functional regions (i.e., semantically coherent areas of interest) such as shopping areas, business districts, educational areas, and tourist zones. To test this claim, we will study the co-occurrence patterns of different

POI categories as well as the associated human activities (i.e., mobility, check-ins, reviews, and comments), and thus employ analytical measures to quantify their differences and conduct classifications of functional regions.

The contributions of this research are as follows:

- We propose a novel framework to study urban functional regions by employing data about Points Of Interest and human activities derived from social media.
- We incorporate location-based social network user check-ins into a probabilistic topic modeling technique to discover functional co-occurrence patterns of different POI types.
- The proposed method can support functional inferences for specific type of regions and thus serve as a new heuristic to enable the search for similar urban places/regions, based on their POI-type distributions and corresponding human activities, and using natural language processing and machine learning techniques.

The remainder of this article is structured as follows. Section 2 introduces background material and related work. Next, Section 3, discusses the datasets used and the selection of study areas. Section 4 introduces the methods used in our framework and specifically LDA. In Section 5, we present the results of topic modeling to characterize, cluster, and compare functional regions. Next, we discuss the broader implications of our work in Section 6 before concluding and pointing out directions of future work in Section 7.

## 2 Related Work

With the increasing popularity of travel blogs, volunteered geographic information (VGI), location-based social networks (LBSN), and so forth, researchers have developed a variety of place-based studies that employ datasets from these various sources. For instance, Adams and Janowicz (2012) presented a topic modeling methodology to estimate geographic regions from unstructured, non geo-referenced text on Wikipedia and travel blogs by computing a density surface of geo-indicative topics over the Earth's surface. The proposed framework combined natural language processing techniques, geostatistics methods, and data-driven bottom-up semantics. In order to evaluate the use of topic modeling techniques on the extraction of thematic characteristics of places, Adams and McKenzie (2013) applied that approach on a set of travel blog entries to identify the themes that are most closely associated with specific places around the world. Their proposed method is capable of measuring the degree to which certain themes are local or global, as well as analyzing thematic changes over time. POI data play an important role for human activity-based land use, transportation, and environmental models. Jiang et al. (2015) utilized the Yahoo online POI data together with publicly available aggregated employment data from census at the block group level to derive fine resolution of disaggregated land use estimates (i.e., employment by category) at the city block level. For the evaluation, they first used a variety of machine learning algorithms to match and cluster POI types into a labeled business establishment taxonomy, and then compared it with ground-truth data from commercial business data vendors. The results demonstrated that their proposed method got a better goodness of fit with a lower relative mean squared error for the estimated employment population across all city blocks than that from the traditional uniform-distribution disaggregation approach. As for LBSN applications, Noulas et al. (2011) proposed a method to classify the geographical areas and LBSN users based on place types and the users' check-in statistics in Foursquare venues. The experiments were conducted in the metropolitan cities of London and New York and identified similar regions and user groups in each city. However, they didn't consider the temporal pattern of user activities. Later on, Yuan et al. (2012) employed both POI type information and the temporal patterns of taxi pick-ups/drop-offs in segmented map regions, utilized a topic modeling method based on latent Dirichlet allocation and Dirichlet multinomial regression techniques, and discovered various urban functional regions in the city of Beijing. The extracted region clusters were annotated as nine different groups: *diplomatic and embassy areas, education and science areas, developed residential areas, emerging residential areas, developed commercial/entertainment areas, developing commercial/entertainment areas, regions under construction, areas of historic interests, and nature & parks.* However, such rich multiple datasets that complement each other in the same city and especially high-precision mobility data are usually hard to fully access. One challenging issue is how to semantically classify and label the regions that are found given only one data source, and how to find similar places and regions across different cities. Adams (2015) proposed a novel observation-to-generalization place model and employed natural language processing techniques to derive place attributes. The proposed methods can support similar-place-search functions and the case studies were conducted using over 600,000 place articles on Wikipedia as a proof of concept. Later, Adams and Janowicz (2015) presented a novel method to enrich the place information on linked knowledge graphs using thematic signatures that are derived from unstructured text through topic modeling. This method can also be used to clean miscategorized places on the linked data cloud. In another study, Hobel et al. (2015) developed a semantic region growing algorithm based on the density of POIs on OpenStreetMap to extract places that afford certain type of human activities, e.g., shopping areas. In their model, four features including the number of *banks & ATM, restaurants, tourist facilities,* *and subcategories of shops* were used to identify the shopping areas/settings. They then compared the similarity of shopping areas/settings based on the four features in two European capital cities: Vienna and London. By incorporating human spatio-temporal activity data from social media, Zhou and Zhang (2016) extracted the spatial distribution hotspots of six types of urban functions (i.e., *Travel & Transport, Education Resource, Shop & Service, Nightlife Spot, Outdoor & Recreation, Food & Restaurant*) in the cities of Boston and Chicago. Zhi et al. (2016) introduced a low-rank-approximation-based model to detect functional regions based on 15 million social media check-in records in the city of Shanghai, China. This method discovered latent spatio-temporal human activity patterns and linked these with different functional regions. Researchers are also interested in the regional differences on discovering thematic characteristics of different POI types. McKenzie and Janowicz (2017) identified the most and least spatially varying place types and compared their thematic signatures internationally. The ongoing trend in this research direction lies in data-synthesis-driven approaches to study places and vague cognitive regions as well as the semantic generalizations of urban settings (Hobel et al., 2015, 2016; Gao et al., 2017).

In summary, there is a variety of research studying places and place types from human data traces, including spatio-temporal human mobility patterns that can reveal the functions of regions. However, only a few studies have simultaneously considered both POI information and human activities on location-based social network to derive urban functional regions. Moreover, to the best of our knowledge, there is no thorough discussion of the robustness of discovered urban and regional functional areas using different numbers of topics and clusters. There has also been no attempt to develop an urban function ontology based on the structure of POIs using a bottom-up approach.

# 3 Study Area and Datasets

## 3.1 Study Area

Urban areas – cities for short – are the highly populated places on the planet and include metropolitan regions, urban districts, towns, and suburbs. In order to explore the thematic characteristics and semantic clusters of urban areas in connection with urban functions, the ten most populated U.S. cities based on the 2015 population census: *New York, Los Angeles, Chicago, Houston, Philadelphia, Phoenix, San Antonio, San Diego, Dallas, and San Jose* and their surrounding metropolitan regions were selected as our study areas. The cartographic boundaries of those ten metropolitan areas are downloaded from the U.S. Census Bureau's TIGER geographic database[1].

## 3.2 Points of Interest Dataset

People usually go to different POIs for different kinds of activities, e.g., studying, working, dining, shopping, and relaxing. We assume that the spatial distributions and interactions of different types of POIs reflect particular urban functions. Location-based social networks such as Foursquare have created traces of social interactions based on the physical location of users. In these LBSN systems, users can check-in to a venue (i.e., a POI), rate it, and share their comments or tips. As shown in Figure 1, we first randomly generated 200 points as search locations in each urban area and then identified the surrounding Foursquare venues with their attribute information including *name, location coordinates, place category, number of check−ins, number of checked users, number of tips,and the rating score* in each search locations. Note that because of the Foursquare developer API limits, we only retrieved at most 50 nearby venues given a random search point. The POI data were collected in De-

[1] `https://www.census.gov/geo/maps-data/data/cbf/cbf_msa.html`

Table 1: The 95% inverse CDF distance thresholds for all the ten urban areas

| City Name | 95% Inverse-CDF Distance Threshold (km) |
|---|---|
| Chicago | 7.389 |
| Dallas | 8.994 |
| Houston | 8.647 |
| Los Angeles | 4.474 |
| New York | 7.123 |
| Philadelphia | 8.894 |
| Phoenix | 7.969 |
| San Antonio | 7.475 |
| San Diego | 7.837 |
| San Jose | 4.822 |
| Average | 7.363 |

cember 2016 and the attribute information for all venues is a historic snapshot at that time. There is a total of 480 different POI types in our data. Figure 2 shows the empirical cumulative density function (CDF) of the distance distribution between each Foursquare venue and the corresponding search location. Steeper curves (with larger slope values) before reaching the relatively steady state (about 95% cumulative probability) show that more POIs are closer to the search locations given the same number of Foursquare venues. In order to generate most 'nearby' POIs around each search location, we further spatially filtered out those venues outside the 95% inverse CDF distance threshold; i.e., we only selected those venues within a relatively small search distance. The distance thresholds differ among cities as shown in Table 1.

# 4 Methods

## 4.1 Popularity-based Probabilistic Topic Model

Probabilistic topic models have been widely used to discover latent thematic characteristics and their structure when analyzing large sources of textual doc-

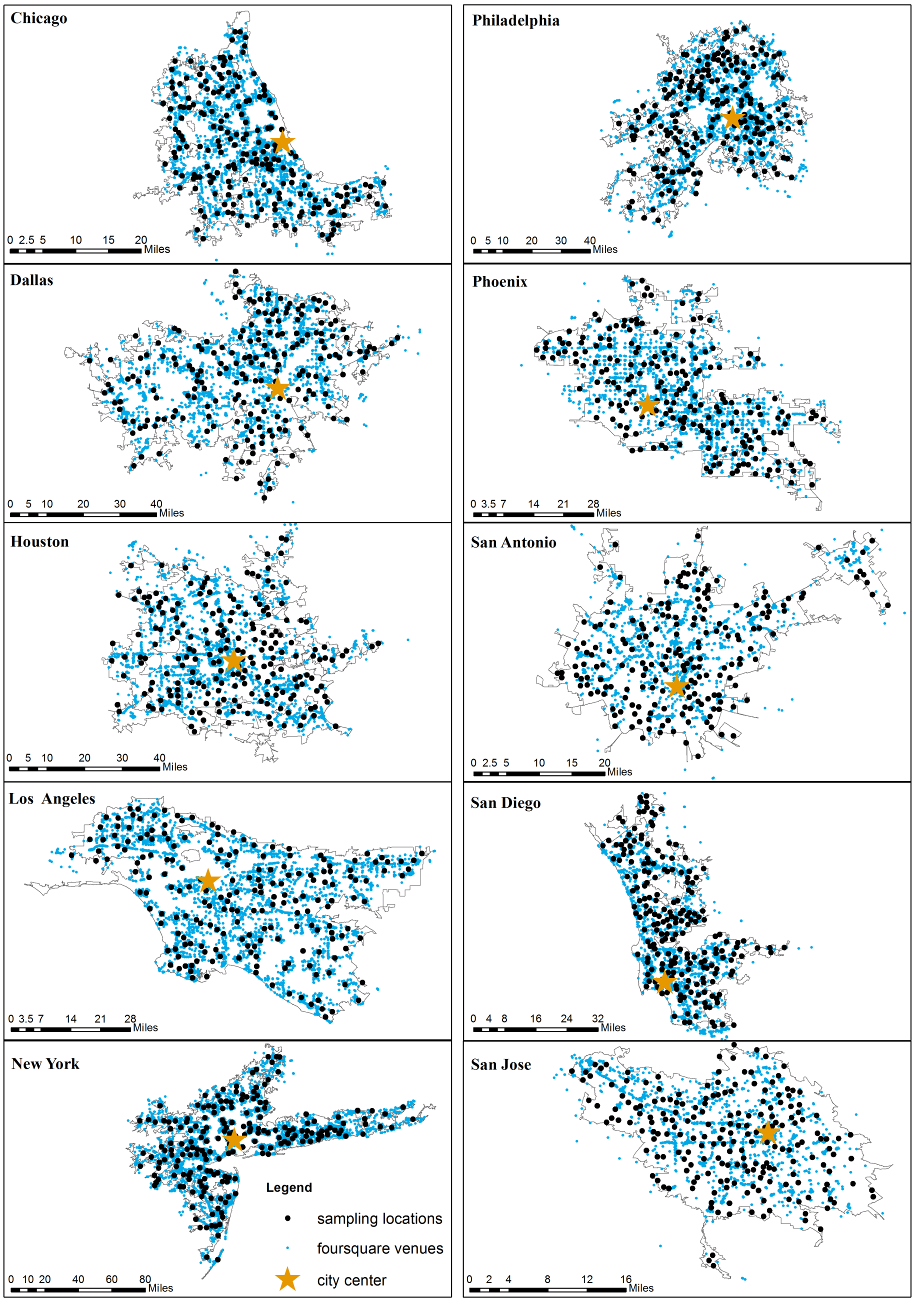


Figure 1: The spatial distributions of sampling locations and the collected Foursquare venues (POIs) in ten urban areas.

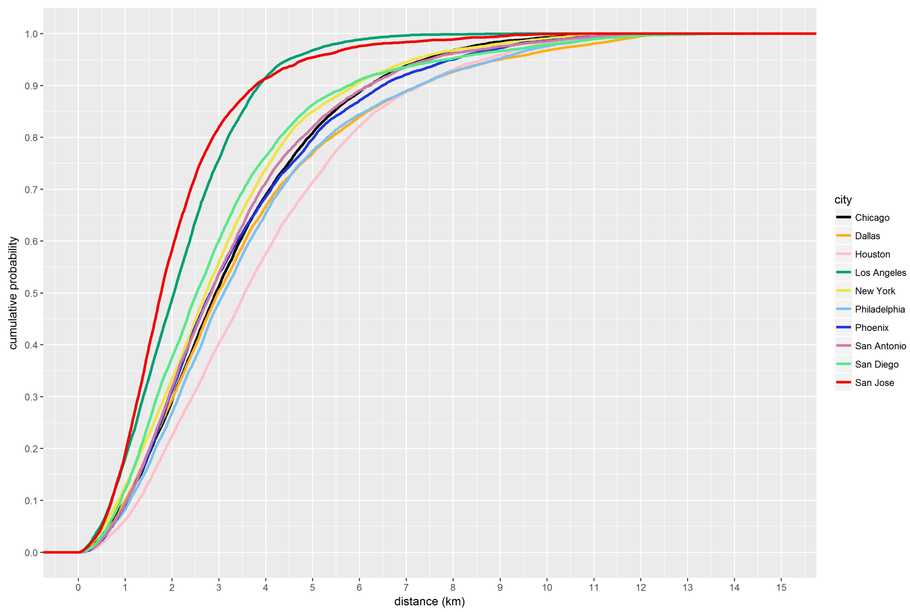


Figure 2: The cumulative density function of the distance distribution between each Foursquare venue and the corresponding search location.

uments (Blei et al., 2003; Steyvers and Griffiths, 2007; Blei, 2012). The *latent Dirichlet allocation (LDA)* is among the most popular topic modeling methods. LDA is an unsupervised generative probabilistic model that takes a bag-of-words approach (which implies that the order of words in the document does not matter) to constructing topics. The key idea of LDA is that documents can be represented as a joint probability distribution over latent topics and each topic is characterized by a distribution over words (Blei et al., 2003). Assume that there are total $K$ number of topics associated with $N$ words in the document corpus $D$, and $\alpha$ and $\eta$ represent the prior parameters for the Dirichlet document-topic and topic-word distribution respectively. The mathematical relationship between the latent variables and the observed variables is described below:

$$p(\beta_{1:K}, \theta_{1:D}, Z_{1:D}, W_{1:D})$$
$$= \prod_{i=1}^{K} p(\beta_i) \prod_{i=1}^{D} p(\theta_d) (\prod_{n=1}^{N} p(Z_{(d,n)}|\theta_d) p(W_{(d,n)}|\beta_{1:K}, Z_{(d,n)})) \quad (1)$$

As shown in Figure 3, the generative process can be described as follows:

I. Let $\beta_k$ denote a probabilistic distribution over the word vocabulary for a given topic $k$, and draw $\beta_k \sim \mathrm{Dir}(\eta)$;

II. Let $\theta_d$ represent the topic proportions for the $d_{th}$ document, and draw $\theta_d \sim \mathrm{Dir}(\alpha)$;

III. Let $Z_{(d,n)}$ denote the topic assignment for the $n_{th}$ word in document $d$ and $W(d, n)$ represent the $n_{th}$ word in document $d$ from a fixed vocabulary, and draw the multinomial distributions: $Z_{(d,n)} \sim \mathrm{Multinomial}(\theta_d)$ and $W_{(d,n)} \sim \mathrm{Multinomial}(\beta_{z_{(d,n)}})$.

In order to compute the conditional distribution of the topic structure given the word observations in documents, the expectation−maximization (EM) algorithm and *Gibbs sampling* are most the commonly used methods. After finishing the computation, two matrices $\theta$ and $\beta$ associated with topic proportions and assignments are generated. More detailed notations, calculations, and explanations can be found in Blei et al. (2003).

In analogy with LDA's use of textual materials, we take the type (e.g., school, park, restaurant) of each POI as a word, the search region that contains those POIs as a document, and an urban function or a land use as a topic that represents thematic characteristics and the semantics of places. By running the LDA topic modeling technique, we can find the posterior probabilistic distribution of each POI type in a certain type of region conditioned on the search

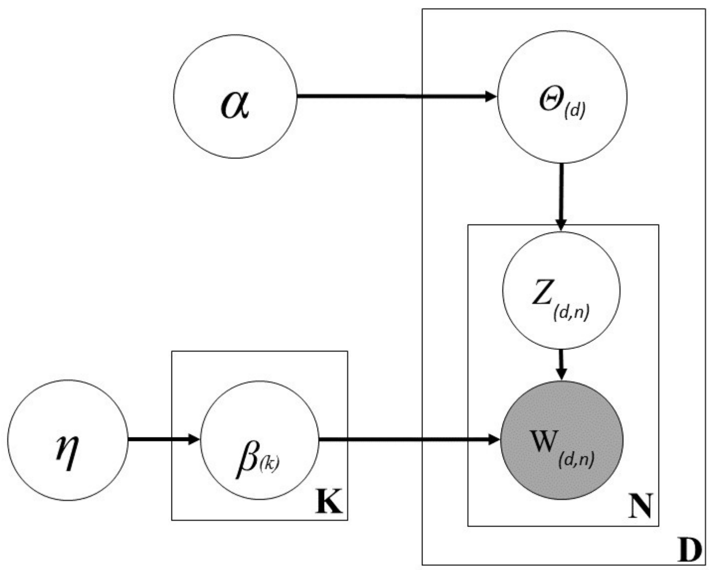


Figure 3: The graphical representation for latent Dirichlet allocation. The topic-related random variables in the generative process are the unshaded nodes while the observed words in documents are represented as a shaded node. The rectangles are "plate" notation that denotes replication.

region's topic assignments. The LDA model running on POI types generates summaries of thematic place topics (e.g., beach promenades, art zones, shopping areas) with a discrete probability distribution over POI types for each topic, and infers per-search-region probability distributions over topics. For example, one would assume that a *beach promenade* topic should contain venues such as *beach*, *seafood restaurants*, and *surfing spots*; while a *shopping area* would more likely contain *clothing stores*, *cosmetics shops*, and *shoe stores*.

Another important concept in our method is the *popularity* of a POI as captured by its LBSN user check-in behaviors. For example, the neighborhood of a football stadium usually has only one instance of *stadium* surrounded by dozens of *sports bars*, *restaurants*, and *parking lots*. However, a stadium usually attracts thousands of visitors and is the dominant feature of its neighborhood. Thus this particular POI type makes said neighborhood distinct from other neighborhoods (e.g., nightlife zone), which also contain *cocktail bars* and *restaurants*. We need to address such an human activity effect during the generation of the document-word frequency matrix. More specifically, we will rescale the POI type occurrences according to their associated POI instance check-in counts. The rescaling process can be represented as follows:

$$Freq_{(d,t)} = \sum_i Log(V_{(d,t,i)}) \tag{2}$$

where $Freq_{(d,t)}$ represents the rescaled occurrence for a POI type $t$ given a search region $d$; and $V_{(d,t,i)}$ is the number of unique users who have contributed their check-ins for a venue $i$ that belongs to the same POI type $t$ in the same search region $d$.

We then test whether such an unsupervised popularity-based LDA topic modeling technique can support the discovery of characteristic semantic regions across different U.S. cities with a similar structure of POI type mix distribution.

Finding the appropriate number for latent topics is important but difficult given a dataset using the LDA topic model. Several metrics and methods have been developed to address this issue. Griffiths and Steyvers (2004) used the Gibbs sampling algorithm to obtain samples from the posterior distribution over topic assignments $Z$ at different choices of the total $K$ number of topics, and then calculated a log-likelihood $P(W|Z,k)$. The value of $K$ at which the log-likelihood gets the maximum and stabilizes after hundreds of iterations will be taken as the right number of topics for a specific document corpus. With the consideration of one issue that sometimes words have too many overlaps across those generated latent topics, Cao et al. (2009) proposed a density-based method for adaptive LDA model selection. The key idea of this algorithm is to maximize the intra-topic similarity while minimize the inter-topic similarity. They calculated the average cosine distance between pairwise topics with their word assignments and then used a heuristic to find the most stable topic structure given the best $K$ value based on the topic density measure. Arun et al. (2010) viewed the LDA topic model as a matrix factorization mechanism and applied the symmetric Kullback−Leibler (KL) divergence (Kullback and Leibler, 1951) on the distributions generated from topic-word and document-topic matrices for finding the right number of topics. The best $K$ value at which the symmetric KL-divergence is the minimum would derive the most discriminative topics and their distributions become orthogonal. In several empirical geographic information studies (Adams and Janowicz, 2015; McKenzie et al., 2015; Gao et al., 2017), different $K$ numbers (e.g., 60, 100, 300) of topics have been deployed to investigate place characteristics. An optimal value of $K$ may vary between different datasets and has influence on the thematic similarity of POI types (McKenzie and Janowicz, 2017).

## 4.2 Functional Region Aggregation

After deriving those latent thematic topics by running the proposed popularity-based LDA model, each region can be represented as a vector of the $K$-dimensional POI type topics. Those regions that are semantically similar in the topic space might contribute to the same urban function and can be aggregated into the same cluster as a functional region. Two clustering approaches are applied in this work: *K-means clustering* and *the Delaunay triangulation spatial constraints clustering methods.*

*K-means clustering* only takes the thematic characteristics of multivariate topic distributions of places into consideration without any spatial constraints (MacQueen et al., 1967). It is an unsupervised clustering approach in which the number $K$ needs to be predefined. The *silhouette* criterion has been widely used for determining an appropriate value of $K$ (Rousseeuw, 1987). The *silhouette* value $s(i)$ quantifies how well an object $i$ is appropriately clustered. The range of *silhouette* value is between -1 and 1. A high $s(i)$ value (close to 1) indicates that an object is appropriately clustered and is very dissimilar from other clusters. In the region clustering process for our POI datasets, we tried different $K$ values ranging from 1 to 30 and identified the maximum average silhouette value across all clusters and chose that $K$ as the optimal K-means clustering parameter for reporting the corresponding results.

*Delaunay triangulation spatial constraints clustering* has been introduced by Assunção et al. (2006). This approach consists of three steps: (1) building a connectivity graph to capture the adjacency relations between points based on Delaunay triangulation spatial constraints; (2) creating a minimum spanning tree (MST) (Gower and Ross, 1969) from the neighboring connectivity graph with minimizing the sum of the dissimilarities over all the edges of the tree; (3) partitioning the derived MST into different subtrees as spatial clusters using a hierarchical division strategy to minimize the intra-cluster square deviations. More implementation details about this clustering algorithm can be found at Assunção et al. (2006).

# 5 Analysis and Results

## 5.1 Topic Modeling Results

As proposed in Section 4, before running the LDA model, we first incorporated the number of visitors for each venue as a popularity score in the rescaling process to generate a new document-word matrix (i.e., a search region−POI type occurrence matrix) across all search regions in the ten urban areas. Next, we evaluated the performance of different choices of $K$ as the total number of topics for the LDA topic model using three introduced measures. As shown in Figure 4, by choosing the value of K from 5 to 200 and then running LDA topic models on our POI data, we derived different topic assignment results. The measure proposed by Griffiths and Steyvers (2004) aims to maximize the log-likelihood of word-topic probability in the documents, while other two measures (Cao et al., 2009; Arun et al., 2010) aim to minimize the proposed criteria. In the ideal case, one would expect those three measures converge at the same value of $K$. Unfortunately, in empirical studies, they do not necessarily present such perfect convergence patterns. In our parameter tuning experiments, the optimal $K$ value for the "CaoJuan2009" and "Arun2010" measures is in the range of $90 - 160$, while the "Griffiths2004" metric gets relatively stable when $K$ reaches 130 topics.

Therefore, we set $K = 130$ as the total number of topics and ran 2000 iterations of the Gibbs sampling process to derive the posterior probabilistic distribution over topic assignments. In Figure 5, we show nine of those inter-

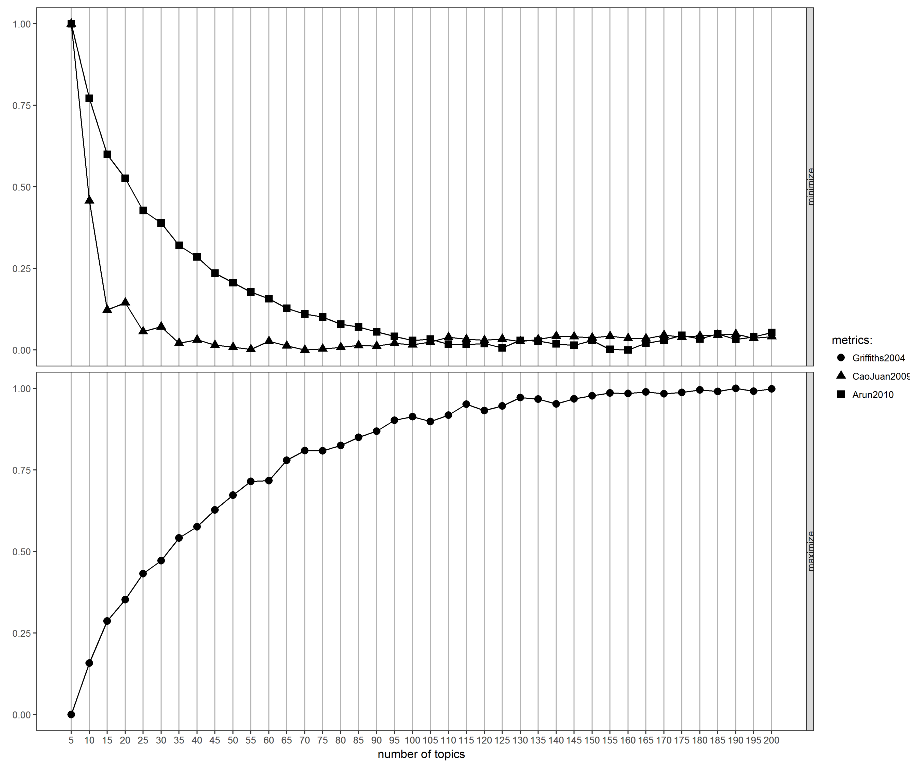


Figure 4: Find an appropriate K value for the total number of topics using three metrics.

esting topics related to urban functions. Note that the probability assignments for those POI types are weighted and ranked by their *term frequency−inverse document frequency* (i.e., POI type frequency−inverse region frequency) so that each topic can display more distinctive and meaningful POI types that are directly proportional to the frequency in documents while inversely proportional to the region frequency at which a POI type occurs in the whole corpus. For instance, *coffee shop* has a very high frequency but also widely exists in most of the regions in our POI data so it plays a less important role than other categories (e.g., *theme park*) in distinguishing the function of a region. Some of those meaningful topics are illustrated as follows.

*Topic 67* is a shopping-plaza topic that consists of various frequent occurring POI types including *shopping mall*, *accessories store*, *chocolate shop*, and a few restaurants. It is one of the most prominent topics across all cities. *Topic 109* is a beach-related topic that consists of *beach, surf spot, island, beach bar, pier, and so on.* In terms of spatial distribution, one would expect such topics should be located in coastal or lake-side cities only. Both *Topic 21* and *Topic 25* contain *history museum* and *art museum*, but *Topic 21* is more related to college/university regions since it also contains *pool, college rec center, tourist information center*, and several other educational facilities; while *Topic 25* is more likely an art museum district since it also consists of *art gallery, antique shop, scenic lookout, and so on. Topics 6, 117, and 119* relate to outdoor sports and leisure activity places, such as *national park, ski lodge, gym, golf course, tennis court*, and various restaurants and studios. *Topic 36* and *Topic 74* describe mix distribution patterns of *bar, restaurant, government building, residential apartment, and business service*, which may suggest central-city areas.

In order to explore the variability of the above discovered nine topics while changing the total number of topics, we further investigate whether we can find exactly matching or most similar topics with different values for $K$. Two evaluation criteria, namely *Cosine Similarity* and *Jaccard Index*, were applied for this purpose (Han et al., 2011). Assume that each topic vector is a sequence of probabilistic values between 0 and 1 for all the 480 POI categories. Considering each pair of one target topic (e.g., Topic 6 when $K$=130) and another one comparing topic (e.g., Topic 1 when $K$=10), the cosine similarity measures the cosine of the angle between two non-zero vectors defined using an inner product. It is well suited to evaluate sparse vectors such as document-word matrices and the topic-POI matrices in our experiments. Unlike the cosine similarity that is frequently used for numeric vectors, the Jaccard index is a popular similarity measure for binary and categorical data, which is defined as the cardinality of the intersection divided by the cardinality of the union of two sets. We use the Jaccard index to quantify the topic structure similarity for their top-fifteen probabilistically ranked POI types. The larger the value, the more similar two topics are, where 1 equals a perfect match while 0 indicates no overlapping top-terms (i.e., POI types) in the comparison of two topics. The comparison batch processing was conducted from $K$=10 to $K$=150 with a step of 10. During each run with a given $K$, the maximum similarity values to each of the nine topics were computed. As shown in Figure 6, the maximum cosine similarities for *Topics 74 and 117* reach almost 1 and remained stable when the total number of topics exceeded 30. As for *Topics 6, 36, 67, and 109*, we can also identify most semantically similar ($\geq$0.9) topics to them with $K$ value equals to 150, 120, 150 or 140 respectively. This indicates the stability of identifying those prominent urban functional topics related to frequently co-occurrent physical facilities and services, a variety of bars and restaurants, and leisure activity places. However, we cannot find very similar ($\geq$0.8) topics to *Topics 21, 25, and 119* when choosing different $K$ values, which implies that these topics may be more characteristic of a specific $K$ value. In a similar manner, we analyze the topic structure similarity using the Jaccard index. As shown in Figure 7, those low similarity values illustrate the large composition variability existed in the top-fifteen probabilistically ranked POI types for all discovered topics with different $K$ values. Their implications will be discussed in the section 6.

In short, rather than a traditional top-down approach for describing urban functions based on familiar compositions of POI types, we demonstrated a bottom-up statistical topic-learning approach for finding underlying co-occurrence relationships among different types of POIs based on data on human activities extracted from location-based social networks.

## 5.2 Searching for Similar Places

Searching for similar places is an important task in geographic information retrieval and also valuable in many applications, such as tourism, real estate, and immigration. People may consider many factors such as job market, affordability, natural environment, and quality of life. When people consider moving into new cities, they may also want to know how they will like these new places and whether they can find similar neighborhoods to the ones they will be leaving. Such places typically contain a mix of types of POIs that people would like to visit. Fortunately, such information can be retrieved from popular location-based social network platforms that have been used as a lens of social sensing to capture human-place interactions. In the following, we illustrate this idea with two scenarios:

(1) Search for similar regions given a dominant theme. We selected the city of Denver as our target city, which was ranked as the best metropolitan area to live in the U.S. according to a survey[2] from the U.S. News in 2016. It has a variety of local attractions and support many activities. Here we aim to find regions that are similar to those represented by *Topic 25*, which is related to art districts. We collected the Foursquare POIs and user check-in data for Denver by randomly sampling search locations and then searching for 50 nearby POIs given each sample location. Based on the aforementioned data processing procedures and the LDA topic model by incorporating the popularity score based on unique Foursquare check-in users, we can infer the probabilistic combination of different topics for a search region given its POI type co-occurrence pattern. As shown in Figure 8, within this search neighborhood, we discovered a high probabilistic topic distribution for *Topic 25*, which consists of a

[2] `http://realestate.usnews.com/places/rankings/best-places-to-live`

| Topic 6 | |
|---|---|
| **Category** | **Prob.** |
| gas station | 0.309651 |
| italian restaurant | 0.028574 |
| flower shop | 0.013235 |
| national park | 0.002077 |
| ski lodge | 0.000968 |
| jewish restaurant | 0.000899 |
| auditorium | 0.000847 |
| southern food | 0.000226 |
| ice cream shop | 0.000123 |
| farmers market | 0.000109 |
| club house | 0.000105 |
| bbq joint | 0.000100 |
| pizza place | 0.000100 |
| winery | 0.000072 |
| grocery store | 0.000059 |

| Topic 21 | |
|---|---|
| **Category** | **Prob.** |
| pool | 0.122166 |
| history museum | 0.047285 |
| historic site | 0.043474 |
| college basketball court | 0.015107 |
| concert hall | 0.012485 |
| art museum | 0.012412 |
| college rec center | 0.010488 |
| park | 0.007814 |
| sculpture garden | 0.007216 |
| outdoor sculpture | 0.005112 |
| college soccer field | 0.004028 |
| college cafeteria | 0.003933 |
| tourist information center | 0.003731 |
| molecular gastronomy | 0.003120 |
| stables | 0.002947 |

| Topic 25 | |
|---|---|
| **Category** | **Prob.** |
| museum | 0.065636 |
| art museum | 0.047585 |
| art gallery | 0.046691 |
| american restaurant | 0.038677 |
| record shop | 0.025063 |
| antique shop | 0.024183 |
| building | 0.002540 |
| cycle studio | 0.001103 |
| health food store | 0.000462 |
| history museum | 0.000430 |
| soup place | 0.000350 |
| concert hall | 0.000281 |
| scenic lookout | 0.000264 |
| animal shelter | 0.000179 |
| burger joint | 0.000179 |

| Topic 36 | |
|---|---|
| **Category** | **Prob.** |
| yoga studio | 0.105001 |
| science museum | 0.065819 |
| boutique | 0.029987 |
| gay bar | 0.015371 |
| sculpture garden | 0.012688 |
| government building | 0.008197 |
| israeli restaurant | 0.004401 |
| apartment / condo | 0.003005 |
| pakistani restaurant | 0.002829 |
| street food gathering | 0.001212 |
| track stadium | 0.000872 |
| college baseball diamond | 0.000602 |
| mexican restaurant | 0.000542 |
| gym / fitness center | 0.000526 |
| cheese shop | 0.000481 |

| Topic 67 | |
|---|---|
| **Category** | **Prob.** |
| shopping mall | 0.207709 |
| accessories store | 0.056738 |
| chocolate shop | 0.013896 |
| shoe store | 0.000288 |
| breakfast spot | 0.000282 |
| gaming cafe | 0.000196 |
| optical shop | 0.000180 |
| post office | 0.000114 |
| bistro | 0.000105 |
| dumpling restaurant | 0.000096 |
| korean restaurant | 0.000090 |
| german restaurant | 0.000080 |
| herbs & spices store | 0.000079 |
| airport terminal | 0.000078 |
| outlet store | 0.000076 |

| Topic 74 | |
|---|---|
| **Category** | **Prob.** |
| bar | 0.511221 |
| board shop | 0.000046 |
| asian restaurant | 0.000046 |
| brewery | 0.000036 |
| ice cream shop | 0.000030 |
| parking | 0.000025 |
| buffet | 0.000025 |
| business service | 0.000019 |
| apartment / condo | 0.000016 |
| fried chicken joint | 0.000012 |
| resort | 0.000007 |
| gourmet shop | 0.000007 |
| lighthouse | 0.000006 |
| indian restaurant | 0.000006 |
| train | 0.000006 |

| Topic 109 | |
|---|---|
| **Category** | **Prob.** |
| beach | 0.285864 |
| surf spot | 0.028952 |
| italian restaurant | 0.015458 |
| island | 0.005920 |
| beach bar | 0.004078 |
| board shop | 0.003793 |
| bridge | 0.001484 |
| indie theater | 0.001235 |
| pier | 0.001187 |
| outdoor sculpture | 0.001121 |
| sri lankan restaurant | 0.001040 |
| bistro | 0.000891 |
| nature preserve | 0.000851 |
| arepa restaurant | 0.000751 |
| neighborhood | 0.000726 |

| Topic 117 | |
|---|---|
| **Category** | **Prob.** |
| italian restaurant | 0.082055 |
| fast food restaurant | 0.000131 |
| gym | 0.000064 |
| golf course | 0.000056 |
| sushi restaurant | 0.000044 |
| salon / barbershop | 0.000043 |
| boutique | 0.000034 |
| café | 0.000030 |
| szechuan restaurant | 0.000030 |
| japanese restaurant | 0.000030 |
| paella restaurant | 0.000027 |
| men's store | 0.000023 |
| caribbean restaurant | 0.000017 |
| deli / bodega | 0.000016 |
| massage studio | 0.000014 |

| Topic 119 | |
|---|---|
| **Category** | **Prob.** |
| french restaurant | 0.090092 |
| cocktail bar | 0.072534 |
| lounge | 0.035774 |
| tennis court | 0.005389 |
| whisky bar | 0.003636 |
| american restaurant | 0.000697 |
| dry cleaner | 0.000168 |
| pizza place | 0.000118 |
| café | 0.000117 |
| art museum | 0.000110 |
| bakery | 0.000106 |
| jazz club | 0.000082 |
| chinese restaurant | 0.000074 |
| neighborhood | 0.000068 |
| cycle studio | 0.000040 |

Figure 5: Nine interesting topics with their top-15 ranked POI types related to urban functions.

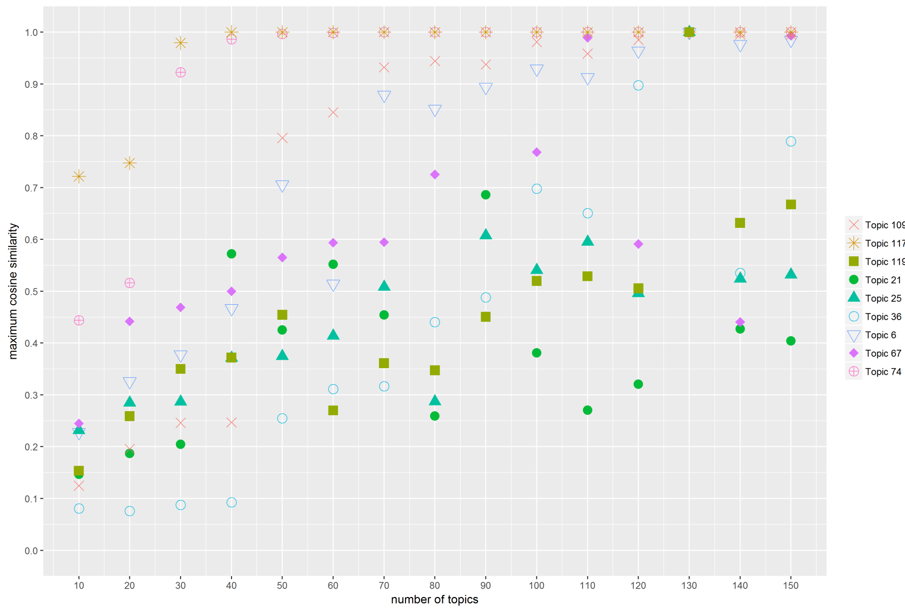


Figure 6: Maximum cosine similarity between the selected nine topics and the resulting topic models by choosing different total number of topics.

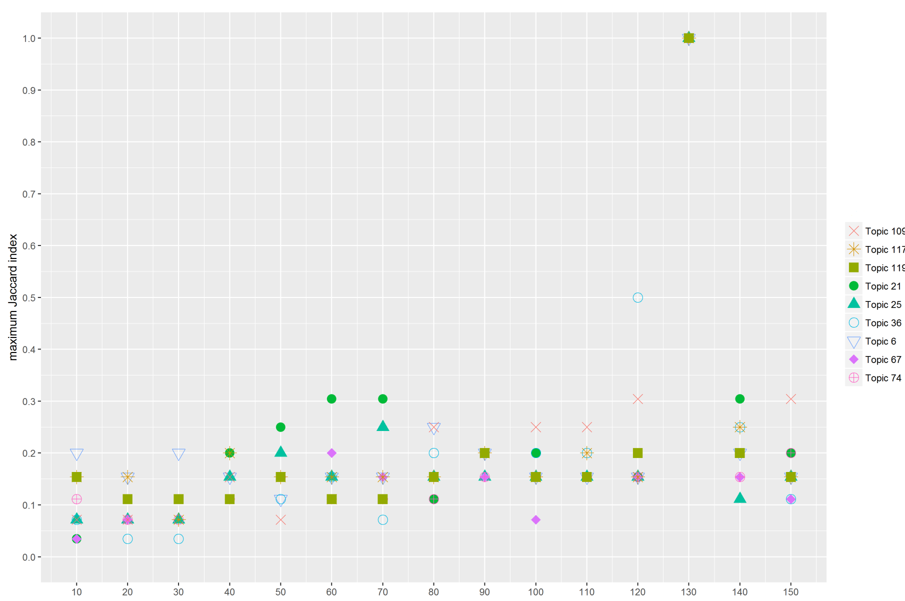


Figure 7: Maximum Jaccard similarity between the top-15 POI types of the selected nine topics and that from the resulting topic models by choosing different total number of topics.

variety of prominent POI types such as *art museum, art gallery, history museum, concert hall and American restaurant*. Such a place may serve multiple functions. The second largest probabilistic topic in this search neighborhood is *Topic 121* that contains a large percentage composition of *brewery places*. By looking up other geographic background information and Web pages, we realize that local people actually also identify this region near the *"Santa Fe Dr."* as an "Art District" in Denver, which attracts many local residents, artists and tourists[3]. This example illustrates the inference capability of our method to identify similar neighborhoods given certain thematic characteristics.

(2) Search for similar places considering all themes. After running the LDA topic model, each place can be represented as a multinomial distribution of $K$-dimensional POI type topics, denoted as a probability vector $[p_1, p_2, \ldots, p_k]$, where all the probability values sum to one. Thus we can apply a variety of probabilistic distance or similarity measures (e.g., Hellinger distance, cosine similarity, and Jensen-Shannon divergence (JSD) (Lin, 1991) to quantify the pairwise similarity among all search regions in our POI data with regard to their POI type mix distributions. JSD is a symmetric distance measure derived from the Kullback-Leibler divergence (KLD) asymmetric distance measure between two probability distributions $P$ and $Q$ (Kullback and Leibler, 1951).

$$KLD(P|Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \tag{3}$$

$$M = \frac{P+Q}{2} \tag{4}$$

$$JSD(P|Q) = JSD(Q|P) = \frac{KLD(P|M)}{2} + \frac{KLD(Q|M)}{2} \tag{5}$$

The JSD is bounded by 0 and 1 if using the base 2 logarithm for the two

[3] `http://www.denver.org/about-denver/neighborhood-guides/artdistrict-on-santa-fe`

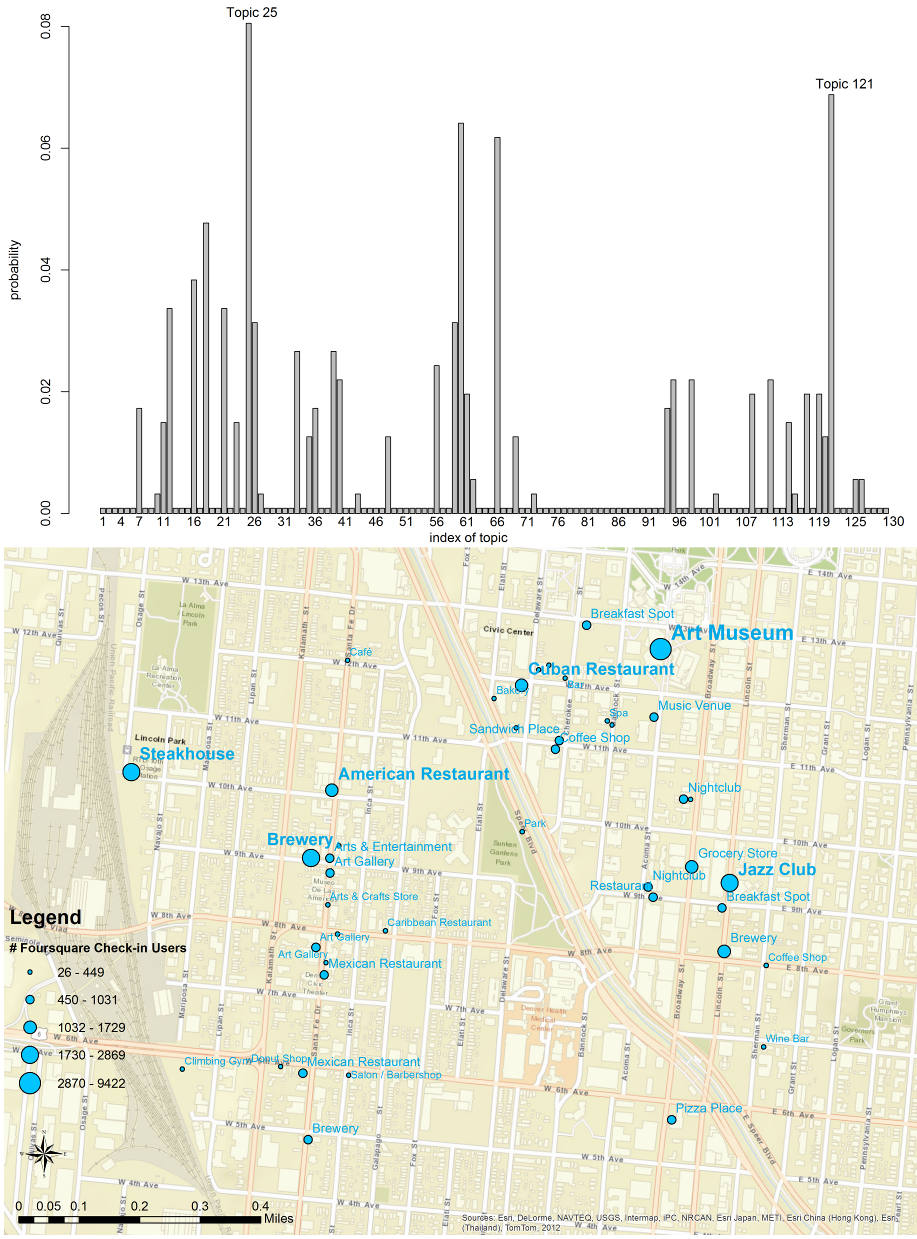


Figure 8: The topic probability distribution and the spatial distribution of Foursquare POIs around the Denver art district and museums.

KLD relative entropy calculation. And thus we can define a JSD-similarity metric ($S_{(JSD)}$) as follows:

$$S_{(JSD)} = 1 - JSD(Q|P) \tag{6}$$

where the base 2 logarithm is used in the KLD and JSD calculations.

Therefore, according to the proposed similarity measure $S_{(JSD)}$, we can analyze the pairwise similarity among our randomly selected search places that contains those POIs. Figure 9 shows a JSD-similarity matrix for 200 randomly selected places in Los Angeles, derived from one part of our whole dataset in Section 3, and each place is represented as a vector of 130-dimensional thematic topics. The similarity score in each grid is between 0 and 1. The higher the value, the more similar the two places are with regard to their topic distributions. The values in the diagonal are all 1. By visualizing this similarity matrix, one can easily identify two anomalous red stripes (i.e., the labeled $R_{th}$ row and the $C_{th}$ column) with relatively low similarity values across the grid cells. Interestingly, as shown in Figure 10, further investigation reveals that this place was sampled at a location inside the *Disneyland Resort* in the Los Angeles metropolitan area, which is very different from all other randomly sampled places and the dominant topic (*Topic 56*) has unusual POI types such as *theme park, theme park ride/attraction, and gift shop.* The frequent co-occurrence of those distinctive types of POIs in this region causes the very low similarity to all other places. Thus, given any place, we can find the most similar or dissimilar places in another geographic region based on this similarity matrix.

## 5.3 Discovering Functional Regions

Another goal for us is to discover urban functional regions where semantically similar places group together. As described in Section 4, two clustering meth-

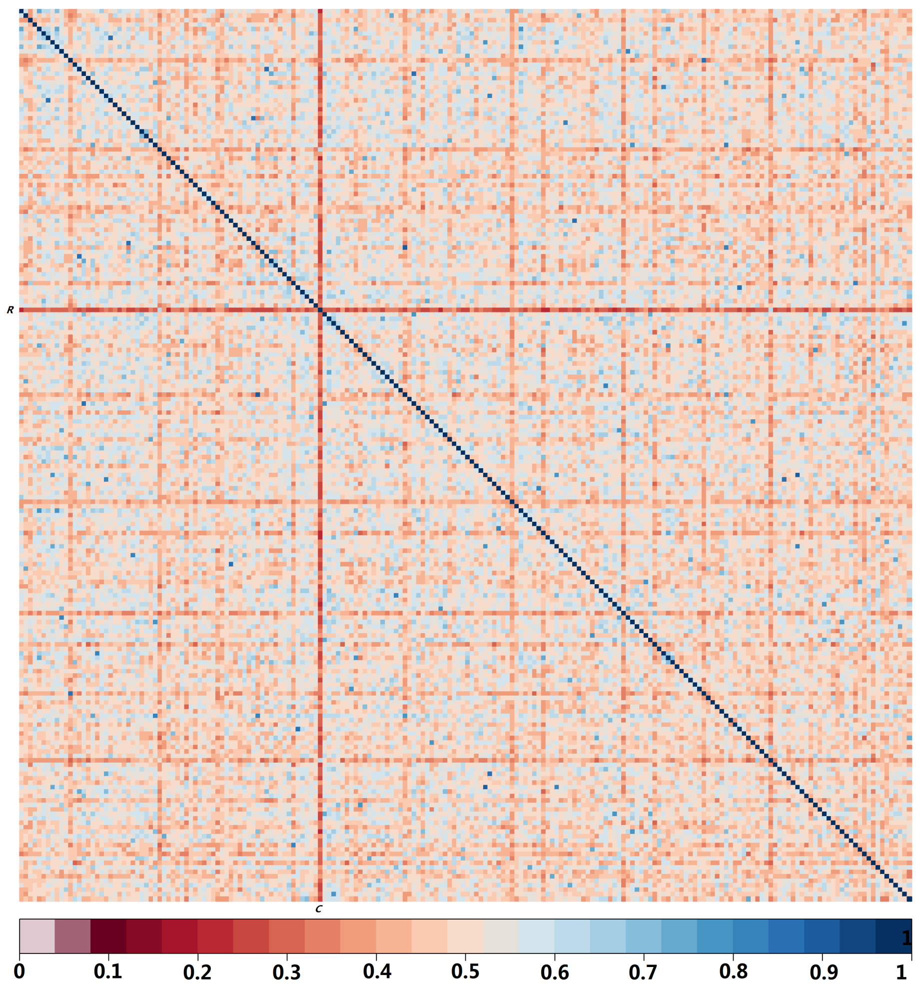


Figure 9: The JSD-similarity matrix for 200 randomly selected places in Los Angeles, where each place consists of 130 dimensional thematic topics.

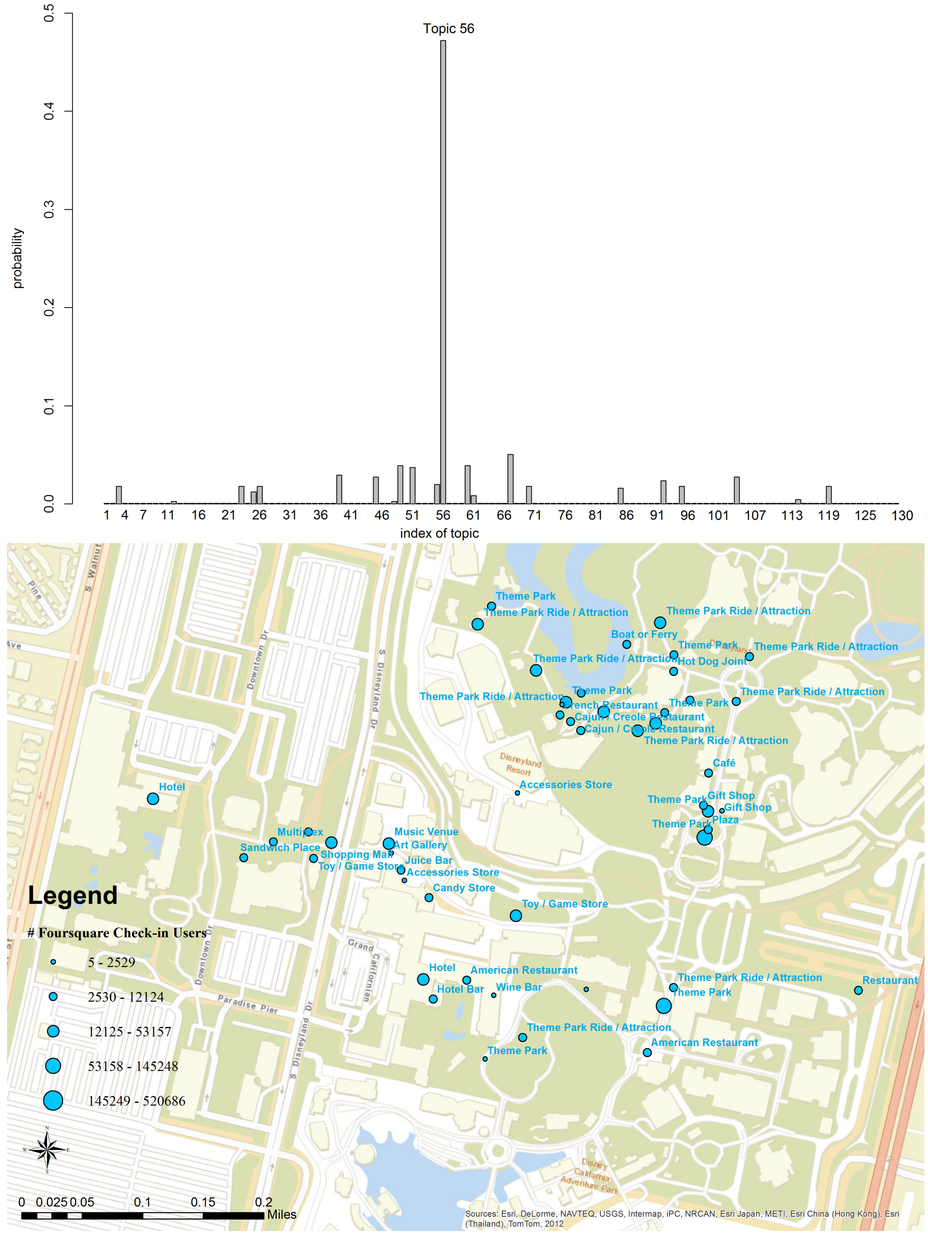


Figure 10: The topic probability distribution and the spatial distribution of Foursquare POIs in the Los Angeles Disneyland Resort.

ods are applied for aggregating similar places into functional regions. Figure 11 shows the K-means clustering result for 200 randomly sampled places in Los Angeles. The *silhouette* value for determining the optimal number of clusters in Los Angeles is 15, and thus we group those places into fifteen clusters. The circles with the same color on the map belong to the same cluster within which POI structures are more similar in their topic space of types. The numeric label on the top of each circle displays the top ranked topic that has the largest probability over the 130-dimensional thematic topic vector in this location. Note that purely K-means clustering doesn't consider any spatial constraints, and thus distant places sharing similar functions or thematic characteristics can also be grouped into the same cluster. For example, several places are dominated by food-related *Topic 30*, which contains frequent distributions of various restaurants such as *Korean restaurant, Mongolian restaurant, Portuguese restaurant, and Polish restaurant* are grouped into *Cluster 8*, although those places are spatially separated. However, if we take spatial constraints into considerations, only places that are semantically similar in the topic space and also located near each other can be aggregated into the same cluster. Figure 12 shows the Delaunay-triangulation-spatial-constraints clustering result for those 200 sampled places in Los Angeles. Note that we keep the same color scheme for the visualization of two clustering results, but those clusters in the same color from two maps are not identical. In the West Coast area, we can see that several places are dominated by the beach *Topic 109* and related leisure activity categories are spatially clustered together into *Cluster 7*. Although *Cluster 7* and *Cluster 3* are spatially close and share beach characteristic, *Cluster 3* tends to have another dominant POI type (*shopping plaza*) in this region, which distinguishes it from *Cluster 7 and Cluster 10.*

By analyzing the spatial distribution of similar places and clusters, researchers can have a better understanding of how certain types of POIs co-locate in order to serve different urban functions from the bottom-up perspective. In addition, urban planners or managers are able to further investigate the needs for complementary physical facilities and services related to the thematic characteristics derived from the human activities on the location-based social networks. This keeps in line with the human-centered and community-oriented perspectives in traditional top-down urban planning and design. Furthermore, we create the bounded functional regions as *convex polygons* derived from those points in the same cluster (Figure 12). This can help geographic information service providers develop topic-related POI search services within certain functional regions. Because the POI type assignments for all topics are semantically interpretable, we can also select multiple dimensions of topics in geographic information queries such as the *beach + shopping plaza* topics. *Cluster 3* (in Figure 12) would be a good candidate since it has a mix of the dominant beach topic and the shopping topic.

In addition, in order to test the robustness of discovered urban functional areas with different probabilistic topics, we perform a series of clustering result comparisons by choosing different numbers of topics ranging from 10 to 150. We use their corresponding probabilistic POI type compositions as clustering features and run both K-means clustering and the Delaunay-triangulation-spatial-constraints clustering. Two popular metrics for comparing clustering results are applied in our tests: the *Rand index* (Rand, 1971) and *normalized mutual information (NMI)* (Strehl and Ghosh, 2002). The Rand index measures the percentage of decision agreements between two clustering results X and Y. It contains two types of decision agreements: (1) the number of pairs of search locations within the same clustering region in X that are also in the same clustering region in Y; (2) the number of pairs of search locations that are in different clustering regions in X and also in different clustering regions in Y. The NMI quantifies the mutual dependence/similarity between two clustering results using information theory. The detailed formula descriptions can be found in the original article (Strehl and Ghosh, 2002). For both the Rand index and NMI, their value range is between 0 and 1 and larger values indicate higher similarity between two clustering results. Figure 13 shows the K-means clustering comparisons using the Rand index and the NMI metric between the target scenario (130 topics and 15 clusters) and other scenarios with different number of topics but with the same total number of clusters. Figure 14 shows the comparison results for the Delaunay-triangulation-spatial-constraints clustering in a similar manner. We find that the Rand index keeps a high value around 0.85 for both clustering methods, which indicates a large percentage of agreements on the clustering membership of those search locations and derived functional areas. But the NMI values show a fluttering pattern that indicates the existence of cluster membership variability. Furthermore, as for both evaluation metrics, the Delaunay-triangulation-spatial-constraints clustering has a higher similarity value in most comparison scenarios and seems to be more stable than the K-means clustering results. It may imply that the spatial constraints play a role in deriving the functional regions.

# 6 Broader Implications and Discussions

Based on the analysis results in this research, we show that several latent topics of POI categories are spatially and semantically related to certain urban functions. For example, the college/university topic that consists of *buildings, pool, sports fields, and apartments*, is also co-located with several *restaurant* and *bar* like topics; the *shopping plaza* topic is often also co-located with the *parking* and *resort* like topics. This reveals the underlying relations of how POI cate-

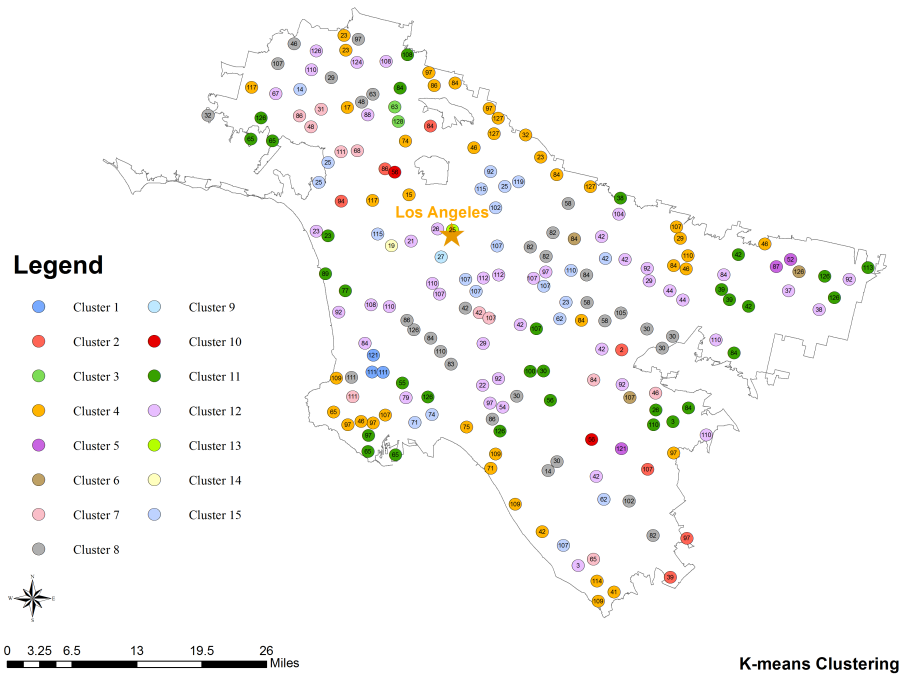


Figure 11: The K-means clustering result for 200 sampled places in Los Angeles.

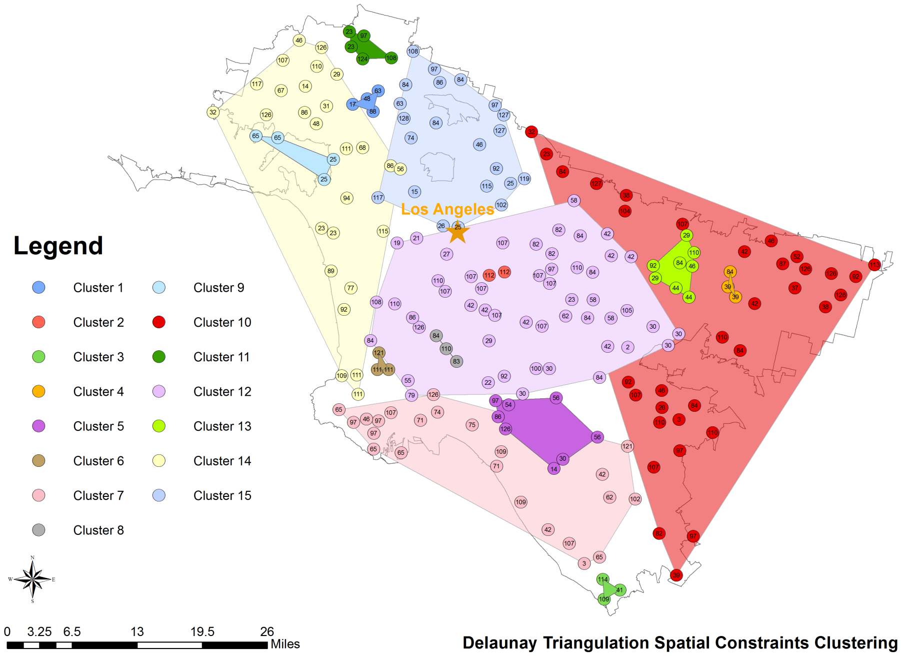


Figure 12: The Delaunay triangulation spatial constraints clustering and convex polygon generation result for 200 sampled places in Los Angeles.

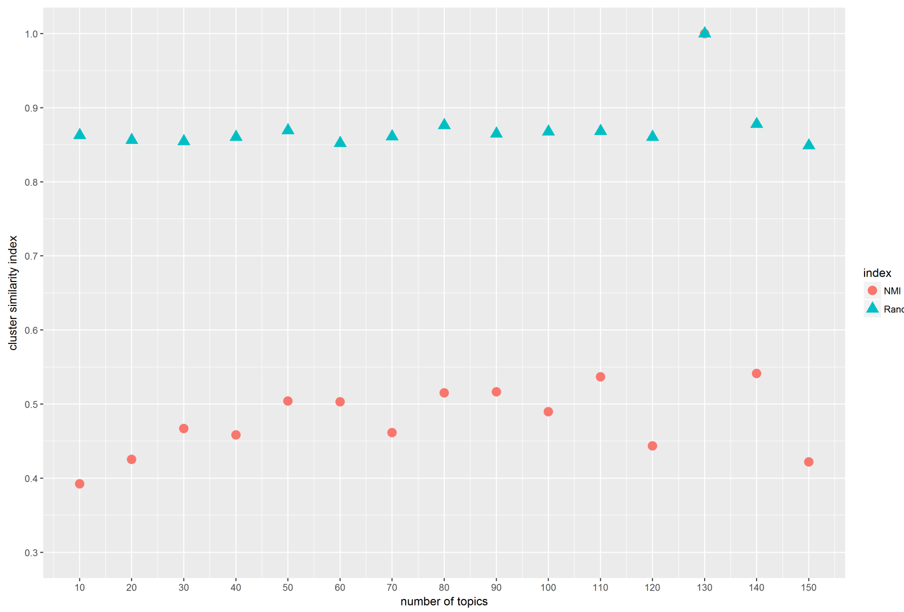


Figure 13: K-means clustering similarity evaluation using the NMI and Rand metrics with different number of topics.

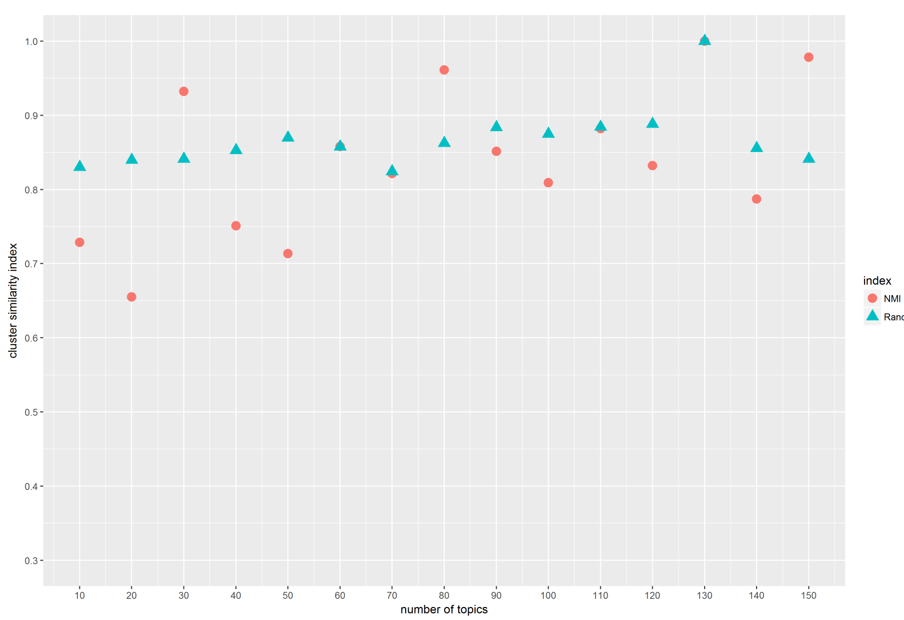


Figure 14: Delaunay triangulation spatial constraints clustering similarity evaluation using the NMI and Rand metrics with different number of topics.

gories function in geographic settings. Occasionally a topic may contain a less meaningful or even an outlier POI category such as the *neighborhood* in *topic 109*. Data cleaning or post-processing may help to eliminate or reduce the noise. This is the nature of crowdsourced data or user-generated content in which the information is not validated by any authority. Also, the coverage and the accuracy of POI data in different cities may vary and the POI categories might also change over time. We need to pay attention to those issues when interpreting findings. In addition, we have also discovered various urban functional regions as clusters of multinomial topic distributions over POI categories. However, one limitation is that we cannot systematically evaluate the accuracy of those derived functional regions without labeled ground truth data or the detailed urban land-use GIS data. But we can test the intrinsic robustness of identifying functional topics with different parameter settings. The variability analyses were carried out at two levels: *the topic-level* and *the cluster-level*. At the topic level, we found the stability in identifying prominent urban functional topics related to frequently co-occurrent physical facilities and services, a variety of bars and restaurants, and leisure activity places regardless of the total number of topics. But the topic composition of top-ranked POI categories varies in different scenarios. It implies the variability of the semantic structure of functional topics. Although choosing an optimal $K$ in topic modeling can either maximize the log-likelihood of the term-topic probability in the training document corpus, or minimize the inter-topic similarity, we may miss the opportunity for discovering some interesting topic composition structures that can only be identified with a different $K$ value or with other model parameter settings. At the cluster level, a series of clustering result comparisons by choosing different numbers of topics were evaluated using the Rand index and the NMI metric. We found a large percentage of agreements on the clustering membership of those search locations with their surrounding POIs and derived functional areas that can be supported by the mix types of POIs.

One broader question is whether we can automatically identify those topological and hierarchical relations in order to support the development of an ontology for urban functional regions. As shown in Figure 15, we applied the Ward hierarchical clustering method (Ward Jr, 1963) on those 130-topics derived from the aforementioned LDA topic model. Each topic is a 480-dimensional probabilistic vector over all POI categories in our datasets. Those semantically related topics are grouped together in each step by minimizing the increment of within-cluster variance after merging. This process repeats until all topic vectors merge into the same group. This tree diagram is derived from a bottom-up approach and can be used as a starting point with regard to constructing the urban functional region ontology. However, it does not yet include the spatial relationships nor the dichotomous relationship among POIs. It may be more promising to combine this bottom-up approach with the top-down approach of the expert urban geographers or planners to develop a more holistic ontology in the future.

# 7 Conclusions and Future Work

In this work, we develop a statistical framework that applies the LDA topic modeling technique and incorporates user check-ins on LBSN in order to help discover semantically meaningful topics and functional regions based on co-occurrence patterns of POI types. The "functions" derived from probabilistic topic modeling techniques can reveal the latent structure of POI mixtures and the semantics of places. Based on a large corpus of about 100,000 Foursquare venues and check-in behavior in the ten most populated urban areas in the U.S., we demonstrate the effectiveness of proposed methodology by identifying distinctive types of latent topics and further, by extracting urban functional

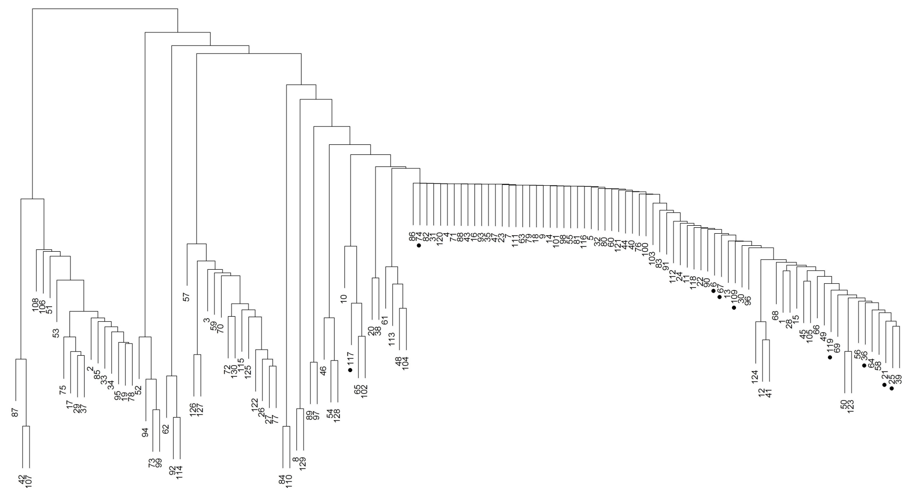

Figure 15: The dendrogram for the hierarchal clustering result on the 130-LDA topics using the Ward clustering method. The topics highlighted with a black filled-in circle are those mentioned in section 5.

regions using the K-means clustering and the Delaunay triangulation spatial constraints clustering methods. A region can have multiple functions but with different probabilities, while the same type of functional region can span multiple geographically non-adjacent locations. Compared with the remote sensing images that mainly uncover the physical landscape of urban environments, results derived from the popularity-based POI topic model can be seen as a complementary social sensing view of urban space based on human activities and the place settings of urban functions. However, there may exist gaps between the real-world business establishments and the online available POI information. Data-fusion and cross-validation relying on multiple sources may help reduce such gaps.

Although we have successfully identified several types of semantically meaningful urban functional topics, LDA topic modeling is an unsupervised approach that has certain limitation with respect to discovering plausible urban functions. In the future, we plan to investigate additional semantic signatures such as those incorporating the spatial patterns of POI distributions and using supervised-versions of probabilistic topic models to compare the performance of two families of topic models (unsupervised or supervised) in discovering urban functional regions. Last but not least, we also aim at developing a functional region ontology by combing the data-driven approach as outlined in this work with the top-down knowledge engineering approach based on our understanding of urban functional regions from human geography and urban planning.


# Acknowledgement

We would like to thank Gengchen Mai for his help and discussion on the evaluation of different clustering results. We also want to thank the editor and three anonymous referees for their valuable comments and suggestions.


# References


Adams, B., 2015. Finding similar places using the observation-to-generalization place model. Journal of Geographical Systems 17 (2), 137–156.

Adams, B., Janowicz, K., 2012. On the geo-indicativeness of non-georeferenced text. In: ICWSM. pp. 375–378.

Adams, B., Janowicz, K., 2015. Thematic signatures for cleansing and enriching place-related linked data. International Journal of Geographical Information Science 29 (4), 556–579.

Adams, B., McKenzie, G., 2013. Inferring thematic places from spatially referenced natural language descriptions. In: Crowdsourcing Geographic Knowledge. Springer, pp. 201–221.

Arun, R., Suresh, V., Madhavan, C. V., Murthy, M. N., 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp. 391–402.

Assunção, R. M., Neves, M. C., Câmara, G., da Costa Freitas, C., 2006. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. International Journal of Geographical Information Science 20 (7), 797–811.

Banzhaf, E., Netzband, M., 2012. Monitoring urban land use changes with remote sensing techniques. Applied Urban Ecology: A Global Framework , 18–32.

Barnsley, M. J., Barr, S. L., 1996. Inferring urban land use from satellite sensor images using kernel-based spatial reclassification. Photogrammetric Engineering and Remote Sensing 62 (8), 949–958.

Blei, D. M., 2012. Probabilistic topic models. Communications of the ACM 55 (4), 77–84.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3 (Jan), 993–1022.

Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S., 2009. A density-based method for adaptive lda model selection. Neurocomputing 72 (7), 1775–1781.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., Ju, Y., Gong, L., Adams, B., Yan, B., 2017. A data-synthesis-driven method for detecting and extracting vague cognitive regions. International Journal of Geographical Information Science , 1245–1271.

Gower, J. C., Ross, G., 1969. Minimum spanning trees and single linkage cluster analysis. Applied Statistics , 54–64.

Griffiths, T. L., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences 101 (suppl 1), 5228–5235.

Han, J., Kamber, M., Pei, J., 2011. Data mining: concepts and techniques (Third Edition). Elsevier, Waltham, MA.

Herold, M., Couclelis, H., Clarke, K. C., 2005. The role of spatial metrics in the analysis and modeling of urban land use change. Computers, Environment and Urban Systems 29 (4), 369–399.

Hobel, H., Abdalla, A., Fogliaroni, P., Frank, A. U., 2015. A semantic region growing algorithm: extraction of urban settings. In: AGILE 2015. Springer, pp. 19–33.

Hobel, H., Fogliaroni, P., Frank, A. U., 2016. Deriving the geographic footprint of cognitive regions. In: Geospatial Data in a Changing World. Springer, pp. 67–84.

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., Prasad, S., 2015. Extracting and understanding urban areas of interest using geotagged photos. Computers, Environment and Urban Systems 54, 240–254.

Janowicz, K., 2012. Observation-driven geo-ontology engineering. Transactions in GIS 16 (3), 351–374.

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J., Pereira, F. C., 2015. Mining point-of-interest data from social networks for urban land use classification and disaggregation. Computers, Environment and Urban Systems 53, 36–46.

Kullback, S., Leibler, R. A., 1951. On information and sufficiency. The Annals of Mathematical Statistics 22 (1), 79–86.

Lin, J., 1991. Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory 37 (1), 145–151.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., Shi, L., 2015. Social sensing: A new approach to understanding our socioeconomic environments. Annals of the Association of American Geographers 105 (3), 512–530.

MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. Oakland, CA, USA., pp. 281–297.

McKenzie, G., Janowicz, K., 2017. The effect of regional variation and resolution on geosocial thematic signatures for points of interest. In: Proceedings of the 2017 AGILE Conference. Springer, pp. 237–256.

McKenzie, G., Janowicz, K., Gao, S., Yang, J.-A., Hu, Y., 2015. Poi pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data. Cartographica: The International Journal for Geographic Information and Geovisualization 50 (2), 71–85.

Noulas, A., Scellato, S., Mascolo, C., Pontil, M., 2011. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. The Social Mobile Web 11, 02.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., Zhou, C., 2014. A new insight into land use classification based on aggregated mobile phone data. International Journal of Geographical Information Science 28 (9), 1988–2007.

Rand, W. M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association 66 (336), 846–850.

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.

Steiger, E., Westerholt, R., Zipf, A., 2016. Research on social media feeds–a giscience perspective. European Handbook of Crowdsourced Geographic Information , 237.

Steyvers, M., Griffiths, T., 2007. Probabilistic topic models. Handbook of Latent Semantic Analysis 427 (7), 424–440.

Strehl, A., Ghosh, J., 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3 (Dec), 583–617.

Ward Jr, J. H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American statistical association 58 (301), 236–244.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., Mai, K., 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. International Journal of Geographical Information Science 31 (4), 825–848.

Yuan, J., Zheng, Y., 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. ACM, pp. 186–194.

Zhi, Y., Li, H., Wang, D., Deng, M., Wang, S., Gao, J., Duan, Z., Liu, Y., 2016. Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data. Geo-spatial Information Science 19 (2), 94–105.

Zhong, C., Huang, X., Arisona, S. M., Schmitt, G., Batty, M., 2014. Inferring building functions from a probabilistic model using public transportation data. Computers, Environment and Urban Systems 48, 124–137.

Zhou, X., Zhang, L., 2016. Crowdsourcing functions of the living city from twitter and foursquare data. Cartography and Geographic Information Science 43 (5), 393–404.