Visualizing Hidden Themes of Taxi Movement with Semantic Transformation

Ding Chu, David A. Sheets, Ye Zhao, Yingyu Wu *
Kent State University

Jing Yang †
UNC Charlotte

Maogong Zheng, George Chen ‡
Shenzhen Institute of Advanced Technologies

Abstract

A new methodology is developed to discover and analyze the hidden knowledge of massive taxi trajectory data within a city. This approach creatively transforms the geographic coordinates (i.e. latitude and longitude) to street names reflecting contextual semantic information. Consequently, the movement of each taxi is studied as a document consisting of the taxi traversed street names, which enables semantic analysis of massive taxi data sets as document corpora. Hidden themes, namely *taxi topics*, are identified through textual topic modeling techniques. The taxi topics reflect urban mobility patterns and trends, which are displayed and analyzed through a visual analytics system. The system integrates interactive visualization tools, including taxi topic maps, topic routes, street clouds and parallel coordinates, to visualize the probability-based topical information. Urban planners, administration, travelers, and drivers can conduct their various knowledge discovery tasks with direct semantic and visual assists. The effectiveness of this approach is illustrated by case studies using a large taxi trajectory data set acquired from 21,360 taxis in a city.

Keywords: Taxi Trajectories, Semantic Transformation, Clustering, Topic Modeling, Latent Dirichlet Analysis

1 Introduction

In the metropolitan areas of big cities, such as New York, London and Beijing, a large amount of taxis move around streets transporting people among urban cores, business centers, tourist attractions, transportation hubs, and residential territories. A modern vehicle GPS device on a taxi can record its realtime moving path (i.e. trajectory) as a series of positions sampled with a small periodic interval. At each location, GPS information and meta information are stored such as time, geographical coordinates (latitude, longitude), speed, direction, and occupied/vacant status regarding whether the taxi has a customer. It remains a far-reaching goal to extensively decipher the information hidden in the complex, dynamic behavior of large populations. Moreover, domain experts and practitioners require effective visualization so that they can intuitively manipulate the iterative, exploratory process and derive pertinent insights.

In this paper, we propose a new visual analytics system that finds and manifests implicit patterns derived from the taxi movement data throughout a city. We creatively study each trajectory as a *document* consisting of the taxi-traversed street names. This is made possible by mapping GPS coordinates (i.e. latitude and longitude) from geometric positions to a meaningful text representation. This method is named semantic transformation since it can replace a spatio-temporal GPS sample with additional contextual information. Such transformation can add semantic knowledge from different aspects, such as street names, Points of Interest (POI) (e.g., restaurants and shops), and other functional information. The transformation can be flexibly manipulated such as applying it only on the segments when the taxi is occupied (i.e. with customers). In this paper we implement the transformation from taxi positions to street names. A street itself implies geographic and cultural information (i.e. semantics) to the citizen of a city. In the future we will study the transformation of more semantic information.

This approach enables analysis of massive taxi datasets as document corpora. Consequently, we are able to introduce the Latent Dirichlet Allocation (LDA) [4] to infer hidden patterns of moving taxi populations. We name such patterns *taxi topics* since LDA is widely used in topic modeling of document collections. Here a taxi topic is a cluster of streets within a probabilistic framework (see details in Sec. 4). This new scheme has several characteristics including:

- **Discovering Group Movement Patterns**: The topics reflect group movement patterns of taxis more than just the spatial division of road networks. For instance, one topic may consist of major streets of a city district and a highway connecting an airport to this district. The highway may concurrently be a part of another topic together with streets of another city district.
- **Integrating Text Analysis Tools**: The transformation facilitates the use of text mining techniques (in this paper we use LDA) on the raw GPS location data. It creates a new method for effective knowledge discovery of trajectories.
- **Tolerating Error/Noise**: A limited number of erroneous GPS mapping or missing GPS records are tolerated in topic detection based on the probabilistic topic models. It significantly reduces the need for data preprocessing and correction in traditional trajectory processing.
- **Reducing Dimensions**: The dimensionality of the trajectories, originally a large and varied number of GPS sampling locations, is reduced to the number of topics. This number is controlled and adjusted by users. A street is also represented by such a vector of probabilities over the topics. Therefore, investigative operations like clustering and outlier detection can be applied to the trajectories or streets via the probability vectors.

We design a prototype VATT (Visual Analytics of Taxi Topics) system aimed at supporting visual analytics tasks. It integrates taxi topic maps with several visualizations including street clouds, street/trajectory parallel coordinates, and topic routes. The system helps users effectively explore the discovered hidden patterns. We illustrate our scheme with several case studies on a large data set acquired from more than twenty thousand taxis in Shenzhen, China.

2 Related Work

2.1 Trajectory Mining

Our method involves using semantics transformation and LDA for mining taxi mobility patterns. Trajectory data mining has been extensively studied [35]. A data-centric framework is presented in [18] to investigate the periodic behaviors along time intervals. Reference spots and Fourier transform are used to enable periodicity discovery [17]. Grouping moving objects according to spatial and temporal closeness defines flock/swarm/convoy patterns or path/sub-trajectory clusters [32].

Clustering and classifying long trajectories are mostly performed over distance measures using Euclidean or shortest-path over networks. For example, partitions of trajectories are used to cluster

*dchu,dsheets,zhao,ywu@cs.kent.edu; corresponding author: Y. Zhao; D. Chu and D. Sheets contributed equally to this work

†jing.yang@uncc.edu

‡mg.zheng,qian.chen@siat.ac.cn

them in rectangular regions, creating fine-grain groups [14]. TRACLUS [15] uses modified DBSCAN, a popular clustering method based on density reachability [8], to group trajectories by the density of their segments in the space. NEAT [22] studies trajectory clustering over road networks, where a similarly modified DBSCAN is used by further considering traffic flows together with segment densities and connectivity. The method finds clusters by looking for contiguous road segments with large traffic flows. Our taxi topics can also be considered clusters of streets but with distinct features described here. First, the topics are not strictly confined by the connections in the road network but rather created by the probabilities that sets of taxi trajectories follow similar streets. Second, our method can naturally incorporate a variety of parameters over the transformed names into the clustering process. We explore this with approaches such as filtering repeated (or unwanted) street names or augmenting the trajectory documents based on taxi speed (i.e. speed compensation). The existing techniques mostly account for spatial-temporal values where such operations cannot be easily applied or need geometric computations. Finally, using LDA, our topics overlap with a probability distribution. A road can belong to multiple topics to better reflect taxi movement patterns within and across the topics. For instance, an airport highway is an important component of several topics (with different probabilities), each linking to one city district. This naturally reflects the taxi groups traveling within roads of one district and the airport. Using existing techniques, the cluster containing the highway connecting the airport might include zero, one, or several districts but does not directly exhibit the fact that the highway is shared by them.

Trajectories are used to extract geographical borders [24]. This method assumes non-overlapping regions which is expected for administrative/spatial borders. In contrast, our technique finds topics without the constraint of specific spatial areas, where the overlap is quantified by a probability defining the strength of multiple topics in a common area.

Semantic trajectories [22] has been studied where trajectories are enriched with semantic meanings. Preprocessing and manual (or semi-automatic) annotation are needed to add semantics to the whole trajectories or their segments. These methods do not directly annotate GPS positions since it "is not efficient as it may generate a large number of repetitive annotations" [22]. In contrast, our direct transformation from GPS samples to street names and the novel use of document mining techniques can overcome this problem and promote automatic processing.

Based on similar taxi dataset, Yuan et al. developed T-Drive [33], which makes recommendations of fastest paths for taxi drivers through a routing algorithm based on historical data. Gao et al. [10] presented a system that can help identify factors affecting taxi drivers income. Our approach based on semantic transformation and dimension reduction can be integrated with such existing methods for better suggestion and planning.

2.2 Trajectory and Taxi Visualization

Many approaches have presented visual tools for exploring geographical information [21, 25, 31, 34]. For visualizing spatial-temporal data, Andrienko et al. presented a review of the key issues and approaches [2, 3]. The techniques involve map-based displays and use information visualization techniques to visualize the spatial attributes of the data over temporal changes. Object trajectories are a key type of movement data where many techniques have been developed, such as GeoTime [13], TripVista [11], FromDaDy [12], vessel movement [30], etc. Wang et al. [28] studied point pattern analysis methods and built an interactive visualization system of hot spots based on a mashup technique. Crnovrsanin et al. [5] visualizes movement trace data based on the proximity to points of interest. The method limits its study on important locations and the specific movement patterns (e.g. concurrence, convergency, fluctuation) are computed on an abstraction space. These methods mostly visualize trajectories using density-based map and spatio-temporal aggregation to provide data overview. Spatial and temporal dimensions are discretized into regions/windows for levels of detail and clutter reduction techniques are employed.

Tominski et al. tackled the problem of visualizing trajectory attributes together with the individual points comprising the trajectories [26]. Landesberger et al. interactively visualized trajectory data with categorical changes over time and spatial data displays [27].

The visual study of taxi data has been conducted in a variety of applications. Liu et al. presented a visual analytics system of route diversity [19]. They developed several visual encoding schemes to display the statistic information for different routes and their importance with a ring map. In particular, entropy is computed for a source/destination pair of locations, and visualized to reveal the diversity of routes. Pu et al. [23] developed a system for users to monitor and analyze complex traffic situations in big cities for regions, roads and from vehicle views. Historical statistical information (named fingerprints) are encoded and placed on the map as ring-map-based radial layouts. Wang et al. presented an interactive system for visual analysis of urban traffic congestion [29] using traffic jam propagation graphs that join spatially and temporally related jam events. These graphs are explored by multiple views for the whole city, together with details of road segments. Ferreira et al. allowed users to visually query taxi trips [9]. The model supports origin-destination queries, different aggregations and visual representations for exploration and comparison of query results. An adaptive level-of-detail rendering strategy can generate clutter-free visualization for large results. Hidden details are shown as summarized overlay heat maps.

The VATT system aims at supporting visual analysis based on the discovered taxi topics with several major differences from previous methods. First, we need to visualize the discovered patterns of the group behavior, which is implemented as multiple colored layers of clusters over maps. Street line transparencies are used to show the associated probabilities for overlapped topics over streets. Second, our method introduces semantics to spatial-temporal data so we use tag cloud to help users directly find the information. Third, the LDA-computed probability and entropy are used in visualizing diversity of streets and taxi trajectories. Parallel coordinates and clouds are used on the reduced topical dimensions. The VATT visualizations need to be further refined. We will provide more visualization functions inspired by the related work, such as better clutter reduction and levels of detail, temporal semantic information with ring maps, and interactive query support.

2.3 LDA and GPS Data

LDA [4] has become a widely-used machine learning technology in organizing and finding patterns in different textual corpora. The complexity of detecting and processing GPS data into a usable form has been widely studied [20]. To the best of our knowledge, our method is the first to use LDA in GPS sampling data processing.

3 Taxi Data Transformation

Taxi Trajectory Data The trajectory data used in our case studies contains daily trajectories of 21,360 taxis in Shenzhen, a big city in southern China bordering with Hong Kong. Shenzhen has over fifteen million residents in a condensed area and taxis are a major means of passenger transportation. Each taxi reports nearly three thousand GPS sample positions per day. Each sample consists of taxi plate, time, status, speed, direction, and latitude and longitude, e.g., (BXXXXX, 06/27/2012 23:59:33, 0, 50, 180, 22.533, 114.044). The accumulated taxi trajectories create a very large data set for investigation with a total of 59,087,230 samples recorded in one day.

Latitude	Longitude	Street Name	Street ID
22.543	113.991	Qiaocheng East Rd	38819
22.546	113.997	Qiaoxiang Rd	20694
22.568	114.066	Beihuan Ave	10192
22.568	114.067	Beihuan Ave	10206

Table 1: *Examples of mapping GPS to streets.*

Mapping Geographical Coordinates to Streets The semantic transformation maps a GPS position to the street name it lies on. Through this process, the position coordinate becomes a "keyword". A taxi trajectory becomes a "document" of the street names traversed in a given time window. The time window can be defined by users to study interesting patterns in different periods. Consequently, the documents from all taxis form a document corpora. In implementation, we use a road matching algorithm [16] to find the closest streets to each GPS position.

Often, a very long street covers a large geographic distance using the same street name. Such streets can be divided into small segments, each represented by a given street ID used as keywords in the topic analysis. The number of segments can be controlled to satisfy different analytical requirements. More segments increase the geographical granularity, while long streets better reveal information of the city's large-scale motion pattern. For example, a 30 mile freeway can be studied as a whole to reveal taxi traveling patterns related to this highway. It can also be studied and visualized as several segments so that in the topics the segments play independent roles regionally. A few examples of the GPS coordinates and their mapped streets can be found in Table 1.

4 Studying Taxi Data with Topic Modeling

4.1 Topic Discovery with LDA

LDA (Latent Dirichlet Allocation) is an effective (semi-)automated probabilistic topic modeling tool that can discover hidden thematic structure in a large corpus of documents [4]. LDA statistically groups keywords into potential topics by studying their occurrences among a large collention of documents. Consequently, each keyword is assigned a probability of its appearance in each topic. Visualizing the topics has been investigated to present the themes inside news, tweets, and other textual documents [6, 1]. To the best of our knowledge, we are the first to apply this powerful tool to geographic trajectory data, enabled by the semantic transformation. In our study we implement the LDA computation using the Stanford Topic Modeling Toolbox. Table 2 and Table 3 illustrate the LDA results of eight taxi topics, which are generated from 21,360 taxi trajectories between 9am to 12pm, Jun. 27, 2012.

Topic Importance: Total frequency of all taxis' occurrences on a topic represents the topic importance. In Table 2, the topics are ranked by importance (i.e., total frequency) from Topic1 to Topic8. For example, Topic1 has a total frequency of 758,188 and Topic8 only has about a quarter of that at 177,316. This means more GPS samples are associated with Topic1 than Topic8.

Street Probability Distribution: For each different topic, a street also has a frequency of occurrence, which we refer to as street importance over the topic. The street importance is used to compute the appearance probability of a street (i.e., keyword) w in a given taxi topic z, which is depicted as $p(w|z)$. Here a street's probabilities over all topics sum up to 1 ($\sum^{z} p(w|z) = 1$). Table 2 displays four example streets with their frequencies over each individual topic, which are used to compute their probability distribution on the topics. For example, Binhe Ave. has a total frequency of 109,076. Topic1 covers that street 39% (i.e. 42,476) of the time it occurs. Binhe Ave. is shared by Topic1, 2 and 4, and in contrast, Fumin Rd. is dominated by Topic1 at 99%. Those streets with a high probability $p(w|z)$ in one topic z are considered representing that topic. A threshold c is assigned to choose streets with $p(w|z) > c$ for topic visualization.

	Topic's	Binhe Ave		Xinzhou Rd		Fuqiang Rd		Fumin Rd		...
	Total Freq	Freq	Prob	Freq	Prob	Freq	Prob	Freq	Prob	...
Topic1	758188	42476	0.39	37432	0.97	36072	0.98	22867	0.99	...
Topic2	717219	54580	0.50	52	0.00	274	0.01	0	0.00	...
Topic3	405737	0	0.00	0	0.00	0	0.00	0	0.00	...
Topic4	404502	11177	0.10	618	0.02	336	0.01	0	0.00	...
Topic5	298811	0	0.00	0	0.00	3	0.00	0	0.00	...
Topic6	273008	831	0.01	366	0.01	153	0.00	128	0.01	...
Topic7	218869	12	0.00	13	0.00	0	0.00	0	0.00	...
Topic8	177316	0	0.00	49	0.00	0	0.00	24	0.00	...

Table 2: *LDA results: Total frequency of all taxis' appearance on a topic represents this topic's importance; A street has different frequencies over different topics, which define the probabilities, $p(w|z)$, of this street's appearance in multiple topics.*

Trajectory	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8
BXXXX1	0.00	0.00	0.00	0.00	0.23	0.69	0.08	0.00
BXXXX2	0.00	0.00	0.00	0.28	0.10	0.48	0.00	0.14
BXXXX3	0.00	0.00	0.15	0.27	0.58	0.00	0.00	0.00
BXXXX4	0.31	0.31	0.00	0.38	0.00	0.00	0.00	0.00

Table 3: *LDA results: probability contribution, $p(z|d)$, of taxi topics to taxi trajectories.*

Trajectory Probability Distribution: Each document has a probability distribution over multiple topics. This contributive probability of a taxi topic (z) to a given taxi trajectory d is represented as $p(z|d)$ where $\sum^{z} p(z|d) = 1$. Table 3 shows the probability distribution of eight topics to four taxi trajectories (real plate numbers are not shown for privacy protection). For example, BXXXX1 travels 69% inside Topic6, 23% in Topic5, and 8% in Topic7.

4.2 Entropy Information

Entropy can quantify the uncertainty of a street or trajectory over the topics to measure the variation or diversity of the variable:

Street Entropy: $H(w) = \sum_{z} p(w|z) \log p(w|z)$. A street with high entropy is shared by multiple topics. Very likely, it is a backbone street connecting different parts of the city. For example, in Table 2, Binhe Ave. has a higher entropy than Fumin Rd. From Shenzhen map, Binhe Ave. is a major street passing city hubs, including the Convention Center and the Huanggang port to Hong Kong.

Trajectory Entropy: $H(d) = \sum_{z} p(z|d) \log p(z|d)$. A specific taxi's trajectory has high entropy if it receives significant contributions from multiple topics. For example, this taxi may travel through a long path in the city across various districts (Fig. 8).

In summary the achieved probabilities and entropies are used to visualize the knowledge hidden in massive trajectories.

4.3 Topic Modeling with Trajectory Data Adjustments

The massive trajectory data include more information than the sequence of geographic coordinates, such as the vehicle's speed and occupancy status. Such information can be integrated into the semantic transformation process to expose more knowledge in the downstream topic modeling results. In Section 3, we have discussed that a long street can be divided into several segments to increase geographic granularity. Next, we show other approaches to preprocessing the data before the transformation and topic modeling.

Taxi Occupancy Applying the semantic transformation to original trajectories discovers topics without considering taxi occupancy. In addition, we can apply the transformation only on those trajectory sections a taxi is occupied by passengers. The resulting topics thus reflect moving behaviors of taxi riders. For example, they imply the mobility between a city's business and residential divisions. Such knowledge can potentially contribute to new public transit routes or traffic rules over specific streets. Moreover, the topics of unoccupied taxi trajectories manifest moving patterns of drivers seeking passengers, e.g., patterns related to busy transportation hubs. The administration can arrange bus/subway stations accordingly.

Speed Compensation The GPS positions that are acquired with regular time intervals create bias toward slower moving roads with more samples. This results in high importance computed for such roads in the topics. This is useful in determining traffic jams and bottlenecks. On the other hand, users may want to reveal and analyze the mobility patterns without such bias. The bias can be removed by injecting pseudo-sampling points based on the taxi speed. In implementation, we set a speed threshold S_t and a speed interval ΔS. If a taxi's speed S on a street exceeds S_t, we add a series of pseudo points, the number of which is $\frac{(S-S_t)}{\Delta S}$.

4.4 Taxi Topics Evolution

Taxi topics are created for a given time window. At different periods of a day they reflect different patterns. For example, people travel from suburbs to city centers in the morning and reversely in the afternoon. It is important and interesting to find the evolutionary trends of the topics along a time line. Thus, we define a *topic similarity* between two topics to find similar topics between consecutive time windows. Given two topics i and j, the similarity is computed as:

$$S_{i,j} = \frac{Size(T_i \cap T_j)}{Min(Size(T_i), Size(T_j))}, \tag{1}$$

where T_i and T_j are the sets of streets with a high probabilities $p(w|z) > c$ in each topic. Based on the similarities, closely related topics are identified in consecutive time windows, which reflect temporal evolution of topics. Such relationships are visualized to identify the variation of topical contents, the emergence of new topics, and their fading out.

5 VATT System

Our VATT system integrates multiple visualizations supporting visual analytics of the hidden knowledge. Figure 1 shows the system interface consisting of several components: (a) taxi topic maps; (b) street clouds; (c) parallel coordinate plot (PCP) of streets or trajectories over topics; (d) topic routes of temporal evolution. These visualizations are orchestrated in a coordinated manner for effective user exploration.

5.1 Taxi Topic Maps

Taxi topics, consisting of multiple streets, are visualized over geographic maps. Various backgrounds, such as topographical, satellite, and transportation maps, can be used to provide visual cues and context of geographic and cultural information. Features such as tourist attractions, subway stations, and government buildings can be placed on the map. The system supports direct use of Google Maps, OpenStreetMaps (openstreetmap.org), and user customized maps. Overlayed with the background, multiple layers are used to present different topics. To visualize one topic, streets belonging to it (defined as $p(w|z) > c$) are drawn with a given topic color. The line width of a street is determined by the street importance in a particular topic. Here a street may be drawn by multiple overlapped lines with different topic colors. To improve visual effects, the transparency of a line can be set (1) by the street importance, so that a street is shown obviously with the color of the topic that has the largest $p(w|z)$; or (2) by the entropy H(w) of the street to the topics such that multiple overlapped lines of this street representing each topic are better depicted, especially with a zoomed in view. Figure 1(a) displays eight LDA-generated topics with distinct colors on a terrain map of Shenzhen. Users can manipulate the appearance of the topics by setting layers visible/invisible and changing the color and transparency.

5.2 Street Cloud

We use a street cloud to display semantic information and the relationships between the streets. The size of a street name reflects the street importance, i.e., the frequency of its occurrence in one topic. Moreover, the probability distribution over multiple topics, $p(w|z)$, of a street is drawn as a graph line and overlaid on the lower part of the street name. This graph line allows users to identify whether the street is solely included in one topic or diversely shared by others. More importantly, the street names are not randomly positioned. Instead, the layout reflects their relationships by introducing attracting and repelling forces computed from pairwise cosine similarities:

$$sim(w_i, w_j) = \frac{\sum_{z=1}^{Z} p(w_i|z)p(w_j|z)}{\sqrt{\sum_{z=1}^{Z} p(w_i|z)^2 \cdot \sum_{z=1}^{Z} p(w_j|z)^2}}, \tag{2}$$

where $w_{i,j}$ are two streets and Z is the number of topics. Streets with similar topic distributions are positioned near each other, and dissimilar streets are positioned further apart.

In the street cloud of a given topic, some streets are aggregated together if they are visited by taxis traveling much inside this topic. Meanwhile, different groups are placed far away from each other. Street cloud can also show the most important streets among all topics, or those selected by users. Figure 1(b) shows the top thirty streets with the largest frequencies in all topics. The street names are thus grouped according to their distribution over different topics.

5.3 Parallel Coordinate Plots of Streets and Trajectories

The PCP visualization helps users to interactively discover knowledge from probability distributions over topics, which was used in exploring document topics [7]. Figure 1(c) visualizes streets over a PCP, where each dimensional coordinate is one topic and each polyline represents one street. To draw a street over multiple dimensions, its probabilities over the topics $p(w|z)$ are used. A logarithmic scale is applied since many probabilities are very small and only a few of them are large. To overcome clutter, streets are not drawn if they have smaller probability values than a threshold P_t over K topics. P_t and K are user controllable parameters. Clicking a topic's axis highlights all the streets which have high probabilities in that topic. In a similar manner, the PCP visualization can also be applied to taxi trajectories with each polyline representing a trajectory (here the probabilities $p(z|d)$ are used). In general, PCP helps users identify distributive patterns of streets or taxis, as well as study outliers and salient features.

5.4 Topic Routes of Temporal Evolution

Figure 1(d) demonstrates the routes of temporal evolution among multiple topics. Each column represents a time window and adjacent columns are consecutive time windows. Each circle stands for one topic at a time slot, whose radius is defined by the topic's importance. For each column, we always show the topics with higher importance on top of lower ones. Two circles in adjacent columns are connected by a link if their topic similarity exceeds a given threshold. The width of the link is related to the similarity, whose value is also shown. Similar topics are given the same color. Figure 1(d) reflects the evolution of topics in a whole day. This visualization also plays a role as a control panel, where users can select time windows and topics so that they are shown in other coordinated views for further investigation.

6 Case Studies: Visual Analytics of Taxi Data

We use the VATT system to explore the massive taxi trajectories of one day (June 27, 2012) in Shenzhen city. The number of topics and time windows are easily controlled by users. We have tried different numbers of topics. The selection should depend on specific knowledge of the city. Meanwhile, the number of time windows can be adjusted by users for interactive analysis. In the following examples, taxi topics are generated at each of the eight time windows of the whole day, i.e., 0am-3am, $\cdots$, 9pm-12am, etc.

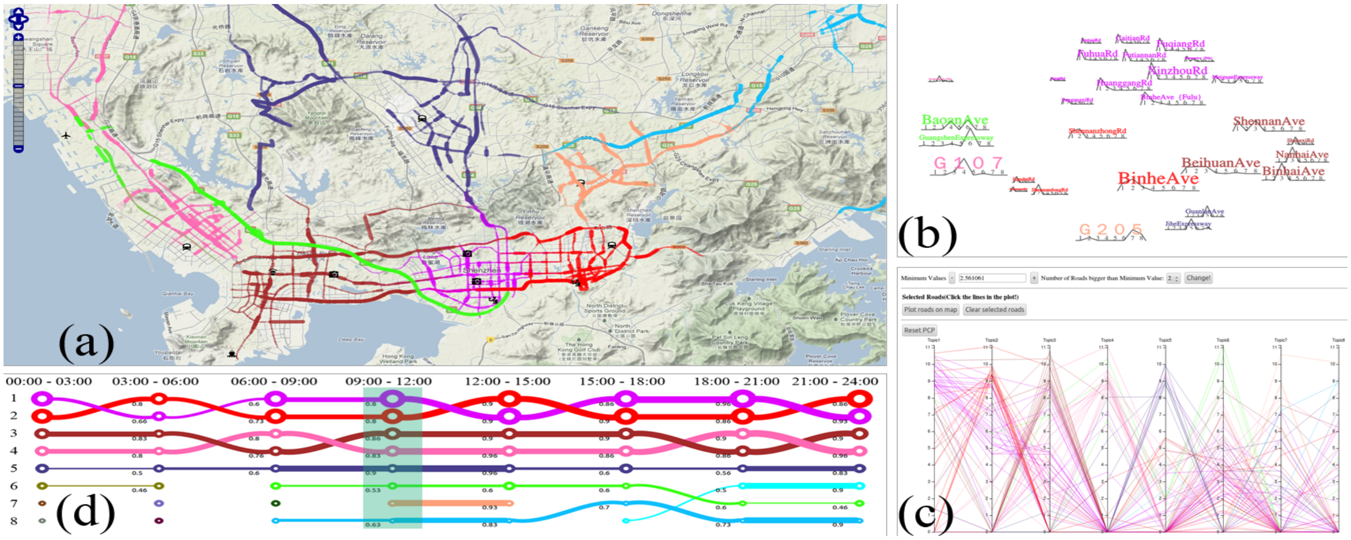


Figure 1: *VATT System Interface. (a) Taxi topic maps; (b) Street cloud; (c) Parallel Coordinate Plots of Streets and Trajectories; (d) Topic Routes.*

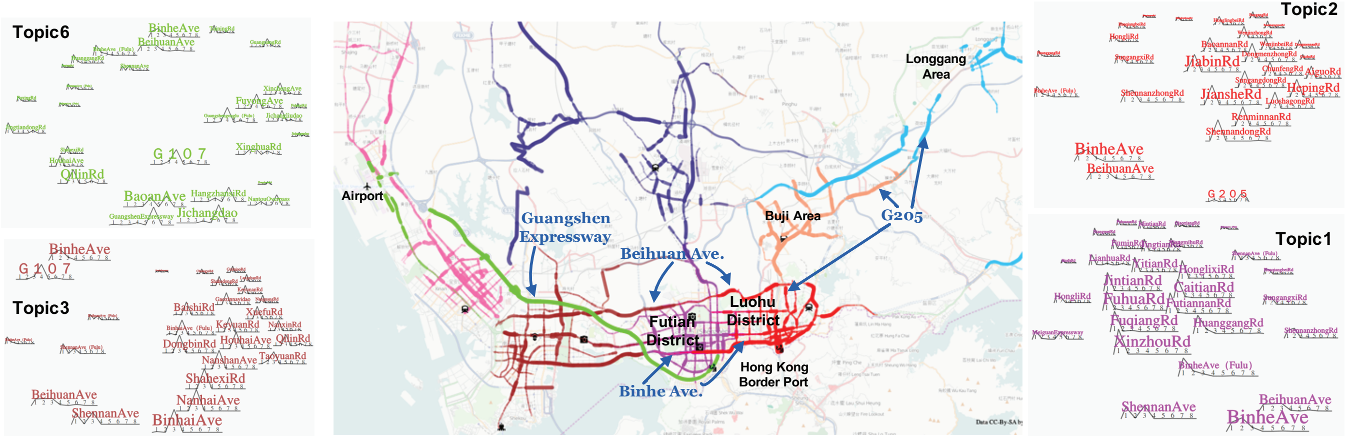
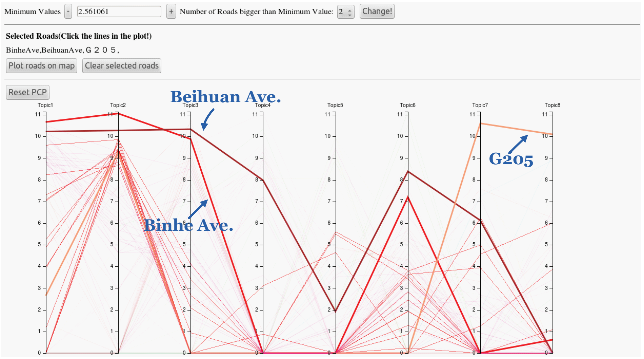


Figure 2: *Eight taxi topics shown on a Shenzhen map. Street clouds of four topics are shown in the corresponding colors.*


Minimum Values - 2.561061 + Number of Roads bigger than Minimum Value: 2 Change!

Selected Roads(Click the lines in the plot!)

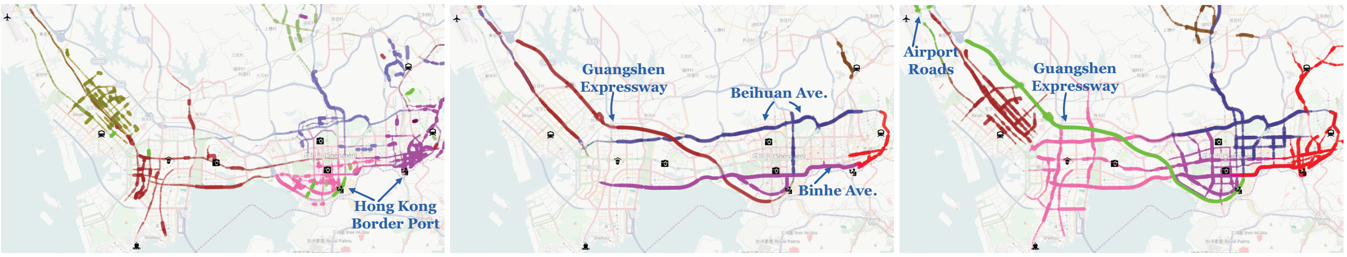
BinheAve,BeihuanAve,G 2 0 5 .

Figure 3: *PCP-based analysis of Topic2 streets of Fig. 2.*

6.1 Taxi Topics and City Mobility Patterns

The topics reveal typical traveling patterns of city cabs. Fig. 2 displays the eight taxi topics created for the time window 9am to 12am. Here the speed compensation is applied and the topics are created from full taxi trajectories including occupied and vacant sections. The streets of each topic with a high probability are drawn on the map in a distinct topic color. Street clouds of four topics are also shown in the figure. They display the top 30 streets of each topic in the corresponding topic color. The topics are ranked and labeled by their importance from Topic1 (high) to Topic8 (low). An immediate finding is that the topics are approximately distributed following the city's administrative divisions. The visualization implies that the districts of Shenzhen are functionally set up: a large portion of taxis can accomplish their movement inside a district. Next we introduce more findings with examples from Fig. 2.

Topic1 in purple roughly overlaps over Futian District, which contains the business and administration center of Shenzhen. Topic1 has the largest importance which indicates this district is the most concentrated area of taxis from 9am to 12am. Observing the street cloud of Topic1, the important streets can be easily recognized according to the font size, such as the Xinzhou Rd., Fuqiang Rd. and Binhe Ave. The first two are in a large group of streets in the middle of the cloud. This group includes the major roads inside this district. Meanwhile, Binhe Ave. is pushed away from this group by the similarity forces. We find it is a city expressway shared by other topics, as demonstrated by the curve below the name. Another interesting insight is about Binhe Ave. (Fulu), which is located in the middle between Binhe Ave. and the large group above. Users can zoom in on the map to study Binhe Ave. (Fulu) which is a side road out of Binhe Ave. From its location in the street cloud, in conjunction with the map, we can derive that it is used by taxis connecting the local roads inside this district as well as by taxis entering Binhe Ave. aimed to other districts. Such revelation on road connectivity might indicate potential traffic improvements to the administration.



(a) Vacant topics: 3am-6am (b) Occupied topics: 3am-6am (c) Occupied topics: 6am-9am

Figure 4: *Occupied and vacant taxi topics.*

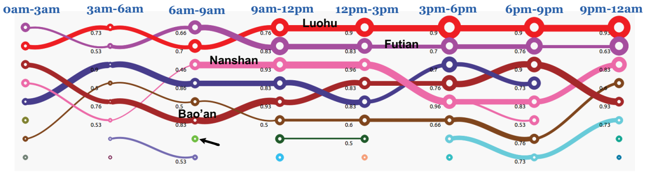


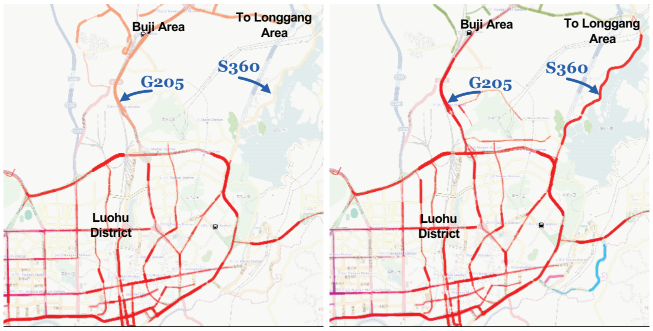
Figure 5: *Topic routes of occupied taxis during the whole day .*

Topic2 in red is the second most important topic which covers Luohu district, another critical business area. On the street cloud, Binhe Ave., Beihuan Ave. and G205 are the names far away from the main group of Luohu's internal streets. Binhe Ave. connects Futian area of Topic1, and Beihuan Ave. is an alternative to Binhe Ave. on its north. They are two major expressways of Shenzhen in the east-west direction. Moreover, G205 is a unique outlier that starts at Luohu and extends towards the two northeastern regional centers, Buji (Topic8 in orange) and Longgang (Topic7 in blue). In the street clouds of those two topics, G205 is also a significant contributor. Such information can also be identified in the PCP plot of streets. In Fig. 3, streets of Topic2 are shown and three outlier polylines are highlighted, Binhe, Beihuan and G205, respectively.

Topic6 in bright green is special in that it connects the city centers with the Bao'an International airport of Shenzhen on the northwest. As shown on its street cloud, the related streets constitute a ring shape. This means that these streets are not grouped into salient groups. In other words, taxis transport passengers around the city to different divisions. The long green path clearly shows the highway (Guangshen Expressway) connecting the airport to Huanggang port from the Hong Kong border. It indicates that taxis are busy servicing people from the border port to the airport, perhaps because domestic flights flying from Shenzhen airport within China are less expensive than those flying from Hong Kong airport. Many travellers from Hong Kong use Shenzhen airport to fly around China; and many tourists use this airport to visit Hong Kong.

6.2 Occupied and Vacant Taxi Topics

We further create eight taxi topics using the data of only passenger-occupied taxi trajectories, and another eight using vacant trajectories. Fig. 4 shows a few examples of these results. For the time slot from 3am to 6am, Fig. 4a and Fig. 4b plot the streets over vacant and occupied topics, respectively. Fig. 4a reveals the activities of those taxis seeking passengers at the very early morning. They travel around the city's main districts while focusing on important streets/sites. For example, we can observe that the taxis are very active around two major ports of the Hong Kong border indicated by arrows. Users can zoom in to study the active streets. Such knowledge may be used by police to design optimal patrol routes. Fig. 4b shows the results of passenger occupied taxis. Obviously, most of them transport passengers to destinations through several major expressways of the city, including the aforementioned Beihuan, Binhe, and Guangshen roads. They are easily accessed and have no traffic jams at this time. In comparison, we draw the occupied topics from 6am to 9am in Fig. 4c. Now the awakened city becomes active at all regions forming obvious topical divisions. In particular, an airport topic related to taxis serving flight passengers is shown in light green including airport roads and the highway. Such a topic with airport roads does not appear in Fig. 4b since flights seldom land between 3am and 6am.



(a) 9am-12pm (b) 3pm-6pm

Figure 6: *Temporal change over one topic.*

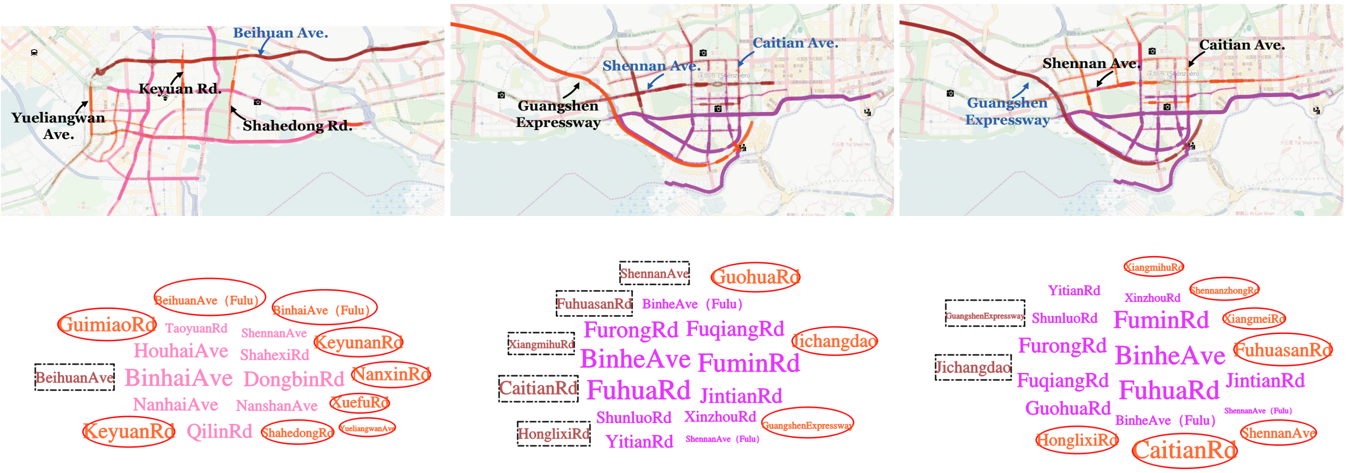
6.3 Taxi Topics over Time Periods

The variation of city traffic patterns at different time slots plays a significant role in finding important information and making smart decisions. For example, city administration can adjust the rule of streets (e.g. one-way) according to the time of day, to help avoid or reduce traffic jams. Tourists can arrange their schedule to plan a smooth travel. Taxi drivers can also plan better routes for their daily activities. This variation can be investigated by studying the temporal changes of taxi topics at different time windows.

6.3.1 Topic routes analysis

Fig. 5 shows the temporal evolution of topics, which are created for each three hour time window from the occupied taxi trajectories. We set the number of topics to eight. Fig. 5 is the view of the topic route view. Fig. 4b and Fig. 4c are the maps of taxi topics for the time slots 3am-6am and 6am-9am, respectively. There are many useful insights that can be observed based on the topic routes, such as:

- The taxi activities are very low during midnight and early morning, as small circles are shown in the first three time windows for all topics. The lowest activity time is 3am-6am. The size of circles increases from daytime to late night showing increasing activity of taxis.
- The circles and routes in red and purple become even bigger during the late night times from 9pm to 12am compared to day time. They mainly represent Luohu and Futian, two major business and entertainment centers. These regions are always the two major



(a) Topic3: 9am-12pm (b) Topic2: 3pm-6pm (c) Topic2: 6pm-9pm

Figure 7: *Visualizing disappearing (brown) and emerging (orange) streets for a selected topic on maps and street clouds.*

taxi topics throughout the day. From this we know, in this southern China city, people show heavy activities for night life at these regions, maybe because of the summer time (the day is Jun. 27) and the high temperature (average at 27.8 C or 82 F for June). In comparison, the brown route oscillates from morning to night. At late night 9pm to 12am, the brown circle is smaller. This topic is mostly about Bao'an district, a large residential region that does not have as much entertainment activity as Futian and Luohu.

- Individual topics not linked to others reflect important patterns related to specific times. At the time window 6am to 9am a special green circle appears as indicated by the arrow in Fig. 5. Checking the map view in Fig. 4c, the bright green topic of occupied taxis is the airport area roads and Guangsheng Expressway linking to Hong Kong border (similar to Topic6 in Fig. 2 created from original data). In other time slots, these roads are accommodated in other topics such as Luohu and Futian. Forming an independent topic implies that during 6am-9am taxis traveling on these roads are significant. This reveals that this time period is the busiest time of flights traveling to and from Shenzhen Airport.

6.3.2 Topical details analysis

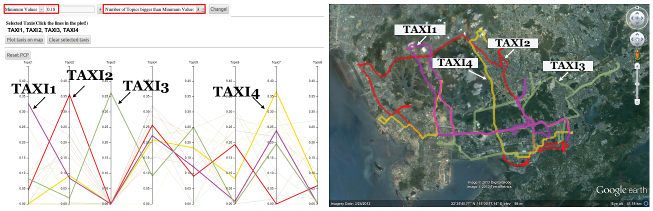
Next we discover time-related knowledge of individual topics with VATT to illustrate its support for detailed investigation. Fig. 6 shows two maps of a topic mostly related to Luohu district. Fig. 6a renders the important streets of this topic in red, which are acquired at the time window from 9am to 12pm, during the normal working time of a day. In contrast, Fig. 6b shows the streets of the topic acquired from 3pm to 6pm, including the afternoon rush hours. While both of them cover the major streets inside downtown Luohu area, clear differences can be observed at G205 and S360, the two major expressways connecting two northeastern residential centers, Buji and Longgang, respectively. They appear at the afternoon rush hours, when taxis are busy on moving workers from the downtown business region to suburban residences. Meanwhile, these roads are not active for taxis moving from Luohu during the morning working hours.

VATT also supports direct visualization of temporal change over the map and street cloud. For a user-selected topic in a time slot, its important (with high probability $p(w|z)$) streets are shown on the map. Meanwhile, the disappearing streets (i.e., those streets that are important at the previous time window for the similar topic, but are no longer important) are also drawn in a distinct color, as shown in the street cloud within rectangular boxes. Furthermore, the emerging streets (i.e., the streets that are not important before but become important now) are also specially colored on the map, and displayed in the cloud within ellipses. Fig. 7 shows the examples of such evolutionary information. Fig. 7a displays Topic3's map and street cloud of the occupied taxi data for 9am to 12pm. One very important disappearing street (shown in brown) is Beihuan Ave., a key road connecting different districts. This figure implies that this street is no longer significant inside this district (Nanshan) at the normal working hours. In the previous time slot, however it is important because in the morning (6am-9am) people travel to their working sites at Nanshan from other districts. On the other hand, the emerging streets (shown in orange) are mainly regional roads inside Nanshan, e.g., Keyuan, Yueliangwan and Shahedong Rd. Their emergence reflects the increasing business activities during the working hours.

Fig. 7b and Fig. 7c depict the maps and clouds of Topic2 in two time windows. This topic relates to Futian district. The topic evolution visualizations lead to interesting findings. For 3pm to 6pm (Fig. 7b), one emerging street is Guangshen Expressway (in orange), from this region to the airport. This reveals that at this time period many people seek air travel. Meanwhile, the disappearing streets (in brown) are regional business streets such as Shennan and Caitian Ave., reflecting the end of working day. In contrast, Fig. 7c shows the opposite of the trend. Traffic toward the airport is reduced as Guangshen Expressway disappears (in brown). And the regional streets are added (in orange), maybe due to the starting of entertainment traffic flows.

6.4 City Topics and Taxi Trajectories

The probabilities of a trajectory over topics represent the trajectory's appearance over multiple topics. Users can identify interesting taxi paths related to the city from the PCP visualization of trajectories. Adminstration and taxi companies can use the tool to find, evaluate and regulate taxi drivers' performance. Fig. 8 illustrates several taxis with long paths during the morning hours from 6am-9am. They diversely travel several topical districts. First, users filter out thousands of taxi trajectories by showing only those taxis having the probabilities greater than 0.18 on at least three topics. Then, users can interactively select and highlight particular taxis of interest. Fig. 8a displays four highlighted taxis on the PCP (plate numbers are not shown for privacy protection). The corresponding trajectories are drawn in Fig. 8b, where Google Earth map is selected as the background. For example, TAXI2 has large frequencies on Topic2, Topic4 and Topic6, which reveals its trajectory connecting three diverse regions Luohu, Longhua, and Bao'an. This taxi may be profitable at the time, and the regulators may investigate whether the driver has taken unnecessary indirect routes.



Minimum Values 0.18

Number of Topics bigger than Minimum Value: 3

Selected Taxis(Click the lines in the plot!)

TAXI1, TAXI2, TAXI3, TAXI4

(a) PCP: 6am-9am

(b) Map: 6am-9am

Figure 8: *Taxi trajectories.*

7 Conclusion and Future Work

We innovate semantic transformation to enable the use of text analysis techniques on massive GPS sampling data. The LDA model is applied to generate taxi topics, which reflect inherent patterns of taxi mobility. We have developed a visual analytics system to explore massive taxi trajectories data based on the discovered topics. The visualizations support various visual analytics tasks.

In the future, we will investigate advanced LDA models (e.g. temporal LDA) for taxi data. We will study semi-automatic parameter definitions for optimal number of topic. We will also integrate more visualization techniques into our system.

Acknowledgements

This work is partially supported by NSF IIS-1352927 and IIS-1352893.

References

[1] J. Alsakran, Y. Chen, D. Luo, Y. Zhao, J. Yang, W. Dou, and S. Liu. Real-time visualization of streaming text with a force-based dynamic system. *IEEE Comput. Graph. Appl.*, 32(1):34–45, Jan. 2012.

[2] G. Andrienko, N. Andrienko, J. Dykes, S. I. Fabrikant, and M. Wachowicz. Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research. *Information Visualization*, 7(3):173–180, June 2008.

[3] G. L. Andrienko and N. V. Andrienko. Visual analytics for geographic analysis, exemplified by different types of movement data. Information Fusion and Geographic Information Systems, Lecture Notes in Geoinformation and Cartography, pages 3–17. Springer, 2009.

[4] D. M. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[5] T. Crnovrsanin, C. Correa, C. W. Muelder, and K.-L. Ma. Proximity-based visualization of movement trace data. In *IEEE VAST*, pages 11–18, October 2009.

[6] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2412–2421, Dec. 2011.

[7] W. Dou, X. Wang, R. Chang, and W. Ribarsky. Paralleltopics: A probabilistic approach to exploring document collections. In *IEEE VAST*, pages 231–240, 2011.

[8] M. Ester, H. peter Kriegel, J. S, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of KDD)*, pages 226–231, 1996.

[9] 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 Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, 2013.

[10] Y. Gao, P. Xu, L. Lu, H. Liu, S. Liu, and H. Qu. Visualization of taxi drivers' income and mobility intelligence. In *ISVC (2)*, pages 275–284, 2012.

[11] H. Guo, Z. Wang, B. Yu, H. Zhao, and X. Yuan. Tripvista: Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection. In *Proceedings of the 2011 IEEE Pacific Visualization Symposium*, pages 163–170, 2011.

[12] C. Hurter, B. Tissoires, and S. Conversy. Fromdady: Spreading aircraft trajectories across views to support iterative queries. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1017–1024, Nov. 2009.

[13] T. Kapler and W. Wright. Geo time information visualization. *Information Visualization*, 4(2):136–146, July 2005.

[14] J.-G. Lee, J. Han, X. Li, and H. Gonzalez. Traclass: trajectory classification using hierarchical region-based and trajectory-based clustering. *Proc. VLDB Endow.*, 1:1081–1094, August 2008.

[15] J.-G. Lee, J. Han, and K.-Y. Whang. Trajectory clustering: a partition-and-group framework. In *Proceedings of the ACM SIGMOD*, pages 593–604, New York, NY, USA, 2007. ACM.

[16] T. Li. A map matching algorithm for public traffic control center based on long space and mass gps data. *Master Thesis, Huazhong University of Science and Technology*, 2012.

[17] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye. Mining periodic behaviors for moving objects. In *KDD*, pages 1099–1108. ACM, 2010.

[18] Z. Li, M. Ji, J.-G. Lee, L.-A. Tang, Y. Yu, J. Han, and R. Kays. Movemine: mining moving object databases. In *SIGMOD*, pages 1203–1206, New York, NY, USA, 2010. ACM.

[19] H. Liu, Y. Gao, L. Lu, S. Liu, L. Ni, and H. Qu. Visual analysis of route diversity. *IEEE Conference on VAST*, pages 171–180, 2011.

[20] X. Liu, J. Biagioni, J. Eriksson, Y. Wang, G. Forman, and Y. Zhu. Mining large-scale, sparse gps traces for map inference: comparison of approaches. In *KDD*, pages 669–677. ACM, 2012.

[21] M. Nöllenburg. Geographic visualization. *Human-Centered Visualization Environments*, pages 257–294, 2007.

[22] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan. Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4):42:1–42:32, Aug. 2013.

[23] J. Pu, S. Liu, Y. Ding, H. Qu, and L. Ni. T-watcher: A new visual analytic system for effective traffic surveillance. In *Proceedings of Mobile Data Management*, pages 127–136, 2013.

[24] S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, and F. Giannotti. Discovering the geographical borders of human mobility. *KI - Knstliche Intelligenz*, 26:253–260, 2012.

[25] B. Speckmann and K. Verbeek. Necklace maps. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):881–889, 2010.

[26] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko. Stacking-based visualization of trajectory attribute data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2565–2574, 2012.

[27] T. von Landesberger, G. Andrienko, N. Andrienko, M. Tekusova, and S. Bremm. Visual analytics methods for categoric spatio-temporal data. In *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, VAST '12, pages 183–192, 2012.

[28] H. Wang, H. Zou, Y. Yue, and Q. Li. Visualizing hot spot analysis result based on mashup. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, LBSN '09, pages 45–48, New York, NY, USA, 2009. ACM.

[29] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. van de Wetering. Visual traffic jam analysis based on trajectory data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2159–2168, 2013.

[30] N. Willems, H. van de Wetering, and J. J. van Wijk. Visualization of vessel movements. In *EuroVis*, pages 959–966, Aire-la-Ville, Switzerland, Switzerland, 2009. Eurographics Association.

[31] J. Wood, J. Dykes, A. Slingsby, and K. Clarke. Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1176–1183, Nov. 2007.

[32] H. Yoon and C. Shahabi. Accurate discovery of valid convoys from moving object trajectories. *Data Mining Workshops, International Conference on*, 0:636–643, 2009.

[33] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. T-drive: driving directions based on taxi trajectories. In *Proceedings of the GIS'10*, pages 99–108, 2010.

[34] J. Zhao, P. Forer, and A. S. Harvey. Activities, ringmaps and geovisualization of large human movement fields. *Information Visualization*, 7(3):198–209, June 2008.

[35] Y. Zheng and X. Zhou. *Computing with Spatial Trajectories*. Springer, 2011.