Visual Analysis of Multiple Route Choices based on General GPS Trajectories

Min Lu, Chufan Lai, Tangzhi Ye, Jie Liang, *Member, IEEE,* and Xiaoru Yuan, *Senior Member, IEEE*

Abstract—There are often multiple routes between regions. Drivers choose different routes with different considerations. Such considerations, have always been a point of interest in the transportation area. Studies of route choice behaviour are usually based on small range experiments with a group of volunteers. However, the experiment data is quite limited in its spatial and temporal scale as well as the practical reliability. In this work, we explore the possibility of studying route choice behaviour based on general trajectory dataset, which is more realistic in a wider scale. We develop a visual analytic system to help users handle the large-scale trajectory data, compare different route choices, and explore the underlying reasons. Specifically, the system consists of: 1. the interactive trajectory filtering which supports graphical trajectory query; 2. the spatial visualization which gives an overview of all feasible routes extracted from filtered trajectories; 3. the factor visual analytics which provides the exploration and hypothesis construction of different factors' impact on route choice behaviour, and the verification with an integrated route choice model. Applying to real taxi GPS dataset, we report the system's performance and demonstrate its effectiveness with three cases.

Index Terms—Route Choice Behaviour, Visual Analysis, Interaction, Route Choice Model

1 Introduction

With the development of sensing technologies, a variety of *big data* has been produced in urban space. Urban computing combines urban sensing, data management, analytics and services as an integral process, which throws light on the rich knowledge of city and improves people lives [1]. Transportation is one of the most essential urban computing applications. Many transportation systems analyse the city-wide human mobility data and other urban data (e.g., weather data) to understand the travel behaviour [2], [3] and improve the travel experience [4].

In modern traffic networks, there are often multiple routes to choose from when travelling from one place to another. Understanding how drivers make route choices, i.e., the route choice behaviour, is an interesting topic in transportation area. It not only assists the city planners in the improvement of route usage, but also helps drivers make wise travelling decisions.

However, route choice behaviour is not an easy problem. Drivers choose different routes considering different factors. The expected time cost is one example. Choosing the route with minimum time cost is what widely experienced in daily life. Some other factors may also influence route decision making, like the number of traffic lights, travelling comfortableness, etc. Meanwhile, the impact of factors may change over time. Drivers who care about travel efficiency on workdays, might trade it off with the travelling comfortableness at weekends. Moreover, the problem is even more complex when various factors interact with each other.

Classically, research efforts have been made to study the influence of different factors on route choices based on Stated Preference (SP) survey data [5]. SP survey collects the route preferences in hypothetical situations from respondents. Different choice considerations, such as travel safety, can be directly captured by the information in questionnaires. With SP data, various route choice models [6], [7] are developed, trying to estimate the impact of different factors on the route choice behaviour. However, such investigations are limited in scale and the surveys need to be carefully designed. Also, information obtained from investigation is quite subjective and not practically reliable enough. In more recent years, some researchers perform the analysis with the help of Global Positioning System (GPS) where GPS receivers are used to collect trajectories from volunteers. Compared to traditional investigations, it takes less effort and is more realistic. But such pilot studies are often conducted among a limited number of users in a restrained spatio-temporal scale, like only collecting morning commute trips [8], [9].

In this work, we explore the possibility of studying route choice behaviour based on more general GPS trajectory data, i.e., taxi GPS trajectories. Taxi trajectories are sampled in real situation and cover a wider spatial and temporal range. However, challenges arise when studying route choice behaviour based on general GPS trajectories:

- *Extract relevant trajectories in the context of multiple routes:* Unlike the experimental GPS trajectories constrained in limited spatial and temporal range, extracting trajectories related to multiple routes from massive trajectories is a challenge to be tackled.
- *Raise hypotheses on factors that significantly influence the route choice:* Different from the verification of predefined factors in hypothesis-oriented experiments, it is a challenge to detect factors that potentially influence route choice from general GPS trajectories and verify the significance.

Visual analytics is proposed as the science of analytical reasoning

- *Min Lu, Chufan Lai, Tangzhi Ye and Xiaoru Yuan are with Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University, Beijing, P.R. China, 100871.*
- *Jie Liang is with Engineer and Information Technology, University of Technology, Sydney. This is a joint work when visiting Peking University.*
- *E-mail: {min.lu, chufan.lai, yetangzhi, xiaoru.yuan}@pku.edu.cn, {jie.liang}@uts.edu.au. To whom correspondence should be addressed, email xiaoru.yuan@pku.edu.cn*

Manuscript received November 30, 2015

facilitated by interactive visual interface [10]. By integrating computational and theory-based tools with innovative interactive techniques and visual representations, visual analytics enables human to participate in problem solving.

In this work, from the perspective of visual analytics, we propose a visual analytics system which leverages human interaction and judgement in the trajectory data mining process [11] to tackle the above challenges: with a suite of graphical filters, trajectories between regions of interest are queried interactively; based on filtered trajectories, feasible routes are constructed automatically; with a list of factors derived from general GPS trajectory data, route choice distributions over those factors are visualized, which supports to explore and raise hypotheses on potential influence; then the hypotheses are further verified by the statistical model to draw reliable conclusions.

The contributions of this work are:

- We explore the possibility of analyzing multiple route choice behaviour based on general GPS data.
- We develop a visual analytic system to explore the route choice behaviour with real GPS data.

For the remaining part, we first report related work in Section 2. Then in Section 3, we give an overview of the data background, analytic tasks and overall system pipeline. Details of route generation and visual design are explained in Section 4 and Section 5. We report the system's performance in Section 6 and demonstrate its usage in Section 7. In Section 8, we have a discussion on the system. Finally it comes the conclusion.

2 Related Work

In this section, we have a discussion on the related work: route choice behaviour analysis in transportation field, research progress in visual analytics of trajectories, rank-based visualization and route visualization.

2.1 Route Choice Behaviour Analysis

Route choice behaviour has been widely studied in the transportation area. In early years, most researches are based on statistical investigations or experiments. By analysing a total of 2182 home-to-work records in Seattle, Mannering et al. [12] find that 26% people do not always use the same route. To find the reasons, Khattak et al. [13] study 700 commute trips collected via questionnaires, and find that both congestion and the perception of alternative routes increase the probability of route changes. With respect to personality, males, young people and experienced drivers are more likely to change routes, as concluded by Xu et al. [14] in a study of 247 morning home-to-work trips. In these works, statistical inquiries play an important role, where questionnaires are carefully designed to obtain problem-related information involving personal details. However, investigations are limited in both the sample range and its validity. Realism is also a problem given the divergence between recalled and observed circumstances.

To obtain more authentic information, some researchers base their studies on GPS data in recent years. Li et al. [8] study morning route choice patterns based on a GPS dataset collected from 182 vehicles in 10 days. Factors like age, departure time and income level are found convincingly influential. More recently, Vacca et al. [9] study route switch behaviour between the same OD (i.e., Origin-Destination) pair by tracking the participants with portable GPS devices. Some dominant factors are revealed, such as traffic light number (per km), highway percentage, perception of time, etc. Compared with investigations, GPS records provide more truthful measurement of route choice behaviour, with lower costs and higher precision. However, subject to the analytical requirement of individual characteristics, the data is still problem-related and range-limited. Instead, our system is designed for general GPS data covering a much larger range (tens of thousands of taxies). One similar work is proposed by Pan et al. [15]. They extract regular routing patterns from the historic taxi trajectories and detect the anomaly routing behaviour that significant differ from the original patterns. Based on social media data, they focus on exploring semantic meaning of the travel anomalies. Different from their semantic exploration, our work focuses on comparing the properties of multiple routes and exploring the regular factors that impact route choice behaviour, such as the departure time. Meanwhile, what's provided in our system can support interactive data customization and real-time processing according to different analytical demands.

2.2 Trajectory Visual Analysis

In trajectory mining field, Zheng [11] survey various mining techinques, including outlier detection, pattern mining, etc. From the apsect of visual analysis, Andrienko et al. [16] present a taxonomy of generic analytic techniques based on possible types of movement data. For trajectories, there are three kinds of visual explorations [17]: direct depiction, pattern extraction and visual aggregation. Direct plotting could simply fail because of visual cluttering. Pattern extraction methods employ automatic analysis to extract underlying data patterns [18], e.g., the traffic jam propagation graph extraction [19]. Aggregation methods visualize movement groups to reveal the high-level movement graph. Guo [20] and Andrienko [21] et al. construct geographical regions and visually aggregate the in-between movements as flows. Lu et al. [22] design *ODWheel* to compare the traffic dynamics from the centeral region from/to its departing/arriving regions. Besides aggregation between regions, travel behaviour within interchange region can also be visualized. Guo et al. [23] provide a circular design to explore movement at a road intersection. Zeng et al. [24] derive a visualization from Circos [25] to display interchange traffic flow at subway transition stations.

Liu et al. [26] study the route diversity between locations and provide a clock like radial layout to display temporal statistic distribution. Different from analysing individual trajectories in Liu et al.'s work, our method provides analysis based on the extracted topology structure. Zeng et al. [27] visualize the mobility of routes starting from a single source in public transportation system and provide the comparison among different routes. Similar to their routes' comparison, our work provides comparison among multiple routes.

Alternative to analyse trajectories as a whole, some works perform local analysis of the filtered trajectories of interest. Andrienko et al.'s book [28, Chapter 4.2] summarizes the different kinds of filtering, including the spatial, temporal filtering, etc. Marios et al. [29] introduce spatial query which specifies a spatiotemporal pattern as a sequence of distinct spatial predicates. Vieira et al. [30] design the trajectory query using regular expression over a spatial alphabet of regions. Different from those textural query languages, visual query languages explore the spatial data graphically. Ferreira et al. [31] propose a visual

TABLE 1
Table of Derived Factors

Object	Attribute	Description	Motivation
Route	Route's Length	Route geographical length	Do drivers prefer shorter route?
	Traffic Light Number	Total number of traffic lights along the route	Do drivers prefer less traffic lights?
	Route Importance	The average road level of the route	Do drivers prefer route with higher level road?
	Time Cost Distribution	Time cost distribution of a route	Do drivers prefer route with less time cost? Do drivers prefer route whose time cost has less variation?
Trajectory	Departure Time in a Day	Departure time in the time scale of day	Do drivers departing at different time make different route choices?
	Departure Day	Departure day in the time scale of a week	Do drivers departing on different days make different route choices?
	Trajectory's Length	Total travel distance of the trip	Do drivers travelling to different distances choose different routes?

query model to filter trajectories by their origins and destinations. TrajectoryLenses [32] supports users to interactively filter the trajectories by manipulating the lenses on the map. Similar to TrajectoryLenses, we design a suite of circular filters in this work. Compared with TrajectoryLenses, our design not only supports more spatial constraints but also allows for the direction assignment.

2.3 Rank-based Visualization

Ranking as an operation to organize data in order is widely used in visualization, especially when comparing data items over multiple attributes. Because of the linear property of ordering, ranking technique is usually integrated into line-based visualizations [33]. Parallel Coordinates [34] visualizes multivariate data by connecting items' actual value over multiple attributes, which embeds the ranking implicitly. Instead of actual value, Bump Charts [35] explicitly visualizes data by order and connect order change with slopes. One more recent ranking design is LineUp [36], which not only visualizes the ranking changes, but also encodes the cause of the rank. Lu et al. [37] aggregate trajectories along a single route and rank them by the time cost along the road segments, to reveal mainstream and outliers. Similar to those ranking techniques, we rank routes over attributes for comparison. However, in our case, we need to deal with dynamic route attributes, e.g., the travel time cost attribute of a route which ensembles the time costs from all trajectories. Some ranking visualizations deal with dynamic changes by expanding the time dimension. Batty [38] designs Rank Clocks to show the change of city population rankings across several centuries, which is similar to Parallel Coordinates but represents different time as axes. On the other hand, keeping time continuous, Shi et al. [39] propose RankExplorer, in which they segment the rankings into several groups and use a ThemeRiver [40] to show their temporal changes. Instead of expanding time, we aggregate the dynamic route attribute samplings by trajectories and propose a ranking visualization for attributes with single value and multiple values.

2.4 Route Visualization

To visualize a path, a well-known technique in geographical application is the space-time cube [41], [42], which visualizes the dynamic changes of geographical of a path in 3D space. Tominski et al. [43] propose stacking bands in hybrid 2D/3D view to visualize the trajectory attributes. With the metaphor of lenses, Karnick et al. [44] place magnified lenses on the significant points along a route, to encode more details. The other way is to do distortion. Agrawala [45] distorts and simplies routes to highlight important features which are similar to human drawing maps. Alternatively, Sun et al. [46] distort the map to broaden the roads of interest so that temporal information can be embedded. In this work, we keep the map view undistorted to maintain easy perception of the geospatial information of routes. With careful design, the topological information of multiple routes is encoded.

For sufficient analysis among routes, Zeng et al. [27] present an isotime flow map view in a parallel isotime fashion. There are similar flow diagrams [47], [48] when broadening the horizontal representation to temporal dimension. We derive the abstract route view from flow diagram to show the topology structure.

3 Overview

In this section, we first introduce the data and tasks. Then we present pipeline of the visual analytics system.

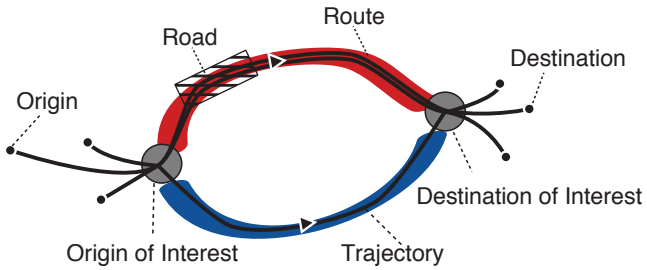

Fig. 1. Illustration of Relevant Concepts in Multiple Routes

3.1 Data

To facilitate our discussion, we list the common terminologies as illustrated in Figure 1:

- *Trajectory* records a list of positions that an object travels in temporal order.
- *Origin/Destination (O/D)* refers to the beginning/ending position of the trajectory.
- *Road* is the physical connection between one location to another, where vehicles can travel on.
- *Route* is a sequence of roads that vehicles travel through.
- *Origin/Destination of Interest (OoI/DoI)* refers to the beginning/ending position of movement that the analyst is interest in.
- *Multiple Routes* are all the travelling routes between a pair of *OoI* and *DoI*.

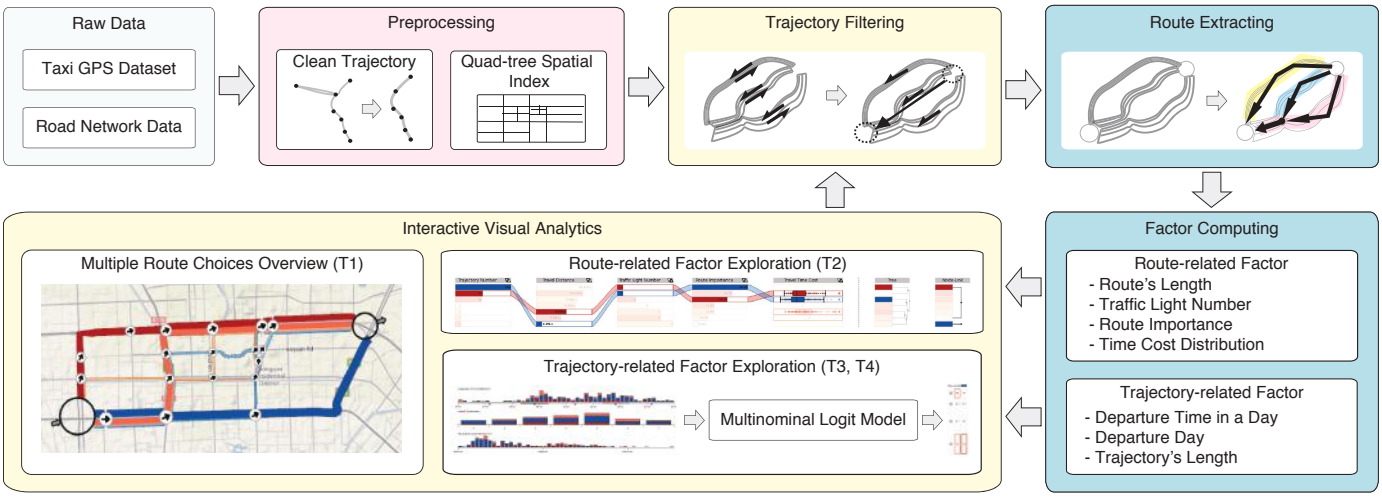

Fig. 2. The System's Pipeline: consists of data preprocessing (pink), automatic computation (blue) and human involved visual analytic module (yellow).

Note that *OoI*/*DoI* is not necessarily *O*/*D*. *OoI*/*DoI* can be placed at the region of interest where multiple route choices are concerned.

Different from the predefined factors in the controlled experiments, factors in this work are directly derived from general GPS dataset. In Table 1, factors are categorised into two groups: *route-related factors* and *trajectory-related factors*. For each route, some inherent attributes probably play as factors in route decision making, e.g., the length of route, the number of traffic lights along the route and route importance. Specifically, the route importance refers to the average road level of the route. The road level is distinguished by what type the road is, like a trunk or residential road, which is implied by the *highway* tag in OpenStreetMap [49]. Besides those static route attributes, the time cost distribution sampled by historical passing vehicles is viewed as another route-related factor, whose average and variance potentially influences the route choice.

On the other hand, each trajectory has individual differences that potentially affect the route choice, e.g., departure time in a day, departure day and trajectory's length. For example, drivers travelling in peak time and off-peak time may make different route decisions. It is also possible for drivers to select different routes when travelling in different distances from O to D.

3.2 Analytic Tasks

In this part, we clarify analytic tasks to explore route choice behaviour with general GPS trajectories. According to the typology of visualization tasks [50], our analytic tasks are designed from high-level to low-level. Given a pair of *OoI* and *DoI*, firstly, an overview is given to present multiple feasible routes. Then the route-related factors and trajectory-related factors are calculated and visualized. It helps to build hypotheses about how factors impact the route choices. With the hypotheses on certain routes and factors, the system should be capable of examining the proposed hypotheses to tell if the impact is significant.

With these considerations, the design tasks are summarized as following:

- *Overview of multiple route choices* ***(T1)***: give an overview of all feasible route choices between *OoI* and *DoI*.
- *Exploration of the route-related factors' impact on route choice* ***(T2)***: describe each route by route-related factors and compare routes in terms of route-related factors.
- *Building hypotheses of the impact of trajectory-related factors on route choice* ***(T3)***: explore the route choice distribution over trajectory-related factors and propose hypotheses on the potential impact.
- *Evaluating the impact of trajectory-related factors* ***(T4)***: build a statistic model to examine whether the impact is significant or not.

3.3 System Overview

To support the above tasks, we propose a visual analytic system integrating automatic processing, visualization and interaction. Figure 2 shows the system's pipeline.

In the preprocessing stage, trajectories are cleaned. A quad-tree spatial index is built to facilitate filtering in massive trajectories.

In run-time stage, trajectories between a pair of *OoI* and *DoI* are filtered using a suite of graphical filters. With those filtered trajectories, all feasible routes are extracted. A topology graph of the routes is constructed using a grid-based algorithm.

For each route or trajectory, related factors (discussed in Section 3.1) are derived. Then those routes and factors are fed as input to the visual analytics module.

The visual analytics module consists of three parts. The spatial visualization gives a geographical overview of the multiple routes (**T1**). The route-related factor view displays the route-related factors in a ranking diagram. Users can compare them across different routes (**T2**). The trajectory-related factor view visualizes different route choices over trajectory-related factors. This view supports the proposal of hypothesis (**T3**). Then users can input their hypotheses. A choice analysis model, i.e., Multinominal Logit model (MNL) [51], is used for the verification (**T4**). After modelling, the results are visually integrated back in the trajectory-related factor view, to tell whether the impact is significant or not.

The three views cooperate in a brushing and linking manner, i.e., entities selected in one view are updated in other views. At last, users can launch a new loop of analysis by resetting the filtering.

4 Multiple Routes Generation

For the massive taxi GPS trajectories, the system provides a suite of graphical fitlers. They support to query trajectories intuitively with spatial and temporal constraints. With the filtered trajectories, a grid-based algorithm is proposed to extract all feasible routes automatically.

4.1 Trajectory Filtering

From the temporal aspect, a two-level temporal filter is provided: date and time. Date range is set in the date filter. Time range in a day is set in the time filter, whose granularity is 10 minutes. With these two different temporal granularities, the temporal filter allows users to query trajectories in a periodic pattern, such as the commute trips in the morning.

From the spatial aspect, we design the filter similar to TrajectoryLenses [32]. The filter covers a circular area and filters trajectories with 6 spatial constraints. The 6 constraints are defined according to the spatial relationship between trajectory and the underlying circular area: *origin*, *destination*, *origin/destination*, *passing*, *inclusive*, and *exclusive*. The concepts are shown in Figure 3(b). For example, a filter with the *origin* constraint filters trajectories starting from the circular area. Besides the spatial constraints, there are some other geometric constraints, e.g., the center position and radius of the circular area. For usage simplicity, constraint configuration is embedded into the circular filter. As Figure 3(a) shows, when hovering on a certain region, certain function is waked and the corresponding handle is shown. For example, hovering in the center of the circle invokes the moving function and a + handle is visible. Clicking and dragging the + changes the center of the filter. Complex queries can be built which combines different filters in an intersection manner. Moreover, for two or more filters, directions can be assigned between filters to select trajectories following certain flow directions. For the ease of constraint perception, constraints are explicitly encoded in the circular filter. Figure 3(b) shows the circular filters with 6 spatial constraints respectively. In this work, the first two filters are detected as the *OoI* and *DoI* by default.

4.2 Multiple Routes Extraction

With the filtered trajectories from *OoI* to *DoI*, we employ a general grid-based algorithm to extract multiple routes automatically. The basic idea is to cover the trajectories by grid and then build up the multiple route graph among cells of the grid.

Figure 4 illustrates the process of route extraction. Figure 4(a) shows the filtered trajectories between *OoI* and *DoI*. At the beginning, a uniform grid is covered over the boundary box of filtered trajectories, which divides the space into cells (Figure 4(b)). Trajectories are segmented by the cells and each of them can be denoted by the sequence of passing cells (Figure 4(c)).

Each cell collects the segments from trajectories which intersect with it. Then for each cell that contains segments, we derive the average direction from trajectory segments inside it. The directions are further approximated as horizontal or vertical ones (Figure 4(d)). The horizontal direction is more likely the left-right going than the up-down going and the vertical one is more likely the up-down going. To remove the zigzag between two cells, two types of ambiguous cells are detected: the neighbour cells with horizontal direction which are side-by-side horizontally; the neighbour cells with vertical direction which are side-by-side vertically. The detected cells are merged (Figure 4(e)).

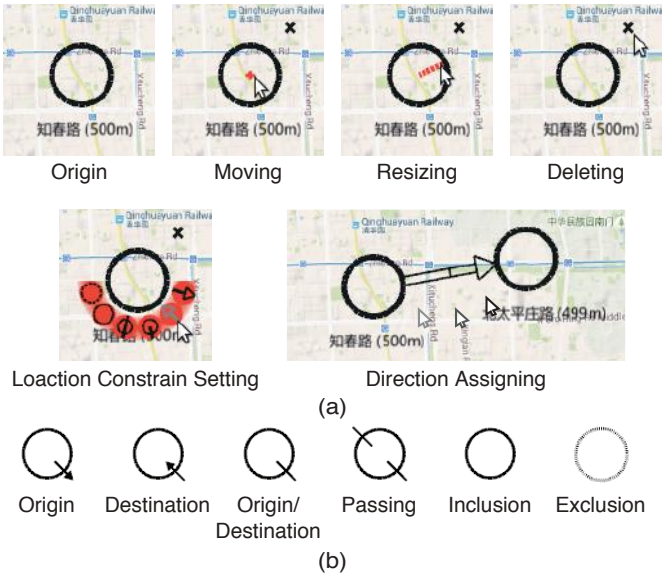

Fig. 3. Circular Filter: (a) different functions are invoked by hovering on corresponding regions. Direction between filters is assigned by dragging from one to another. (b) 6 circular filters with different spatial constraints.

After that, routes are formed by linking the centroids of cells (Figure 4(f)). Cells with more than one in/out degree are detected as the splitting/merging nodes (Figure 4(g)). The multiple route graph is constructed with these nodes and the routes connecting them. Finally, multiple routes are encoded visually (Figure 4(h)), which will be introduced in Section 5.1.

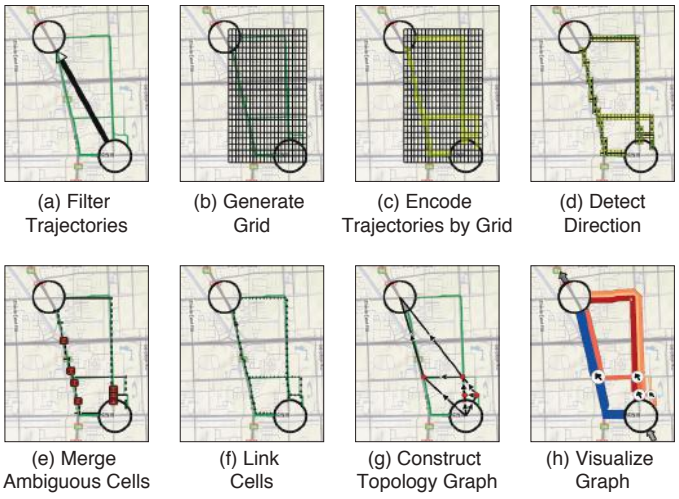

Fig. 4. Multiple Routes Construction: by covering a grid over trajectories, multiple route graph is built upon travelled cells.

5 Visual Design

In this section, we present design of visualizations in our system. Corresponding to tasks introduced in Section 3.2, the interface mainly consists of three parts: the route spatial view, the route-related factor view and the trajectory-related factor view.

5.1 Route Spatial View

With the specified *OoI* and *DoI*, multiple routes are obtained by the algorithm introduced before. To provide an overview of all the

feasible routes (**T1**), the spatial view is designed with following considerations:

- **Represent *OoI* and *DoI* (CI)**: to locate the areas of *OoI* and *DoI*.
- **Display multiple routes (CII)**: to visualize the feasible routes between *OoI* and *DoI*, including both the popular ones and the seldom travelled ones.
- **Indicate traffic flow directions (CIII)**: to show the travelling directions along routes, especially at the intersections.
- **Summarize the routing (CIV)**: to summarize the major route choices by merging similar routes

After defining the filtering conditions (see Section 4.1), the *OoI* and *DoI* circular filters are settled on the map. To indicate *OoI* and *DoI* filters, inward and outward arrows are attached to the circular filters respectively (**CI**) (Figure 5(a)).

Each extracted route is visualized as a band, whose width encodes the number of passing trajectories. A logarithmic mapping is used to maintain the visibility of seldom travelled routes (**CII**). Routes are stacked together when sharing the same roads. When hovering, a tooltip is shown to facilitate selection of the bands (Figure 5(b)). The number of travelled trajectories is also displayed in the tooltip. The current hovered route is highlighted both in the spatial view and the tooltip. Users can easily switch the focus in the tooltip, in case that some routes are too small to choose on the map.

Considering that directions in straight roads are self-evident, we only indicate the traffic direction at the crossings of roads (**CIII**) using glyphs. The size of glyph encodes the volume of passing traffic flows. The arrow inside the glyph implies the average traffic flow direction at the crossing.

In order to summarize the complex routing, we divide the routes into a few groups (**CIV**). Each group contains a mainstream route and some alternative routes. We first choose the popular routes as the mainstreams. Specifically, the route whose traffic volume is larger than third quartile Q_3 of the whole traffic volume distribution are regarded as the mainstreams. The maximum number of mainstream routes is limited to 5 in order to avoid excessive dividing. With mainstreams determined, the remaining are assigned to the mainstream routes according to the topology similarity. In our case, we denote a route as a sequence of its road crossings, and use the edit distance [52] to measure the similarity between routes, which counts the minimum amount of switches required to transform from one sequence to the other. We show the grouping results in a topology graph to help understand the routing (Figure 6(a)). Each mainstream with its similar alternative routes are considered as a group. Qualitative colors [53] are used to differentiate different groups. Within each group, all routes are colored similarly, with the lightness inversely proportional to the route popularity. Figure 6(b) shows the color legend of the two groups in Figure 6(a). The color legend is consistent over all views.

5.2 Route-related Factor View

Inspired by ranking visualizations (e.g., LineUp [36]), we design a ranking-based visualization to support exploration on route-related factors' impact on route choice behaviour (**T2**). The ranking-based visualization helps users interpret how the factors affect route choices. There are several considerations we have taken in the design:

- **Accommodate different factor types (CI)**: to visualize both static and dynamic factors.

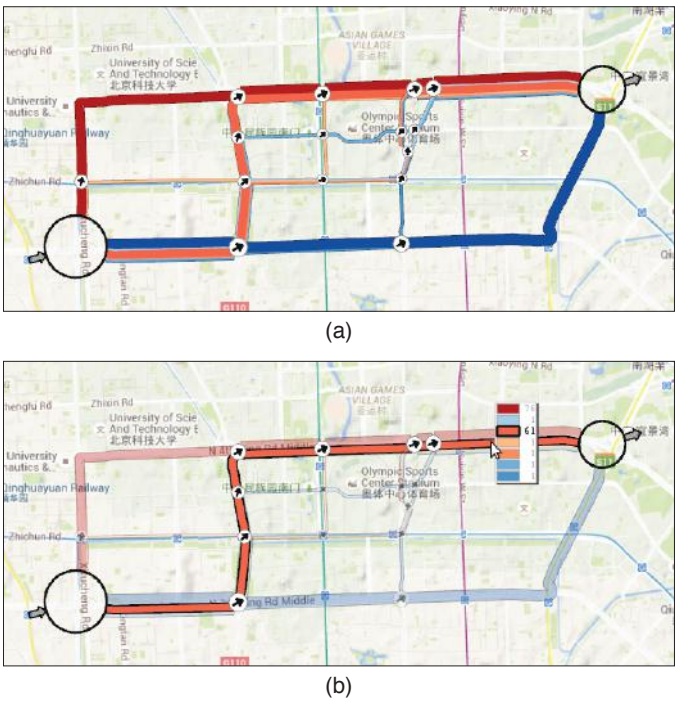

Fig. 5. Route Spatial View: (a) geographical overview of multiple routes: the route width encodes the amount of traffic flow. The arrow glyph at each intersection indicates the average flow direction. (b) one highlighted route in the road segment tooltip: all stacked routes are displayed in the tooltip to facilitate selection.

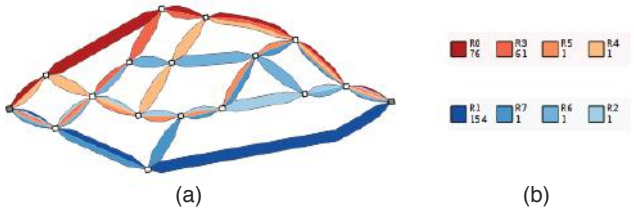

Fig. 6. Topology-based Color Scheme: (a) topological graph of multiple routes: all routes are categorized into a few groups based on topology similarity. (b) the global color scheme for routes based on topology-based grouping.

- **Compare factors among multiple routes (CII)**: to enable the comparison of route factors.
- **Explore the routes in topological relationship (CIII)**: to support exploration on those routes in topological relationship, i.e., similar routes.

Figure 7(a) displays the design of route-related factor view. The view mainly contains two parts: route-related factors in the rank list (left part) and topological relationships in the dendrogram (right part). In the factor part, each row represents a route, whose color is consistent with the route spatial view. In the left part of Figure 7(a), each column represents one route-related factor. The static factor, e.g. route's length, is depicted by a bar whose horizontal width encodes its value. The dynamic factor, e.g., time cost distribution of the route, is represented by a horizontal-aligned box plot (**CI**). The box plot preserves the outliers as dots. Each factor can be ranked for comparison (**CII**). The rank list can be sorted in the increasing/decreasing order by clicking the triangle/inverted triangle buttons beside to the label. It is easy to

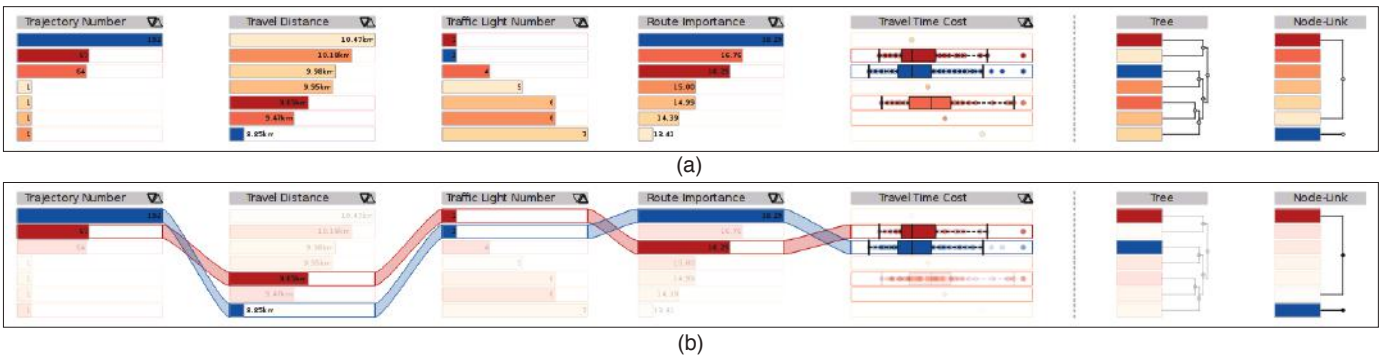

Fig. 7. Route-related Factor View: (a) with ranked factors. (b) with two selected routes.

rank the static factors because of single value. However, it is not straight forward to rank the dynamic factor, because each of them is a distribution. Considering median as the one of the typical representatives of a distribution, we choose to rank the dynamic factor by median in our case. Moreover, to maintain an intuitive visual tracking of routes, factors belonging to the same route are connected by the semi-transparent links (see Figure 7(b)).

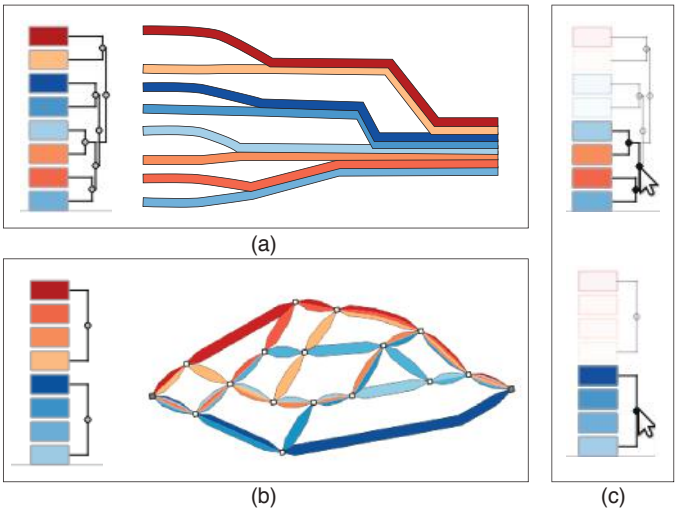

Fig. 8. Two Types of Topological Relationships: (a) encodes how the routes split from others. (b) encodes the similar groups in terms of the edit distance. (c) by clicking the corresponding node, a group of routes in certain topological relationship can be selected.

It is interesting to compare routes in close topological relationship, which is driven by the curiosity of how the similar routes (i.e., routes with overlapped parts) differ in route-related factors. Hence, to support topological exploration (**CIII**), two types of topological relationships are integrated in the right of the view (Figure 7(a)): the tree structure and the node-link structure. These two different types measure the topological similarity between routes from different perspectives. Figure 8(a)(b) illustrate them and their visual encoding respectively. In the tree structure (Figure 8(a)), from right to left, the hierarchical structure shows how the routes split apart. In the node-link structure (Figure 8(b)), it shows the groups of similar routes according to the edit distance, which has been introduced in Section 5.1. Uesrs can not only select each single route in the views, but also select a group of routes. As Figure 8(c) shows, the nodes in the topological structure can be clicked to select several routes. Especially, in the tree structure, routes in coarse to fine similar relationship can be selected by the nodes from right to left.

5.3 Trajectory-related Factor View

As discussed in Section 3.1, trajectory-related factors are important to explain the route choice diversity. Three trajectory-related factors are derived from general GPS trajectories. In this section, we first introduce the statistical model used to validate the factor impact. Then we present the visualization and interactions that help with hypotheses construction (**T3**) and verification (**T4**). To simplify the discussion, we call the trajectory-related factors as 'factors' in this section.

5.3.1 Multinomial Logit Model

To verify factors' impact on route choices, we adopt the Multinomial Logit (MNL) model [51]. It is simple, understandable, and widely used in the transportation area for route choice analysis [54, Chapter 7.3]. The basic assumption of MNL is that people always choose the option with the maximum utility. Assume that there are M people choosing from N routes, the utility is measured as follows:

$$U_{ij} = \boldsymbol{\beta}_i \boldsymbol{X}_j + \varepsilon_i, i = 1, \cdots, N, j = 1, \cdots, M \tag{1}$$

The U_{ij} here represents utility of the i-th option assessed by the j-th person. It consists of an observable part $\boldsymbol{\beta}_i \boldsymbol{X}_j$ and an unknown part ε_i. The $\boldsymbol{X}_j$ vector denotes observable factors of the j-th person, like the age, weight, height, etc. $\boldsymbol{\beta}_i$ is the coefficient vector of option i, a major output of the model. With the maximum utility assumption, the probability of person j choosing option i can be derived as follows:

$$P_j(i) = \Pr(U_{ij} > U_{kj}), \forall k \neq i \tag{2}$$

To eliminate the unknown term, probability is expressed explicitly in the MNL model as:

$$P_j(i) = \frac{e^{\boldsymbol{\beta}_i \boldsymbol{X}_j}}{\sum_{k=1}^{N} e^{\boldsymbol{\beta}_k \boldsymbol{X}_j}} \tag{3}$$

From the equation (3), we can see that the probability varies monotonously with the observable term:

$$\boldsymbol{\beta}_i \boldsymbol{X}_j = \sum_{t=1}^{D} \beta_{it} X_{jt} \tag{4}$$

Here, D denotes the number of observed factors, and the coefficient β_{it} depicts the impact of factor X_{jt} on option i. The

model can be further simplified by setting a basic option, e.g., option m. Other probabilities are made relative to the base:

$$\ln(P_j(i)) - \ln(P_j(m)) = \ln(\frac{e^{\boldsymbol{\beta}_i \boldsymbol{X}_j}}{e^{\boldsymbol{\beta}_m \boldsymbol{X}_j}}) = (\boldsymbol{\beta}_i - \boldsymbol{\beta}_m)\boldsymbol{X}_j = \boldsymbol{\beta}_i^{'}\boldsymbol{X}_j \quad (5)$$

The new coefficient $\boldsymbol{\beta}_i^{'}$ shows the impact of factors $\boldsymbol{x}_j$ when choosing between option i and m. This term can be expanded into $\boldsymbol{\beta}_i^{'}\boldsymbol{X}_j = \sum_{t=1}^{D} \beta_{it}^{'} X_{jt}$. Specifically, if $\beta_{it}^{'}$ is positive, the person is more likely to choose option i over the base option m, when X_{jt} increases. Otherwise, the person prefers the base option m when X_{jt} increases. This exactly explains how the factor influences the final choice. The result is only meaningful when it is tested significant. So we also derive the confidence level (p-value) to verify the significance of the impact. All in all, the coefficient matrix $\boldsymbol{\beta}_i^{'}$ along with the p-values, can validate the influence of factors on route choices.

In our context, it is worth studying why some people did not choose the major route. Hence, we set the most popular route (i.e., the route followed by most trajectories) as the base option m. There are 3 observable factors, i.e., departure time in a day, departure day and trajectory's length (discussed in Section 3.1). However, not all factors X_{jt} are numerical. The trajectory's length is numerical, so it is easy to explain what happens when the trip is longer. But the other two factors, i.e., the departure time in day and departure day are ordinal. There is intrinsic chronological order but not the numeric order. For example, Saturday is chronologically later than Friday, but not numerically larger than Friday. To solve this problem, those two ordinal factors are divided into C categories. $C-1$ binary variables are introduced to indicate the absence or presence of certain category. Hence, when X_{jt} comes to the departure time in a day or departure day, it becomes a vector of dummy variables $[D_{jt_1}, D_{jt_2}, ...]$:

$$D_{jtc} = \begin{cases} 1, & x_{jt} \in [r_{c-1}, r_c] \\ 0, & x_{jt} \notin [r_{c-1}, r_c] \end{cases} \quad , c = 1, 2, ..., C-1 \quad (6)$$

Each binary variable acts as an independent sub-factor and its impact on route choices are studied. Different configurations of C and $[r_{c-1}, r_c]$ of each category investigate potential different impact. In our system, users are allowed to interactively divide a factor and customize dummy variables. This part will be introduced in Section 5.3.2.

Figure 9 shows an example of the input and output of the MNL model. Suppose the blue route is the basic route option and the departure time in a day is categorized into three categorizes. Then two binary variables are imported: 7:00 9:00 and 16:00 19:00. The matrix of $\boldsymbol{\beta}_{it}^{'}$ shows the impact of dummy variables. We can see that when departuring between 7:00 to 9:00, people are more likely to choose the orange route over the blue route.

		Orange VS Blue	Red VS Blue
Departure Time in a Day	⓪ 7:00~9:00	0.8 *	- 0.1
	② 16:00~19:00	- 0.3	- 0.1

* p - value < 0.05

Fig. 9. An Example Coefficient Matrix Output of MNL Modelling

5.3.2 Visual Design

Besides the model, we design the trajectory-related factor view. Users can observe factors' distributions and route choices, build hypotheses. Specifically, the trajectory-related factor view takes charge of two analysis tasks: one is to support the exploration of trajectory-related factors, and help users build hypotheses (**T3**); the other one is to verify those hypotheses, to see whether the impact is significant or not (**T4**). For the latter task, MNL statistical analysis model is applied, which can be configured interactively via the interface.

In general, the view is designed with following considerations:

- **Compare factors across multiple route choices (CI)**: to help users see how factors changes affect route choices, which facilitates the raising of hypotheses on potential impact.
- **Configure the statistical analysis model (CII)**: to support customization of the model, so as to validate different assumptions.
- **Indicate the factors' impact on route choices (CIII)**: to show the credible conclusions about the factors' impact on multiple route choices.

Figure 10 shows the designed visualization for one factor, the departure time in a day. Given the above three considerations, it is composed of three parts: the stacked bar chart, the factor configuration panel, and the factors' impact matrix. The stacked bar chart visualizes the distribution. Factor configuration panel configures the modelling and finally the factors' impact matrix shows the analysis result.

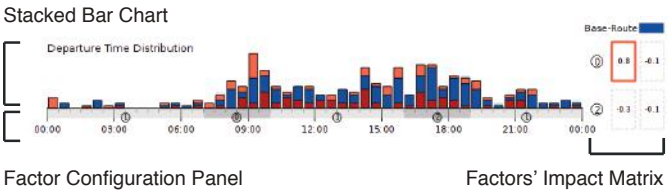

Fig. 10. Trajectory-related Factor Views: the stacked bar chart (left-top) visualizes distribution; factor configuration panel (left-bottom) supports factor customization; factors' impact matrix shows the output.

Dividing factor into interval bins, the population of certain route choice within each bin are counted. To visualize the population over bins, we choose bar chart which is widely used to display the distribution over the discrete bins. Shared space technique is more efficient for comparison in small visual spans than separate space [55]. Hence, bars of multiple route choices are stacked to share the space, i.e., the stacked bar chart. We use it to support the comparison over different route choices and explore the population changes. Meanwhile, stacking bars of different route choices together makes it easy to check the ratio of route choices in the same bin. The left-top part of Figure 10 shows an example of the stacked bar chart, which visualizes the distribution of three route choices over departure time in a day. The color legend is consistent with the other two views. Stacking multiple routes in a superposition layout makes it intuitive to locate the dramatic change of trajectory volume with certain route choice, which probably indicates a potential impact (**CI**). For example, trajectories choosing the orange route dramatically increase on 9 oclock, compared to other time in the day (Figure 10). Based on stacked bar chart, several interactions are developed. The number of trajectories is visible when hovering on a certain bar. Bars of the same route are shifted to horizontal axis when a certain route is selected.

With hypotheses raised by visual exploration in stacked bar chart, the factor configuration panel supports to customize ordinal factors into categories, serving as the input of MNL model (**CII**). In the left-bottom part of Figure 10, a gray band marks the range of a category, which is labelled by an index. As discussed in Section 5.3.1, by changing the number of categories C and value range of certain category $[r_{c-1}, r_c]$, users can investigate different hypotheses. As Figure 11 shows, several interactions in the panel enable users to directly manipulate the C and $[r_{c-1}, r_c]$. Similar to the interaction design in trajectory filtering, hovering in different regions invokes the different functions. Hovering on the gray band highlights current category. Hovering near the left or right boundary of the band invokes the range adjusting function. By clicking and dragging, users are able to change the $[r_{c-1}, r_c]$ of current category. Hovering on the central index, a menu is popped out to provide editing options of C. Users can add, delete or merge the categories. For example, by clicking the '+', a new category is created.

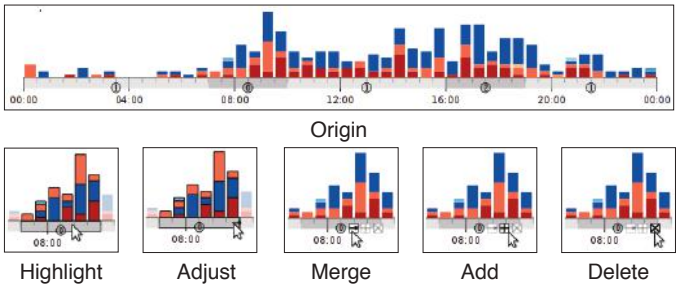

Fig. 11. Interactions to Configure Factors: hovering on different regions invoke different functions, e.g., adding a new category.

After factor configuration, the MNL model is used for statistical evaluation. Results are returned as a coefficient matrix. To keep the results precise as well as intuitive (**CIII**), we visualize them right in the matrix. As the right of Figure 10 shows, a matrix displayed aside the stacked bar chart, which visually encodes the coefficient outputs of MNL in Figure 9. The blue route with maximum trajectory number selected as base option is visualized at the top-right corner. Indices of categories are marked on the left of each row. For each cell in the matrix, the coefficient is directly printed to preserve the precision of result. Those cells with significant impact (95% certain) are highlighted in corresponding route colors to make it more distinct from others, which are drawn in dashed frame.

6 Implementation and Performance

A prototype system is developed to verify the effectiveness of our method. In this section, we first introduce the experiment dataset and implementation detail. Then we report the system's performance.

6.1 Input Data

We take the GPS dataset recorded in Beijing as the experiment data. The data is collected from 28,519 taxis in 24 days, from March 2^{nd} to 25^{th}, 2009. The data size is 34.5 GB in total and consists of 379,107,927 sampling points. The sampling rate is every 30 seconds. Each sampling point contains the following attributes: *time*, *latitude*, *longitude*, *speedmagnitude*, *direction* as well as a boolean *CarryPassengerState*. *CarryPassengerState* is a tag indicating whether the taxi carries passengers or not. In this work, we only use the trajectories with passengers, each of which can be identified by ID.

Besides the taxi GPS dataset, the road network data is collected from OpenStreetMap's jXAPI [49]. Following an existing paper [19], trajectories are cleaned and matched to the road network in the data preprocessing step. The final data size is 12.1 GB.

6.2 Implementation

The system is mainly written in C++, with Qt framework. The rendering is performed with both OpenGL and Qt GraphicsView framework. A third-party library Graphviz [56] is used to do the topological graph layout in route visualization. A MatLab [57] extension is integrated in the system to perform the MNL analysis.

As introduced in Section 4, trajectory filtering supports to narrow the scope of analysis down to trajectories related to certain *OoI*/*DoI* pair, which are fed into the extraction of multiple routes and the further visual analysis. Several strategies are adopted to facilitate the filtering.

In the preprocessing stage, trajectories are indexed by a spatial quadtree, which divides the 2D spatial region recursively and adaptively based on the distribution of trajectories' sampling points. Each quadtree node stores the IDs of trajectories that intersect it. In the run-time stage, a filtering operation is conducted in three steps: in the coarse filter step, the system fetches trajectories from the quadtree nodes where the circular filter locates; in the intermediate filter step, the top N trajectories (e.g., N = 100 in this work) which satisfy the filtering constraints are returned and rendered; in the fine filter step, trajectories satisfying the filtering constraints in the whole dataset are filtered.

To ensure interactive filtering, the fine filtering step is not performed during dynamic filtering. For example, during the procedure of moving or resizing the circular filter, only the top N trajectories are returned and rendered. Once the filter is settled down, the fine filtering step is performed. Meanwhile, when multiple filters are applied, the filtering is conducted based on the previous filtered trajectories recursively, where the query space is much smaller than the whole data set.

6.3 Performance

The system's performance is tested on a Dell T3400 workstation with an Intel(R) Core(TM)2 2.66 GHz, 4 GB RAM and a NVIDIA Geforce GTX 470 GPU. The performance of filtering serves as the basis of the system so that it is crucial for the overall performance.

As the basis for more complex filtering, performance of the first filter is essential. The filtering performance under different days are tested and Figure 12 gives the result. Filtering are sampled at seven different locations to alleviate the bias caused by spatial locations. As Figure 12 shows, the time cost of the first filter scales well over 24 days in the experiment dataset. The average time cost of filtering trajectories per day is around 0.76s. Variation of time cost arises among different locations. The extreme outlier of time cost is a filter in a busy traffic area, almost 400 trajectories filtered per day. Notice that Figure 12 gives the time cost when the first filtering is settled down. During users' dynamic filtering (introduced in Section 6.2), top 100 trajectories can be returned immediately in the system. Users can set filtering smoothly independent from the temporal range. When filtering settled down, time cost is inevitable as the time range increases.

However, the system handles the latency in predictable response time.

Further filtering by multiple filters or advanced filtering setting is conducted based on the result by the first filter. Because filtering space is greatly narrowed down as well as trajectories are loaded in memory, this step costs greatly less. Averagely, it costs less than 1s to filtering in hundreds of trajectories.

For the multiple route extraction, the computational complexity is $O(n^3)$. Given hundreds of trajectories, it averagely costs no more than 5s to extract the routes. Taking Case 2 (Section 7.2) as an example, it takes around 7s to filter trajectories in 7 days and 3.6s to extract the routes.

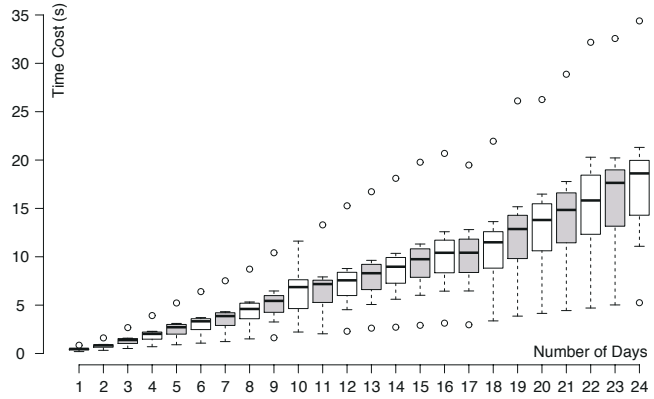

Fig. 12. Performance of the First Filtering over Different Days: it costs averagely 0.76s to filter trajectories per day, which scales well as the number of days increases.

7 Case Study

In this section, with the Beijing GPS dataset (introduced in Section 6.1), we report three cases to demonstrate how the visual interface facilitates the exploration of multiple route choice behaviour.

7.1 Case I: Overview of Multiple Route Choices

In this case, we demonstrate some examples to explore multiple routes in Beijing (**T1**), in which trajectories are filtered in the whole 24 days. Beijing adopts the ring and radial highway system. The ring roads (ranked as 1^{st}, 2^{nd}, etc.) provide rapid access around the city, while the radial highways provide rapid access between ring roads. Basically, there are two types of travelling. One is made along the ring road and the other is travelling between different ring roads. Multiple route choices containing these two different travellings are explored in this case. As the Figure 13(a)(b) shows, we set the *OoI* and *DoI* filters on different ring roads. To be specific, the ring roads are shown horizontal here, while the radial ones are shown vertical. In Figure 13(a), the *OoI* is on the 4^{th} ring and the *DoI* is on the 2^{nd} ring road respectively. Similarly, in Figure 13(b), the *OoI* is set at a business district on the 3^{rd} and the *DoI* is set at the transportation hub on the 2^{nd}. For both two examples, lots of multiple routes are extracted. Besides routes choosing the main ring road to travel, there are many alternative routes travelling in the radial byways connecting the ring roads. In Figure 13(c), *OoI* and *DoI* are both set on the 2^{nd} ring road, but at two transportation hubs respectively.

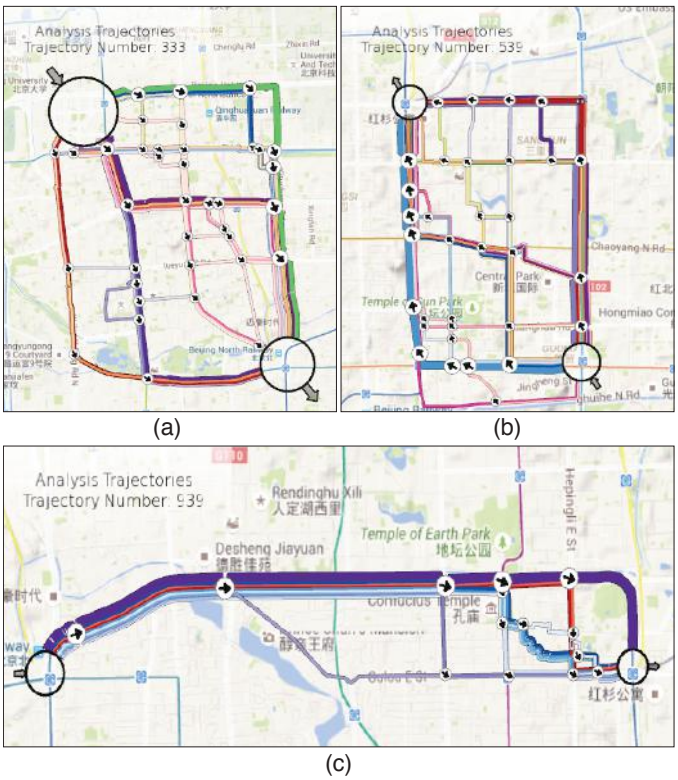

Fig. 13. Case Study #1 Overview of Multiple Route Choices: (a) multiple routes from the 4^{th} to 2^{nd} ring roads (b) multiple routes from 3^{rd} to 2^{nd} ring roads. (c) multiple routes when travelling among the same 2^{nd} ring road.

The route choices between these two places are much less than the above two. The majority of drivers choose to follow the main ring road without branching. However, a few drivers make different choices at some road segments.

7.2 Case II: Exploring the Route-related Factors of Multiple Routes

In this case, we demonstrate how the system supports to explore the route-related factors of multiple route choices between interested regions (**T2**). With the route-related factor view, users are able to compare different route-related factors among multiple routes, and study how those factors impact the route choices.

As Figure 14(a) shows, the *OoI* and *DoI* is set at Beijing Airport and a central business district in down-town area respectively. With several filters with exclusion constraint, a few dirty trajectories caused by the error or misreport by GPS devices are filtered out. After that, there are 192 trajectories travelling from Beijing Airport to the business center, from March 2 to March 8 2009.

Figure 14(a) shows all the feasible routes between these two regions extracted from the filtered trajectories. Routes are categorised into two groups according to topological similarity. Figure 14(c) gives the color legend of the two groups. The group of routes in warm color chooses the upper highway, while the group in cold color chooses the bottom one. For both groups, there is one or two mainstream route choices with dominant popularity, and several alternative choices with small amount of trajectories. Those alternative routes leave the main choice to other seldom chosen roads somewhere. To obtain a general understanding of

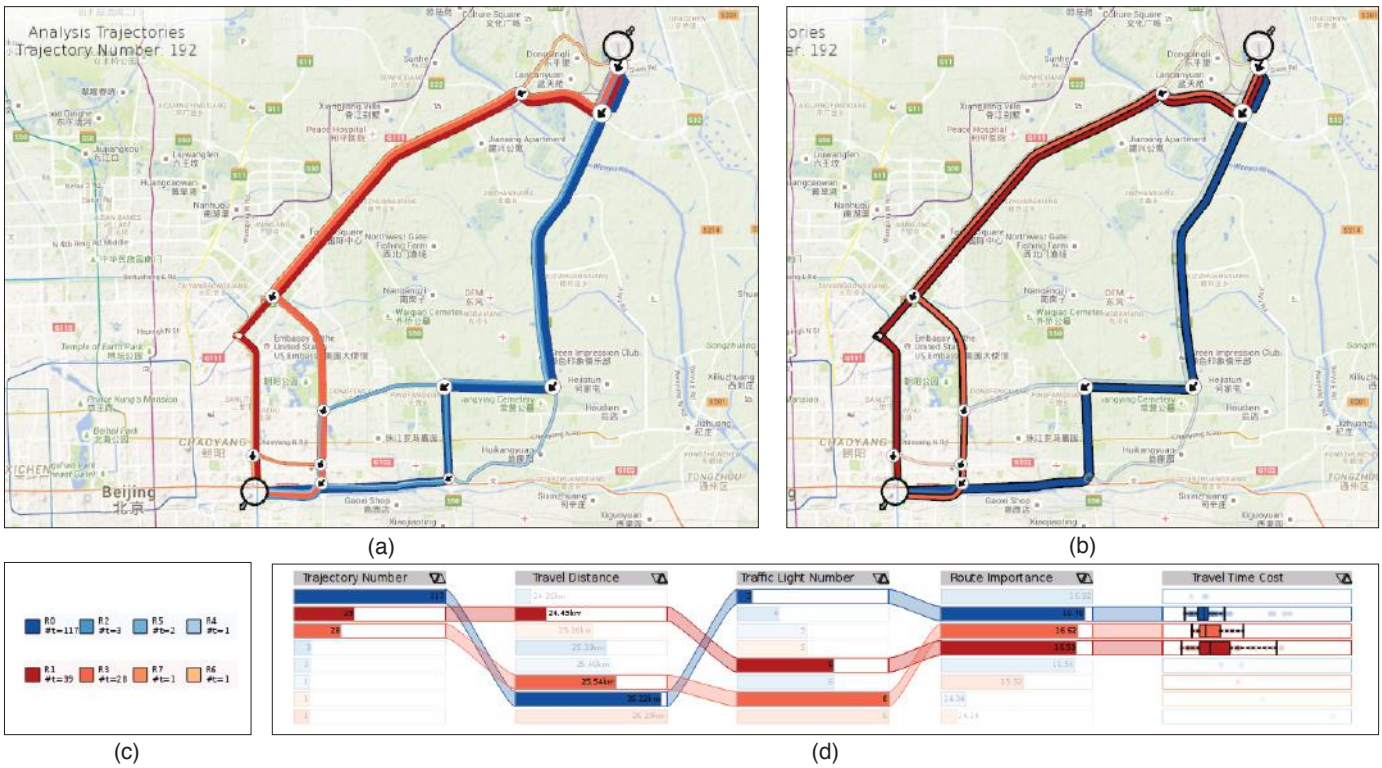

Fig. 14. Case Study #2 Exploring Route-related Factors of Multiple Routes: exploring route choices from Airport to a commercial district: (a) the spatial overview. (b) three selected popular routes. (c) route-related factors' comparison among the selected routes

the route choices, the three most passed routes are selected. As Figure 14(b) shows, the three routes share some common roads in the beginning but soon the blue one splits from the other two. Subsequently, after travelling some distance, the red and orange routes head into different directions respectively. Finally, the blue and orange routes meet with each other. Among the three route choices, nearly 60% taxi drivers choose the blue one, 20% choose the orange one and nearly 15% choose the red one. The left 5% choose the other less travelled routes.

By ranking factors in the route-related factor view (Figure 14(d)), the three selected routes are compared. Although the blue route has the longest travel distance, it has the least traffic light number and highest route importance. The blue route is the one occupying the largest ratio of highway roads comparing to the red and orange routes. Ranking the factor 'Travel Time Cost' in the ascending average time cost order, we can see that the blue one has the smallest average value. The second comes to the orange one and finally the red one. Meanwhile, the blue one gets the smallest variance, i.e., time cost of the blue route is more reliable and predictable because of its small variability. Although their advantage in travel distance, time cost distribution of the red and orange routes are more stretched, which may be caused by the bad traffic when travelling on the ring roads in down-town. Hence, in this case, it is observable that when travelling from Airport to the business district, most of drivers would like to choose less and more predictable time cost rather than short travel distance.

7.3 Case III: Exploring the Trajectory-related Factors' Impact on Multiple Route Choices

In this case we show how the system helps to analyze the trajectory-related factors' impact on route choices, from hypothesis construction (**T3**) to statistical verification (**T4**).

Travelling between regions at different ring roads is very common in Beijing. Taking it as an example, we place the *OoI* filter on the 3^{rd} north ring road and the *DoI* filter at the 4^{th} north ring road. 296 trajectories travelling through are filtered from March 2 to March 8, 2009. Figure 15(a) shows that the top three most popular routes, i.e., the blue, red and orange. The blue one has much larger popularity than the other two. Their route-related factors are ranked in the ranking view (Figure 15(d)). Overall, the top three routes have advantages over those seldom travelled routes in static route-related factors, i.e., the route distance, traffic light number as well as the route importance. Furthermore, the blue route ranks top within the three routes. In the column of travel time cost, the ranking order by median time cost is consistent with that of route's popularity. That is, the more chosen route has the smallest average time cost. This indicates that drivers tend to choose the route with less time cost.

To explore the trajectory-related factors' impact on choice among the three major routes, their distributions over factors are visualized in Figure 15(b). In this view, it can be obviously observed that the population of the orange route increases dramatically around 9 o'clock in the morning when comparing to other time in a day. Similarly, the blue route gains a lot of traffic volume during early evening, at around 18:00. These observations give a rise to the hypothesis that drivers may have larger probability to choose the orange route in the morning, and the blue route in the evening. Based on the hypothesis, we configure the factor, i.e., the departure time in a day, into three corresponding categories, 7:00 - 10:00, 16:00 - 19:00 and the remainder. Together with default configuration of the other two factors, we run MNL to verify the hypothesis. The output of route choice model is given in factor

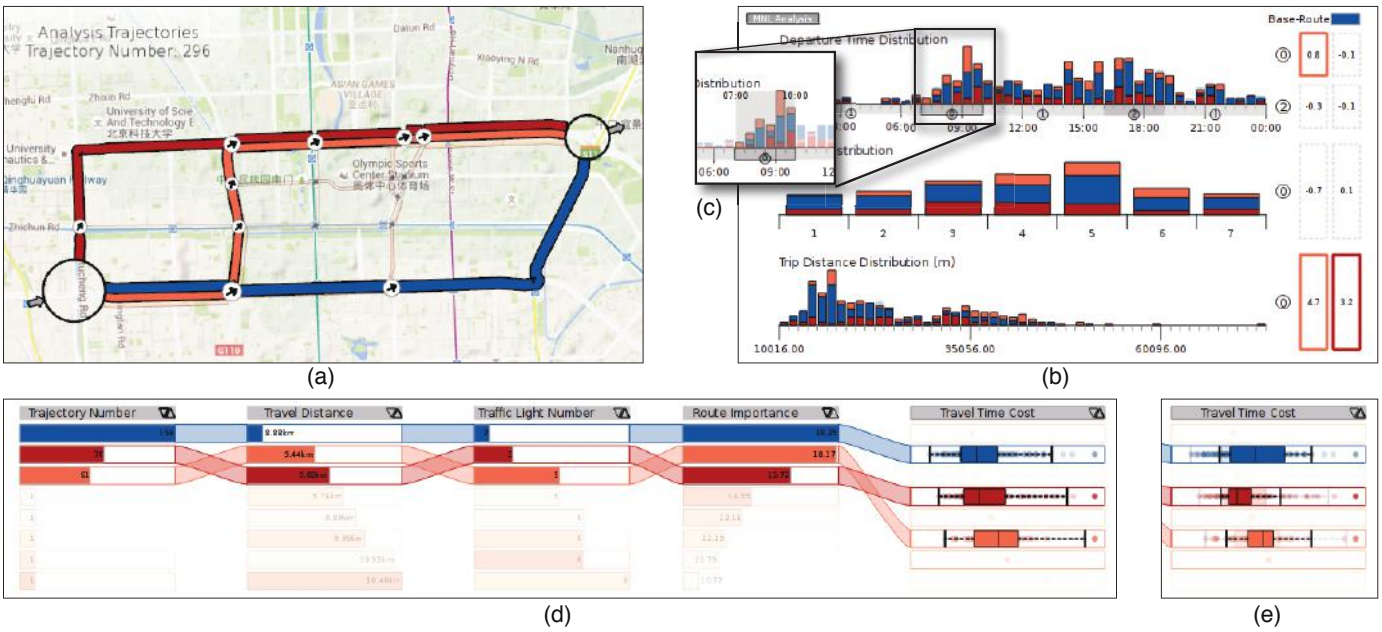

Fig. 15. Case Study #3 Exploring Trajectory-related Factors' Impact on Multiple Route Choices: setting the *OoI* filter on the 3^{rd} north ring road and the *DoI* filter at the 4^{th} north ring road, (a) route spatial view of three selected major routes. (b) trajectory-related factor view. (d) route-related view with selected three routes; (c) selected departure time range; (e) travel time cost distribution of the three selected routes during departure time range in (c).

matrix in Figure 15(b). Taking the blue route as the base route, the orange rectangle indicates that departing in the morning (7:00 - 10:00) significantly improves the odds of choosing orange route than the blue one. That is, when travelling in the morning peak, drivers have larger probability to choose orange route than the blue one. As Figure 15(c) shows, by selecting the trajectories travelling during this time period, the distribution of time cost in the route-related factor view is updated as Figure 15(e) shows. During this period of time, the average time cost and its variance of blue route increases. Comparing with the blue one, although the average time cost of the red and orange are larger than the blue one, their time costs are more reliable during this period. It might be the reason why drivers give up the blue and choose those two. Another interesting point shown by the model is that the odds for both the red and orange routes is tested increasing as the trajectory's length increases by the model. It can be explained by either the difference of route length or the impact of trajectory's length, which needs to be further explored.

8 Discussion

So far we have clarified the data background and tasks of route choice analysis. We have also introduced the visual analytic system, which allows users to interactively explore route choice behaviour with real taxi GPS data. In this section, we discuss the limitations and further improvements.

First of all, compared to other research based on experimental data, the scope of performing route choice analysis based on general GPS data is different. It is mainly defined by the factors which can be derived from uncustomed general GPS data. As discussed in Section 3.1, we propose two types of factors (i.e., the route-related factor and trajectory-related factor) from the core properties of general GPS dataset. Basically, the trajectory-related factors are derived from the spatial and temporal properties of GPS data, which are universal in other movement dataset. And the route-related factors are derived from the route properties, which are independent from what GPS dataset are used. Hence, the scope of using general GPS dataset to study route choice behaviour are explored at its minimum to ensure the universality in this work. However, it is not limited to this minimum scope. Our method can be easily extended to include more properties with richer semantics.

For example, speed is a common attribute in GPS data. We take speed as the inverse attribute to time cost, which has been already considered in this work. More advanced information, such as traffic jams, can be derived from speed and serves as a new factor, which is interesting for future work. Another example is that some taxi GPS dataset carries the property of cab fare for each journey, which also acts as a potential trajectory-related factor. New factors can be integrated in the trajectory-related as the three trajectory-related factors do in this work. Similarly, if there are more route properties, e.g., the score of landscape along the route, they can be appended to the ranking-based visualization as new factor columns. One of the future work is to make it more flexible to plug in new factors, which can use either XML/JSON or visual language for the factor configuration.

Except the properties derived from trajectories or routes, social events (e.g., road construction, concert event, etc.) which possible influence drivers' route choices are not included in this work. Fusing our system with other datasets from different sources such as social media will be potentially interesting. Also, the impact of subjective factors are not included in our work. For example, the impact of the drivers' travelling experience on route choice is out of the scope because it is unavailable in the general dataset. If there is this kind of drivers' profile data, it would be interesting to fuse the information of drivers into our system. And then, how to protect the privacy may be one of the critical challenges.

Another problem is about the route choice model introduced in Section 5.3.1. One limitation is that the preference is only

compared between other routes and the base routes. Currently, it is quite hard for users to compare two arbitrary route choices. In the next step, instead of preference over a basic option, we plan to provide the preference analysis among several route choices. Currently, our system integrates the MNL model. Although it is one of the most widely used choice analysis models, there are some other discrete choice modelling methods that can be embedded, such as the Mixed Logit. In our system, the modelling computations are loosely plugged by a flexible Matlab Engine [57], which is easy to be replaced if necessary.

9 Conclusion

In this paper, we explore the possibility of studying route choice behaviour based on taxi GPS trajectories. Compared to classical route choice analysis method, our general GPS based solution covers larger temporal-spatial range as well as larger number of samples. In this work, we list the factors that can be derived from trajectories, which defines the boundary of this general GPS data based solution. With this, we present a visual analytic system which supports tasks from route choice overview to verify factors' impact on route choice. The system's visualizations and interactions are designed carefully according to task-oriented considerations. The system allows interactive visual exploration in massive trajectories and factors exploration with route choice model. With Beijing taxi GPS trajectory dataset, we demonstrate three case studies to show the system's effectiveness.

In the future, we would like to apply the system to more datasets. For example, applying to trajectory datasets in different areas, we probably are able to compare the route choice behaviour of drivers over different regions. Meanwhile, we would like to improve and extend our system regarding the current limitations. Besides what is discussed in Section 8, there are two possible research directions. Considering that the input factors are fixed, we will improve the system to support the creation of factors. For example, OD distribution can be one of the possible trajectory-related factors. Another interest point is to extend to the system with route advisory function. By taking the analysis of route choice, it is possible to recommend routes by taking different factors into consideration and measure the fitness of route.

Acknowledgments

The authors would like to thank Dr. Zhang Wen for his help on the route choice model and Dr. Wang Zuchao for his help on writing. The authors wish to thank the anonymous reviewers for their valuable comments. This work is supported by NSFC No. 61170204. This work is also partially supported by NSFC Key Project No. 61232012 and the National Program on Key Basic Research Project (973 Program) No. 2015CB352503. This work is also funded by PKU-Qihoo Joint Data Visual Analytics Research Center.

References

[1] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 38, 2014.

[2] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing, "Discovering spatio-temporal causal interactions in traffic data streams," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1010–1018.

[3] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 186–194.

[4] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-Drive: enhancing driving directions with taxi drivers' intelligence," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 220–232, 2013.

[5] W. Adamowicz, P. Boxall, M. Williams, and J. Louviere, "Stated preference approaches for measuring passive use values: choice experiments and contingent valuation," *American journal of agricultural economics*, vol. 80, no. 1, pp. 64–75, 1998.

[6] E. Cascetta, A. Nuzzolo, F. Russo, and A. Vitetta, "A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks," in *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, 1996, pp. 697–711.

[7] V. Henn, "Fuzzy route choice model for traffic assignment," *Fuzzy Sets and Systems*, vol. 116, no. 1, pp. 77–101, 2000.

[8] H. Li, R. Guensler, and J. Ogle, "Analysis of morning commute route choice patterns using global positioning system-based vehicle activity data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1926.1, pp. 162–170, 2005.

[9] A. Vacca and I. Meloni, "Understanding route switch behavior: An analysis using gps based data," *Transportation Research Procedia*, vol. 5, pp. 56–65, 2015.

[10] J. J. Thomas and K. A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005.

[11] Y. Zheng, "Trajectory data mining: An overview," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, pp. 29:1–29:41, May 2015.

[12] F. Mannering, S.-G. Kim, W. Barfield, and L. Ng, "Statistical analysis of commuters' route, mode, and departure time flexibility," *Transportation Research Part C: Emerging Technologies*, vol. 2, no. 1, pp. 35–47, 1994.

[13] A. J. Khattak, J. L. Schofer, and F. S. Koppelman, "Effect of traffic information on commuters' propensity to change route and departure time," *Journal of Advanced Transportation*, vol. 29, no. 2, pp. 193–212, 1995.

[14] C. Xu, W. Wang, Z. Li, and C. Yang, "Comparative study on drivers' route choice response to travel information at different departure time," *2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR)*, vol. 3, pp. 97–100, 2010.

[15] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2013, pp. 344–353.

[16] G. Andrienko, N. Andrienko, P. Bak, D. Keim, S. Kisilevich, and S. Wrobel, "A conceptual framework and taxonomy of techniques for analyzing movement," *Journal of Visual Languages & Computing*, vol. 22, no. 3, pp. 213 – 232, 2011.

[17] 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*, vol. 7, no. 3, pp. 173–180, 2008.

[18] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 330–339.

[19] 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*, vol. 19, no. 12, pp. 2159–2168, 2013.

[20] D. Guo, "Flow mapping and multivariate visualization of large spatial interaction data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1041–1048, 2009.

[21] N. Adrienko and G. Adrienko, "Spatial generalization and aggregation of massive movement data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 2, pp. 205–219, 2011.

[22] M. Lu, J. Liang, Z. Wang, and X. Yuan, "Exploring od patterns of interested region based on taxi trajectories," *Journal of Visualization*, vol. 19, no. 4, pp. 811–821, 2016.

[23] 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 IEEE Pacific Visualization Symposium (PacificVis)*, 2011, pp. 163–170.

[24] W. Zeng, C.-W. Fu, S. M. Arisona, and H. Qu, "Visualizing interchange patterns in massive movement data," *Computer Graphics Forum*, vol. 32, pp. 271–280, 2013.

[25] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: an information aesthetic for comparative genomics," *Genome research*, vol. 19, no. 9, pp. 1639–1645, 2009.
[26] H. Liu, Y. Gao, L. Lu, S. Liu, H. Qu, and L. M. Ni, "Visual analysis of route diversity," in *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2011, pp. 171–180.
[27] W. Zeng, C.-W. Fu, S. Arisona, A. Erath, and H. Qu, "Visualizing mobility of public transportation system," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1833–1842, 2014.
[28] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel, *Visual analytics of movement*. Springer, 2013.
[29] M. Hadjieleftheriou, G. Kollios, P. Bakalov, and V. J. Tsotras, "Complex spatio-temporal pattern queries," in *Proceedings of the 31st international conference on Very large data bases*, 2005, pp. 877–888.
[30] M. R. Vieira, P. Bakalov, and V. J. Tsotras, "Querying trajectories using flexible patterns," in *Proceedings of the 13th International Conference on Extending Database Technology*, ser. EDBT '10, 2010, pp. 406–417.
[31] 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 cab trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2149–2158, 2013.
[32] R. Krüger, D. Thom, M. Wörner, H. Bosch, and T. Ertl, "TrajectoryLenses - a set-based filtering and exploration technique for long-term trajectory data," *Computer Graphics Forum*, vol. 32, no. 3, pp. 451–460, 2013.
[33] M. O. Ward, G. Grinstein, and D. Keim, *Interactive data visualization: foundations, techniques, and applications*. CRC Press, 2010.
[34] A. Inselberg, "Multidimensional detective," in *Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97)*, 1997, pp. 100–107.
[35] E. R. Tufte, "Envisioning information." *Optometry & Vision Science*, vol. 68, no. 4, pp. 322–324, 1991.
[36] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, "LineUp: Visual analysis of multi-attribute rankings," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2277–2286, 2013.
[37] M. Lu, Z. Wang, and X. Yuan, "TrajRank: Exploring travel behaviour on a route by trajectory ranking," in *2015 IEEE Pacific Visualization Symposium (PacificVis)*, 2015, pp. 311–318.
[38] M. Batty, "Rank clocks," *Nature*, vol. 444, pp. 592–596, 2006.
[39] C. Shi, W. Cui, S. Liu, P. Xu, W. Chen, and H. Qu, "RankExplorer: Visualization of ranking changes in large time series data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2669–2678, 2012.
[40] S. Havre, E. G. Hetzler, P. Whitney, and L. T. Nowell, "ThemeRiver: Visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 9–20, 2002.
[41] R. Eccles, T. Kapler, R. Harper, and W. Wright, "Stories in geotime," *Information Visualization*, vol. 7, no. 1, pp. 3–17, 2008.
[42] T. Kapler and W. Wright, "Geotime information visualization," *Information Visualization*, vol. 4, no. 2, pp. 136–146, 2005.
[43] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko, "Stacking-based visualization of trajectory attribute data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2565–2574, 2012.
[44] P. Karnick, D. Cline, S. Jeschke, A. Razdan, and P. Wonka, "Route visualization using detail lenses," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 2, pp. 235–247, 2010.
[45] M. Agrawala, "Visualizing route maps," Ph.D. dissertation, Stanford University, 2002.
[46] G. Sun, Y. Liu, W. Wu, R. Liang, and H. Qu, "Embedding temporal display into maps for occlusion-free visualization of spatio-temporal data," in *Proceedings of IEEE Pacific Visualization Symposium*, 2014, pp. 185–192.
[47] K. Wongsuphasawat and D. Gotz, "Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2659–2668, 2012.
[48] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman, "LifeFlow: visualizing an overview of event sequences," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 1747–1756.
[49] I. Dees, "OpenStreetMap jXAPI," http://wiki.openstreetmap.org/wiki/Xapi.
[50] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2376–2385, 2013.
[51] T. A. Domencich and D. McFadden, "Urban travel demand-a behavioral analysis," *American Economic Association*, 1975.
[52] G. Navarro, "A guided tour to approximate string matching," *ACM computing surveys (CSUR)*, vol. 33, no. 1, pp. 31–88, 2001.
[53] M. Harrower and C. A. Brewer, "Colorbrewer.org: An online tool for selecting colour schemes for maps," *The Cartographic Journal*, vol. 40, pp. 27–37, 2003.
[54] J. de Dios Ortúzar and L. G. Willumsen, *Modelling transport*. John Wiley & Sons, 2011.
[55] W. Javed, B. McDonnel, and N. Elmqvist, "Graphical perception of multiple time series," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 927–934, 2010.
[56] E. R. Gansner and S. C. North, "An open graph visualization system and its applications to software engineering," *Software Practice and Experience*, vol. 30, no. 11, pp. 1203–1233, 2000.
[57] MATLAB, *R2013a*. The MathWorks Inc., 2010.

Min Lu received BS degree in Computer Science from Beijing Normal University in 2011. She is a senior PhD candidate on computer science at school of EECS, Peking University. Her major research interests include visualization and visual analytics of movement data, especially urban GPS trajectory.

Chufan Lai received his BS degree in Electronic Engineering from University of Science and Technology of China in 2013. Since then, he started pursuing a PhD degree on computer science at school of EECS, Peking University. His major research interests include high-dimensional data visualization and visual analytics of multivariate data.

Tangzhi Ye received BS degree in Computer Science from Shanghai Jiao Tong University in 2014. He has been a master student on computer science at school of EECS, Peking University since fall 2014. His major research interests include visualization and visual analytics of traffic data.

Jie Liang has a joint appointment with full-time academic in the faculty of engineering and IT, University of Technology Sydney and an honorary researcher at Qihoo-PKU joint Lab, Peking University. She received the university medal and bachelor of science with first class honours from the University of Technology Sydney, obtained a master of business information systems from the University of Sydney and was a recipient of CMCRC Honours scholarship and Australian Postgraduate Awards, with a Doctorate in Data Visual Analytics from the University of Technology Sydney.

Xiaoru Yuan received BS degree in computer science and BA degree in law from Peking University in 1997 and 1998 respectively. In 2005 and 2006, he received MS degree in computer engineering and PhD degree in computer science at University of Minnesota at Twin Cities. He is now a professor at Peking University, in the Laboratory of Machine Perception (MOE). His primary research interests lie in the field of scientific visualization, information visualization and visual analytics with emphasis on large data visualization, high dimensional data visualization, graph visualization and novel visualization user interface.