CoPFun: an urban co-occurrence pattern mining scheme based on regional function discovery

Article *in* World Wide Web · May 2018

DOI: 10.1007/s11280-018-0578-x

CITATION
1

READS
33

6 authors, including:

Xiangjie Kong
Dalian University of Technology
95 PUBLICATIONS **844** CITATIONS
SEE PROFILE

Jianxin Li
Swinburne University of Technology
38 PUBLICATIONS **379** CITATIONS
SEE PROFILE

Feng Xia
Dalian University of Technology
255 PUBLICATIONS **3,877** CITATIONS
SEE PROFILE

World Wide Web
https://doi.org/10.1007/s11280-018-0578-x

CoPFun: an urban co-occurrence pattern mining scheme based on regional function discovery

Xiangjie Kong[1] · Menglin Li[1] · Jianxin Li[2] · Kaiqi Tian[1] · Xiping Hu[3] · Feng Xia[1]

Received: 31 October 2017 / Revised: 7 March 2018 / Accepted: 24 April 2018

Abstract Analysis of mobile big data enables smart cities from aspects of traffic pattern, human mobility, air quality, and so on. Co-occurrence pattern in human mobility has been proposed in recent years and sparked high attentions of academia and industry. Co-occurrence pattern has shown enormous values in aspects of urban planning, business, and social applications, such as shopping mall promotion strategy making, and contagious disease spreading. What's more, human mobility has strong relation with regional functions, because each urban region owns a major function to offer specialized services for city's operations and such location-based services attract massive passenger flow, which is exactly the essence of urban human mobility pattern. Therefore, in this paper, we put forward a co-occurrence pattern mining scheme (CoPFun) based on regional function discovery utilizing various mobile data. First, we do traffic modeling to map trajectory data into population groups, which include temporal partition and map segmentation. Then we employ a frequent pattern mining algorithm to mine co-occurrence event data. Meanwhile, we exploit TF-IDF method to process POI data and LDA algorithm to process trajectory data to discover urban regional functions. We apply CoPFun to real mobile data to extract co-occurrence event data and compare it with OD data to analyze urban co-occurrence pattern from a perspective of regional functions. The experiment results verify the effectiveness of CoPFun.

Keywords Co-occurrence pattern · Human mobility · Regional function · Smart cities

This article belongs to the Topical Collection: *Special Issue on Geo-Social Computing*
Guest Editors: Guandong Xu, Wen-Chih Peng, Hongzhi Yin, Zi (Helen) Huang

✉ Feng Xia
f.xia@ieee.org

[1] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China

[2] School of Computer Science and Software Engineering, University of Western Australia, Perth, Australia

[3] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

1 Introduction

With rapidly increasing availability of sensing technology and telecommunication infrastructure, huge volumes of mobile data tracking human mobility can be acquired, which are likely to pave the way for mobile big data-driven analysis of human mobility [19, 25]. Human mobility analysis is an essential part of mobile big data analysis and offers support for urban planning in terms of passenger flow prediction [8, 18], abnormal traffic events detection [23], and functional regions mining [11], etc. Co-occurrence pattern in human mobility denotes people from two regions visit an urban place during the same time span, which is of great significance due to its extensive applications with high value of urban planning, business, and social activity in modern society [30]. Take Figure 1 as an example of urban co-occurrence pattern. In the figure two persons from residential area A and four persons from residential area B visit the school at same time interval respectively so we can say there is an co-occurrence event between residential area A and residential area B. Based on the analysis of co-occurrence pattern in human mobility, we can capture the information of when and where people from the same functional regions prefer to co-occur and where people present at certain functional regions frequently come from. Furthermore, co-occurrence pattern has significant potential business value, such as shop owners can make well-targeted promotions by identifying which kind of people will visit shopping centers.

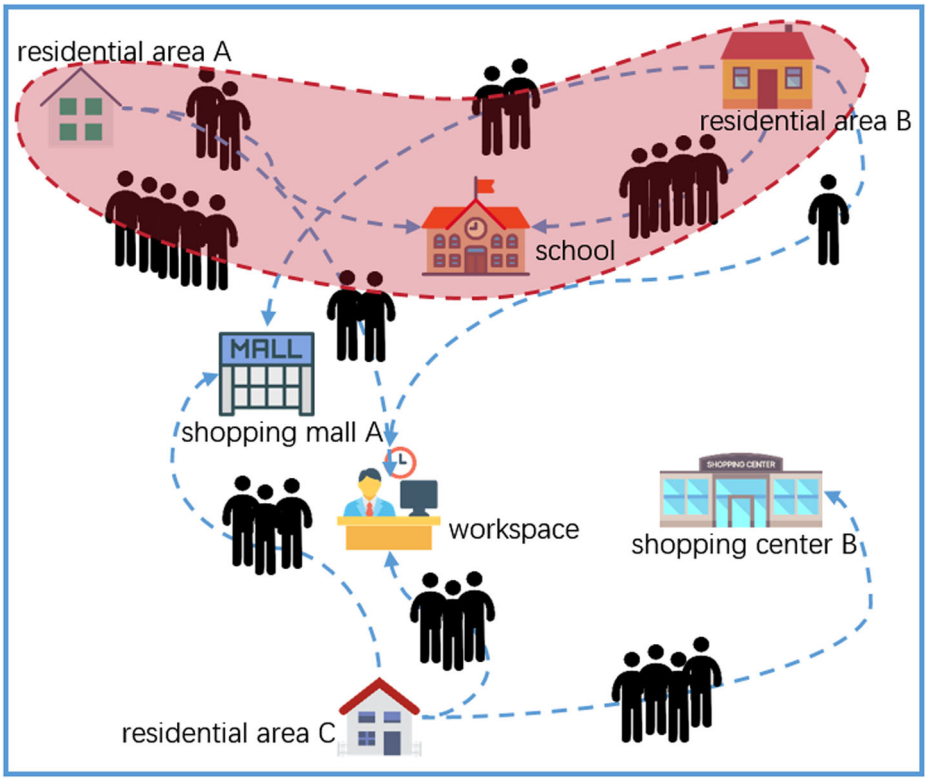


Figure 1 An example of co-occurrence pattern

For regional functions, two kinds of information play a crucial role in human mobility pattern analysis. The first is when people arrive at a region and when people leave a region. For example, people tend to leave from residential areas in the morning and go back home in the evening, while they are likely to go to entertainment only in daytime of non-working days and evening of working days. The second is which region people come from and which region people go to. For instance, people often go to entertainment from workspace (working days) or residential areas (non-working days). Therefore, if people come from similar regions, their arriving regions may also be similar. What's more, if people's destinations are similar, their origins are likely to be similar. People visit different functional regions based on their own needs, which generates urban human mobility pattern. That is, human mobility pattern is strongly related to regional functions [31]. Hence, we aim at analyzing urban co-occurrence pattern based on regional functions.

However, understanding co-occurrence pattern in human mobility is rather challenging. First, the deluge and the complexity of mobile big data make it difficult for digging up the value of the data [10]. Mobile data contains speed, fee, and even semantic information apart from temporal and spatial properties and abundant properties interfere with the extraction of truly valuable information. What's more, as the above example shows, co-occurrence pattern is a kind of complicated human mobility. Due to its intrinsic features, mining co-occurrence pattern is computationally highly expensive [6]. Furthermore, the set of candidate patterns is exponential in the number of co-occurrence events. The study of co-occurrence evolves explosively in various research fields, including microbial communities, gene mutation, scholars cooperation, and so on. Based on telco data, biclustering techniques are adopted to detect co-occurrence pattern and rich visualization forms are utilized to reduce the difficulty of analyzing co-occurrence pattern [30]. An infectious disease model on empirical networks of human contact is built to bridge the gap between dynamic network data and contact matrices [22]. In this paper, we creatively adopt frequent pattern algorithm to mine co-occurrence pattern and analyze it from the perspective of urban regional functions.

We propose a co-occurrence pattern mining scheme (CoPFun) to analyze urban co-occurrence pattern based on regional functions using various mobile data. We first do traffic modeling to map taxi trajectory data into different population groups, including temporal partition and map segmentation. Based on population group data, we extract co-occurrence event data utilizing frequent itemset mining algorithm. Meanwhile, we mine static functions from POI data using TF-IDF method, discover dynamic functions from traffic trajectories utilizing LDA algorithm, and obtain urban actual regional functions by combining the above results. Then we carry out experiments on real Shanghai mobile dataset to demonstrate the effectiveness of our method and do statistic analysis and visual analysis of urban co-occurrence pattern. To the best of our knowledge, CoPFun is the first to mine co-occurrence event data utilizing frequent pattern algorithm based on traffic trajectory data and to analyze co-occurrence pattern from a perspective of regional functions.

The major contributions of this work can be summarized as follows:

- We propose a co-occurrence event data mining scheme (CoPFun) utilizing frequent pattern algorithm based on traffic trajectory data.
- We discover urban actual regional functions by combining static functions extracted by TF-IDF method and dynamic functions mined by LDA algorithm.
- We evaluate our scheme using Shanghai taxi trajectories, road network, and POI data to demonstrate that CoPFun can extract co-occurrence event data effectively. We analyze urban co-occurrence pattern from a perspective of regional functions.

The rest of our paper is organized as follows. In Section 2, we review the related work about co-occurrence pattern mining and regional functions discovery. Section 3 presents details of the proposed scheme CoPFun. Data description and experiment results are displayed in Section 4. Following that, we show urban co-occurrence pattern analysis in Section 5. Finally, we conclude our work and give further discussions on open research issues in Section 6.

2 Related work

This section provides an overview of the related research work. We focus on two most relevant topics: co-occurrence pattern mining and regional functions discovery.

Co-occurrence pattern mining Co-occurrence pattern is a ubiquitous topic with high research value in various fields. As an interesting and significant pattern, co-occurrence pattern in the field of computer vision [27], biological symbiosis [9], mobile phone user application mode analysis [28] and many other aspects have been in-depth researched and domain experts have put forward targeted effective models and methods, apart from spatial-temporal co-occurrence in urban human mobility mainly covered in this paper. Spatial-temporal co-occurrence pattern is an essential issue with numerous applications.

However, co-occurrence pattern mining is computationally expensive and data set is over large, which cause great resistance of co-occurrence pattern analysis. Celik et al. propose a monotonic composite interest measure for discovering mixed-drove spatiotemporal co-occurrence pattern (MDCOP) and novel MDCOP mining algorithms to improve computational efficiency [6, 7]. Spatio-temporal co-occurrence pattern represents subsets of event types that occur together in both space and time. Pillai et al. [26] present a general framework to identify spatio-temporal cooccurrence patterns for continuously evolving spatio-temporal events that have polygon-like representations. Aydin et al. [3] investigate using specifically designated spatiotemporal indexing techniques for mining co-occurrence patterns from spatiotemporal datasets with evolving polygon-based representations. Data visualization technology is an effective tool to reduce the difficulty of data analysis. In recent years, data visualization has been widely used in data analysis. Wu et al. [30] present TelCoVis, an interactive visual analytics system, which helps analysts leverage their domain knowledge to gain insight into the co-occurrence in urban human mobility based on telco data. Sun et al. [29] present a five-level design framework for bicluster visualizations to provide a potential solution to ease the process of exploring and identifying coordinated relationships (e.g., four people who visited the same five cities on the same set of days) within some large datasets for sensemaking.

Co-occurrence pattern analysis can be utilized to solve practical problems. Hong et al. [13] propose a two-step black hole detection algorithm to detect urban black holes based on human mobility data, and black holes/volcanos are special cases of co-occurrence pattern. An infectious disease model on empirical networks of human contact is built to bridge the gap between dynamic network data and contact matrices and the high-resolution dynamic contact network is based on co-occurrence laws of individuals [22]. Akbari et al. put forward a new method to extract implicitly contained spatial relationships algorithmically, to deal with different feature types that is with point, line and polygon data, and to mine a spatio-temporal co-occurrence pattern simultaneously in space and time [1]. What's more, this method is applied on a real case study for air pollution.

Co-location pattern discovery searches for subsets of spatial features whose instances are often located at close spatial proximity [5]. The spatial co-location rule problem is different from co-occurrence pattern since there is no natural notion of transactions in spatial data sets which are embedded in continuous geographic space. Co-location pattern is highly similar to but different from co-occurrence pattern and co-location pattern is also widely studied. Huang et al. [15] provide a transaction-free approach to mine colocation patterns by using the concept of proximity neighborhood and Huang et al. [14] address the problem of mining co-location patterns with rare spatial features. Barua et al. [5] propose a pruning strategy for computing the prevalence measures to discover subsets of spatial features which are co-located due to some form of spatial dependency but not by chance.

In this paper, we focus on spatial temporal co-occurrence pattern in urban human mobility and aim to mining co-occurrence pattern based on an improved Apriori algorithm to analyze it from the perspective of regional functions.

Regional functions discovery Urbanization and modern civilization fosters various urban functional zones, such as workspace and business areas [17]. Discovering the functions of urban space is highly important for detecting urban problems, evaluating planning strategies, and supporting policy making [39]. The evolution of urban regional functions is influenced by many factors. Therefore, scholars evaluate regional functions from different perspectives based on a variety of data. Yuan et al. [35] infer the functions of each region using a topic-based inference model, which regards a region as a document, a function as a topic, categories of POIs (e.g., restaurants and shopping malls) as metadata (like authors, affiliations, and key words), and human mobility patterns (when people reach/leave a region and where people come from and leave for) as words. Zhong et al. [38]proposes a centrality index and attractiveness indices for detecting the urban spatial structure of functional centers and their spatial impacts using travel survey data. Besides survey data, Zhong et al. [39] integrate smart card data to infer urban functions at the building level utilized their proposed method in light of the potential of data mining and spatial analysis techniques for urban analysis. By combining smart card data and POI data, Han et al. [21] exploit discovering zones of different functions model and cluster analysis based on dimensionality reduction and expectation-maximization algorithm to identify functional zones that well match the actual land uses in Beijing. Yuan et al. [37] introduce the concept of Latent Activity Trajectory (LAT) to capture socioeconomic activities conducted by citizens at different locations in a chronological order and cluster the segmented regions into functional zones leveraging mobility and location semantics mined from LAT. Assem et al. [2] not only yield a deeper understanding of a complex city but also offer finer personalized recommendations based on regions' functionality that changes over space and time using the data collected from Location-Based Social Networks (LBSNs) [32]. In this paper, we mine static functions from POI data using TF-IDF method, discover dynamic functions from traffic trajectories utilizing LDA algorithm, and obtain urban actual regional functions by combining the above results.

3 Design of CoPFun

3.1 Overview

We display the framework of our proposed scheme CoPFun in Figure 2. We first perform general data preprocessing operations, including data cleaning, data mapping, and data organizing. Then we do temporal partition and map segmentation to extract population

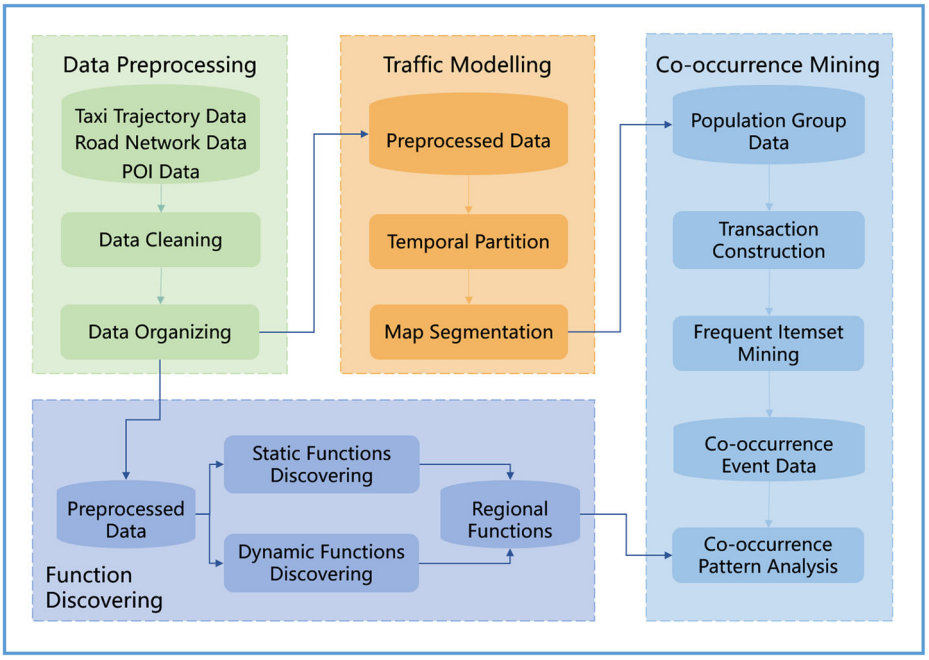


Figure 2 Framework of CoPFun

group data from preprocessed traffic data to carry out traffic modeling. Utilizing population group data, we construct transaction data and employ an improved Apriori algorithm to mine co-occurrence event data. Meanwhile, we discover urban regional functions using TF-IDF method and LDA algorithm. We have a detailed description of these parts in the following subsections.

Before introducing our scheme, we emphasize the definition of co-occurrence event once again. From human mobility perspective, the definition of a co-occurrence event is as follows:

If people from region A and region B visit region C at the same time interval, we say that "region A co-occurs with region B at region C".

We notice that the definition refers to time intervals and regions. Therefore, in order to get co-occurrence event data from raw trajectory data, we need to map continuous time to time intervals and map continuous longitude and latitude values to regions, which are temporal partition and map segmentation in traffic modeling.

3.2 Traffic modeling

3.2.1 Temporal partition

Temporal partition methods include regular time interval partition and irregular time interval partition [4, 20]. The former is to partition a day into same time intervals and the duration of each time interval should be chosen according to time distribution law of data. Irregular time interval partition is based on passengers flow peaks distribution. Taking into account of the characteristics of co-occurrence pattern, we divide days into same time intervals. The granularity of temporal partition is essential for co-occurrence pattern

analysis. A co-occurrence event refers to the origin and destination of a trajectory. So if the length of a time interval is too short, most travel will be cut apart. Whereas, if the length of a time interval is too long, we cannot obtain fine-grained co-occurrence pattern analysis results. The distribution of travel time can be a good reference to the duration of each time interval. We suggest the shortest time slot which covers most travel time as the duration of each time interval. In our experiments, we choose 30 minutes as a time slot. And we use (1) to do temporal partition of taxi trajectory data,

$$T_k = [k\theta, (k+1)\theta), k = 0, 1...47 \tag{1}$$

where T_k is the number of time intervals and θ is the duration of each time interval.

3.2.2 Map segmentation

Map segmentation is to divide the whole urban area into different small regions and map taxi trajectory data to OD data among small regions, which can provide an intuitive distribution laws of co-occurrence pattern in urban space. There are two kinds of region segmentation methods: regular segmentation and irregular segmentation. Regular segmentation is to divide study area into regular squares. Irregular segmentation includes using Voronoi tessellation to divide area based on particles [16], and segmentation based on urban road network framework [37]. Proper methods should be chosen according to data characteristics and analysis requirements. To achieve the goal we mentioned above, map segmentation needs two functions: divide urban space in a 2D plane, number each region and map a given longitude and latitude to a region. As we know, urban roads are designed according to urban planning and construction, which map the city into structured blocks and these blocks tend to show the city's functional bias, that is, blocks aggregate similar functions [36]. Consequently, it's reasonable to do map segmentation of urban space on city road network.

To carry out map segmentation and number regions, based on binary image processing method of mathematical morphology, map segmentation algorithm includes following four steps:

- **Dilation.** To acquire a fine-grained map segmentation result, we extract the main framework of urban road network to generate a binary image, in which black pixels stands for roads. Then we do dilation operation on the image to remove small gaps among road crossings. Iteration operation is required for a good dilation effect.
- **Thinning.** Perform thinning operation on the dilated image to thin the road width as a pixel. It's important to keep roads smooth during the thinning process. But the remove of pixels tends to generate non-smooth jagged lines.
- **Number.** Then we number the thinned image. We number each pixel of the image and pixels at same region have same number.
- **Delete.** Pixels representing roads in numbered image need to be removed. Our solution is to divide these pixels into their adjacent regions.

Based on temporal partition and map segmentation, we obtain population group data, which paves the way for co-occurrence event data mining.

3.3 Co-occurrence mining

Co-occurrence pattern refers to the phenomenon of traffic flow running to some regions at the same time, which coincides with frequent pattern in data mining. We regard co-occurrence pattern as the promotion and application of frequent pattern. Frequent pattern

is the pattern that appears frequently in a dataset. The set of items that are frequently present in the dataset at the same time is called frequent itemset. The discovery of such frequent pattern plays a crucial role in mining correlation, relevance, and many other interesting relations of data. Besides, it's also helpful for data classification, aggregation, and other data mining tasks. Therefore, frequent pattern mining is an important data mining task and one of the topics concerned by data mining researches. We utilize abundant research results of frequent pattern to mine co-occurrence event data effectively.

3.3.1 Transaction construction

Based on above operations, we map trajectory data to different population groups and then we need to do transaction construction to adapt for co-occurrence event data extraction. Population group data contains serial numbers of time intervals and serial numbers of origins and destinations. Co-occurrence pattern analysis focuses on relation of regions. So we construct transactions of regions at each time interval and mine co-occurrence pattern at each interval. We aggregate the destinations of traffic travel and such a transaction represents the set of regions that arrive at a same destination at a time interval. Examples of transaction set are displayed in Table 1. Line 1 of the table means that people from 18th region, 22nd region, and 33rd region go to 50th region at 9th time interval.

3.3.2 Frequent pattern

Before mining frequent itemset, we need to understand basic knowledge of frequent pattern. There are two primary statistic indexes for frequent pattern: support and confidence, which form the basic support-confidence framework to measure the interest of the rules for frequent patterns to reflect the usefulness and certainty of the rules found respectively [12]. Support count or support of the itemset is defined as the frequency of an itemset, which is the number of transactions containing the itemset in dataset. In co-occurrence transaction set, support is the number of destinations where a set of regions arrive at a time interval. Taking Table 1 for example, the support of region set containing 18th region, 22nd region, and 33rd region is 2 at 9th time interval. In co-occurrence pattern analysis, we define co-occurrence degree as the support of a region set. Another important statistic index of frequent pattern is confidence. It is defined as follows: rule A to B has confidence in a transaction set, where confidence is the percentage of transactions that contain A and also contain B in the transaction set.

$$confidence(A \Rightarrow B) = \frac{support(A \bigcup B)}{support(A)} \tag{2}$$

Table 1 Examples of constructed transaction set

TimeID	DestinationID	RegionID
9	50	18, 22, 33
32	343	359, 379, 381, 395
9	10	18, 22, 33
9	4	18, 22
1	74	22, 194
24	210	273, 348, 357, 370, 402, 428, 431
9	18	18, 22

Take Table 1 for example again. We define region set of 18th region and 22nd region as set A, region set of 33rd region as set B. Then $confidence(A \Rightarrow B)$ is 0.5. If the support of an itemset satisfies the predefined minimum support threshold, then the itemset is a frequent itemset. We define F as a frequent itemset. If there is no set larger than F, which has the same support as F in dataset, then F is closed and is called a closed frequent itemset. Analogously, if there is no set larger than F, which is frequent in dataset, then F is a maximal frequent itemset. Note that the difference between closed frequent itemsets and maximal frequent itemsets is that the support count of all subsets of closed frequent itemsets is datum and is the same as the closed frequent itemsets, while the maximal frequent itemsets only guarantee that all subsets are frequent. The essence of frequent itemset mining is to mine closed frequent itemsets. It also involves an important property of frequent itemsets: transcendental nature, which means, all non-empty subsets of frequent itemsets must be frequent. This property will be used to mine co-occurrence event data.

3.3.3 Frequent itemset mining

We employ a classical frequent itemset mining algorithm Apriori to mine co-occurrence event data. Agrawal and R. Srikant proposed the original algorithm for mining frequent patterns in 1994. The name of the algorithm is based on the fact that the algorithm uses transcendental nature of frequent itemsets. The transcendental nature means that all non-empty subsets of frequent itemsets must also be frequent, which is exactly what we mentioned above. Apriori algorithm uses an iterative method of layer-by-layer search, where the $k-itemsets$ is used to search for $(k+1)-itemsets$. First, by scanning the database, accumulate the count of each item, collect the items that satisfy the minimum support, and find frequent 1-itemsets. The set is denoted by L_1. Then, use L_1 to find the set L_2 of frequent 2-itemsets, and so forth, until you can not find the frequent $k-itemsets$. Finding each L_k needs a full scan of the database. The transcendental nature is used in the algorithm by connecting step and pruning step operations. The connecting step is to find a set of candidate $k-itemsets$ by connecting L_{k-1} with itself to find L_k. We can get from the transcendental nature that if a $(k-1)-item$ subset of a $k-itemset$ is not in L_{k-1}, then the $k-itemset$ can not be frequent and such $k-itemset$ is removed in the pruning step.

We have introduced details of co-occurrence event data mining algorithm. As a classical mining algorithm, Apriori algorithm has the problem of inefficiency in time consumption. Using L_k to generate $k+1$ candidate itemsets, it is too much to judge the connection conditions. The frequent itemsets set L_k whose itemsets' number is n, has the time complexity of $O(k*m^2)$ for performing comparison conditions and where the value of n can be large, especially when the minimum support threshold is set small. Definite the candidate $k-item$ set as C_k and element of C_k as c. In generating L_k from C_k, we need to determine if k $(k-1)$ subsets of c in C_k are all in $L_(k-1)$. In this process, $L_(k-1)$ only need to be scanned once for c in best case, ie the first $k-1$ subset is not in $L_(k-1)$. In the worst case, we need to scan k times. Thus, in the average case, for any c belonging to the C_k, its number of times scanning $L_(k-1)$ is $|L_(k-1)|*k/2$; then the number of scans required for all candidate itemsets is $|C_k|*|L_(k-1)|*k/2$. In order to get the support of all candidate frequent itemsets of $C_k(k=1,2,...,m)$, the database needs to be scanned m times. Considering the above deficiencies of Apriori algorithm, we take some measures to improve mining efficiency and optimize algorithm structures.

Transcendental nature: Subsets of frequent itemsets are frequent itemsets. When generating candidate sets, we first judge if all subsets of the itemset are frequent itemsets. If one

of its subsets is not frequent itemset, then the candidate set is discarded. In this way, we can compress the number of candidate sets and save the time of transaction scanning.

Hash map: We utilize hash map to store frequent itemsets and candidate sets. Hash map can support quick search of large amount of data and save searching time of algorithms.

Algorithm 1 CoPFun Approach

input:
Co : co-occurrence transaction data set
Min_{sup} : minimum support threshold
output:
co-occurrence events in Co.

$L_1 = find\_frequent\_1\_itemsets(Co)$;
for $(k = 2; L_{k-1} \neq \emptyset; k++)$ **do**
  $C_k = co - occurrence\_gen(L_{k-1})$;
  for each transaction $t \in Co$ **do**
    $C_t = \textbf{subset}(C_k, t)$;
    for each candidate $c \in C_t$ **do**
      $c.count++$;
    end for
  end for
  $L_k = c(C_k | c.count \geq Min_{sup})$;
end for
return $L = \bigcup_k L_k$

procedure $co - occurrence\_gen$ (L_{k-1}:$frequent$ $(k-1)$ $itemset$);
for each itemset $l_1 \in L_{k-1}$ **do**
  for each itemset $l_2 \in L_(k-1)$ **do**
    if $(l_1[1] = l_2[1]) \wedge ... \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] = l_2[k-2])$ **then**
      $c = l_1 \bowtie l_2$;
      if $has\_infrequent\_subset(c, L_{k-1})$ **then**
        delete c;
      else
        add c to C_k;
      end if
    end if
  end for
end for
return C_k

procedure $has\_infrequent\_subset$ (c : candidate k itemset;L_{k-1}: frequent$(k-1)$ itemset);// use transcendental nature
for each $(k-1)$ subset s of c **do**
  if $s \notin L_{k-1}$ **then**
return TRUE
  end if
end for
return FALSE

As shown in the above pseudocode , $co - occurrence\_gen$ has two actions: connecting and pruning. In connecting part, L_{k-1} connects with L_{k-1} to generate possible candidates

(step 2 step 5). Pruning part (step 6 step 8) utilize transcendental nature to delete candidates with infrequent itemsets. The test of infrequent itemsets is presented in the process of $has\_infrequent\_subset$. We store the frequent itemsets and their support of the extracted co-occurrence event data according to data and time mark. The minimum support threshold we used in our experiment is 4.

3.4 Discovering regional functions

We introduce how to extract co-occurrence event data from taxi trajectory data utilizing our proposed scheme CoPFun above. Based on co-occurrence event data, we can analyze urban co-occurrence pattern. Region functions have great impact on human mobility [31]. Therefore, in this section, we will describe how to discover region functions to pave the way for analyzing urban co-occurrence pattern from the perspective of different region functions. Although regional functions are largely determined by urban planning, actual regional functions or strength of actual functions can be changed due to the development of city and the impact of human activities. Actual regional functions are influenced by the factors of static semantic and dynamic activity. So we study on static functions, dynamic functions, and actual functions respectively and discover urban regional functions based on multiple factors.

3.4.1 Discovering static functions

To understand regional static functions, we use point of interest data (POI data) to discover the status of regions assuming functions. POI data includes names, locations, and classifications of physical buildings [33, 34]. Proper mining of all kinds of POI data at each regions can get static semantic functions [37]. We consider the relation of POI data and regional functions from two aspects. On one hand, if the frequency of certain kind of POI data appearing at a region is high, that is, absolute quantity is large, then the percentage of the region to assume such kind of function should be high; on the other hand, if the frequency of one kind of POI data appearing at other regions is low, while it appears at the region to a degree, that is, relative quantity is large, then such kind of POI data may reflect the functional characteristics of the region. This idea is quite similar to the thought of TF-IDF (Term Frequency-Inverse Document Frequency) [24] in information retrieval. Therefore, in this paper, we employ TF-IDF method to process POI data of regions to dig up the POI distribution of regions. TF-IDF method evaluate the importance of a word from its two properties: First, the frequency of the word appearing in a document (Term Frequency, TF). Generally speaking, when one word appears in a document repeatedly, it can reflect document content, that is, the value of its TF bigger, the word more important. Second, inverse document frequency (IDF). If one word appears repeatedly in the whole document set, then the ability to distinguish documents' content is poor and the value of IDF is low. TF-IDF algorithm multiply the value of TF by the value of IDF to be the importance measure of a word. If and only if the term frequency of a word in a document is high and it rarely appears in other documents of the document set, the value of TF-IDF is high.

First, we calculate the frequency of all kinds of POI data at each region, that is term frequency. We use symbol $TF_{i,j}$ to present the frequency of jth POI data at region r_i. As shown in (3), S represents the total number of POI categories.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_{k-1}^{S} n_{i,k}} \tag{3}$$

Then, we use IDF_j to present the IDF of jth POI data. The formula is as follows:

$$IDF_j = log \frac{R}{|i|n_{i,j} \neq 0, i = 1, 2, ..., R| + 1}, \tag{4}$$

where R is the number of regions. The multiplication of the two is the TF-IDF value of region r_j for jth POI data, denoted as $TF - IDF_{i,j}$.

$$TF - IDF_{i,j} = TF_{i,j} * IDF_j \tag{5}$$

From this formula, we can get that $TF - IDF_{i,j}$ is proportional to the number of occurrences of jth POI data at region r_i and is inversely proportional to the number of occurrences of jth POI data at all regions. That is the importance of one kind of POI data or function at a region that we first want to measure. At last, POI's distribution status can be represented by vector $\vec{Y}_i$ in formula (6).

$$\vec{Y}_i = (TF - IDF_{i,1}, TF - IDF_{i,2}, ..., TF - IDF_{i,S}) \tag{6}$$

After these calculations, we can get the distribution of POI data at each region and understand the importance and percentage of all functions at regions from the perspective of static semantic.

3.4.2 Discovering dynamic functions

Only from the distribution of POI data, we cannot distinguish quality status of different regions and interactions between regions. For example, the city is full of all kinds of food service agencies, but their roles in regional function are different, small snack bars may only meet daily needs of local residents, while some restaurants are well-known to attract the whole city and even the whole world to enjoy, then the possibility that this place has entertainment function may be high. Such information can be extracted from the number and patterns of trajectory data and other human mobility. In this paper, we utilize Latent Dirichlet Allocation (LDA) model to process trajectory data to obtain regional dynamic functions affected by human mobility.

LDA is a topic model, which can offer the topic of each article in document set in the form of probability distribution and then do similarity analysis, text clustering, and so on of documents based on topic distribution. LDA considers each document in document set contains multiple topics and each word in documents belongs to a topic. When given all words appearing in each document of the document set, LDA can infer the implied topic distribution of documents. LDA model solving is a quite complex optimization problem and its common methods are Gibbs sampling-based solution, variational method-based EM solution, and method based on expectation advance.

In this paper, we do analogy on function mining of regions and topic mining of documents and the validity of such analogy method has been proved [37]. Specifically, we treat all regions as a document set, a region as a document, a regional function as a topic of a document, then a region has various functions as a document implies a series of topics. Meanwhile, we regard a travel mode appearing in a region as a word in a document.

For region r_i, we define a origin matrix L^i of R rows and T columns, where R represents the total number of regions and T represents the number of time intervals; the element $L^i[j, k]$ in jth row and k column of the matrix denotes the travel trajectory from the region to region r_j at time interval t_k, value of the element is the number of corresponding trajectory. Similarly, define a destination matrix A^i of R rows and T columns for region r_j and the element $A^i[j, k]$ in jth row and kth column represents the travel trajectory from region

r_j to the region and its value is the number of corresponding trajectories. Then we regard region r_i as a document and R regions make up the document set; each element $L^i[j,k]$ and $A^i[j,k]$ in two matrices (origin matrix and destination matrix) of region r_i stand for a kind of travel mode, which can be treated as words in documents; the value of elements $L^i[j,k]$ and $A^i[j,k]$ is the number of travel mode, which can be regarded as the frequency of words. Based on above analogy, we can obtain travel modes at each region and use LDA algorithm to infer the function distribution implied in regions and the similarity of regions.

3.4.3 Hybrid mining of static function and dynamic function

We gain regional static functions from POI data and regional dynamic functions from taxi trajectory data. Regional actual functions are affected by various factors. If we only consider one of them to mine regional functions, the results are not enough to objectively reflect real situation. Therefore, we combine the two kinds of functions to mine actual functions. We construct a cost function and solve it utilizing gradient descent method.

Inspired by machine learning methods, we define cost function J, that is, objective function, to represent the deviation of actual functions and static and dynamic functions. Since actual regional functions doesn't deviate greatly from its inherent functions, that is, the static functions excavated from POI data using TF-IDF method, we initialize actual functions with static functions when initializing. Under different conditions, we define the cost as the deviation of actual situation and the result of TF-IDF method or the deviation between actual situation and the result of LDA algorithm. Then we need to find the minimum point of cost function, which means the minimum difference between actual functions distribution we expect and the dynamic and static functions displayed. We utilize gradient degree method to iteratively update independent variables until the function's value changes slowly or reaches the maximum number of iterations. At this point, we think that the minimum value of cost function is obtained and it is regarded as the final function proportion of regions. The function distribution takes into account of the static semantic factors and dynamic human activities, which can reflect actual regional function situation objectively.

4 Experiments

In this section, we first give details of datasets and then present experiment results from the aspects of traffic modeling, co-occurrence mining, and regional functions discovering. We utilize real taxi trajectory data, road network data, and POI data of Shanghai to display the process of and demonstrate the effectiveness of our scheme.

4.1 Experiment datasets

With rapidly increasing availability of sensing technology and telecommunication infrastructure, huge volumes of mobile data tracking human mobility can be acquired. Multi-source heterogeneous traffic data make analysis of human mobility pattern precisely. In this work, we employ three types of traffic datasets, which are taxi GPS trajectory, road network data, and POI data. Taxi dataset is generated by Qiangsheng taxi GPS trajectories from Shanghai in China. As shown in Table 2, the dataset consists of 7 fields such as state, speed, date, time, geographical coordinates, and so on. Meanwhile, Table 2 shows the statics in detail. The unit of speed in Table 2 is km/h. We carry out general data preprocessing operations on raw GPS data, including data cleaning, data mapping, and data organizing. Based

Table 2 Record examples and statics of taxi GPS trajectory data

Name	Field	Annotation	Example
GPS data	TaxiId	Taxi Id	2201252167
	Latitude	Coordinates	121.465545
	Longitude	Coordinates	31.224068
	State	1: occupied, 0: vacant	1
	Date	Date that GPS record was sent	2015-04-25
	Time	Time that GPS record was sent	13:17:00
	Speed	Taxi running speed	6.613991
Statics	time	April, 2015	1th-30th
	days	30	21 weekdays
	number of taxis	13,695	about 25% of the taxis in Shanghai
	dataset size	34 billion GPS records	619GB

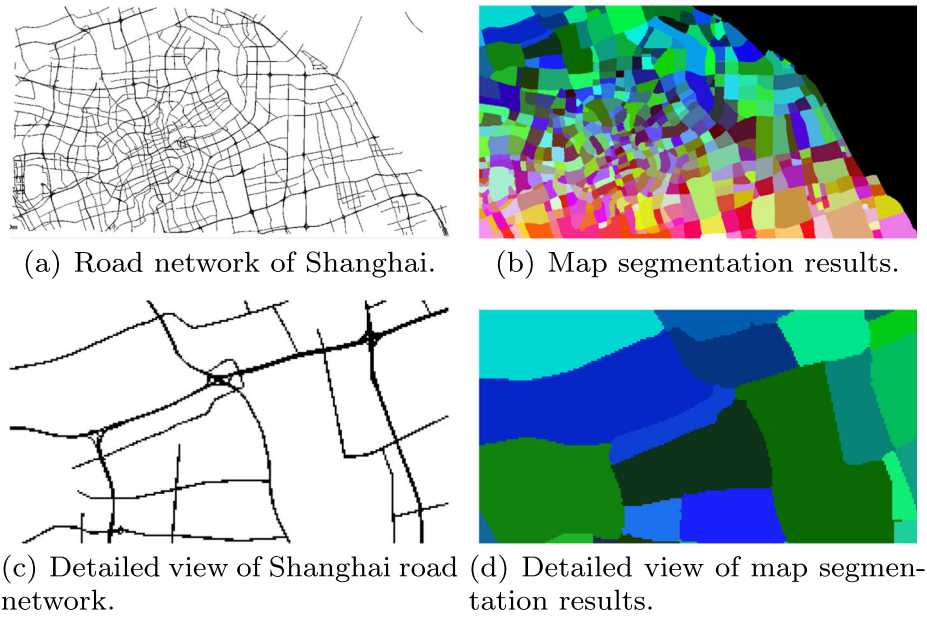
on preprocessed data, we extract time and location of boarding and alighting according to the requirements of co-occurrence pattern analysis.

We use 486,815 Shanghai POI data and each POI data contains six properties: ID, name, latitude, longitude, explanation, and type, as shown in Table 3. To pave the way for regional function mining, we refine the classification of POI data into six categories: residential area, workspace, education, business area, public service, and scenery spot and filter out the records which are not in above categories. Table 3 also presents basic statistic results of Shanghai POI data.

We obtain Shanghai road network data from OpenStreetMap (OSM), which includes road ID, road type, information of traffic signals, and longitude and latitude of all points on the road. We extract information of highways, primary roads, and secondary roads from the road network data and map it on a binary image, as shown in Figure 3a. We focus on urban area in Shanghai, which is located at $31.15°N$ to $31.37°N$ and $121.31°E$ to $121.84°E$.

Table 3 Record examples and statics of POI data

Name	Field	Example
POI data	ID	81166
	Name	China Postal Savings Bank (Southern Drive Branch)
	Longitude	121.415172
	Latitude	31.688554
	Explanation	Financial and insurance services; banks; China Postal Savings Bank
	Type	4
Statics	Residential area	26,729
	Workspace	59,092
	Education	9,752
	Business area	53,793
	Public service	74,298
	Scenery spot	1,454



(a) Road network of Shanghai. (b) Map segmentation results.

(c) Detailed view of Shanghai road network. (d) Detailed view of map segmentation results.

Figure 3 Comparison of raw road network and map segmentation results in Shanghai

4.2 Traffic modeling results

After data preprocessing operations, we do traffic modeling to map vehicles into different population groups, which contains temporal partition and map segmentation.

4.2.1 Travel rule analysis

The granularity of temporal partition is hard to control and has great impact on human mobility analysis. To determine the length of time intervals, we do travel rule analysis on taxi trajectory data.

After data preprocessing, we extract OD data from trajectory data. However, there are still abnormal data, such as a taxi drives three kilometers in four hours. For such abnormity, our cleaning method is to calculate average travel speed and to filter out records according to minimum speed threshold. Speed calculation depends on distance and driving time. We use Manhattan distance to calculate travel distance, which is also known as taxi distance. As urban streets generally have southern, northern, western, and eastern layout rules, the distance of taxis from one point to another point is about equal to the value that distance from the north-south direction adds distance from the east-west direction, which is exactly Manhattan distance. Then we store two more fields: duration and speed, which are duration of travel time and average travel speed respectively. Based on calculation results, we do statistics of duration of travel time and display the cumulative proportion of four typical days in Figure 4, which are 5th, 7th, 9th, and 10th of April, 2015. They are two days in the Qingming Festival and two weekdays. We can get from the figure that 80% of the duration of travel time is less than 30 minutes, and with the increase in duration of travel time, the growth rate of cumulative proportion is getting slower. Therefore, in order to obtain fine-grained co-occurrence analysis results, we set the length of time intervals as 30 minutes,

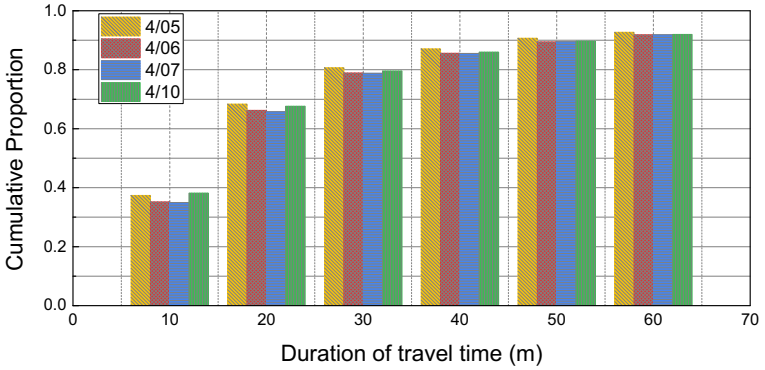


Figure 4 Duration of taxi travel time

number time intervals, and convert the temporal information of OD data into time interval number.

4.2.2 Map segmentation

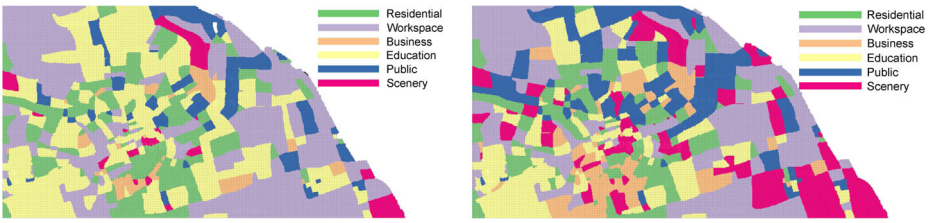
After temporal partition, we divide the study area utilizing map segmentation based on the binary image of road network. Though we just select highways, primary roads, and secondary roads to make up the urban road network, the lines of the image are still intensive and there are a lot of small crossings. After dilation operation, we remove unnecessary details from the image and there is an example in Figure 3. The whole marked results of map segmentation is presented in Figure 3b. We divide the whole study area into 541 regions and number them. We then apply map segmentation results to OD data and convert its location information into region number.

4.3 Co-occurrence event data

In the process of co-occurrence event data extraction, we first traverse the entire transaction set obtained by transaction construction operation to calculate the support of each transaction, that is, co-occurrence degree. Then we use frequent pattern mining algorithm to dig up frequent itemsets upwards layer by layer, and end iterations when we cannot find more frequent itemsets. We store all frequent itemsets, that is, co-occurrence event data we expect. Table 4 displays several examples of extracted co-occurrence event data, in which a record stands for a co-occurrence event. The co-occurrence event data has four fields including time id, item, support, and key. We mine co-occurrence pattern at each time interval. Field

Table 4 Examples of co-occurrence event data

TimeId	Item	Support	Key
41	422, 443, 445, 474, 483, 506, 510, 516	4	8
41	357, 431, 458, 459, 475, 515	4	6
13	357, 431	53	2
11	10,22	8	2
12	334, 340, 357, 373, 431	4	5



(a) Static functions (b) Actual functions

Figure 5 Distribution of functional regions

item denotes a region set in which regions co-occur with each other, field key is the number of regions in item, and support is the value of item support. Take the line1 in Table 4 for example. It means a region set of 8 regions, which contains 422nd region, 443rd region, 445th region, 474th region, 483rd region, 506th region, 510th region, and 516th region, co-occurs with each other at 4 destinations at 41st time interval.

4.4 Regional functions distribution

In Section 3, we refine the classification of POI data categories, map it into different regions, and employ TF-IDF method to mine static regional functions. The process of dynamic reginal functions mining is relatively complex. We first extract effective information from taxi trajectory data, including temporal and spatial information of alighting and boarding. Then we organize the above information in the form of matrices. We apply IDA algorithm to matrices to do topic mining and set k as 6, which is equal to the number of regional functions and gain probability distribution of topics at each region. Cost function we defined helps to combine TF-IDF results and LDA results. We store probability distribution of topics at each region when cost function reaches its minimum. Then we define the topic with largest probability as the regional function. The distribution of functional regions is displayed as Figure 5. In the distribution of static functions, the number of business areas is quite large, while the recognition ability of other several important functional regions is poor. We combine static functions and dynamic functions to generate actual functions, and it has a clear advantage on identifying workspace, scenery spot, education and other functional regions. Table 5 presents the statistic results of functional regions.

Table 5 Regional functions distribution

Category	No.	Number of static functions	Number of actual functions
Residential area	0	121	100
Workspace	1	145	123
Education	2	28	65
Business area	3	183	105
Public service	4	45	81
Scenery spot	5	20	68

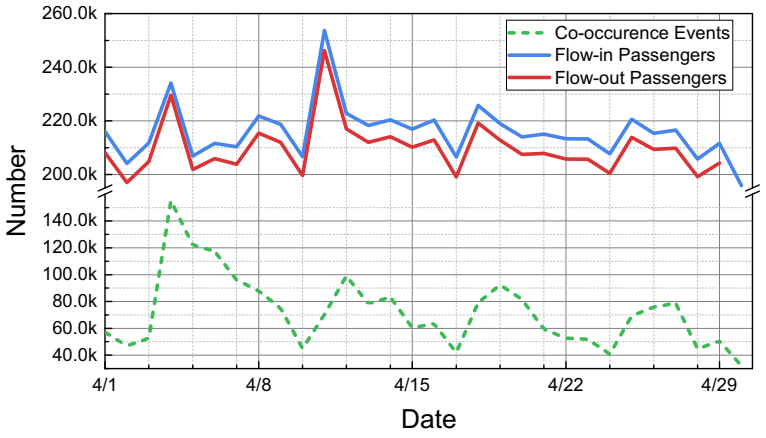


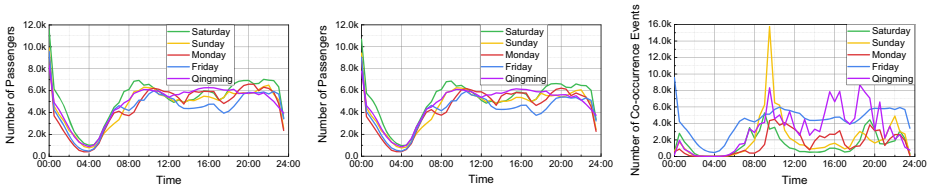
Figure 6 Passenger inflow, passenger outflow, and co-occurrence events' number in April 2015

5 Co-occurrence pattern analysis

To acquire urban co-occurrence pattern, we carry out statistic analysis and visual analysis of extracted co-occurrence event data and make comparison with OD data. Functional regions are essential for urban human mobility. Therefore, we utilize regional function mining results to analyze co-occurrence pattern.

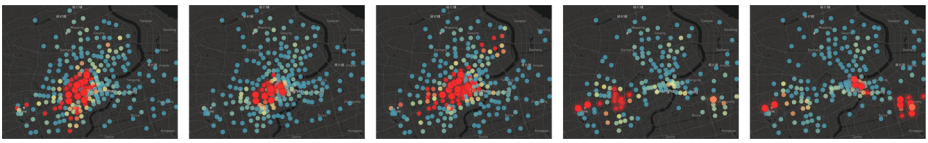
5.1 Urban global co-occurrence pattern

We first focus on changing trend of passenger inflow, passenger outflow, and co-occurrence events' number in April 2015, which is shown in Figure 6. We can get from the figure that periodic law of passenger flow is obvious and weekly changes are very similar. Flow peak is Saturday and flow trough is Friday. Passenger inflow is slightly higher than passenger outflow on the whole. However, compared with passenger flow, the quantity of co-occurrence events is obviously smaller. Its periodic law is basically similar to the periodic law of passenger flow. For example, their troughs are both Friday. Difference also exists. Peak of co-occurrence events' number is Sunday, rather than Saturday. After the first peak in Figure 6, the trend of passenger flow is significantly different from co-occurrence events' number. The peak date is April 4, which is the first day of Qingming Festival, a statutory holiday. It means co-occurrence pattern is more significantly influenced by holidays, which may be related to its magnitude but also shows a characteristic of co-occurrence pattern.



(a) Inflow (b) Outflow (c) Co-occurrence

Figure 7 Passenger inflow, passenger outflow, and co-occurrence events' number of 5 typical days in April 2015



(a) Qingming (b) Saturday (c) Sunday (d) Monday (e) Friday

Figure 8 Co-occurrence heat distribution at 09:30 a.m. of 5 typical days in April 2015

Then we choose five typical days to analyze co-occurrence pattern further in Figure 7. They are April 4, April 11, April 12, April 13, and April 17, corresponding to the first day of Qingming Festival, Saturday, Sunday, Monday, and Friday. Through Figure 7 we can see that there exists a high degree of similarity between passenger inflow and passenger outflow, while co-occurrence event data is significantly different from the above two. For passenger flow, the change over time is not obvious and the change of co-occurrence event data has great fluctuation. Passenger flow on weekend has a morning peak at around 9 and a night peak at about 22. In addition to the above two peaks, passenger flow on weekdays also has a peak at 4:00 p.m. However, Qingming Festival has high passenger flow all day. Both on weekdays and weekends, co-occurrence event data has similar changing law and peaks with passenger flow, but fluctuates strongly. The number of co-occurrence events is significantly more than other days and the co-occurrence events; number is extremely large from 9:00 to 9:30. To obtain more detailed co-occurrence pattern distribution, we map co-occurrence event data from 9:00 to 9:30 of the five days to heatmap to display geographical distribution in Figure 8.

Co-occurrence events of Qingming Festival and weekends are crowded and concentrate on urban center. Co-occurrence heat on weekdays is smaller and scatter in the west and east of the city. In Figure 9, we use dots to present regions and use lines to present interactions, that is, co-occurrence events. The color of dots indicates corresponding regional functions. Urban area with great co-occurrence heat on weekdays is concentrated with orange dots, which denote workspace. Workers need to work on weekdays to form a great many co-occurrence events.

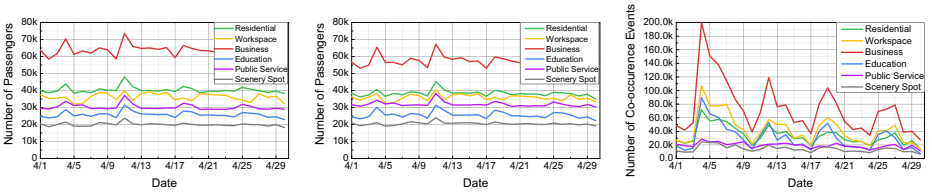
5.2 Functional regions' co-occurrence pattern

We divide regional functions of the whole city into six categories: residential area, workspace, business area, education, public service, and scenery spot and employ POI data and taxi trajectory data to mine regional functions from 541 regions in Shanghai. Based on regional functions mining results, we utilize co-occurrence event data and OD data to analyze urban co-occurrence pattern from a perspective of regional functions.

First we display total passenger flow and total co-occurrence events' number of all regions belonged to each kind of functional regions over April 2015 in Figure 10. The

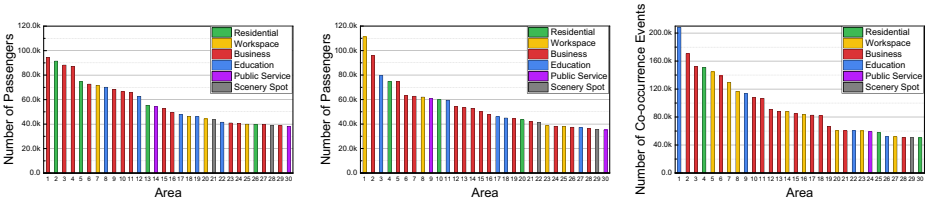
(a) Qingming (b) Saturday (c) Sunday (d) Monday (e) Friday

Figure 9 Co-occurrence interactions at 09:30 a.m. of 5 typical days in April 2015



(a) Inflow (b) Outflow (c) Co-occurrence

Figure 10 The total number of inflow passengers , outflow passengers , and co-occurrence events of 6 functional regions in April 2015

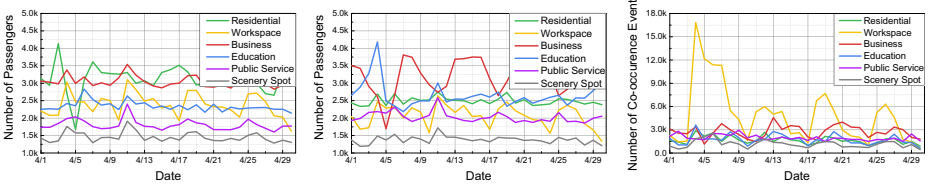


(a) Inflow (b) Outflow (c) Co-occurrence

Figure 11 Top 30 ranking regions of passenger inflow, passenger outflow, and co-occurrence events' number in April 2015

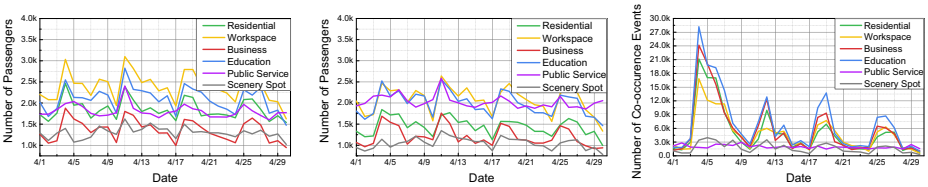
Table 6 Statistics of regional functions in top 30 regions

Function	Inflow	Outflow	Co-occurrence
Residential area	4	3	3
Workspace	4	4	8
Business area	13	14	13
Education	5	5	4
Public service	2	2	1
Scenery spot	2	2	1



(a) Inflow (b) Outflow (c) Co-occurrence

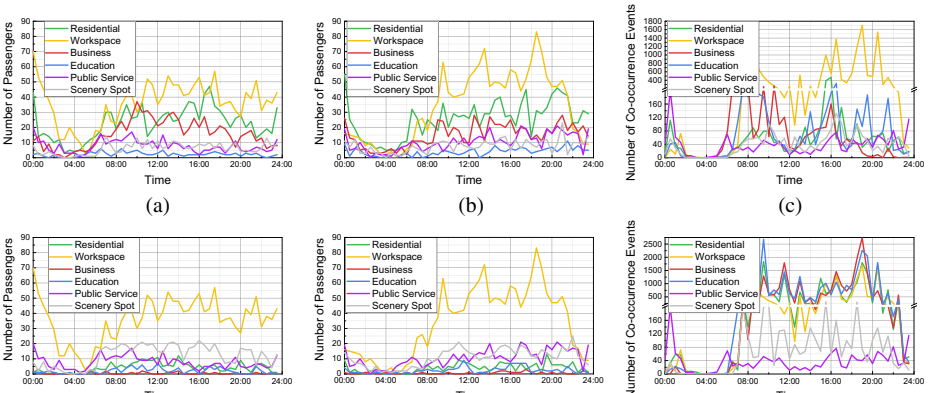
Figure 12 The number of inflow passengers , outflow passengers, and co-occurrence events of 6 top functional regions sort by passenger flow in April 2015



(a) Inflow (b) Outflow (c) Co-occurrence

Figure 13 The number of inflow passengers, outflow passengers, and co-occurrence events of 6 top functional regions sort by co-occurrence events' number in April 2015

passenger flow of business area, residential area, workspace, public service, education, and scenery spot goes down in turn. The number of co-occurrence events on business areas is significantly more than other functional regions and workspace follows. Co-occurrence events' number of residential areas and education is almost the same. The quantity of co-occurrence events on public service is slightly higher than scenery spot, but they both have a small number of co-occurrence events. Besides the quantitative relation above, passenger flow and co-occurrence events' number have significant weekly change law, which is consistent with the total passenger flow and total co-occurrence events' number. Then we sum up passenger flow and co-occurrence events' number in 30 days for each region and rank regions. We select top 30 regions of passenger inflow, passenger outflow, and co-occurrence events' number to display in Figure 11 respectively. We can get that both for passenger flow and co-occurrence events' number, proportion of business areas is the largest, which may be related to the number of business areas. The difference between passenger flow and co-occurrence events' number is mainly reflected in the number of workspace, that is, the co-occurrence events' number of workspace stands out. Compared with original OD data, co-occurrence event data can reflect traffic clustering situation, which is indicated in Figure 11. In top 10 regions, the number of education and workspace of co-occurrence events' number is significantly more than that of passenger flow (Table 6).



(a) (b) (c)

(d) (e) (f)

Figure 14 Flow trend on April 4, 2015. Passenger inflow (**a**), passenger outflow (**b**), and co-occurrence events' number (**c**) of 6 top functional regions sort by passenger flow. Passenger inflow (**d**), passenger outflow (**e**), and co-occurrence events' number (**f**) of 6 top functional regions sort by co-occurrence events' number

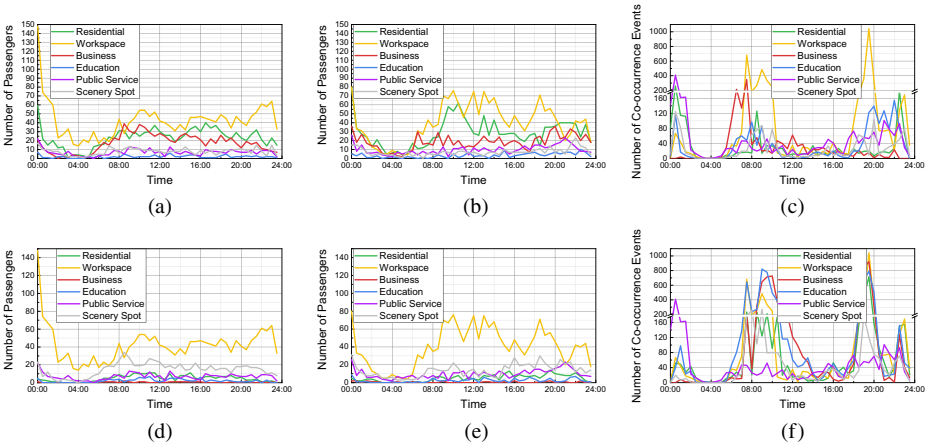


Figure 15 Flow trend on April 11, 2015. Passenger inflow (**a**), passenger outflow (**b**), and co-occurrence events' number (**c**) of 6 top functional regions sort by passenger flow. Passenger inflow (**d**), passenger outflow (**e**), and co-occurrence events' number (**f**) of 6 top functional regions sort by co-occurrence events' number

In order to explore co-occurrence pattern of a single region from the aspect of regional functions, we sort regions belonged to each type of functions according to passenger flow and co-occurrence events' number respectively, and select top 1 regions of passenger flow and co-occurrence events' number respectively. Figures 12 and 13 display periodic laws in passenger flow and co-occurrence events' number. Regions with large flow may have a small number of co-occurrence events, which indicates that the travel time of these regions is relatively dispersive. Co-occurrence events' number on workspace is quite great due to concentrated travel time of workers. This law is reflected in Figure 13, passenger flow of scenery spots and public service is quite impressive, while their co-occurrence events' number is relatively small, which means the travel time involved in such two functional regions is scattered. In contrast, co-occurrence events' number of other education, business area,

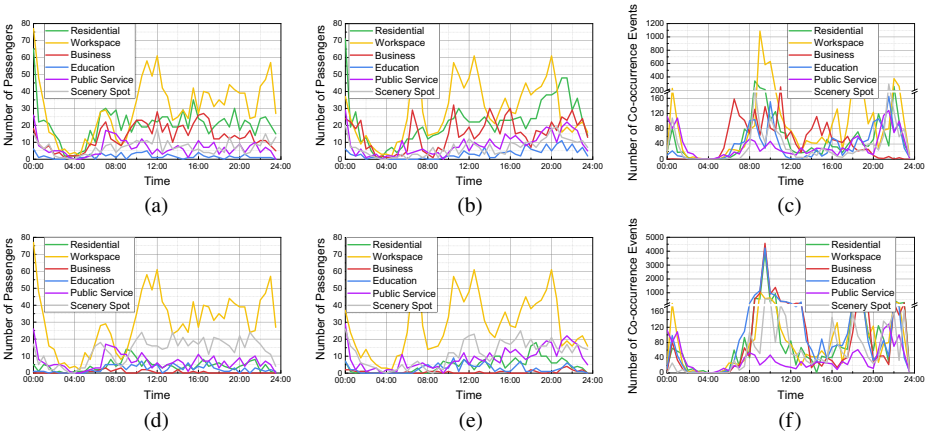


Figure 16 Flow trend on April 12, 2015. Passenger inflow (**a**), passenger outflow (**b**), and co-occurrence events' number (**c**) of 6 top functional regions sort by passenger flow. Passenger inflow (**d**), passenger outflow (**e**), and co-occurrence events' number (**f**) of 6 top functional regions sort by co-occurrence events' number

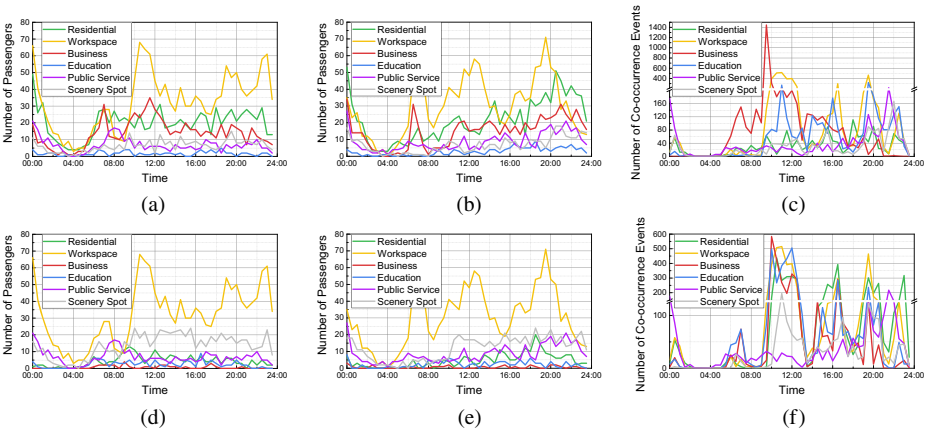


Figure 17 Flow trend on April 13, 2015. Passenger inflow (**a**), passenger outflow (**b**), and co-occurrence events' number (**c**) of 6 top functional regions sort by passenger flow. Passenger inflow (**d**), passenger outflow (**e**), and co-occurrence events' number (**f**) of 6 top functional regions sort by co-occurrence events' number

residential area, and workspace is huge. Travel of these functional regions is sensitive to time and crucial for urban operating normally. In addition, we can also get that holidays have great impact on co-occurrence events' number and co-occurrence pattern is sensitive to human mobility and has strong exploring ability.

We further explore co-occurrence pattern of functional regions over one day. Figure 17 presents the results of April 13rd. We can see that there are three distinct peaks of workspace in the change law of passenger flow and the change trend of other five functional regions is relatively slow. In contrast, co-occurrence events' number has significant fluctuations throughout the day. It has three peaks at 10, 16, and 20 respectively and many other sub-peaks, which shows that although flow over the day changes flat, people are concentrated to go out at ten o'clock, sixteen o'clock, and twenty o'clock around. Results of other days are shown in Figures 14, 15, 16, 18.

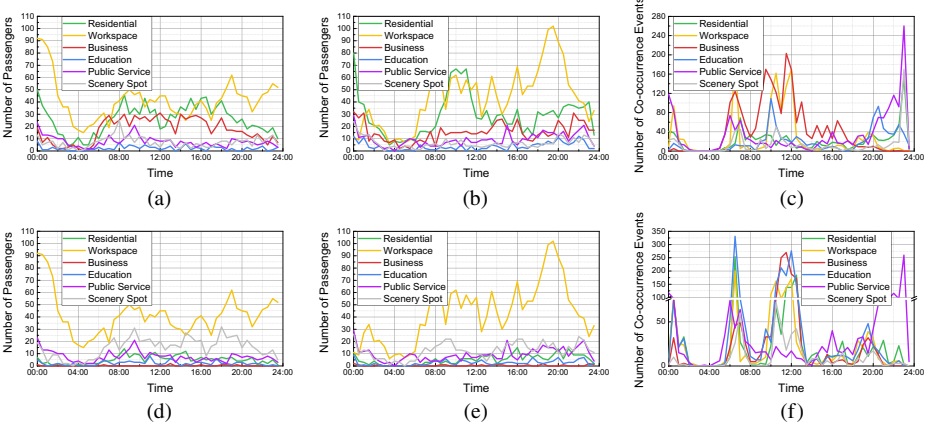


Figure 18 Flow trend on April 17, 2015. Passenger inflow (**a**), passenger outflow (**b**), and co-occurrence events' number (**c**) of 6 top functional regions sort by passenger flow. Passenger inflow (**d**), passenger outflow (**e**), and co-occurrence events' number (**f**) of 6 top functional regions sort by co-occurrence events' number

6 Conclusion and future work

In this paper, we put forward a co-occurrence pattern mining scheme CoPFun to extract co-occurrence event data from a variety of traffic data, which embraces traffic modeling, co-occurrence mining, and function discovering. Case studies on real taxi trajectory data, POI data, and urban road network data of Shanghai demonstrate our proposed scheme can mine co-occurrence event data effectively. Then we present co-occurrence pattern analysis by comparing co-occurrence event data with OD data from a perspective of urban regional functions. We find that co-occurrence events' number owns obvious periodic changing law but great fluctuation over days. Holidays have significant impact on co-occurrence events' number. Co-occurrence pattern is sensitive to human mobility and has strong exploring ability.

There are multiple venues for future work. First, we plan to optimize algorithms to mine co-occurrence events and passenger flow at the same time. In addition, we intend to develop a suite of visualization forms to explore urban co-occurrence pattern intuitively and flexibly and provide valuable suggestions for planning and development of cities from the aspect of co-occurrence pattern.

Acknowledgments This work was partially supported by the National Natural Science Foundation of China under Grant no. 61572106, the Natural Science Foundation of Liaoning Province, China under Grant no. 201602154, Fundamental Research Funds for the Central Universities under Grant no. DUT18JC09, and China Scholarship Council under Grant no. 201706060067.

References

1. Akbari, M., Samadzadegan, F., Weibel, R.: A generic regional spatio-temporal co-occurrence pattern mining model: a case study for air pollution. J. Geogr. Syst. **17**(3), 249–274 (2015)
2. Assem, H., Xu, L., Buda, T.S., O'Sullivan, D.: Spatio-temporal clustering approach for detecting functional regions in cities. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI), pp. 370–377. San Jose, USA (2016)
3. Aydin, B., Kempton, D., Akkineni, V., Gopavaram, S.R., Pillai, K.G., Angryk, R.: Spatiotemporal indexing techniques for efficiently mining spatiotemporal co-occurrence patterns. In: 2014 IEEE international conference on big data (Big Data), pp. 1–10. Washington, DC, USA (2014)
4. Bao, J., Zheng, Y., Wilkie, D., Mokbel, M.: Recommendations in location-based social networks: A survey. Geoinformatica **19**(3), 525–565 (2015)
5. Barua, S., Sander, J.: Sscp: Mining statistically significant co-location patterns. In: International symposium on spatial and temporal databases, pp. 2–20. Berlin, Germany (2011)
6. Celik, M., Shekhar, S., Rogers, J.P., Shine, J.A.: Mixed-drove spatiotemporal co-occurrence pattern mining. IEEE Trans. Knowl. Data Eng. **20**(10), 1322–1335 (2008)
7. Celik, M., Shekhar, S., Rogers, J.P., Shine, J.A., Yoo, J.S.: Mixed-drove spatio-temporal co-occurence pattern mining: A summary of results. In: 6th international conference on data mining, 2006. ICDM'06, pp. 119–128. Hong Kong, China (2006)
8. Chen, D.: Research on traffic flow prediction in the big data environment based on the improved rbf neural network. IEEE Trans. Ind. Inf. **13**(4), 2000–2008 (2017)
9. Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., Huttenhower, C.: Microbial co-occurrence relationships in the human microbiome. PLoS computational biology **8**(7), e1002,606 (2012)
10. Ferreira, N., Poco, J., Vo, H.T., Freire, J., Silva, C.T.: Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. IEEE Trans. Vis. Comput. Graph. **19**(12), 2149–2158 (2013)
11. Gao, S., Janowicz, K., Couclelis, H.: Extracting urban functional regions from points of interest and human activities on location-based social networks. Trans. GIS **21**(3), 446–467 (2017)

12. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier, New York (2011)
13. Hong, L., Zheng, Y., Yung, D., Shang, J., Zou, L.: Detecting urban black holes based on human mobility data. In: Sigspatial International Conference on Advances in Geographic Information Systems, p. 35 (2015)
14. Huang, Y., Pei, J., Xiong, H.: Mining co-location patterns with rare events from spatial data sets. Geoinformatica **10**(3), 239–260 (2006)
15. Huang, Y., Shekhar, S., Xiong, H.: Discovering colocation patterns from spatial data sets: a general approach. IEEE Trans. Knowl. Data Eng. **16**(12), 1472–1485 (2004)
16. Kong, X., Song, X., Xia, F., Guo, H., Wang, J., Tolba, A.: Lotad: long-term traffic anomaly detection based on crowdsourced bus trajectory data. World Wide Web. https://doi.org/10.1007/s11280-017-0487-4 (2017)
17. Kong, X., Xia, F., Ning, Z., Rahim, A., Cai, Y., Gao, Z., Ma, J.: Mobility dataset generation for vehicular social networks based on floating car data. IEEE Transactions on Vehicular Technology. https://doi.org/10.1109/tvt.2017.2788441 (2018)
18. Kong, X., Xia, F., Wang, J., Rahim, A., Das, S.K.: Time-location-relationship combined service recommendation based on taxi trajectory data. IEEE Trans. Ind. Inf. **13**(3), 1202–1212 (2017)
19. Li, F., Li, Z., Sharif, K., Liu, Y., Wang, Y.: Multi-layer-based opportunistic data collection in mobile crowdsourcing networks. World Wide Web. https://doi.org/10.1007/s11280-017-0482-9 (2017)
20. Liu, Y., Liu, C., Yuan, N.J., Duan, L., Fu, Y., Xiong, H., Xu, S., Wu, J.: Exploiting heterogeneous human mobility patterns for intelligent bus routing. In: 2014 IEEE International Conference on Data Mining, pp. 360–369 (2014)
21. Long, Y., Shen, Z.: Discovering functional zones using bus smart card data and points of interest in Beijing, pp. 193–217 (2015)
22. Machens, A., Gesualdo, F., Rizzo, C., Tozzi, A.E., Barrat, A., Cattuto, C.: An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices. BMC Infect. Dis. **13**(1), 185 (2013)
23. Ning, Z., Xia, F., Ullah, N., Kong, X., Hu, X.: Vehicular social networks: enabling smart mobility. IEEE Commun. Mag. **55**(5), 16–55 (2017)
24. Paik, J.H.: A novel tf-idf weighting scheme for effective ranking. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 343–352. New York, USA (2013)
25. Palchykov, V., Mitrovic, M., Jo, H.H., Saramaki, J., Pan, R.K.: Inferring human mobility using communication patterns. Sci. Report. **4**, 6174 (2014)
26. Pillai, K.G., Angryk, R.A., Banda, J.M., Schuh, M.A., Wylie, T.: Spatio-temporal co-occurrence pattern mining in data sets with evolving regions. In: 2012 IEEE 12th international conference on data mining workshops (ICDMW), pp. 805–812. Brussels, Belgium (2012)
27. Qi, X., Xiao, R., Li, C.G., Qiao, Y., Guo, J., Tang, X.: Pairwise rotation invariant co-occurrence local binary pattern. IEEE Trans. Pattern Anal. Mach. Intell. **36**(11), 2199–2213 (2014)
28. Srinivasan, V., Moghaddam, S., Mukherji, A., Rachuri, K.K., Xu, C., Tapia, E.M.: Mobileminer: Mining your frequent patterns on your phone. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing, pp. 389–400. New York, USA (2014)
29. Sun, M., North, C., Ramakrishnan, N.: A five-level design framework for bicluster visualizations. IEEE Trans. Vis. Comput. Graph. **20**(12), 1713–1722 (2014)
30. Wu, W., Xu, J., Zeng, H., Zheng, Y., Qu, H., Ni, B., Yuan, M., Ni, L.M.: Telcovis: Visual exploration of co-occurrence in urban human mobility based on telco data. IEEE Trans. Vis. Comput. Graph. **22**(1), 935–944 (2016)
31. Yang, Q., Gao, Z., Kong, X., Rahim, A., Wang, J., Xia, F.: Taxi operation optimization based on big traffic data. In: 2015 smart world congress, pp. 127–134. Beijing, China (2015)
32. Yin, H., Cui, B., Chen, L., Hu, Z., Zhang, C.: Modeling location-based user rating profiles for personalized recommendation. ACM Trans. Knowl. Discov. Data **9**(3), 1–41 (2015)
33. Yin, H., Zhou, X., Cui, B., Wang, H., Zheng, K., Nguyen, Q.V.H.: Adapting to user interest drift for poi recommendation. IEEE Trans. Knowl. Data Eng. **28**(10), 2566–2581 (2016)
34. Yin, H., Zhou, X., Shao, Y., Wang, H., Sadiq, S.: Joint modeling of user check-in behaviors for point-of-interest recommendation. In: ACM international on conference on information and knowledge management, pp. 1631–1640, New York, USA (2015)
35. Yuan, J., Zheng, Y., Xie, X.: Discovering regions of different functions in a city using human mobility and pois. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 186–194. New York, NY, USA (2012)

36. Yuan, N.J., Zheng, Y., Xie, X.: Segmentation of urban areas using road networks. MSR-TR-2012–65, Tech. Rep. (2012)
37. Yuan, N.J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., Xiong, H.: Discovering urban functional zones using latent activity trajectories. IEEE Trans. Knowl. Data Eng. **27**(3), 712–725 (2015)
38. Zhong, C., Huang, X., Arisona, S.M., Schmitt, G.: Identifying spatial structure of urban functional centers using travel survey data: A case study of singapore. In: COMP@ SIGSPATIAL, pp. 28–33. New York, USA (2013)
39. Zhong, C., Huang, X., Arisona, S.M., Schmitt, G., Batty, M.: Inferring building functions from a probabilistic model using public transportation data. Comput. Environ. Urban. Syst. **48**, 124–137 (2014)