AL-DOHUKI, SHAMAL, Ph.D., May 2019 Computer Science

INTERACTIVE VISUAL QUERYING AND ANALYSIS FOR URBAN TRAJECTORY DATA

Dissertation Advisor: Ye Zhao

Advanced sensing technologies and computing infrastructures have produced a variety of trajectory data of moving objects in urban spaces. One type of this data is taxi trajectory data. It records real-time moving paths sampled as a series of positions associated with vehicle attributes over urban road networks. Such data is big, spatial, temporal, unstructured and it contains abundant knowledge about a city and its citizens. Exploratory visualization systems are needed to study taxi trajectories with efficient user interaction and instant visual feedback. The extracted information can be utilized in many important and practical applications to optimize urban planning, improve human life quality and environment. As the primary novelty contribution, this thesis presents a set of visual analytics solutions with different approaches to interacting with massive taxi trajectory data to allow analysts to look at the data from different perspectives and complete different analytical tasks. Our approaches focus on how people directly interact with the data store, query and visualize the results and support practitioners, researchers, and decision-makers to advance transportation and urban studies in the new era of the smart city.

First, we present SemanticTraj, a new method for managing and visualizing taxi trajectory data in an intuitive, semantic rich, and efficient means. In particular, taxi trajectories are converted into taxi documents through a textualization transformation process. This process maps global positioning system (GPS) points into a series of street/POI names and pickup/drop-off locations. It also converts vehicle speeds into user-defined descriptive terms. Then, a corpus of taxi documents is formed and indexed to enable flexible semantic queries over a text search engine.

Second, we present a visual analytics system, named as QuteVis, which facilitates domain users to query and examine traffic patterns from large-scale traffic data in an urban transport database. QuteVis supports a different type of data query and analytical tasks. It helps users discover those specific times and days in history that have similar traffic patterns as they speculate on multiple, spatially-diverse city locations.

Third, we present a web-based software, named TrajAnalytics, for the visual analytics of urban trajectory datasets. It allows users to interactively manage, analyze, and visualize the massive taxi trajectories over urban spaces. The software offers data preprocessing and management capability and enables various visual queries through a web interface.

Finally, a set of visual exploration tools have been implemented to be utilized in the systems above for visual exploration of trajectory data. These visual exploration tools are considered as an effective approach to providing material for human's perception and plays a vital role in analyzing and visualizing trajectory data.

INTERACTIVE VISUAL QUERYING AND ANALYSIS FOR URBAN TRAJECTORY DATA

A dissertation submitted to
Kent State University in partial
fulfillment of the requirements for the
degree of Doctor of Philosophy

by

Shamal Mohammed Ameen Taha AL-Dohuki

May 2019

Dissertation written by

Shamal Mohammed Ameen Taha AL-Dohuki

B.Sc., University of Duhok, 2005

M.Sc., University of Duhok, 2008

Ph.D., Kent State University, 2019

Approved by

______________________________, Chair, Doctoral Dissertation Committee
Ye Zhao

______________________________, Members, Doctoral Dissertation Committee
Xiang Lian

____
Cheng-Chang Lu

____
Xinyue Ye

____
Xiaoling Pu

Accepted by

______________________________, Chair, Department of Computer Science
Javed Khan

______________________________, Dean, College of Arts and Sciences
James L. Blank

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

I want to express my sincere gratitude to my advisor Prof. Ye Zhao for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study. He always offers me his kindness, support, guidance, and friendship over the last five years. Besides my advisor, I would like to thank many excellent colleagues and my paper coauthors, both internally and externally. This thesis would not be able to reach this level without their contribution.

I want to thank all of the jury members of my thesis defense, including Prof. Cheng-Chang Lu, Prof. Xiang Lian, Prof. Xinyue Ye, and Prof. Xiaoling Pu. I am appreciative of their time and efforts in reading and assessing this thesis work.

I thank my fellow labmates for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last five years. I want to thank the Kurdistan Regional Government (KRG) and the HCDP team for their support. Last but not least, I would like to thank my family for supporting me spiritually throughout writing this thesis and my life in general.

Dedication

Dedicated to My Precious Family:

Father (Mohammed Ameen AL-Dohuki), Mother (Hiyam Kheder)

Father-in-law (Najeeb Alnomani), Mother-in-law (Dilbar Saed)

Wife (Hingav Alnomani)

Daughters (Shanar & Shan) & Son (Shiyan)

Brothers (Haval, Hawar, Huger, Hayman, Araz)

Ye Zhao, my adviser, my guide and my friend.

KRG & HCDP Team

Shamal AL-Dohuki Kent, Ohio, USA

May, 2019

CHAPTER 1

Introduction

In this chapter, we present the background and the motivation of this thesis work, as well as the contributions and organization of this dissertation.

1.1 Visual Data Analysis

In the era of big data, while the world is full of complex data that are collected by many organizations, companies, and researchers, they lack the capability to automatically analyze and comprehend the purport of such data especially with massive amounts of data. A navigator to get through and discover meaningful relationships between its pieces is demanded. Visual data analysis is the key to sailing smoothly through the sea of such big and complex data. There are two approaches to visual data analysis:

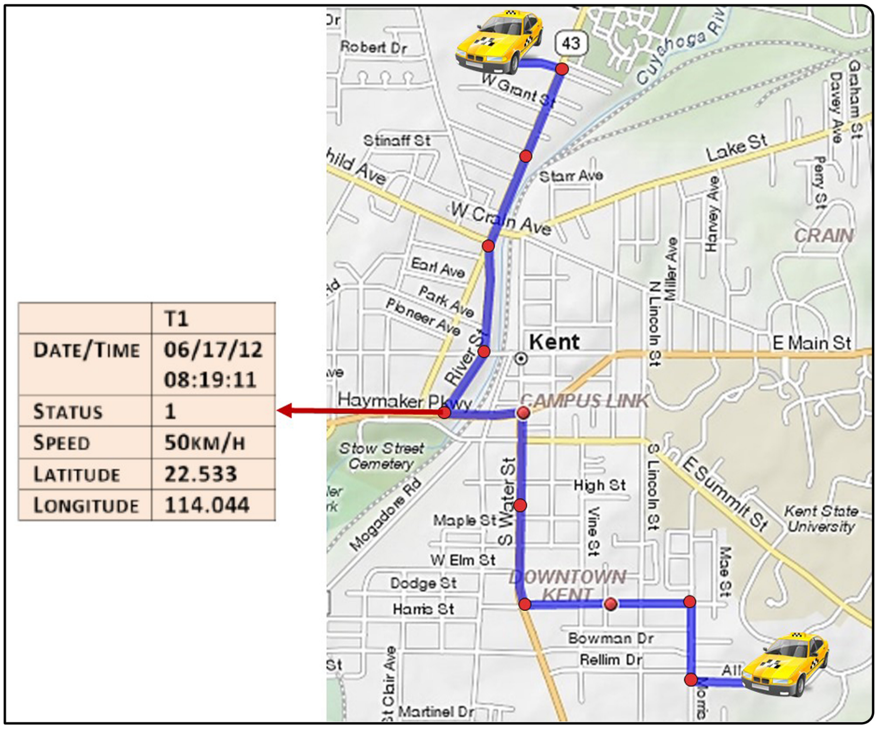
Data Visualization is a general term which refers to the graphical representation of the data (statistical, geographical, etc.). By placing data in a visual representation, viewers can detect patterns, trends or anomalies that can be subsequently useful in the analysis that may not be apparent when presenting data as a set of numbers.

Visual Analytics is the key technology that helps people to make decisions through explore, understand, find answers and build stories in the data. It is more than just a visualization. It is an integral approach to decision-making that combines visualization, human factors, and data analysis. Visual analytics combines automated analysis techniques with interactive visualizations to enable effective understanding, reproducible reasoning, and defensible decision-making in the context of large and complex data sets [2]. The visualization aspects allow individuals to interact with a massive disparate set of data in ways that enable them to see multiple types of connections and relevancies among those data entities and extract them all in one place and one time as opposed to doing it linearly or serially or one after the other. The goal of visual analytics is the creation of tools and techniques to enable the user to (1) synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data; (2) detect the expected and discover the unexpected; (3) provide timely, defensible, and understandable assessments; (4) communicate assessment effectively for action [3].

1.2 Urban Trajectory Data

Movement is the change in the physical position of an object concerning a set reference system, which is in most spaces a two or three-dimensional representation of geographic space. A trajectory represents the path made by the moving object through space and time. Typically, trajectories are recorded through a series of positions in time, so a trajectory can be viewed as a function that attaches a location in (geographical) space to each time window. Time window related features are the position at a particular time, speed, acceleration or direction at a point in space, accumulated distance, etc.

Location acquisition technologies are producing massive spatial trajectory data of a diversity of moving objects, such as people, vehicles, and animals in urban spaces at an unprecedented scale and speed. With the prevalent low-cost global positioning system (GPS) devices, smartphones, wearable chips with an embedded global positioning system (GPS), and radio frequency identification (RFID) devices, population mobility information is accurately recorded as trajectories of people, taxis, fleets, public transits, and mobile phones. The availability of trajectory data has fostered a diversity of applications, calling for algorithms that can manage, explore and extract knowledge from the data in an efficient way. The obtained information has a significant advantage in the studies of the urban system, environment, energy, economy, and citizens to optimize urban planning, improve human life quality and environment, and amend city operations. In particular, the emerging urban trajectory data such as taxi trajectory data provides real situations from which the statistics of real traffic flow can be extracted, and city-wide transport patterns can be discovered [4]. Exploiting the emerging urban trajectory data can play a transformative role in transportation-associated research by offering domain experts, researchers, and decision-makers unprecedented capability to conduct data-driven studies based on real-world information. Robust, easy-to-use visual analytics system enabling effective exploration of the data is direly needed and will contribute to building capacity in seeking solutions for the social, economic, and environmental challenges facing our communities [4].



	T1
Date/Time	06/17/12 08:19:11
Status	1
Speed	50km/h
Latitude	22.533
Longitude	114.044

Figure 1: An example of a Taxi Trajectory Data.

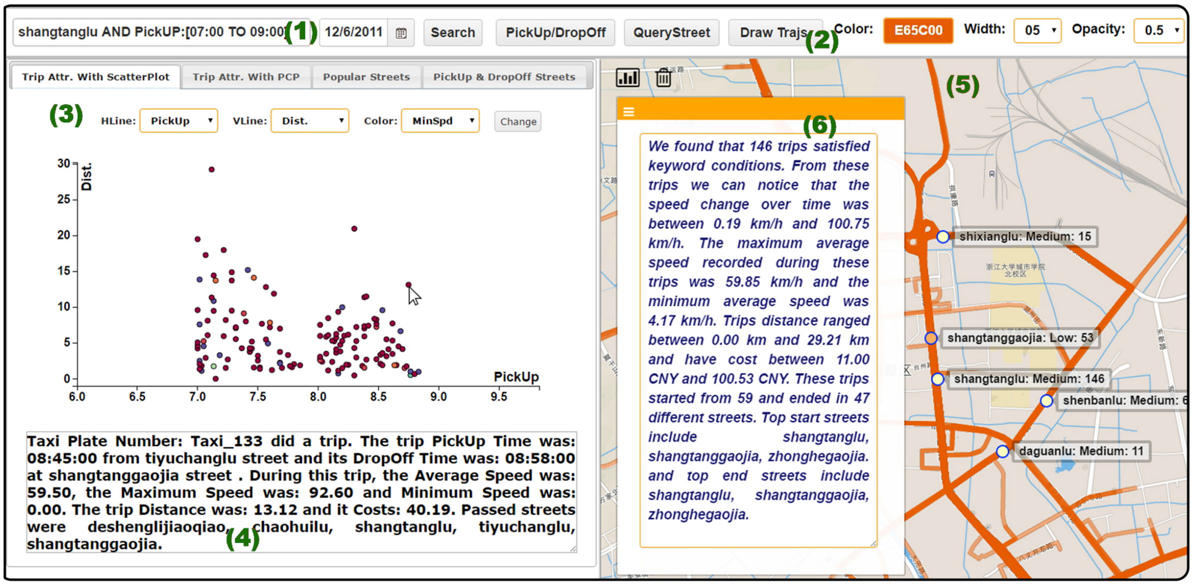
1.3 Motivation

The taxi trajectory data records real-time moving paths sampled as a chronologically ordered position over urban space (see Figure 1). Rich and heterogeneous information can be linked at each location, including vehicle and human attributes, geographical features, business/urban information, and more. Such data is big, spatial, temporal, dynamic, and unstructured. For example, Microsoft's T-drive system [5] collected the trajectories from over 33,000 Beijing taxis for three months. The total length of the trajectories is more than 400 million kilometers, and the total number of global positioning system (GPS) points reaches 790 million. The associated data of the point samples further increases the data scale.

To extract deep insights from the data, researchers must conduct iterative, evolving information foraging and sense-making and guide the process using their domain knowledge. Iterative visual exploration is one key component in the processing, which should be supported by efficient data management and visualization tools. Therefore, practitioner, researchers, and decision-maker demand a handy and effective visual analytics software system which integrates scalable data management and interactive visualization with powerful computational capability. Implementing such a system is not an easy task because even a simple function, such as smoothly displaying the heat maps of thousands of taxis' average, maximum, or minimum speed, cannot be easily completed with active user interactions (e.g., frequent zoom-in/out of regions and changes of time periods). Such operations require temporal and spatial data aggregations and visualizations with random access patterns, where advanced computational technologies should be exploited.

1.4 Contributions

Towards the motivation, this dissertation presents a set of visual analytics solutions with different approaches for computing, interacting, and understanding massive mobility data such as taxi trajectory data. This thesis formulates four major contributions, corresponding to focus on *people directly interact with the data store, query, analyze, and visualize the results* and support practitioners, researchers, and decision-makers to advance transportation and urban studies in the new era of the smart city.



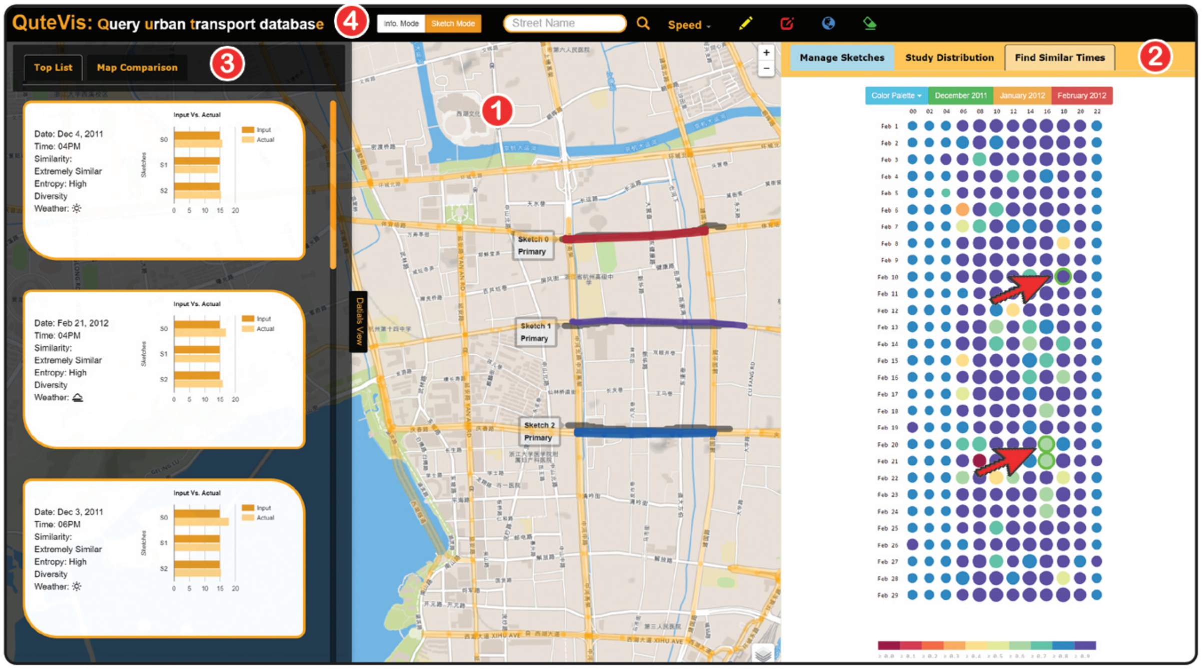
Taxi Plate Number: Taxi_133 did a trip. The trip PickUp Time was: 08:45:00 from tiyuchanglu street and its DropOff Time was: 08:58:00 at shangtanggaojia street . During this trip, the Average Speed was: 59.50, the Maximum Speed was: 92.60 and Minimum Speed was: 0.00. The trip Distance was: 13.12 and it Costs: 40.19. Passed streets were deshenglijiaoqiao, chaohuilu, shangtanglu, tiyuchanglu, shangtanggaojia.

We found that 146 trips satisfied keyword conditions. From these trips we can notice that the speed change over time was between 0.19 km/h and 100.75 km/h. The maximum average speed recorded during these trips was 59.85 km/h and the minimum average speed was 4.17 km/h. Trips distance ranged between 0.00 km and 29.21 km and have cost between 11.00 CNY and 100.53 CNY. These trips started from 59 and ended in 47 different streets. Top start streets include shangtanglu, shangtanggaojia, zhonghegaojia. and top end streets include shangtanglu, shangtanggaojia, zhonghegaojia.

Figure 2: Using SemanticTraj to visualize taxi trips which passed Shangtanglu street of Hangzhou, China in the morning (7am-9am) of Dec 6, 2011. See details in Sec. 3.7. (1) Query input box accepting semantic query conditions as Shangtanglu AND PickUp:[7:00-9:00]; (2) Visualization control panel for adjusting the appearance; (3) Scatterplot view for users to study search results. Other visual tools can be selected in this view; (4) Meta-summary of a selected trip which automatically summarizes the trip fact; (5) Map view showing trip trajectories. Text labels are displayed on critical streets about its role in these trips; (6) A meta-summary of the group of all 146 result trips. Users can interact with the name tags to filter trips.

First, we present SemanticTraj (Figure 2), a new method for managing and visualizing taxi trajectory data in an intuitive, semantic rich, and efficient means. Existing tools typically require users to select and brush geospatial regions on a map when retrieving and exploring taxi trajectories and passenger trips. With SemanticTraj, domain and public users can find answers to seemingly simple questions such as "What were the taxi trips starting from Main Street and ending at Wall Street in the morning?" or "Where are the taxis arriving at the Art Museum at noon typically coming from?", easily through direct queries based on the terms. They can also interactively explore the retrieved data in visualizations enhanced by semantic information of the trajectories and trips. In particular, taxi trajectories are converted into taxi documents through a textualization transformation process. This process maps global positioning system (GPS) points into a series of street/POI names and pick-up/drop-off locations. It also converts vehicle speeds into user-defined descriptive terms. Then, a corpus of taxi documents is formed and indexed to enable flexible semantic queries over a text search engine. Semantic labels and meta-summaries of the results are integrated with a set of visualizations in a SemanticTraj prototype, which helps users study taxi trajectories quickly and easily. A set of usage scenarios are presented to show the usability of the system. We also collected feedback from domain experts and conducted a preliminary user study to evaluate the visual system.

Second, we present QuteVis (Figure 3), a visual analytics system that supports the study of urban traffic patterns on transport databases. Unlike most existing approaches which investigate traffic patterns in user specified spatial regions and temporal periods, QuteVis supports a different type of data query and analytical tasks. Multi-sketch query and visualization helps users discover those specific times and days in history, which have specified joint traffic patterns distributed on different city locations. Users can use touch input devices to define, edit, and modify multiple sketches on a city map. Sketch recognition algorithm is reliable and flexible which supports sketches of a street, a path, or a region. The sketches specify combined spatial traffic conditions such as speed, volume,



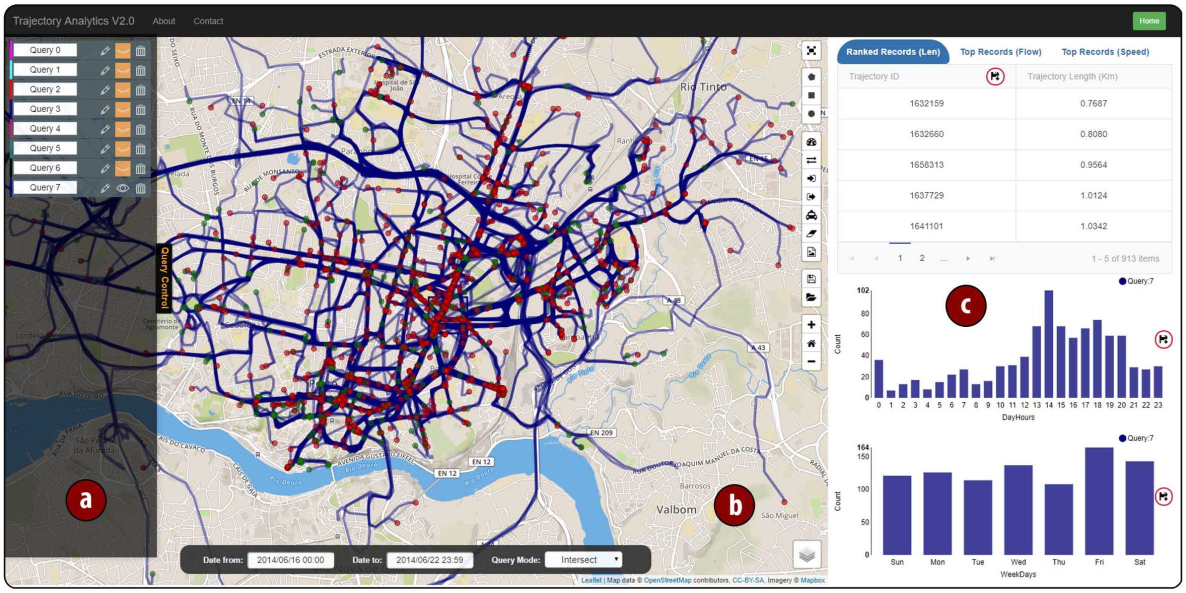
Date: Dec 4, 2011
Time: 04PM
Similarity:
Extremely Similar
Entropy: High
Diversity
Weather: ☼

Date: Feb 21, 2012
Time: 04PM
Similarity:
Extremely Similar
Entropy: High
Diversity
Weather: ☁

Date: Dec 3, 2011
Time: 06PM
Similarity:
Extremely Similar
Entropy: High
Diversity
Weather: ☼

Figure 3: QuteVis interface consisting of (1) a map canvas, (2) a control panel, (3) a detail study panel, and (4) a top panel. It supports visual queries and analytics of historical days and times that have similar traffic patterns as speculated by users.

and taxi pickups/dropoffs, on streets, regions, or paths. Then, weighted similarities between user speculated traffic pattern and actual historical traffic situations are computed. The weights are defined with respect to the hierarchical levels of different streets or users' preference. A set of visualizations and interactions are provided in QuteVis to help users browse and compare the retrieved traffic situations. Users can discover potential influential factors, such as weekdays and weather, through visual comparisons. They can also test hypotheses about the relationships of traffic behaviors between different locations in a city. In order to facilitate the sketch-based querying and exploration with interactive responses, we construct a transport database from heterogeneous data sources. It supports fast spatio-temporal data retrieval with an optimized spatial indexing and weighted similarity computation. We conduct several case studies with real world data and domain experts in transportation and urban planning. The studies demonstrate how QuteVis is useful in addressing many major problems in modern cities, such as traffic jams, unbalanced traffic flows, and unsatisfied travel demands.



Trajectory ID	Trajectory Length (Km)
1632159	0.7687
1632660	0.8080
1658313	0.9564
1637729	1.0124
1641101	1.0342

1 2 ... 1 - 5 of 913 items

Date from: 2014/06/16 00:00 Date to: 2014/06/22 23:59 Query Mode: Intersect

Figure 4: The visualization interface of TrajAnalytics: It contains convenient functions for effective and efficient knowledge discovery: (a) Control panel of queries, (b) Interactive map view, (c) Visual report view.

Third, we present a web-based software, named TrajAnalytics (Figure 4) that provides exploratory data visualization tools for researchers, administrations, practitioners and general public to understand the data and to reveal knowledge intuitively. Advanced technologies in sensing and computing, the mobility patterns and dynamics of urban cities and their citizen are recorded and manifested in a variety of urban trajectory datasets, which include the moving paths of human, taxi, bus, fleets, cars, and so on. Understanding and analyzing such large-scale, complex data is of great importance to enhance both human lives and urban environments. TrajAnalytics allows users to interactively manage, analyze, and visualize the massive taxi trajectories over urban spaces. The software offers data processing and management capability and enable various visual queries through a web interface. A set of visualization widgets and interaction functions are developed to promote easy user engagement. It contains three major modules: (1) Data loading and transformation; (2) TrajBase database; (3) TrajVis visual analytics interface. The system is developed in two forms: a cloud based software and a local version, to fulfill the requirements of many real world users. It has been released for open access.

Finally, a set of visual exploration tools have been implemented to be utilized in the above visual analytics solutions for visual exploration of trajectory data. It enables users to perform progressive visual exploration, hypothesis generation and evaluation with real-time interactive visualization. These visual exploration tools are considered as an effective approach to providing material for human's perception and plays a vital role in analyzing and visualizing trajectory data.

1.5 Thesis Organization

After the general discussion about motivation and contribution of this thesis, the remainder of this dissertation is organized as follows:

- Chapter 2 presents a broad view of related works towards trajectory data from several different perspectives, including trajectory data visual analytics, management, processing, mining, and semantic.

- Chapter 3 presents *SemanticTraj*, a new method for managing, analyzing, and visualizing taxi trajectory data in an intuitive, semantic rich, and efficient means. It utilizes the power of text search engine to supports various queries by keyword-based conditions and no particular data structure is needed to manage trajectories.

- Chapter 4 presents *QuteVis*, which facilitates domain users to query and examine traffic patterns from large scale traffic data in an urban transport database by using multi-sketch query.

- Chapter 5 presents *TrajAnalytics*, the exploratory data visualization tools for researchers, administrations, practitioners and general public to understand the data and to reveal knowledge intuitively, that designed and developed in close collaboration with domain experts.

- Chapter 6 provides concluding remarks. Based on the work in this dissertation, many interesting problems for further theoretical study and practical application exist. We present some interesting open issues and future research directions.

1.6 Publications and Research Activities

- The first research contribution (Chapter 3) has been published in the IEEE Transactions on Visualization and Computer Graphics, 23(1), 11 - 20, Jan. 2017 [**Shamal AL-Dohuki**, Farah Kamw, Ye Zhao, Chao Ma, Yingyu Wu, Jing Yang, Xinyue Ye, Fei Wang, Xin Li, and Wei Chen, "SemanticTraj: A New Approach to Interacting with Massive Taxi Trajectories"]

- The second research contribution (Chapter 4) is accepted to be published in the IEEE Computer Graphics and Applications (CG&A), 2019 [**Shamal AL-Dohuki**, Ye Zhao, Farah Kamw, David Sheets, Jing Yang, Xinyue Ye, Wei Chen, "QuteVis: Visually Studying Transportation Patterns Using Multi-Sketch Query of Joint Traffic Situations"]

- The third research contribution (Chapter 5) is submitted to the 22nd Intelligent Transportation Systems Conference (ITSC), 2019 [**Shamal AL-Dohuki**, Farah Kamw, Ye Zhao, Xinyue Ye, and Jing Yang, "An Open Source TrajAnalytics Software for Modeling, Transformation and Visualization of Urban Trajectory Data"]

  Moreover, a short paper has been published in the IEEE Workshop on Visualization in Practice : Open Source Visualization and Visual Analytics Software, IEEE Visualization Conference 2016, Baltimore, Oct, 2016 [Ye Zhao, **Shamal AL-Dohuki**, Thomas Eynon, Farah Kamw, David Sheets, Chao Ma, Yueqi Hu, Xinyue Ye, Jing Yang, "TrajAnalytics: A Web-Based Visual Analytics Software of Urban Trajectory Data"]

- Another research contribution has been published in the IEEE Transactions on Intelligent Transportation Systems, Jan. 2019 [Farah Kamw, **Shamal AL-Dohuki**, Ye Zhao, Thomas Eynon, David Sheets, Jing Yang, Xinyue Ye, Wei Chen, "Urban Structure Accessibility Modeling and Visualization for Joint Spatiotemporal Constraints"]

- Another research contribution is to be submitted to the IEEE Conference on Visual Analytics Science and Technology (IEEE VAST 2019) [Chao Ma, Ye Zhao, Andrew Curtis, Farah Kamw, **Shamal AL-Dohuki**, Jing Yang, Suphanut Jamonnak, Ismeal Ali, "Semantic Driven Visual Analysis of Community-Level Events"]

- Another research contribution is to be submitted to the IEEE Conference on Visual Analytics Science and Technology (IEEE VAST 2019) [Suphanut Jamonnak, Ye Zhao, **Shamal AL-Dohuki**, Andrew Curtis, Farah Kamw, Jing Yang, and Xinyue Ye, "GeoVisuals: A Visual Analytics Approach to Studying Geospatial Videos and Narratives"]

- Another research contribution is to be submitted to the IEEE Conference on Visual Analytics Science and Technology (IEEE VAST 2019) [Chao Ma, Ye Zhao, **Shamal AL-Dohuki**, Jing Yang, Xinyue Ye, Farah Kamw, and Andrew Curtis, "SeMapLens: Semantic Map Lens for Interactive Visual Exploration of Geo-Text Data"]

- A tutorial has been given in the IEEE VIS 2018, Berlin, Germany: Urban Trajectory Data Visualization [Ye Zhao (Kent State University, Kent, Ohio, United States), Jing Yang (UNCC, Charlotte, North Carolina, United States), Wei Chen (Zhejiang University, Hangzhou, China), **Shamal AL-Dohuki** (Kent State University, Kent, Ohio, United States)]

CHAPTER 2

State of the Art

This chapter presents a literature review that related to this thesis work, in particular, studies on how to interact with urban trajectory data including processing and management, querying, analyzing, and visualizing such data.

2.1 Trajectory Data Cleaning

In this chapter, many studies are built upon an assumption that "*the positioning of the moving objects is precisely provided*". However, the trajectory data in real-life is more unreliable than expected to be used by the various applications. For instance, TELeCOmmunications (TELCO) data that collected by mobile sensors are often imprecise and incorrect due to the limitation of positioning techniques (e.g., inaccurate global positioning system (GPS) measurement and sampling errors, indoor signal loss, smartphone battery runs out) [6,7]. Another reason resides on the privacy aspect of users (e.g., people do not want to disclose their precise or private locations, and deliberately expose only an approximation of positioning data) [8]. Therefore, to build more useful trajectory databases and a better understanding of mobility data, trajectory data cleaning is needed and cannot be overlooked.

Different errors occur in the trajectory data (e.g., GPS errors, Sampling errors, Transformation errors, Missing data). The main focus of trajectory data cleaning is to remove or reduce the number of records that have errors such as GPS errors. GPS errors are wrong measurements of the spatial position (i.e., not exactly the correct location). Jun et al. summarizes two types of GPS errors: systematic errors and random errors [9]. The systematic errors are the large wrong values that are totally invalid GPS positioning from the actual location, which may be caused by the low number of satellites in view and Horizontal Dilution Of Position (HDOP). The random errors are the small wrong values up to ±15 meters that are caused by the satellite orbit, clock or receiver issues. Both types of errors refer to the spatial domain, as the temporal information is considered precise due to the high calibration clocks that the satellites are equipped with.

To detect and remove systematic errors, some researchers have recourse to visual inspection with automated filtering methods which is not practical for large datasets. Marketos et al. proposed a parametric online approach that filters noisy positions by taking into advantage of the maximum allowed speed of a moving object. A given threshold or parameter is used to determine whether a reported time-stamped position from the GPS stream, must be considered as noise and consequently discarded, or kept as a normal record [10]. Random errors are small distortions from the actual values, and their influence can be decreased by smoothing methods. Different techniques are presented by various researchers (e.g., [9, 11, 12]). Yan et al. presented an approach based on the idea of nearest neighbor smoothing and locally weighted regression models [7]. Lashley et al. smoothed data points by recursively modifying error values by using a Kalman filter, which uses measurements observed over time (the positions coming in the GPS receiver), and predicts positions that tend to be closer to the true values of the measurements. The Kalman filter smoothes a position by computing a weighted average of the predicted position and the measured position [11].

In addition to removing or reducing errors, there are also missing values due to the signal loss (e.g., in tunnel or indoor) of GPS trajectories, which need to be interpolated in some applications. An example of such interpolation is provided in [13] where the authors use polynomial interpolation to estimate the missing data. In this thesis, different existing approaches are used to handle random errors and clean the data including smoothing and interpolation.

2.2 Trajectory Data Map Matching

Moving objects can be classified into two groups. First, moving objects that are allowed to move freely, without any constraints in their movement such as animal movement. Second, moving entities such as people or vehicles are by their nature restricted to move within a given road network. For example, taxis are restricted to move on a road network, public buses have their own planned routes, and people are not allowed to walk on the highway. This type of trajectory data is called network-constrained trajectories [14]. Transportation network represented as a directed graph G=(V, E) , where V is a set of nodes corresponding to the road intersections, E is a set of edges corresponding to the road links. Direction of the edge is defined by the driving direction at the corresponding road link [15]. For analyzing network-constrained trajectory data, there is a very active research area in the Geographic Information System (GIS) with many algorithms called “map matching”. The objective of map matching is to determine (or estimate) the actual trajectory traces in a given road network from the raw global positioning system (GPS) tracking points with errors. It maps the positional measurements onto a network which it supposes to travel on, estimating the correct position in a node or edge of the network graph [16, 17].

Map matching methods can be classified into three types: (1) *Local Method* focuses on individually matching a single GPS point to a path in the road network based on a given distance measurement [18]; (2) *Incremental Method* improves the local method and matches a portion of the trajectory (i.e., both current point and previous points) [17]; and (3) *Global Method* analyzes the entire view of a trajectory (i.e., both past and coming GPS points) and find the best matching road segment [19].

These various map matching approaches include several post-processing techniques to calibrate and correct the initial matching results. For example, Krüger et al. presented Visual Interactive Map Matching, a visual and interactive approach for map matching. It allows users to interactively and visually clean trajectory data, adjust matching parameters, and edit the underlying road network structure [20]. Obviously, this could decrease the algorithm's efficiency (i.e., computation time). Therefore, besides the effectiveness of a map matching algorithm that can be measured by the final matching accuracy, a map matching algorithm should also consider the algorithm's efficiency, in particular for working on massive trajectory datasets. In this thesis, we used Local Method for matching POI, trip pickup, and trip drop-off to the nearest street segment and Global Method to match trips or trajectories to street segments in the road network.

2.3 Trajectory Data Management

Over the past decades, with the rapid development of consumer electronics (e.g., GPS-equipped gadgets, smartphones, smart watches, and vehicles, as well as radio frequency identification (RFID) tag tracking and sensor networks), the database community started to face the more extensive availability of mobility data that is generated by moving objects and such data is typically called *trajectories*. Similar to other data types, trajectories are required to be accurately modeled, structured, and queried efficiently. A lot of relevant database researches have been established and become hot topics in the data management field such as *A Survey of Spatio-Temporal Database Research* [21]. Traditional database technology has been extended into Moving Object Databases (MODs) to handle modeling, indexing, and query processing issues for trajectories. Different researchers introduce several prototypes of MODs. Trajcevski et al. introduced DOMINO (Database fOr MovINg Objects) system for managing moving objects databases [22].

Mokbel et al. presented PLACE (Pervasive Location-Aware Computing Environments), a scalable location-aware database server developed at Purdue University [23].

Nowadays, different extensions introduced for existing Relational Database Management Systems (RDBM) such as MySQL extensions for Spatial Data [24]. This extension allows the user to store spatial data (e.g., point, line, polygon) and uses R-tree for indexing. The index is built using the minimum bounding rectangle (MBR) of a geometry. This extension provides two types of queries. Point queries that search for all objects that contain a given point and Region queries that search for all objects that intersect a given region. PostGIS [25] is a spatial database extender for PostgreSQL object-relational database. It adds support for geographic objects allowing location queries to be run in Structured Query Language (SQL). It provides several index types (B-tree, R-tree, Hash, and Generalized Search Tree (GIST)) and each index type uses a different algorithm that is best suited to different kinds of queries. By default, the CREATE INDEX command in PostgreSQL will create a B-tree index, which fits the most common situations. Emerging Not-only Structured Query Language (NoSQL) databases are more and more replacing relational ones for their ability to easily scale by distributing data over different servers while providing a flexible data model [26]. Different than the tabular relations used in traditional relational databases, NoSQL databases have a dynamic schema for unstructured data. They use a variety of data models, including document, graph, key-value, in-memory, and search [27]. Some NoSQL databases are currently being used to manage geospatial data include MongoDB (open source, used in Foursquare), BigTable (developed by Google, used in Google Earth), Cassandra (developed by Facebook, now open source and maintained by Apache), CouchDB (open source, Apache).

In this thesis, different approaches are used to manage different types of urban trajectory data for different tasks. In SemanticTraj (Chapter 3), we utilized the power of Lucene text search engine as a NoSQL spatial database where a trajectory/trip record is considered as a document. Then a corpus of taxi documents is formed and indexed by pertinent indexing schemes of the engine. The engine is well built and optimized so that large trajectory/trip data can be processed with outstanding query performance. In QuteVis (Chapter 4) and TrajAnalytics (Chapter 5), we utilized PostGIS over PostgreSQL to construct a transport database from heterogeneous data sources.

2.4 Trajectory Data Indexing

The design of an efficient and effective trajectory indexing structure is a significant and challenging task, which can ensure high performance for trajectory data management, particularly for achieving efficient data querying. Different approaches introduced by various researchers. Two surveys have summarized these trajectory and spatio-temporal indexing techniques [28,29]. In order to handle spatial data efficiently in Geo-data applications, Guttman et al. described a dynamic index structure called R-tree [30]. It helps to retrieve data items quickly according to their spatial location. Based on the R-tree more and more indexes are developed to add efficiency to spatial queries including those involving moving objects such as R^+-tree, R^*-tree, Quad-tree, TPR-tree, TPR^*-tree (Time

Parametric Rectangle), B-tree, and (uniform, parallel, hierarchical) grid, along with attributes such as dimensional (3D), temporal (history and prediction), etc [31]. Indexes with variant structures are experimented for different goals to improve performance. In general, two index approaches are used by existing database [32]:

Augmented R-tree: This type uses any multidimensional access method like R-tree indexes with augmentation in temporary dimensions such as 3D R-tree, or STR-tree [33]. R-tree is a height-balanced data structure. It is efficient and straightforward and has been used in the spatial database research community and commercial spatial database management systems such as MySQL and PostgreSQL [34]. Each node in R-tree represents a region which is the minimum bounding box (MBB) of all its children nodes. Each data entry in a node contains the information of the MBB associated with the referenced children node. The search key of an R-tree is the MBB of each node. R-tree can be used in range, point and nearest neighbor queries to retrieve all spatial objects within a specified query region.

R-tree has two disadvantages: (1) The execution of a point location query in an R-tree may lead to multiple searching paths followed from the root to the leaf level. This characteristic may lead to performance deterioration, especially when the overlap of the MBRs is significant; (2) A few large rectangles may increase the degree of overlap significantly, leading to performance degradation during range query execution, due to empty space; A variant of R-tree, called R^+-tree and R^*-tree have been developed to minimize node access by reducing overlapping of MBBs [32].

Grid-Based Indexing: It is a space partitioning indexing techniques and handling the overfull region by the split operation. The grid index and the quad-tree are very closely related. There are some differences between them: the grid index as 2D cells is a height-balanced structure while not in a quad-tree. Thus the grid index is more efficient in searching with the shorter average path, especially for skewed data. SETI indexing approach (Scalable and Efficient Trajectory Index) is the first grid-based index proposed by [35]. SETI decouples the indexing of the spatial and the time dimensions which leads to greater search and update efficiencies.

Existing spatial indexing methods (Augmented R-tree and Grid-Based Indexing) are designed mainly for static data and suffer from intensive updates due to the following reasons: (1) Most of the indexing approach process one single update at a time, which block mostly the update ability of the index; (2) Most of the existing approaches locate and remove the old object entry upon an update; This process leads to significant disk overhead [36].

Different from the existing methods, in SemanticTraj (Chapter 3) a text search engine maintains and manages an inverted index structure based on the indexing scheme of documents. It is designed to efficiently construct, compress, and manipulate all the document indexes, as well as the inverted index structures. It is also optimized over memory and disk to handle data indexing and queries. Utilizing the engine gives us the power and flexibility of data management and interactive retrieval for efficient visualization. In QuteViswe (Chapter 4), All the sample points from taxi traces are indexed by their spatiotemporal attributes. Each GPS point is also map-matched to a specific street segment. Next, we build a traffic data cube by aggregating the raw data records to compute and store the traffic attributes such as speed, taxi pickups, taxi drop-offs, and taxi flow volume, on each of these street segments. Unlike existing work such as Nanocube [37], our data aggregation and caching is conducted on street segments instead of spatial grid cells (of a tree subdivision). This is a unique feature since we seek to retrieve traffic attributes only on streets. In TrajAnalytics (Chapter 5), Three types of indexes have been added to the different tables to improve the performance of the queries over these tables. B-TREE is used over the Trip IDs to enhance the search by the Trip ID. To improve the spatial queries, Generalized Search Tree (GIST) indexing is used to index the geometric data types such as trajectory/trip path, pickup, and drop-off. Generalized Inverted Index (GIN) indexing is used for handling cases where the items to be indexed are composite values such as arrays, and the queries to be handled by the index need to search for element values that appear within the composite items such as trajectory/trip day, time, speeds, average speed, etc.

2.5 Trajectory Data Query Models

Existing query models of the trajectory data interested in searching and finding trajectories with respect to a given space and time. It includes several traditional spatial data querying techniques, such as point, region, and k nearest neighborhood (kNN) queries. A search inherited from spatial and temporal databases. (e.g., “find all objects within a given area (or at a given point) sometime during a given time interval” or “find the k-closest objects with respect to a given point at a given time interval") [38]. In general, we can summarize trajectory data queries into two types [33]:

Coordinate-Based Queries: This type of queries are similar to traditional spatial database queries that focus on offering information about spatial coordinates, which include [39]:

1. **Point Queries:** This type of query aims to find a set of trajectories that passed through a specific location. (e.g., find the location of specific object between 1:00pm-1:30pm). Pfoser et al. introduced VAIT, the visual analytics system for urban transportation. The system provides many queries based on grid indexing. The user can query the location or trajectory of a given taxi for a given time or a given time period [33].

2. **Region Queries:** This type of query aims to find a set of trajectories that passed through certain space or time intervals. (e.g., find all trajectories passed through R region between 1:00pm-1:30pm). [40] Proposed region query based on the k-d tree to find all trips that picked up or dropped off from the selected region. [41] Introduced TrajStore, the adaptive storage system for huge trajectory datasets, which provided range queries via a very limited number of I/Os.

3. **K- Nearest Neighbor Queries:** This type of query aims to find a set of trajectories that share similarities with a given set of trajectories. (e.g., find all trajectories within 500m of a gas station between 1:00pm-1:30pm). Hu et al. presented Taxi-Viewer system that shows nearest neighbor taxis based on the user

Geo-location [42].

Trajectory-Based Queries: This type of queries can be classified into two categories [43]:

1. **Topological Queries:** This type of queries aims to answer questions like "*When did vehicle X enters street Y most recently.*" It is interested in the data stored in whole or part of a trajectory. Operations involved are like enter, leaves, bypass, and crosses.

2. **Navigational Queries:** This type of queries aims to answer questions like "*What is the current speed of vehicle X.*" Overhead of computing the answers are paid because the information needed is not directly stored. Operations involved are like speed, heading, travel distance, etc.

In TrajAnalytics (Chapter 5), TrajQuery provides both Coordinate and Trajectory based queries. It supports users to conduct spatial queries combined with temporal constraints. It is defined to complete the four query modes (Pass, Start, End, Contain). Moreover, it supports joint queries based on Boolean operations. Unlike TrajAnalytics (Chapter 5) which investigate traffic patterns in user specified spatial regions and temporal periods, QuteVis (Chapter 4) supports a different type of data query and analytical tasks. It helps users discover those specific times and days in history that have similar traffic patterns as they speculate on multiple, spatially-diverse city locations. However, in many real-world tasks, domain and public users want to query urban trajectory data by combined conditions using street or POI names, and visually study the retrieved data. In SemanticTraj (Chapter 3), we proposed a new data query model, *name query*, as part of the framework of our visual query system of taxi trajectories to search massive trajectory data by utilizing the power of text search engine. It directly supports various queries by text-based conditions and no particular data structure is needed to manage trajectories. Users can intuitively conduct various information retrieval tasks through flexible query conditions, which are enabled by the powerful functions of boolean, wildcard, range and proximity query. This model is different from the conventional region query model of trajectories, which provides one or multiple spatial regions (or simply points) to find trajectories traversing them, or picking-up or dropping-off on them.

2.6 Trajectory Data Mining

In the field of data mining, trajectories of human and vehicle motions are used to discover knowledge from large-scale datasets [44]. A pattern can represent the aggregated abstraction of many individual trajectories of moving object [45]. Three major categories of patterns that can be discovered from a single trajectory or a group of trajectories [32]. These categories are:

Traffic Flow Monitoring: Application requiring classification based on monitored movement patterns such as targeted advertising, or resource allocation needs to discover traffic flow patterns [32]. The problem of maintaining hot motion paths, i.e., routes frequently followed by multiple objects over the recent past, has been investigated in recent years [46, 47]. A density-based algorithm named FlowScan which finds hot routes in a road network was proposed by [47].

Behavior Mining: In many applications, Large-scale mobile phone data with geographic information system (GIS) information is used to uncover hidden patterns in urban road usage (e.g., objects follow the same routes (approximately) over regular time intervals) [48]. Hidden patterns can be viewed as a short description of common behaviors, in terms of both space (i.e., the regions of space visited during movements) and time (i.e., the duration of movements). For instance, the trajectory of a person frequently remaining at a location from 9.00am to 5.00pm usually means that they work at that location [32].

Path Planning: Effective path planning or route recommendation is beneficial to travelers who are looking for directions or planning a trip. Historical trajectory data can reveal valuable information on how other people usually choose routes between locations [32]. For instance, T-Drive is a system that provides personalized driving directions that adapt to weather, traffic conditions, and a person's driving habits. Two versions of the T-Drive system were introduced by [5, 49]. The first version of this system [5] only suggests the practically fastest path based on historical trajectories of taxicabs. The second version of T-Drive [49] mines taxis' historical trajectories and weather condition records to build four landmark graphs corresponding to different weather and days. The system also calculates the real-time traffic according to the recently received taxi trajectories and predicts future traffic conditions based on the real-time traffic and the corresponding landmark graph.

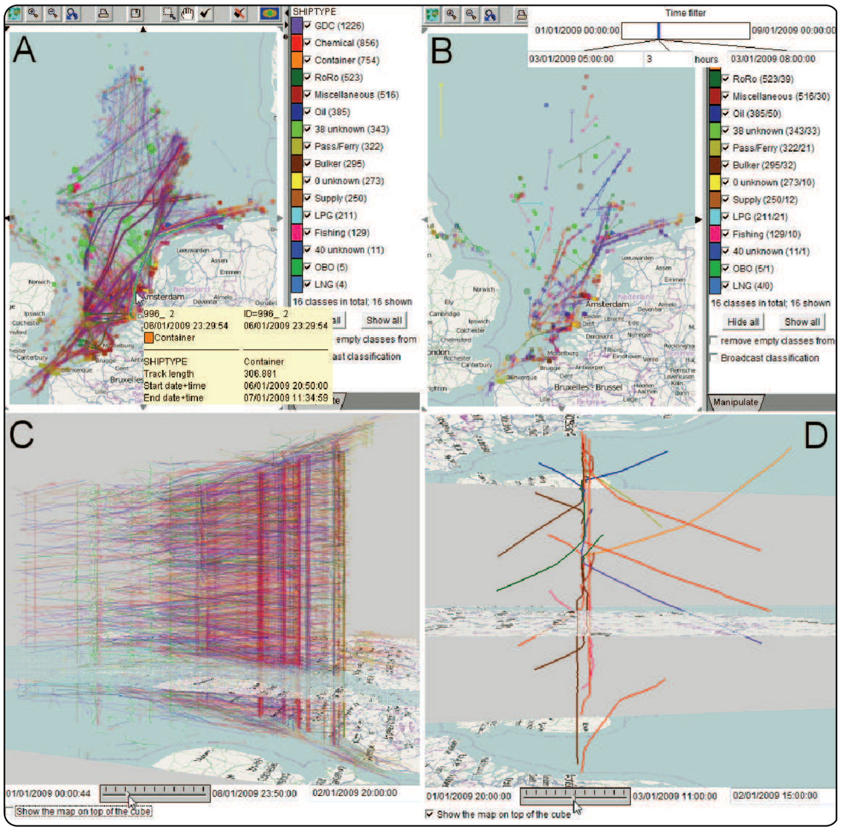
2.7 Visual Analytics of Trajectory Data

Many approaches have been proposed to visually explore the movement data [1, 50, 51]. They allow analysts to look at the data from different perspectives and complete different analytical tasks. Many of them are focused on the origins and destinations of the trajectories, such as flow maps [52], Flowstrates [53], OD maps [54], and visual queries for the origin and destination data [40]. Other work visualizes trajectories using various visual metaphors and interactions, such as GeoTime [55], TripVista [56], FromDaDy [57], vessel movement [58], route diversity [59]. Some of these approaches coordinate multidimensional visualization and map views [40,56,60]. These visualization approaches are mostly designed for specific tasks in understanding the data. Many of them are highly integrated with data mining algorithms to help users find hidden patterns.

According to [1, 50, 51], we can classify existing visual analytics approaches of the trajectory data into two categories: (1) Looking at Trajectories and (2) Looking Inside Trajectories.

2.7.1 Looking at Trajectories

The focus is on trajectories of moving objects considered as wholes. The methods support exploration of the spatial and temporal properties of individual trajectories and comparison of several or multiple trajectories [1, 50, 51]. In this category, we should take three points into consideration. The first point is how to visualize trajectories and how to interact with the representations. The second point is to use one of the



SHIPTYPE
☑ GDC (1226)
☑ Chemical (856)
☑ Container (754)
☑ RoRo (523)
☑ Miscellaneous (516)
☑ Oil (385)
☑ 38 unknown (343)
☑ Pass/Ferry (322)
☑ Bulker (295)
☑ 0 unknown (273)
☑ Supply (250)
☑ LPG (211)
☑ Fishing (129)
☑ 40 unknown (11)
☑ OBO (5)
☑ LNG (4)
16 classes in total; 16 shown

996_ 2	ID=996_ 2
06/01/2009 23:29:54	06/01/2009 23:29:54
Container	
SHIPTYPE	Container
Track length	306.881
Start date+time	06/01/2009 20:50:00
End date+time	07/01/2009 11:34:59

Time filter
01/01/2009 00:00:00 09/01/2009 00:00:00
03/01/2009 05:00:00 3 hours 03/01/2009 08:00:00

☑ RoRo (523/39)
☑ Miscellaneous (516/30)
☑ Oil (385/50)
☑ 38 unknown (343/33)
☑ Pass/Ferry (322/21)
☑ Bulker (295/32)
☑ 0 unknown (273/10)
☑ Supply (250/12)
☑ LPG (211/21)
☑ Fishing (129/10)
☑ 40 unknown (11/1)
☑ OBO (5/1)
☑ LNG (4/0)
16 classes in total; 16 shown
Hide all Show all
□ remove empty classes from
□ Broadcast classification

01/01/2009 00:00:44 08/01/2009 23:50:00 02/01/2009 20:00:00
□ Show the map on top of the cube

01/01/2009 20:00:00 03/01/2009 11:00:00 02/01/2009 15:00:00
☑ Show the map on top of the cube

Figure 5: A: An interactive map with trajectories of ships shown as lines with 10% opacity. B: Map animation through a time filter. Positions and movements of ships during a 3-hours' time interval are visible. C: An interactive space-time cube (STC) showing all ship trajectories with 20% opacity; the vertical dimension represents time. The cube is seen from the southwest. D: A STC with several selected trajectories is seen from the east [1].

clustering methods to do comparative studies of multiple trajectories. The last point is the time transformations supporting exploration of temporal properties of trajectories and comparison of dynamic properties of multiple trajectories.

Visualizing Trajectories: The common visualization method used for trajectories are static and animated maps [55, 61, 62] and interactive space-time cubes [50, 63, 64] with linear symbols representing trajectories. These methods provide some interaction techniques as well as illustrated in figure 5.

From the example in figure 5, we noticed that multiple trajectories might suffer from visual clutter and occlusions. The clutter can be reduced by decreasing the opacity of the symbols, but the occlusion problem remains. Beside occlusion, space-time cubes (STC) suffers from a distortion of both space and time due to projection [1, 50, 51]. In general, visualizing trajectories in three-dimensional space (e.g., in the air or underwater) are harder than on a surface. Different researchers proposed different solutions for these drawbacks. For instance, Willems et al. suggested a display called Trajectory Contingency Table that supports the exploration of attributes of trajectories [65]. Interaction with any visualization methods helps the user to filter the data by hiding some parts and focus on others. Common interaction techniques include [1, 50, 51]:

1. Manipulations of the view (zooming, shifting, rotation, changing the visibility and rendering order of different information layers, changing opacity levels, etc.).

2. Manipulations of the data representation (selection of attributes to represent and visual encoding of their values, e.g., by coloring or line thickness).

3. Manipulations of the content (selection or filtering of the objects that will be shown), and interactions with display elements (e.g., access to detailed information by mouse pointing, highlighting, selection of objects to explore in other views, etc.).

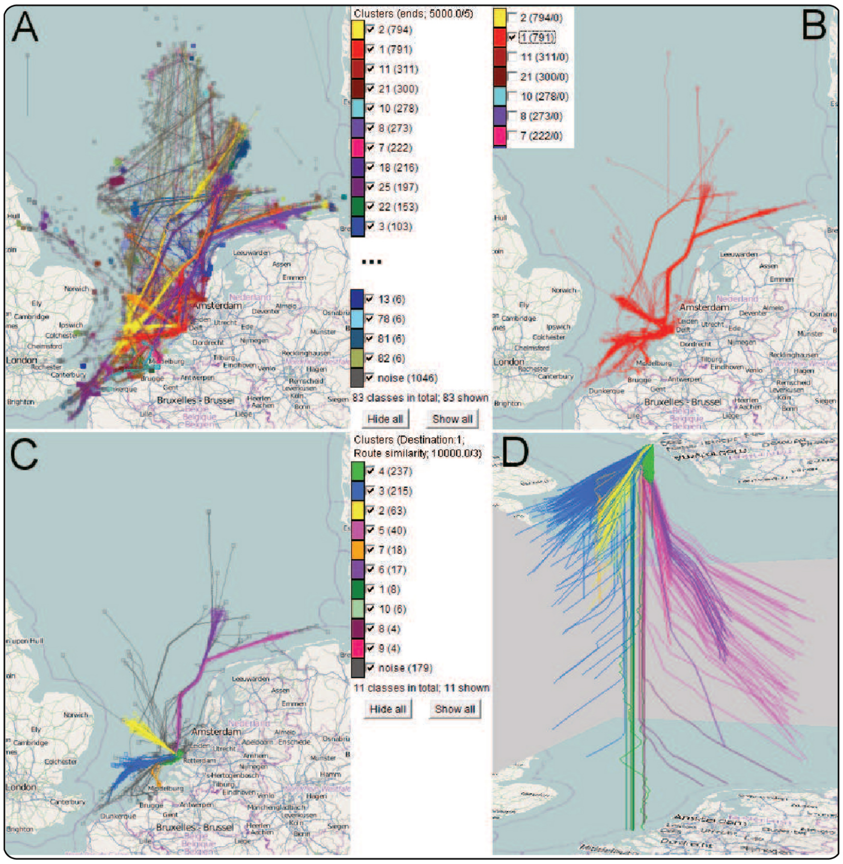
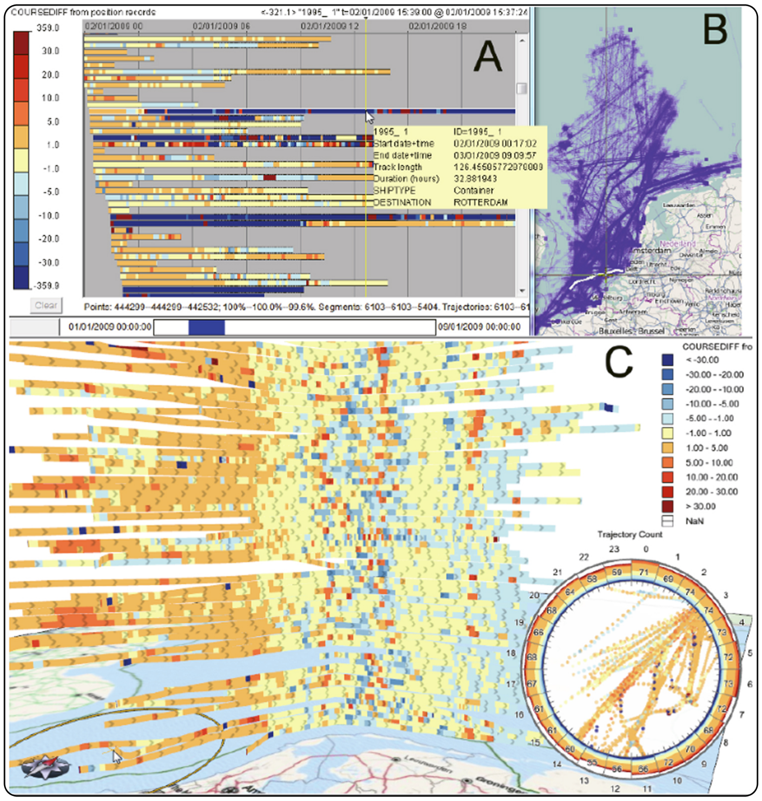


Figure 6: Interactive progressive clustering of trajectories. A: The ship trajectories have been clustered according to the destinations. B: One of the clusters is selected. C: Clustering by route similarity has been applied to the selected cluster. D: The clusters by route similarity are shown in an STC; the noise is excluded [1].

Clustering Trajectories: To handle large amounts of data in visual analytics, clustering methods are used to do comparative studies of multiple trajectories. Existing clustering methods supporting not only visual inspection but also an interactive refinement of clustering results. Trajectories of moving objects are quite complex spatio-temporal constructs. Their potentially relevant characteristics include the geometric shape of the path, its position in space, the lifespan, and the dynamics, i.e., the way in which the spatial location, speed, direction and other point-related attributes of the movement change over time. Rinzivillo et al. proposed different distance functions to address different properties of trajectories and introduced the procedure of progressive clustering that allows analysts to combine these distance functions in the process of analysis. The analysis is done in a sequence of steps. In each step, clustering with a single distance function is applied either to the whole set of trajectories or to one or more of the clusters obtained in the preceding steps. This helps the user to build a comprehensive understanding of different aspects of the trajectories [66]. Andrienko et al. suggested several distance functions suitable for clustering of trajectories [67]. The procedure of progressive clustering is illustrated in figure 6. The drawback of clustering tools is that it can only work with the data loaded in the computer random access memory (RAM) and are therefore not scalable to very large datasets. Andrienko et al. suggested a way to overcome this problem by supporting the analyst by interactive visual and computational tools in the process of creating, testing, and refining a classification model for assigning trajectories to the clusters after defining clusters of trajectories on the basis of a subset of trajectories [68]. Then the classifier is used to supplement the clusters with new trajectories, which are loaded from a database by portions fitting in the main memory [1, 50, 51].

2.7.2 Looking Inside Trajectories

This approach focus on visualizing and analyzing the variation of movement characteristics and other dynamic attributes associated with trajectories. Trajectories are considered at the level of segments and points. Visualization methods support detecting and locating segments with particular movement characteristics or derived attributes



COURSEDIFF from position records

<-321.1> "1995_ 1" t=02/01/2009 15:39:00 @ 02/01/2009 15:37:24

1995_ 1 ID=1995_ 1
Start date+time 02/01/2009 00:17:02
End date+time 03/01/2009 09:09:57
Track length 126.46505772070009
Duration (hours) 32.881943
SHIPTYPE Container
DESTINATION ROTTERDAM

Points: 444299--444299--442532; 100%--100.0%--99.6%. Segments: 6103--6103--5404. Trajectories: 6103--61

01/01/2009 00:00:00 09/01/2009 00:00:00

Figure 7: A: A time bars display shows temporal variation of values of a dynamic attribute within trajectories. The bars represent trajectories; the attribute values are color-coded. B: A trajectory selected in the time bars display is highlighted on a map. The crossing of the vertical and horizontal lines marks the spatial position corresponding to the position of the mouse cursor in the time bars display. C: A trajectory wall represents trajectories by segmented bands stacked on top of a cartographic map [1].

and sequences of segments representing particular local patterns of individual movement. Derived attributes may express instant, interval, and cumulative aspects of trajectories [69]. Instant movement characteristics include instantaneous speed, direction, acceleration (change of speed), and turn (change of direction). Interval characteristics are computed for time intervals of a chosen constant length before, after, or around a given time moment. They include traveled distance, displacement, average speed, as well as statistics of the instant characteristics. Cumulative characteristics are computed for the interval from the start of the trajectory to a given time moment or for the remaining interval to the end of the trajectory [1]. The common way to visualize position-related attributes is by dividing the lines or bands representing trajectories on a map or in a 3D display into segments and varying the appearance of these segments. The user can apply clustering methods to join similar segments with similar characteristics. This allows the user to detect and locate different movement behaviors, for instance, day migration and night migration. Different coloring of line segments shows these behaviors on a map. Due to display clutter and occlusions, visualizing dynamic attributes of trajectories on a map or in STC may be ineffective, especially when trajectories or their parts overlap in space. Therefore, position-related dynamic attributes are often visualized using additional displays [1,50,51]. To avoid overlapping, the idea of stacking is used by [60]. Trajectories are represented by bars or bands stacked within a temporal or spatial display. The bars or bands are divided into segments, which are colored according to values of some dynamic attributes. Examples are shown in figure 7(A) and (C).

In this thesis, a set of visual exploration tools have been implemented to be utilized in the presented systems for visual exploration of trajectory data and help users to look at and inside trajectories. All presented systems are designed as a map-centric application where the map plays the role as a canvas in the middle for users to define the query regions and show the results with a set of coordinated views. In SemanticTraj (Chapter 3), a set of elements such as labels and meta-summary have been added to the map view to reveal the semantic information of the acquired taxi trips over the map. The paths of taxi trips are drawn with the same opacity, so that the color density on a road reflects how often it is used by these trips. Users can change the drawing attributes such as color, line width, or opacity for their favorite appearance. They can also show the pickup and drop-off locations as map markers. These display options are used to decrease visual clutter and occlusions. In QuteVis (Chapter 4), the map view supports touch-based selection of multiple streets, regions, or paths and plays the role as a canvas in the middle for users to sketch on. Users can select any region/street to study its traffic behavior and weather in a specific time. The map view can be shown in different styles and it supports smooth zooming and panning operations. Furthermore, Map Comparison allows users to observe and compare the traffic behaviors at different times on a multi-map view. This view is coordinated with the major map canvas. Fully functioned map operations on either the canvas map or any of the small maps are well synchronized. Therefore, users can study details of any specific area of the city. In TrajAnalytics (Chapter 5), an interactive map with different visualization layers are provided that users can toggle between them. Trajectories or trips are directly shown as connected polylines on the map. Users can select speed or flow to be visualized on each trip, which is represented by the line width and the color. The start (pickup) and end(drop-off) locations are visualized as red and green points, respectively.

CHAPTER 3

SemanticTraj: A New Approach to Interacting with Massive Taxi Trajectories

With the development of sensor-based technologies, more and more trajectory data can be collected, which represent the mobility of a diversity of moving objects, such as people, vehicles, and animals. One kind of this data is taxi trajectory data that can help experts to carry out knowledge discovery in transportation and urban planning. However, existing visual analytics systems of trajectory data require users to select and brush geospatial regions on a map to interact with GPS points and trajectory paths. The analyst needs to learn the system and use complex operations to complete some straightforward tasks and analysis. This chapter presents a visual analytics system called SemanticTraj, a new method for interacting, managing and visualizing taxi trajectory data in an intuitive, semantic rich, and efficient means. By converting the trajectory data into text description data, the trajectory data is analyzed through flexible semantic queries over a text search engine. Translating trajectory data into textual descriptions provides high-level information that makes it easier for users to summarize the results of the analysis. The query method through a text search engine does not require analysts to conduct professional training because most people are familiar with the process of phrasing queries as keywords to dig up information through search engines such as Google and Baidu.

3.1 Introduction

Advanced sensing technologies and computing infrastructures have produced a variety of trajectory data of humans and vehicles in urban spaces. Taxi trajectory data records realtime moving paths sampled as a series of positions associated with vehicle attributes over urban road networks. Massive trajectory data contains abundant knowledge about a city and its citizens which has been widely used in urban computing [44]. Exploratory visualization systems are demanded to study taxi trajectories with efficient user interaction and instant visual feedback. However, users often need to select, brush, and filter regions on maps to interact with GPS points and trajectory paths. Complex operations are needed to complete some straightforward tasks. We give two scenarios as examples:

- Scenario 1: A shopping mall has a plan to open shuttle buses for their customers. Its manager wants to investigate where and when visitors take taxis to the mall.

  Task 1: "For the taxi trips arriving at the shopping mall, what are their major pick-up locations? "

- Scenario 2: Two witnesses reported a criminal suspect taking a taxi passing South Street and North Street between 3pm and 3:20pm. A policeman wants to find suspicious taxi paths.

Task 2: "What are the taxi trips passing South Street and North Street between 3pm and 3:20pm? What are the other streets/POIs they visited?"

For Task 1, the manager needs to:

1. Select a region enclosing the mall on the map to display drop-off points.
2. Select a given time period.
3. Brush to filter those points related to this mall inside this region.
4. Display the corresponding pick-up points of passenger trips on the map.
5. Observe the map to find hot locations with a high density of pick-up points.
6. Find the streets and POIs of these hot locations. Then, the candidate streets/POIs for shuttle stops are found.

For Task 2, the policeman needs to:

1. Select a region around North Street on the map.
2. Select the time period from 3pm to 3:20pm in a time selection tool (e.g., a slider).
3. Brush to select all GPS sample points residing on North street.
4. Display the GPS points in the same taxi trips with the selected points from (3).
5. Repeat steps (1)-(3) to select points on South Street. After these steps, the required taxi trips are found.

6. Find other streets they also passed.

Apparently, these interaction steps require non-professional users to be trained for the map-based operations. Advanced trajectory filtering tools [1] are designed where users can visually operate specific filters such as lens on maps (e.g., [70,71]). The learning curve may deter some domain or public users. Can we help the users complete the query tasks in an alternative way, akin to querying over keywords in a search engine? For example, can the policeman simply input ”Select trips on North Street AND South Street AND [3pm, 3:20pm]” for Task 2? Then, the resulting trips are visualized on maps together with intuitive text labels, or through a short textual summary, so that the policeman can immediately find the requested street names. Such interactions are natural and easy to conduct for city residents and general practitioners.

In this chapter, we propose SemanticTraj to enable such new interactions for studying taxi trajectory data. It supports non-professional users to retrieve taxi trips by giving street or POI (Point of Interest) names, speed descriptions, and their boolean combinations. These terms are referred to as *semantic information* in this chapter. Then, users can visually study the results with a set of visual representations enhanced by the semantic information.

Most people are familiar with the process of phrasing queries as keywords to dig up information. In SemanticTraj, the experience is extended to taxi trajectory data, leading to a new paradigm of the interaction between users and trajectory data, in addition to selection or brushing over maps. In particular, taxi trajectory data is transformed into “taxi documents” by a process of “textualization”, which projects a GPS point to a street or POI name, and maps a numerical speed value into a descriptive term (e.g., slow, normal, fast). The semantic information gives direct cues of geographical and transportation information for users living in the city. Users are allowed to retrieve matching data from keywords of interest using text search engine (Apache Lucene in our experiments). Moreover, the data management and query performance is highly efficient in the engine with well-designed index schemes. Therefore no specific data structures are needed to manage trajectory data over grids and trees. Interactive visualization is well supported by the fast computation over big taxi data. Furthermore, users can flexibly query taxi documents by wildcard, boolean, fuzzy, proximity and range search with semantic hints. To the best of our knowledge, our approach is the first to employ text search engine in managing and querying taxi trajectories.

The semantic information of query results is integrated with visual representations and interactions to promote easy understanding. First, textual labels are added on taxi traces and streets in the map view. It is well known that text/title information is the most effective visual cues for users [72]. Users reading the labels can immediately get ideas about the search results. Second, a meta-summary of the taxi trajectories is provided to present their mobility features. Each individual trajectory also has a meta-summary of its own behavior. Users can directly refine the query results by interacting with keywords in the meta-summaries.

In summary, our major contribution is a novel approach for effective interaction with big taxi data. First, our approach uses textualization to map trajectory data to taxi documents. Second, our approach manages and queries the taxi documents in a text search engine. Third, we have developed a fully working prototype for this approach. It allows users to visually explore trajectories with enriched semantic information. A set of usage scenarios are presented to illustrate its usability.

3.2 Related Work

Our project is related to the mining and visual analysis of trajectory data. It also provides a new way of spatial data management. Meanwhile, semantic information is utilized in our system. Therefore, we introduce the related work in these three directions.

3.2.1 Trajectory Data Mining and Visual Analysis

Trajectories of human and vehicle motions are one of the most important data used in urban data analytics [73]. Utilizing the trajectory data has been categorized into three main categories: the study of the collective behavior of a city's population, the traffic flow, and the operators (e.g. drivers) [74]. For instance, vehicle trajectory data has been used in traffic monitoring and prediction [75], personalized route recommendation [76], urban planning [77], driving routing [5, 49], extracting geographical borders [78], service improvement [79], and energy consumption analysis [80]. Public transit trajectories are used in bus arrival time predictions [81], user's transportation mode inference [81], and travellers' spending optimization [82].

The analytical tasks are supported by visualizing the spatial-temporal data [1], where the techniques combine map-based displays and information visualization techniques [83]. Andrienko et al. [84] transformed GPS-tracked trajectories into aggregated flows between areas to depict important moving patterns over a city. Wang et al. [85] explored traffic data recorded on sparsely distributed cells in a city. Liu et al. [86] presented a visual analytics system of route diversity. Wang et al. [60] presented an interactive analysis system for urban traffic congestion. Pu et al. [87] visually monitored and analyzed complex traffic situations in big cities. Wang et al. [88] presented a visual reasoning approach for the data-driven transport assessment on urban roads. It supports dynamic query of traffic situations within a bi-directional hash structures on a spatial grid. Ferreira et al. [89] allowed users to visually query taxi trips on spatio-temporal constraints, which was designed only for queries over the pick-up and drop-off pairs.

3.2.2 Spatial Trajectory Management

Spatial indexing mechanisms such as minimum bounding rectangles (MBR) and hierarchical trees [90] are often used over a spatial division. Indexing trajectory data requires extra work in maintaining the links between sampling points. R-tree [30] and its extensions (e.g., [32]) build MBRs based on clusters of trajectory segments. Other methods maintained the linkage of samples and time indexes over fixed grid cells (e.g. over quadtree) (e.g. [35, 41]). Sample points can be hashed [88, 91] on a cell to achieve a fast query. Our method is distinguished from them by managing trajectories as textual documents and performing queries over the documents where no spatial MBRs or trees are needed. Managing massive trajectories with traditional methods requires careful design and optimization of computer systems. For instance, SETI [35] and TrajStore [41] store trajectory segments that are spatially close on the same disk partitions and memory pages to accelerate query processing. Our method instead utilizes the data management capability of the text search engines, where the system level data optimization is well-developed including memory and disk use, caching, encoding and compression [92].

3.2.3 Semantic Trajectory Management

Many approaches enrich trajectories by associating locations with semantic meanings such as labelling user activities or place names (e.g., Restaurant, Theater). A survey paper well discussed current approaches for modeling and analysis of semantic trajectories [93]. The concepts of semantic subtrajectories, points, geographical places, events, etc. were proposed in a concept model [94]. A knowledge-based framework was also presented for semantic enrichment and analysis of moving data [95]. Activity trajectory similarity query (ATSQ) [96] supported searching activities through a specific grid-index combined with inverted index of a predefined set of activities. A hybrid index structure (GIKI) further organized the trajectories with both spatial and textual similarity [97]. These methods partitioned trajectories into segments by considering spatial, temporal and semantic features. They were designed for finding trajectories with predefined activity terms in regions. Our aim is different which is to visualize trajectories by exact street names, status, and time ranges. Streets become the major query units instead of regions and the search is enabled by text search engine after mapping GPS points to street names. Su et al. [98] generated a short descriptive text of a trajectory by selecting features of trajectory segments. Annotated semantics was also used for automatic insight externalization [99]. Chu et al. [100] studied massive trajectories as a document corpus based on similar transformation where LDA topic modeling is employed. However, their work found hidden themes of population movement and did not provide query functions to retrieve data.

3.3 Overview

3.3.1 Taxi Trajectory Data

Massive trajectory datasets are acquired by taxis traveling over cities. A taxi trajectory records consecutive samples at an interval of a few seconds in a given time period. Each sample includes the GPS location (longitude and latitude), vehicle ID, speed, time, occupancy status, direction, and possibly other attributes. A taxi trip consists of those samples when the taxi was hired by customers. It then refers to one taxi service trip for passengers.

3.3.2 Visual Exploration Tasks and Our Approach

SemanticTraj allows users to retrieve and visually explore a taxi trajectory dataset such as:

- *Task1: Taxi trips passing a street in a given time period;*

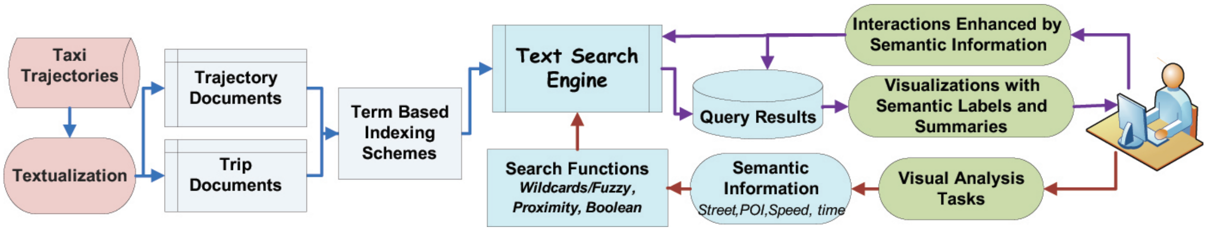


Figure 8: SemanticTraj framework of processing, searching and visualizing taxi trajectory data.

- *Task2: Taxi trips passing multiple streets in given time periods with different logic conditions;*
- *Task3: Taxi trips with single or multiple POIs;*
- *Task4: Taxi trajectories passing given streets and POIs;*
- *Task5: Taxi trajectories with specific behaviors in travel speed (e.g., slow, fast, change from slow to fast).*

Figure 8 shows the framework of SemanticTraj. First, raw taxi trajectories are processed and transformed into taxi documents through “textualization”, which projects geographic locations to streets, and maps speed values to descriptions. A taxi document can be a trip document consisting of the streets a taxi passed with passengers. It can also be a trajectory document including all streets a taxi traversed in a time period. The trajectory dataset of all taxis is converted to a taxi document corpus. Indexing schemes are designed for the corpus to manage the taxi documents using a text search engine. Then with an interactive visual system, users query the data by a variety of convenient text search functions, and the query results are visually presented. Rich semantic information is conveyed through semantic labels and meta-summaries, which enable easy and fast understanding of the query results. Users can operate on the visual interface to refine and explore the results for insight discovery.

3.4 Textualization of Taxi Trajectories

We refer the process of converting an attribute in the raw data to a text term as "textualization". In general, the aim is to create semantic-rich forms by externalizing associated contextual information of the original data. Domain knowledge and user input are then incorporated into the transformed data. We transform massive taxi data in the following ways:

1. Each geographical location of latitude and longitude is mapped to the street name it resides. Such a transformation process is implemented through road matching to find the closest street of a given position where we follow the implementation in [100]. As a result, a sequence of GPS points over a trajectory are mapped into a taxi's traversed streets. Such a location mapping can also be applied by mapping the GPS point to the city's POIs or regions.

2. The associated attributes are transformed to semantical information. The numeric travel speed is converted to a descriptive term as:

    - speed$<$0.01Km/h $\mapsto$ *Stop*,
    - 0.01Km/h$\leq$speed$<$20Km/h $\mapsto$ *Slow*,

- 20Km/h≤speed<60Km/h ↦ *Normal*,
- 60Km/h≤speed<100Km/h ↦ *Fast*,
- 100Km/h≤speed ↦ *Very Fast*.

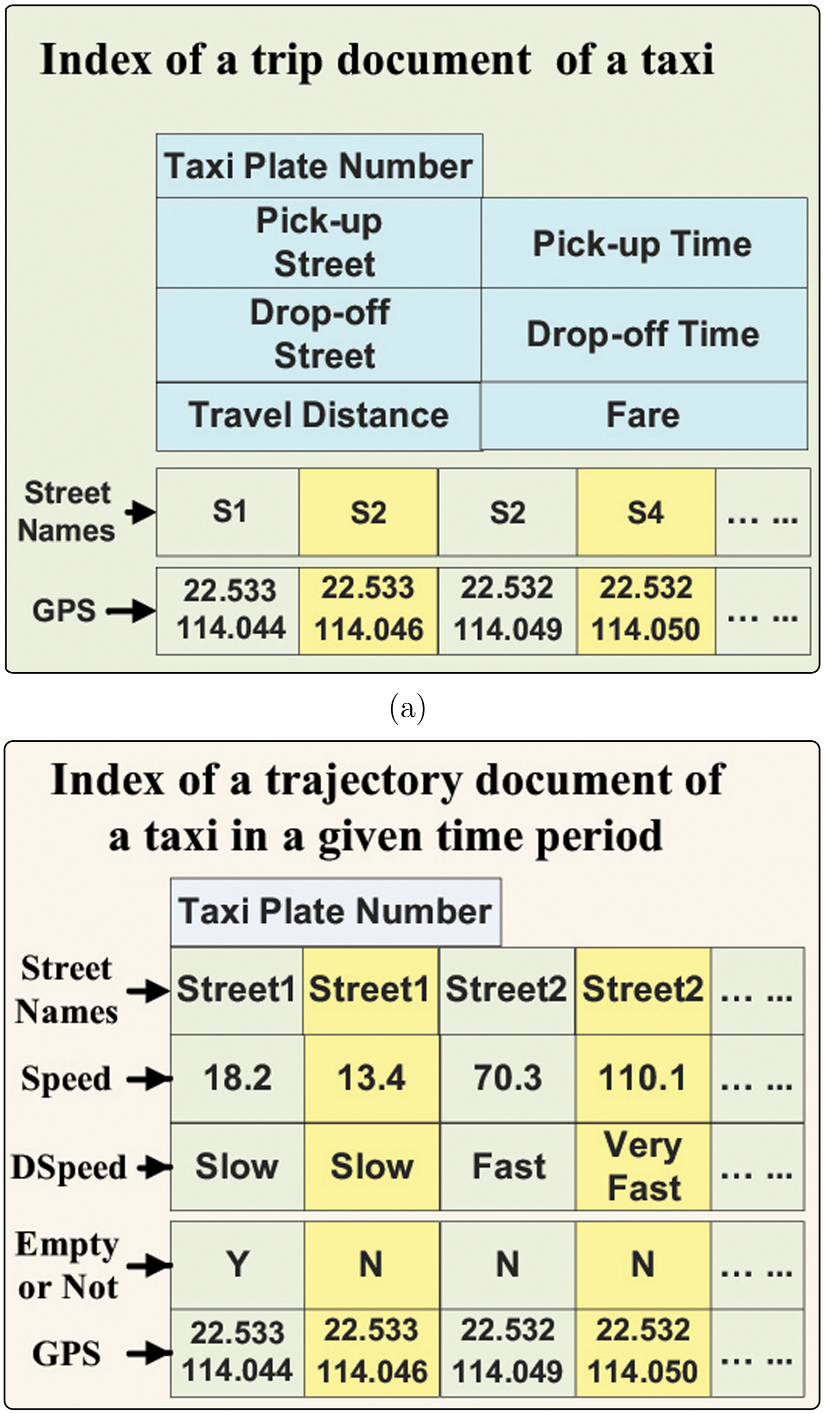
In this way, text search can quickly find specific behavior described by the terms, such as identifying speeding drivers with *Very Fast*. Such a mapping is currently predefined after consulting local drivers. It can be configured by domain users based on their specific knowledge and aims in practical applications.

3.5 Taxi Documents

By introducing semantic meanings into the trajectory data, we create *taxi documents* from massive trajectories on which text based searches are applied. Taxi documents enable us to manage and search big trajectory data using text search engines. In an engine, each document of a corpus is represented by an *index* consisting of multiple fields, where each field stores a term. A term usually refers to words but may also be a date, number, etc, whereas they are represented in a textual form. The query criteria combine a set of given terms in the corresponding fields. Next we discuss the trajectory and trip documents in Sec. 3.5.1 and 3.5.2, respectively. Then we summarize the search functions in Sec. 3.5.3.

3.5.1 Trip Documents

Fig. 9a shows an index of one *trip document* related to one trip of a taxi's service to passengers. The index includes the fields of Pick-up and Drop-off streets and Times,



Index of a trip document of a taxi

Taxi Plate Number	
Pick-up Street	Pick-up Time
Drop-off Street	Drop-off Time
Travel Distance	Fare

Street Names →	S1	S2	S2	S4	… …
GPS →	22.533 114.044	22.533 114.046	22.532 114.049	22.532 114.050	… …

(a)

Index of a trajectory document of a taxi in a given time period

Taxi Plate Number

Street Names →	Street1	Street1	Street2	Street2	… …
Speed →	18.2	13.4	70.3	110.1	… …
DSpeed →	Slow	Slow	Fast	Very Fast	… …
Empty or Not →	Y	N	N	N	… …
GPS →	22.533 114.044	22.533 114.046	22.532 114.049	22.532 114.050	… …

(b)

Figure 9: Indexing of taxi documents. (a) An index of a trip document. (b) An index of a trajectory document.

and the attributes associated with this taxi trip such as its Travel distance and Fare. The field of Street Names stores the sequence of streets the taxi traveled. In particular, the field of GPS is used to store the sequence of GPS points, which is used to draw the trajectory in visualization. If the whole trip path is not considered, it might only store the pick-up and drop-off GPS locations. A set of such indexes are generated for a collection of trip documents in a time period (e.g., one day). These indexes are stored in one index file, C. Multiple index files can be queried simultaneously to search trips in multiple periods (days). The analysis tasks in Sec. 3.3.2 are conducted by text search engine using specific query conditions. Two examples are:

- *Task1 Question*: What are the pick-up and drop-off locations for the trips passing S?

  Query Condition: {*S in Street Names in C*}.

- *Task2 Question*: What are the trips picking up passengers at S_1 during T_1 AND dropping them off at S_2 during T_2?

  Query Condition: {{*S_1 in Pick-up Street AND Pick-up Time in T_1*} *AND* {*S_2 in Drop-off Street AND Drop-off Time in T_2*}}.

If the tasks are about a POI P (*Task3*), we first find the pre-computed street segments close to P, and then apply similar street-based queries. More complex queries with combined constrains and logic operators can be implemented in a similar manner. For example, we can add {*Fare:[30 TO 50]*} in the above queries to get trips earning between 30 and 50.

3.5.2 Trajectory Documents

Fig. 9b shows an index of one *trajectory document* related to one taxi's trajectory in a given time period T. The field of Street Names includes the sequence of streets the taxi traveled during T. The field of DSpeed stores the description terms (e.g., slow, fast) of textualized numeric speeds. The Occupation Status is represented by Y or N at each point.

Using such an index for each trajectory, a document corpus is created and stored in one index file $C(T)$. Using the length of T as 10 minutes, a set of 144 indexing files jointly represent the data in a whole day. Two example tasks they support are:

- *Task4 Question*: Given all the taxi trajectories passing S during T, what is their average speed? How many are occupied?

  Query Condition: $\{S$ *in Street Names in* $C(T)\}$.

- *Task5 Question*: What are the taxis whose speed shows an abrupt speed-up from slow to very fast (across normal) during T?

  Query Condition: $\{$*Slow VeryFast* ∼*1 in DSpeed in* $C(T)\}$.

Here a proximity search *Slow VeryFast* ∼*1* is utilized to find taxis whose speed suddenly changes from slow to very fast (see the search functions and syntax in Sec. 3.5.3). In a similar manner, we can search a long phrase *Slow Slow ... Slow* whose appearance may reflect potential traffic congestion, or search *VeryFast VeryFast ... VeryFast* for finding taxis who may have excessive speeding. Another interesting example is to search consecutive street names *Road1West Road2North*, which reflects a left turn. The returned trajectories may violate the traffic law if such a left turn is not allowed. It is the textualization and text search engine which provide a novel tool for users to conduct these jobs.

The indexing scheme of trajectory documents is different from that of trip documents, because of their different analytical focuses. Users study trajectories to learn traffic situations, while they are mostly interested in taxis' behaviors when investigating trips. Due to such differences, we use the time-period based indexing approach $C(T)$ for trajectory documents. Users can query the trajectories in multiples of 10 minutes. Surely T can be further reduced for more refined data search which increases the number of index files. The benefit is that this indexing scheme supports quick data feedback and visualization. Instead, if one big index file for 24 hours is used, we will need to read through all the returned documents and filter the results for a given small time period, where the extra computation reduces time efficiency. In summary, the design of indexing schemes of taxi documents can vary according to the analytics tasks to be performed.

A text search engine maintains and manages an inverted index structure based on the indexing scheme of documents. It is designed to efficiently construct, compress, and manipulate all the document indexes, as well as the inverted index structures (see [92] for details). For example, the Apache Lucene engine is deliberately tuned for scalable, high-performance indexing [101]. The index size is roughly 20-30% of the size of documents indexed. It is also optimized over memory and disk to handle data indexing and queries. Utilizing the engine gives us the power and flexibility of data management and interactive retrieval for efficient visualization.

3.5.3 Flexible Search Functions

Taxi documents enable a new means of interaction between users and taxi trajectories. Users can utilize flexible search functions combining street/POI names and speed descriptions. Boolean, wildcard, fuzzy, proximity and range queries are supported over taxi documents.

- *Boolean Query:* A boolean query combines multiple queries of individual terms with boolean conditions. It allows users to quickly conduct a task (e.g., *Task2*) which retrieves trajectories by multiple conditions.

- *Range Query:* A range query matches the documents whose terms fall into the supplied range. For example, we can query taxi pick-up time between *[07:00:00 TO 10:00:00]*, or query fare of taxi trips larger than 30.

- *Wildcard and Fuzzy Query:* A wildcard query supports users with single and multiple missing characters in query terms. Users can query a street without clearly remembering the name in full, such as *north** for *north ave* or *north str.* A fuzzy query finds any matched terms with Levenshtein Distance. For example, *north*∼ can find terms like lorth or nortn.

- *Proximity Query:* A proximity query supports matching words within a specific distance in text. For example, a query of "*slow veryFast ∼1*" finds those documents where *slow* and *fast* happen within two consecutive words. The query helps users immediately find a speed change event from *slow* to *fast* beyond *normal*.

These functions flexibly support visual analysis tasks, which may require complex operations in traditional region queries with geo-spatial indexing. Table 1 summarizes the query syntax. The complete syntax and examples can be found in [102]. Users with knowledge of Boolean operations need a small effort to become familiar with the query language. For novice users, we further design a form to fill in the query fields without phrasing the queries.

The indexing model also facilitates ranking query results by customized scoring. Taxi trips can be ranked by the term frequency of a street name (S), which implies how many samples are on S in a taxi trip. This score reflects a trip spending lots of time on S. The ranking can also be completed by the length or the fare of a taxi trip, so that users immediately get the longest trip or the trip achieving the highest fare. The ranking scheme is very useful in visualization tasks to provide users preferred information of query results.

3.6 Data Processing and Query Performance

We use the taxi trajectory data of Hangzhou, China. Hangzhou has a population of about 2.5 million and taxi is one of its major transportation methods. The dataset of a whole month (Dec. 1-31, 2011) has a raw size of 77GB acquired by 8,120 taxis. Each day there is a raw data size around 2.5GB in the format of raw GPS sample points and

Table 1: Syntax of Queries.

Type	Syntax	Description
Term Query	t	Find term t
Phrase Query	”$t_1 ... t_n$”	Find ordered terms in a given phrase
Wildcard Query	s_1*s_2	Find terms leading by string s_1 and ending with s_2
Fuzzy Query	$t \sim$	Find term t approximately
Proximity Query	”$t_1 ... t_n$” $\sim d$	Find phrase within distance d
Field Query	f:Q	Find a query Q in field f
Range Query	f:[t_1 TO t_2]	Find terms lexicographically between t_1 and t_2
Boolean Query	Q_1 OR(AND,NOT) Q_2	Find Q_1 or (and, not in) Q_2

Table 2: Query performance on trip documents.

Time Period	Index Size	Indexing Time (sec)	Q1 Time (sec)	Q1 Hits	Q2 Time (sec)	Q2 Hits	Q3 Time (sec)	Q3 Hits
One Day	297MB	9	0.15	25k	0.14	6k	0.19	39k
One Week	1.00GB	336	0.21	90k	0.19	19k	0.32	137k
One Month	6.55GB	2,012	1.75	590k	1.32	141k	2.85	918k

associated attributes (ID, speed, time, etc.). Data preprocessing consists of several steps: (1) Remove erroneous (duplicated) records. (2) Add street names and speed descriptions to the points after performing textualization (which needs map matching). (3) Create trajectories from the points with the same taxi ID. Here the points should be sorted by time since in a trajectory they need to be stored in sequence. For fast computing, we sort and generate trajectory segments for each small time interval (10 mins) of a day, and then join these segments for complete daily trajectories. (4) Find trips from the trajectories according to the occupancy status. The total processing time is around 9 hours for the dataset. This speed can be accelerated by using an advanced map-matching method and parallel processing of the points and trajectories. In results, the size of the trajectory documents is about 38GB, in which 10% is used for names. The size of the trip documents is around 8.5GB with 6,734,497 trips. Next we show the performance of name queries using Lucene with this dataset tested on a 64bit Windows 7 workstation (Intel Xeon E5620 2.40GHz, 24GB, 1TB).

Three queries are applied to the trip documents: (Q1) Search trips passing Shangtanglu street; (Q2) Search trips passing Shangtanglu AND Zhonghegaojia streets; (Q3) Search trips passing Shangtanglu OR Zhonghegaojia streets. Table 2 is the query performance on the trip documents for one day, one week, and one month period. It shows that the query time on the whole month data is very fast at 1.75 seconds for 590k hits (Q1), 1.32 seconds for 141k hits (Q2), and 2.85 seconds for nearly 0.9 million (918k) hits (Q3). These times are proportional to the number of hits. When users query one day (or one week), which are the cases for most analysis tasks, the queries can be done within 0.5 seconds. The indexing is fast, e.g., one month data is completed in around 22 minutes. Here the size of the index file is 6.55GB, which is 21% smaller than 8.5GB of the raw trip documents. This is achieved by the data compression of the engine.

We further conduct Q1-Q3 queries on the trajectory documents. The performance is shown in Table 3 for different time periods. Please note that we use each index file for a ten-minute period as discussed before. The query performance is fast, at less than 1 second for one day/week data. For one month data that has 23GB index files, the queries can be done in less than 5 seconds (i.e., 4.69 seconds for Q3 with nearly 2 million hits). The index size of 23GB is about 40% smaller than the raw trajectory documents at 38GB due to the compression.

Table 3: Query performance on trajectory documents.

Time Period	No. of Files	Index Size	Indexing Time (sec)	Q1 Time (sec)	Q1 Hits	Q2 Time (sec)	Q2 Hits	Q3 Time (sec)	Q3 Hits
One Day	144	794MB	77	0.17	38k	0.21	3k	0.28	63k
One Week	1008	4.98GB	840	0.89	257k	0.65	19k	1.0	427k
One Month	4464	23GB	3,951	3.66	1,169k	2.14	84k	4.69	1,921k

Table 4: Query performance on proximity and fuzzy queries.

Time Period	Proximity Query		Fuzzy Query	
	No. of Hits	Time (sec)	No. of Hits	Time (sec)
One Day	5k	0.24	27k	0.25
One Week	16k	0.29	96k	0.36
One Month	118k	0.69	630k	1.03

Table 4 shows the performance of proximity and fuzzy queries. A proximity search *"shangtanglu zhonghegaojia"∼10* and a fuzzy search *shangtanglu∼0.8* are applied over trip documents for one day, one week, and one month data, respectively. The table shows that these queries are completed with fast performance too.

3.6.1 Comparison with Region Query

In our approach, users find trajectories or trips by simply giving street/POI names or speed descriptions. It is different from the conventional method which provides spatial regions to find trajectories. The region queries are usually built up on spatial databases where the hierarchical structures, (e.g., R-trees, B-trees, K-d tree) of spatial cells. Please note that the region query model cannot easily implement queries by street names. It needs to conduct geospatial region search and filter the results according to the geometry of the streets. For example, an approximate method [88] used a set of small cells on the surface of a street to support users brushing a street for querying. Each cell was handled as a region to complete the query. This approach cannot easily handle name queries, especially for search conditions involving multiple streets. On the other hand, our model also cannot directly answer queries given a geospatial region, since extra computing is needed to find the streets inside the region. In the situations that users do not know street names, the map-based query is still needed. In summary, the two query models are *complementary* to each other, which can be combined to fulfill various visual query tasks of trajectories.

We compare our performance with the existing visual system of taxi data by Wang et al. [88]. Following their method, we first find a set of regions on the surface of the street, Shangtanglu, used in Q1 above. Then we perform region queries using their method which combines spatial trees with hash tables for indexing trajectories. Wang's system uses 712 seconds to create indexing from trip files of one week. Then Q1 costs 0.53 seconds. In comparison, Lucene is faster using around 336 seconds for indexing and 0.21 seconds for the query. Note that Wang's system cannot handle the whole month data in one workstation, because it uses a great amount of memory in creating spatial data structures. Moreover, its region query does not support searches with two street names by logic operations, such as Q2 and Q3. Extra join and filter algorithms are needed.

3.7 Visualization System of SemanticTraj

3.7.1 Design Rationale

Trajectory data dynamically evolves over geospatial-temporal dimensions. Users exploring the data conduct interactions over both spatial and time dimensions. In the introduction we have exemplified complex interactions people need to conduct for different tasks. The major goal of SemanticTraj is to facilitate an easier and more intuitive way for users interacting with trajectory data, which complements to map-based visualizations and interactions. Therefore, the design of SemanticTraj focuses on

- Helping users easily externalize their ideas of data queries: In SemanticTraj, the visual interface allows direct input of semantic names and terms as in a familiar text search interaction.

- Promoting prompt understanding of query results: The query results are shown on a map with geographic context with zooming and panning operations. Visually studying them, such as finding what streets they traversed, is often confounded by cluttering while drawing many trajectories on road network. Semantic information, including text labels and meta-summaries, is then added to the visualization. The reason is that text is a very effective tool to enhance user understanding. Text labels are used to show important information over the map, and meta-summaries are used to automatically describe the behavior of individual trips or a group of trips.

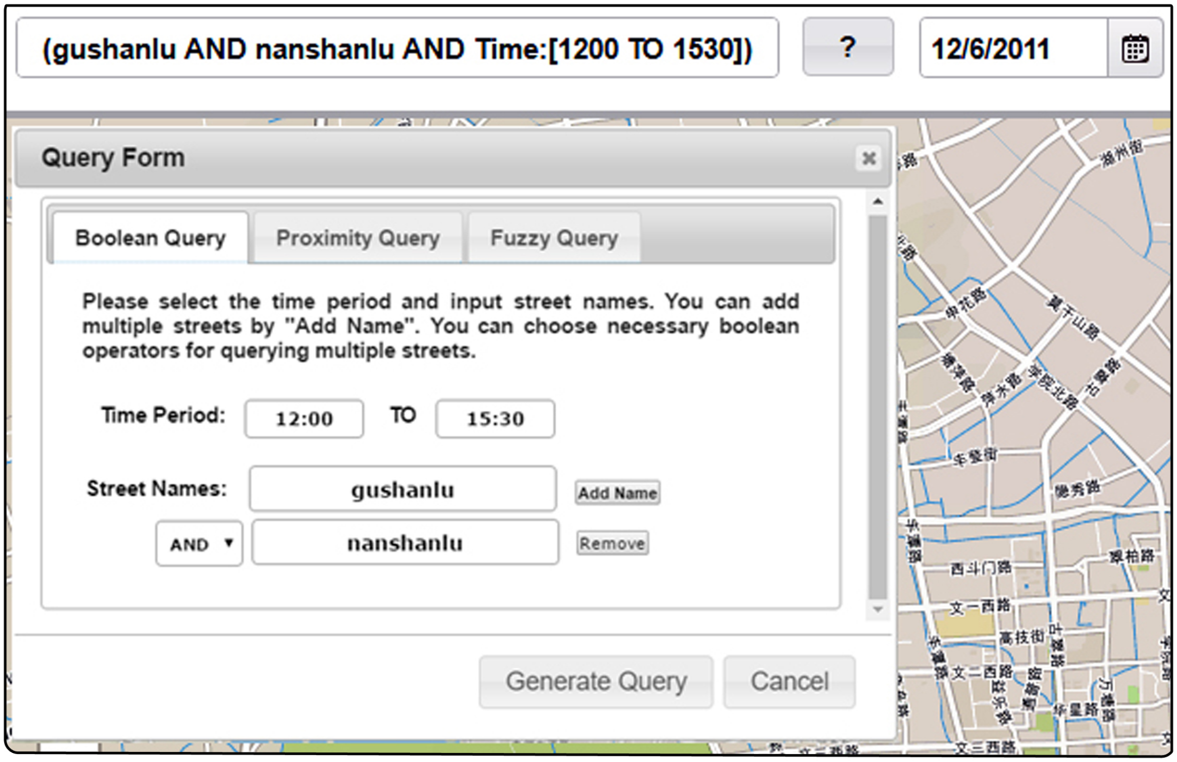
- Guiding users in data explorations with easy interactions: SemanticTraj helps users discover insights with guided data refinement and drilling down. A table view allows users to look through all retrieved trajectories and select individual ones for examination. Scatterplots, parallel coordinates, and parallel sets are presented for users to find interesting results. Users can interact with meta-summaries to further drill down according to interesting streets.

3.7.2 Visual Exploration of Taxi Trajectories

Fig. 2 shows the visualization interface of SemanticTraj. It consists of widgets for query construction and showing a set of coordinated views. These views are synchronized when users make selection or filtering on each of them. Next, we describe how the visual system facilitates effective and efficient visual exploration of taxi trajectories. We use an example of querying taxi trips passing Shangtanglu street in the morning (7am-9am) of Dec 6, 2011.

3.7.2.1 Inputting Semantic Query Sentence

Fig. 2(1) is the input box of a semantic query. Users can write a textual search sentence with name and time conditions. Auto-complete provides search suggestions with similar names. Lucene's query parser will process the input. For the example task, users input *Shangtanglu AND PickUp:[07:00:00 TO 09:00:00]*, where the AND combines the condition of street name and the range query of pick-up time.



(gushanlu AND nanshanlu AND Time:[1200 TO 1530])

12/6/2011

Query Form

Boolean Query | Proximity Query | Fuzzy Query

Please select the time period and input street names. You can add multiple streets by "Add Name". You can choose necessary boolean operators for querying multiple streets.

Time Period: 12:00 TO 15:30

Street Names: gushanlu

AND nanshanlu

Figure 10: Users can fill in a form to automatically generate query sentence in the input box.

Users can alternatively open a form and fill in it to create a query sentence. The sentence is automatically generated from the input fields. Fig. 10 shows the form where users choose the time period, fill in two street names, and select the AND operation. The auto-generated query sentence is displayed in the input box. Proximity and fuzzy queries can be generated in a similar way.

3.7.2.2 Displaying Query Results with Semantic Information

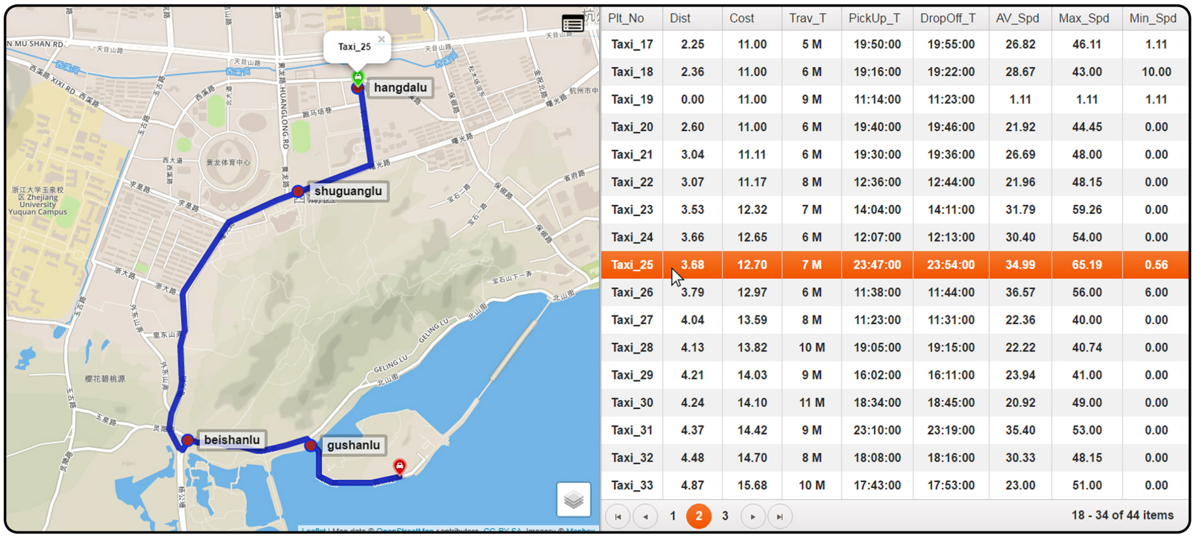
Map View: Fig. 2(5) is a zoom-in view of the acquired taxi trips on the map. The paths of these trips are drawn with the same opacity, so that the color density on a road reflects how often it is used by these trips. Users can change the drawing attributes such as color, line width, or opacity for their favorite appearance. They can also show the pick-up and drop-off locations as map markers. These display options are controlled by the panel shown in Fig. 2(2).

Text Labels: Text labels are given on a few streets, which show the street names, the speeds of the trips, and the taxi counts on the streets. For example, the top most label in Fig. 2(5) is "shixianglu: medium: 15". It indicates 15 taxi pick-ups happened on Shixianglu street and the average travel speed on this street is medium. The system provides a variety of labeling options to show the top streets where pick-ups (or drop-offs) happen, or the top streets with a fast (slow) speed. The number of labels is controllable by users to reduce clutter. They can toggle on/off different items (name, speed, etc.) in the labels. Users can also click on a street to show its label.

Meta-summary: Fig. 2(6) displays the meta-summary of the taxi trips in textual description. It summarizes the total number of trips, the speed information of this group, and the popular streets. From this report, users can get an idea of the retrieved taxi trips immediately. They can also click on a name to refine the trips passing the corresponding street. The summary can be toggled on and off by users.

3.7.2.3 Interactively Refining Query Results

Table View: Users can study the attributes of all the result trips in a table shown in Fig. 11. It displays the plate number (anonymized for privacy protection), trip distance (length of the path), trip cost (computed by the taxi fare rule of Hangzhou), travel time,



Plt_No	Dist	Cost	Trav_T	PickUp_T	DropOff_T	AV_Spd	Max_Spd	Min_Spd
Taxi_17	2.25	11.00	5 M	19:50:00	19:55:00	26.82	46.11	1.11
Taxi_18	2.36	11.00	6 M	19:16:00	19:22:00	28.67	43.00	10.00
Taxi_19	0.00	11.00	9 M	11:14:00	11:23:00	1.11	1.11	1.11
Taxi_20	2.60	11.00	6 M	19:40:00	19:46:00	21.92	44.45	0.00
Taxi_21	3.04	11.11	6 M	19:30:00	19:36:00	26.69	48.00	0.00
Taxi_22	3.07	11.17	8 M	12:36:00	12:44:00	21.96	48.15	0.00
Taxi_23	3.53	12.32	7 M	14:04:00	14:11:00	31.79	59.26	0.00
Taxi_24	3.66	12.65	6 M	12:07:00	12:13:00	30.40	54.00	0.00
Taxi_25	3.68	12.70	7 M	23:47:00	23:54:00	34.99	65.19	0.56
Taxi_26	3.79	12.97	6 M	11:38:00	11:44:00	36.57	56.00	6.00
Taxi_27	4.04	13.59	8 M	11:23:00	11:31:00	22.36	40.00	0.00
Taxi_28	4.13	13.82	10 M	19:05:00	19:15:00	22.22	40.74	0.00
Taxi_29	4.21	14.03	9 M	16:02:00	16:11:00	23.94	41.00	0.00
Taxi_30	4.24	14.10	11 M	18:34:00	18:45:00	20.92	49.00	0.00
Taxi_31	4.37	14.42	9 M	23:10:00	23:19:00	35.40	53.00	0.00
Taxi_32	4.48	14.70	8 M	18:08:00	18:16:00	30.33	48.15	0.00
Taxi_33	4.87	15.68	10 M	17:43:00	17:53:00	23.00	51.00	0.00

1 2 3 18 - 34 of 44 items

Figure 11: Using a table view to study individual taxi trajectories.

and pick-up and drop-off streets. The table also shows the max, min, and average speed of each trip. The trips can be ordered using different scoring functions. Users can choose single or multiple trips from the table. Selected trips will be displayed on the map (Fig. 11), while the major streets are labeled so that users can efficiently study the paths.

Scatterplot View: Fig. 2(3) is a scatterplot which draws each taxi trip as a dot. Users can choose the attributes mapped to the axes and the color of dots. The attributes are trip distance, cost, traverse time, pick-up time, drop-off time, max speed, min speed, and average speed. Users can brush and choose a subset of trips for further study by selecting a set of dots. Moreover, users can hover over a dot to show its individual meta-summary. The behavior of this trip is displayed in a descriptive sentence (Fig. 2(4)).

Pick-up and Drop-off Relations: A parallel sets view [103] shows the relationships of the pick-up and drop-off streets (Fig. 12c). The top streets used by pick-ups are drawn on the top horizontal line and the top streets used by drop-offs are drawn on the bottom.

The width of ribbons linking two streets is proportional to the number of trips traveling between them. A ribbon's color is the same as the color of pick-up streets. Fig. 12c shows that the trips starting from Gushanlu end on a variety of streets. Users can click on a ribbon to investigate the trips it represents in other views.

Parallel Coordinates Plot: A parallel coordinates plot (PCP) helps users identify trends as well as outliers of the taxi trips. For example, Fig. 12b shows that most taxi pick-ups/drop-offs happen between 8am and 10pm. Users can hover over a line to highlight a trip on the map and see the meta-summary of the trip.

Circular Heatmaps: Circular heatmaps (Fig. 15) visualizes the statistical data such as the average speed of streets (see Sec. 3.8.4). The whole day is divided into 24 hours and each hour is represented by 6 arcs so that each unit reflects the average speed (or other attributes) in a 10-minute period.

We use a qualitative color spectrum from ColorBrewer [104]. More spectrums can be selected such as colorblind-safe palettes.

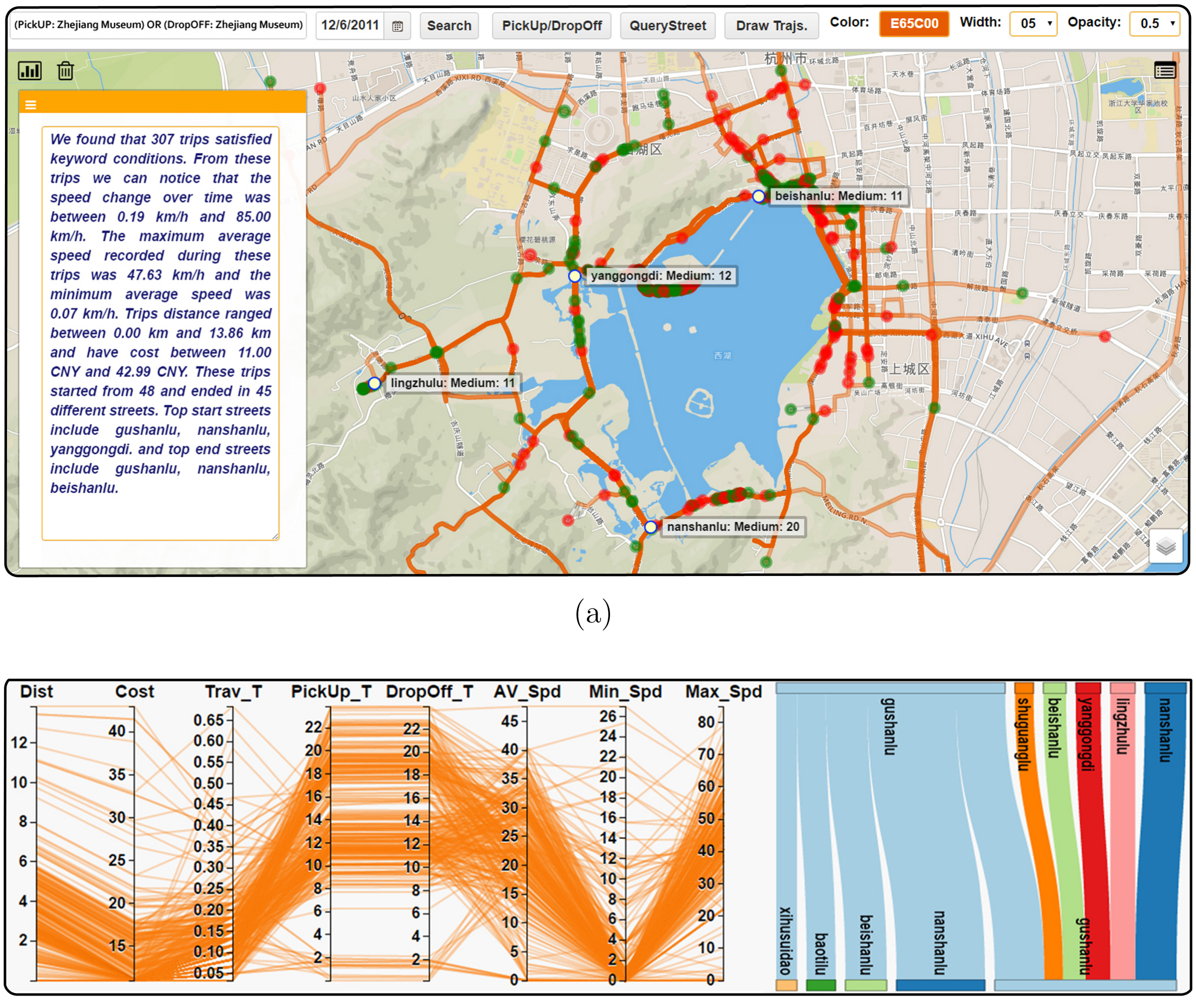
3.8 Usage Scenarios

We show the usability of our prototype in a set of illustrative usage scenarios of taxi data visual analytics.

3.8.1 Provide shuttle buses for tourists

Zhejiang Museum wanted to improve their service to customers. The museum clerks used SemanticTraj to query taxi trips with their own POI name (i.e., Task 3). By inspecting the results, they found the paths, popular pick-up and drop-off locations, and when passengers took taxis to reach the museum. The museum would provide complementary shuttle buses based on the results. In particular, the clerks formed the query input *Pick-up:Zhejiang Museum OR Drop-off:Zhejiang Museum* over a week from Dec 5 to Dec 11. Fig. 12a showed the results on the map where the drop-off/pick-up locations were shown in green/red dots. The visualization provided a general overview of their distribution, but the clerks need more information to decide the locations of candidate bus stations. They looked at the meta-summary, where basic facts of the 329 taxi trips were summarized. They toggled on the text labels for the top pick-up streets (Nanshanlu, Beishanlu, Lingzhulu and Yanggongdi). The street names and pick-up counts were displayed on the map. These streets were good candidates. Users clicked on the meta-summary and labels to filter trips passing a specific street.

The clerks further filtered the result trips in the morning and afternoon (not shown here). They found that Nanshanlu and Beishanlu were used for pick-up and drop-off in the morning, but in the afternoon Beishanlu was mostly used only as drop-off locations. This fact helped them schedule the buses. Moreover, there was a far-away suburban visitor center of Lingyin mountain at Lingzhulu. The system showed that most passengers took taxis from it to the museum in the morning. So shuttle buses could be arranged there in the morning. The PCP in Fig. 12b allowed the clerks study attributes of the trips. For example, the trips with a pick-up time of (14-17) were less than that of (10-14) and (17-19), which can be used to decide shuttle bus schedules.



(PickUP: Zhejiang Museum) OR (DropOFF: Zhejiang Museum) | 12/6/2011 | Color: E65C00 | Width: 05 | Opacity: 0.5

We found that 307 trips satisfied keyword conditions. From these trips we can notice that the speed change over time was between 0.19 km/h and 85.00 km/h. The maximum average speed recorded during these trips was 47.63 km/h and the minimum average speed was 0.07 km/h. Trips distance ranged between 0.00 km and 13.86 km and have cost between 11.00 CNY and 42.99 CNY. These trips started from 48 and ended in 45 different streets. Top start streets include gushanlu, nanshanlu, yanggongdi. and top end streets include gushanlu, nanshanlu, beishanlu.

(a)

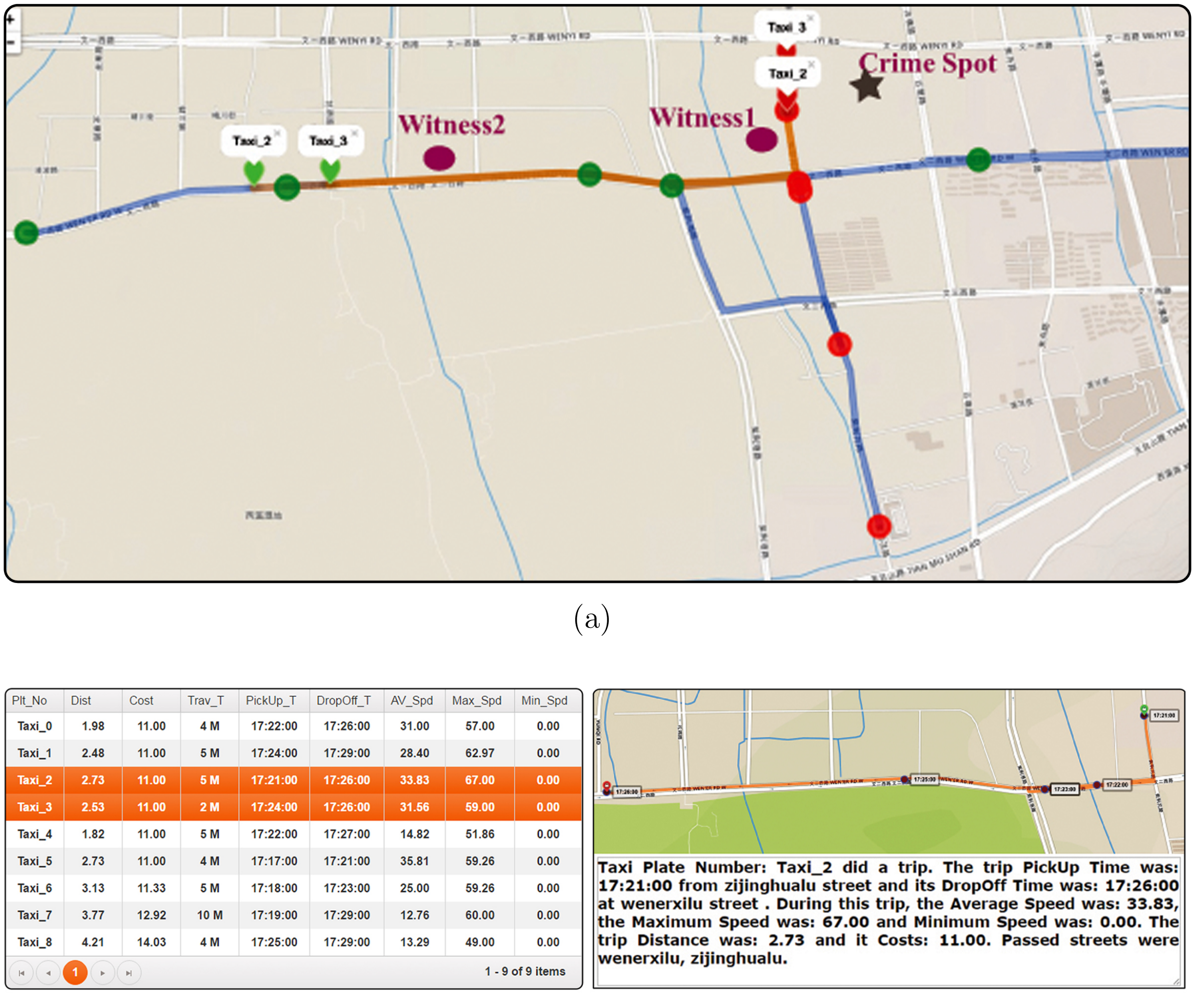
(b)

(c)

Figure 12: Provide shuttle buses for tourists of Zhejiang Museum after query taxi trips leaving or arriving at the museum. (a) Taxi trips are shown where street names with frequent pick-ups (potential bus stops) are easily found by the text labels; (b) The PCP view of the trips show their attributes. (c) The parallel sets view reveals the pick-up and drop-off relations.

3.8.2 Find criminal suspect taxi passenger

A crime happened in Hangzhou at around 5pm, Dec, 5. The characteristics of the suspect was described by the victim and then broadcasted on TV. One eyewitness reported to the police that he saw the suspect on a taxi traveling at Zijinghualu Street at around 5:10-5:30pm, and another citizen said that he witnessed the taxi at Wenerxilu Street at around 5:20-5:30pm. To find potential taxi trips of the suspect, a police officer queried the trip documents by the two street names (i.e., Task 2) and the time periods. The query condition was {*Zijinghualu AND Wenerxilu AND* {*Pick-up Time:[17:00:00 TO 17:30:00]*} *AND* {*Drop-off Time* > *17:20:00*}}. Please note the times given by the witnesses were not accurate so the query covered a larger time range. Fig. 13a showed the acquired trips on the map, where their pick-up and drop-off locations were shown as markers. The officer also marked the positions where the witnesses were and the criminal site. Fig. 13b showed the list of these trips on the table view. The officer browsed the list and interactively clicked on each trip to inspect its path on the map. The officer quickly identified two trips of Taxi_2 and Taxi_3 which passed the site of the witnesses. Fig. 13c showed one of the two suspicious trips. The meta-summary described its travel behavior, including pick-up/drop-off streets, passed streets, and speeds. The information allowed the officer to further investigate the trip.



(a)

Plt_No	Dist	Cost	Trav_T	PickUp_T	DropOff_T	AV_Spd	Max_Spd	Min_Spd
Taxi_0	1.98	11.00	4 M	17:22:00	17:26:00	31.00	57.00	0.00
Taxi_1	2.48	11.00	5 M	17:24:00	17:29:00	28.40	62.97	0.00
Taxi_2	2.73	11.00	5 M	17:21:00	17:26:00	33.83	67.00	0.00
Taxi_3	2.53	11.00	2 M	17:24:00	17:26:00	31.56	59.00	0.00
Taxi_4	1.82	11.00	5 M	17:22:00	17:27:00	14.82	51.86	0.00
Taxi_5	2.73	11.00	4 M	17:17:00	17:21:00	35.81	59.26	0.00
Taxi_6	3.13	11.33	5 M	17:18:00	17:23:00	25.00	59.26	0.00
Taxi_7	3.77	12.92	10 M	17:19:00	17:29:00	12.76	60.00	0.00
Taxi_8	4.21	14.03	4 M	17:25:00	17:29:00	13.29	49.00	0.00

1 - 9 of 9 items

Taxi Plate Number: Taxi_2 did a trip. The trip PickUp Time was: 17:21:00 from zijinghualu street and its DropOff Time was: 17:26:00 at wenerxilu street . During this trip, the Average Speed was: 33.83, the Maximum Speed was: 67.00 and Minimum Speed was: 0.00. The trip Distance was: 2.73 and it Costs: 11.00. Passed streets were wenerxilu, zijinghualu.

(b) (c)

Figure 13: Find a criminal suspect passenger on taxi trips. (a) All result trips shown on the map; (b) A list of the trips on the table; (c) One of the two suspicious trips.

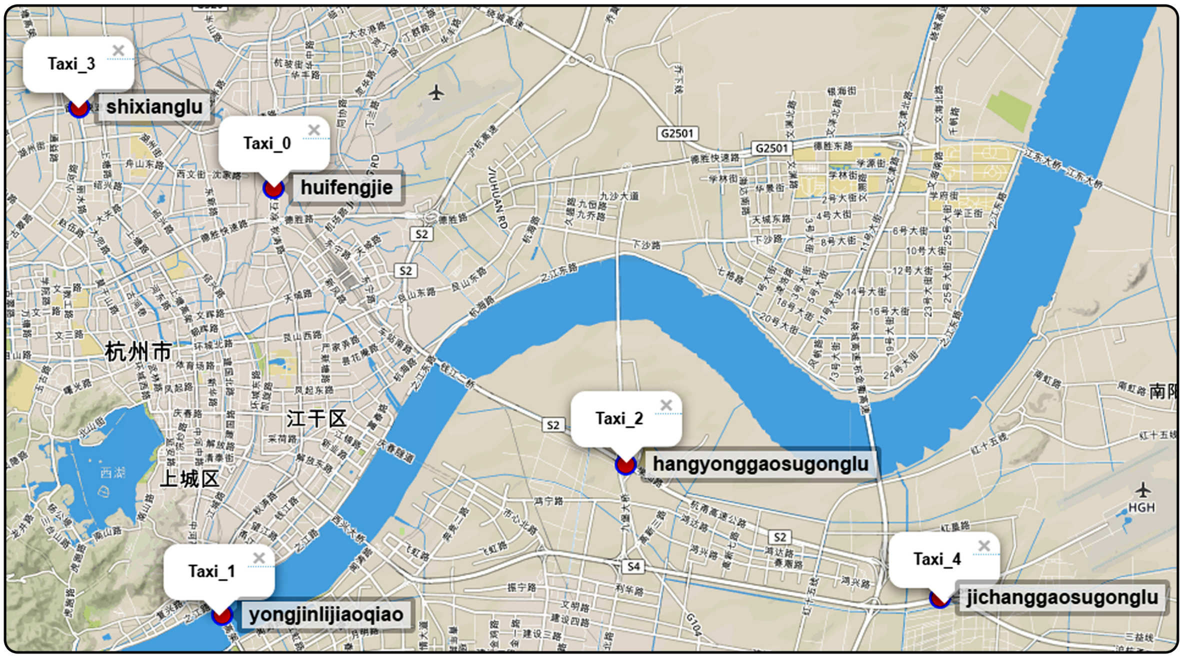
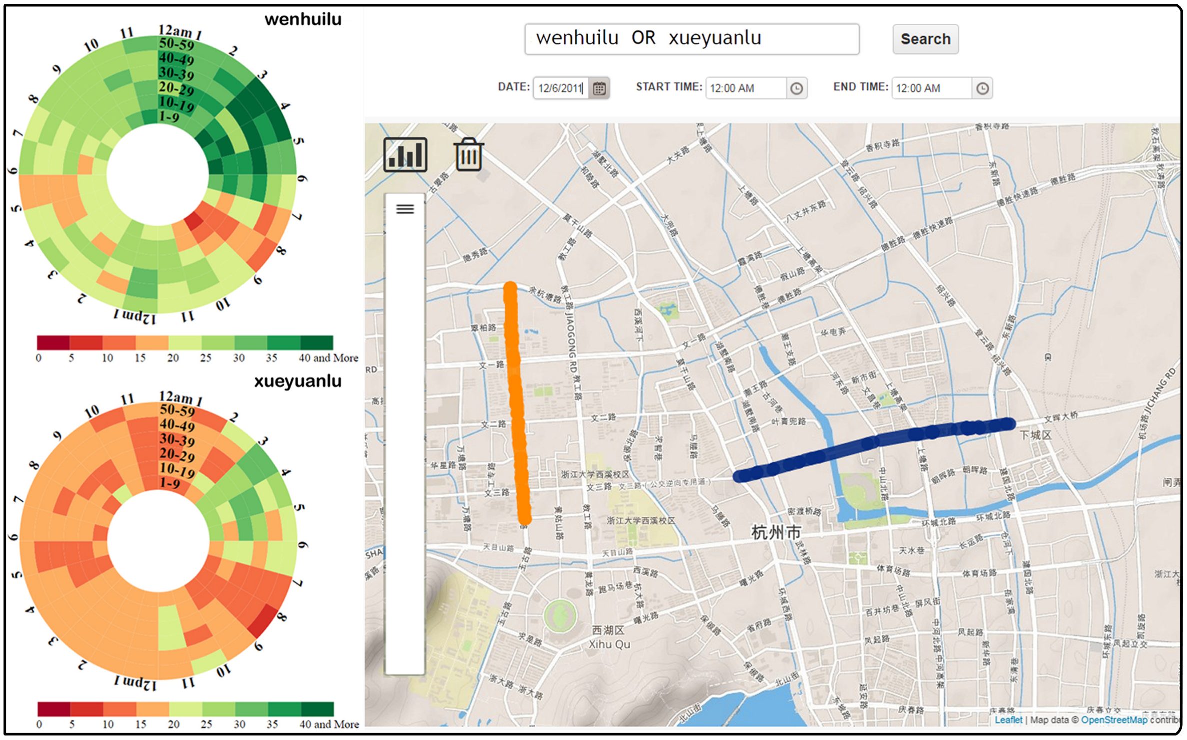


Figure 14: Identify taxi drivers' abnormal behavior.

3.8.3 Identify taxi drivers' abnormal behavior

A taxi company manager wanted to find drivers who had improper driving performance in a day. The manager queried the trajectory documents to find taxis having an abrupt change of speed (i.e., Task 5). She observed Fig. 14 which showed the results using the query condition $\{\{\mathit{Slow\ VeryFast} \sim \mathit{1}\}\ \mathit{in\ DSpeed}\}$, which implied speeding since such a quick acceleration from less than 20 km/h to more than 100 km/h was not proper in the city center. The taxi IDs and the streets related to these abnormal trajectories were labeled on the map.



wenhuilu OR xueyuanlu

Search

DATE: 12/6/2011

START TIME: 12:00 AM

END TIME: 12:00 AM

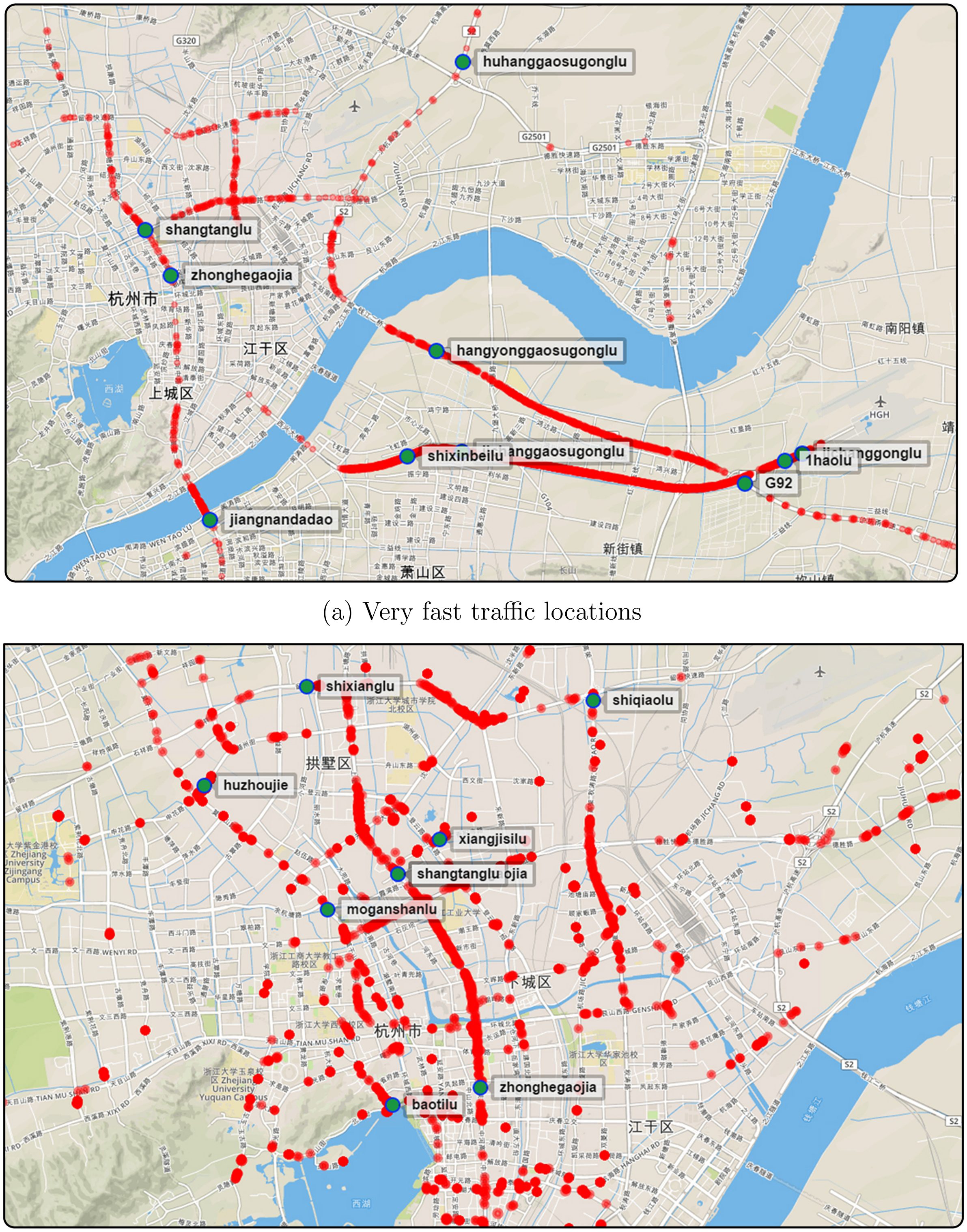
Figure 15: Assess traffic situations of two streets.

3.8.4 Study traffic information over streets

The residents of Hangzhou wanted to learn the traffic information of the city. In Fig. 15, a resident queried the trajectory documents of a day with the names of two street close to his home, namely Wenhuilu (blue) and Xueyuanlu (orange). The two streets were highlighted on the map. Two circular heatmaps showed the average speed of them throughout the day. The resident compared the two heatmaps. She found that both of them had rush hours between 7am and 9am. She also found that Xueyuanlu was much busier than Wenhuilu in most time periods even in midnight between 12am and 3am.

In another scenario, a practitioner in the Bureau of Transportation wanted to assess urban traffic situations. He queried the trajectory documents by {{*VeryFast VeryFast*} *in DSpeed*}. Fig. 16a showed the result locations, where taxis drove very fast were indicated in red color. They manifested many major roads in Hangzhou. However, many parts of these roads had intermittent red dots, which indicated slow traffic.

He queried again with the query condition using ten continuous Slows as {{*Slow Slow ... Slow*} *in DSpeed*}. This query sentence is easily created in the proximity query on the form (Fig. 10). Fig. 16b showed the result locations, where a traffic jam possibly happened since the retrieved trajectories had consecutive slow speeds. Visual labels in the two figures helped the practitioner quickly identify street names.



(a) Very fast traffic locations

(b) Traffic jam locations

Figure 16: Assess city-wide traffic information.

3.8.5 Discussion: Benefits of Using Our System

In the illustrative usage scenarios, users are diversified including police officers, museum clerks, city residents, and taxi company managers. Our system is a good choice in completing their work for the following reasons:

(1) It is very convenient for these general users to conduct searches by simply filling in street/POI names (or other descriptive terms). It would cost them extra effort to locate and select (e.g. by brushing) the streets over a map.

(2) Our system works fast in giving results through a variety of visualizations.

(3) The summaries and labels are read quickly for semantic knowledge.

3.9 Evaluation

3.9.1 Expert Feedback

We conducted in-person interviews with two domain experts: one is an urban transportation researcher and the other one is a criminologist. Both of them hold a Ph.D. degree in their fields. First, we introduced our method. Then we demonstrated our system with several examples usage scenarios. Then, the experts used our system for interactive exploration. Finally, they were provided a document describing our methodology and system. A few days later, the experts fed back to us with written documents of their opinions.

The criminologist has been working with police departments on fighting urban crime. She provided the following feedback: “Visual query using text search engine is an innovation for policing strategy. Street or POI names are more straightforward to criminologists than abstract geometry or spatial coordinates. This tool is *very useful because it is closer to our understanding of the real place featured with names instead of a virtual location represented by numbers.* With great volumes of trajectories data by virtue of everyday technology, the sharing and utilization of these massive data have presented great challenges for research and application in crime mapping. Moreover, the concepts and technologies of Web 2.0 represent demands for user-oriented interaction and collaboration. Most criminologists are not geographers or geographic information scientists. Hence, there is a huge gap between spatial technology advances and policing needs. Most current crime mapping tools rely on spatial query through geometric regions or buffers on a map to search trajectories, which is not as intuitive and easy as visual query for domain scientists and common users. This new tool addresses an emerging need to provide non-technical users with an integrated platform for spatial exploration and visualization. This practice will open up a rich empirical context for interdisciplinary studies and policy interventions.”

The urban transportation researcher provided the following feedback: “As a former urban planner and currently an urban geographer, I have spent years of efforts on understanding the vehicle flow on the urban street network and its uneven traffic distribution across space and over time. Many methods have been tried, ranging from surveys and field observation to spatial statistical analysis. However, these methods do not recognize the difficulty of big transportation data and impossibility of visualizing such data using conventional approaches. It warrants notice that spatial distribution of real traffic flow in the street network may vary from time to time, which poses a challenge to store and query the data for real-time use. The developed tool can solve two major issues: store and query the increasing data size especially real-time feature; analyze and identify the network structure of transportation flow in a fast and visual platform. This tool is a good example of spatial intelligence, which allows efficient data integration for large-volume and near real time spatial and non-spatial data (multi-source data). In addition *it allows users who have very few spatial data handling skills to conduct space-time data analysis easily and effectively.* Hence, this tool can serve as a practice, research and education platform by delivering geographic information service in the data-rich age which is featured by the unprecedented terabytes of digitized data."

They suggested that we improve the work mostly by making the web-based system remotely operated by multiple users in real-time, and providing more interactions for users to find more city data. These comments provide several directions of our future work.

3.9.2 User Study of Visual System

We performed a preliminary user study with 15 CS graduate students (6 females and 9 males with ages from 23 to 34). The goals of the study were to evaluate whether the visual system is easy to use and whether the semantic meta summaries and labels, is helpful in visual analytics. The subjects conducted the study one by one. First, the subjects were given an introduction describing the system and the tasks. Then, they practiced on the system for 5 minutes. After that, they were asked to search for a street, Gushanlu, to find all trips passing it using a name query. Next, they were asked to write down the answers to the following questions: "How many trips passed this street?"; "What is the maximum average speed?"; "What is the top two pick-up streets?"; "What is the maximum fare cost?"; and "How many trips start from (nanshanlu) street?". Each subject conducted the above task in two sequential sessions. Group 1 include half of all subjects (picked randomly) who answered the questions using the full system (S1) in the first session. Then they answered the same questions again on new answer sheets using the system without meta-summaries and labels (S2) in the second session. Group 2 including another half of the subjects used S2 in the first session and S1 in the second session. At the end of the study, they answered the following questions:

(A) Overall, do you think the system can help complete such tasks intuitively?

(B) To complete such tasks, which visual tools were more useful (Meta-summary, visual labels, or other tools)?

(C) To what extent the visual system is easy to use: Very Easy, Easy, Fair, or Poor?

(D) What is your suggestion for improvement?

The average task completion time using S1 and S2 were 2 minutes and 5 minutes, respectively. 93% of subjects achieved correct answers compared to the ground truth. There was no significant difference on the completion time and correctness between Group

1 and Group 2. The results indicated that the meta-summaries and the labels were helpful in accelerating the visual analytics process. For question (A), all subjects agreed that the system was qualified in completing the tasks. For (B), 67% of the subjects preferred to use Meta-summary and Labels to complete the tasks. 33% of the subjects preferred to use other visual representations (e.g., PCP and scatterplot). For (C), 73% of the subjects agreed that this system was very easy to use and 27% of the subjects said it was easy. In addition, the subjects were very excited about our prototype. They suggested many ideas for improving our prototype, such as showing more interesting information that reflects the semantic of trips on the map; using the system for real-time data; etc. This study showed that our prototype system is easy to use and can be of great interest for urban trajectory study.

3.10 Chapter Summary

We have proposed a new approach to interacting with massive taxi trajectory data sets. It utilizes textualization and taxi documents so that the interactions can be applied through semantic rich operations. In this chapter, semantic information refers to the "meaning in language", where we map the numeric values of GPS points and speeds to names and descriptions in natural language for easy query, visualization and user understanding. Semantics may also imply high-level, summarized information describing the pattern and knowledge hidden in languages. The underlying text search engine provides a new way of data management and fast query support for various Boolean conditions, which is complementary to existing region queries. General users are thus provided easy, intuitive tools with enhanced guidance in their visual exploration for a set of analytical tasks. We developed a prototype visual analytics system with a set of visualization tools for users to conduct interactive visual exploration.

CHAPTER 4

QuteVis: Visually Studying Transportation Patterns Using Multi-Sketch Query of Joint Traffic Situations

QuteVis is a visual analytics system that supports the study of urban traffic patterns. Unlike most existing approaches which investigate traffic patterns in user specified spatial regions and temporal periods, QuteVis supports a different type of data query and analytical tasks. Multi-sketch query and visualization helps users discover those specific times and days in history, which have specified joint traffic patterns distributed on different city locations. Users can use touch input devices to define, edit, and modify multi-sketch on a city map. A set of visualizations and interactions are provided to help users browse and compare the retrieved traffic situations and discover potential influential factors. To facilitate these operations, we construct a transport database from heterogeneous data sources with an optimized spatial indexing and weighted similarity computation. Several case studies with real world data and domain experts demonstrate how QuteVis is useful in addressing many major problems in modern cities.

4.1 Introduction

Nowadays, a large amount of real time datasets of urban transportation behaviors are collected by administrations, companies, and domain researchers. In these urban transport databases, traffic speed, volume, human/vehicle movement traces, etc. are recorded in different time periods of different days. They have been utilized by researchers and analysts to assess the situations and patterns related to major problems in modern cities, such as traffic jams, unbalanced capacities, and frequently occurring accidents. Therefore, efficient and user-friendly tools for visually querying and examining large urban transport database are important in modern transportation studies and practices.

Existing work has presented many visual analytics (VA) systems so that users can define spatial constraints (by drawing a spatial region, brushing a road, or giving road names) and temporal conditions (by selecting dates, times, or weekdays) to retrieve and examine the traffic information with interactive visual tools. A typical task is to visually study the traffic speed and volume in a region (or street) at different time periods (or week days), which is supported by the "*where+when⇒what*" query. Similarly, commercial map services such as Google Maps and Baidu Maps allow users to visually identify traffic facts or patterns by giving spatio-temporal conditions.

However, there are no existing VA systems devoted to the "*what+where⇒when+how*" queries with *multiple spatial constraints* over the transport database. For example, assuming a highway bridge H and a street M are two critical traffic nodes, in two separate regions of a city. An analyst often wants to find the traffic situations when both of them have traffic jams, with the following "what if" questions:

- Q1: **What** are the specific days and times **if**, approximately, H has a traffic speed around 30 KM/hour AND M has around 10 KM/hour?
- Q2: **What** are the distributions of the matched results on different periods (e.g., morning, afternoon, night), different days (e.g., Mon, Tue, etc.), and different weather conditions, **if** H and M have the given traffic speeds?
- Q3: **What** is the traffic situation in a nearby residential district, **if** H and M have the given traffic situations?

These questions speculate a joint traffic pattern of "*what+where*" situations at *multiple and spatially diverse* geo-locations. A visual system is demanded by domain users to investigate "*when+how*" this pattern may happen from a transport database including: (1) the similar patterns happened in different time periods of different days (Q1); (2) the influential factors that possibly determine this traffic pattern (Q2); (3) the traffic behaviors at other locations (Q3). Such "*what+where⇒when+how*" tasks are common in the practices of utilizing transport data, which can contribute to urban infrastructure improvement and planning as well as business development.

In this chapter, we present a visual analytics system, named QuteVis, which helps users interactively (q)uery (u)rban (t)ransport databas(e) (Qute) to conduct these tasks over large scale datasets. The system combines several features in supporting efficient visual analytics:

- *Multi-sketch query at separate locations on the map*: Users can define joint conditions by specifying traffic attributes (e.g., traffic speed) as multiple input sketches through direct drawing on a city map by hand or by mouse. Reliable sketch recognition interface is designed to allow free style *sketches* on different geo-structures: a street, a path, or a region. The system further allows users to edit and modify sketches with easy interactions.

- *Heterogeneous data integration*: The traffic information (e.g., traffic volume and speed over time) on street segment level is integrated with spatial information (e.g., city roads and POIs), mobility information (e.g., vehicle pickups/dropoffs and trajectories), and other information (e.g., weather condition). These data is efficiently stored and retrieved by utilizing specific spatial indexes, together with the similarity computation, to locate user-selected streets, paths, and regions. These operations are supported by (1) a new street segment optimization algorithm, (2) a data cube designed and built for accelerated data retrieval, and (3) geo-spatial indexing for fast geographical locating.

- *Weighted traffic similarity for joint conditions*: To quickly find matching results in massive traffic data records at different time periods, a similarity value is computed between the given value on a sketch and each data record in the database. A global "weighted similarity" is computed for multiple sketches, where a weight is assigned to each spatial sketch. For example, a large weight may be given to a major street, while small weights are assigned to secondary streets. The weights thus can be customized to user priority. These similarities on multiple sketches are visualized to provide cues for users to investigate their preferred results from the joint conditions.

- *Sketch+visualization interface*: A set of visualizations are integrated into a sketch + visualization interface. Then, QuteVis facilitates users to overview the distribution of joint traffic patterns from their sketch inputs. They can further investigate top matched time periods, and select any interesting time periods or days for drill-down analysis. Moreover, a multi-map comparison view allows them to compare traffic situations in the whole city and/or at any specific locations.

We have developed a fully working prototype of QuteVis with coordinated views and abundant interactions. Case studies and domain expert feedback are presented to illustrate its usability and efficiency.

In summary, the main contributions of this chapter include:

- We present QuteVis system to study urban transportation data based on a different type of investigation model: "*what+where⇒when+how*". To the best of our knowledge, it is the first visual analytics system devoted to such tasks.

- Multiple types of transportation data are integrated and stored with efficient geo-indexing, data cubes and optimal street network. The data queries are supported by a weighted traffic similarity and similarity categorization.

- Joint traffic situations can be specified on multiple locations that can distribute separately in space, which have not been supported in existing transportation application tools.

- Multi-sketches are supported by reliable sketch recognition algorithms for different types of sketches on street, path and region.
- A sketch+visualization interface allows users to conduct interactive query input and drill-down study.

4.2 Related Work

4.2.1 Spatio-Temporal Data Visualization and Interaction

There has been extensive research on visual analytics of traffic and various movement data [50]. Typically, map-based displays and information visualization techniques were combined in visualizing the spatio-temporal data [50]. With interactive query of traffic/trajectory data, a large amount of papers were presented. For example, Wang et al. [85] explored traffic data recorded on sparsely distributed cells in a city. Liu et al. [86] presented a visual analytics system of discovering route diversity. Wang et al. [60] presented an interactive analysis system for urban traffic congestion. Pu et al. [87] visually monitored and analyzed complex traffic situations in big cities. Wang et al. [88] presented a visual reasoning approach for the data-driven transport assessment on urban roads. It supports dynamic query of traffic situations within a bi-directional hash structures on a spatial grid. Ferreira et al. [89] allowed users to visually query taxi trips on spatio-temporal constraints, which was designed only for queries over the pick-up and drop-off pairs. Al-Dohuki et al. [105] transformed taxi trajectories into texts consisting of street names and text labels denoting taxi speeds for users to query a taxi trajectory database using a text search engine. Zeng et al. [106] unveiled passenger mobility on public transportation system and assisted transportation planning. Chen et al. [107] presented Visual Analyzer for Urban Data (VAUD), that focuses on linking a variety of urban data together for analysis and includes similar concepts on when+where queries. Poco et al. [108] defined a time-varying vector-valued function on the road network graph, and visualized this vector field to reveal interesting mobility patterns. Some related work combined visual analytics with computational and data mining tools (e.g., [109, 110]) to discover temporal and spatial patterns of traffic dynamics. For example, topological computing over time varying scalar functions in different data slices can automatically find events (e.g., road block occurrence) in an urban region and retrieve similar events [111].

These approaches performed spatio-temporal query of transport data by giving both spatial and temporal constraints. In contrast, our approach alternatively helps users examine similar times and days that approximately have the user-specified values on given streets, paths and regions.

4.2.2 Spatio-Temporal Data Query and Management

To enable interactive data query and visual analysis, a spatial database needs to be quickly accessed to retrieve matching records and then visualized at interactive rates. There has been extensive research in spatial indexing mechanisms designed for efficiently finding spatial data inside given regions and time periods, such as the popular R-tree [30] and its variants (e.g., [32, 112]) for data points inside a spatial domain. A few methods extended R-trees to index moving trajectories such as the Spatio-Temporal R-tree [43] and the global heap [113]. Moreover, visual system developers have proposed data indexing techniques such as tree-based indexing [89,114], hash structures [88] and text search [105]. Many multidimensional and spatiotemporal datasets were pre-aggregated and indexed in Nanocube [37] based on data cube techniques. Hashedcubes [115] further used a more compact representation and a simpler implementation over Nanocube to avoid a large number of aggregations. imMens [116] allowed interactive data querying among multivariate data tiles enabled by GPU-based parallel query processing. A GPU-based indexing scheme [117] efficiently supported spatio-temporal queries over massive point data which facilitates simultaneous filtering over multiple dimensions. All these methods were designed for quickly retrieve and process massive data, such as millions of spatial points and trajectories. Regional events (e.g., road block occurrence) are mined and discovered by automatical computation over time varying scalar functions in different data slices [111], so that similar events can be retrieved. But their aim is not for users to interactively speculate arbitrary traffic values at different locations, and find whether or not, when, and how these situations happen in history.

In this chapter, our focus is to study traffic patterns on multiple city streets selected by user sketches (regions and paths are represented as sets of streets). We query a transport database including traffic and other attributes of street segments at different time periods. The spatial indexing similar to R-tree is adopted for quickly locating user-selected streets in the interactive sketching operations.

4.2.3 Sketch-Based Visual Analytics

Sketch-based interfaces and modeling has been studied to enable users to interact with a computer through sketching [118]. Browne et al. [119] have described the design and realization of SketchVis, a sketch-based proof-of-concept application that leverages hand-drawn input for exploring data through simple charts. Visualization-by-Sketching [120] enabled artists and other visual experts to create accurate and expressive data visualizations by painting on top of a digital data canvas, sketching data glyphs, and arranging and blending together multiple layers of animated 2D graphics. SketchStory [121] helps the presenter stay focused on telling her story by eliminating the burden of manual data binding. It allows the presenter to record a sequence of charts along with example icons before the presentation, and to invoke them with simple sketch gestures in real-time. In GIS domain, sketch-based query concept, model, and prototype were studied in a dissertation [122]. SpaceSketch [123] provided interactive tools of sketching paths and regions for spatiotemporal data visualization. Sketch based queries were supported to find locomotion pattern of human trajectories [124]. Blaser et al. [125] highlighted a set of interaction methods and sketch interpretation algorithms that are necessary for pen-based querying of geographic information systems. Users of nuSketch [126] sketched on maps in reasoning about a hostile battlespace. Spatial data characteristics was visually studied along a line drawn across a map [127]. Malik et al. [128] discussed regional correlations and allows users to sketch time periods of interest. HotSketch [129] allowed police officers to sketch a path on the map by specifying control points that define a polyline to query point-based crime data within a distance. In contrast, our system uses sketches over map canvas to select the road network geometries, paths and regions, instead of points, and then retrieves their traffic records in the database. Our sketch recognition algorithm is different from the existing methods as we identify and match sketch inputs in the units of our optimized the street segments, and we accept free drawings for multiple paths and regions on the map (Sec. 4.7). In addition, our system allows users to perform easy editing of the sketch results based on street segments.

4.3 Overview

QuteVis aims to help users "see" and examine the traffic patterns in large volume of historical data by giving their speculated values in separate locations all over the city, so that they can test their hypotheses with real datasets. They can quickly sketch and speculate traffic behaviors and immediately get visual feedback from large scale historical time-varying traffic data together with related attributes such as weather and taxi activities. This goal is different from many existing works either mining the data to discover hidden patterns or studying traffic information by selecting locations and times.

4.3.1 Goal and Tasks

During our system design, we have interviewed a group of five experts (named as 5ExP) who have been working on urban planning and transportation analysis. We discussed our objectives and the experts provided us a set of tasks that are important for urban transportation analysts and planners. One example task has been described in the introduction section about a highway bridge H and a major street M. Other example tasks include:

- ***Example task*** : When the street M is jammed, a street around an office building in its vicinity may have a smaller number of taxi pickups or dropoffs. An analyst can use our system to interactively investigate: (1) Whether such situation happens often? (2) Is there any difference over various weather conditions? and (3) Whether other related locations have a large number of taxi pickups/dropoffs? The analysts can make new recommendations for improving transportation service.
- ***Example task*** : A user can select a path or multiple paths, e.g., from home to working place. Our system can be used to visualize traffic conditions on different times and days on the streets along these paths. The information can be used to arrange the means and times for their travel.

4.3.2 System Requirements

We further discussed system requirements with the experts (5ExP), who guided our design of QuteVis from domain users' perspective. Several major requirements of the QuteVis system were identified as:

- **Geographical context**: Users should be facilitated to conduct visual queries over urban geographical context. Information such as road networks, street hierarchy (e.g., highway, major, or secondary roads), and Points of Interest (POIs) should be provided.

- **Flexible multiple selections and editing**: Users should be able to quickly locate and select their preferred urban structures (including streets or regions), and then easily set up query conditions such as the values of traffic speed or taxi pickups. Users should also be able to edit their selections and values. It would be better if hints can be given about historical values at specific locations for users to set up query conditions.

- **Fast response and interactive exploration**: The system should provide immediate response for queries and the results should be visualized quickly so that users can iteratively explore the large data.

- **Easy perception and understanding**: Visualization of the query results should be straightforward for users to discern and understand. Visualization may not focus on small differences in numerical values (e.g., speeds of 20 or 22 km/hours are not distinctive in traffic analysis).

- **Visual comparison**: Visual comparison of the situations in multiple times/days and at different locations is an essential function which should be supported. Multiple locations diversely distributed in a city should be compared together.

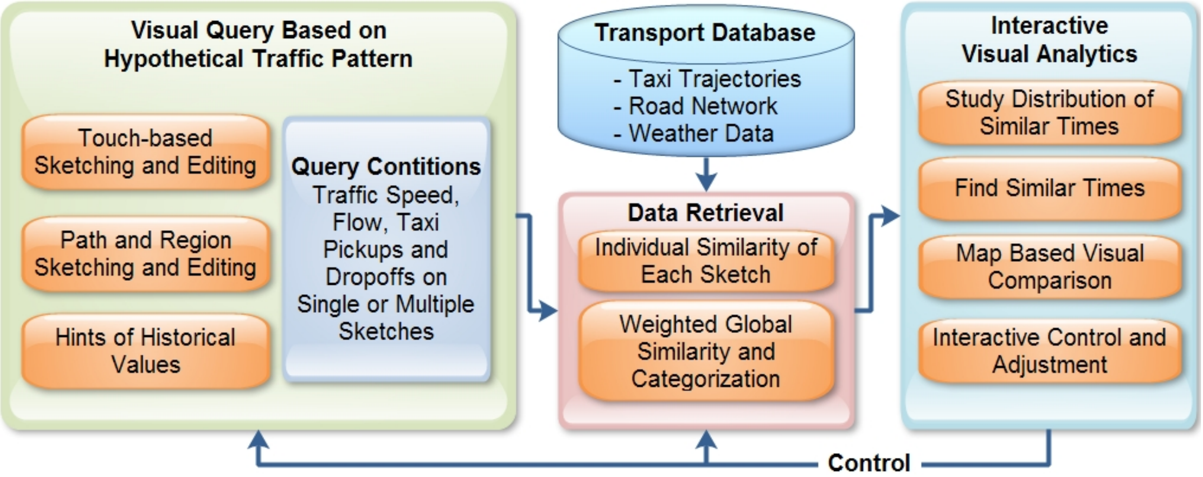

Visual Query Based on Hypothetical Traffic Pattern
Touch-based Sketching and Editing
Path and Region Sketching and Editing
Hints of Historical Values
Query Contitions
Traffic Speed, Flow, Taxi Pickups and Dropoffs on Single or Multiple Sketches
Transport Database
- Taxi Trajectories
- Road Network
- Weather Data
Data Retrieval
Individual Similarity of Each Sketch
Weighted Global Similarity and Categorization
Interactive Visual Analytics
Study Distribution of Similar Times
Find Similar Times
Map Based Visual Comparison
Interactive Control and Adjustment
Control

Figure 17: QuteVis System Framework. QuteVis is an interactive visual analytics system integrating user sketching and visualization. It has efficient data integration and management.

4.3.3 System Design and Modules

QuteVis is an interactive VA system integrating user sketching and visualization, which is sup-ported by efficient data management. Its framework and major functional modules are illustrated in Fig. 17:

- **Multi-sketch query based on hypothetical traffic patterns**: Users can define their speculated patterns by (1) selecting *multiple, spatially diverse* locations by drawing on a city map. The selected objects (i.e., sketches) can be streets or paths, as well as arbitrarily shaped regions in a city. (2) Specifying query values at these locations about transport attributes (e.g., traffic speed), with visual hints of their historical values to guide their input. The selected sketches can be conveniently adjusted and edited. Users can also manage multiple sketches by enabling or disabling them in querying. In addition to desktop environment, we also implement QuteVis on touch input devices for the convenience to domain users, which is especially helpful in the field environment.

- **Data retrieval from transport database**: Different times and days in the database are categorized into *not similar*, *little similar*, *very similar* and *extremely similar* by their similarities to the given traffic pattern of multiple query locations (sketches). This global similarity is jointly computed from individual similarities of the multiple sketches by given weights. User can customize these weights of importance to adjust the contributions from the sketches.

- **Interactive sketch + visual analytics**: The interactive sketches are seamlessly with visual interface for immediate visual analysis. The similarities of different times at different days to the sketched queries are visualized for users to (1) find similar times having the given values of transport attributes at these times, (2) interactively study distributions of top-matching results in different periods (e.g., morning, afternoon), different weekdays, and different weather conditions. Moreover, QuteVis provides a convenient multi-map view as an effective tool for comparative study of top matching times or any interesting times. Users can examine transportation features on any location of the city. Users can easily conduct iterative exploration and visual queries with effective interaction.

Next, we first introduce the urban transport database generated from large, heterogeneous data sources, and then discuss the weighted matching approach for data retrieval.

Then we discuss QuteVis visual interface with sketching, querying, and visual analytics functions.

4.4 QuteVis Transport Database

QuteVis develops a spatial database integrating urban structures (e.g., street geometry, POI locations, and their names), traffic information (e.g., volume and speed over time), mobility information (e.g., taxi pickups/dropoffs and trajectories), and other information (e.g., weather condition).

4.4.1 Incorporating Heterogeneous Data

These heterogeneous data is well organized to well support fast interaction of sketch selection, data querying, and visualization. The street-level traffic data such as the real time traffic speed on urban roads are now provided by many commercial services including Google Maps, HERE Maps, and Baidu Maps. At beginning, we tried to generate traffic databases from their open APIs. However, we found that currently these APIs only provide limited access to traffic information. They do not allow users to download a large volume of historical data to create a transport information database for many months. We hope and also foresee that in the future these data sources will enable public access for more traffic information, so that our system can be integrated with them and used in many cities over the world. Therefore, we use the taxi data sets, which are widely used in urban computing and visualization works, to generate traffic information for QuteVis database. The taxi data also provides human movement information that is important for transport study. We also extract historical weather conditions of the city which are, fortunately, available from many weather services.

In our prototype, we use the datasets from Hangzhou city in China, which has about 9.1 million residents as the capital of Zhejiang province. Traces of 8,120 taxis in three months, Dec. 2011 to Feb. 2012, are utilized. Each trace includes a set of raw GPS sample points and associated attributes (ID, speed, time, etc.). The large volume of the raw data sets are 46.1 GB for Dec., 42.8 GB for Jan., and 44.2 GB for Feb. The city's road network including the geometries of street segments is retrieved from OpenStreetMap. Moreover, the historical weather data of the Hangzhou city is downloaded from an online portal (forecast.io) at each hour of these days. The weather attributes include summary (cloudy, rain, etc.), and other properties such as pressure, visibility, humidity, temperature, etc. They are saved in an independent table in the database. Next, we introduce our approach to process these data and build the QuteVis database.

4.4.2 Generating Optimal Urban Road Network

In the sketch recognition algorithms, the scale of street segments defines the sketch selection resolution. Street segments are also the basic units storing traffic data records for user selection. Therefore, they should not only represent the city road network but also (1) facilitate accurate sketch selection and editing and (2) enable high resolution traffic data storage and management. However, the "raw" road segments retrieved directly from an existing geo-database (e.g., from OpenStreetMap) cannot be directly used in the sketch operations. It is a significant challenge that needs to be addressed. The raw street segments have many problems. First, a raw street segment may be very long and pass multiple road intersections or have a complex geometric shape. Thus, a sketch for selecting a short street part may retrieve a long path or a path with a complex shape. Second, many raw street segments have numeric errors. These problems make it hard for users to interactively select desired streets or paths on the map, leading to frustrating sketching experience. Our goal is to provide satisfying sketching accuracy and experience by regenerating optimized street segments. Our algorithm has two stages. First, we define new street segments to make sure that their endpoints are located at neighbor intersections. This is implemented by (1) finding all geometric intersections from the raw street segments, and (2) defining one new segment between each pair of the neighbor intersections. In computer graphics, T-junction detection algorithms perform similar operations on images [130], while our intersection detection algorithm is applied directly on raw street geometries. Second, in the new set of street segments, we divide a long segment into multiple shorter ones. The purpose is to make each street segment short enough so that people can sketch to select them with high accuracy. This is very important for ensuring accurate object selection in visual studies. In our implementation, a total number of 14,639 street segments are generated for Hangzhou, while the raw road network from OpenStreetMap has 9,764 segments.

4.4.3 Building Traffic Data Cube

All the sample points from taxi traces are indexed by their spatio-temporal attributes. For the time dimension, we assign each point to time windows using two-hour periods of each day (hourly or other finer resolution can be used too). It is also assigned to a specific weekday. Each GPS point is also map-matched to a specific street segment. Consequently, the raw data is stored in a database table whose records are in the form of (p, h, d, w, S_i), where p is the GPS point, h is the hour, d is the day, w is the weekday, and S_i is the corresponding street segment where p resides on. Next we build a traffic data cube by aggregating these raw data records to compute and store the traffic attributes such as speed, taxi pickups, taxi dropoffs, and taxi flow volume, on each of these street segments. Unlike existing work such as Nanocube [37], our data aggregation and caching is conducted on street segments instead of spatial grid cells (of a tree subdivision). This is a unique feature since we seek to retrieve traffic attributes only on streets.

The traffic data cube is constructed by the *GROUP-BY* aggregation of p over the dimensions of (h, d, S_i), and (h, w, S_i), respectively. Then we get the traffic data cubes of (($a_1 ... a_K$), h, d, S_i) and (($a_1 ... a_K$), h, w, S_i). Here ($a_1 ... a_K$) represent the traffic attributes achieved by the *Cube* operation such as Count, Sum, Average, etc. In particular, *Average()* is used to compute average speed from the taxi speeds of a group of points p, and *Max()* and *Min()* are also used to get maximum and minimum speeds. Meanwhile, *Count()* is used to get the numbers of taxi pickups and dropoffs, as well as the volume of taxi flows. In this way, finding these attributes over any time periods of any day (or weekday) can be directly retrieved from the data cube when users specify a sketch condition.

4.4.4 Indexing Street Segments

All the database tables and operations are implemented by a PostgreSQL database. Finding corresponding street segments in the database should be fast enough for prompt response. This is accelerated by utilizing an spatial indexing structure, so that the system can immediately find them, and then retrieve corresponding data records from the traffic cubes. We utilize PostGIS over PostgreSQL, which implements a tree-based spatial indexing structure.

4.5 Similarity Based Data Retrieval

The data query model is defined as:

Definition 1 ***Data Query based on Multiple Spatial Constraints*** *Given urban transport database U, a set of user-specified spatial constraints Q, a query computes the* ***similarities*** *of all data records in U and retrieves them ranked by the similarities.*

In Definition 1, each data record $r(h, d, w, c, T) \in U$ reflects the transport information (T) at a specific time (h) of a date (d) which also refers to a specific week day (w). c is the attributes of weather conditions in the city. The set $T = \{S_i\}, i = 1..N$ is formed by S_i at each of the N street segments in the city, where i is the ID of a street segment. Each S_i is an array of K attributes, $(a_1...a_K)$, representing traffic speed, taxi flow, taxi pickups, taxi dropoffs, etc. Q includes L constraints while each of them $c_j(a_m)$ has the input value c_j at a street segment j, in one of the attributes, $a_m \in (a_1...a_K)$.

To implement the query, a global similarity value S is defined for each data record r.

It is a value between 0 to 1 (or percentage 0% to 100%), which is computed according to the given Q as:

$$S(r|Q) = \frac{\sum_{i \in L} \omega_i d_i}{\sum_{i \in L} \omega_i}, \tag{1}$$

Here ω_i is the weight given to the sketch i. d_i is the similarity between the actual value, $S_i(a_m)$, and the given value, $c_i(a_m)$, of attribute a_m. It is computed as:

$$d_i = 1 - \frac{|c_i(a_m) - S_i(a_m)|}{max(c_i(a_m), S_i(a_m))}, \tag{2}$$

where the rightmost term is the percentage difference of the two values. d_i is the similarity of one spatial constraint (i.e., one sketch). In practice, users can give one constraint by selecting one path, one intersection, or one region, which includes multiple small street segments. The average of actual values over these segments defines S_i.

The weighted global similarities provide hints about how the traffic attributes in different times close to the given pattern. The top matched results with large similarities can form recommendations to users. On the other hand, each d_i is also useful for users to identify the similarity of each sketch. In our visual design, we visualize both S and their constituent factors d_i.

There have been different methods that deliberately define similarity over attribute values varying across temporal cycles including hours of days and days of weeks. These methods cluster transport values and mine the hidden patterns of transport dynamics (see [109, 110, 131] and others in [50]). However, our goal of QuteVis is to provide data records close to users's input values, rather than find the hidden clusters or temporal patterns. Moreover, these relatively complex methods cannot easily response quickly for the sketch inputs when processing big street level traffic data. Therefore, we use the weighted similarity defined by Eqn. 1 and Eqn. 2.

Similarity Categorization: At the beginning of our design, the numeric similarity values are directly color-encoded in visualization charts. However, in our user study, most domain users are not satisfied with the complex visual representations from various similarity values. 5ExP suggested that in real scenarios, most users would like to quickly identify the distributions of similar times to their query. *Meaningful and qualitative* visualizations are preferred as quantitative color mapping requires more effort to read and understand. Based on their suggestion, we predefined categorical metrics of similarities as:

- $0\% \leq S < 50\% \mapsto$ *"Not Similar"*
- $50\% \leq S < 70\% \mapsto$ *"Little Similar"*
- $70\% \leq S < 90\% \mapsto$ *"Very Similar"*
- $90\% \leq S \leq 100\% \mapsto$ *"Extremely Similar"*

In the visual interface, the distribution of these categories in different time windows, weekdays, etc are visualized in different charts (e.g., Fig. 20). Users were happy to see this change of visual design. Moreover, the category descriptions are heuristic and users can set up them by their preference.

4.6 QuteVis System Interface

The interface is designed as a map-centric application where the map plays the role as a canvas in the middle for users to sketch on. This design gives users maximal capacity to draw and edit. Fig. 3 shows the visual interface which includes the following major views:

- A canvas over city map, Fig. 3(1), to facilitate context-aware operations on the city. In **Sketch Mode**, it supports the selection of multiple streets, regions, or paths, where three *sketches* are shown (in the following, we also refer a *sketch* as one selection). To help users to differentiate between multiple sketches, different colors are used to encode multiple sketches on the canvas. Users can also edit the sketches interactively. In **Info Mode**, the map view visualizes traffic data, POIs, and other urban information, and also users can interacts with the map view for their exploratory study. Users can select any region/street to study its traffic behavior and weather in a specific time. The map view can be shown in different styles and it supports smooth zooming and panning operations.

- A control panel with multiple tabs (Fig. 3(2)). The tabs include: (a) **Manage Sketches**: which allows users to define query conditions on multiple sketches, and to control whether the sketches are active in a query; (b) **Study Distribution**: which helps users investigate the distributions of similar times over weekdays, hours, and weather conditions; (c) **Find Similar Times**: which shows a grid heatmap view for users to find and study the similarities of available historical times of days. Those candidate times with the largest similarities are highlighted and put into a top list by default, while users can also select any arbitrary times of their interest and add them to the top list.

- A detail study panel (Fig. 3(3)) to investigate selected times in the Top List. **Top List** are visualized for users to study detail information of each time in the list, including the input value and actual value on each sketch, similarity categories, and the weather condition. Furthermore, **Map Comparison** allows users to observe and compare the traffic behaviors at different times on a multi-map view. This view is coordinated with the major map canvas. Fully functioned map operations on either the canvas map or any of the small maps are well synchronized. Therefore, users can study details of any specific area of the city.

- A top panel (Figure 3(4)) allows users to toggle between Info Mode and Sketch Mode, quickly locate a street/POI by name, re-initialize the system, and set up sketching parameters.

4.7 Multi-Sketch Visual Query on Geo-Structures

Ben Shneiderman predicted that “we will be seeing touch screens used for more applications than ever before” [132]. However, only in recent years we see this prediction becomes true and touch input techniques are being used in visual analytics projects.

Many GIS software (e.g., ArcGIS) and practical map applications (e.g., [133]), now provide touch-based systems. Touch screen device is portable and easy to interact, especially for non-professional domain users and senior users [134]. In our requirement analysis and user study, the urban planners and geographers are all enthusiastic on implementing the visual traffic querying with sketch based operations, compared with traditional mouse-keyboard interface.

QuteVis need to support users to easily define and edit multiple sketches on the map. Its sketch recognition algorithm, unlike general methods, is built up on the geographical or urban structures. Users want to sketch on three different types of entities including street segments, path, and region which are targeted by the recognition algorithms. A unique feature of our approach is that the geometry of our optimized street segments is integrated into the recognition algorithm to promote accuracy and reliability.

4.7.1 Sketch Recognition and Editing Algorithms

Users can draw on the central map canvas (Fig. 3(1)) with a mouse, a touch pen or hand drawing (if supported). In Fig. 3(1), three sketches are selected, which represent three major streets (categorized as primary in Hangzhou's street hierarchy) in downtown Hangzhou. One sketch is a set of small street segments selected from the optimal street network of a city (Sec. 4.4).

A **street sketch recognition algorithm** is used to map the user input on screen to the street segments, which consists of several steps:

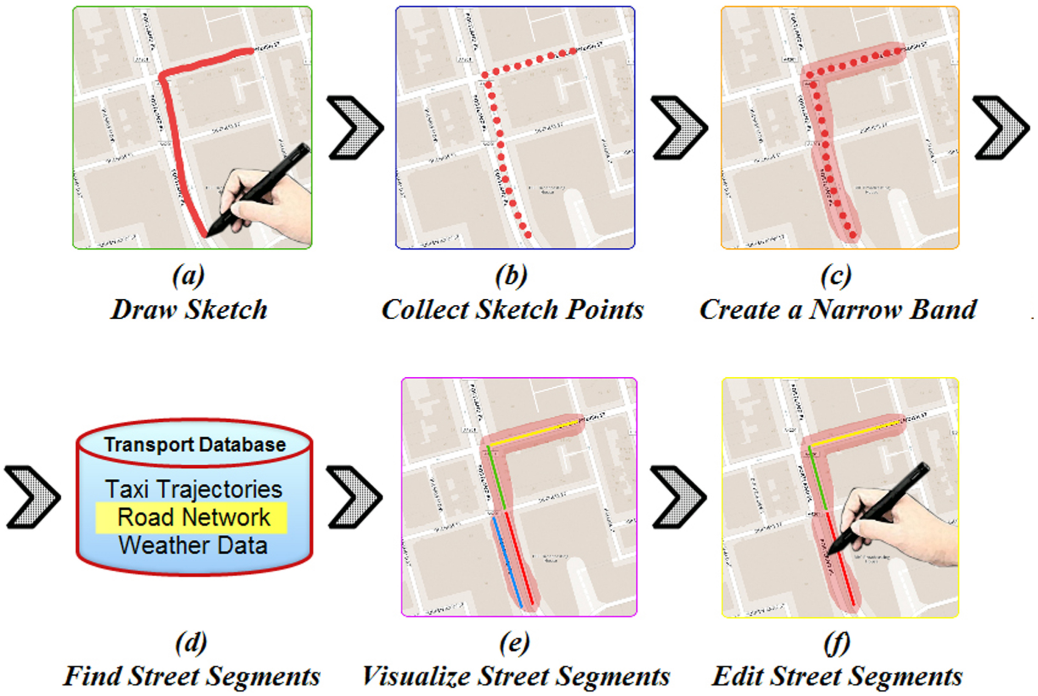


Figure 18: Illustration of the process of path selection and editing.

1. When a user draws on the map (Fig. 18(a)), their touch operation is collected as a sequence of sketch points (Fig. 18(b)). These points are processed to remove obvious errors and noises.

2. A polyline is formed from the points by connecting consecutive input points. Then, a narrow band enclosing this polyline, while following its shape, is generated (Fig. 18(c)). This narrow band mimics the stroke of this sketch. The width of the stroke is defined adaptively according to the zooming level of the map, which plays an important role in defining the accuracy of the sketch selection. In implementation, this stroke band is geometrically represented as a bounding polygon of the polyline.

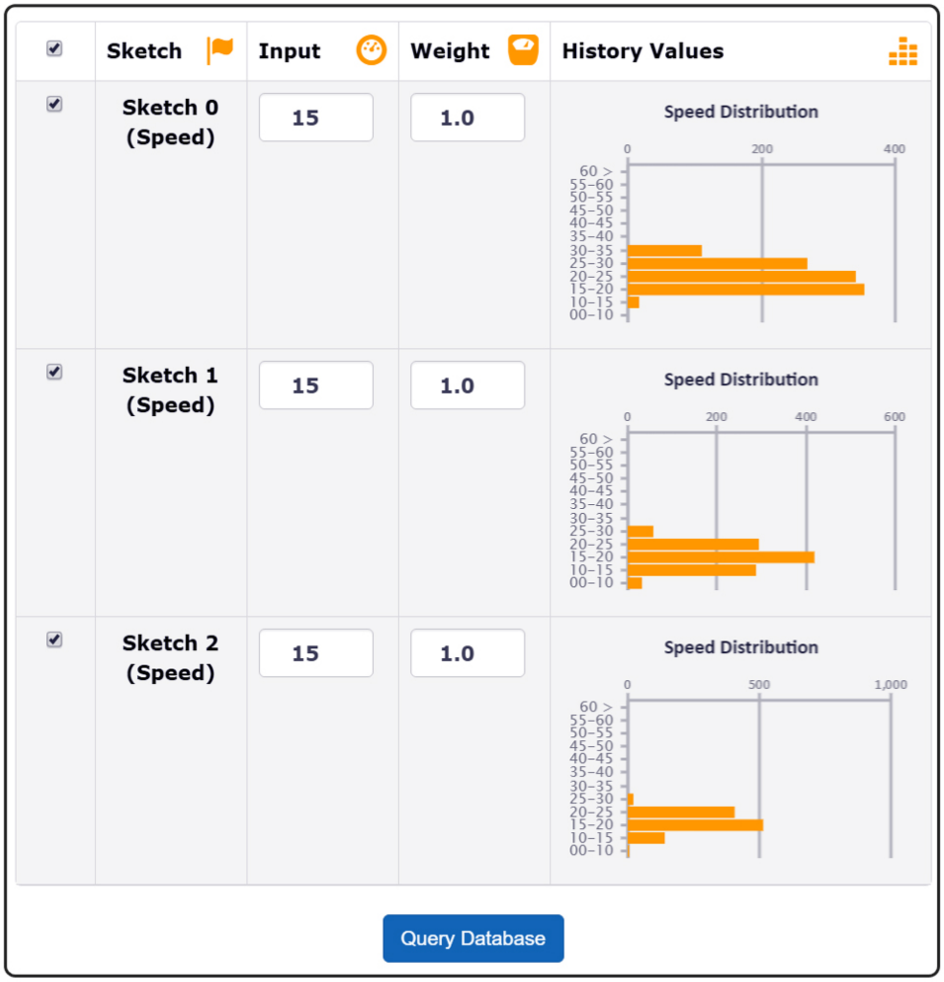
3. This stroke bounding polygon is used to apply a region-based spatial query on the road network (Fig. 18(d)), to retrieve all the inside street segments. For example, two segments in red and blue are selected in Fig. 18(e), which represent two separate lanes in opposite directions on the same road.

4. However, when users drawing on the map, the input stroke may collect unintended street segments, such as the blue segment in Fig. 18(e). On the other hand, sometimes a stroke may not cover a necessary segment which is needed. Then editing operations (Remove and Add) is required for users to edit any sketch by adding and removing related street segments. As shown in Fig. 18(f), users can click to remove the blue segment.

Finally, the selected street segments form the result sketch. This algorithm allows users to choose a single street or a long path between two locations as well, which is flexible for traffic study.

In addition, a sketch can also be defined as a region in the city. In a **region sketch recognition algorithm**, the user-input points still create a polyline. Then we make this polyline as a closed curve by connecting all points, so that it forms a polygon. We then use this polygon to query the road network and all the street segments inside are selected. Similar editing operations are available to refine the sketch results.

All these operations are available on traditional devices as well as touch input devices.



☑	Sketch	Input	Weight	History Values
☑	Sketch 0 (Speed)	15	1.0	Speed Distribution
☑	Sketch 1 (Speed)	15	1.0	Speed Distribution
☑	Sketch 2 (Speed)	15	1.0	Speed Distribution

Figure 19: Sketch Manager: Users manage the list of sketches.

4.7.2 Managing Multiple Sketches

After users draw multiple sketches on the canvas map as in Fig. 3(1), the sketches are managed on the Sketch Manager view as shown in Fig. 19. For visual queries, users need to specify an input for each sketch with the value of their preferred traffic attributes (speed, volume, pickups, and dropoffs). For weighted matching, each sketch has a default weight which can be adjusted by users. For instance, a large weight may be given to a major street, while small weights are assigned to secondary streets. The default weights are predefined by our experts based on the hierarchical levels of sketch streets (e.g., 1.0 for highway and primary roads and 0.7 for secondary streets). For each sketch a bar chart is shown on the right, which provides visual cues to users about the historical values of the corresponding attribute. For example, users become aware of the normal traffic speed or outliers speed of a sketch, so that they can easily speculate the input value. Fig. 19 shows the query traffic speed at 15 km/h for the three primary roads, which are low to medium speed for them based on the histograms.

4.8 Visual Analytics Functions

We describe the visual representations and functions by assuming that an urban analyst, named Amy, makes three sketches on the canvas as shown in Fig. 3(1). Then she defines the slow traffic speeds (15 km/h) for these sketches (Fig. 19). She wants to study how such patterns occur in history traffic database of Hangzhou.

4.8.1 Studying Distribution of Similarities in Bar Charts

First, Amy discovers the distributions of historic times on the Study Distribution view (Fig. 20). It shows three different types of distributions. The colors in the bars represent the categories of similarity. Here a set of popular color palettes are made available for Amy to choose from a menu. In each chart, Amy can click on each of the icons of these categories to make them visible or invisible. In Fig. 20, Amy wants to see the *extremely similar* times only shown as blue bars.

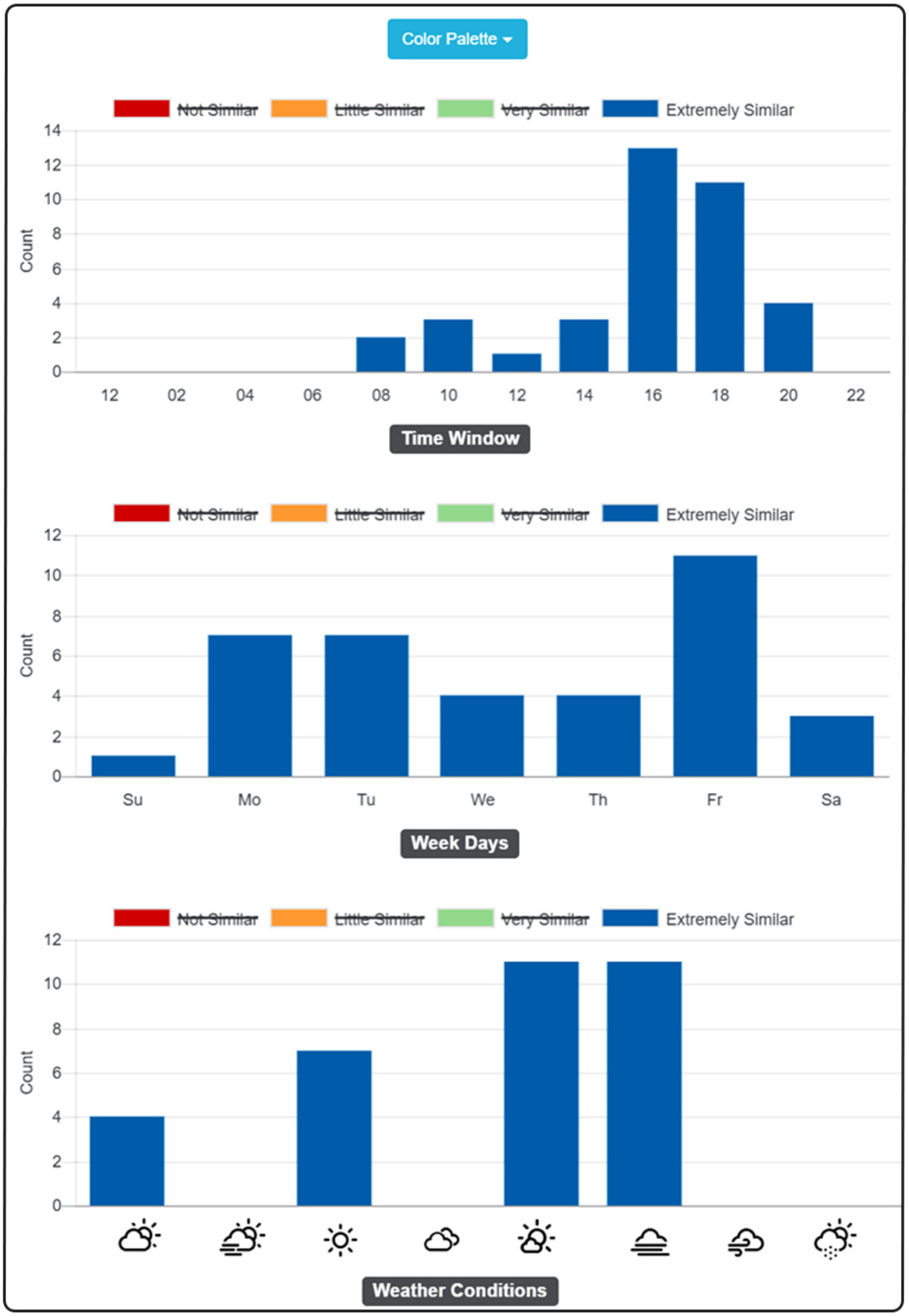


Figure 20: Study the distribution of query results over day-times, weekdays, and weather conditions.

On the top bar chart, the x-axis represents the times (in two-hour time windows) of a day, which is 12am-2am, 2am-4am, and so on. The y-axis shows the count of data records that falls into these time windows. The Hangzhou database has 91 data records for each time window, that is, 91 days in the three months. Amy can find the distribution of similarities in each time window. In her case, those times when the three sketched streets all have a low speed 15 km/h only appear during the daytime. The afternoon rush hours (4pm-8pm) are the most frequent times such situation happens. The morning rush hours are not comparable to the afternoon, which hints a mobility pattern of residents.

On the middle bar chart, the distribution over weekdays from Sunday to Saturday is perceived by Amy. Each weekday has 156 different data records in the database (12 time windows per day in 13 weeks). Amy finds that her input traffic situation on the three streets does not appear often on Saturday and Sunday, while Sunday has the fewest counts. Meanwhile, Friday is the weekday that the three streets suffer most from low speed.

Finally, in the bottom, Amy reads the distribution of data records on different weather conditions. For example, she can find the highest bars represents cloudy and foggy (and polluted) days.

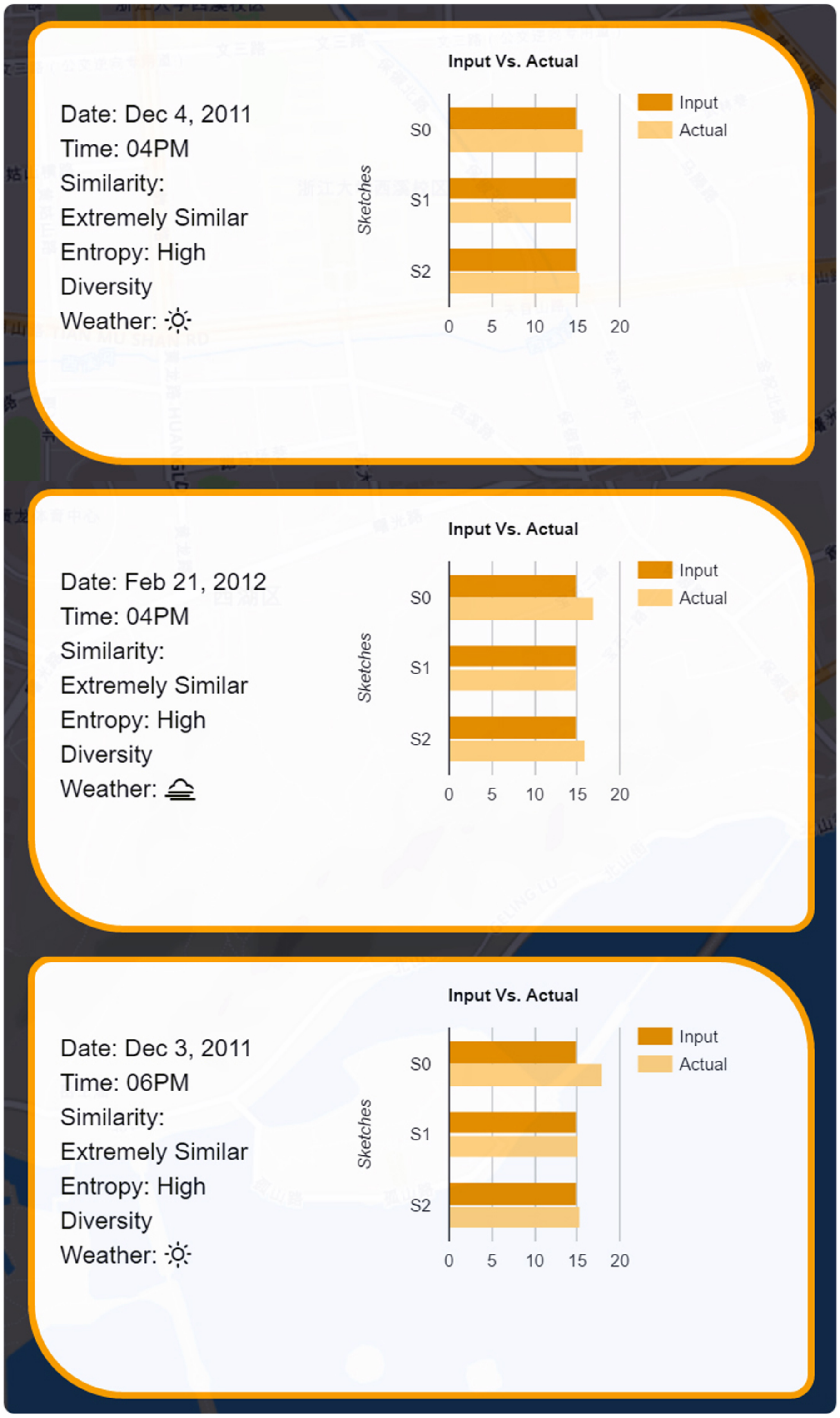
4.8.2 Finding Similar Times in a Grid Heatmap

Amy then wants to find details of the matching times in different days. As shown in Fig. 3(2), a grid of dots on the right panel shows a heatmap of similarities. Each dot presents one time window in a specific day in a specific month (a related calendar view has been used in [135]). Its size represents the similarity value, S in Eqn. 1, of this time to the query. The top ten matched times are highlighted with a green boundary so that Amy can immediately find what times are in the Top List of matching.

In Eqn. 1, S is indeed computed from a set of d_i, each of which is the similarity of the actual value compared to the given query value on one sketch. Two dots may have the same size, but they have different d_i values over different sketches. For example, one dot may have small differences on Sketch1 and Sketch2, and the other dot has a larger difference on Sketch1 but no difference on Sketch2. On the heatmap, we use the color of these dots to give users some hint of this diversity. In particular, we map the entropy, H, of d_i values to the dot color, which is computed as:

$$H = \sum_{i \in L} p(d_i) \log p(d_i) \quad where \;\; p(d_i) = \frac{d_i}{\sum\limits_{i \in L} (d_i)}, \tag{3}$$

Here L is the number of sketches (constraints). As shown in Fig. 3(2), two top matched dots are indicated by two arrows in red. Their corresponding times are 6pm at Feb. 10 and 4pm at Feb. 21, both of which have a large similarity considering Amy's query. However, the visualization shows that the entropy H is lower in the former time than the latter time. Amy may pay special attention to study the dot for Feb. 21 to find the differences. In particular, Amy can hover the mouse over the dots to find such details. Please note users can freely change the color scheme with their preference. She can also click on any dot to add it to the Top List. The details of Top List is given in a list view on the left panel.



Date: Dec 4, 2011
Time: 04PM
Similarity:
Extremely Similar
Entropy: High
Diversity
Weather: ☼

Date: Feb 21, 2012
Time: 04PM
Similarity:
Extremely Similar
Entropy: High
Diversity
Weather: 🌫

Date: Dec 3, 2011
Time: 06PM
Similarity:
Extremely Similar
Entropy: High
Diversity
Weather: ☼

Figure 21: Study the Top List for details of interesting matches.

4.8.3 Examining Top List of interesting matches

After Amy overviews top matches and possibly selects her interesting times, she can further investigate their details in the Top List view. As shown in Fig. 21, this view shows a set of cards of all items in the list. She can read one card about the specific time, the date, and the weather. On the right, she can find the differences between the input values and the actual values of each active sketch. Amy can further click one card to show the traffic information on the central map canvas. She can go to the multi-map view for map comparison of traffic behaviors in any specific location.

4.8.4 Comparing Query Results with Multi-maps

In Fig. 22(a), the multi-map view is shown in the left panel. Each map shows the traffic information of one specific time in Top List. The colored lines indicate the traffic information on the roads. Roads with smooth traffic flows are marked in green, roads where congestion is moderate are in orange, while those that are congested are shown in red. Amy can select different attributes to be shown on these maps. She can click one map to duplicate (enlarge) it in the central map canvas. Zooming, panning and other map operations can be performed on any of these maps, including the central canvas. All the maps are coordinated so that Amy can compare any area and street in different times. She can also give a street name or POI name, in the panel of Fig. 3(4), to quickly locate them in these maps. Moreover, Amy can go back to the sketch mode so that she can change or adjust sketch conditions, activate more sketches in a new query, and make other investigations in an exploratory manner. The QuteVis system thus helps domain users as Amy to perform urban data analysis tasks.

4.9 Evaluation

4.9.1 Evaluation Team

After QuteVis was fully implemented, we invited a team (named CTeam) of 23 users (15 males and 8 females) to help us with case studies. CTeam had users who were familiar with Hangzhou city in China to provide meaningful use cases. Among these users, there was an active urban geography researcher who worked as an urban transportation planner in Hangzhou city for several years and two local residents who lived in Hangzhou for more than 20 years. Moreover, to further validate the QuteVis system, we interviewed a group (named FTeam) of eight domain experts (5 males and 3 females) whose ages ranged from 27 to 60. Five of them were experts in the areas of urban planning and transportation, GIS, regional economy, and geography, while three of the five worked as an urban planner before. One of them had 25 years of experience as a transportation planner working with private firms and public agencies. The other three domain experts were the researchers and scholars with earned Ph.D. degree in the areas of GIS, urban planning, and geography. FTeam used QuteVis and provided subjective feedback in an interview.

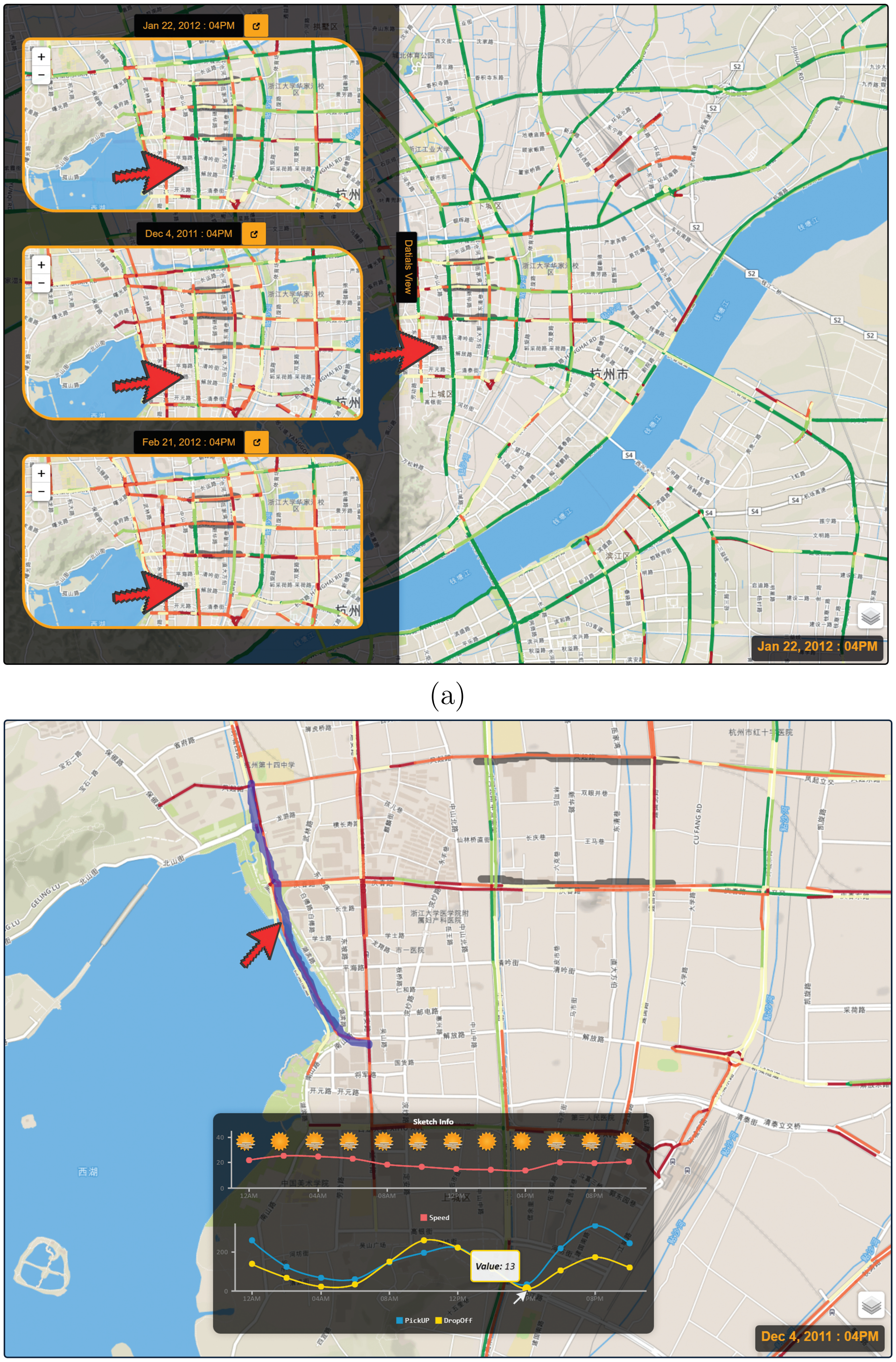
4.9.2 Case Studies

Three use cases from CTeam are presented below. To better illustrate them, we continue to use the name Amy for description.

4.9.3 Traffic Speed on Three Major Roads

While introducing the system interface in Sec. 4.6, we have shown one case where Amy defines **Pattern 1** involving slow traffic speeds, *the what*, on three major downtown streets, *the where* (Fig. 3). Fig. 22(a) further shows the multi-map view, where three maps are used by Amy to compare the traffic situations, *the how*, at 4pm-6pm of Jan. 22, Dec. 4, and Feb. 21, respectively. The first one (Jan. 22) is enlarged in the central canvas. For example, four red arrows point to the same street, Zhonghe road in these maps. The traffic on this road at Jan. 22 is smooth shown in green. But at the other two days, this road has slow traffic speeds shown in red/yellow. Amy discovers the reason: Jan. 22 is in the holiday period of Chinese New Year in 2012, when many residents stayed at home or went back to their home towns.

In Fig. 22(b), she further zooms in to an area on the east bank of the West Lake. She sketches on the canvas map to select one specific street (highlighted in purple) she wants to study. In the Info Mode, an information visualization window pops up to show the traffic behavior of this road. Three lines on the two line charts display the varying traffic speed (red), taxi pickups (blue), and dropoffs (yellow) in a whole day, respectively, together with the weather conditions. For Dec. 4, Amy finds that the numbers of pickups/dropoffs decrease at 4pm-6pm to the lowest values during the day, when the three sketch roads have slow speeds as she speculates. This shows to her that the taxi activities on this road are potentially affected by the three major roads in downtown Hangzhou.



(a)

(b)

Figure 22: Investigate traffic behaviors at different times: (a) the multi-map view that helps users to compare and study different traffic situations; (b) a pop-up window to display the traffic behavior of the specified road.

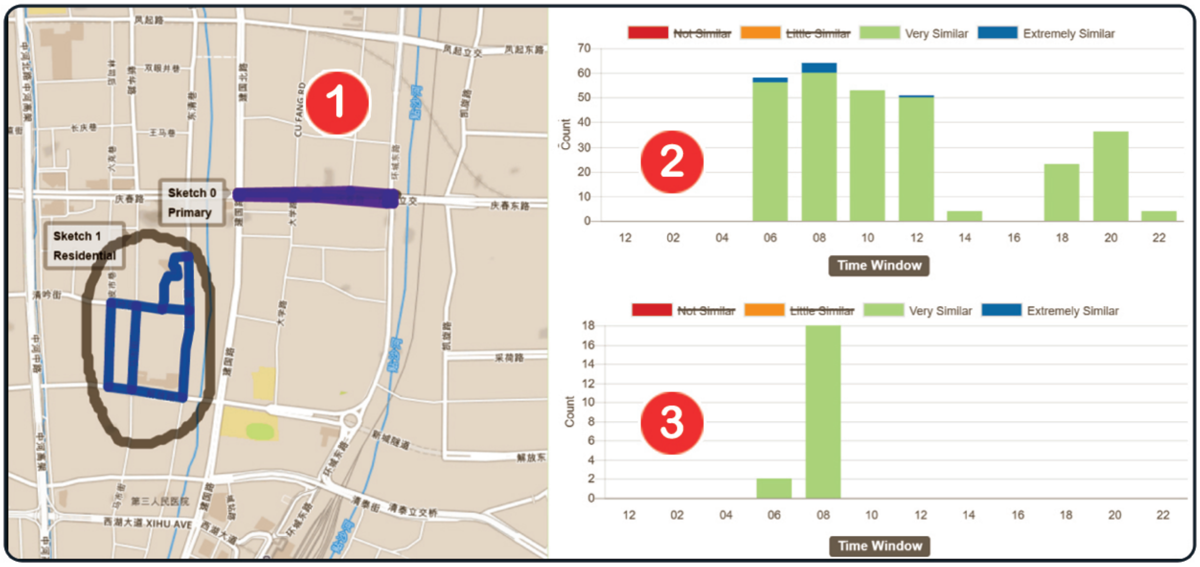
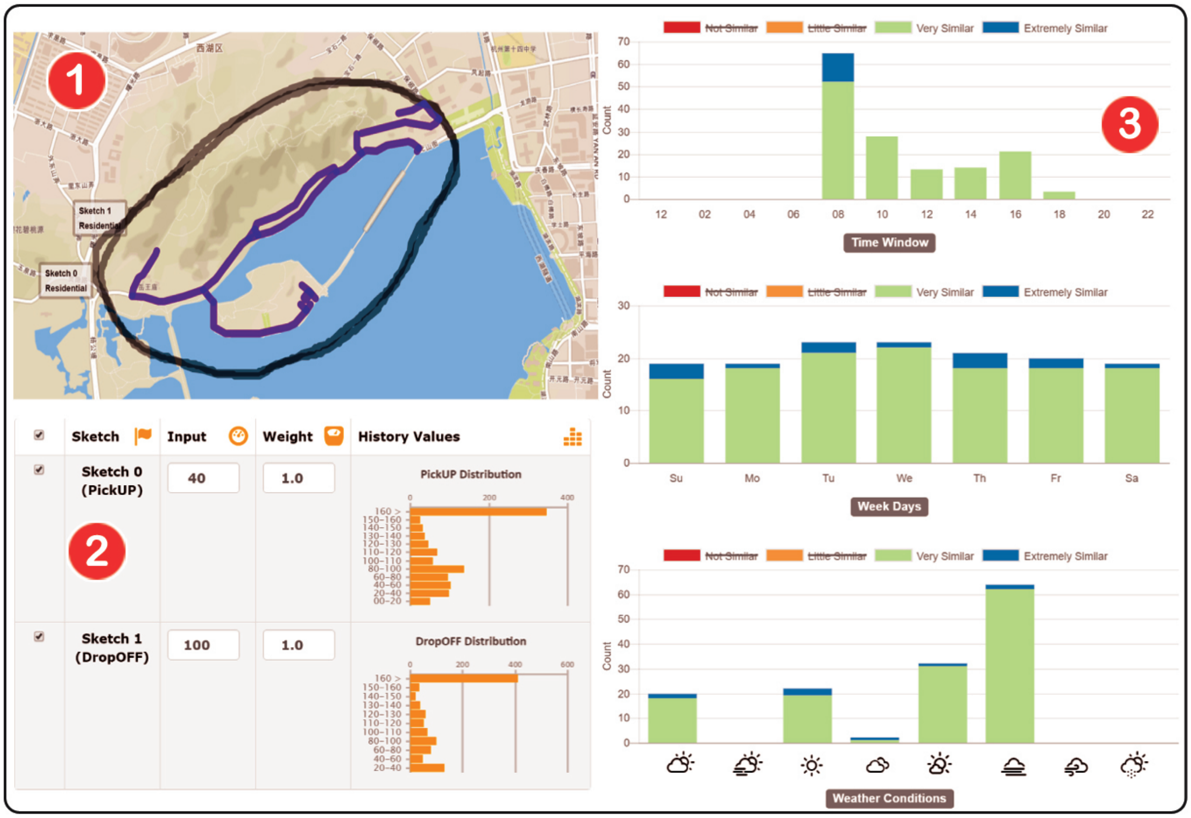


Figure 23: Study traffic pattern involving the speed of a major road and taxi activities in a hospital region. (1) User specified two sketches; (2) The queried pattern happened mostly from 8am to 2pm; (3) After changing query values, the pattern happened only during morning hours 8am-10am.

4.9.4 Traffic Speed and Taxi Activities of a Hospital Region

Fig. 23 shows Amy studying **Pattern 2** which involves traffic speed of a major road and taxi activities in a residential area close to it. In Fig. 23(1), Sketch0 shows a primary street segment, Qingchun road, which connects the west downtown area to a primary street, East Loop road, of the city. Sketch1 defines a residential region which mainly includes two big hospitals of the city. Amy specifies a normal speed, 30 km/h, on Qingchun road, and a value of taxi dropoffs as 120, which indicates a lot of passengers arriving this region in a 2-hour window. Then, Amy observes the distribution charts of the query results. As shown in Fig. 23(2), this queried pattern happens mostly during the daytimes from 8am to 2pm. For comparative study, Amy adjusts her query by changing the input speed on Qingchun road to a slow value at 12 km/h. Fig. 23(3) shows that this situation happens only during morning hours 8am-10am. Amy realizes that the counts of records having such a situation are much smaller (e.g., 18 at 8am-10am) than the counts of her first query (e.g., around 50-60 from 8am to 2pm). It means when Qingchun road has a slow speed, taxi dropoffs can rarely have a large value. In comparison, when Qingchun road has fast speed, taxi dropoffs happen a lot in the region. By this comparison, Amy may concludes that the taxi activities in the hospitals are affected by the speed of Qingchun road. This finding confirms Amy's suspicion at the beginning. She can further use map views to study several specific times for traffic behaviors on other surrounding locations.



☑	Sketch	Input	Weight	History Values
☑	Sketch 0 (PickUP)	40	1.0	PickUP Distribution
☑	Sketch 1 (DropOFF)	100	1.0	DropOFF Distribution

Figure 24: Find a traffic pattern involving taxi activities in a tourism area. (1) User specified two sketches for taxi pickups and drop-offs in a tourism area; (2) Taxi pickups and drop-offs had similar histograms; (3) A pattern was frequent in the morning 8am-10am and happened for all weekdays.

4.9.5 Taxi Activities of a Tourism Region

Fig. 24(1) shows another task to investigate **Pattern 3** involving taxi pickups and dropoffs in a tourism area. Sketch0 and Sketch1 include the same region, West Lake Scenic Area, which is one of the most famous tourist sites in China. After sketching, Amy finds that the taxi pickups and dropoffs have similar histograms from Fig. 24(2). She wants to find a situation when the number of pickups is low at 40, and the number of dropoffs is high at 100. Fig. 24(3) shows that such a pattern is frequent in the morning

8am-10am. It is meaningful as the morning time is when visitors arrive for their tour. Meanwhile, this pattern happens for all weekdays. This is also reasonable for this famous place, whose visitors come from all over China and do not show big weekday/weekend differences. Amy further study specific days of interest through other visualizations.

4.10 Domain User Study and Feedback

To validate QuteVis system, We conducted a preliminary study with the domain experts in FTeam. Our key goal was to justify the usability of QuteVis, with its sketch+visualization interface, and to gain knowledge about the limits and to identify future directions. First, we explained the system to our subjects (i.e., FTeam members) with a detailed description of the functions and interface. Then, we allowed each subject to use and explore the system on a touch screen for about 20 minutes. After the preparation, we asked them to draw two sketches on the map, and then answer the following questions for four different VA tasks:

- **Q1:** *Find similar traffic times:* Which weekday had the largest similarities?
- **Q2:** *Identify associated attributes:* Which weather condition had the largest similarity?
- **Q3:** *Ranking:* What were the top 3 times and dates having your given conditions?
- **Q4:** *Comparison:* In the top 3 times, which one had the worst traffic conditions?

These questions were designed because (1) they included major VA tasks supported by QuteVis, (2) they required the subjects to use all QuteVis visual functions including multi-sketch and all visual charts and map views. In particular, Q1 and Q2 were linked to the distribution charts and the grid heatmap, Q3 required the use of the Top List for ranking, and Q4 involved the multiple map comparison. Based on these tasks, all subjects gained knowledge about QuteVis and provided their evaluation. After these tasks, we conducted a subjective feedback interview with the FTeam experts. They were asked to fill a survey form about the system with respect to the following three aspects:

- **A1:** The usefulness of QuteVis system in transportation planning and traffic management studies.

- **A2:** The convenience of the sketch functions for supporting users to query and investigate times and days that have "similar" traffic patterns to user input.

- **A3:** The effectiveness of our sketching, visualization and interaction functions to show, filter, and compare the results.

Based on their answers in the survey, all experts agreed on its potential use in the field of urban planning, transportation, business marketing and education. Some specific comments wrote: "The input of focal speed is very useful for planners because any given segment has a corresponding speed limit (can be used here) or a safe speed according to the user", "The system reflects congestion of traffic in the city that helps the investor to plan investment locations", "The similar traffic patterns are useful to compare two different areas, for example downtown vs suburban areas". One expert mentioned "I want to call it a platform instead of a tool because domain users and researchers can get many hints of how to use the trajectory data along with the urban streets in a more intuitive way. In other words, the current default setting can allow the user to become familiar with the data from multiple perspectives of transportation such as speed, streets layout, urban structure, and day/time in a comparative context." Then, they unanimously agreed that the system interface is intuitive, very friendly and easy to use. They were more excited about the tools that allow them to do multi-sketch on the map especially by using a touchscreen. One expert wrote "The sketch function plays the major role in this system, not only because it is convenient to use given so many touching screen facilities, but also it can be useful in different contexts, from novice users (e.g. education environment) to domain expert (e.g. urban planners)." They also liked the multi-map view to compare and study the traffic situations in different time windows.

Meanwhile, the experts pointed out some drawbacks and gave us valuable suggestions to improve the system, such as allowing users to visualize similarity categories from *not similar* to *extremely similar*, which we have implemented in the system. In addition, one useful comment was to "use more dimensions of trajectory data or related urban streets" where users can provide more diverse constraints in their queries. Another comment was "it will be great to have other transportation data such as public bus and subway data", which will be one of our future works. One expert suggested to add an interactive tutorial to the system to teach first time users. Another suggestion was to add more descriptions to the icons in the top panel. They also suggested to add a view to study the real-time traffic situation on the city if data is available. They would like to see this tool to be combined with traffic predictions, which is our future work to combine the system with commercial map service APIs.

4.11 Chapter Summary

We have developed a visual analytics approach, QuteVis, for urban planners, traffic analyzers, and other practitioners to visually query a transport database with their speculated patterns. They can visually study the times and days such patterns may occur. The system supports hypothetical study to answer users' questions like: "does this traffic pattern happen? what are the times it happens? what affects the pattern? and when it happens, what are the situations in other locations?" The visualization system provides intuitive, easy-to-use interface and tools so that users can perform their investigation easily and efficiently. We have also addressed data management issues with spatio-temporal data and constructed a transport database from heterogeneous data sources with optimized spatial indexing. We have collected feedbacks from a group of domain experts that not only justify the usefulness and efficiency of the system, and provide valuable suggestions for our future work.

CHAPTER 5

An Open Source TrajAnalytics Software for Modeling, Transformation and Visualization of Urban Trajectory Data

Nowadays, a large amount of urban trajectory data sets are collected by transportation administrations, companies, and researchers. Some of them are available for public use in research, such as the Beijing city taxi data [136], the Rome Italy taxi data [137], and the New York City taxi data [138]. More trajectory datasets are not publicized but used by researchers in their studies and publications including taxis [87], public transits [139], human paths [140], etc. In the long run, we will see more and more such data with the widespread use of trajectory recording devices and systems. Understanding and analyzing such large-scale, complex data is of great importance to enhance both human lives and urban environments. Therefore, exploratory visual analytics software is needed to study taxi trajectories with efficient user interaction and instant visual feedback. In this chapter, we present TrajAnalytics, a system that integrates scalable data management and interactive visualization with a powerful web-based computing platform. TrajAnalytics provides exploratory data visualization tools for researchers, administrations, practitioners and general public to understand the data and to reveal knowledge intuitively.

5.1 Introduction

Advanced sensing technologies and computing infrastructures are producing massive trajectory data of people and vehicles in urban spaces at an unprecedented scale and speed. With the prevalent GPS, Wi-Fi, Cellular, and RFID devices, population mobility information is accurately recorded as the moving paths of taxis, fleets, public transits, and mobile phones. The information can be utilized in the studies of urban system, environment, economy, and citizens to optimize urban planning, improve human life quality and environment, and amend city operations. In particular, the urban trajectory data can play an important role in the assessment and planning of transportation infrastructures and policies. For example, major problems in modern cities, such as traffic jams, unbalanced capacities, and frequently occurring accidents, can be attributed to improper road planning, maintenance, and traffic control. Researchers and analysts need to assess such situations in transportation studies. Conventionally, the tasks are conducted by (1) identifying the factors that influence transportation and studying their effects by qualitative approaches through empirical models [141] or survey methods [142], and (2) using simulation products (e.g., EMME [143]) which provide quantitative results of modeling equations to evaluate road networks, where users have to specify complex road attributes and trial-and-error processes are demanded. In contrast, the emerging urban trajectory data provides real situations from which the statistics of real traffic flow can be extracted and city-wide transport patterns can be discovered. Exploiting the emerging data can play a transformative role in transportation-associated research by offering domain experts, researchers, and decision-makers unprecedented capability to conduct data-driven studies based on real-world information. Robust, easy-to-use software enabling effective exploration of the data is direly needed and will contribute to building capacity in seeking solutions for the social, economic, and environmental challenges facing our communities.

5.1.1 Visual Analytics Software is Needed:

The trajectory data records realtime moving paths sampled as a series of positions over urban networks. Rich and heterogeneous information can be associated at each position, including human and vehicle attributes, geographical features, business/urban information, and more. Such data is big, spatial, temporal, dynamic, and unstructured. For example, a taxi-tracking system (with 5,000 cabs in San Francisco) collects 7.2 millions GPS points each day [144]. To extract deep insights from the data, researchers must conduct iterative, evolving information foraging and sense making and guide the process using their domain knowledge. Iterative visual exploration is one key component in the processing, which should be supported by efficient data management and visualization tools. Therefore, transportation researchers demand a handy and effective visual analytics software system which integrates scalable data management and interactive visualization with powerful computational capability. In order to support general users, we have developed an open source software system named TrajAnalytics. It offers data management capability and support various data queries by leveraging web-based computing platforms. It allows users to visually conduct queries and make sense of massive trajectory data.

5.1.2 No Publicly Available Visual Analytics Software Exists:

Conventional transportation design software, such as TransCAD [145], Cube [146] and EMME [143], provides platforms for urban transportation forecasting, planning and analysis. Domain researchers can build transport models, perform simulations, and create simple visual representations of the results for analysis and presentation. However, these software packages are not developed for data-driven analysis utilizing real-world trajectory data. Urban computing has emerged recently in the data mining community to advance discovery of knowledge from a variety of data including trajectories [44]. However, there is no visual analytics software devoted for domain researchers to utilizing the data. As stated in the latest survey [44], "when facing multiple types and huge volume of data, how exploratory visualization can provide an interactive way for people to generating new hypothesis becomes even more difficult. This is calling for an integration of instant data mining techniques into a visualization framework, which is still missing in urban computing." General-purpose information visualization software (e.g., Tableau) does not specifically support the trajectory data. Visual analytics approaches have been conducted for spatio-temporal data. For instance, Drs. Natalia and Gennady Andrienko have contributed many technologies with some prototype systems like V-Analytics [147]. In addition, a few visualization projects have been developed (e.g., [40, 60, 86, 87]) based on the trajectory datasets. However, a significant gap remains between the usability of these approaches and what is needed in transportation studies because:

- These approaches are not developed to provide publicly accessible and easy-to-use software for domain researchers with big urban trajectory datasets.
- The commonly used data models are not designed for scalable and interactive visualization of urban trajectories. For very big datasets, visual analytics demands a distributed computing environment and the power of parallel computing. Specific data management and computing models are necessary in such environments.
- The urban networks of roads and transits are not specifically utilized. Traditional grid-based spatial databases are commonly used to manage the data in Euclidean space [30,112]. However, the trajectories are built on the networks, which distribute sparsely and unevenly over the 2D Euclidean space. For example, a city hub of roads recording a large amount of trajectories falls only in a very small grid cell. The grid-based methods become inefficient in such cases. More importantly, transportation studies are mostly based on the networks which should be easily accessed in data models.

TrajAnalytics is a open source visual analytics software, which integrates scalable data modeling, transformation, management, and interactive visualization within a powerful web-based computing platform. TrajAnalytics provides exploratory data visualization tools for researchers, administrations, practitioners and general public to understand the data and to reveal knowledge intuitively. It contains three major modules: (1) Data loading and transformation; (2) TrajBase database; (3) TrajVis visual analytics interface. The system is developed in two forms: a cloud based software and a local version, to fulfill the requirements of many real world users. It has been released for open access.

Our major aims in this chapter are two-fold: (1) presenting a new software system for practitioners of urban trajectory data, which includes specific designs in data modeling, transformation, and visual metaphors; (2) sharing our experience during the design, implementation, and evaluation processes, which can contribute to open-source visualization software design practices in the future.

We have conducted comprehensive requirement analysis from many domain users to identify their need for a trajectory visualization software. Different from developing a new and specific visual analytics technique, the key challenge for the TrajAnalytics software is to combine visual analytics functions with data engineering algorithms and tools that can fulfill the demands of novice visualization users. The challenge is addressed in TrajAnalytics through the design and seamless integration of data modeling and visualization functions. The major contributions include:

- A *Geo+Trajectory model* is developed to incorporate different types of raw data and match them to geographic structures defined as street geometry, region geometry, or Points of Interest. The modules of data loading and map matching enable users to upload and transform their own data easily.

- A *TrajBase database* is developed to store both the trajectory data and the geographic data. We design effective spatio-temporal indexing algorithms of these data to support interactive visual queries.
- A *TrajVis visual interface* is designed which allows users to perform interactive queries and visualizations of multiple trajectory attributes in geo-contexts. It supports common tasks for urban trajectory study and is extensible for adding more visual functions.

Designing a useful toolkit for domain users cannot be successful if the users are not completely involved. Many domain experts have participated the development in multiple phases. In particular, we have conducted a comprehensive user evaluation which focuses on the usability of the system. To support widespread test, we have developed a web-based system to collect feedbacks. We report our process of software evaluation and share our gains.

5.2 Related Work

5.2.1 Data Driven Urban Transportation Studies

Transportation studies develop and evaluate strategies that improve safety, mobility, and sustainability in transportation systems and enhance the ability to construct, maintain and operate transportation infrastructure [148, 149]. It is a multi-disciplinary field comprising researchers from engineering, geography, and social and behavioral studies.

Understanding urban transportation systems is a fundamental task for a variety of research directions. Exploring patterns and trends of intra-urban human mobility advances the understanding of urban dynamics and reveals socioeconomic driving forces [150–152]. Location-aware devices are widely applied in urban studies [153–155]. Location data from cell phones is used in urban analysis in Milan [155] and Rome [156], Italy. The increasing availability of GPS data has greatly facilitated the study of street networks [157–159]. Urban traffic flow can be viewed as transportation demands which are aggregately distributed in street networks [160]. The floating car technique has been used by intelligent transportation systems to obtain its positional information [161, 162]. Hence, taxis often serve as floating cars to obtain human mobility data and examine real-time traffic status and individual behaviors [157, 163–165].

5.2.2 Data Mining with Urban Trajectories

In the field of data mining, trajectories of human and vehicle motions are used to discover knowledge from large-scale datasets [44]. Utilizing the trajectory data has been divided into three main categories: the study of the collective behavior of a city's population, the traffic flow, and the operators (e.g. drivers) [74]. In particular, vehicle trajectory data has been used in traffic monitoring and prediction [75], urban planning [77], driving routing [5, 49], extracting geographical borders [78], service improvement [79], energy consumption analysis [80], and dynamic travel time estimation [166]. Large-scale mobile phone data with GIS information is used to uncover hidden patterns in urban road usage [159], find privacy bounds of human mobility [167], estimate travel time [113] and infer land use [168]. Public transit trajectories are used in bus arrival time predictions [81], user's transportation mode inference [81], and travelers' spending optimization [82]. NEAT [169] studies trajectory clustering over road networks by considering traffic flows together with segment densities and connectivity. In the aforementioned studies, researchers have successfully used the emerging trajectory data from different aspects. Data visualization is a necessary component in their research, which usually costs them much effort. TrajAnalytics further advance the state of the art by providing an interactive visual analytics system, so that they can explore the data and generate/test new hypotheses more conveniently.

5.2.3 Trajectory Visualization

A large number of approaches have been proposed to visually explore movement data (see [1] for a recent survey). Many of them are focused on the origins and destinations of the trajectories, such as flow maps [52], Flowstrates [53], OD maps [54], and visual queries for origin and destination data [40]. Other work visualizes trajectories using various visual metaphors and interactions, such as GeoTime [55], TripVista [56], FromDaDy [57], vessel movement [58], route diversity [59], and more [40, 60, 170–174]. Some of these approaches coordinate multidimensional visualization and map views [40, 56, 60]. These visualization approaches are mostly designed for specific tasks in understanding the data. Many of them are highly integrated with data mining algorithms to help users find hidden patterns. For example, Drs. Natalia and Glennady Andrienko and their colleagues in Europe have proposed multiple approaches to extracting meaningful clusters from trajectories and used them in visualization [84,175]. More importantly, the existing work has not created an available software system for wider users. Instead, TrajAnalytics is general-purpose and aimed to provide researchers and practitioners with a convenient platform, where they can conduct basic visual analytics tasks based on data queries and spatial/time aggregations. The existing techniques can be included in our software, so that the achievements in visualization research can be quickly and widely utilized by domain researchers. On the other hand, our system can be employed by visualization experts to study the nature and patterns of the trajectory data, so that their advanced visual design for specific applications can be expedited.

5.3 *Geo+Trajectory* Data Model

5.3.1 Point and Trajectory

Urban trajectory data records human or vehicle traces, as a group of positioning points, in a geographical space. Each point usually includes spatial location, time, and other attributes. Here the spatial location usually is a pair of longitude and latitude in trajectories recorded by GPS devices. In some cases, it may also be a street address or an urban area acquired from other means. These locations may be transferred to GPS location by mapping them to their center points. As shown in Fig. 25, each point p_j has the attributes Latitude, Longitude, Time Stamp, and Speed, Type and other associated attributes. The first three are required attributes for TrajAnalytics.

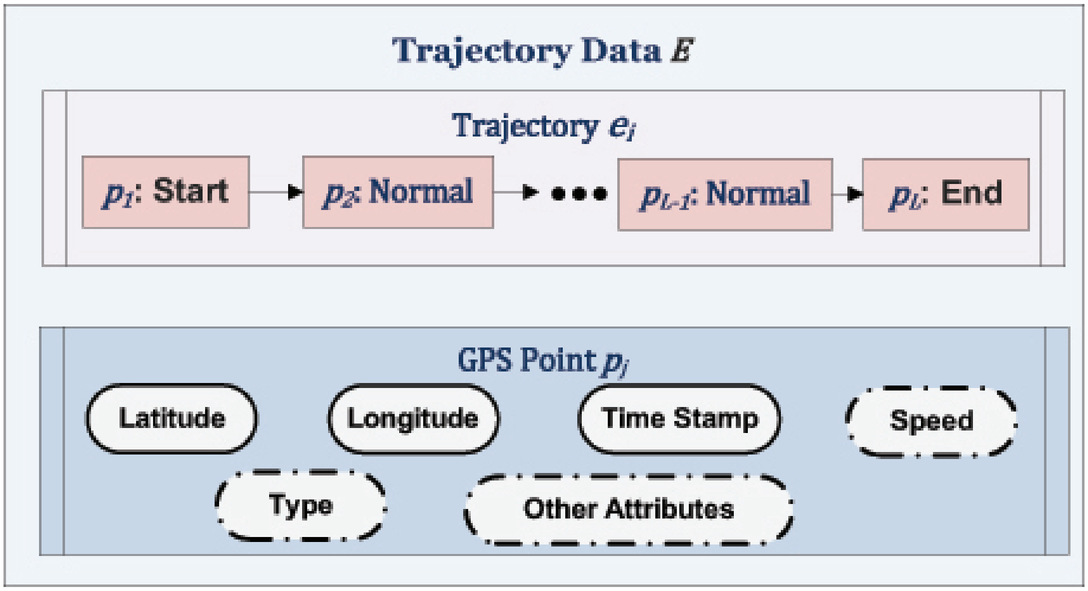


Figure 25: *Urban trajectory data.*

A continuous trajectory may have a sequence of consecutive points which are sampled within a small time interval. For example, a taxi trip is sampled for its GPS locations every several seconds. Sometimes, a trajectory may not include consecutive sample points but instead consists of unevenly distributed points. This often happens when the points are samples by Wifi or checkin/out devices. Moreover, some trajectories only have the start/end points, such as the O/D (origin/destination) data of taxi trips. Fig. 25 shows a trajectory e_i consisting of linked points $p_1, ..., p_L$.

5.3.2 Geographical Structure

Urban space consists of multiple types of geographical structures. In analytical tasks, most commonly used structures include streets, regions, and POIs (Points of Interest), as illustrated in Fig. 26. A region may be a grid cell, a zipcoded region, or a user-defined area. It has the name and boundarying polygon. A street, usually represented as a

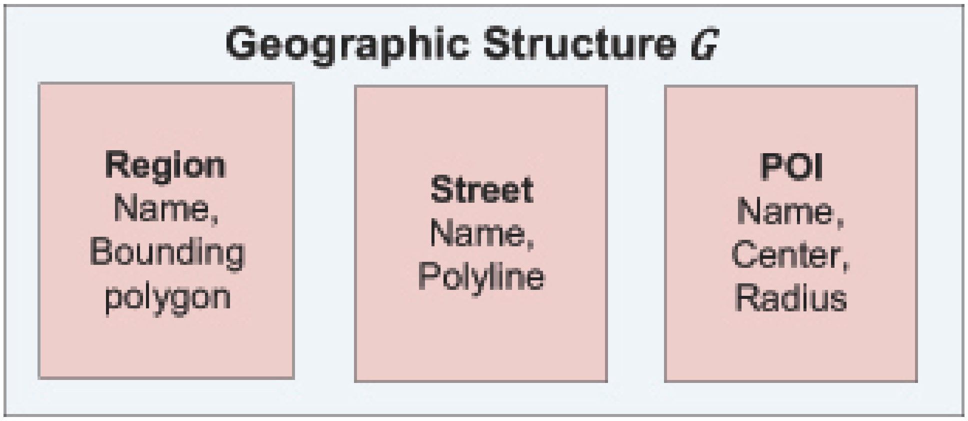


Figure 26: *Geographical Structure.*

geometric polyline, is part of urban street network. A POI is the common expression for a place which may have accurate geometric borders. But most cases, it is considered as a centroid with a center point and small radius. Demographics, business, social and other information can be linked to these geographic structures.

5.3.3 Data Registration and Matching

Urban trajectory data should be registered to the geo-structures. This is a fundamental task in GIS systems. Map-matching algorithms are developed to match trajectories to street networks. As a result, each GPS point in Fig. 25 is assigned to one street segment in Fig. 26. Moreover, they can also be matched to regions and POIs. These important operations connect geo-structures with urban trajectories, which are basic steps in TrajAnalytics for data preprocessing.

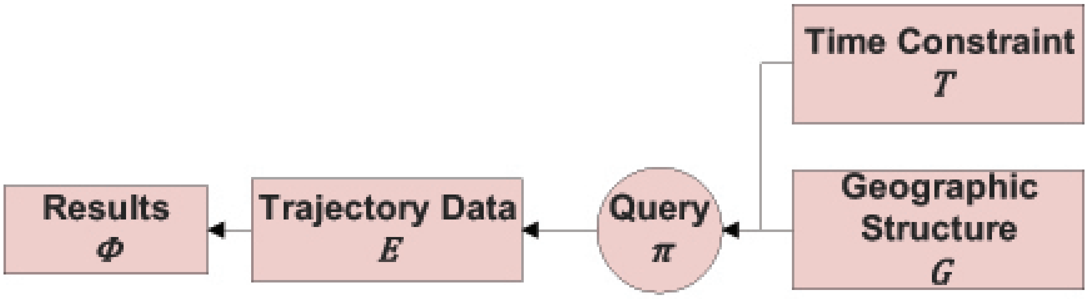


Figure 27: *Data Query Model*

5.3.4 Data Query and Joint Query

In visual analytics tasks, trajectory data need to be accessed by given geospatial and temporal constraints, which can be described as

$$E(G, T, \pi) \longrightarrow \Phi. \tag{4}$$

As depicted in Fig. 27, the query conditions involve the geographic structure G and the time constraint T applied with the query mode π to the trajectory data E. The result subset Φ are then sent to visual analytics module. Here, π usually can be of multiple modes such as:

- Pass mode: Find trajectories traversing G in T;
- Start mode: Find trajectories starting from G in T;
- End mode: Find trajectories ending G in T;
- Contain mode: Find trajectories whose path are contained inside G in T;

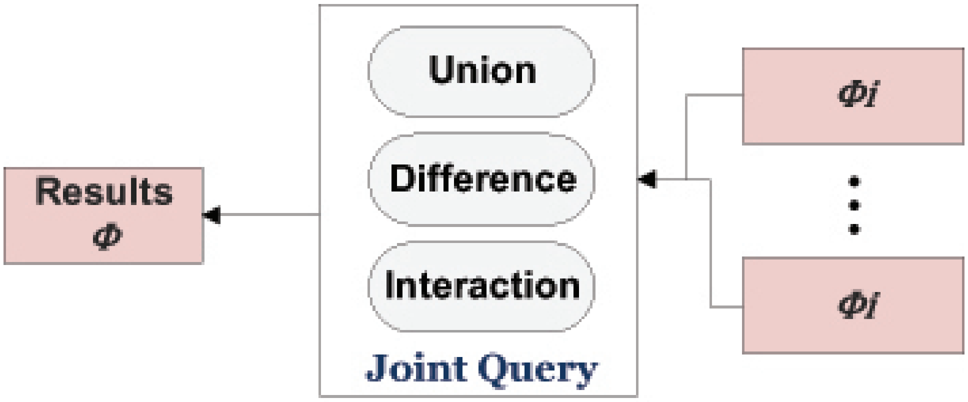


Figure 28: *Joint Data Query*

The modes can be of different conditions based on application requirements.

In most real-world tasks, data query tasks often have multiple tasks, where the individual queries π_i are combined with joint conditions. As illustrated in Fig. 28, Union, Difference and Intersection can be used to generate final query results.

5.4 Visual Analytics Tasks

The query results should be well utilized in a variety of data analytics tasks with visual tool. We summarize the tasks in several categories:

- **Trajectory Study:** Users query trajectories with given spatial-temporal conditions, and then (1) visualize them (in different ways) on map, which is integrated with (2) different types of charts, glyphs and diagrams for their temporal or spatial distributions. One special case of the trajectory study is the study of O/D data, which refers to original and destination of the trajectories. Similarly, O/D points are visualized on the map and combined with visual charts.

- **Street Study:** After querying trajectories, the results are used to acquire a variety of attributes on street segments. These attributes mostly are about traffic information such as traffic speed and traffic flow (amount of passing, starting, ending trajectories). Aggregation values such as average, maximum, minimum of the speed may be computed. The computed attributes may also be related to other types of information incorporated in the raw trajectory data.

- **Region Study:** Urban regions, such as zip-code areas, are also the target for urban study from trajectories. The query results thus can be mapped to given spatial regions to compute multiple attributes, such as the amount of trips starting/ending in a region, the average speed in given times, etc. They can be visualized by region view on the map and by different visual metaphors.

- **POI Study:** POIs are important so that the query result trajectories can be projected to POIs and visualized to show different attributes.

- **Flow Study:** In addition to region/street/POI attributes, the query results are often used to compute values related to pairwise relationships between them, such as the population flow from one region to another region, or from one type of POIs to another type of POIs. In some cases, the relationships may involve multiple geo-structures, such as starting from A, passing B and arriving C.

Time dimension is always an important factor in these studies. In visualization tools, interactive brushing and filtering over times is indispensable.

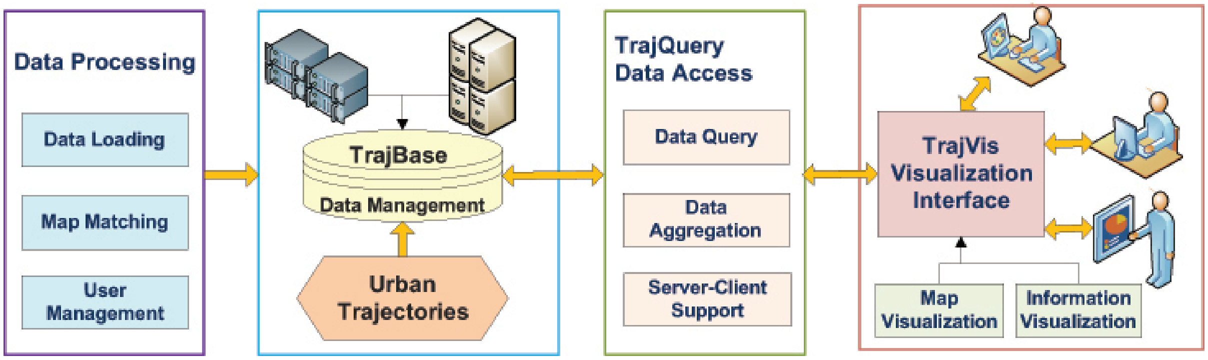


Figure 29: *TrajAnalytics software framework.*

TrajAnalytics is developed to provide an accessible software tool for practitioners of these tasks. It is a framework including fundamental functions, while facilitates users to add more complex visualization functions for specific applications.

5.5 TrajAnalytics Software Components

TrajAnalytics is designed based on the Geo+Trajectory model. It currently released as a client-server platform, where users can access the software through our server (cloud-version). Moreover, a local version is provides for software test and for private use. TrajAnalytics consists of four major components including Data Processing, TrajBase, TrajQuery, and TrajVis, as shown in Fig. 29.

1. **Data Processing**: This logistical module is demanded by domain users and their usage scenarios to define TrajAnaltyics to a practical software platform. First, it supports users to upload their raw trajectory datasets through a web browser. Second, the trajectories can be matched to geo-structures including streets, regions, and POIs. Here the software provides a convenient function to directly download

street geometry data from OpenStreetMap based on the uploaded trajectories. Moreover, the cloud-based system provides a fully functioned user management module for users all over the world to manage their data and analytical work.

2. **TrajBase**: A scalable database is specifically designed for storing and managing big trajectory data. This database is designed to accommodate consecutive trajectories and O/D trajectories. They are optimized to answer spatial and temporal queries through pre-computed aggregations. In implementation, PostgreSQL platform is used and its spatial extension PostGIS is utilized to accelerate spatial queries. TrajBase supports fast computation over various data queries in a remote and Web-based computing environment.

3. **TrajQuery**: TrajQuery supports users to conduct spatial queries combined with temporal constraints. It is defined to complete the four aforementioned query modes. Moreover, it supports joint queries based on Boolean operations. Data aggregation over geo-structures, such as computing speed and other attributes, is also completed by this module. A technical challenge is to facilitate fast query response and transfer of the query results over the internet connection. TrajQuery manipulates the data through a server-client data transfer channel.

4. **TrajVis**: To support online visual analysis with fast speed and easy user interaction, we design the system with a set of coordinated views. In implementation, we use the open source libraries of D3.js for information visualization and leaflet.js for

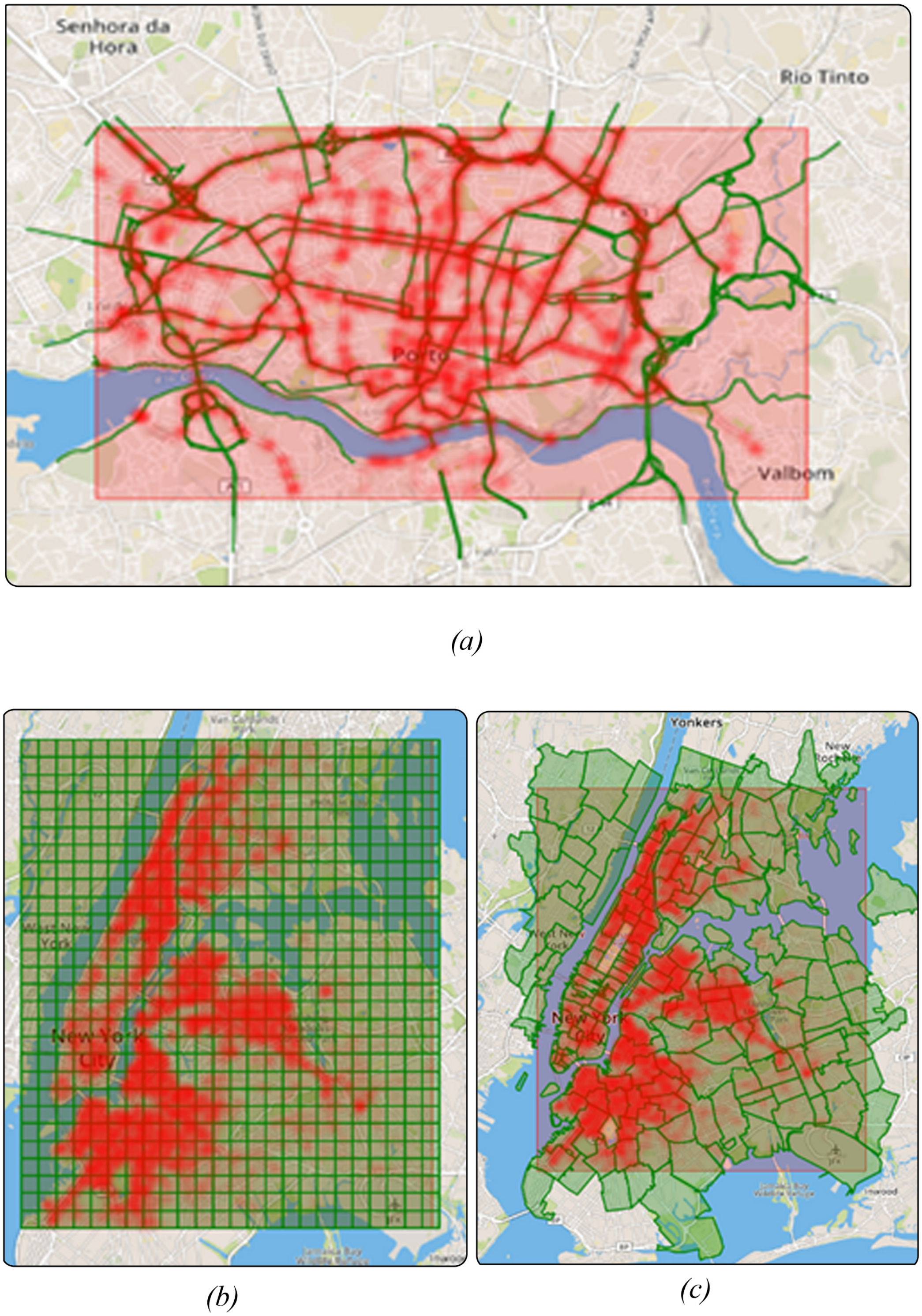
map-based visualization. Users can conduct exploratory visual analysis through the informative and intuitive interactions. In particular, users can choose different types of maps for geographic contexts. They can conduct region selections and time period selections through easy mouse operations. The queried results are visualized by a set of visual metaphors accessible from any Web browse.

In the following sections, we describe the specific designs and algorithms in these modules.

5.6 Integrating Trajectory Data with Geo-Structures

5.6.1 Raw Data Processing and Uploading

TrajAnalytics supports users to upload their raw trajectory datasets in the comma-separated values (CSV) format. Each data item refers to one sampling point. The values should include Geo-locations (longitude, latitude), trip ID, timestamp, and speed (optional). Here trip ID indicates which trip this point belongs to. Users can upload the raw CSV data through our data uploading interface. It will convert the raw data to a Table Data (TD) table, which is imported into our TrajBase database. TD table stores the trip data, instead of the raw sampling points. These points are aggregated by Trajectory ID into trips. TrajBase creates three types of spatial indexes (B-Tree, GIST, and GIN) to enhance the query speed of data. A TD tables stores trips as groups of spatiotemporal points. However, they are not managed in a appropriate geographical context. TrajAnalytics further provides two different ways to match the trip data with geographical units, streets and regions, respectively.



(a)

(b)

(c)

Figure 30: TrajBase Map Matching Modules. (a) Street Map Matching. (b) Grid Cells Regions Map Matching. (c) Zipcode Regions Map Matching.

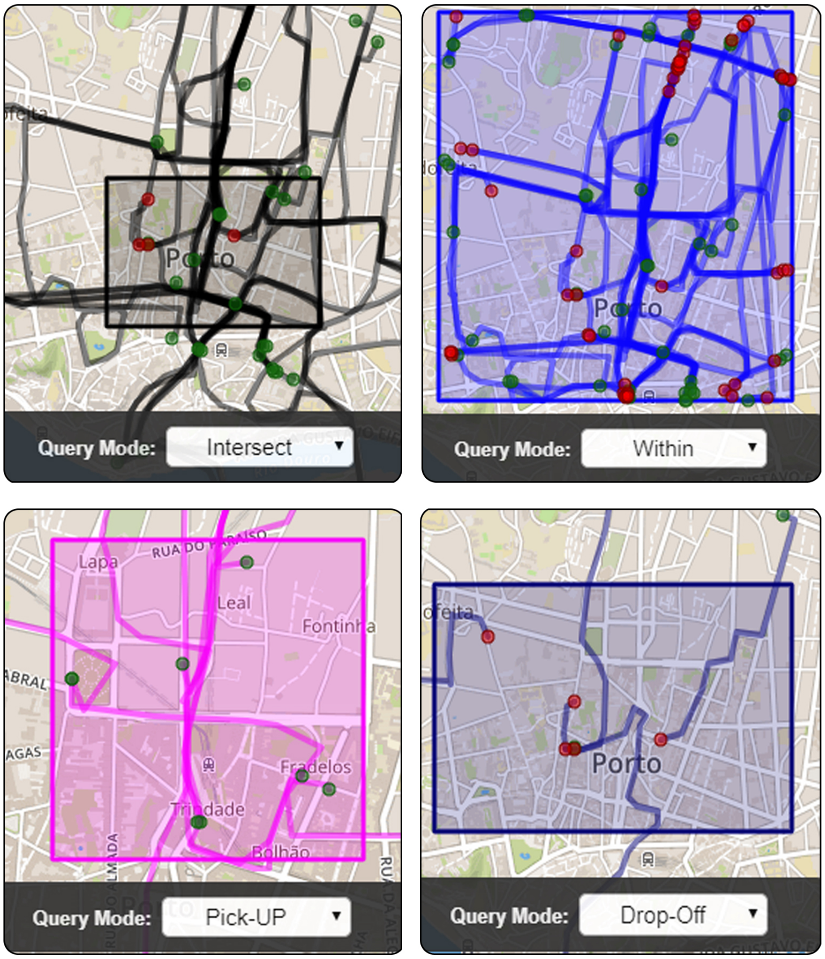
5.6.2 Map Matching

5.6.2.1 Street Map Matching Module

TrajAnalytics provides a module that enables users to match their dataset with a street network of the corresponding geographical area. The street network is automatically downloaded from OpenStreetMap, if available. A table data mapped to the street network table (TDS) is created in TrajBase which has a list of street segments each including a unique ID and the geometry (polylines) of the segment. TDS tables stores trips similarly to a TD table, while in addition, a street ID links each point to a street segment. Figure 30(a) shows the sampling points in red and the downloaded street network in green in preparation for the street map matching process.

5.6.2.2 Region Map Matching Module

When a street network is not available or not to be used, TrajAnalytics supports map-matching of trips into spatial regions in the corresponding area. Currently, two types of regions are supported: Zipcode regions (the US only) and grid cells regions. We found Zipcode regions are available freely in the US only. So, we allow users to divide the area into a grid with arbitrary resolution, then match trips into the grid cells. In TrajBase, a table data mapped to regions (Zipcode or grid cells regions) table (TDR) is created to store the list of cells including a unique ID and the geometric boundaries of the regions/cells. TDR tables store trips similarly to a TD table, while in addition, a region ID links each point to a Geo region. Figure 30(b) shows the sampling points in red and the spatial grid cells for the corresponding area in green, and Figure 30(c) shows the sampling points in red and the zipcode regions for the relevant area in green in preparation for the region map matching process.


Query Mode: Intersect

Query Mode: Within

Query Mode: Pick-UP

Query Mode: Drop-Off

Figure 31: Spatial Queries provided by TrajQuery.

5.7 TrajQuery

Answering data queries fast is vital for building visual encodings in an interactive system. TrajQuery supports the user to conduct spatial queries combined with temporal constraints. Regardless of the table type (TD, TDS, TDR), the spatial queries extract trips or trajectories with (1) pick-up (Start) regions; (2) drop-off (End) regions; (3) traversed (Pass) regions; and (4) within (Contain) a region. Figure 31 shows the spatial queries provided by TrajQuery. The green and red dots refer to the trips pickup and drop-off location respectively. A map-based interface is provided to define the regions, including rectangular, circular and free form selections. The four types can be conducted individually or jointly. They are also integrated with time period selections. Figure 32 shows different forms of defining query region in Porto city downtown in Portugal to find all taxi trips passed the specified region in July 1, 2013.

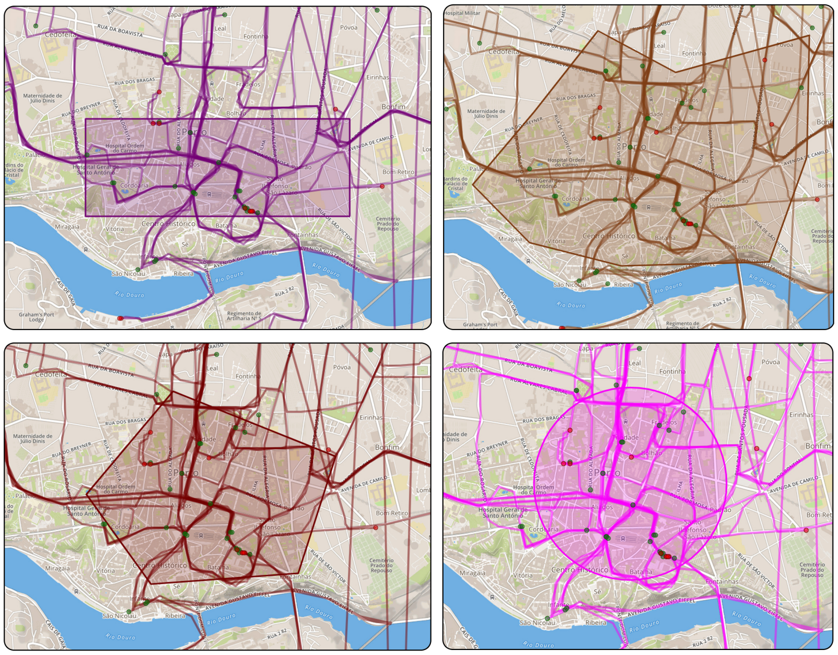


Figure 32: Different forms of defining query region.

In general, TrajQuery supports two types of data queries: (1) Coordinate-Based Queries; (2) Trajectory-Based Queries.

Coordinate-Based Queries: The queries retrieve the statistical information on given coordinates. The given coordinates could be a street segment, a grid cell, or a zipcode region. Retrieved information includes average, sum, minimum/maximum, and count. The results are major visualization objects.

Trajectory-Based Queries: Trajectory queries are different from the above queries looking for the locations and features of moving objects at a time. They search for a set of trajectories over a given time period. The query is transferred to TrajBase and all trips in the corresponding table is retrieved, processed and returned to TrajVis. With respect to different types of data tables, a set of attributes are computed and returned as follow:

- Trajectory or trip attributes (average speed, maximum speed, and minimum speed of each trajectory or trip).
- Trajectory or trip start (origin) and end (destination) locations.
- Street attributes (for TDS data) (average speed, maximum speed, minimum speed, and the number of trajectories or trips (i.e., flow) passing each street segment).
- Region attributes (for TDR data) (average speed, maximum speed, minimum speed, and the number of trajectories or trips (i.e., flow) passing each region).
- Time-varying trajectory or trip attributes including: (1) Weekly averages of the attributes such as the average speed of trips on Mondays, Tuesdays, etc; (2) Hourly averages of the attributes such as the average speed of trips in 8am, 10am, etc; (3)

Daily averages of the attributes such as the average speed of trips in each day in the specified time period.

- Time-varying street attributes (for TDS data) including: (1) Weekly averages of the attributes such as the average speed of each street segment in Mondays, Tuesdays, etc; (2) Hourly averages of the attributes such as the average speed of each street segment in 8am, 10am, etc; (3) Daily averages of the attributes such as the average speed of each street segment in each day in the specified time period.
- Time-varying region attributes (for TDR data) including: (1) Weekly averages of the attributes such as the average speed of each region in Mondays, Tuesdays, etc. (2) Hourly averages of the attributes such as the average speed of each region in 8am, 10am, etc. (3) Daily averages of the attributes such as the average speed of each region in each day in the specified time period.

5.8 TrajVis Visualization

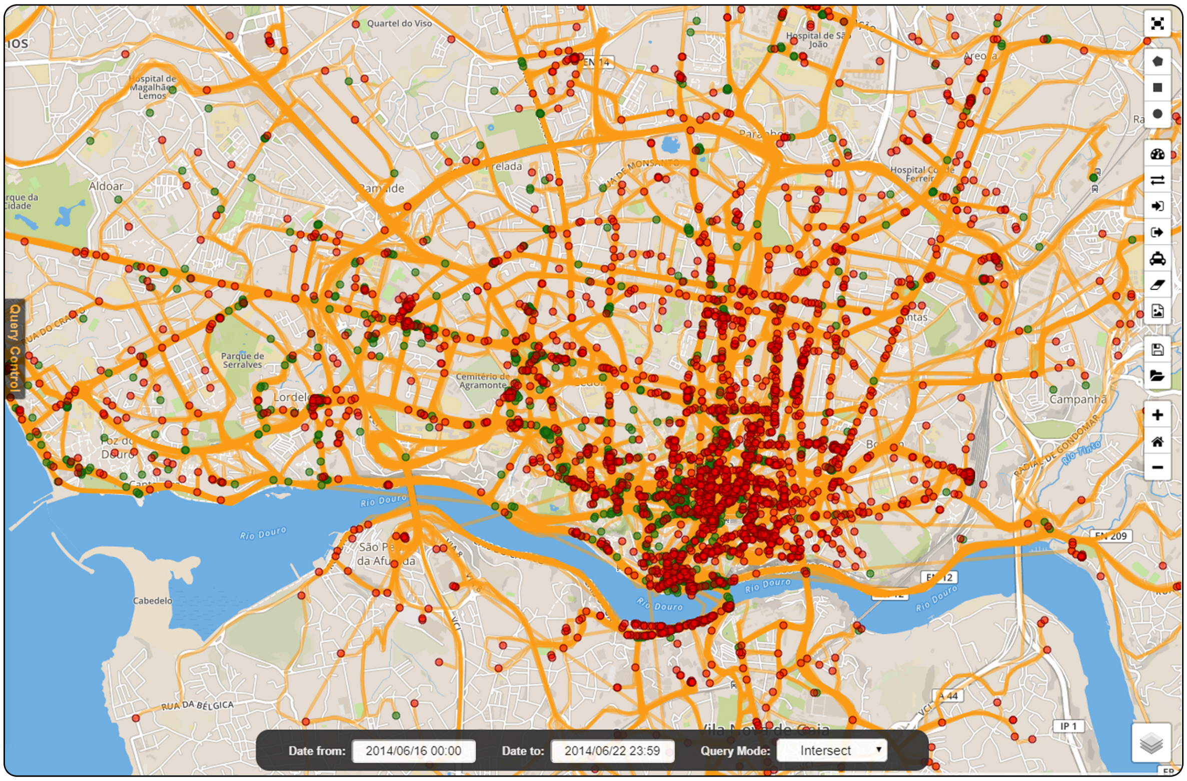
To support online visual analysis with fast speed and easy user interaction, we have designed a system with a set of coordinated views. The interface is illustrated in Figure 4 with the interactive map view (Figure 4 (b)), manageable list of multiple queries (Figure 4 (a)), and interactive lists, charts, and diagrams for data analytics (Figure 4 (c)). Users can conduct exploratory visual analysis through the informative and intuitive interactions. They can conduct region selections and time period selections through easy mouse operations. The queried results are shown in points, heat maps, or trajectories while users can alter the display methods. To make analysis easier, the list view of queries help users manage multiple queries. The visual report view (Figure 4 (c)) shows a set of charts and diagrams of the query results for quantitative attribute analysis. The online visual system also support multiple users to conduct their work from different sites.

5.8.1 Interactive Map View

TrajVis is designed as a map-centric application where the map plays the role as a canvas in the middle for users to define the query regions and show the results. Different map layers are provided that users can toggle between them. Figure 4 shows the visual interface which includes the interactive map view (Figure 4 (b)) which has the following major views:

Trajectory or Trips Visualization: The trajectories or trips are directly shown as connected polylines on the map. Users can select speed, flow, or other attributes to be visualized on each trajectory or trip, which is represented by the line width and the color. The start (pickup) and end (drop-off) locations are visualized as red and green points, respectively. Figure 33 shows all taxi trips that passed a specified region in Porto city downtown in Portugal between June 16, 2014, and June 22, 2014.

Heatmap Visualization: The start (Pickup) and end (Drop-Off) points are aggregated and visualized as a heatmap on the map. Their densities are shown by the colors. Figure 34 shows the heatmap of start (Pickup) and end (Drop-Off) of all trips that passed a specified region in Porto city downtown in Portugal between June 16, 2014, and June 22, 2014.



Date from: 2014/06/16 00:00

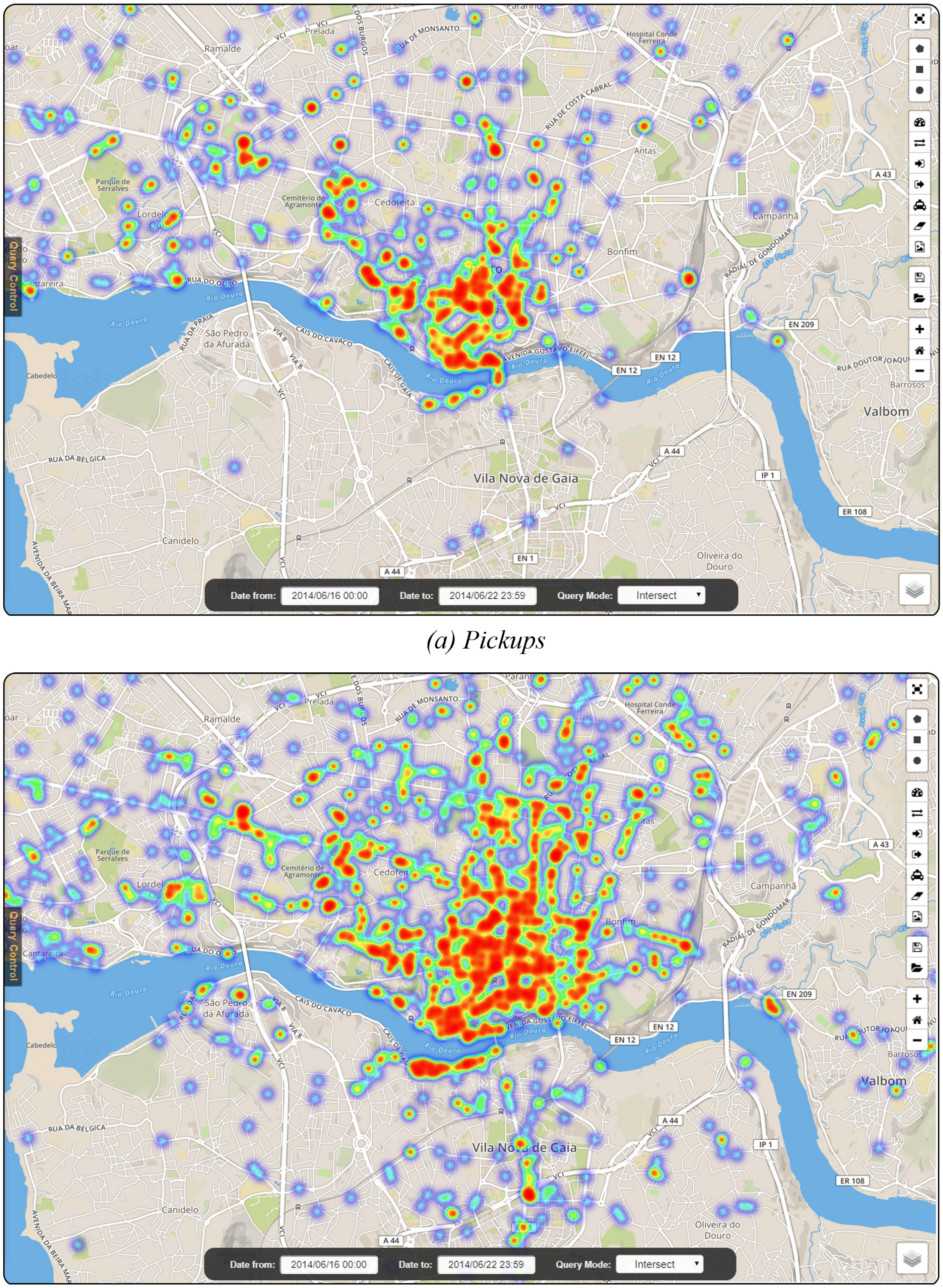
Date to: 2014/06/22 23:59

Query Mode: Intersect

Figure 33: Trajectory or Trips Visualization.

Street Attributes Visualization (for TDS data): The trajectories or trips pass a group of roads. A set of attributes are computed on these roads by aggregating the information from these trajectories or trips. Users can select an attribute such as speed, flow, or any other attributes to be visualized on the street segments, which is represented by the line width or color of the streets. Figure 35 shows the traffic speed (a) and the traffic flow (b) over the road network of Porto city downtown in Portugal based on the taxi trips that passed a specified region between June 16, 2014, and June 22, 2014.

Regions Attributes Visualization (for TDR data): Similarly, a set of attributes are computed on regions (zipcode or grid cells regions) by aggregating the information from trajectories or trips from a query. Users can visualize an attribute by the color of



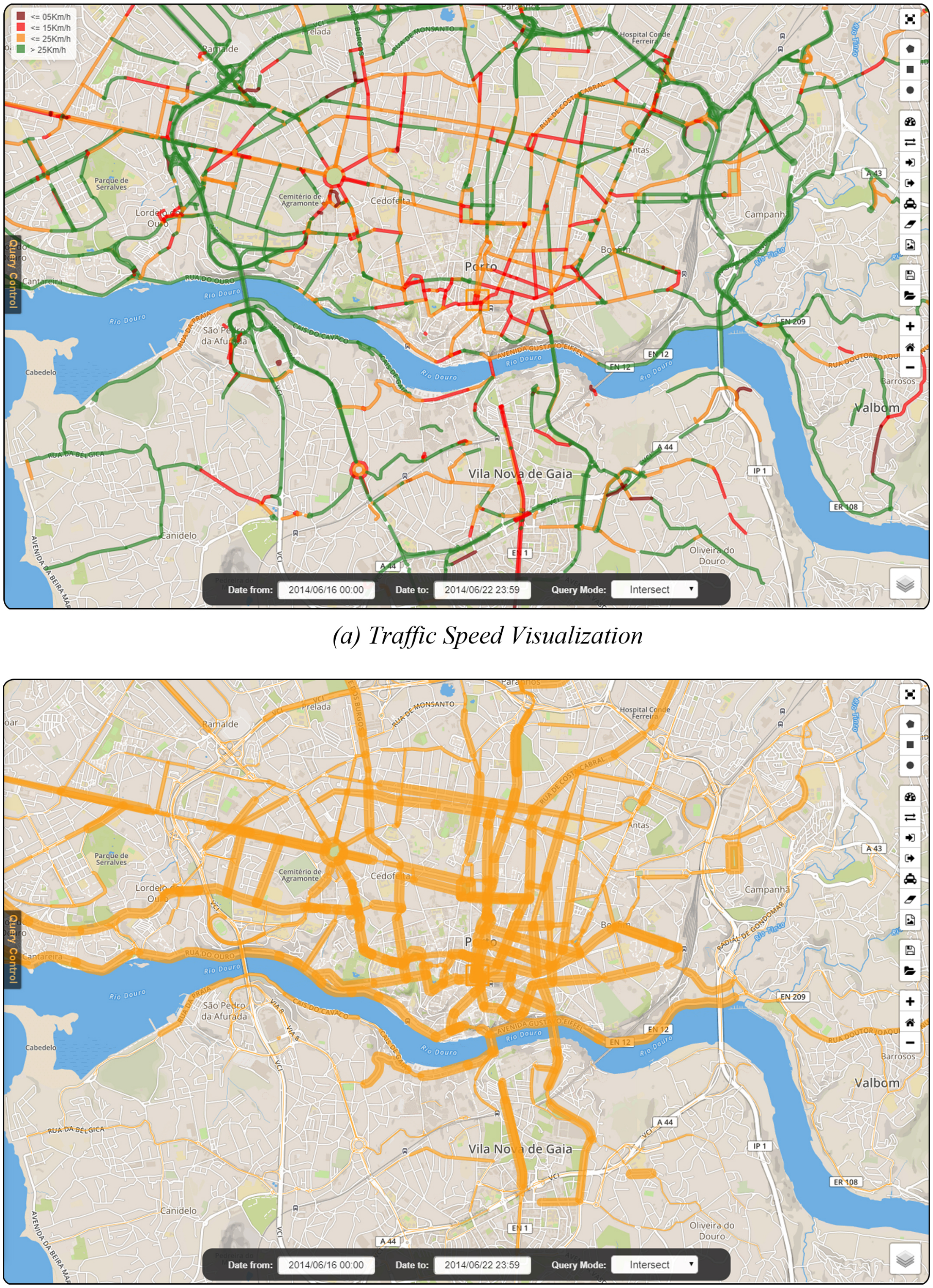
Date from: 2014/06/16 00:00 Date to: 2014/06/22 23:59 Query Mode: Intersect

(a) Pickups

Date from: 2014/06/16 00:00 Date to: 2014/06/22 23:59 Query Mode: Intersect

(b) Drop-Offs

Figure 34: Trajectory or Trips Start (Pickup) and End (Drop-Off) Locations Heatmap Visualization.



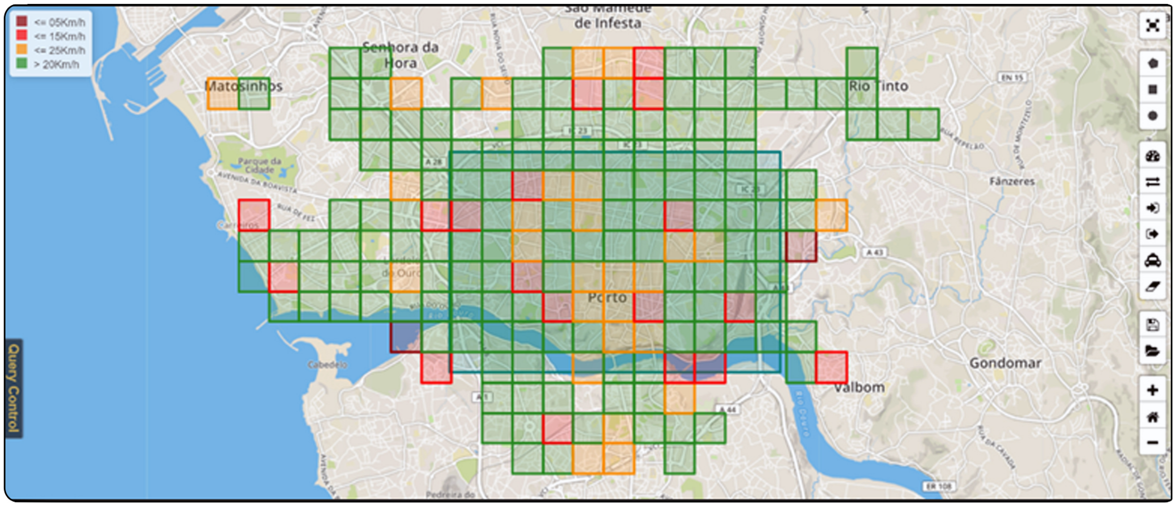
Date from: 2014/06/16 00:00 Date to: 2014/06/22 23:59 Query Mode: Intersect

(a) Traffic Speed Visualization

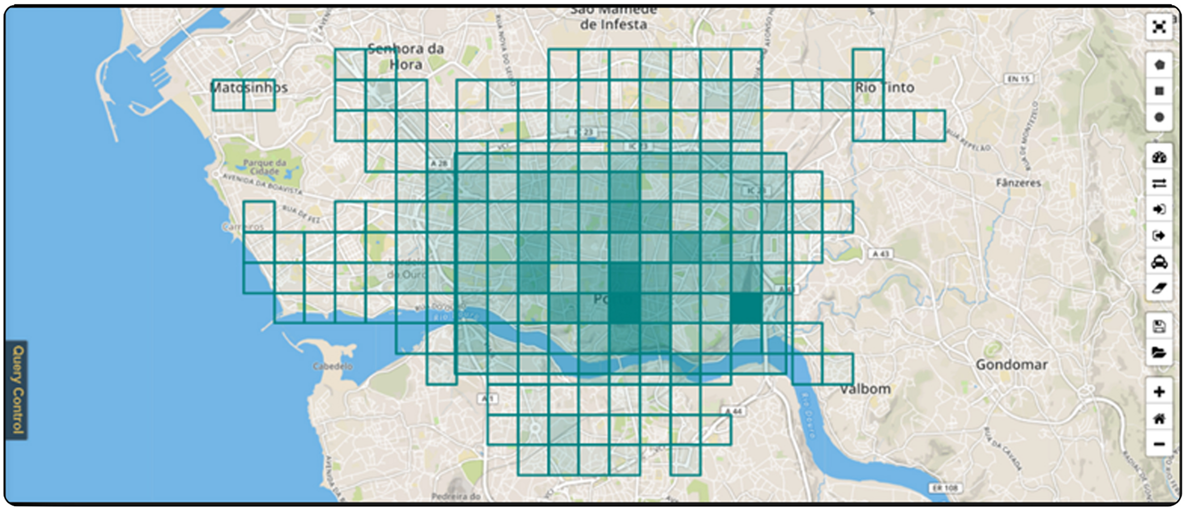
Date from: 2014/06/16 00:00 Date to: 2014/06/22 23:59 Query Mode: Intersect

(b) Traffic Flow Visualization

Figure 35: Visualize Attributes on Streets (for TDS data).



Traffic Speed Visualization



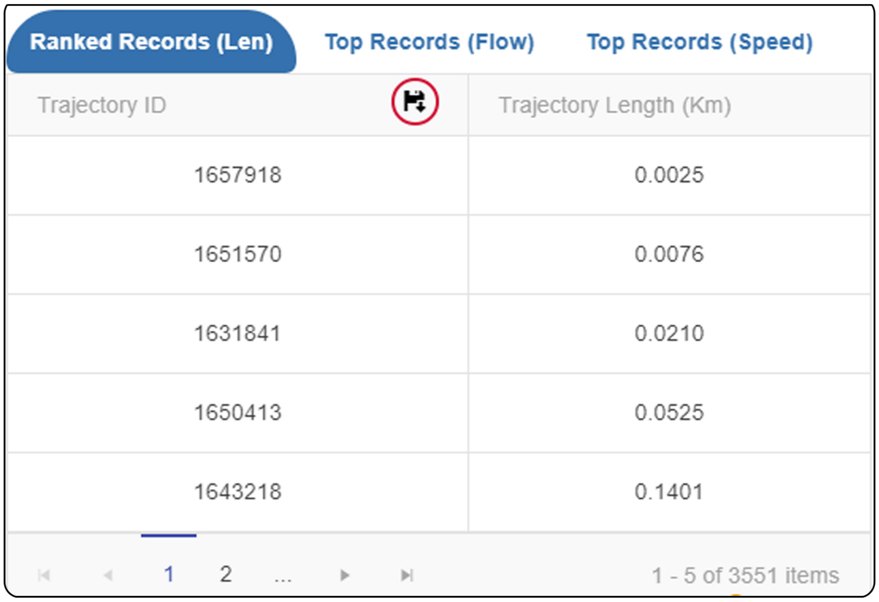
Traffic Flow Visualization

Figure 36: Visualize Attributes on Grid Cells Regions (for TDR data).

these regions on the map. Figure 36 shows the traffic speed (a) and the traffic flow (b) over the grid cells regions of Porto city downtown in Portugal based on the taxi trips that passed a specified region in July 1, 2013.

5.8.2 List View

List view as shown in Figure 37 enables users to study the attributes extracted from both trajectory and coordinate-based queries. Trajectories or trips are visualized in a maneuverable list with their attributes as columns. In case of TDS or TDR, the streets or regions are visualized in a maneuverable list with their attributes as columns. Users can click on a column title to change the ranking order (descending or ascending) by the corresponding attribute. Users can also hover over a trajectory or trip, streets, or region to highlight it on the map.



Trajectory ID	Trajectory Length (Km)
1657918	0.0025
1651570	0.0076
1631841	0.0210
1650413	0.0525
1643218	0.1401

Figure 37: List View.

5.8.3 Chart View

Chart view is shown in Figure 38. Trajectories or trips are distributed in the given time period. Two types of interactive charts are used to visualize the facts related to the time windows. In the daily chart, the number of trajectories or trips are visualized in the bar charts along different days of weeks. In the hourly chart, the number of trajectories or trips are visualized in the bar charts along different hours of days. Users can click on a bar to filter and investigate the trajectories or trips it represents in other views.

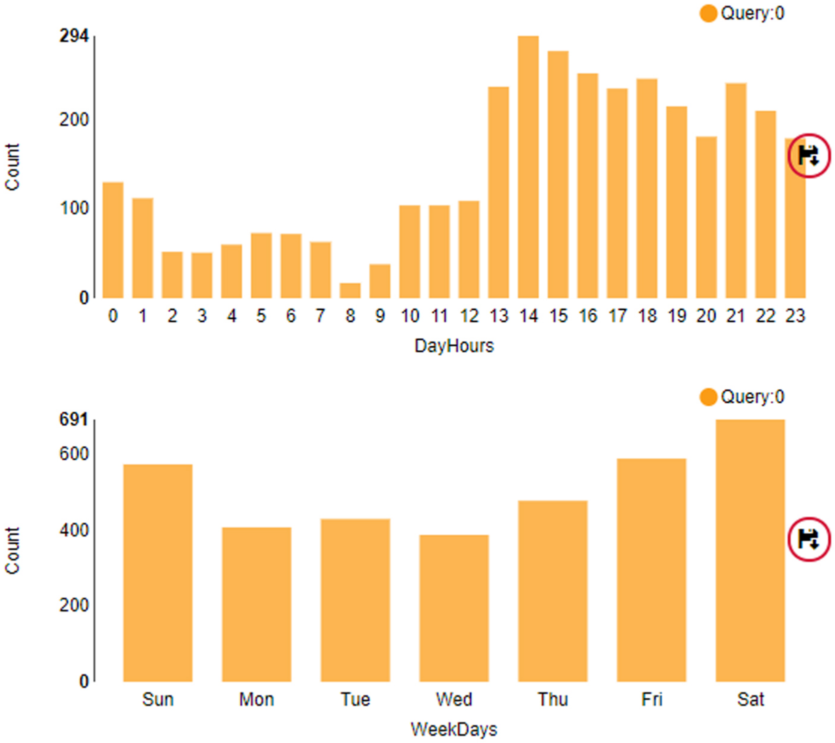


Figure 38: Chart View.

5.9 TrajAnalytics Software Implementation

TrajAnalytics is developed as a cloud-based software which can be accessed through web browsers. It supports users to upload raw trajectory data, and perform road based or region based map matching. The generated data tables are stored in TrajBase database on cloud server for remotely realtime access. TrajVis interface is implemented using multiple Javascript libraries such as D3.js for information visualization and leaflet.js for map-based visualization. TrajAnalytics is also provided as a local version for users to work with the software on their own machines after configuring their own server. The software is published online for free access at `http://vis.cs.kent.edu/TrajAnalytics/` with full tutorials and user guides.

5.10 Evaluation and User Study

We conducted a user study of TrajAnalytics with a group of 35 domain experts and practitioners from different countries. Their ages range from 21 to 63 years old and most have Ph.D. degrees. Their working experience ranges from 1 to 30 years in a variety of fields related to human and vehicle trajectory study, as shown in Fig. 40. They were asked to use TrajAnalytics with help of our online tutorials and videos. Then, they filled a questionnaire. The questions included: if TrajAnalytics is generally useful in your field? and how different modules are useful respectively? They gave a score (0-10) for the system and modules. Fig. 39 shows the average scores from 35 domain users. Overall, They gave a positive evaluation with all scores above 7.8 out of 10. The TrajVis interface got the highest usability score at 8.9.

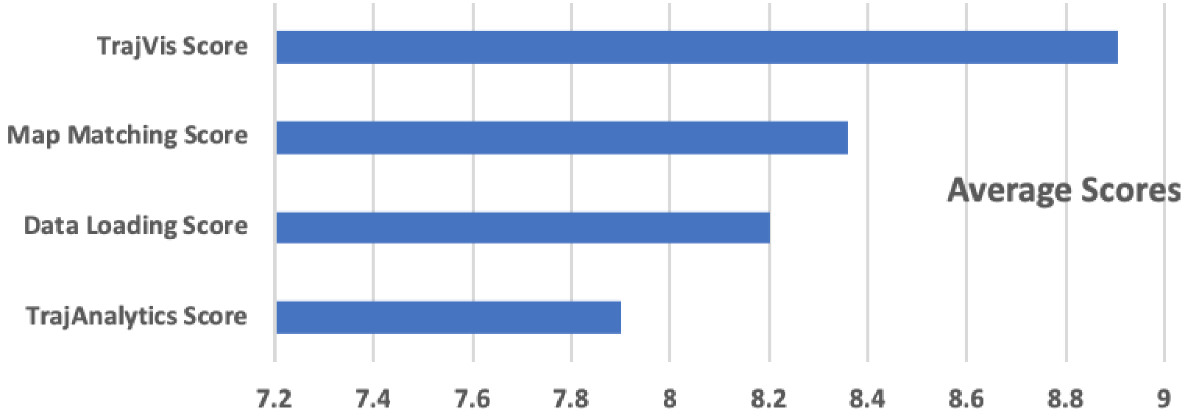


Figure 39: *Average scores from domain users about TrajAnalytics.*

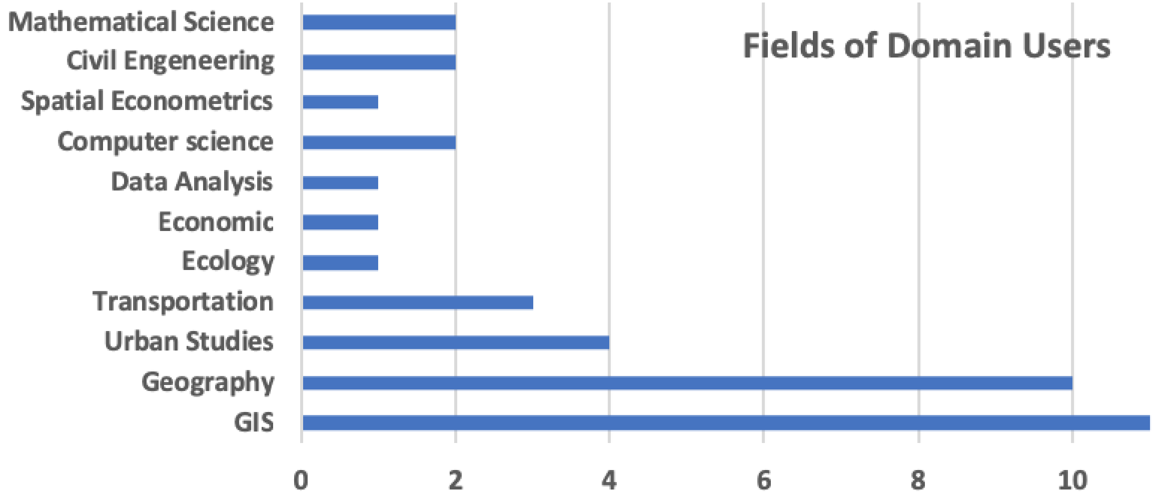


Figure 40: *Domain users from different fields.*

In addition, we asked the users to provide comments and suggestions of the system. They mostly agreed that TrajAnalytics is a good tool for domain study, such as ".... is easy to learn and use and it needs less memory"; "... is useful for analyzing traffic jam, some geographic phenomena, like migration, trading, and behavior"; "... provides data support for studying the smoothness of the road and evaluating the rationality of road planning.”. Most of the researchers agreed that the data loading and processing modules including the roads and regions map matching are effective and very useful tools, since “...there is only few software providing the function of map matching”.

Moreover, both cloud version and local version are evaluated. Almost half of domain users preferred the cloud version as it does not require installation and configuration, while the others liked the local version for its stability and because sometimes the data needs to be kept confidential.

The users also pointed out some limitations and gave valuable suggestions to improve the system such as:

- Enriching the visualization and analysis functions by adding some extended features such as data analysis, spatial analysis, and machine learning methods.
- Supporting comparative study of trajectory data from different countries and regions.
- Providing a simpler way to install and configure the local version.
- Migration to a distributed platform to accommodate larger datasets.
- Providing extended programming support.
- Creating a complete guide book.

We will enhance the software based on these suggestions.

5.11 Chapter Summary

The mobility and behavior of moving humans and transportation vehicles form the basic component in human society. Iterative visual exploration is one key component in processing urban trajectory data, which should be supported by efficient data management and visualization tools. In this chapter, we present TrajAnalytics, a free software for visually exploring urban trajectories. Our software facilitates easy access gateway so that the urban trajectory data can be retrieved, analyzed and visualized by different researchers, and their results can be shared and leveraged by others. TrajAnalytics support users with a scalable data storage and management which support a variety of data queries with immediate responses. It also provides exploratory visualizations that are informative, intuitive, and facilitate efficient interactions. It allows simultaneous operations by many users from different places. It will advance a broad spectrum of applications by enabling researchers, administrations, practitioners and general public to visually analyze the emerging trajectory data.

CHAPTER 6

Conclusion and Future Work

Thanks to advanced technologies in sensing and computing, the mobility patterns and dynamics of urban cities and their citizen are recorded and manifested in a variety of urban trajectory datasets, which include the moving paths of the human, taxi, bus, fleets, cars, and so on. The availability of such data has fostered a diversity of applications, calling for algorithms that can manage, explore and extract knowledge from the data in an efficient way. The obtained information can be utilized in many relevant and practical applications to optimize urban planning, improve human life quality and environment, and amend city operations. These applications include:

- Urban planning (Identifying behaviors, citywide human mobility for detecting urban anomalies, the city's functional regions and urban planning).
- Transportation (Traffic monitoring and prediction, Urban planning, routing, and energy consumption analysis, POI, land use).
- Environment (Information about urban air quality and distribution of air and noise

pollution).

- Energy (City's energy consumption: obtained directly from sensors or inferred from data sources implicitly, e.g., from the GPS trajectory of a vehicle).
- Social (Social structure: a graph denoting relationship, interdependence, or interaction between users. User-generated social media, texts, photos, and videos, which contain user's behavior/interests).
- Economy (City's economic dynamics: transaction records of credit cards, stock prices, housing prices, and people's incomes).
- Public safety and security (Loop sensors, surveillance cameras, and floating cars, floating car data).

Abundant data that is collected for analysis, whether acquired from GPS records, WiFi/RFID check-ins, or extracted from geo-tagged messages, is mostly in the form of a raw list of data points and their corresponding attributes, but not ready to use for visual analytics that requires integrated trajectories and geo-context. This requires domain analysts to implement their own model of trajectory data transformation and storage from the raw data, before even utilize a visualization toolkit, which unfortunately has not been easily obtained in any free, open-source system. Despite myriad visualization systems for *ad hoc* urban trajectory data analytics tasks, there remains a gap between the demand of domain researchers and a general visualization software integrating raw trajectory data processing, storage, and interactive visualization. In this thesis, we have presented a set of visual analytics solutions with different approaches to interacting with massive taxi trajectory data to allow analysts to look at the data from different perspectives and complete different analytical tasks. Our approaches focus on how people directly interact with the data store, query and visualize the results and support practitioners, researchers, and decision-makers to advance transportation and urban studies in the new era of the smart city.

6.1 Detailed Conclusion

Firstly, in SemanticTraj (Chapter 3) we have proposed to use the text search engine in a visual query and analysis of taxi trajectory data. It utilizes textualization and taxi documents to develop a new way to study the emerging urban data. Users can perform keyword or name queries in different ways for a set of useful analytical tasks. The text search engine provides efficient query support with optimized data management. It also has excellent capabilities to conduct various queries. It can be complementary to the existing region and range queries. We also develop a prototype visual query system with a set of visualization tools for users to conduct interactive trajectory study.

Secondly, Querying and studying of urban traffic patterns from a transport database is supported by a sketch based visual analytics system named as QuteVis (Chapter 4). Unlike most existing approaches that investigate traffic data in specified spatial regions and temporal periods, QuteVis supports a different type of data query and analytical tasks. It helps users discover those specific times and days in history that have similar traffic patterns as they speculate on multiple, spatially-diverse city locations. Users can specify arbitrary spatial traffic conditions such as speed, volume, or taxi pickups/dropoffs, on streets, regions, or paths, where touch input devices can be used for fast sketch-based interaction on a city map. Weighted similarities between user speculated traffic pattern and actual traffic situations in historical times are computed. The weights are defined with respect to the hierarchical levels of different streets or users' preference. Then, a set of visualization and interaction functions further help users overview and compare the traffic situations on different time periods of different days. They can discover potential influential factors (e.g., weekdays, weather) and test hypotheses about the relationships of traffic behaviors at different locations in a city from massive historical data. The system thus can help researchers and analysts assess urban transport features related to major problems in modern cities, such as traffic jams, unbalanced traffic flows, and unsatisfied travel demands.

Thirdly, we have introduced TrajAnalytics system (Chapter 5) which has been developed to help different users including researchers, administrations, practitioners and the general public to understand their data and to reveal knowledge intuitively. These domain users are from all over the world in a variety of fields of urban planning, geography, transportation, and social studies. TrajAnalytics is an open source visual analytics software, which integrates scalable data modeling, transformation, management, and interactive visualization within a powerful web-based computing platform. It contains three major modules: (1) Data loading and transformation; (2) TrajBase database; (3) TrajVis visual analytics interface. The system is developed in two forms: a cloud-based software and a local version, to fulfill the requirements of many real world users. It has been released for open access. Many domain experts have participated in the development in multiple phases. In particular, we have conducted a comprehensive user evaluation which focuses on the usability of the system. To support widespread test, we have developed a Web-based system to collect feedbacks. We report our process of software evaluation and share our gains.

Finally, to support all the above approaches, we have developed a set of visual exploration tools to be utilized in these systems. In SemanticTraj (Chapter 3), we have focused on developing tools that reveal the semantic information of the taxi trajectory data such as the meta-summary. In QuteVis (Chapter 4), we have developed multi-sketch interaction on a city map to help users to conveniently specify multiple conditions at separate locations on the map. A reliable sketch recognition algorithm is developed so that freestyle drawing is automatically matched to street geometry. The system further allows users to edit sketches with easy interactions. Furthermore, the coordinated multi-map view helped the user to compare the traffic information of different time window. In TrajAnalytics (Chapter 5), different map layers are provided to show different attributes such as speed, flow, pickup, and drop-offs. These visual exploration tools are considered as an effective approach to providing material for human's perception and plays a vital role in analyzing trajectory data.

6.2 Future Work

These approaches can be improved in several aspects:

- In SemanticTraj (Chapter 3), Semantics may imply high-level, summarized information describing the pattern and knowledge hidden in languages. We will further investigate in this direction where text mining tools can be used to analyze the taxi documents for more profound insights.

- The QuteVis (Chapter 4) system has its limitations to be improved. First, the correlation and the cause-effect relationship between different locations cannot be directly discovered over the interface. We will introduce data mining tools to automatically detected such information and provide profound interaction with users. Second, statistical functions may be included for users to analyze the query results, such as statistical significance testing and A/B testing; Third, other important types of urban data can be integrated with the traffic data, such as human mobility and social media information, for an enhanced intelligent system.

- TrajAnalytics (Chapter 5) system can be enhanced by performing more extensive user studies, providing more functions as a cloud service, employing more interaction techniques, and support more dataset format such as migration dataset.

BIBLIOGRAPHY

[1] N. Andrienko and G. Andrienko, "Visual analytics of movement: An overview of methods, tools and procedures," *Information Visualization*, vol. 12, no. 1, pp. 3–24, 2013.

[2] Y. Schulz, "Achieving value from visual analytics," 2016. [Online]. Available: https://www.corvelle.com/achieving-value-from-visual-analytics/

[3] D. A. Keim, G. L. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, "Visual analytics: Definition, process, and challenges," in *Information Visualization*, 2008.

[4] X. Ye, "Big flow data visual analytics through trajanalytics," 2018. [Online]. Available: https://cga-download.hmdc.harvard.edu/publish_web/CGA_Conferences/2018_DataScience/slides/CGA_2018_THU_XinyueYe.pdf

[5] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: Driving directions based on taxi trajectories," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '10. New York, NY, USA: ACM, 2010, pp. 99–108. [Online]. Available: http://doi.acm.org/10.1145/1869790.1869807

[6] M. F. G. Jingxiong Zhang, *Uncertainty in Geographical Information.* CRC Press, 2002.

[7] Z. Yan, C. Parent, S. Spaccapietra, and D. Chakraborty, "A hybrid model and computing platform for spatio-semantic trajectories," in *The Semantic Web: Research and Applications*, L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 60–75.

[8] B. Bamba, L. Liu, P. Pesti, and T. Wang, "Supporting anonymous location queries in mobile environments with privacygrid," in *Proceedings of the 17th International Conference on World Wide Web*, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 237–246. [Online]. Available: http://doi.acm.org/10.1145/1367497.1367531

[9] J. Jun, R. Guensler, J. Ogle, and A. Professor, “Smoothing methods designed to minimize the impact of gps random error on travel distance, speed, and acceleration profile estimates,” 12 2018.

[10] G. Marketos, E. Frentzos, I. Ntoutsi, N. Pelekis, A. Raffaetà, and Y. Theodoridis, “Building real-world trajectory warehouses,” in *Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, ser. MobiDE ’08. New York, NY, USA: ACM, 2008, pp. 8–15. [Online]. Available: http://doi.acm.org/10.1145/1626536.1626539

[11] M. Lashley, “Kalman filter based tracking algorithms for software gps receivers.” *Master’s thesis, Auburn University, the Netherlands*, 2006.

[12] N. Schuessler and K. Axhausen, “Processing gps raw data without additional information,” 01 2008.

[13] M. Horemuž and J. V. Andersson, “Polynomial interpolation of gps satellite coordinates,” *GPS Solutions*, vol. 10, no. 1, pp. 67–72, Feb 2006. [Online]. Available: https://doi.org/10.1007/s10291-005-0018-0

[14] R. H. Güting, T. de Almeida, and Z. Ding, “Modeling and querying moving objects in networks,” *The VLDB Journal*, vol. 15, no. 2, pp. 165–190, Jun. 2006. [Online]. Available: http://dx.doi.org/10.1007/s00778-005-0152-x

[15] I. Abdulganiev and A. Agafonov, “Automatic checking of road network models,” 01 2016, pp. 249–255.

[16] M. A. Quddus, W. Y. Ochieng, and R. B. Noland, “Current map-matching algorithms for transport applications: State-of-the art and future research directions,” 2007.

[17] S. Brakatsoulas, D. Pfoser, R. Salas, and C. Wenk, “On map-matching vehicle tracking data,” in *In Proc. 31st VLDB Conference*, 2005, pp. 853–864.

[18] D. Bernstein and A. L. Kornhauser, “An introduction to map matching for personal navigation assistants.”

[19] Y. Meng, “Improved positioning of land vehicle in its using digital map and other accessory information,” *Hong Kong Polytechnic University Ph.D. Dissertation*, 2006.

[20] R. Krüger, G. Simeonov, F. Beck, and T. Ertl, “Visual interactive map matching,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 6, pp. 1881–1892, June 2018.

[21] N. Pant, M. Fouladgar, R. Elmasri, and K. Jitkajornwanich, “A survey of spatio-temporal database research,” in *Intelligent Information and Database Systems*, N. T. Nguyen, D. H. Hoang, T.-P. Hong, H. Pham, and B. Trawiński, Eds. Cham: Springer International Publishing, 2018, pp. 115–126.

[22] G. Trajcevski, O. Wolfson, H. Cao, H. Lin, F. Zhang, and N. Rishe, "Managing uncertain trajectories of moving objects with domino," in *ICEIS*, 2002.

[23] M. F. Mokbel, X. Xiong, W. G. Aref, S. E. Hambrusch, S. Prabhakar, and M. A. Hammad, "Place: A query processor for handling real-time spatio-temporal data streams," in *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, ser. VLDB '04. VLDB Endowment, 2004, pp. 1377–1380. [Online]. Available: http://dl.acm.org/citation.cfm?id=1316689.1316838

[24] MySQL, "Mysql spatial extension documentation," https://dev.mysql.com/doc/refman/8.0/en/spatial-types.html, 2018.

[25] F. Warmerdam, "Postgis: A standards based geographic extension for postgresql," http://www.pgcon.org/2008/schedule/attachments.

[26] M. Orru, R. Paolillo, A. Detti, G. Rossi, and N. B. Melazzi, "Demonstration of opengeobase: The icn nosql spatio-temporal database," in *2017 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*, June 2017, pp. 1–2.

[27] Amazon, "Dynamodb documentation," https://aws.amazon.com/nosql/, 2018.

[28] M. F. Mokbel, T. M. Ghanem, and W. G. Aref, "Spatio-temporal access methods," *IEEE Data Engineering Bulletin*, vol. 26, pp. 40–49, 2003.

[29] L.-V. Nguyen-Dinh, W. G. Aref, and M. F. Mokbel, "Spatio-temporal access methods: Part 2 (2003 - 2010)," *IEEE Data Eng. Bull.*, vol. 33, pp. 46–55, 2010.

[30] A. Guttman, "R-trees: A dynamic index structure for spatial searching," *SIGMOD Rec.*, vol. 14, no. 2, pp. 47–57, Jun. 1984. [Online]. Available: http://doi.acm.org/10.1145/971697.602266

[31] G. Li and J. Tang, "A new r-tree spatial index based on space grid coordinate division," in *Proceedings of the 2011, International Conference on Informatics, Cybernetics, and Computer Engineering (ICCE2011) November 19–20, 2011, Melbourne, Australia*, L. Jiang, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 133–140.

[32] K. Deng, K. Xie, K. Zheng, and X. Zhou, "Trajectory indexing and retrieval," in *Computing with Spatial Trajectories.* Springer, 2011, pp. 35–60.

[33] D. Pfoser, C. S. Jensen, and Y. Theodoridis, "Novel approaches in query processing for moving object trajectories," in *Proceedings of the 26th International Conference on Very Large Data Bases*, ser. VLDB '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 395–406. [Online]. Available: http://dl.acm.org/citation.cfm?id=645926.672019

[34] J. E. B. Blomsøy, "Evaluating algorithms for nearest neighbor searches in spatial databases using r-trees," *NTNU Master thesis*, p. 107p, 2017.

[35] V. P. Chakka, A. C. Everspaugh, and J. M. Patel, "Indexing large trajectory data sets with seti," *Ann Arbor*, vol. 1001, pp. 48 109–2122, 2003.

[36] X. Xiong, M. F. Mokbel, and W. G. Aref, "Lugrid: Update-tolerant grid-based indexing for moving objects," in *7th International Conference on Mobile Data Management (MDM'06)*, May 2006, pp. 13–13.

[37] L. Lins, J. T. Klosowski, and C. Scheidegger, "Nanocubes for real-time exploration of spatiotemporal datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2456–2465, Dec. 2013. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2013.179

[38] Y. Theodoridis, T. K. Sellis, A. Papadopoulos, and Y. Manolopoulos, "Specifications for efficient indexing in spatiotemporal databases," in *Proceedings of the 10th International Conference on Scientific and Statistical Database Management*, ser. SSDBM '98. Washington, DC, USA: IEEE Computer Society, 1998, pp. 123–132. [Online]. Available: http://dx.doi.org/10.1109/SSDM.1998.688117

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

[40] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2149–2158, 2013.

[41] P. Cudre-Mauroux, E. Wu, and S. Madden, "Trajstore: An adaptive storage system for very large trajectory data sets," in *IEEE International Conference on Data Engineering (ICDE)*. IEEE, 2010, pp. 109–120.

[42] K. Hu, Z. He, and Y. Yue, "Taxi-viewer: Around the corner taxis are!" in *2010 7th International Conference on Ubiquitous Intelligence Computing and 7th International Conference on Autonomic Trusted Computing*, Oct 2010, pp. 498–500.

[43] D. Pfoser, C. S. Jensen, and Y. Theodoridis, "Novel approaches in query processing for moving object trajectories," in *Proceedings of the 26th International Conference on Very Large Data Bases*, ser. VLDB '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 395–406. [Online]. Available: http://dl.acm.org/citation.cfm?id=645926.672019

[44] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology*, 2014.

[45] Y. Zheng, "Trajectory data mining: An overview," *ACM Transaction on Intelligent Systems and Technology*, September 2015. [Online]. Available: https://www.microsoft.com/en-us/research/publication/trajectory-data-mining-an-overview/

[46] D. Sacharidis, K. Patroumpas, M. Terrovitis, V. Kantere, M. Potamias, K. Mouratidis, and T. Sellis, "On-line discovery of hot motion paths," in *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, ser. EDBT '08. New York, NY, USA: ACM, 2008, pp. 392–403. [Online]. Available: http://doi.acm.org/10.1145/1353343.1353392

[47] X. Li, J. Han, J.-G. Lee, and H. Gonzalez, "Traffic density-based discovery of hot routes in road networks," in *Proceedings of the 10th International Conference on Advances in Spatial and Temporal Databases*, ser. SSTD'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 441–459. [Online]. Available: http://dl.acm.org/citation.cfm?id=1784462.1784487

[48] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. Gonzalez, "Understanding road usage patterns in urban areas," in *Scientific reports*, 2012.

[49] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-drive: Enhancing driving directions with taxi drivers' intelligence," *IEEE Trans. on Knowl. and Data Eng.*, vol. 25, no. 1, pp. 220–232, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2011.200

[50] G. Andrienko, N. Andrienko, W. Chen, R. Maciejewski, and Y. Zhao, "Visual analytics of movement and transportation: State of the art and further research directions," *IEEE Transactions on Intelligent Transporation System*, p. To appear, 2017.

[51] W. Chen, F. Guo, and F.-Y. Wang, "A survey of traffic data visualization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 2970–2984, 2015.

[52] D. Phan, L. Xiao, R. Yeh, and P. Hanrahan, "Flow map layout," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, Oct 2005, pp. 219–224.

[53] I. Boyandin, E. Bertini, P. Bak, and D. Lalanne, "Flowstrates: An approach for visual exploration of temporal origin-destination data," in *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*, ser. EuroVis'11. Chichester, UK: The Eurographs Association &#38; John Wiley &#38; Sons, Ltd., 2011, pp. 971–980. [Online]. Available: http://dx.doi.org/10.1111/j.1467-8659.2011.01946.x

[54] J. Wood, J. Dykes, and A. Slingsby, "Visualisation of origins, destinations and flows with od maps," *The Cartographic Journal*, vol. 47, no. 2, pp. 117–129, 2010. [Online]. Available: https://doi.org/10.1179/000870410X12658023467367

[55] T. Kapler and W. Wright, "Geo time information visualization," *Information Visualization*, vol. 4, no. 2, pp. 136–146, Jul. 2005.

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

[57] C. Hurter, B. Tissoires, and S. Conversy, "FromDaDy: Spreading Aircraft Trajectories Across Views to Support Iterative Queries," *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, vol. 15, no. 6, pp. 1017–1024, 2009.

[58] N. Willems, H. van de Wetering, and J. J. van Wijk, "Visualization of vessel movements," in *EuroVis*. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2009, pp. 959–966. [Online]. Available: http://dx.doi.org/10.1111/j.1467-8659.2009.01440.x

[59] H. Liu, Y. Gao, L. Lu, S. Liu, L. Ni, and H. Qu, "Visual analysis of route diversity," *Proceedings of IEEE Conference on Visual Analytics Science and Technology*, October 2011.

[60] 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.

[61] M.-P. Kwan, "Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set," *Transportation Research Part C: Emerging Technologies*, vol. 8, no. 1, pp. 185 – 203, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X00000176

[62] M. J. Kraak, "The space-time cube revisited from a geovisualization perspective," *Proceedings of the 21st International Cartographic Conference*, vol. 1995, 1988. [Online]. Available: http://www.itc.nl/library/Papers_2003/art_proc/kraak.pdf

[63] I. Vasiliev, "Mapping Time," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 34, no. 2, pp. 1–51, 1997.

[64] N. Andrienko, G. Andrienko, and P. Gatalsky, "Supporting visual exploration of object movement," in *Proceedings of the Working Conference on Advanced Visual Interfaces*, ser. AVI '00. New York, NY, USA: ACM, 2000, pp. 217–220. [Online]. Available: http://doi.acm.org/10.1145/345513.345319

[65] N. Willems, R. Scheepens, H. van de Wetering, and J. J. van Wijk, *Visualization of Vessel Traffic.* New York, NY: Springer New York, 2013, pp. 73–87. [Online]. Available: https://doi.org/10.1007/978-1-4614-6230-9_5

[66] S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, and G. Andrienko, "Visually driven analysis of movement data by progressive clustering," *Information Visualization*, vol. 7, no. 3, pp. 225–239, Jun. 2008. [Online]. Available: http://dx.doi.org/10.1057/palgrave.ivs.9500183

[67] G. Andrienko, N. Andrienko, and S. Wrobel, "Visual analytics tools for analysis of movement data," *SIGKDD Explor. Newsl.*, vol. 9, no. 2, pp. 38–46, Dec. 2007. [Online]. Available: http://doi.acm.org/10.1145/1345448.1345455

[68] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti, "Interactive visual clustering of large collections of trajectories," in *2009 IEEE Symposium on Visual Analytics Science and Technology*, Oct 2009, pp. 3–10.

[69] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel, "From movement tracks through events to places: Extracting and characterizing significant places from mobility data," in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct 2011, pp. 161–170.

[70] 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, pp. 451–460, 2013.

[71] M. Lu, Z. Wang, and X. Yuan, "Trajrank: Exploring travel behaviour on a route by trajectory ranking," *Proceedings of IEEE Pacific Visualization Symposium*, pp. 14–17, 2015.

[72] M. Borkin, Z. Bylinskii, N. Kim, C. Bainbridge, C. Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond memorability: Visualization recognition and recall," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 22, no. 1, pp. 519–528, 2016.

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

[74] P. S. Castro, D. Zhang, C. Chen, S. Li, and G. Pan, "From taxi gps traces to social and community dynamics: A survey," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 17:1–17:34, Dec. 2013. [Online]. Available: http://doi.acm.org/10.1145/2543581.2543584

[75] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *Proceedings of SIGSPATIAL'13.* New York, NY, USA: ACM, 2013, pp. 344–353. [Online]. Available: http://doi.acm.org/10.1145/2525314.2525343

[76] J. Dai, B. Yang, C. Guo, and Z. Ding, "Personalized route recommendation using big trajectory data," in *Data Engineering (ICDE).* IEEE, 2015, pp. 543–554.

[77] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proceedings of the 13th International Conference on Ubiquitous Computing*, ser. UbiComp '11. New York, NY, USA: ACM, 2011, pp. 89–98. [Online]. Available: http://doi.acm.org/10.1145/2030112.2030126

[78] S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, and F. Giannotti, "Discovering the geographical borders of human mobility," *KI - Künstliche Intelligenz*, vol. 26, pp. 253–260, 2012.

[79] N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-finder: A recommender system for finding passengers and vacant taxis," *IEEE Trans. on Knowl. and Data Eng.*, vol. 25, no. 10, pp. 2390–2403, Oct. 2013. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2012.153

[80] F. Zhang, D. Wilkie, Y. Zheng, and X. Xie, "Sensing the pulse of urban refueling behavior," in *Proceedings of UbiComp '13.* New York, NY, USA: ACM, 2013, pp. 13–22. [Online]. Available: http://doi.acm.org/10.1145/2493432.2493448

[81] J. Zimmerman, A. Tomasic, C. Garrod, D. Yoo, C. Hiruncharoenvate, R. Aziz, N. R. Thiruvengadam, Y. Huang, and A. Steinfeld, "Field trial of tiramisu: Crowd-sourcing bus arrival times to spur co-design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 1677–1686. [Online]. Available: http://doi.acm.org/10.1145/1978942.1979187

[82] N. Lathia and L. Capra, "Mining mobility data to minimise travellers' spending on public transport," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1181–1189. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020590

[83] S. Liu, W. Cui, Y. Wu, and M. Liu, "A survey on information visualization: Recent advances and challenges," *Vis. Comput.*, vol. 30, no. 12, pp. 1373–1393, Dec. 2014. [Online]. Available: http://dx.doi.org/10.1007/s00371-013-0892-3

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

[85] Z. Wang, T. Ye, M. Lu, X. Yuan, H. Qu, J. Yuan, and Q. Wu, "Visual exploration of sparse traffic trajectory data," *IEEE Trans. on Vis. Comp. Graph.*, vol. 20, no. 12, pp. 1813 – 1822, 2014.

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

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

[88] F. Wang, W. Chen, F. Wu, Y. Zhao, H. Hong, T. Gu, L. Wang, R. Liang, and H. Bao, "Visual reasoning approach for data-driven transport assessment on urban road," in *IEEE Conference on Visual Analytics Science and Technology.* IEEE, Oct. 2014, pp. 103–112.

[89] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2149–2158, Dec. 2013. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2013.226

[90] B. C. Ooi, R. Sacks-Davis, and J. Han, "Indexing in spatial databases," http://www.comp.nus.edu.sg/ ooibc/spatialsurvey.pdf, 1993.

[91] J. Zhou, A. K. Tung, W. Wu, and W. S. Ng, "R2-d2: a system to support probabilistic path prediction in dynamic environments via semi-lazy learning," *Proceedings of the VLDB Endowment*, vol. 6, no. 12, pp. 1366–1369, 2013.

[92] S. Büttcher, C. Clarke, and G. V. Cormack, *Information Retrieval: Implementing and Evaluating Search Engines.* The MIT Press, 2010.

[93] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan, "Semantic trajectories modeling and analysis," *ACM Comput. Surv.*, vol. 45, no. 4, pp. 42:1–42:32, Aug. 2013.

[94] V. Bogorny, C. Renso, A. R. de Aquino, F. de Lucca Siqueira, and L. O. Alvares, "Constant – a conceptual data model for semantic trajectories of moving objects," *Transactions in GIS*, vol. 18, no. 1, pp. 66–88, 2014.

[95] R. Fileto, C. May, C. Renso, N. Pelekis, D. Klein, and Y. Theodoridis, "The baquara^2 knowledge-based framework for semantic enrichment and analysis of movement data," *Data Knowl. Eng.*, vol. 98, pp. 104–122, 2015.

[96] K. Zheng, S. Shang, N. J. Yuan, and Y. Yang, "Towards efficient search for activity trajectories." in *IEEE Data Engineering (ICDE).* IEEE Computer Society, 2013, pp. 230–241.

[97] B. Zheng, N. J. Yuan, K. Zheng, X. Xie, S. Sadiq, and X. Zhou, "Approximate keyword search in semantic trajectory database," in *IEEE Data Engineering (ICDE).* IEEE, 2015, pp. 975–986.

[98] H. Su, K. Zheng, K. Zeng, J. Huang, S. Sadiq, N. J. Yuan, and X. Zhou, "Making sense of trajectory data: A partition-and-summarization approach," in *IEEE Data Engineering (ICDE).* IEEE, 2015, pp. 963–974.

[99] Y. Chen, S. Barlowe, and J. Yang, "Click2annotate: Automated insight externalization with rich semantics," in *Proceedings of IEEE VAST*, 2010, pp. 155–162.

[100] D. Chu, D. A. Sheets, Y. Zhao, Y. Wu, J. Yang, M. Zheng, and G. Chen, "Visualizing hidden themes of taxi movement with semantic transformation," in *Pacific Visualization Symposium (PacificVis), 2014 IEEE*. IEEE, 2014, pp. 137–144.

[101] A. Bialecki, R. Muir, G. Ingersoll, and L. Imagination, "Apache lucene 4," in *Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval*, 2012.

[102] Apache, "Apache lucene 4.0.0 documentation," https://lucene.apache.org/core/4_0_0/, 2016.

[103] R. Kosara, F. Bendix, and H. Hauser, "Parallel sets: Interactive exploration and visual analysis of categorical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 4, pp. 558–568, 2006.

[104] M. A. Harrower and C. A. Brewer, "ColorBrewer.org: An Online Tool for Selecting Color Schemes for Maps," *The Cartographic Journal*, vol. 40, no. 1, pp. 27–37, 2003.

[105] S. Al-Dohuki, Y. Wu, F. Kamw, J. Yang, X. Li, Y. Zhao, X. Ye, W. Chen, C. Ma, and F. Wang, "Semantictraj: A new approach to interacting with massive taxi trajectories," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 11–20, Jan 2017.

[106] W. Zeng, C.-W. Fu, S. Arisona, A. Erath, and H. Qu, "Visualizing mobility of public transportation system," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 1833–1842, 2014.

[107] W. Chen, Z. Huang, F. Wu, M. Zhu, H. Guan, and R. Maciejewski, "Vaud: A visual analysis approach for exploring spatio-temporal urban data," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, no. 99, pp. 1–1, 2018.

[108] J. Poco, H. Doraiswamy, H. T. Vo, J. a. L. D. Comba, J. Freire, and C. T. Silva, "Exploring traffic dynamics in urban environments using vector-valued functions," in *Proceedings of the 2015 Eurographics Conference on Visualization*, ser. EuroVis '15. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2015, pp. 161–170. [Online]. Available: http://dx.doi.org/10.1111/cgf.12628

[109] G. Andrienko, N. Andrienko, S. Bremm, T. Schreck, T. von Landesberger, P. Bak, and D. Keim, "Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns," in *Proceedings of the 12th Eurographics / IEEE - VGTC Conference on Visualization*, ser. EuroVis'10. Chichester, UK: The Eurographs Association John Wiley Sons, Ltd., 2010, pp. 913–922. [Online]. Available: http://dx.doi.org/10.1111/j.1467-8659.2009.01664.x

[110] G. Andrienko, N. Andrienko, G. Fuchs, and J. Wood, "Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data," *IEEE Transactions on Visualization & Computer Graphics*, 2017.

[111] H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C. Silva, "Using topological analysis to support event-guided exploration in urban data," *IEEE Transactions on Visualization & Computer Graphics*, vol. 20, no. 12, pp. 2634–43, 2014.

[112] I. Kamel and C. Faloutsos, "Hilbert r-tree: An improved r-tree using fractals," in *Proceedings of the 20th International Conference on Very Large Data Bases*, ser. VLDB '94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 500–509. [Online]. Available: http://dl.acm.org/citation.cfm?id=645920.673001

[113] L.-A. Tang, Y. Zheng, X. Xie, J. Yuan, X. Yu, and J. Han, "Retrieving k-nearest neighboring trajectories by a set of point locations," in *Proceedings of the 12th International Conference on Advances in Spatial and Temporal Databases*, ser. SSTD'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 223–241. [Online]. Available: http://dl.acm.org/citation.cfm?id=2035253.2035272

[114] S. Liu, J. Pu, Q. Luo, H. Qu, L. Ni, and R. Krishnan, "Vait: A visual analytics system for metropolitan transportation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 4, pp. 1586–1596, 2013.

[115] C. A. L. Pahins, S. A. Stephens, C. Scheidegger, and J. L. D. Comba, "Hashedcubes: Simple, low memory, real-time visual exploration of big data," *IEEE Transactions on Visualization & Computer Graphics*, vol. 23, no. 1, pp. 671–680, 2017.

[116] Z. Liu, B. Jiang, and J. Heer, "immens: Real-time visual querying of big data," in *Proceedings of the 15th Eurographics Conference on Visualization*, ser. EuroVis '13. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2013, pp. 421–430. [Online]. Available: http://dx.doi.org/10.1111/cgf.12129

[117] H. Doraiswamy, H. T. Vo, C. T. Silva, and J. Freire, "A gpu-based index to support interactive spatio-temporal queries over historical data," *IEEE 32nd International Conference on Data Engineering (ICDE)*, pp. 1086–1097, 2016.

[118] J. Joaquim and S. Faramarz, *Sketch-based Interfaces and Modeling.* Springer-Verlag London, 2011.

[119] J. Browne, B. Lee, S. Carpendale, N. Riche, and T. Sherwood, "Data analysis on interactive whiteboards through sketch-based interaction," in *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, ser. ITS '11. New York, NY, USA: ACM, 2011, pp. 154–157. [Online]. Available: http://doi.acm.org/10.1145/2076354.2076383

[120] D. Schroeder and D. F. Keefe, "Visualization-by-sketching: An artist's interface for creating multivariate time-varying data visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 877–885, Jan 2016.

[121] B. Lee, R. H. Kazi, and G. Smith, "Sketchstory: Telling more engaging stories with data through freeform sketching," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2416–2425, Dec 2013.

[122] A. D. Blaser, *Sketching Spatial Queries.* Dissertation, University of Munster, 2000.

[123] A. Godwin and J. Stasko, "Drawing data on maps: Sketch-based spatiotemporal visualization," in *Poster in IEEE Visualization*, 2015.

[124] G. C. D. Silva, T. Yamasaki, and K. Aizawa, "Sketch-based spatial queries for retrieving human locomotion patterns from continuously archived gps data," *IEEE Transactions on Multimedia*, vol. 11, 2009.

[125] A. D. Blaser and M. J. Egenhofer, "A visual tool for querying geographic databases," in *Proceedings of the Working Conference on Advanced Visual Interfaces*, ser. AVI '00. New York, NY, USA: ACM, 2000, pp. 211–216. [Online]. Available: http://doi.acm.org/10.1145/345513.345318

[126] K. D. Forbus, J. Usher, and V. Chapman, "Qualitative spatial reasoning about sketch maps," *AI Mag.*, vol. 25, no. 3, pp. 61–72, Sep. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1045744.1045752

[127] C. Turkay, A. Slingsby, H. Hauser, J. Wood, and J. Dykes, "Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data," *IEEE Transactions on Visualization & Computer Graphics*, vol. 20, no. 12, pp. 2033–2042, 2014.

[128] A. Malik, R. Maciejewski, N. Elmqvist, Y. Jang, D. S. Ebert, and W. Huang, "A correlative analysis process in a visual analytics environment," in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct 2012, pp. 33–42.

[129] A. Godwin and J. Stasko, "Hotsketch: Drawing police patrol routes among spatiotemporal crime hotspots," *Proceedings of the International Conf. on Systems Sciences*, pp. 1372–1380, 2017.

[130] Q. Zeng, W. Chen, H. Wang, C. Tu, D. Cohen-Or, D. Lischinski, and B. Chen, "Hallucinating stereoscopy from a single image," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 1–12, May 2015. [Online]. Available: http://dx.doi.org/10.1111/cgf.12536

[131] L. Shao, S. Mittelstädt, R. Goldblatt, I. Omer, P. Bak, and T. Schreck, "StreetExplorer: Visual Exploration of Feature-based Patterns in Urban Street Networks," *Proc. Int. Conference on Information Visualization Theory and Applications*, 2016.

[132] B. Shneiderman, "Touch screens now offer compelling use," *IEEE Software*, vol. 8, no. 2, pp. 93–94, 1991.

[133] M. Ronneberg, T. Sarjakoski, and L. T. Sarjakoski, "Developing a multi-touch map application for a large screen in a nature centre," *Nordic Journal of Surveying and Real Estate Research*, vol. 10, no. 1, pp. 47–62, 2014.

[134] M. Ziat, H.-Y. Yao, R. Schmitt, and V. Hayward, "Frontpanel: Tangible user interface for touch-screens dedicated to elderly," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '16. New York, NY, USA: ACM, 2016, pp. 3808–3811. [Online]. Available: http://doi.acm.org/10.1145/2851581.2890266

[135] J. J. Van Wijk and E. R. Van Selow, "Cluster and calendar based visualization of time series data," in *Proceedings of the 1999 IEEE Symposium on Information Visualization*, ser. INFOVIS '99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 4–. [Online]. Available: http://dl.acm.org/citation.cfm?id=857189.857665

[136] B. C. L. T. Trajectories., "Taxi trajectory data," Website, 2013, http://www.beijingcitylab.com/data-released-1/data1-20/.

[137] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, , and A. Rabuffi, "Crawdad roma/taxi data set," Website, 2014, http://crawdad.org/roma/taxi/.

[138] D. Donovan, Brian; Work, "New york city taxi trip data (2010-2013) ," 2016. [Online]. Available: https://doi.org/10.13012/J8PN93H8

[139] L. Sun, D.-H. Lee, A. Erath, and X. Huang, "Using smart card data to extract passenger's spatio-temporal density and train's trajectory of mrt system," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, ser. UrbComp '12. New York, NY, USA: ACM, 2012, pp. 142–148. [Online]. Available: http://doi.acm.org/10.1145/2346496.2346519

[140] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 791–800. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526816

[141] W. Guan, "A qualitative model of cross-lane inhomogeneities in traffic flow," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 3, pp. 188–199, 2004. [Online]. Available: http://dx.doi.org/10.1109/TITS.2004.833760

[142] C. Musselwhite and G. Lyons, "Exploring the public acceptability of road pricing," *Proc. 41st Universities Transport Study Group Conferences, University College London*, 2009.

[143] INRO, "EMME," Website, 2018, https://www.inrosoftware.com/en/products/emme/.

[144] Z. Li, J.-G. Lee, X. Li, and J. Han, "Incremental clustering for trajectories," in *Proceedings of the 15th International Conference on Database Systems for Advanced Applications - Volume Part II*, ser. DASFAA'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 32–46. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12098-5_3

[145] Caliper, "TransCAD Transportation Planning Software," Website, 2014, http://www.caliper.com/tcovu.htm.

[146] Citilab, "Cube," Website, 2014, http://www.citilabs.com/.

[147] N. Andrienko and G. Andrienko, "Research homepage," Website, http://geoanalytics.net/and/.

[148] E. Weiner, *Urban Transportation Planning in the United States: History, Policy, and Practice*, 01 2008.

[149] M. Picone, S. Busanelli, M. Amoretti, F. Zanichelli, and G. Ferrari, *Advanced Technologies for Intelligent Transportation Systems.* Springer Publishing Company, Incorporated, 2014.

[150] G. Chowell, J. M. Hyman, S. Eubank, and C. Castillo-Chavez, "Scaling laws for the movement of people between locations in a large city." *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 68 6 Pt 2, p. 066102, 2003.

[151] S. Sang, M. O'Kelly, and M.-P. Kwan, "Examining commuting patterns: Results from a journey-to-work model disaggregated by gender and occupation," *Urban Studies*, vol. 48, pp. 891–909, 04 2011.

[152] Y. Liu, C. Kang, S. Gao, Y. Xiao, and Y. Tian, "Understanding intra-urban trip patterns from taxi trajectory data," *Journal of Geographical Systems*, vol. 14, no. 4, pp. 463–483, Oct 2012. [Online]. Available: https://doi.org/10.1007/s10109-012-0166-z

[153] P. Pavithra, "Network Structure and Metropolitan Mobility," *Journal of Transport and Land Use*, vol. 7, no. 2, 2014.

[154] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti, "Activity-aware map: Identifying human daily activity pattern using mobile phone data," in *Human Behavior Understanding*, A. A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 14–25.

[155] C. Ratti, D. Frenchman, R. M. Pulselli, and S. Williams, "Mobile landscapes: Using location data from cell phones for urban analysis," *Environment and Planning B: Planning and Design*, vol. 33, no. 5, pp. 727–748, 2006. [Online]. Available: https://doi.org/10.1068/b32047

[156] A. Sevtsuk and C. Ratti, “Does urban mobility have a daily routine ? learning from the aggregate data of mobile networks,” 2010.

[157] B. Jiang, “Street hierarchies: a minority of streets account for a majority of traffic flow,” *International Journal of Geographical Information Science*, vol. 23, pp. 1033–1048, 2009.

[158] B. Jiang and C. Liu, “Street-based topological representations and analyses for predicting traffic flow in gis,” *International Journal of Geographical Information Science*, vol. 23, pp. 1119–1137, 2009.

[159] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González, “Understanding Road Usage Patterns in Urban Areas,” *Scientific Reports*, vol. 2, no. arXiv:1212.5327, p. 1001. 47 p, Dec 2012, comments: 47 pages, 24 figures.

[160] S. Gao, Y. Wang, Y. Gao, and Y. Liu, “Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality,” *Environment and Planning B: Planning and Design*, vol. 40, no. 1, pp. 135–153, 2013. [Online]. Available: https://doi.org/10.1068/b38141

[161] X. Dai, M. A. Ferman, and R. P. Roesser, “A simulation evaluation of a real-time traffic information system using probe vehicles,” in *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*, vol. 1, Oct 2003, pp. 475–480 vol.1.

[162] D. Tong, C. J. Merry, and B. Coifman, “New perspectives on the use of gps and gis to support a highway performance study,” *Transactions in GIS*, vol. 13, no. 1, pp. 69–85. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9671.2009.01144.x

[163] L. Liu, C. Andris, and C. Ratti, “Uncovering cabdrivers’ behavior patterns from their digital traces,” *Computers, Environment and Urban Systems*, vol. 34, no. 6, pp. 541 – 548, 2010, geoVisualization and the Digital City. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0198971510000773

[164] Q. Li, T. Zhang, H. Wang, and Z. Zeng, “Dynamic accessibility mapping using floating car data: a network-constrained density estimation approach,” *Journal of Transport Geography*, vol. 19, no. 3, pp. 379–393, 2011, special Issue : Geographic Information Systems for Transportation. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0966692310001158

[165] G. Qi, X. Li, S. Li, G. Pan, Z. Wang, and D. Zhang, “Measuring social functions of city regions from large-scale taxi behaviors,” in *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, March 2011, pp. 384–388.

[166] D. Pfoser, S. Brakatsoulas, P. Brosch, M. Umlauft, N. Tryfona, and G. Tsironis, "Dynamic travel time provision for road networks," in *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '08. New York, NY, USA: ACM, 2008, pp. 68:1–68:4. [Online]. Available: http://doi.acm.org/10.1145/1463434.1463513

[167] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, Mar. 2013. [Online]. Available: http://dx.doi.org/10.1038/srep01376

[168] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, ser. UrbComp '12. New York, NY, USA: ACM, 2012, pp. 1–8. [Online]. Available: http://doi.acm.org/10.1145/2346496.2346498

[169] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan, "Semantic trajectories modeling and analysis," *ACM Comput. Surv.*, vol. 45, no. 4, pp. 42:1–42:32, Aug. 2013.

[170] H. Wang, H. Zou, Y. Yue, and Q. Li, "Visualizing hot spot analysis result based on mashup," in *Proceedings of the 2009 International Workshop on Location Based Social Networks*, ser. LBSN '09. New York, NY, USA: ACM, 2009, pp. 45–48. [Online]. Available: http://doi.acm.org/10.1145/1629890.1629900

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

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

[173] 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.

[174] J. J. Thomas and K. A. Cook, "A visual analytics agenda," *IEEE Comput. Graph. Appl.*, vol. 26, pp. 10–13, January 2006. [Online]. Available: http://portal.acm.org/citation.cfm?id=1110637.1110648

[175] J. Wood, J. Dykes, A. Slingsby, and K. Clarke, "Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup." *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1176–1183, Nov. 2007.