Visual Analysis of Route Choice Behaviour based on GPS Trajectories

Min Lu
Peking University
Beijing, China
lumin.vis@gmail.com

Chufan Lai
Peking University
Beijing, China
chufan.lai.1990@gmail.com

Tangzhi Ye
Peking University
Beijing, China
yetangzhi66@gmail.com

Jie Liang
Peking University
Beijing, China
christy.jie@gmail.com

Xiaoru Yuan
Peking University
Beijing, China
xiaoru.yuan@gmail.com

ABSTRACT

There are often multiple routes between regions. Many factors potentially affect driver's route choice, such as expected time cost, length etc. In this paper, based on taxi GPS trajectory data, we propose a visual analysis system to explore driver's route choice behaviour among multiple routes, i.e., how it is influenced by factors. With interactive trajectory filtering, the system constructs real feasible routes between regions of interest. Three complementary visualizations are designed to explore different routes and potential factors' impact on route choice behaviour. Applying to real trajectory dataset, the effectiveness of the system is demonstrated by two cases.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Visual Analytics

General Terms

Visual Analytics

Keywords

Multiple Route Choice, Visual Analytics, Visualization

1. INTRODUCTION

As modern traffic road network develops, there are often multiple routes to choose from when travelling from one place to another. Drivers may make different route choices in different considerations. For example, expected time cost is one of the most dominant factors. Some web mapping services, e.g. Google Map [1], offer route planning which mainly considers travel manner and expected time cost. However, there are some other factors that potentially influence route decision making, such as the number of traffic light, travelling comfortableness etc. Various factors interact with each other so that it is not straight-forward to figure out what drives route decision making.

Recently, many efforts have been made to study how driver selects route with different factors i.e., route choice behaviour. Classically, research is performed based on Stated Preference (SP) survey data [2]. SP collects route preferences in hypothetical situations from respondents. Different choice considerations can be directly measured by the information in questionnaires. With SP data, various route choice models [7] [18] are developed, trying to capture impacts of different factors on the route choice behaviour. However, such investigations are limited in range and the surveys need to be carefully designed. Also, what obtained from investigation is not practically reliable enough. In recent years, some researchers perform the analysis with the help of Global Positioning System (GPS). Compared to traditional investigations, GPS receivers are used to collect trajectories of volunteers, which takes less effort and is more realistic. But such pilot studies are often conducted among a limited number of users in restrained spatio-temporal scale, like, only collecting morning commute trips [22] [29].

In this work, rather than experiment dataset, we explore the possibility of studying route choice behaviour based on more general GPS trajectory data, specifically, taxi GPS trajectories. Relieved from customizing, taxi GPS trajectory is more general. However, at the same time, there are two main challenges introduced by using general GPS trajectory dataset:

- *Extract relevant trajectories in the context of multiple routes:* Unlike the experimental GPS trajectories which are constrained in certain spatial and temporal range, how to extract trajectories travelling through multiple routes of interest from massive trajectories needs to be tackled.
- *Raise hypotheses on factors that significantly influence the route choice behaviour based on taxi trajectories:* Different from data collecting from hypothesis-oriented experiments, how to indicate the impact of factors on route choice to facilitate hypothesis establishment is crucial.

From the angle of visual analytics field [28], we propose a visual analytics system by leveraging human interaction and judgement in the analysis process to tackle the above challenges: with a suite of graphical filters, trajectories travelling 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

UrbComp'15, August 10, 2015, Sydney, Australia
.

hypotheses on potential influence; then the hypotheses are further verified by the statistical model to draw reliable conclusions.

2. RELATED WORK

In this section, we have a discussion of related work on route choice behaviour analysis in transportation field and then report the research progress in visual analytics using trajectory data set.

2.1 Route Choice Behaviour Analysis

Route choice behaviour has been widely studied in the transportation analysis 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. [25] find that 26% people do not always use the same route. Concerning the reason, Khattak et al. [19] study 700 commute trips from 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. [32] 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. [22] 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, Alessandro et al. [29] study route switch behaviours between the same OD 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 behaviours, 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). 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

Andrienko et al. [4] present a taxonomy of generic analytic techniques based on possible types of movement data. For trajectories, there are three kinds of explorations [6]: 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 pattens [13], e.g. the traffic jam propagation graph extraction [30]. Aggregation methods visualize movement groups to reveal the high-level movement graph. Guo [15] and Adrienko [3] et al. construct geographical regions and visually aggregate the in-between movements as flows. Besides aggregation between regions, travel behaviour within interchange region are visualized. Guo et al. [16] provides a circular design to explore movement at a road intersection. Zeng et al. [34] derive a visualization from Circos [21] to display interchange traffic flow at subway transition stations. Lu et al. [24] aggregate trajectories along a single route and rank them by their time cost along the road segments, to reveal mainstream and outliers. Liu et al. [23] study the route diversity between locations and provide a clock like radial layout to display temporal statistic distribution. Different from analysing individual trajectories, our method provides analysis based on extracted topology structure. Zeng et al. [33] visualize the mobility of routes starting from a single source in public transportation system and provide comparison among different routes. Similar to their routes' comparison, our work provides comparison among multiple routes.

Alternative to analyse global trajectory as a whole, trajectories of interest can be filtered to perform local analysis. Andrienko et al.'s book [5, Chapter 4.2] summarize different kinds of filtering. In this work we implement fully interactive filters similar to TrajectoryLenses [20]. It allows users to extract trajectories from a common origin and a common destination. The origin and destination are defined by interactive lenses. Ferreira et al.'s system [11] for New York taxi exploration also has a similar design to filter trajectories by their origins and destinations.

3. OVERVIEW

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

3.1 Data Background

In the following, we list down common terminologies in this work to facilitate the discussion, which are illustrated in Figure 1:

- *Trajectory* is a list of positions in the temporal order, recording the movement.
- *Origin/Destination (O/D)* refers to the beginning/ending position of movement.
- *Route* is a sequence of physical roads that vehicles travel through.
- *Origin/Destination of Interest (OoI/DoI)* refers to the interested beginning/ending position of movement.
- *Multiple Routes* are the alternative travelling *Routes* between the same pair of *OoI* and *DoI*.

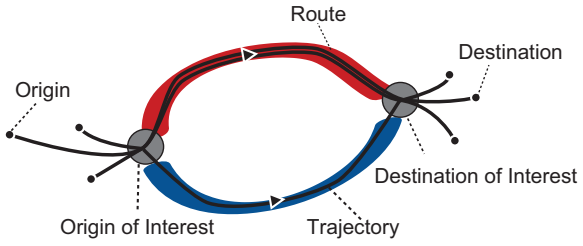


Figure 1: Illustration of Related Traffic Concepts

Note that *OoI/DoI* is not necessarily the *O/D*. In Figure 1, *OoI/DoI* is set at the common position where trajectories begin/end, to study the choice behaviour where multiple routes exist.

Different from designed experiment collecting factors on purpose, factors in our case are derived directly from general GPS dataset. Table 1 summarises the derived factors, which consists of two categories: *route-related factors* and *trajectory-related factors*. For each route, with road network dataset, its traffic related attributes can be derived, e.g. the length of route, the total number of traffic lights along the routes. Those attributes probably play a role in the drivers' route choice. On the other hand, for each trajectory, it has individual difference which drivers probably consider to

Table 1: Table of Derived Factors

Object	Attribute	Description	Motivation
Route	Route Length	The route geographical length	Do drivers prefer shorter route length?
	Traffic Light Number	The total number of traffic light along the route	Do drivers prefer less traffic light number?
	Route Significance	The significant metric based on the road level	Do drivers prefer route with more significant?
	Time Cost Distribution	The potential time cost distribution of a route	Do drivers prefer route with less time cost? Do drivers prefer route whose time cost is with less variation?
Trajectory	Departure Time in a Day	The departure time in the time scale of day	Do drivers departing at different time make different route choices and how?
	Departure Day	The departure day in the time scale of a week	Do drivers departing on different day make different route choices and how?
	Trajectory's Length	The total travel distance of the trip	Do drivers travelling in different length make different route choices and how?

adjust the route choice. For example, drivers travelling between the same pair of *OoI* and *DoI* in peak time and off-peak time probably make different route decisions.

3.2 Analytic Tasks

In this part, we clarify analytic tasks to explore route choice behaviour with general GPS trajectories. Given a pair of *OoI* and *DoI*, firstly, the system gives an overview of the possible routes. Then the route-related factors are explored to acquire an overall understanding of choice preference among route-related factors. And then route choice distribution over trajectory-related factors are studied, to raise hypotheses on factors that potentially impact on the route choice behaviour which are otherwise supposed unchanged if only considering route-related factors. Finally, the system should be capable to examine the proposed hypotheses to tell if the impact is significant. Based on these requirements, the design tasks are summarized as following:

- *Overview of multiple route choices* ***(T1)***: give an overview of the possible feasible route choices between *OoI* and *DoI*.
- *Exploration on 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 on 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 on route choice behaviour* ***(T4)***: build a statistic model to examine the impacts on route choice significant or not.

3.3 System Overview

We propose a visual analytics system integrating visualization, interactions and statistical modelling to support the above tasks. Figure 2 shows the pipeline of the system.

In the preprocessing stage, trajectories are cleaned. To facilitate filtering in massive trajectories, a quad-tree spatial index has been built. In run-time stage, with a suite of graphical filters integrated in the system, trajectories between a pair of *OoI* and *DoI* are filtered. With those filtered trajectories, the feasible multiple routes are extracted by a grid-based algorithm and the topology graph is constructed. Then trajectories and extracted routes are fed as input to visualization and visual analytic module. The module mainly consists of three parts: spatial view gives a geographical overview of extracted multiple routes to show how those routes travel; route-related factor view visualizes the route-related factors in a ranking diagram which supports exploration and comparison among different routes; trajectory-related factor view displays the distributions of route choices over trajectory-related factors to help propose and verify hypothesis. For the trajectory-related view in detail, the distribution of multiple routes over those factors are visualized based on stacked-bar chart, in which great interest arises where dramatic volume change of route choices. Then by interactive hypothesis configuration, an statistical analysis model is customized to verify the hypothesis. After modelling, the results are integrated visually back in the trajectory-related factor view, to tell if the impact is significant.

Among the three views, users are able to correlate the route-related factors and trajectory-related factors by cross-filter interaction strategy [31]. Meanwhile, from trajectory filter to modelling, the visual visual analytics procedure supports iterative exploration.

4. MULTIPLE ROUTES GENERATION

For the massive taxi GPS trajectories, the system integrates graph-

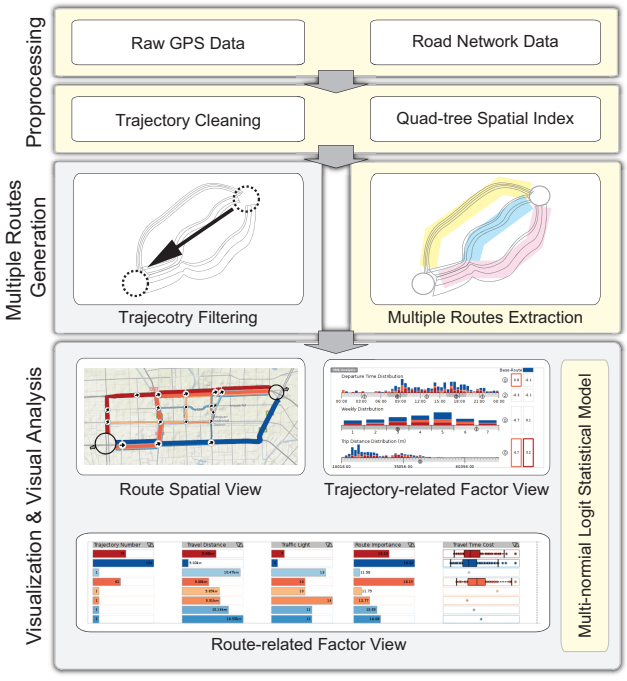


Figure 2: The System's Pipeline: the yellow background indicates the automatic processing and gray background involves human interactions and judgements.

ical filtering which supports 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

The filter model consists of atomic spatial and temporal queries. From the temporal aspect, a two-level temporal filter supports to query from date and time level.

From the spatial aspect, similar to TrajectoryLenses [20], we design circular filter, which defines a circular area in the spatial space to query trajectories with certain constraints. For usage simplicity, parameter tuning is embedded into the circular filter. As Figure 3 shows, when hovering on certain region, certain function is waked and corresponding handle is visible. For example, hovering in the center invokes the moving function. Besides modifications on filter's geometric parameters, six possible location constraints are provided: *origin*, *destination*, *origin/destination*, *passing*, *inclusive*, and *exclusive*. Moreover, for two or more filters, direction can be assigned between filters to select trajectories following certain direction. Complex filtering can be built by combining filters, which cooperate with each other in intersection operation. In this work, the first two filters are detected to define the *OoI* and *DoI* by default.

Figure 3: Interactions on Circular Filter: different functions are invoked by hovering on corresponding regions. Direction between filters is assigned by dragging from one to another.

4.2 Multiple Routes Extraction

With filtered trajectories from *OoI* to *DoI*, we employ an automatic grid-based algorithm to extract multiple routes, regardless of road-network data.

Figure 4 illustrates the process of route extraction. First, a grid is covered within the boundary box of filtered trajectories(Figure 4(b)). Then each trajectory is denoted by the sequence of passing cells (Figure 4(c)). For each cell, we derive its average direction from trajectory segments inside it. Then the directions are further approximated as horizontal or vertical ones (Figure 4(d)). Neighbouring cells with the same directions are merged to avoid ambiguity (Figure 4(e)). After that, routes are formed by linking cells (Figure 4(f)). Cells with more than one in/out degree are detected as the splitting/merging nodes(Figure 4(g)), based on which the multiple route graph is constructed. Finally, multiple routes are encoded visually (Figure 4(h)), which will be introduced in Section 4.2.

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

To provide an overview of multiple routes (**T1**), the route spatial view is designed with following considerations:

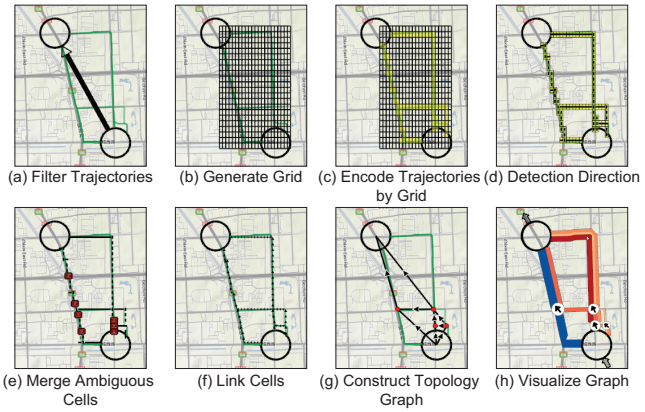


Figure 4: Multiple Route Graph Constructing Process: by covering a grid over trajectories, multiple route graph is built upon travelled cells.

- **Representation of *OoI* and *DoI* (CI)**: to locate the areas where are *OoI* and *DoI*.
- **Visibility of multiple routes (CII)**: to make all the feasible routes between *OoI* and *DoI* visible, including both the popular ones and the seldom travelled ones.
- **Indication of traffic flow directions (CIII)**: to show the travelling directions along routes, especially at the intersections.
- **Summary on routing (CIV)**: to summarize several major route choices by merging similar routes

After interactively manipulating circular filter (discussed in Section 4.1), the *OoI* and *DoI* circular filters are settled down and others are hidden for a clean visual style. To indicate *OoI* and *DoI* filters, inward and outward arrows are attached respectively (**CI**).

The extracted routes are visualized as bands, whose width encodes the number of passing trajectories. A logarithmic mapping is used to enhance the visibility of seldom travelled routes (**CII**). Routes sharing the same road segments are stacked, on which a tooltip is shown when hovering, to facilitate selection (Figure 5(b)). In the tooltip, the currently hovering route with its number of travelled trajectories is highlighted. Hence, users are able to select routes easily by moving up and down over the staked bands, especially the unpopular ones. Considering directions in straight roads are self-evident, we design a glyph to embed the traffic directions at crossings. In the glyph, its size encodes the volume of passing traffic flow and the arrow inside implies the average flow direction (**CIII**).

We summary the routing between a specific *OoI* and *DoI* by mainstream routes with their alternatives (**CIV**). Initially, routes with the top 25% traffic volume are regarded as the mainstreams. Others are assigned to their closest mainstream routes by topology similarity. In our case, we choose the edit distance [27] to measure the topology similarity, which counts the minimum amount of switches required to transform from one sequence to the other by denoting route as a sequence of its crossings. Considering mainstream and its similar alternative routes as a group, qualitative colors [17] are used to differentiate different groups. Within each group, all routes are colored similarly, with the lightness inversely proportional to the route popularity. Additionally, to minimize the

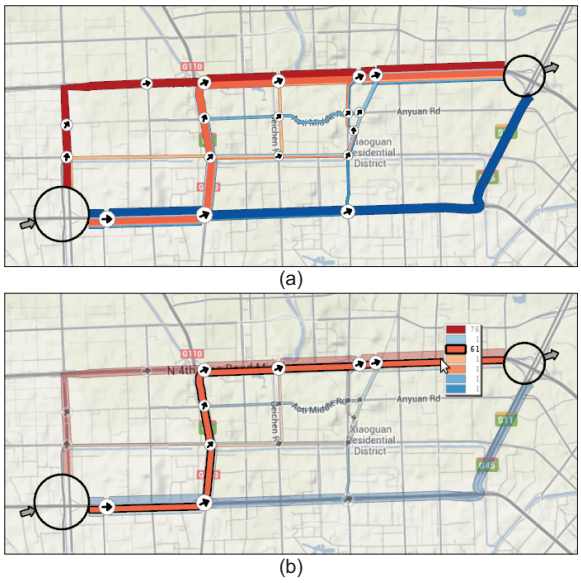


Figure 5: Route Spatial View: (a) geographical overview of multiple routes: the route width encodes the amount of traffic flow. the arrow glyph at each intersection indicates the average flow direction. (b) one highlighted route in the road segment tooltip: all stacked routes are displayed in the tooltip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

mental gap, a topological graph in the node-link diagram (Figure 6(a)) and corresponding color scheme (Figure 6(b)) is given. Note that the color scheme is used globally across all views.

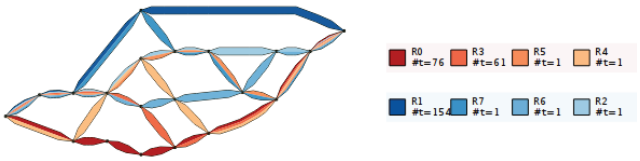


Figure 6: Topology-based Color Scheme: (a) topological graph of multiple routes: all routes are categorized into a few groups based on topology similarities. (b) the global color scheme based on topology-based grouping.

5.2 Route-related Factor View

Inspired by ranking visualizations (e.g., LineUp [14]), we design a ranking-based visualization to support exploration on route-related factors' impact on multiple routes (**T2**). The ranking-based visualization supports users to have an insight into the preference on factors by comparing them. There are several considerations we have taken in the design (for simplicity, factors specifically refer to the route-related factors in this subsection):

- **Accommodation of different factor types (CI)**: to visualize both static and dynamic factors.
- **Comparison of multiple routes among factors (CII)**: to enable comparison among route factors.
- **Exploration on routes in topological relationship(CIII)**: to support exploration on those routes in topological relationship, e.g., similar routes.

Figure 7(a) illustrates the design of route-related factor view. The view mainly contains two parts: route-related factors in the left and topological relationships in the right. In the view, each row depicts a route whose color is consistent with that in route spatial view. Static factor, e.g. route length, is presented by a bar whose horizontal width encodes its value. Dynamic factor, e.g. time cost distribution of the route, is represented by a horizontal-aligned box plot with outliers preserved (**CI**). By default, routes are ranked from top to bottom in decreasing order of the number of passing trajectories. Each factor can be ranked to support comparison (**CII**). Specifically, dynamic factors are ranked by the medians of distributions in our case. Transparent links are used to facilitate visually tracking of routes.

Comparing routes in close topological relationship is especially of great interest. Hence, besides selecting route by directly clicking rows in the view, two more route selection modes are integrated in the right side (**CIII**): hierarchical structure among routes whose root node is the *OoI* and all leaf nodes are the *DoI* to select routes sharing the common road segments in different degrees (named as 'Tree' in the view); similar group structure supports to select a group of similar routes which is discussed in Section 5.1 (named as 'Node-Link' in the view). Figure 8 illustrates the selecting functions of the two modes.

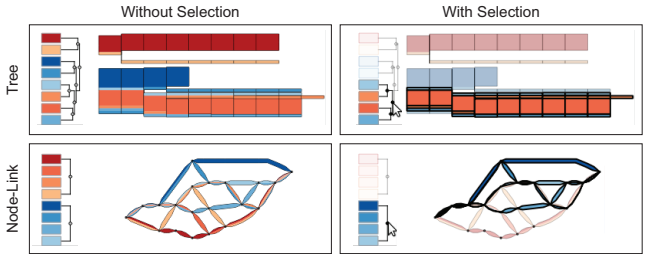


Figure 8: Two Route Selection Modes: by clicking corresponding node, a group of routes in certain topological relationship are selected.

5.3 Trajectory-related Factor View

For each trajectory, it has individual differences from others, i.e. the trajectory-related factors in this work. As discussed in Section 3.1, three trajectory-related factors are derived from general GPS trajectories. In this section, we first introduce the statistical model used to model the impact. Then we present the visualization and interactions that help with hypotheses construction (**T3**) and verification (**T4**). For similarity, the factors refer to the trajectory-related factors in this section.

5.3.1 Multinomial Logit Model

To verify factors' impact on route choice, we adopt the Multinomial Logit (MNL) model [10] which is simple, understandable, and widely used in the transportation area for route choice analysis [8, 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:

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

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, while $\boldsymbol{\beta}_i$ is the coefficient vector of option i, a major output of the

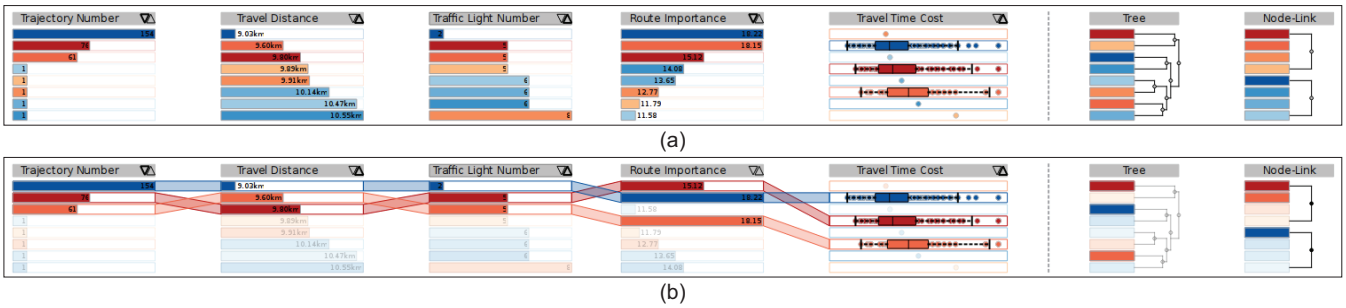


Figure 7: Route-related Factor View: (a) initial view. (b) view after ranking on several factors.

model. From 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 this formula, we can see that the probability varies monotonously with the observable term $\boldsymbol{\beta}_i \boldsymbol{X}_j = \sum_{t=1}^{D} \beta_{it} X_{jt}$. Here, D denotes the number of observed factors, and the coefficient β_{it} depicts the impact of factor t on option i. The model is usually further simplified by setting a basic option, e.g. option m. Choice probabilities are made relative to the base:

$$\ln(P_j(i)) - \ln(P_j(m)) = \ln\left(\frac{e^{\boldsymbol{\beta}_i \boldsymbol{X}_j}}{e^{\boldsymbol{\beta}_m \boldsymbol{X}_j}}\right) = (\boldsymbol{\beta}_i - \boldsymbol{\beta}_m)\boldsymbol{X}_j = \boldsymbol{\beta}_i' \boldsymbol{X}_j \tag{4}$$

The new coefficient shows the impact of factors when choosing between option i and m. Specifically, when β_{it} gets positive, it means that people prefer option i to the base option m when X_{jt} increases. Inversely, when β_{it} gets negative, people prefer the base option m to the option i when X_{jt} increases. The result is only meaningful when it is tested significant. In our context, the route with maximum trajectory number is set as the base option m. We set this because what persuades people to leave the mainstream to another route is always a research interest. The confidence level (p-value) is also derived to verify significance of the impacts. Given the coefficient matrix $\boldsymbol{\beta}_i'$ along with p-values, users are able to validate the influence of each factor on each route choice.

In our case, the potential factors X_{jt} are not all numerical. The departure time in day and week scale are nominal so that there is no intrinsic increasing. For example, Saturday is not increased from Friday. To solve this problem, the nominal factors are allowed to be factorized into numerical sub-factors which take 1 if during the value range and 0 if outside the value range.

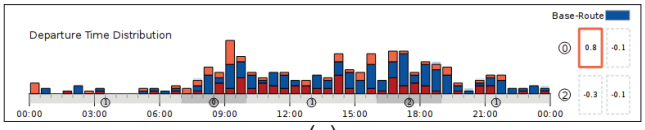
Notice that our system uses but doesn't limit to MNL model. Modelling computations are done by using the flexible Matlab Engine [26], which is highly replaceable by other route choice models.

5.3.2 *Visual Design*

To explore trajectory-related factors' impact of on route choice behaviour, trajectory-related factor view is designed with following considerations:

- **Factor distribution of multiple routes (CI)**: to allow for the comparison of factor distributions when different route choices are made and facilitate raising hypotheses on potential impact.
- **Configuration of statistic model (CII)**: to support users in refining the factors.
- **Impact of factors on route choices (CIII)**: to show credible conclusions about the factor impacts on route choice behaviours.

In Figure 9(a), the factor view is composed of three parts: the stacked bar chart, the factor configuration panel, and the factor impact matrix.



	Orange VS Blue	Red VS Blue
⓪ 7:00~9:00	0.8 *	- 0.1
② 16:00~19:00	- 0.3	- 0.1

* p - value < 0.05

(b)

Figure 9: Trajectory-related Factors View: (a) the stacked bar chart (left-top) visualizes distribution; factor configuration panel (left-bottom) supports factor customization; impact matrix shows the output. (b) the underlying coefficient matrix output of MNL model

Stacked bar chart visualizes the distribution of different route choices over factors, whose color legend is consistent with other 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, in Figure 9(a), the number of trajectories with the route choice in orange color increases dramatically on 9 o'clock.

After raising hypotheses, as discussed in Section 5.3.1, factor configuration panel supports to customize nominal factors into numerical sub-factors, serving as the input of MNL model(**CII**). Gray bands mark the range of sub-factors, each of which is labelled by an index. As Figure 10 shows, several interactions are developed in the panel to modify the configuration. For example, hovering on the band, a menu pops out for basic editing operations, including factor creation, deletion, adjusting and merging.

After factor configuration, the MNL model is used for statistical evaluation. A matrix displayed aside the bar chart (in the right of Figure 9(a)) visually encodes the coefficient outputs of MNL in Figure 9(b). The blue route with maximum trajectory number selected as base option is visualized at the top-right corner. For each

cell in the matrix, the coefficient is directly printed to preserve precision of result. Those cells with significant (95% certain) impacts are highlighted in corresponding route colors.

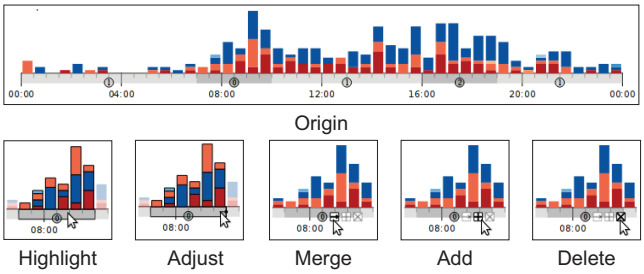


Figure 10: Factor Configuration: operations to customize factors.

6. CASE STUDY

In this section, we first introduce the set-up of system. Based on a prototype system with Beijing taxi GPS trajectories, we report two cases to demonstrate how the visual interface facilitates the exploration of multiple route choice behaviour.

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 2nd to 25th, 2009. The data size is 34.5 GB in total and consists of 379,107,927 sampling points every 30 seconds. Each sampling point contains *time*, *latitude*, *longitude*, *speedmagnitude*, *direction* as well as a boolean *CarryPassengerState*, which is a tag indicating whether the taxi carries passengers or not. In this work, we only use the trajectories with passengers.

Along with taxi GPS dataset, the road network data is collected from OpenStreetMap's jXAPI [9]. Following an existing paper [30], in the data preprocessing step, trajectories are cleaned and matched to the road network. The final data size is 12.1 GB.

6.2 Implementation

We have implemented a prototype system to test the effectiveness of our method. The system is written in C++, with Qt framework. The rendering is performed with both OpenGL and Qt GraphicsView Framework. Third-party libraries Graphviz [12] and MatLab [26] are integrated to do the topological graph layout and perform MNL model. We run the system on an Intel(R) Core(TM)2 2.66 GHz Laptop with 4 GB RAM and a NVIDIA Geforce GTX 470 GPU.

6.3 Case I: Exploring Multiple Route Choices

With the route spatial and route-related factor view, the system supports to explore the multiple route choices between interested regions, by fulfilling the tasks (**T1**)(**T2**).

In this case, the *OoI* and *DoI* are set at Beijing Airport and a central business district respectively. In total, 369 trajectories passing through are filtered from March 2 to March 25, 2009.

As Figure 11(a) shows, multiple routes are extracted from these trajectories, which are categorised into three groups of route choices. For each group, there is a mainstream route choice with dominating popularity and several alternative choices with small number of trajectories. To obtain a general understanding of route choices, the three mainstream route choices are selected. As the route spatial view in Figure 11(b) shows, the three share some common roads

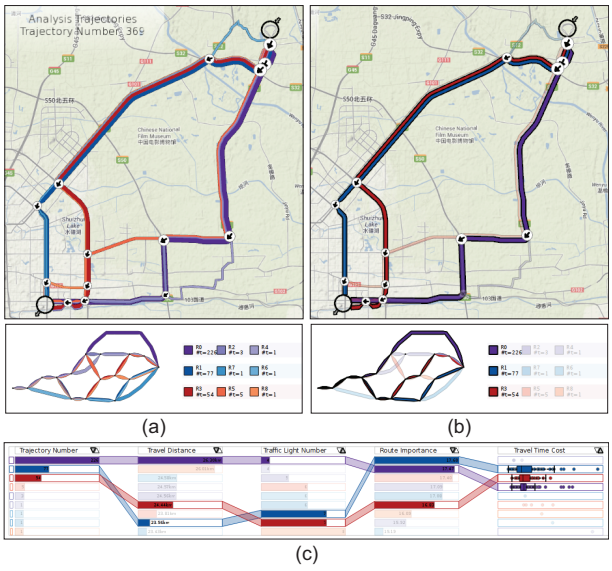


Figure 11: Case Study #1: exploring route choices from Airport to a commercial district: (a) the spatial overview. (b) three selected popular routes. (c) route-related factors' comparison among the selected routes

in the beginning and then the purple one splits from the other two. Subsequently, after travelling a distance, the blue and red heads into different roads respectively. Among the three route choices, nearly 60% taxi drivers choose the purple one and 20% choose the blue one and nearly 15% for the red one. The left 5% choose other less travelled routes.

By ranking factors in the route-related factor view, the three selected routes are compared over multiple attributes. The purple route has the longest travel distance but the least traffic light number, and the average route importance is in-between the blue and red one. Comparing the box plots in the travel time cost column, the blue and red have smaller median travel time cost than the purple one, which probably owes to their undeniable advantage in travel distance. However, the width of their boxes are larger than that of the purple one, i.e. their time cost distributions are more stretched. It indicates that the time cost of the purple route is more reliable and predictable because of its small variability,

6.4 Case II: Exploring Factors' Impact on Multiple Route Choices

In this case we show the system's capability on exploring trajectory-related factors' impact on route choices from hypothesis construction(**T3**) to statistics verification(**T4**).

Choosing route to cross different ring roads is a common multiple route choice problem 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 12(a) shows that the top three most popular route choices, i.e. the blue, red and orange, have much larger popularity than others. To compare their attributes, the top three are selected and ranked by attributes in the ranking view (Figure 12(d)). Overall, the top three routes have advantages over those seldom travelled routes in route distance, traffic light number as well as the route importance level. Within the three routes, the blue route ranks top. The ranking order of median time cost is consistent with that of route's popularity, which indicates

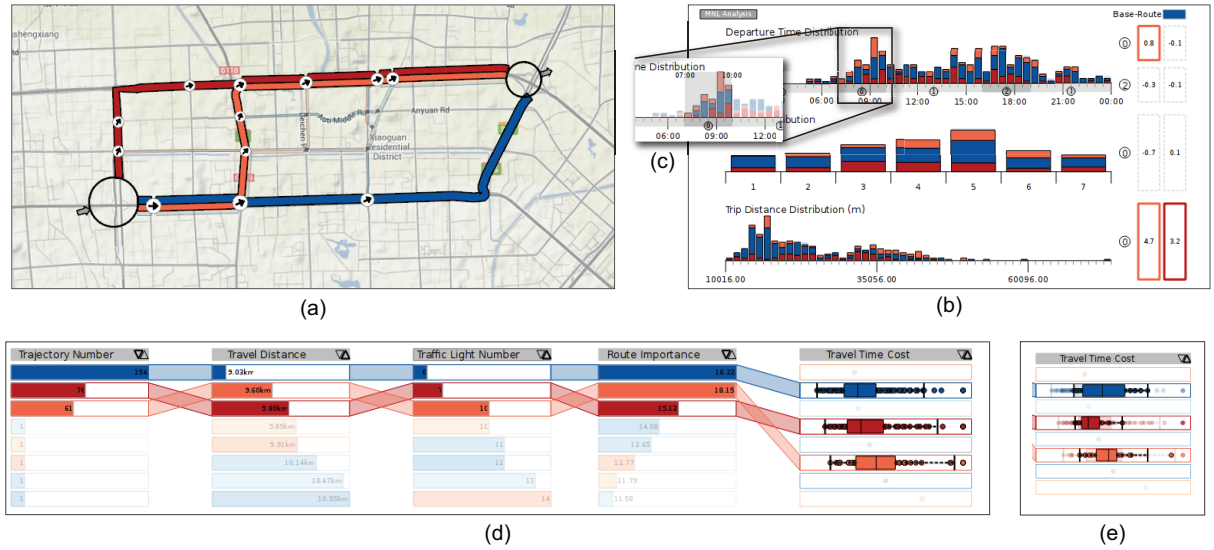


Figure 12: Case Study #2: (a) route spatial view of three selected major routes. (b) trajectory-related factor view. (d) route-related view with selected three routes; (e) travel time cost distribution of the three selected routes during departure time range in (c).

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 12(b). It is observed that the orange route increases dramatically around 9 o'clock in the morning, which raises hypothesis that drivers has larger probability to choose the orange road in the morning. By configuring model with three sub-factors in departure time factor, together with default configuration of the other two factors, output of route choice model is given in factor matrix in Figure 12(b). Taking the blue route as base route, the orange rectangle indicates that departure time from 7:00 to 10:00 in the morning has significant impact on the odds of orange route than the blue route. That is, when travelling in the morning peak, drivers have larger probability to choose orange route than the blue one. As Figure 12(c) shows, by selecting the trajectories travelling during this time period, the distribution of time cost in the ranking attribute view is updated as Figure 12(e) shows. The average time cost of blue route increases and the variance increases. Comparing with the blue one, the red and orange ones have more reliable time cost during this time period. Interestingly, as the trajectory's length increases, the odds for both the red and orange routes increase, which may indicate that drivers tend to drive to the 4th right as soon as possible if travelling far.

7. 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 sampling number. 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 based on four proposed 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. At last, with the historical Beijing taxi GPS trajectory dataset, we demonstrate two case studies to show its effectiveness.

In the future, we would like to improve and extend our system regarding to the current limitations. Currently, the input factors are fixed. Following, the system will be customized to accept creation of factors. For example, *OD* distribution can be one of the possible trajectory-related factors. Another interest point is to extend to system with route advisory function. It is possible to recommend routes by taking different factors into consideration and measure the fitness of route.

8. ACKNOWLEDGMENTS

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. 2015CB352500. This work is also funded by PKU-Qihu Joint Data Visual Analytics Research Center.

9. REFERENCES

[1] Google map. http://maps.google.com/.

[2] 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*, 80(1):64–75, 1998.

[3] N. Adrienko and G. Adrienko. Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics*, 17(2):205–219, 2011.

[4] 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*, 22(3):213 – 232, 2011.

[5] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel. *Visual analytics of movement*. Springer, 2013.

[6] 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, 2008.

[7] 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. pages 697–711, 1996.

[8] J. de Dios Ortúzar and L. G. Willumsen. *Modelling transport*. John Wiley & Sons, 2011.

[9] I. Dees. Openstreetmap jxapi. http://wiki.openstreetmap.org/wiki/Xapi.

[10] T. A. Domencich and D. McFadden. Urban travel demand-a behavioral analysis. *American Economic Association*, 1975.

[11] 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*, 19(12):2149–2158, 2013.

[12] E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *Software Practice and Experience*, 30(11):1203–1233, 2000.

[13] 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*, pages 330–339, 2007.

[14] 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*, 19(12):2277–2286, 2013.

[15] D. Guo. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1041–1048, 2009.

[16] 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)*, pages 163–170, 2011.

[17] M. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40:27–37, 2003.

[18] V. Henn. Fuzzy route choice model for traffic assignment. *Fuzzy Sets and Systems*, 116(1):77–101, 2000.

[19] 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*, 29(2):193–212, 1995.

[20] 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*, 32(3):451–460, 2013.

[21] 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*, 19(9):1639–1645, 2009.

[22] R. G. Li, Hainan 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*, 1926.1:162–170, 2005.

[23] H. Liu, Y. Gao, L. Lu, S. Liu, H. Qu, and L. Ni. Visual analysis of route diversity. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 171–180, 2011.

[24] M. Lu, Z. Wang, and X. Yuan. Trajrank: Exploring travel behaviour on a route by trajectory ranking. In *Proceedings of IEEE Pacific Visualization Symposium*, 2015.

[25] 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*, 2(1):35–47, 1994.

[26] MATLAB. *version 8.2.0.701 (R2013a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

[27] G. Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

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

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

[30] 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.

[31] C. Weaver. Cross-filtered views for multidimensional visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):192–204, 2010.

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

[33] 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*, pages 1833–1842, 2014.

[34] W. Zeng, C.-W. Fu, S. M. Arisona, and H. Qu. Visualizing interchange patterns in massive movement data. *Computer Graphics Forum*, 32(3pt3):271–280, 2013.