Reprinted from DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519
Journal of Imaging Science and Technology® 60(2): 020403-1–020403-18, 2016.

Discovering and Visualizing Underlying Traffic Regions from Vehicle Trajectories with Multi-Features

Dongjin Yu, Ruiting Wang, Jiaojiao Wang, and Wanqing Li
College of Computer, Hangzhou Dianzi University, Hangzhou, China
Key Laboratory of Complex Systems Modeling and Simulation, Ministry of Education, China
E-mail: yudj@hdu.edu.cn

Abstract. *City traffic often exhibits regional characteristics, such as large trucks frequently appearing in the suburbs, and the paths to playgrounds on weekends generally being congested. Discovering and visualizing these hidden traffic regions inside which roads share similar characteristics of traffic conditions simplifies the modeling complexities of whole city traffic conditions and therefore contributes significantly toward city planning. Unfortunately, such traffic regions always have irregular shapes and are time varying, which makes their discovery extremely complicated. In addition, establishing a method to visualize and explore the traffic regions interactively still remains challenging. In this article, the authors propose a latent Dirichlet allocation (LDA)-based approach to the discovery of underlying traffic regions (or region topics) from vehicle trajectories captured by surveillance devices installed along roadsides. They treat vehicle trajectories as documents and the values of different traffic features, such as locations, directions, speeds and vehicle types, as the corresponding words. After applying the LDA model, they obtain a list of region topics with combined feature values, in which the different feature values are clustered with probabilistic assignments. Meanwhile, they build a prototype system to explore the surveillance-device-based vehicle trajectories according to the discovered region topics. The prototype system, which consists of map view, cloud view, treemap view and matrix-table view, visualizes the feature values of hidden traffic regions. The authors finally research a real case based on the traffic data in Wenzhou City, a large city in eastern China with a population of more than nine million. They investigate approximately 157 surveillance devices and 750,000 moving vehicles. The case demonstrates the effectiveness of both their proposed approach and the prototype system.* © *2016 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2016.60.2.020403]

INTRODUCTION

Nowadays, as increasing volumes of urban traffic data are captured and become available, opportunities increasingly arise for data-driven analysis that can lead to improvements in traffic conditions. By using such huge amounts of traffic stream data, decision makers can now understand the patterns and trends of traffic flow in different parts of a city. However, the discovery of a method to effectively visualize and investigate the wide range of complex traffic data still remains as one of the great challenges.[1–4]

In this article, we focus on how to explore traffic regions by analyzing massive vehicle trajectories. Here, we use the phrase *traffic region* to represent an area inside which roads share similar traffic features or patterns statistically—these areas typically have irregular shapes and are also time varying. For example, many large trucks appear in some areas of the suburbs, while areas close to popular theme parks are generally congested on weekends. To explore these traffic regions, we examine one new type of traffic trajectory data that is recorded by surveillance devices, such as loop sensors and surveillance cameras. These surveillance devices are now quite common in cities in China. They are installed along roadsides, generally every few hundred meters. They capture the traffic records of each vehicle passing through them, including the plate number, passing speed, passing direction and vehicle type (large vehicle or small vehicle), etc. In a typical medium sized city in China, generally hundreds of such surveillance devices are in use. Each day, millions or tens of millions of passing records are captured, which involve hundreds of thousands of moving vehicles. Differently from the traffic data collected by floating taxis or buses (i.e., the traditional GPS data that are common in the United States), such surveillance-device-based traffic data collected at fixed locations have at least the following advantages when exploring urban traffic conditions. (1) They cover almost all vehicles running on the major roads of the city. In contrast, traditional GPS data are typically restricted to taxis and buses, which are only a small subset of moving vehicles.[4] (2) Because drivers are reluctant to pick up passengers in congested roads or during rush hour, traditional GPS data only reflect the traffic conditions of regions where taxis move frequently. (3) The GPS trajectory data are recorded in a quasi-continuous manner, which can create additional noise due to moving locations, directions and speeds. Unfortunately, it is not always a straightforward task to identify and remove such noise.[5]

To obtain insight into massive surveillance-device-based trajectory data, we propose an analysis method based on the latent Dirichlet allocation (LDA) model, called the traffic LDA (TLDA). The LDA model was originally used in text analyses to study topics from a large corpus of documents that were naturally decomposed into words.[6] In TLDA, we define trajectories as documents and the values of different traffic features as the corresponding words. Here, the traffic features include locations represented by IDs of surveillance devices, vehicles' passing directions, types and speeds. After

Received June 30, 2015; accepted for publication Nov. 16, 2015; published online Jan. 7, 2016. Associate Editor: Song Zhang.
1062-3701/2016/60(2)/020403/18/$25.00

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

applying the TLDA model, we obtain a list of topics with combined feature values defined earlier. We call these topics *region topics*. In these region topics, the different features' values are clustered with probabilistic assignment. In this way, we can discover the hidden traffic regions and rank them by the frequency of the passing records within a time slot.

The discovered traffic regions based on region topics are valuable for city planning. For example, the industrial function of a certain region can be revealed based on the distribution of different vehicle types. More specifically, a region with a greater number of large vehicles verifies its role of carrying cargo transportation. Additionally, the probabilistic distribution of speeds and directions can be used to discriminate the traffic conditions of different regions. Moreover, by ranking the region topics in different periods of time, the traffic trends can be discovered to support decision making, such as designing reversible lanes and selecting sites of subway stations.

Exploration of traffic trajectory data to discover the hidden messages of urban traffic conditions has always been a hot topic in both academia and industry. To the best of our knowledge, to date only Wang et al. have studied how to visualize surveillance-device-based traffic trajectory data.[4] Hong et al. first applied LDA to unsteady flow fields,[7] whereas Chu et al. first used LDA to find the hidden topics from taxi GPS data.[8] These related works inspired us to effectively use the surveillance-device-based trajectory data based on LDA. Differently from the above works, however, we propose a novel LDA-based multi-feature approach that adds the features about the vehicles' directions, types and speeds in addition to the geographic positions of the monitoring surveillance devices. In addition, we develop a set of comprehensive visualization techniques for better understanding of the extracted region topics, such as the map view that shows multiple region features, the treemap view that compares the values of distinct features and the matrix-table view that explores the feature evaluations.

The contributions of our work are fourfold. (1) To the best of our knowledge, this is the first work to use the traffic data captured by surveillance devices to mine the traffic regions (or region topics), which leads to a more accurate result. Unlike the GPS data collected from floating vehicles, the surveillance-device-based data record the trajectories of nearly all moving vehicles on the major roads of a city. On the contrary, methods based on GPS data are prone to being inaccurate to some extent due to the limited coverage of driving routes and the frequent missing of floating GPS data. (2) We consider multiple features that characterize the traffic regions, such as locations, directions, speeds and vehicle types. Compared with traditional approaches based on one single feature such as passing speeds, our approach can reveal traffic regions that hold much more semantic meaning. (3) According to the probability distribution of different feature values, we divide the city areas into several traffic regions with irregular shapes. Such partitioning is much more consistent with the real situation compared with partitioning relying on mandatory approaches. (4) We demonstrate the components for interactive visualization of traffic regions and their corresponding features, which include map view, cloud view, treemap view and matrix-table view. The most extensive result yet reported for the real case verifies the effectiveness of our approach.

The remainder of this article is organized as follows. After presenting the data profile and problem definition in the second section, we propose the traffic latent Dirichlet allocation (TLDA) model that can be used to discover hidden region topics in the third section. We then give some details about how to visualize the results based on TLDA in the fourth section. Next, the fifth section introduces the implementation of a prototype system for discovering and visualizing traffic regions. A case study is shown in the sixth section to demonstrate the effectiveness of our approach and prototype system. Following the discussion and related work given in the seventh and eighth sections, the last section summarizes our work and outlines future research directions.

DATA PROFILE AND PROBLEM DEFINITION

In the past few years, numerous surveillance devices have been set up along the roadsides for traffic monitoring in China. These surveillance devices continuously record vehicles passing through. In our approach, we use two types of traffic datasets: the trajectory dataset and the surveillance device dataset. The trajectory dataset contains a list of vehicle passing records captured by surveillance devices, with attributes of surveillance device ID, plate number, vehicle type (large vehicle or small vehicle), passing speed, passing direction and passing time. Alternatively, the surveillance devices dataset contains the information on all installed surveillance devices, such as IDs, names and their geographic positions represented by longitudes and latitudes.

This article focuses on the traffic datasets in Wenzhou, a large coastal city in the Zhejiang province of southeast China. Each day, our trajectory dataset increases by approximately 550 million passing records, involving about 750,000 moving vehicles in Wenzhou City. In addition, 157 surveillance devices are installed every few hundred meters along the major roads in Wenzhou City. All surveillance devices can identify the vehicle's plate number, passing direction and speed. The vehicle type (large vehicle or small vehicle) can be obtained from the vehicle database by the vehicle's plate number.

To obtain a glimpse of the data profile, we have made some preliminary analyses on the trajectory dataset. We choose Thursday June 5, 2014 as the sampling day. Figure 1 plots the positions of surveillance devices. For each device, we calculate its traffic flow volume as the number of records in the trajectory dataset. This is mapped to the circle size in the map. Figure 2 shows that the total traffic flow volume for every ten minutes reached a high level at 8:00 am, then began to drop down until noon, and again rose slowly to a second peak between 5:00 pm and 6:00 pm. For some roads, the traffic volume changed dramatically with direction. Figure 3(a) presents such a case, in which the surveillance

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

Figure 1. Locations of all surveillance devices in Wenzhou City on June 5, 2014. Each circle represents one device, with its size being proportional to the volume of traffic flow passing through it.

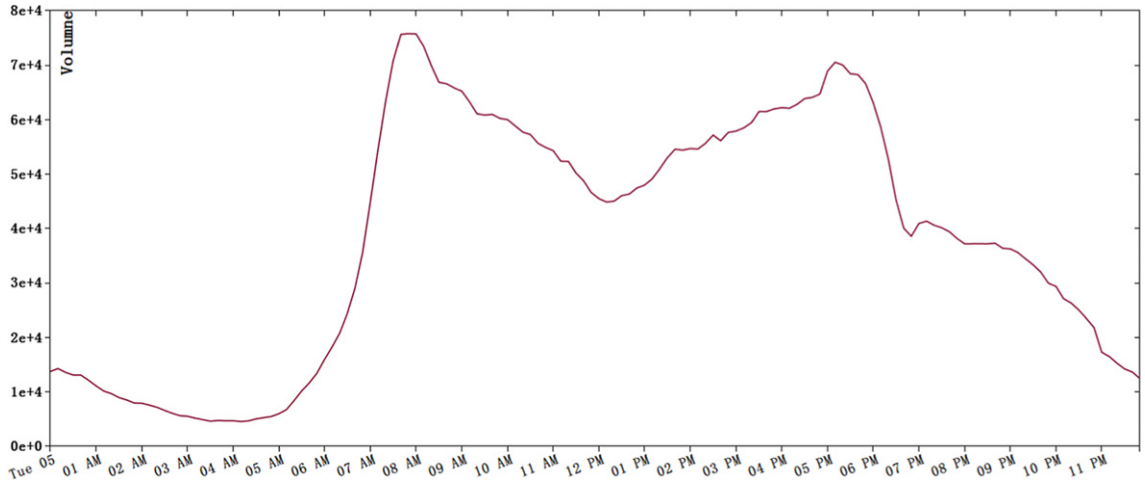
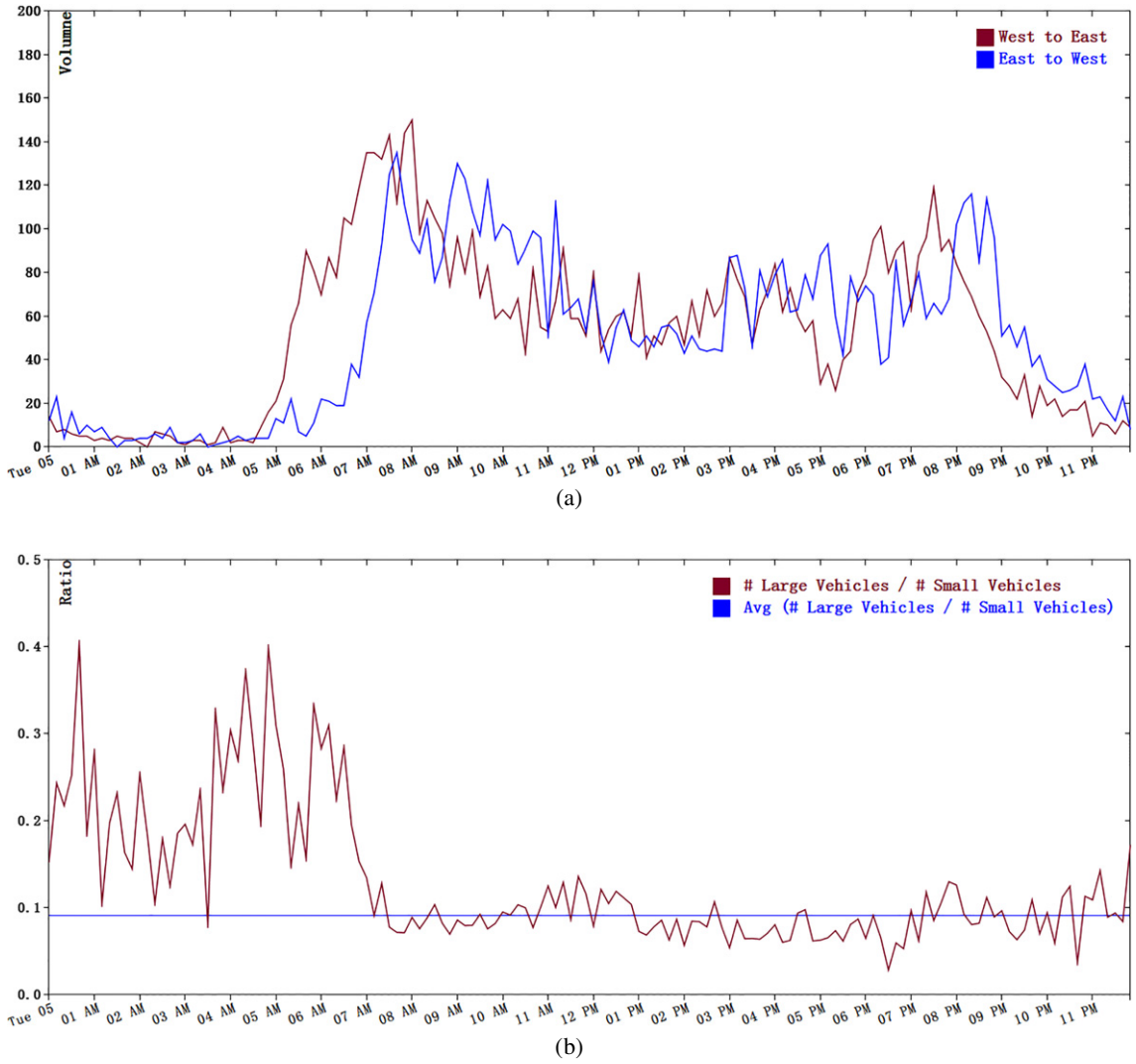


Figure 2. Total traffic flow volume calculated every ten minutes in Wenzhou City on June 5, 2014.

device captured a higher volume from west to east than from east to west between 5:00 am and 8:00 am, and vice versa between 8:00 pm and 10:00 pm. Finally, we choose a surveillance device near the highway off-ramp to observe the ratio between the number of large vehicles and that of small vehicles. Fig. 3(b) shows that the average ratio is 0.096. However, many more large vehicles passed through in the early morning than at any other time. This is possibly because the drivers were busy carrying the cargoes with their large vehicles in the morning.

Although we can obtain the geographic distribution of surveillance devices that have greater volumes of vehicles from the preliminary analysis, we cannot manage to discover the distribution of trajectories, which is much more important for city planning. Here, the trajectory data involve the changing volumes of vehicles in each direction for each vehicle type passing through one specific surveillance device. The regional traffic characteristics may not be obviously revealed from one surveillance device. Instead, a group of surveillance devices may reveal similar characteristics of traffic trajectories passing through them, thus forming a traffic region statistically. Such traffic regions are always hidden with certain probabilities and therefore are very difficult to discover. In summary, the problem of our work can be described as follows: *how to discover the hidden traffic regions that share similar characteristics from surveillance-device-based trajectory data and further explore them in a visualized way.*

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519



(a)

(b)

Figure 3. Traffic volume captured by one surveillance device every ten minutes in Wenzhou City, on June 5, 2014. (a) One-day traffic flow volume in each direction. (b) One-day traffic flow volume ratio for large vehicles to small vehicles.

DISCOVERING TRAFFIC REGIONS

LDA Basics

In natural language processing, LDA is a technique that automatically discovers topics that documents contain. It posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. As an example of a topic model, LDA was first presented by Blei et al. in 2003.[6] Here, we present an intuitive explanation to show how LDA works. The detail can be found in Layman's "Explanation of Topic Modeling with LDA."[9]

Suppose we have the following set of sentences.

Sentence 1: An **apple** contains lots of **vitamins**.

Sentence 2: I like to **eat** an **apple**, but now I prefer *Apple mobile phones*.

Sentence 3: Since *Jobs* has left us, will *Apple* reduce the price?

Given the set of sentences, LDA might classify the bold words under the Topic F, which we might label as "fruit." Similarly, words in italics might be classified under a separate Topic T, which we might label as "technology." Further, it can be inferred from the word count in each sentence that Sentence 1 is of 100% Topic F, Sentence 2 is of 40% Topic F and 60% Topic T, and Sentence 3 is of 100% Topic T.

LDA achieves this in the following three steps.

Step 1: You tell the algorithm how many topics you think there are. You can either use an informed estimate, or simply trial and error.

Step 2: The algorithm will assign every word to a temporary topic in a semi-random manner (according to a Dirichlet distribution, to be exact). This also means that if a word appears twice, each word may be assigned to different topics. Note that in analyzing actual documents, function words (e.g., "an," "I," "but") are removed and not assigned to any topics.

Step 3 (iterative): The algorithm will check and update topic assignments, looping through each word in every document. For each word, its topic assignment is updated by

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

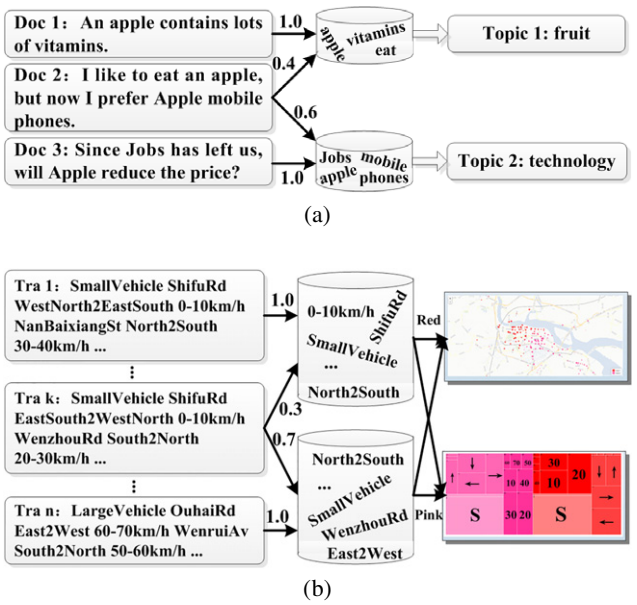


Figure 4. Comparing the traditional LDA model (a) with our TLDA model (b).

weighing conclusions from the following two criteria. How prevalent is that word across topics? How prevalent are topics in the document?

The process of checking topic assignment is repeated for each word in every document, cycling through the entire collection of documents multiple times. This iterative updating is the key feature of LDA that generates a final solution with coherent topics.

The TLDA Model

To use the LDA model for multi-feature analysis in traffic trajectory data, we need to define equivalent LDA concepts at the beginning (Figure 4). We consider vehicle trajectories as the central subjects, which play a similar role to documents in the topic model. We then consider the values of each feature as words, with trajectories being bags of feature values, which is analogous to the concept of documents being bags of words in the topic model. For any facet x of the traffic trajectory data, such as location, direction, speed or vehicle type, we assign a feature value set F_x to describe its possible values. In this way, any trajectories sharing similar characteristics would frequently have the same feature values. All feature value sets F_x can then be united to generate a feature value vocabulary.

Having the defined feature value vocabulary, we can transform the trajectories at a time slot into sequences that contain multiple feature values. Afterwards, the sequences are used as the input to estimate the underlying TLDA model. In this way, the region topics, the distribution of region topics per trajectory, and the distribution of features per region topic can be generated.

For the convenience of the reader, Table I summarizes all important notations used in this article.

As mentioned earlier, the LDA model typically can be used to analyze topics in a corpus of documents. Similarly,

Table I. Key notations and their descriptions.

Notation	Definition and brief description
D	The total number of trajectories.
K	The total number of region topics.
N	The total number of iterations.
V	The set of all possible feature values.
d_j	The jth trajectory.
$\vec{d}$	The corpus of trajectories, i.e., $\vec{d} = \{d_j\}_{j=1}^{D}$.
W_j	The total number of feature values in d_j, i.e., $W_j = \|d_j\|$.
W	The total number of feature values in all trajectories, i.e., $W = \sum_{j=1}^{D} W_j$.
$\omega_{i,j}$	The jth feature value in the trajectory d_i.
$z_{i,j}$	The region topic assignment for the jth feature value in the trajectory d_i.
θ_j	The distribution probability of region topics for the jth trajectory.
$\vec{\theta} = \{\theta_m\}_{m=1}^{D}$	The $D \times K$ distribution probability matrix for D trajectories and K region topics.
φ_k	The distribution probability of feature values for the region topic k.
$\vec{\varphi} = \{\varphi_k\}_{k=1}^{K}$	The $K \times V$ distribution probability matrix for K region topics and V feature values.
α	The Dirichlet prior for θ_j.
β	The Dirichlet prior for φ_k.
x	One certain traffic feature, such as passing direction, passing speed, etc.
F_x	The set of possible feature values on feature x.
$\vec{\varphi}_x$	The $K \times \|F_x\|$ distribution probability matrix for K region topics and $\|F_x\|$ feature values on feature x.
S_k	The sampling frequency for the region topic k.
$\vec{S}$	The sampling frequencies for all region topics, i.e., $\vec{S} = \{S_k\}$.

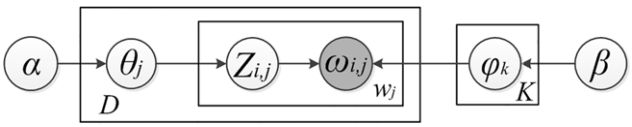


Figure 5. The model of the region topic.

any trajectory d_j can be modeled as a mixture of K region topics, while any region topic k can be characterized by a multinomial distribution φ_k over feature values V. Here, among all variables, only $\omega_{i,j}$ is observable, while others like $z_{i,j}$, θ_j and φ_k are all latent variables. We generate the observation of these latent variables using the following process presented in Figure 5 for the related model.

Step 1: For every trajectory d_j, draw a region topic distribution θ_j from a Dirichlet prior with parameter α, i.e., $\theta_j \sim Dir(\alpha)$, where $j \in \{1, \ldots, D\}$.

Step 2: For every region topic k, draw a feature value distribution φ_k from a Dirichlet prior with parameter β, i.e., $\varphi_k \sim Dir(\beta)$, where $k \in \{1, \ldots, K\}$.

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

Step 3: For a feature value position i in the jth trajectory, where $i \in \{1, \ldots, W_j\}$ and $j \in \{1, \ldots, D\}$, choose a region topic for this position $z_{i,j} = k \sim Multinomial(\theta_j)$ and a feature value from this chosen region topic $\omega_{i,j} = \|V\| \sim Multinomial(\varphi_{z_{i,j}})$.

After the generative process is determined, the joint ability of the model can be described as

$$P\left(\omega_{i,j}, z_{i,j}, \theta_{i,j}; \alpha, \beta\right) = \prod_{k=1}^{K} P\left(\varphi_k; \beta\right) \prod_{j=1}^{M} P\left(\theta_j; \alpha\right) \prod_{i=1}^{W_j} P\left(z_{i,j} \mid \theta_j\right) P\left(\omega_{i,j} \mid \varphi_{z_{i,j}}\right). \tag{1}$$

Therefore, the whole model estimation process serves simply to maximize the above likelihood function by *Bayesian* inference with parameters α and β. In our work, we adopt an implementation using Gibbs sampling.[10]

Algorithm 1 (TLDA) can be divided into the following three steps.

(1) Model initialization (from line 1 to 31).

We first randomly initialize the region topic ID for each feature value about trajectories, so that the ID comes from the uniform distribution in 1 to K. Afterwards, we count the total number of feature values about each trajectory that appears in each region topic (*trajectory2topic*), the total number of each feature value appearing in each region topic (*value2topic*), the number of feature values appearing in each region topic (*values2topic*) and the number of feature values appearing in each trajectory (*values2trajectory*).

(2) Gibbs sampling (from line 45 to 72).

This step runs N times to obtain the converging results about the sampling frequency on each region topic (*values2topic*) and the feature value probability distribution on each region topic (*phi*). One iteration consists of the following four substeps.

- Obtain the region topic ID of every feature value from each trajectory in the previous iteration. Then update the counting number of each feature value in each region topic (*value2topic*), the counting number of each trajectory in each region topic (*trajectory2topic*), the total counting number of feature values in each region topic (*values2topic*) and the total counting number of feature values in each trajectory (*values2trajectory*) (from 47 to 55).
- Calculate the probability of each feature value belonging to a different region topic, and use the polynomial probability distribution about each region topic to sample the new region topic ID of one certain feature value (lines 56–58).
- Use the topic ID generated in the previous substep to update *value2topic*, *trajectory2topic*, *values2topic* and *values2trajectory* (lines 59–64).
- Update the probability distribution of each trajectory in each region topic (*theta*) (lines 65–67), and the probability distribution of each value in each region topic (*phi*) (lines 68–70).

(3) Results processing (lines 35–43).

The feature values are divided by different features (i.e., dividing $\vec{\varphi}$ into different features of $\vec{\varphi}_x$ whose rows are normalized).

The Evolution of Region Topics

To link the region topics of different time slots, we first divide the K areas according to our experience that one surveillance device is only located in one area. The region topics are then created for given time slots, which reflect the different traffic patterns for the different periods of a day. For example, a significant difference in traffic patterns usually exists between the rush hour and other times. Thus, we define a similarity between the region topics discovered through TLDA and the areas defined in advance at one time slot. Given one *Region Topic i* and one area *j*, their similarity can be computed as

$$S_{i,j} = \frac{Size\left(T_i \cap T_j\right)}{Size\left(T_i \cup T_j\right)}, \tag{2}$$

where T_i is the set of surveillance devices with a higher probability over a predefined threshold in area i, or $T_i = \{w \mid P(w \mid i) > c\}$, and T_j is the set of surveillance devices in area j. Based on the similarities, each region topic can be assigned to one most closely related area. The sum of similarities of such most closely related area pairs can therefore reflect the abnormal condition of a time slot with which we are actually concerned.

VISUALIZING TRAFFIC REGIONS

As described in the previous section, the visual interface consists of four components: the map view, the cloud view, the treemap view and the matrix-table view, as presented in Figure 6. These visualization components are orchestrated in a coordinated manner for effective user exploration.

Map View

The location feature represented by the surveillance device ID indicates the geographic characteristic of region topics. Various backgrounds, such as topography, satellite and transportation maps, can be attached to provide visual cues and context to geographic and cultural information. Messages such as tourist attractions, subway stations and government buildings can also be placed on the map. To visualize one region topic's location feature, a surveillance device that belongs to a certain region topic with a probability value greater than a predefined threshold is drawn as a circle with a given region topic color, whose size is determined by its importance, or the proportion of traffic volume passing through it over the whole traffic volume in a particular region topic. For simplicity, we use the words *surveillance device* or simply *device* to represent its corresponding circle on the map.

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

Algorithm 1: TLDA

Input : $\vec{d}$, α, β, and N
Output : $\vec{\varphi}_x$: $x \in \{cell-id, type, direction, type\}$, $\vec{S}$

```
//initialize the model
for i = 1 to V, k = 1 to K do
  // initialize the feature-value/region-topic matrix
  value2topic[i][k] = 0;
end for
for j = 1 to D, k = 1 to K do
  //initialize the trajectory/region-topic matrix
  trajectory2topic[j][k] = 0;
end for
for k = 1 to K do
  //initialize the feature-value/region-topic array
  values2topic[k] = 0;
end for
for j = 1 to D do
  //initialize the feature-value/trajectory array
  values2trajectory[j] = 0;
end for
for j = 1 to D, k = 1 to K do
  theta[D][K] = 0;
end for
for k = 1 to K, i = 1 to V do
  phi[K][V] = 0;
end for
for j = 1 to D, i = 1 to W_j do
  k = rand()%K + 1;
  //obtain the index in feature value set about the i^th
  feature value in j^th trajectry
  valueIndex = getIndexFromValues(i, j);
  value2topic[valueIndex][k] + +;
  trajectory2topic[j][k] + +;
  values2topic[k] + +;
  values2trajectory[j] + +;
  //Gibbs sampling process
  Call GibbsSampling(value2topic, value2topic,
       values2topic, values2trajectory, theta, phi);
end for
//results processing
for each feature x do
  for i = 1 to V do
    if valueIndex(i) ∈ F_x then
      add the i^th row of phi into φ_x
    end if
    normalize each row in φ_x
  end for
end for
return S (total_values2topic), φ_x ;
```

GibbsSampling
Input : $value2topic$, $value2topic$, $values2topic$, $values2trajectory$, $theta(\vec{\theta})$, $phi(\vec{\varphi})$
Output : $value2topic$, $value2topic$, $values2topic$, $values2trajectory$, $theta(\vec{\theta})$, $phi(\vec{\varphi})$

```
for n = 1 to N do
  for j = 1 to D, i = 1 to W_j do
    //obtain topic ID in previous iteration on i^th feature value
    in j^th trajectory
    topicID = getTopicIDFromPreviousIteration(i, j);
    //obtain index in feature value set on i^th feature value in
    j^th trajectory
    valueIndex = getIndexFromValues(i, j);
    value2topic[valueIndex][topicID]--;
    trajectory2topic[j][topicID]--;
    values2topic[topicID]--;
    values2trajectory[j]--;
  end for
  for k = 1 to K do
  probablity[k] = (value2topic[i][k]+β)/(values2topic[k]+V×β) × (trajectory2topic[j][k]+α)/(values2trajectory[j]+K×α);
  end for
  //obtain topic ID from polynomial probability dist. on
  i^th value in j^th trajectory
  topicID = getValueTopicID(probablity);
  value2topic[valueIndex][topicID]++;
  trajectory2topic[j][topicID]++;
  values2topic[topicID]++;
  values2trajectory[j]++;
  for j = 1 to D, k = 1 to K do
    theta[j][k] = (trajectory2topic[j][k]+α)/(values2trajectory[j]+K×α) ;
  end for
  for k = 1 to K, i = 1 to V do
    phi[k][i] = (value2topic[i][k]+β)/(values2topic[k]+V×β);
  end for
end for
return value2topic, value2topic, values2topic,
       values2trajectory, theta(θ), phi(φ)
```

Additionally, a surveillance device may be indicated by multiple overlapped circles with different region topic colors. To improve the visual effects of such overlapped topics, the transparency of a circle can be set to 0.8 so that it can be easily observed. Fig. 6(a) displays the location features of four region topics generated from TLDA on the map of Wenzhou City. Users can manipulate the appearance of the region topics by setting layers as visible/invisible and changing the color and transparency.

Cloud View

We use a word cloud to display semantic information and the relationships between different surveillance devices. The size of a surveillance device name reflects its importance (i.e., the frequency of its occurrence in all trajectories of one region

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

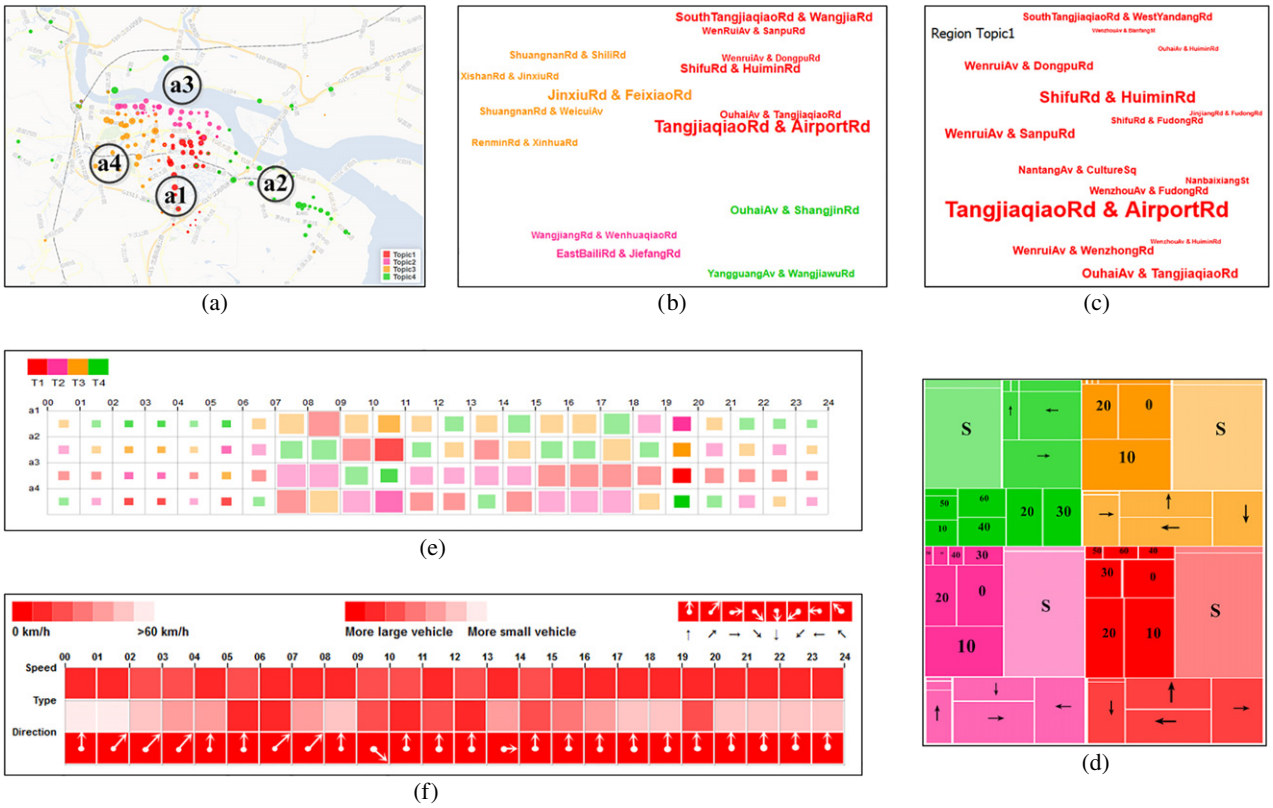


Figure 6. Visual interface. (a) Map view of region topics. The circles marked with "a1," "a2," "a3" and "a4" are mandatory selected areas used for comparison with discovered region topics. (b) Cloud view of the 15 surveillance devices with the greatest traffic volumes captured in all region topics. (c) Cloud view of the 15 surveillance devices with the greatest traffic volume in *region topic 1*. (d) Treemap view of region topics. (e) Matrix-table view showing the change of importance of different areas during the entire day. (f) Matrix-table view showing the change of feature values in area a3 during the entire day.

topic). The names are not randomly positioned. Instead, the distance between two names reflects their relationship by introducing the attracting and repelling forces computed from pairwise cosine similarities:

$$sim(c_i, c_j) = \frac{\sum_{k=0}^{K-1} p(c_i \mid t_k) \cdot p(c_j \mid t_k)}{\sqrt{\sum_{k=0}^{K-1} p(c_i \mid t_k)^2} \cdot \sqrt{\sum_{k=0}^{K-1} p(c_j \mid t_k)^2}}, \quad (3)$$

where c_i or c_j denotes one surveillance device, K is the total number of region topics and $p(c \mid t)$ is the conditional distribution probability of the device of c given the region topic of t. In other words, surveillance devices with similar region topic distributions are clustered together.

Furthermore, in the cloud of a given region topic, the names of surveillance devices are aggregated together if they are frequently passed through by vehicles traveling inside this region topic. The cloud view can also show the most important, or the most frequently passed through, surveillance devices among all region topics, or those selected by users. Fig. 6(b) shows the top 15 surveillance devices with the largest frequencies in all region topics, whereas Fig. 6(c) shows the same but only in *Region Topic 1*. Here, we use the names of roads where surveillance devices are installed to indicate the devices' names.

Treemap View

The treemap view helps users to discover the hierarchical data,[11] while exploring the proportion of sampling frequency in each region topic and the values of each feature. Fig. 6(d) visualizes all feature values for all region topics over a treemap, where each tone denotes one specific region topic and its color represents one specific feature; each rectangle shows one specific feature value for different features. In other words, the treemap has three hierarchies, as follows. The first hierarchy (i.e., the region topic hierarchy) has the same tones as those of the region topic indicated in the map view. The area sizes of region topics with different tones represent the total sampling frequency of all vehicle records, which are assigned as busy degree to this region topic. The second hierarchy that shows region topic features is subdivided equally and distinguished by different colors. The last hierarchy shows different values for each feature. It consists of many small rectangles marked with their corresponding values. Here, the "L" and "S" signs indicate *large* vehicles and *small* vehicles, respectively, regarding the vehicle type feature. In addition, the arrow signs denote the direction feature, whereas the numbers represent the ranges of the speed feature. Generally, the treemap helps users to acquaint themselves with the proportion of vehicle volumes appearing in each region topic, and the proportion of various

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

feature values in each region topic. It also helps to facilitate the comparison of various feature values under different region topics intuitively.

To create a treemap, one must define a tiling algorithm, that is, a way to divide a rectangle into subrectangles of specified areas. Ideally, a treemap algorithm would create rectangles with aspect ratios (the proportional relationship between the width and the height) close to one, which means squares; furthermore the algorithm would preserve some sense of the ordering in the input data. Unfortunately, these two properties have an inverse relationship. As the aspect ratio is optimized, the order of placement becomes less predictable. Conversely, as the order becomes more stable, the aspect ratio is degraded. It is commonly known that human vision is not good at recognizing area sizes, especially for long and thin (non-square) rectangles. Moreover, in our treemap view, we need to label the feature values on the leaf node rectangles, which is a difficult task for long and thin rectangles. In our approach, therefore, we prefer optimized aspect ratios, or more square rectangles, to a stable placement order. We apply the tiling algorithm Squarified.[12] Differently from many other ordered tiling algorithms such as Slice and Dice, BinaryTree and Ordered, Squarified tends to have square leaf nodes in the treemap view, although its order of placement is less predictable. To compensate for its unordered placement, we use different tones and colors for different rectangles in our approach.

Matrix-Table View

The matrix-table view shows the temporal patterns for specific region topics, as illustrated in Fig. 6(e) and (f). It has two types of tables: the region topic evolution table and the feature evolution table. Fig. 6(e) visualizes the region topic evolution table, where each row represents one specific area that is described in Fig. 6(a) and each column shows the areas' corresponding region topics at different time slots with importance in different tones which are the same as those in the map view. To emphasize the exceptional patterns related to correlative time slots, we use transparency to improve the visual effect. A more opaque column represents that the region topics discovered at this time slot (column) change more dramatically than other region topics discovered at other time slots. In addition, the size of the unit stands for the sampling frequency of the region topic at a time slot.

In a similar way, Fig. 6(f) shows the feature evolution table, in which the columns represent time slots and the values for the speed and type features are encoded into the matrix with different tone scales. Here, a deeper color represents a lower speed or a greater number of large vehicles, and a lighter color a higher speed or a greater number of small vehicles. Meanwhile, the periodic values of the direction feature are distinguished by the arrows pointing in eight different directions.

Furthermore, in the matrix-table view, the user can manipulate the region topics in the region evolution table, whose corresponding feature evolutions can be displayed simultaneously in the feature evolution table.

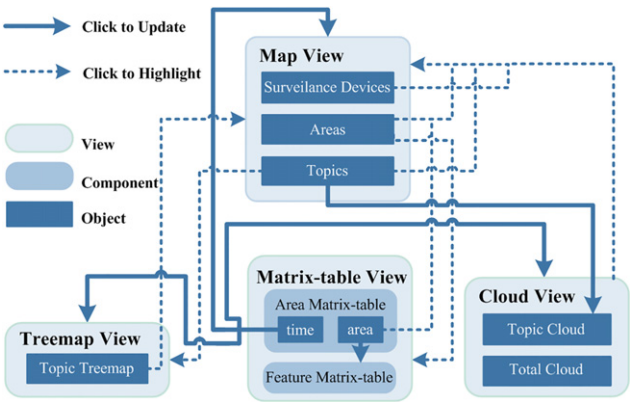


Figure 7. The interaction among different views in our system.

Interactions among Views

The map view, cloud view, treemap view and matrix-table view are fully interactively related (Figure 7). By default, all of the views show the traffic condition for the most recent time slot. If the user clicks on one certain time slot in the past in the matrix-table view, all other views are automatically updated to reflect the traffic condition for that time slot. If the user clicks on one certain area in the region topic evolution table of the matrix-table view, such as a1, a2, a3, or a4 in Fig. 6(e), this area will be automatically highlighted in the map view with its geographic information attached, and its related feature evolution table will be updated simultaneously. In addition, if the user clicks on one certain traffic region in the treemap view, the circles representing the surveillance devices in this traffic region will be automatically highlighted in the map view. Moreover, if the user clicks on one certain traffic region in the map view, the cloud view will automatically show the top 15 most frequently passed-by surveillance devices for this traffic region. Alternatively, if the user clicks on the name of a surveillance device in the cloud view, the circle representing this surveillance device will be automatically highlighted in the map view.

SYSTEM PIPELINE

Based on the TLDA model, we implement a prototype system to support the visualization and exploration of traffic trajectory data (Figure 8).

Meanwhile, we determine four facets (F_{location}, $F_{\text{direction}}$, F_{type} and F_{speed}), whose possible values are listed in Table II.

During our preprocessing phase, we obtain each vehicle's trajectory from the raw trajectory dataset at different time intervals, including the ID of the surveillance device, passing direction, passing speed and vehicle type.

We implement the computation of TLDA based on JGibbLDA, a Java implementation of latent Dirichlet allocation (LDA) using Gibbs sampling for parameter estimation and inference.[13] To use the TLDA model, a set of parameters must be preset, including the number of topics K, the Dirichlet prior on per-document topic distribution α, the Dirichlet prior on per-topic word distribution β and

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

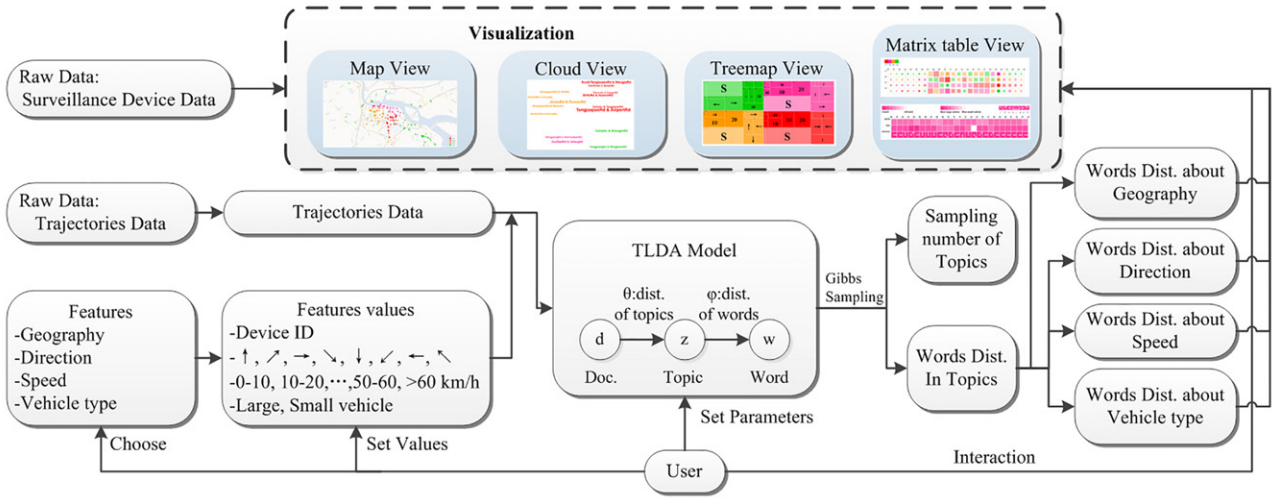


Figure 8. Prototype of our system. The user chooses the features and initially sets the value for each feature. Then, trajectories are extracted with these values. After setting the TLDA parameters, the Gibbs sampling result of the model, including the sampling frequency of region topics and their distributions of feature values, can be generated. The system offers multiple interactive views to visualize the results.

Table II. Possible values for each facet.

Facet	Possible values
$F_{location}$	The ID of the surveillance device
$F_{direction}$	West to east, east to west, south to north, north to south, southeast to northwest, northwest to southeast, northeast to southwest and southwest to northeast (denoted by →, ←, ↑, ↓, ↖, ↘, ↙, ↗, respectively)
F_{type}	Large vehicle, small vehicle
F_{speed}	0–10 km/h, 10–20 km/h, 20–30 km/h, 30–40 km/h, 40–50 km/h, 50–60 km/h and >60 km/h

the number of iterations N. We specify these particular parameters as Tang et al. describe in Ref. 14. By applying TLDA on the trajectory data, we can acquire both the sampling frequencies of the appearances of feature values and their distribution in each region topic. These sampling frequencies can be further used to rank the importance of different region topics.

The visualization part consists of four subparts. The first subpart, or *map view*, presents the surveillance devices whose probabilities in one region topic are greater than a predefined threshold, with the map attached to show their locations. In the second subpart, we use the *cloud view* to analyze the relationship between surveillance devices. The third subpart supplies the *treemap view* to compare the traffic importance among different region topics, and the proportion of different feature values in each region topic. In the last subpart, we use the *matrix-table view* to display the changes in the importance of each region and the overall values of different features with time changing.

A CASE STUDY

To evaluate the effectiveness of our approach and the prototype system, we conducted an extensive experiment in Wenzhou City, a large city in eastern China with a population of more than nine million. We used the real traffic trajectory data collected from 157 surveillance devices in Wenzhou City and 436 surveillance devices in its neighboring counties on June 5, 2014. We set the number of region topics to six in Wenzhou City and its neighboring counties, and to four in Wenzhou City, and we set the time slot to one hour. In other words, region topics are generated every 60 min for the entire day (i.e., 0:00 am–1:00 am, 1:00 am–2:00 am, . . . , 11:00 pm–00:00 am). This section illustrates four cases in detail. The first case shows the discovered region topics, whereas the second reveals the vehicle moving patterns. The last two cases show the features of the discovered region topics and their evolution.

Discovered Region Topics

Tables III and IV illustrate four discovered region topics from 157 surveillance devices in Wenzhou City between 8:00 am and 9:00 am, June 5, 2014.

Region Topic Importance

The total frequency of all vehicles' occurrences in a given region topic represents this region topic's importance. In Table III, the region topics are ranked by importance (i.e., the total number of moving records) from *Region Topic 1* to *Region Topic 4*. *Region Topic 1*, the busiest one, has 295,823 moving records, whereas *Region Topic 4*, the least busy one, has 218,150 moving records. This result indicates that these four topics actually have small differences in their feature values.

Probability Distribution of Feature Values

For each region topic, a feature value also has its distinct frequency of occurrence, which we refer to as *value importance* over the topic. In other words, a value importance, $P(w \mid z)$, represents the appearance probability of a feature value w

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

Table III. TLDA results: the total number of moving records in a given region topic represents this region topic's importance; every feature value has a different frequency of appearance in a given region topic, defined by the probability $P(w \mid z)$.

	# Moving records	The appearance probability of feature values				
		Device ID: 330302023	Speed: 10–20 km/h	Direction: north–south	Vehicle type: small vehicle	...
Region topic 1	295,823	0.019	0.071	0.099	0.109	...
Region topic 2	272,364	0.0009	0.079	0.046	0.110	...
Region topic 3	258,064	0	0.068	0.095	0.155	...
Region topic 4	218,150	0	0.014	0.034	0.156	...

Table IV. TLDA results: the probability distributions of region topics for a given vehicle trajectory $P(z \mid d)$. Real plate numbers are not shown for protection of privacy.

Trajectory	Region topic 1	Region topic 2	Region topic 3	Region Topic 4
C#####1	0.484375	0.484375	0.015625	0.015625
C#####2	0.015625	0.171875	0.171875	0.640255
C#####3	0.294117	0.382354	0.294118	0.029411
C#####4	0.018518	0.018518	0.018518	0.944446

(in one feature x) over a given *Region Topic* z. Here, the sum of *value importances* of all feature values belonging to one certain feature is normalized to 1. Table III displays the distribution of feature values {device ID = 33030223; speed = 10–20 km/h; direction = north–south; vehicle type = small vehicle}, or value importance, over four region topics, as an example.

Trajectory Probability Distribution

Each trajectory has its own probability distributions over multiple region topics. We use $P(z \mid d)$ to denote the distribution probability for a *Region Topic* z over a given vehicle trajectory d. Table IV shows the probability distributions of four region topics for four vehicle trajectories, which can be further used for trajectory clustering.

Vehicle Moving Patterns

The discovered region topics demonstrate some typical traveling patterns of vehicles. Figure 9 displays four region topics generated for the time slots of 3:00 am to 4:00 am and 8:00 am to 9:00 am

One can observe the surveillance devices in the circle marked with "I" in Fig. 9(a). The green devices with large sizes indicate that they are the most frequently passed-by ones in *Region Topic 4*. Although these green devices are separated by two pink devices (*GaoxiangRd* marked with "GX" and *LiuhongqiaoRd* & *ZhanqianRd* marked with "LZ") in geography, the devices within the circle marked with "I" turn out to have the same color in Fig. 9(b). Combined with their cloud views in Fig. 9(c) and (d), the *GaoxiangRd* and *LiuhongqiaoRd* & *ZhanqianRd* devices were far away from those devices at 3:00 am to 4:00 am, but the distances decreased at 8:00 am to 9:00 am. Consequently, we can infer that vehicles seldom traveled between the eastern and western green devices through the pink devices from 3:00 am to 4:00 am, but did travel from 8:00 am to 9:00 am. Meanwhile, from their relevant treemaps on the direction feature, we can see that the passing directions in *Region Topic 4* between 3:00 am and 4:00 am were mainly from north to south or from south to north, as shown in Fig. 9(a), but were almost equally distributed between 8:00 am and 9:00 am, as presented in Fig. 9(b) for the same area, which is consistent with the corresponding map views.

A large residential zone in the northeastern part of Wenzhou City is identified by a circle marked with "II" in Fig. 9. This zone in Fig. 9(a) has a smaller number of devices with smaller sizes compared with that in Fig. 9(b), from which we can draw the conclusion that people rarely went out in the early morning in this zone. In addition, the surveillance devices with larger sizes in Fig. 9(b) prove that they were just experiencing the morning rush hour. The circle marked with "III" in Fig. 9(a) corresponds to the zone for entering and leaving Wenzhou City. From the color of the devices, we can see the presence of two routes (i.e., DongOu Avenue and Kanghua Road). The large sizes of the devices indicate that Kanghua Road was more occupied compared with other routes at that time.

Let us focus on the surveillance devices with various colors on the map. First, the device overlaid by three colors, with a symbol "a" in Fig. 9(a), indicates that it is an important conjunction of three groups of vehicles. Second, a large device in both *Region Topic 1* and *Region Topic 2* at the intersection of Airport Avenue and Jinxiu Road is marked with "b" in Fig. 9(b). The regions for *Region Topic 1* and *Region Topic 2* are adjacent to each other, at the intersection of which many devices are geographically near. Among these devices, only that marked with b is overlaid with two colors (i.e., red and pink); it connects different functional regions.

Analysis of Passing Direction, Speed and Vehicle Type

We further analyze the sampling frequency of each topic, the indicator of busy degree of regions, and the proportion of each of the feature values. Fig. 6(a) and (d) illustrate the location feature on the map, and the speed, direction and type features on the treemap at the time slot from 8:00 am to 9:00 am, respectively. Little difference is seen for the sizes of the red, pink and orange rectangles, whereas the green rectangle has the minimum size, as Fig. 6(d) illustrates. From the related map view in Fig. 6(a), we know that the green topic covers the areas around the airport and the roads linking the airport and downtown, in addition to the Wenzhou expressway. These areas are less busy compared with other traffic regions at morning rush hour.

Meanwhile, Fig. 6(d) uses three different colors, or different scales of tone, to display the features of passing

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

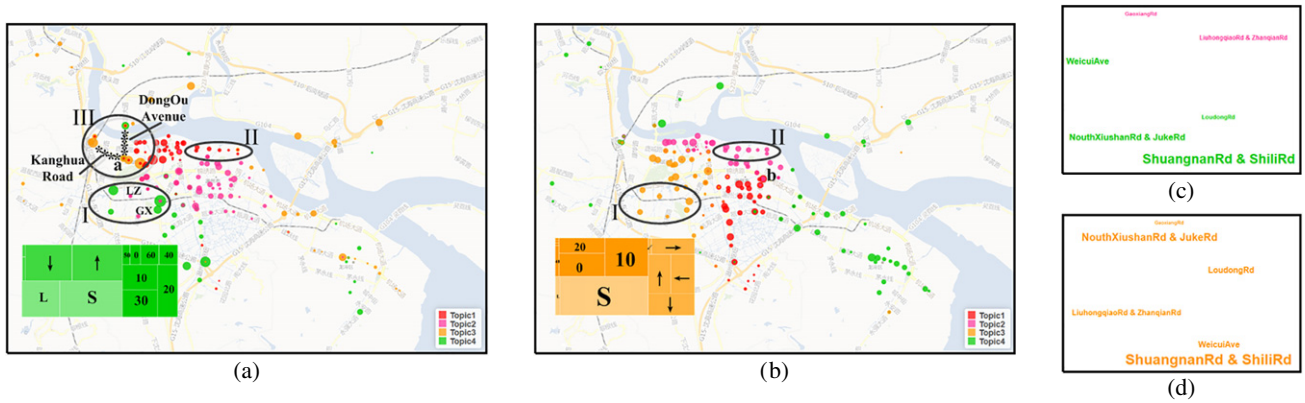


Figure 9. Change of traffic conditions between (a) 3:00 am and 4:00 am and (b) 8:00 am and 9:00 am. Their cloud views about surveillance devices in "I" are given in (c) and (d), respectively.

direction, speed and vehicle type. The darkest one shows the values of the speed feature. The values of "10," "20" and "30" occupy the main and similar proportions in the red, pink and orange regions. However, the value of high speed has a larger proportion in the green region. We can therefore conclude that the regions in the central part of Wenzhou City were all congested except for the suburbs at that time. For the direction feature, from the bottom right rectangle in Fig. 6(d), we can find that the number of vehicles passing from east to west is roughly the same as that from west to east, but more vehicles are passing from north to south than south to north. This shows that the direction from north to south was the main traffic flow direction at this time slot. For the type feature, large vehicles have the maximum number in the green region topic. This is consistent with what would commonly be expected (i.e., large vehicles always appear more frequently in expressway or suburban areas).

Evolution of Region Topics

Fig. 6(f) shows the change of importance of different areas through related region topics during the entire day. From Fig. 6(f), we find horizontally that the traffic volume of each area reaches a high level at the time slot from 8:00 am to 9:00 am, and maintains a high level until the time slot from 6:00 pm to 7:00 pm. The minimum occurs between 4:00 am and 5:00 am, whereas a small decrease occurs between 11:00 am and 2:00 pm. The area marked with 'a3 in Fig. 6(a) can be regarded as the most important area because it is covered by more red and pink colors. Combined with the geographic information about the area a3 we know that this area is the major business center. In addition, observing the transparency of each column, we find that five columns have a smaller transparency, and the column at the time slot from 7:00 pm to 8:00 pm has the smallest.

Our system also supports the visualization of overall feature values for each region topic about the most similar area at different time slots. If a user selects an area in the matrix-table view, the evolution of its feature values will appear simultaneously. Figures 10(b) and (c) show the feature evolution about the areas a1 and a2, respectively. They indicate that the areas a1 and a2 were less busy compared with the other two areas (i.e., the areas a3 and a4), if considering the entire day. This is the reason why their features in the feature evolution tables are illustrated in yellow and green. In the first line of Fig. 10(b), a deeper tone at the time slot from 7:00 am to 8:00 pm than other time slots indicates that the increasing traffic volume reduces the overall speed in daytime. By contrast, the first line in Fig. 10(c) has much shallower tones, indicating that a2 always remains unblocked, simply because it is relatively remote from the downtown area. In addition, the second lines in Fig. 10(b) and (c) represent the proportion of the number of small vehicles to that of large vehicles. The tone at the time slot from 8:00 am to 9:00 am in Fig. 10(b) looks extremely pale, which verifies the fact that it is the time for driving to work. On the contrary, the tone at the time slots from 3:00 am to 5:00 am in Fig. 10(c) looks very deep, which indicates the presence of many large vehicles at that time. Finally, the last lines in Fig. 10(b) and (c) demonstrate the directions of traffic flow. It can be concluded that the direction from south to north in these two areas has greater traffic volume than other directions by considering the arrow directions.

DISCUSSION

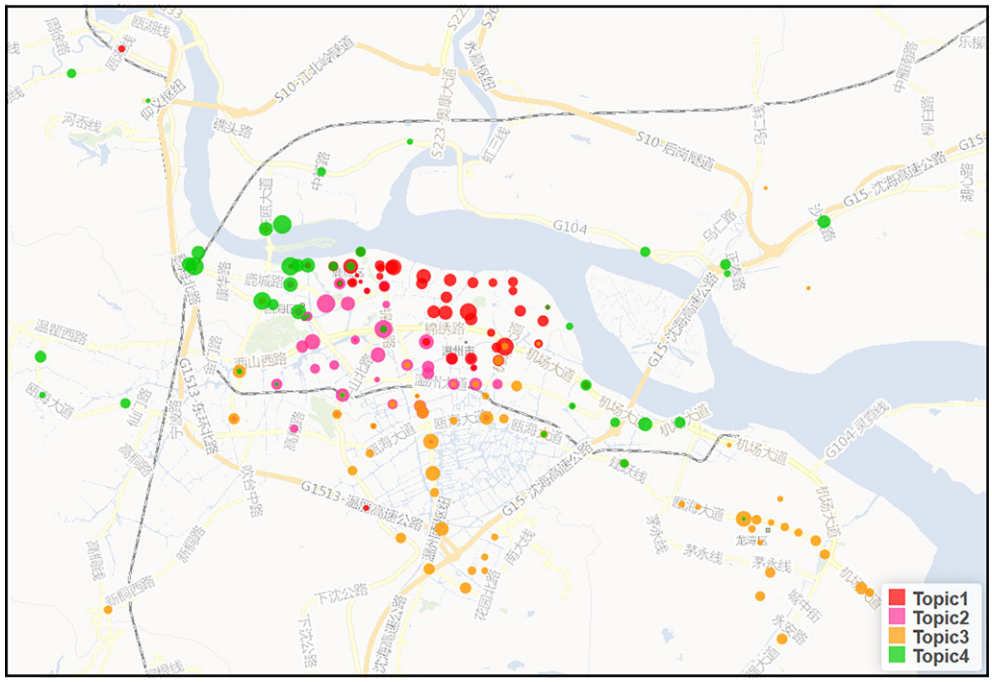
In this section, we discuss the influence of setting different TLDA parameters on final results, evaluate the effectiveness and compare our TLDA model with other approaches from two aspects.

Setting TLDA Parameters

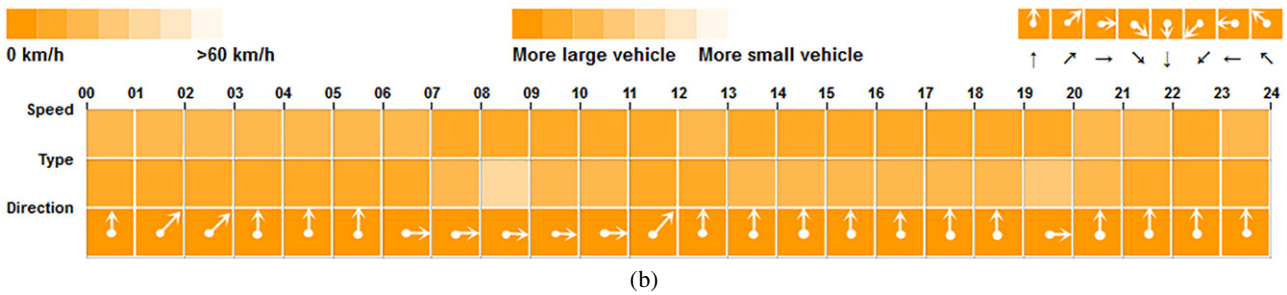
In the TLDA model, the set of parameters can be adjusted to tune the final results. These parameters include the number of region topics (K), the Dirichlet prior on the distribution probability of trajectories over region topics (α), the Dirichlet prior on the distribution probability of region topics over feature values (β) and the number of iterations (N).

Generally, the number of region topics, or K, is difficult to determine since it greatly depends on the actual traffic data. Too few region topics will miss some important patterns, while too many region topics may produce redundant,

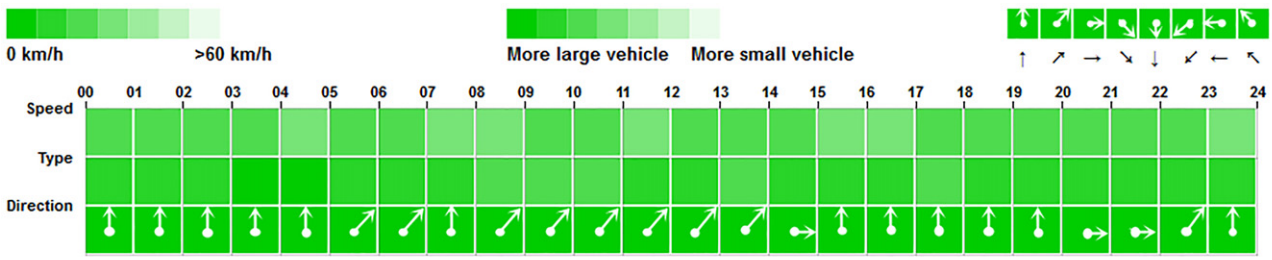
DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519



(a)



(b)



(c)

Figure 10. Map view at 7:00 pm to 8:00 pm. (b) Summary of feature values in area a1 in Fig. 6(a) during the entire day. (c) Summary of feature value in area a2 in Fig. 6(a) during the whole day.

meaningless or trivial results. Meanwhile, the Dirichlet priors α and β influence the region topic distribution per trajectory and the feature value distribution per region topic, respectively. A smaller value would typically make the distribution more concentrated, and vice versa. However, in our work, we do not observe significant differences in the results. Finally, the number of iterations, or N, also affects the quality of results. We found that the output converges quickly within thousands of iterations. In our case, we set the number of iterations to 1000, which is a balance between time efficiency and the quality of results.

We set $K = 4$, $\alpha = 0.1$, $\beta = 0.01$ and $N = 1000$ as the default settings. From the results shown in Figure 11, we can see that as the number of topics increases, the large region topics are divided into small ones with similar importance. However, the results with smaller numbers of region topics are less stable than those with larger ones due to the short trajectories. It is of note that our test shows that the Dirichlet

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

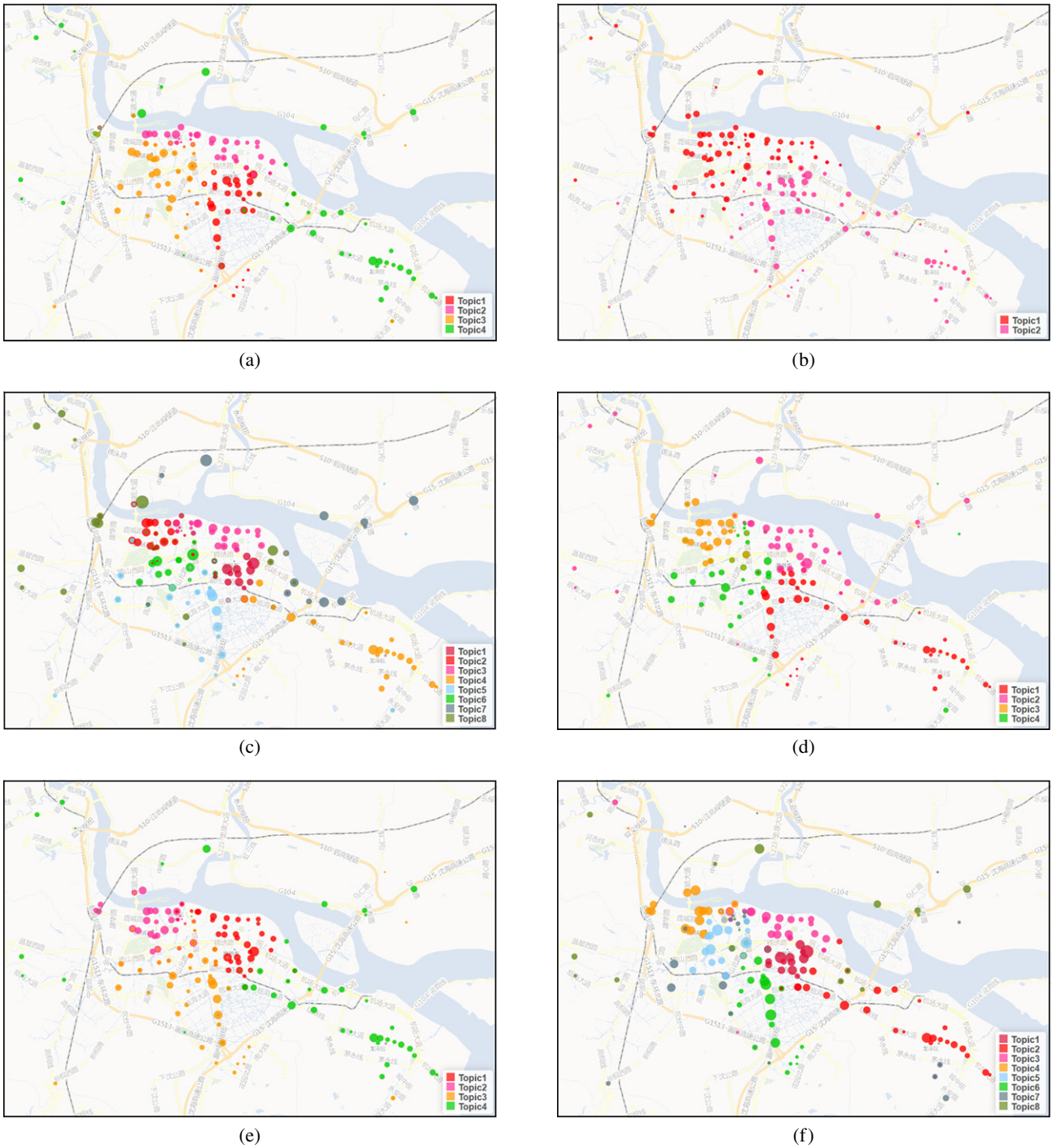


Figure 11. Visualization of TLDA results with different parameter settings: (a) $K = 4$, $\alpha = 0.1$, $\beta = 0.01$; (b) $K = 2$, $\alpha = 0.1$, $\beta = 0.01$; (c) $K = 8$, $\alpha = 0.1$, $\beta = 0.01$; (d) $K = 4$, $\alpha = 1$, $\beta = 0.01$; (e) $K = 4$, $\alpha = 0.1$, $\beta = 0.1$; (f) $K = 8$, $\alpha = 1$, $\beta = 0.1$.

priors α and β have a relatively small sensitivity on the results of our TLDA model.

Evaluation of Effectiveness

Nowadays, the traffic conditions in many large cities are becoming increasingly worse and complicated. Therefore, modeling and exploring the changing traffic conditions at the level of the entire city are extremely challenging. However, some traffic regions always exist inside which roads share similar characteristics of traffic conditions. Discovering such traffic regions can significantly simplify the complexities of modeling whole city traffic conditions.

To evaluate the effectiveness of TLDA, we use the traffic data in the city of Wenzhou along with its neighboring Yueqing County, Ruian County, Pingyang County and Cangnan County. As Figure 12 indicates, we observe that

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

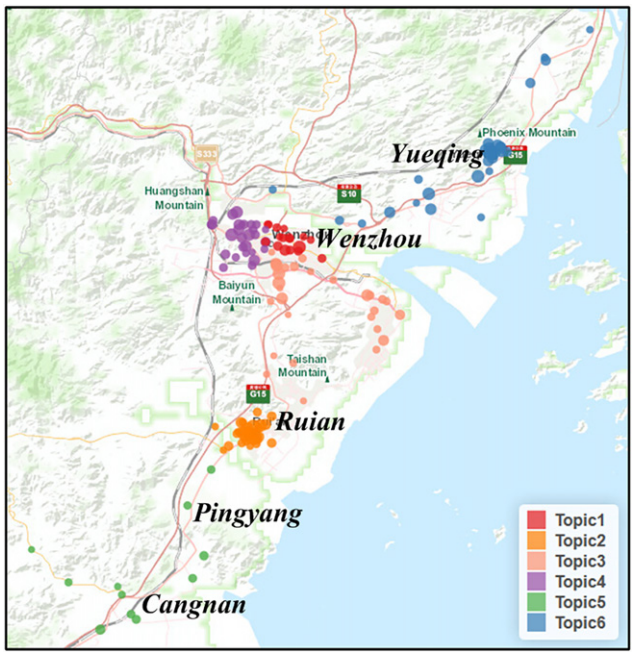


Figure 12. The region topics obtained from TLDA for Wenzhou City and its neighboring counties Yueqing, Ruian, Pingyang and Cangnan. The region topics are distinguished by different colors.

the city of Wenzhou can be divided into three region topics, roughly consistent with its inner three districts, whereas Yueqing County and Ruian County form two different region topics, and Pingyang County and Cangnan County are combined together into one region topic, with the number of topics being set to six. Pingyang County and Cangnan County are merged into one region topic because many vehicles frequently drive along the expressway between Pingyang County and Cangnan County, thus forming a strong connectivity between them. It is obviously concluded that vehicles move inside their respective counties (or districts) more frequently than they do between counties and districts, which is consistent with the real situation.

In addition, our research can be extremely useful in helping the city to improve traffic. We obtained many valuable inspirations from the practice of our approach in Wenzhou City. For example, we identified one zone and its two routes (i.e., DongOu Avenue and Kanghua Road) that are primarily used for entering and leaving Wenzhou City. This suggests that these two routes play a very important role in connecting the inside and outside of the city and need to be broadened when necessary. We also found that the intersection of Airport Avenue and Jinxiu Road belongs to both *Region Topic 1* and *Region Topic 2*, which suggests that this intersection must be paid much attention during traffic control because it connects two different functional regions. Additionally, we concluded that the number of vehicles passing from east to west is roughly the same as that from west to east, but more vehicles are passing from north to south than south to north between 8:00 am and 9:00 am. This shows that the direction from north to south is the main traffic flow at this time slot, and supports the decision of designing reversible lanes in some main roads that run in the north and south direction.

Comparison with Other Approaches

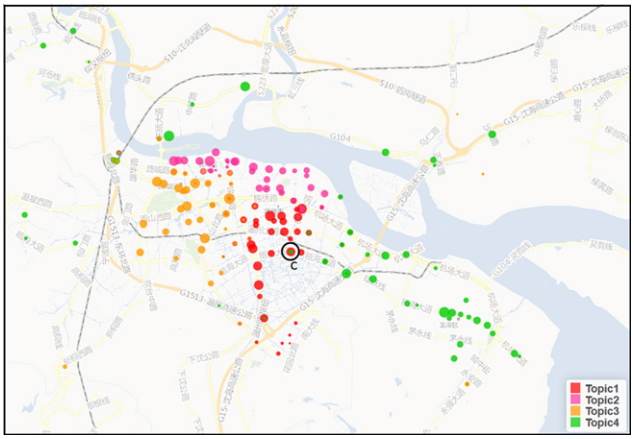
The TLDA model brings a novel perspective to exploring traffic data. Previous exploration methods are more about clustering and statistics, while some are texture based. Compared with these methods, TLDA has two major differences. (1) It not only clusters trajectories using a fuzzy assignment based on probabilities, but also produces the meaningful region topics by incorporating multi-features, such as vehicle speed and direction. These region topics reveal complex inherent traffic flow behaviors, which may be difficult to define for detection and extraction without prior knowledge. (2) It is much easier to combine the features of traffic flow since it treats every trajectory as a bag of feature values. Moreover, these features could be totally heterogeneous from very different fields, which enables users to explore the data in a more flexible way.

In clustering, traditional clustering algorithms such as *k-Means* must determine many influence factors, such as the differences of passing-vehicle numbers and road distances, in advance for the calculation of distances between surveillance devices. In fact, the final results depend heavily on these influence factors. Unfortunately, analysts generally have difficulty in choosing and determining the most suitable ones if they do not have strong domain knowledge.

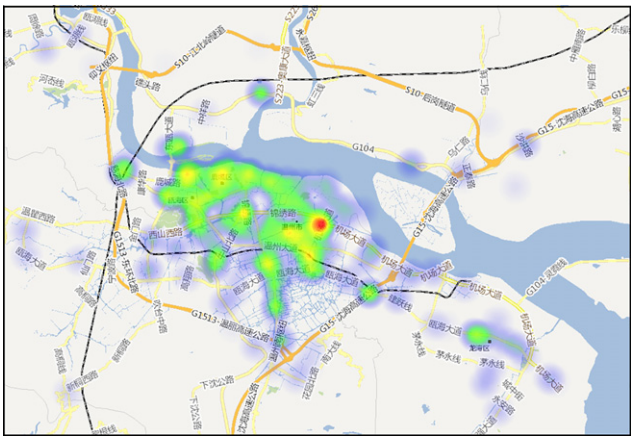
To show the advantages of our approach over traditional clustering algorithms, we compared the visual results of our approach with those of heatmap, which visualizes the clusters of traffic regions based on the traffic volumes passing by surveillance devices and the distances between surveillance devices. Figure 13 shows two diagrams that are based on the same trajectory data between 8:00 am and 9:00 am on June 5, 2014 in Wenzhou City. Clearly, our approach presents not only the frequently passed-by surveillance devices approximately along the south border of *Region Topic 2* and the north border of *Region Topic 3*, indicated by large pink and orange circles, respectively, but also some infrequently passed-by surveillance devices between *Region Topic 2* and *Region Topic 3*, indicated by small circles. In other words, the reason for generating two different regions (*Region Topic 2* and *Region Topic 3*) in our approach is because only a small number of vehicles move between them. However, heatmap fails to differentiate between *Region Topic 2* and *Region Topic 3*. Thus, our approach can discover the relationship among different surveillance devices and the traffic volume of vehicles passing by the same surveillance devices, which heatmap fails to do.

It is worth mentioning that our approach allows one item (surveillance device) to belong to more than one cluster (region topic or traffic region), whereas some traditional clustering approaches allow one item to belong to only one cluster. For example, the surveillance device marked by "c" in Fig. 13(a) is located in both *Region Topic 1* and *Region Topic 4*, meaning that its location connects two different regions, which is consistent with the real situation.

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519



(a)



(b)

Figure 13. Comparing our approach with heatmap in visualizing traffic regions based on traffic data between 8:00 am and 9:00 am on June 5, 2014 in Wenzhou City: (a) TLDA-based traffic regions; (b) heatmap-based traffic regions.

Finally, although some research works focus on the study of clustering of trajectories, determining how to cluster long trajectories still remains a difficult task. Some split long trajectories before clustering,[15] whereas others employ DBSCAN based on the density in the space.[16] Unfortunately, these approaches neither consider the whole complete trajectory nor employ the relationship among the trajectories. In addition, they usually have a higher computational complexity. By contrast, our approach transforms the trajectories into plain text and considers each complete trajectory and its relationship. Thus, it is able to obtain the traffic regions with a reduced computational complexity.

RELATED WORK

Exploring the traffic stream helps in acquiring knowledge of the traffic conditions. Currently, many related works exist, in both academia and industry. In this section, we briefly review current advances discussed in the literature regarding traffic data visualization, topic modeling and LDA, and the design of visualization.

Traffic Visualization

Due to the growing rate at which traffic data are being collected, traffic visualization analysis is becoming a very active research field, combining techniques and expertise from many other fields, including GIS, computational movement analysis, computational geometry, databases and data mining.[17] Three major types of traffic data are origin–destination (OD) data, GPS data and sparse traffic trajectory data. OD data are collected by public transportation systems such as subway and bus systems, bicycle systems, etc., whereas GPS data are primarily used for traffic monitoring concerning specific vehicles. In addition, sparse traffic trajectory data are usually reconstructed from images or videos by surveillance devices.

For OD data, Zeng et al. visualize and explore passenger mobility in a public transportation system with a family of analytical tasks based on inputs from transportation researchers.[18] Furthermore, Beecham et al. use visual analytics techniques to identify, describe and explain cycling behavior within a large and attribute-rich transactional dataset.[19] By applying visual analytics techniques to vehicle traffic data, Andrienko et al. find a way to visualize and study the relationships between traffic intensity and movement speed on links of a spatially abstracted transportation network.[20]

In previous studies, GPS data are frequently used. Ferreira et al. study taxi GPS data to understand trends in movement patterns on *k-means*.[21] Wang et al. present a data-driven solution by leveraging a visual analysis system to evaluate the real traffic situation based on taxi GPS data.[22]

In our article, we focus on sparse traffic trajectory data, which record the movement of vehicles at fixed locations. These trajectory data combine both location-based and movement-based data. To the best of our knowledge, only Wang et al. have studied this type of data in the visualization community.[4] They present a visual analysis system to allow users to check how traffic congestion at one site is correlated with traffic flows on neighboring links, and with route selection in its neighborhood.

Topic Model and LDA Analysis

The topic model is widely used in text analysis. In 1998, Landauer et al. first proposed the concept of Latent Semantic Analysis,[23] which has been a frequently used topic model since then. Latent Semantic Analysis adds a latent semantic layer between documents and words. The extracted latent semantics represent the contextual-usage meaning of words by statistical computations applied to a large corpus of text. With pLSI/pLSA,[24,25] the statistical analysis and generative model based on LSA is introduced. pLSA solves the problem of synonyms and polysemy, but it suffers from overfitting. Blei et al. propose the concept of the topic model and the related LDA model.[6] LDA is a multi-layer Bayesian model, including the layers of words, topics and documents. In LDA, every topic is a mixture of words, while every document is a mixture of topics. By introducing the Dirichlet distribution,

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

the LDA model can avoid overfitting, from which pLSA suffers.

In text analysis, many visualization approaches are proposed for the results derived by the LDA model. Choo et al. propose UTOPIAN to represent the keywords about topics, using the sizes of circles to indicate the probabilities. They use distinct colors to distinguish different topics.[26] To visualize the evolution of topics over time, Wei et al. proposed a visual exploratory text analytic system (TIARA), which encodes the hotness of topics using the width of rivers in ThemeRiver.[27] In addition to applications in the text analysis field, the LDA model has also been adopted in areas such as classification,[28] pattern recognition[29] and segmentation.[30] Hong et al. support the exploration of unsteady flow fields with their proposed LDA-based method.[7] For traffic streams, Chu et al. use the LDA model to discover hidden themes from trajectory data,[8] and Zheng et al. use mobility data as the document, and POIs (Points of Interest) as metadata to discover regions of different functions in a city[31]

In this article, we build a topic model for sparse surveillance-device-based trajectory data to discover hidden knowledge. To obtain more knowledge, we refer to Hong et al.'s work but supplement various features (e.g., speed feature, direction feature, vehicle type feature), from which we can not only obtain the hidden region topics, but also acquire messages of speed, direction and type of related regions.

The Design of the Visualization

To date, many approaches have presented visual tools for exploring geographical information.[32,33] These tools typically have map-based displays and employ information visualization techniques to visualize the spatial attributes of the data over temporal changes. Word clouds are often used to highlight the important words. Treemaps are also used to display multi-class data. For example, NewsMap uses a treemap to display topics by changing the direction of the rectangle and nesting rectangles to represent different levels, and through the size of the rectangle showing the importance of nodes.[34] Alternatively, many technologies exist to visualize time series data. For example, Saito et al. show the rate of data with the evolution of time, and use color to enhance the effect of the change.[35] Chu et al. use a timeline to display the evolution of topics over time.[8] Wang et al. use a pixel table to show different features' cyclical changes.[4,36]

Differently from previous works, we provide a map view for geographic information on topics, a cloud view for the relationship of surveillance devices, a treemap view for other detailed features, together with a matrix-table view to show related messages with temporal evolution.

CONCLUSION AND FUTURE WORK

As numerous surveillance devices have been installed along the roadsides in China, more and more vehicle trajectories are captured and gathered to form the Big Traffic Dataset. In this article, we introduce an LDA-based approach to explore traffic regions based on massive traffic data. A real case in Wenzhou City demonstrates the effectiveness of our proposed approach and prototype system.

In the future, we would like to continue our work in the following two aspects. First, the effects of different TLDA parameters on results must be studied thoroughly to provide explorative guidance. In addition, we will investigate how to support the discovery of region topics about location feature over a long period of time. If some traffic issues exist for some period of time, the city needs to take some action to solve the issues. It is of note that our current approach only considers four features (location, direction, speed and vehicle type) when applying the TLDA model. When exploring the temporal patterns for specific region topics, we have to use a fixed duration of time slot, such as one hour in our case. We hope to be able to obtain clear insight into drastic changes of traffic conditions with smaller and even changeable durations of time slots, such as 5, 10 or 15 min, if we incorporate the time feature when applying the TLDA model in the future.

ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation of China (No. 61100043), Zhejiang Provincial Natural Science Foundation (No. LY12F02003), the Key Science and Technology Project of Zhejiang (No. 2012C11026-3) and Science and Technology Innovation Activity Plan for University Students of Zhejiang Province (No. 2015R407062).

REFERENCES

[1] F. Wang, W. Chen, F. Wu, Y. Zhao, H. Hong, T. Gu, L. Wang, R. Liang, and H. Bao, "A visual reasoning approach for data-driven transport assessment on urban roads," *IEEE Conf. of Visual Analytics Science and Technology* (IEEE, Piscataway, NJ, 2014), pp. 103–112.

[2] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: a study of New York City taxi trips," IEEE Trans. Vis. Comput. Graphics 2149–2158 (2013).

[3] R. Krueger, D. Thom, and T. Ertl, "Visual analysis of movement behavior using web data for context enrichment," *IEEE Pacific Visualization Symposium* (IEEE, Piscataway, NJ, 2014), pp. 193–200.

[4] Z. Wang, T. Ye, M. Lu, X. Yuan, H. Qu, J. Yuan, and Q. Wu, "Visual exploration of sparse traffic trajectory data," IEEE Trans. Vis. Comput. Graphics 1813–1822 (2014).

[5] Y. Zheng, "Trajectory data mining: an overview," ACM Trans. Intell. Syst. Technol. (2015).

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res. 993–1022 (2003).

[7] F. Hong, C. Lai, H. Guo, E. Shen, X. Yuan, and S. Li, "FLDA: Latent Dirichlet allocation based unsteady flow analysis," IEEE Trans. Vis. Comput. Graphics 2545–2554 (2014).

[8] D. Chu, D. Sheets, Y. Zhao, Y. Wu, J. Yang, M. Zheng, and G. Chen, "Visualizing hidden themes of taxi movement with semantic transformation," *IEEE Pacific Visualization Symposium* (IEEE, Piscataway, NJ, 2014), pp. 137–146.

[9] https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation.

[10] G. Heinrich,"Parameter estimation for text analysis," Technical Report (2004).

[11] B. Johnson and B. Shneiderman, "Tree-maps: a space-filling approach to the visualization of hierarchical information structures," Los Alamitos, IEEE Computer Society Press (IEEE, Piscataway, NJ, 1991), pp. 284–291.

[12] M. Bruls and K. Huizing Wijk, "Squarified treemaps," *Proc. Joint Eurographics and IEEE TCVG Symposium on Visualization* (IEEE, Piscataway, NJ, 1999), pp. 33–42.

DOI: 10.2352/ISSN.2470-1173.2016.3.VSTIA-519

[13] X. H. Phan and C. T. Nguyen, "JGibbsLDA: A Java implementation of latent Dirichlet allocation (LDA)," (2008).
[14] J. Tang, Z. Meng, X. L. Nguyen, Q. Mei, and M. Zhang, "Understanding the limiting factors of topic modeling via posterior contraction analysis," *Proc. 31st Int'l Conf. on Machine Learning* (2014), pp. 190–198.
[15] J. G. Lee, J. Han, X. Li, and H. Gonzalez, "Traclass: trajectory classification using hierarchical region-based and trajectory-based clustering," *Proc. VLDB Endow* (2005), pp. 136–146.
[16] J. G. Lee, J. Han, and K. Y. Whang, "Trajectory clustering: a partition-and-group framework," *Proc. ACM SIGMOD* (2007), pp. 593–604.
[17] J. Han, Z. Li, and L. Tang, "Mining moving object, trajectory and traffic data," *Int'l Conf. on Database Systems for Advanced Applications* (2010), pp. 485–486.
[18] W. Zeng, C Fu, S. M. Arisona, A. Erath, and H. Qu, "Visualizing mobility of public transportation system," IEEE Trans. Vis. Comput. Graphics 1833–1842 (2014).
[19] R. Beecham, J. Wood, and A. Bowerman, "A visual analytics approach to understanding cycling behavior," Vis. Anal. Sci. Technol. 207–208 (2012).
[20] N. Andrienko, G. Andrienko, and S. Rinzivillo, "Exploiting spatial abstraction in predictive analytics of vehicle traffic," *IEEE Visualization Conf.* (IEEE, Piscataway, NJ, 2014).
[21] N. Ferreira, J. T. Klosowski, C. E. Scheidegger, and C. T. Silva, "Vector field k-means: clustering trajectories by fitting multiple vector fields," *The Eurographics Conf. on Visualization* (2013), pp. 201–210.
[22] F. Wang, W. Chen, F. Wu, Y. Zhao, H. Hong, T. Gu, L. Wang, R. Liang, and H. Bao, "A visual reasoning approach for data-driven transport assessment on urban roads," IEEE Vis. Anal. Sci. Technol. 103–112 (2014).
[23] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," Discourse Process. 259–284 (1998).
[24] T. Hofmann, "Probabilistic latent semantic indexing," *Proc. 22nd Annual Int'l ACM SIGIR Conf. Research and Development in Information Retrieval* (1999), pp. 50–57.
[25] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," Mach. Learn. 177–196 (2001).
[26] J. Choo, C. Lee, and C. K. Reddy, "Utopian: user-driven topic modeling based on interactive nonnegative matrix factorization," IEEE Trans. Vis. Comput. Graphics 1992–2001 (2013).
[27] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang, "Tiara: a visual exploratory text analytic system," *Proc. 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining* (2010), pp. 153–162.
[28] L. Cao and F. Li, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," *Proc. Int. Conf. on Computer Vision* (2007), pp. 1–8.
[29] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition* (2006), pp. 2169–2178.
[30] X. Wang and E. Grimson, "Spatial latent Dirichlet allocation," *Proc. Neural Information Processing Systems Conf.* (2007).
[31] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," *Proc. ACM SIGKDD Conf. on Knowledge Discovery & Data Mining* (2012), pp. 186-1.
[32] M. Nöllenburg, "Human-centered visualization environments," Geographic Vis. 257–294 (2007).
[33] J. Zhao, P. Forer, and A. S. Harvey, "Activities, ringmaps and geovisualization of large human movement fields," Inf. Vis. 198–209 (2008).
[34] NewsMap: http://newsmap.jp/.
[35] T. Saito, H. N. Miyamura, and M. Yamamoto, "Two-tone pseudo coloring: compact visualization for one-dimensional data," *Proc. IEEE Symposium on Information Visualization* (IEEE, Piscataway, NJ, 2005), pp. 173–180.
[36] Z. Wang, M. Lu, X. Yuan, J. Zhang, and Huub Van De Wetering, "Visual traffic jam analysis based on trajectory data," IEEE Trans. Vis. Comput. Graphics 2159–2168 (2013).