Technische Universität München

Department of Civil, Geo and Environmental Engineering

Chair of Cartography

Prof. Dr.-Ing. Liqiu Meng

Spatiotemporal Visual Analysis of Traffic Flow Patterns Related to Transport Hubs from Floating Car Data

Shirui, WANG

Master's Thesis

Duration:	01.07.2015 – 31.12.2015
Study Course:	Cartography M.Sc.
Supervisor:	Linfang, Ding

2015

DECLARATION OF AUTHORSHIP

Last name: First name:

I declare that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations, except as acknowledged, and has not been submitted, either in part or whole, for a degree at this or any other University.

Formulations and ideas taken from other sources are cited as such. This work has not been published.

(Location, Date) (Signature)

ACKNOWLEDGMENT

Foremost, I would like to express my deep and sincere gratitude to my academic advisor, Linfang Ding, M.Sc, for supporting me to successfully complete my thesis, and for her kind assistance, insightful comments, and knowledge which added considerably to my academic experience. My research would not have succeeded without her help.

Equally, I would like to express my sincere gratitude to all people who helped me through this Master study period. It would not be a fun and joyful experience for me to attend this Master study without support from all of the people in the third batch of Master of Cartography and all the lecturers, especially Juliana who help me a lot at the end of my study. It is a really great experience to find a new international family and study with all of you. My deep gratitude and love to each one of you.

In addition, a special thank you goes to my parents, Ying, and Wenjun, for their constant encouragement and financial support and without the help of whom I could not have pursued this Master's degree.

ABSTRACT

Transport hubs such as airports and railway stations are places where plenty of passengers are exchanged between vehicles or between transport modes. Analyzing the patterns of traffic flows in/out of the transport hubs may help traffic engineers to better understand passenger behaviors and to improve the transportation planning. However, dealing with the movement data, is very challenging due to their large data volume, implicit spatiotemporal relationship, and uncertain semantics.

The goal of this thesis is for visual analysis of the traffic flow patterns related to transport hubs using floating car data. We propose a visual analysis workflow incorporating computational algorithms, data mining approaches, and visualization techniques for exploration of the spatiotemporal patterns of taxi flows. More specifically, we preprocess a large amount of movement data, reconstruct trajectories and extract the starting and ending points especially related to the transport hubs. Secondly, we identify appropriate spatial and semantic clustering methods to derive and categorize transport hubs related significant places. Finally, we design appropriate spatial and temporal visualization techniques, e.g. dot maps, proportional symbol maps, pie chart maps, to visually analyze those significant places.

We use one-week Floating Car Data (FCD) in Shanghai as our test dataset and select Hongqiao international airport as the test transport hub. By applying our framework to the test dataset, the experiment results reveal significant spatiotemporal traffic flow patterns related to Hongqiao airport, which demonstrates the feasibility of the proposed workflow.

Keywords: Visual Analytics, Traffic Flow, Transport Hubs, FCD

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

FCD	Floating Car Data
FTD	Floating Taxi Data
GPS	Global Positioning System
ITS	Intelligent Transportation Systems
FCE	Floating Car Evaluator
KDD	Knowledge Discovery and Data mining
3D	Three-Dimensional
KDE	Kernel Density Estimation
ROI	Region of Interest

1. INTRODUCTION

1.1 BACKGROUND

A transport hub is a place where passengers and cargo are exchanged between vehicles or between transport modes. Public transport hubs include train stations, rapid transit stations, bus stops, tram stop, airports and ferry slips. In particular, airports and railway stations play an important role in the transportation system and urban development. For example, airports have a twofold hub function. First they concentrate passenger traffic into one place for onward transportation. This makes it important for airports to be connected to the surrounding transport infrastructure, including roads, bus services, railway and rapid transit systems. Secondly some airports function as intramodular hubs for the airlines, or airline hubs. This is a common strategy among network airlines if the passengers want to commute between two cities when the airline does not fly directly. Then they usually make their change at one of the hubs.

Therefore, these transport hubs absorb and reflect huge amounts of traffic flows from or to the road network and have considerable social and economic impacts on the surrounding regions. Understanding the traffic flows of the transport hubs could help traffic engineers and policy makers to improve the transportation planning and support passengers to efficiently plan their trips.

For the analysis of traffic flow patterns, accurate traffic movement data are necessary. In recent years, with the development of the Global Positioning System (GPS) technology, a novel technology called Floating Car Data (FCD) was developed. FCD is a method to gather road information where cars act as mobile sensor nodes that are equipped with a location detecting device such as a GPS unit and a communication device such as a cellular phone. This system is based on the exchange of information between a fleet of floating cars traveling on a road network and a central data system (Fabritiis, Ragona, and Valenti, 2008). The data provided by different sensors are spatiotemporal data which contain both the spatial locations of the vehicles and the time recorded. By collecting and processing this kind of data, traffic flows could be analyzed in a different period of time.

1.2 RESEARCH GOALS

Analyzing the FCD is very challenging due to their large data volume, implicit spatiotemporal relationship, and uncertain semantics. Therefore, the goal of this study is to investigate effective visual analytics for exploring spatiotemporal, semantic patterns in the traffic flows in or out of the transport hubs.

To achieve this, this thesis tries to answer the following questions:

- How to propose an appropriate visual analysis workflow for the exploration of the patterns of traffic flows in or out of the transport hubs?
- Which spatial clustering method is suitable to extract and categorize significant places?
- Which visualization methods are appropriate to present the research results?

1.3 THESIS STRUCTURE

This thesis is structured into 6 chapters. Chapter 1 introduces the topic of this thesis and shows an overview of the work with the research purpose. The state of the art in Chapter 2 presents the state of art research in visual analytics of movement FTD. Chapter 3 describes the attributes and properties of the initial test data and the study area. Chapter 4 shows the detailed workflow of preprocessing, spatial clustering, semantic classification, and visualization of FTD. In chapter 5 we concentrate on analyzing the research results. Chapter 6 concludes this thesis and proposes further research ideas of possibilities for the use of FTD.

2. STATE OF THE ART

2.1 FLOATING CAR DATA

2.1.1 The Background and principle of FCD

In recent years, the Intelligent Transportation Systems (ITS) has become a popular issue in the transportation management, especially in terms of traffic monitoring systems by using traffic information data derived from the traffic sensors. ITS is built to bring innovation to improve transportation systems so that transportation problems such as traffic congestion could be solved. It deals with data information and communication technology in vehicles, between vehicles (e.g. car-to-car), and between vehicles and fixed location (e.g. car-to-infrastructure) that could be used to provide road information to guide transportation systems users and traffic monitoring.

There are many technologies that have been developed to provide traffic data in ITS. The conventional and most common method is by using sensors, such as inductive loops. An inductive-loop detector senses the presence of a conductive metal object by inducing currents in the object, which reduces the loop inductance. Inductive loop detectors are installed in the roadway surface. This sensor gathers traffic information from vehicles which pass the sensor. Therefore, this information has a big limitation as it could only provide information traffic estimation in a certain road segment which could not be an accurate representation for all road segments. Another limitation is that these sensors are quite expensive and need to be placed on the road which prone to be broken because of heavy vehicles. In addition, for a long road segment, the sensors need to be placed in several places in a certain interval to maintain an accurate measurement (Jain, Sharma, and Subramanian, 2012).

Another method that can be used to gather traffic information is by using CCTV camera. This method used to monitor real-time traffic by utilizing CCTV camera images for measuring the density of the vehicles. The advantage of using this method is that the installation of cameras does not involve breaking up pavement, which is a necessity for installing ground sensors. However this method also has some disadvantages, such as low camera resolution, which results in highly noisy images, traffic camera’s limited field of view, and light illumination from multiple reflecting sources, which distorts the capabilities of vehicle classification.

The floating car data (FCD) technology is a new and better approach to gather traffic information for most ITS. It is a method to determine the traffic speed on the road network. It is based on the collection of localization data, speed, the direction of travel and time information from mobile phones in vehicles that are being driven. This means that every vehicle with an active mobile phone acts as a sensor for the road network. Based on these data, movement data can be monitored, travel times can be calculated, and traffic reports can be rapidly generated. In contrast to traffic cameras and induction loops embedded in the roadway, no additional hardware on the road network is necessary. This method is quite useful because it could represent a full coverage of monitored areas automatically in real time with minimum cost and still generate a high quality of data. The precision of the vehicle location is relatively high. In the case of urban traffic, taxi fleets are particularly useful due to their large dataset and their on-board communication system (Leduc, G., 2008).

Basically, there are two types of FCD, GPS, and cellular-based systems.

A. GPS-based FCD System

GPS system utilizes the GPS receiver system which is already attached on the car, for example, taxis or courier services, to gather information about the vehicles. With this technology, the floating data is derived from a different type of devices. Then the data is communicated with the service provider using the regular on-board radio unit or via cellular network data. Therefore, the system can locate the exact location and movement of that specific car, for instance, calculating the instantaneous speed. The working principle is shown in Figure 1.

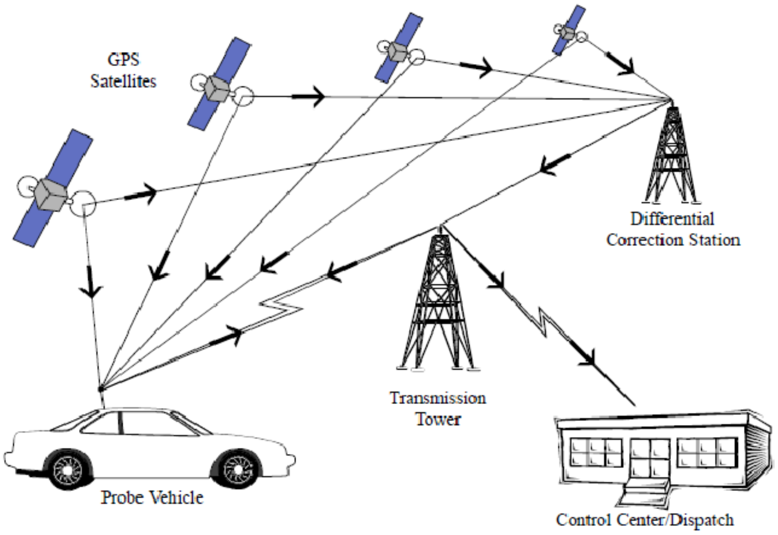


Figure 1 Communication from GPS (FHWA, 1998)

B. Cellular-based FCD (CFCD) System

CFCD is derived from cellular networks. The main benefit is that no special devices or hardware are needed and every mobile phone becomes, in fact, a sensor. The location and movement of the mobile phone are determined using one of several location technologies available by the mobile network. A large number of mobile handsets that are constantly on the move makes it possible to extract high-quality data from the network. Then the data will be sent to the data center through the regular on-board radio unit or via cellular network data. However, more complicated algorithms are required to extract the information. We can see how it works in Figure 2.

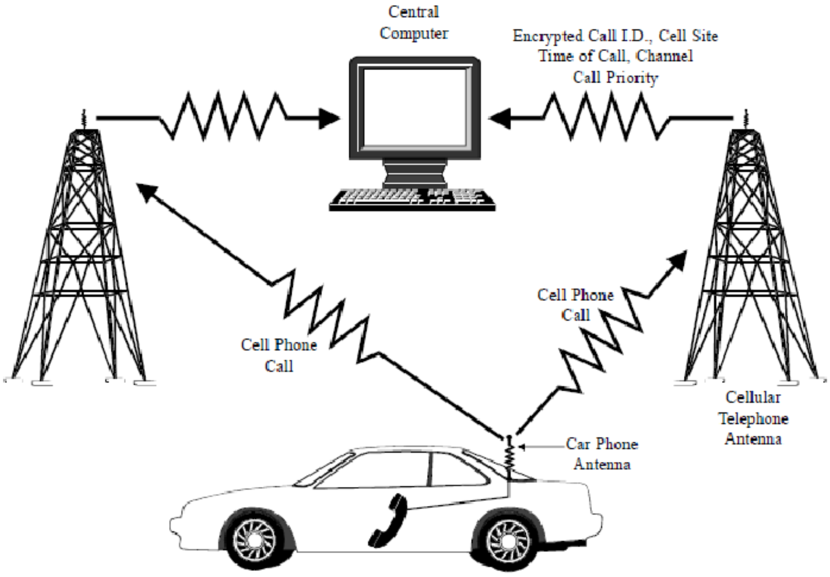


Figure 2 Communication from cellular phone (FHWA, 1998)

2.1.2 The Application of FCD

There are many benefits using FCD as FCD provides a network-wide, accurate and real-time information, which is less cost and constantly accessible. These characteristics make this method gain popularity to provide data for traffic management system. FCD also has some disadvantages such as complex data processing, the massive volume of data storage, no direct information on traffic flow or density, and privacy issues. However, there are many applications that could benefit from the improvement of FCD, especially in the transportation field. Congestion monitoring, bottleneck analysis, traffic growth and route choice analysis for traffic simulation are some examples of the potential application areas.

Actors	**Applications**
Government/public authorities	Congestion monitoring; local transport plans; journey time studies; planning studies; air pollution studies; OD matrices
Logistic and fleet operators	Logistic and fleet operators
Location based service providers	Location based service providers
Consultants	Congestion monitoring; journey time studies; planning studies; air pollution studies; transport studies
Map providers	Predictive journey times
Marketing	Optimized Traffic Systems - Static mobile sites; campaign planning; site planning
Automotive manufacturers	RDS-TMC live data for mobility portal; NavTrack GPS tracking solutions
Telecommunications	Real-time traffic information; short dial telephone traffic service

Table 1 Potential applications derived from the FCD Technology
Source: *ITIS Holdings in Leduc, G (2008)*

There have been many research projects related to FCD applications. In Europe, several FCD projects have been conducted during the last decade. OPTIS in Sweden uses FCD to collect data of traffic condition to provide traffic information for travelers. From the trials, FCD is proven to be a cost-effective method to provide accurate real-time traffic information for the user. Media-mobile in France is using FCD data gathered from 1,700 taxis operating in Paris to provide live traffic information on motorway congestion and traffic congestion for Paris.

Reinthaler, et al (2007) proposed a project called Dmotion which aims to provide an effective traffic management strategy for regional and local authorities. This project tries to provide an efficient way for estimating speeds and traffic state based on FCD data from taxi fleets and public transport. The data is provided by 1,200 taxis, then a multi-stage algorithm (Floating Car Evaluator, FCE) is employed to calculate the speed per road link. Data from public transport are used to complete missing data from FCD. As a result, the estimated path speed based on FCD follow the trend of the average path speed measured by ANPR (Automated Number Plate Recognition) but the level of estimated speeds is lower than average path speeds. However, this result shows that temporal trends of average speeds during a day could be captured by FCD data and this data prove to be a reliable and cheap additional data source for urban traffic state estimation.

Another application is proposed by Tang, et al (2012) which tries to utilize FCD data to detect and update changes in the road network. They choose taxi fleets data because taxis travel all over the city every day. The road network change detection and update follow these steps: data preprocessing, map matching, incremental detection, new data sampling, road network update and new road network detection. The most crucial part in this study is the map matching process as this step will detect whether the position of GPS point data match with the existing road network or not which means that an addition to the road network is needed.

2.2 VISUAL ANALYTICS OF FCD

2.2.1 Data Mining

Computer utilization has become a key role in improving data acquisition and processing to extract relevant information from data sources. Knowledge discovery and data mining (KDD) is the computational process of discovering patterns in large data set involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. KDD methods are mostly suitable for evaluating the quality of the proposed solutions of the problem. Therefore, these methods do not take analyst‘s knowledge into account to provide a solution for a problem.

The main idea of KDD is to extract information from large datasets. In KDD, there are some stages of the process to transform data into various model or representation to obtain the pattern that represents the implicit information within the data. The KDD process consists of data selection, data pre-processing, data transformation, data mining, and interpretation or evaluation. In the pre-processing stage, the dataset will be cleaned from the noises and missing data and formatted to suit the data mining algorithms. This stage is an essential process as the data could be heterogeneous (e.g. textual data, data stored in the database, satellite imagery, etc). Therefore, it requires effective methods for data cleaning and integration. The output of this stage then will be transformed into a form that can be understood by the analyst.

The current KDD methods are not directly applicable in visual analytics scenarios and only support limited user interaction. The models and patterns extracted by traditional KDD from a larger dataset could also be difficult to interpret in which the information within the dataset might be still hidden in the large volume of data. Hence for data mining method to be useful in visual analytics, the KDD methods should be (Keim et al, 2010):

- ✓ Fast enough: sub-second response is needed for efficient interaction.
- ✓ Parameters of the method should be representable and understandable using visualizations.
- ✓ Parameters should be adjustable by visual controls.

An example of data mining is spatial data mining. Spatial data mining could be interpreted as a process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases (Sumathi, et al., 2008). Spatial databases contain spatial and non-spatial attributes of the areas under the study, in which spatial data mining could be done to find implicit rules or patterns hidden in spatial databases that could be helpful for some fields such as geo-marketing or traffic control. There are many techniques that could be used in this method, such as clustering, association and co-location and trend detection. Trend detection or frequent pattern mining is a technique to find existing patterns in data which then could be used to predict trends of the attribute changes with respect to the neighborhood of some spatial objects. This function is the most basic function in data mining. Frequent patterns are those patterns that occur frequently in the database. Those patterns could be used to find predictive trends of the data.

As above mentioned, KDD methods are useful but still limited. Therefore, an integration with data visualization methods is needed to support pattern identification for the spatial dataset.

2.2.2 Spatial Clustering Methods in Data Mining

Spatial Clustering is the process of grouping a set of objects into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters. As a branch of statistics, cluster analysis has been studied extensively for many years, focusing mainly on distance-based cluster analysis. Cluster analysis tools based on Hierarchical cluster, K-means and several other methods have also been built into many statistical analysis software packages or systems. Clustering has also been studied in the field of machine learning as a type of unsupervised learning because it does not rely on predefined class and class-labeled training examples. However, efforts to perform effective and efficient clustering on large databases only started in recent years with the emergence of data mining.

Spatial clustering is a process of grouping a set of spatial objects which have a high degree of similarity into one cluster group. In general, clustering could be classified into four categories: partitioning method, hierarchical method, density based method and grid based method. In partitioning method, the data is considered as one big cluster which then will be classified into certain predefined numbers of the group (*k*). For each cluster, we then find the *k*-cluster centers and assign objects to the nearest cluster center. In the hierarchical method, the dataset is hierarchically decomposed based on the ideas that nearby objects have more similarity than far away objects which means that the distance between objects is used to describe each cluster. In density based method, clusters are defined by the density of objects on the area which separates the areas of high density and of low density. In grid-based methods, a finite number of cells of grid structure need to be built firstly in the area. The cells which contain more than a certain points or objects are treated as dense area.

In order to choose a clustering algorithm that is suitable for a particular application, many factors have to be considered. These include:

1) Application goal

   The goal of an application will often affect the type of clustering algorithms being used. For example, when trying to discover some good locations for setting up their stores, a supermarket chain may like to cluster their customers such that the sum of the distance to the cluster center is minimized. For such application where the distance to the cluster center is desired to be short, partitioning algorithms like K-means and K-medoids are often used. On the other hand, in applications like raster data analysis and image recognition, it is often desirable to find natural clusters, i.e., clusters which are perceived as crowded together by the human eye. In such a case, clusters discovered should have certain uniformity in density, colors, etc. and can be of arbitrary shape and size. Since algorithms like k-means and k-medoid tend to discover clusters with a spherical shape and similar size, density-based algorithms will be more suitable for these applications.

2) Trade-offs between quality and speed

As a rule of thumb, there will usually be tradeoffs between the speed of a clustering algorithm and the quality of the clusters it produces. A suitable clustering algorithm for an application must satisfy both the quality and speed requirements. Often, the size of the data being clustered plays an important factor in the running time of a clustering algorithm. A clustering algorithm that produces good quality clusters may, however, be unable to handle large datasets, making it suitable only for applications with a small database. To handle large datasets, a common approach is to perform some form of compression is usually lossy, the quality of the resulting clusters will often drop. The challenge in this approach is to compress the data in such a way that the speed of clustering increases substantially with minimal drop in cluster quality.

3) Characteristics of the data

The characteristics of the data being clustered serve as another factor in determining the clustering algorithm being applied. These characteristics include:

- The types of data attribute

  The similarity between two data objects is judged by the difference in their data attributes. When all these attributes are numeric, distance measures like Euclidean distance and Manhattan distance can be easily computed. However, when binary, categorical and ordinal attributes are involved, the computation of distance measure is much more complicated. Presently, most clustering algorithms are based on numeric attributes.

- Dimensionality

  The dimensionality of the data refers to the number of attributes in a data object. Many clustering algorithms which perform well on low-dimensional data degenerate when the number of dimensions increases. The degeneration can come in two forms: increase in running time or decrease in cluster quality. In order to choose a clustering algorithm for high-dimensional data clustering, the running time and cluster quality produced by the algorithm at high dimension must first be assessed to meet the requirement of the application.

- Amount of noise in data

  As some of the clustering algorithms are very sensitive to noise and outliers, a careful choice must be made if the data in the application contains a large amount of noise.

Here we would like to focus on the main two clustering algorithms: K-means and Hierarchical clustering, which work reasonably well for large geographical databases.

A. K-means clustering

K-means is one of the unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes a different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done.

At this point, we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result, of this loop, we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move anymore. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$\mathrm{J(V)} = \sum_{\mathrm{i=1}}^{\mathrm{c}} \sum_{\mathrm{j=1}}^{\mathrm{c_i}} \left(\left\| \mathrm{x_i} - \mathrm{v_j} \right\| \right)^2$$

where,

‘$\|x_i - v_j\|$’ is the Euclidean distance between xi and vj.

‘C_i’ is the number of data points in i^{th} cluster.

‘C’ is the number of cluster centers.

As we can see, the advantages of K-means clustering are:

1) Fast, robust and easier to understand.
2) Relatively efficient: n is the number of objects, k is the number of clusters, d stands for the dimension of each object, and t is the number of iterations. Normally, k, t, d << n.
3) Gives the best result when data set is distinct or well separated from each other.

However there are some disadvantages of this algorithm:

1) The learning algorithm requires apriori specification of the number of cluster centers.
2) The learning algorithm is not invariant to non-linear transformations i.e. with different representation of data we get
3) Euclidean distance measures can unequally weight underlying factors.
4) The learning algorithm provides the local optima of the squared error function.
5) Randomly choosing of the cluster center cannot lead us to the fruitful result.
6) Applicable only when the mean is defined, i.e. fails for categorical data.
7) Unable to handle noisy data and outliers.
8) Algorithm fails for the non-linear data set. See an example of Figure 3.

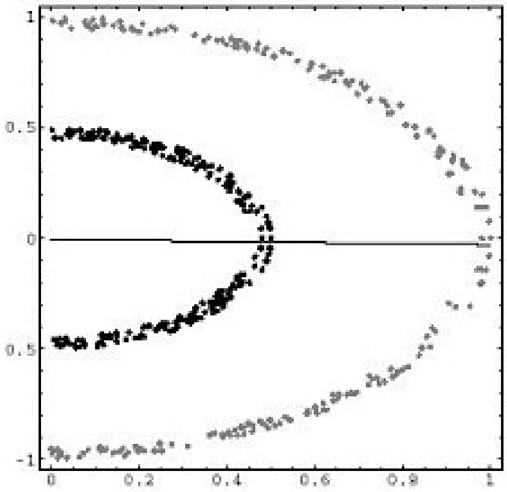


Figure 3 Showing the non-linear data set where k-means algorithm fails

Figure 4 is an example of how K-means clustering algorithm works on discrete geographical points. The '+' sign indicates the changes in cluster representative locations and data assignments is indicated by color.

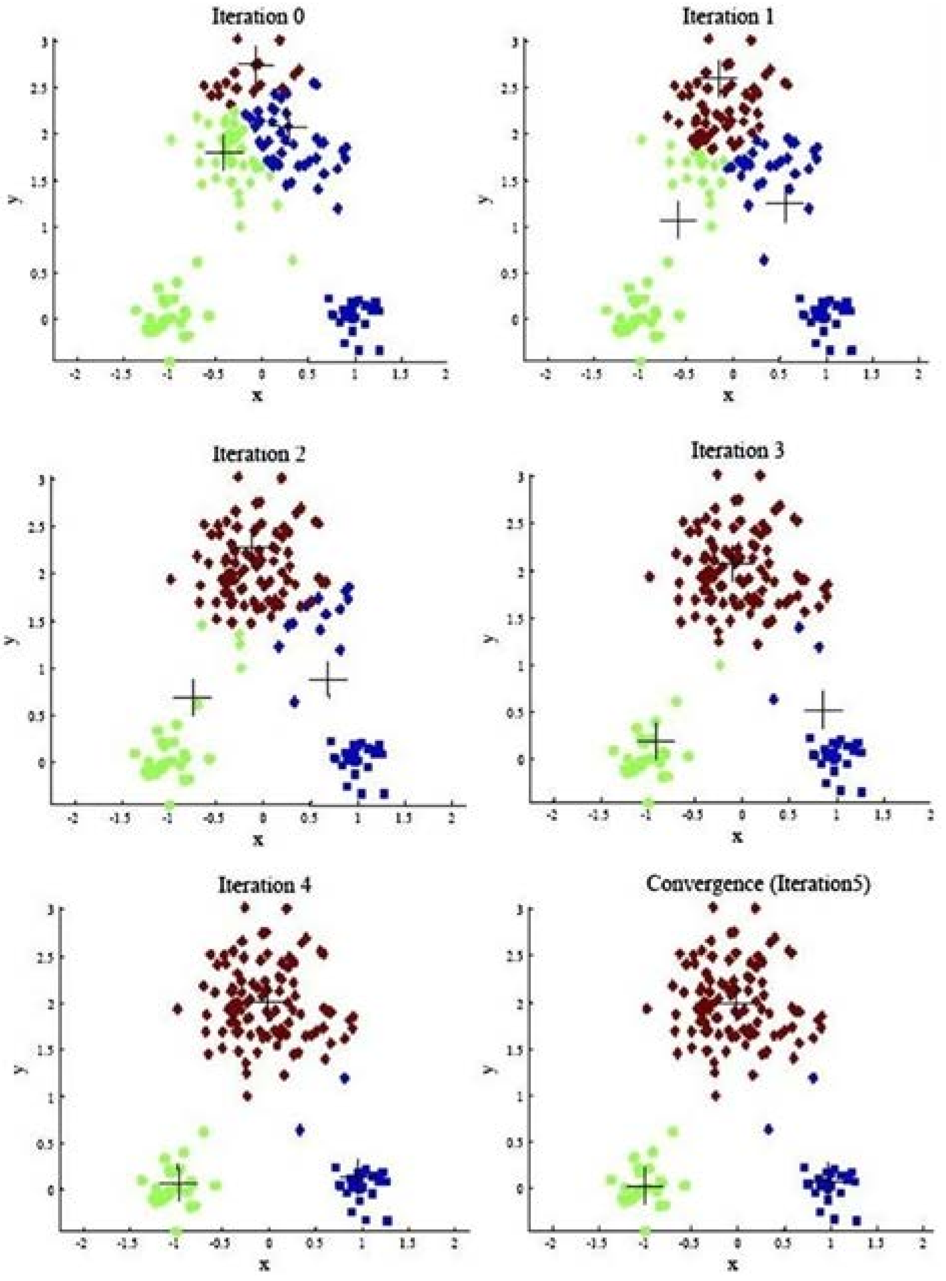


Figure 4 An execution of the k-means algorithm

Source: *https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm/k-means-figure*

B. Hierarchical Clustering

A hierarchical method creates a hierarchical decomposition of the given set of data objects forming a dendrogram – a tree which splits the database recursively into smaller subsets. The dendrogram can be formed in two ways: 'bottom-up' or 'top-down', also known as 'agglomerative' and 'divisive' approaches.

- Agglomerative: this is a 'bottom-up' approach. It starts with each object forming a separate group. It successively merges the objects of two groups and this is done until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds.
- Divisive: this is a 'top-down' approach. It starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters according to some measures until eventually each object is in one cluster or until a termination condition holds.

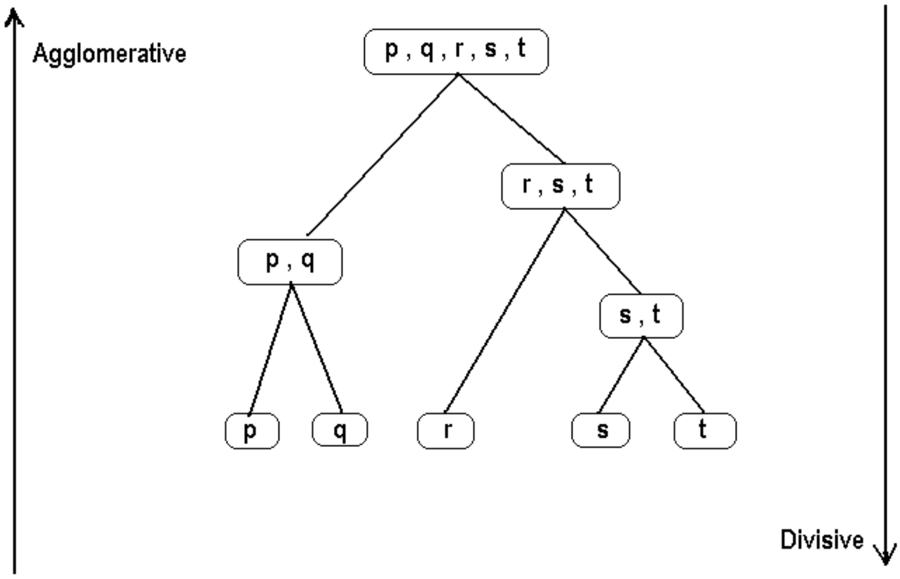


Figure 5 An example of 'agglomerative' and 'divisive' approach

In hierarchical clustering, the data is not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to *n* clusters that each contains a single object. Hierarchical clustering may be represented by a two-dimensional diagram known as a dendrogram, which illustrates the fusions or divisions made at each successive stage of analysis. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows to decide the level or scale of clustering that is most appropriate for the application.

Agglomerative techniques are more commonly used. To perform agglomerative hierarchical cluster analysis on a data set, the following procedure is required:

1) Find the similarity or dissimilarity between every pair of objects in the data set.
2) Group the objects into a binary, hierarchical cluster tree.
3) Determine where to cut the hierarchical tree into clusters.

The following is an example of a dendrogram.

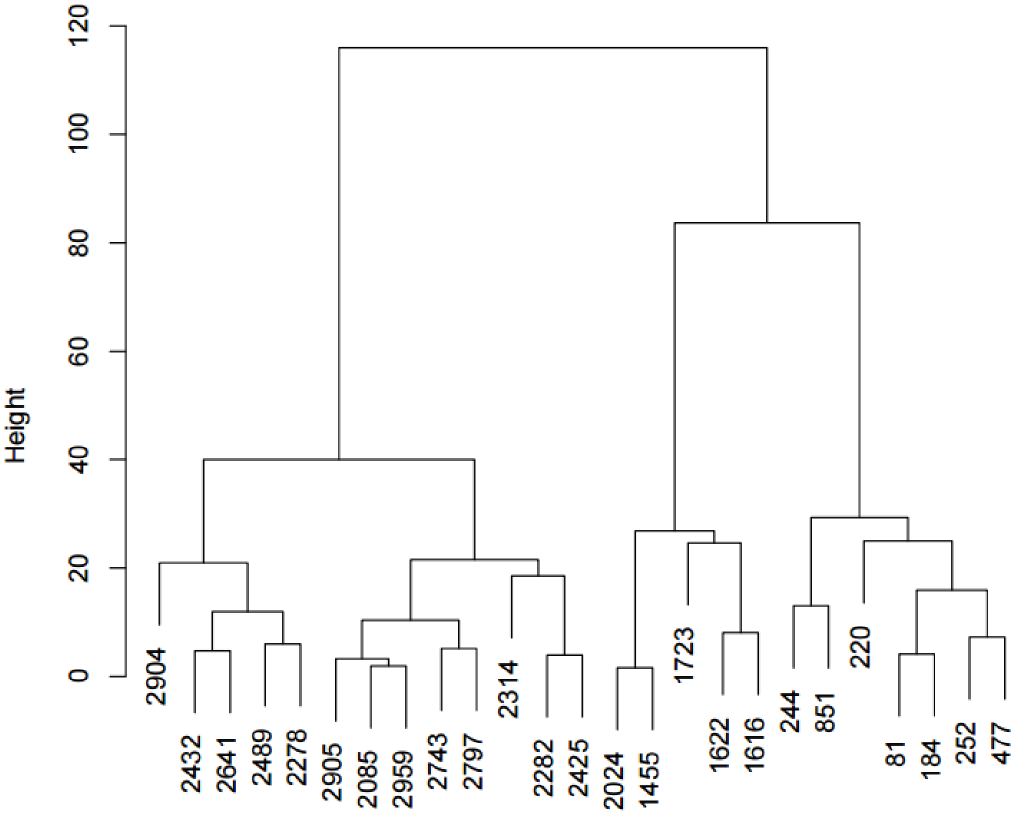


Figure 6 An example of cluster dendrogram. (Tibshirani et al., 2001)

To sum up, the advantages of hierarchical clustering are:

1) No apriori information about the number of clusters required.
2) Easy to implement and gives the best result in some cases.

However there are some disadvantages of this algorithm:

1) Algorithm can never undo what was done previously.
2) The time complexity of at least O(n^2log^n) is required, where 'n' is the number of data points.

3) Based on the type of distance matrix chosen for merging different algorithms can suffer from one or more of the following:
    - Sensitivity to noise and outliers
    - Breaking large clusters
    - Difficulty handling different sized clusters and convex shapes
4) No objective function is directly minimized
5) Sometimes it is difficult to identify the correct number of clusters by the dendrogram.

2.2.3 Spatiotemporal Data Visualization

Visual analytics is the science of analytical reasoning supported by interactive visual interfaces. A more specific definition would be: ‘Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets’ (Thomas, J., Cook, K., 2005). It is an integral approach combining visualization, human factors, and data analysis, which allow users to combine their knowledge with the automation data processing and analysis which done by computer to explore and gain more information from the data.

The visual analytics process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data. Different data sources need to be integrated first in the preprocessing steps (e.g. data cleaning, normalization, etc) before the analytics step could be executed. Then the analyst could apply automatic analysis methods using data mining techniques to generate models of the original data which then could be combined with visualization methods where the analyst interacts with visualization of the data model to define which data model could generate a better result based on certain parameters. The analyst could also do the visual analysis first to reveal an insightful of information which then could be used as a base to build the model to do the automatic analysis. Figure 7 shows the overview of visual analytics process.

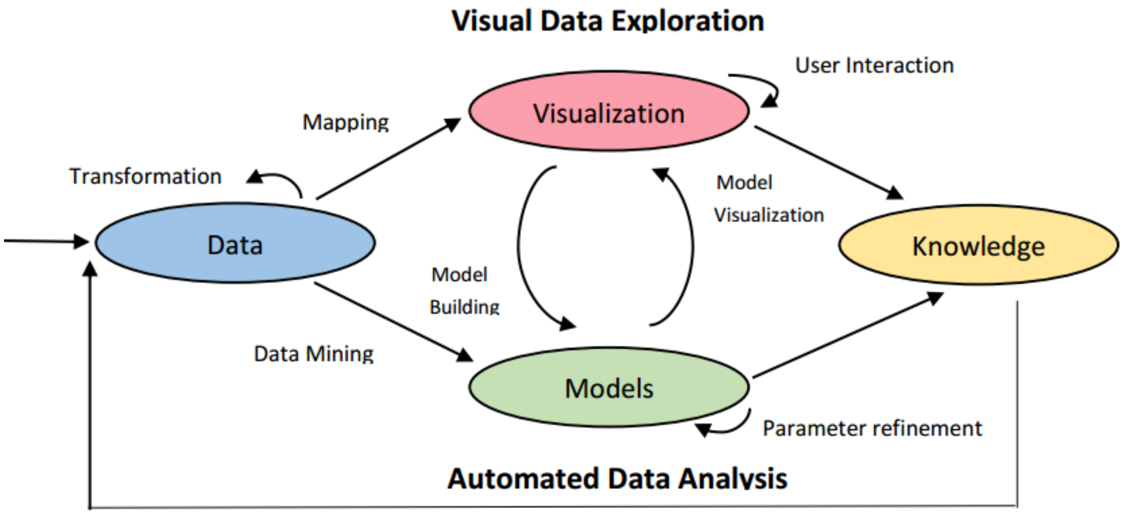


Figure 7 Feedback loop of visual analytics process
Source: *http://www.visual-analytics.eu/faq/*

Nowadays, the visual exploring and analyzing of big movement data becomes more and more important in the disciplines such as location-based services, urban planning, and transportation modeling, which deal with urban and road-based traffic data. The purpose for the visualization of movement data is mainly in understanding the data properties. Accordingly the visual analysis of movement data could be seen as the main component for detecting trends and patterns inside the data, which can hardly be done by simple statistical analysis.

Maps for presenting an overview of the spatial distribution of moving objects within a selected time interval are commonly used. The reason is that the moving objects, which are often shown in their raw form like points, are very numerous and consequently their positions have to be presented in an aggregated way (Andrienko N & Andrienko G, 2007). When dealing with traffic data, aggregation is needed for efficient understanding and analysis.

A. Dot Mapping

One common method to display FCD data pattern is dot mapping. A straightforward dot mapping method is to use one dot representing one data observation, such as the position of the taxi, which also includes some attributes that can be chosen to show different kinds of information like velocity or taxi status. This method is quite simple and accurate to represent the spatial distribution of data. However, the interpretation of spatial patterns and hotspots could be very difficult due to the clutter effect especially with a large amount of data.

A dot symbol can use visual variables like color to show the instantaneous velocity by the particular attribute value, which is associated for each movement position. This allows a quite accurate representation of these attribute values. It is a simple visualization technique, which enables to use the different coloration possibilities and varying amounts of velocity ranges. The values of the velocity attribute, which are listed in each point of the dot map, can be given for long time periods or for only short time windows. One problem for this kind of visualization is the appearance of overlapping when dealing with numerous points. An example for this kind of visualization based on FCD velocities delivers (Stanica R et al., 2013) with a dot map for the road network of Cologne, which is shown in Figure 8.

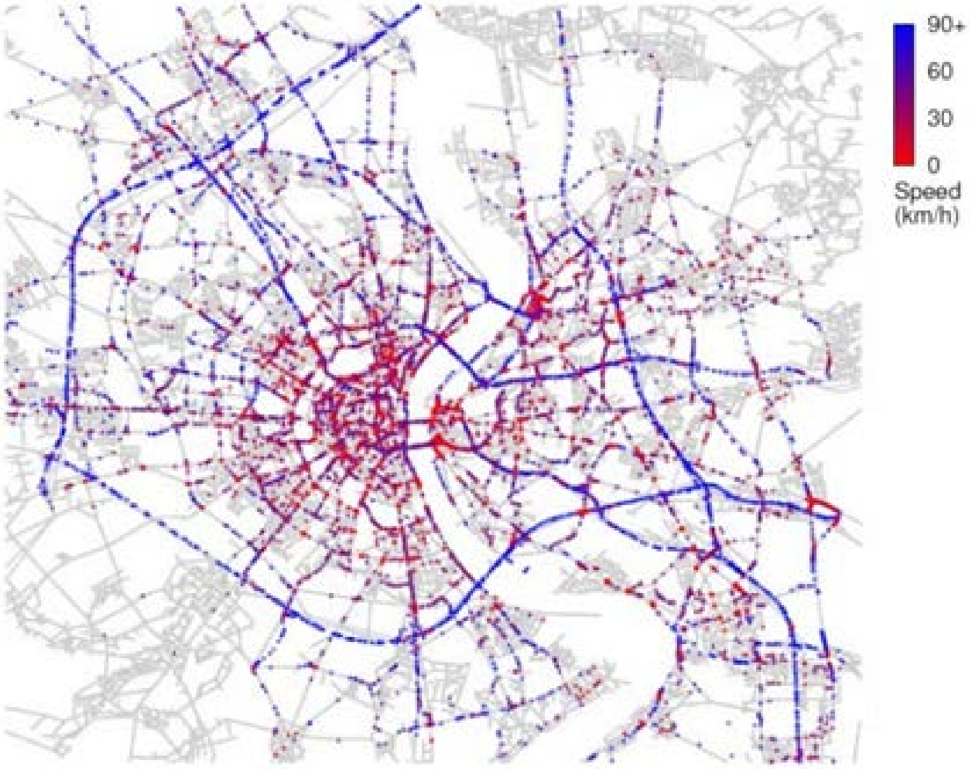


Figure 8 Dot map for the velocities of the road traffic in Cologne (Stanica R et al., 2013), modified

B. Three-dimensional (3D) Spatio-temporal Mapping

Another possibility for the representation of traffic flows is to visualize data in a 3D space. The use of a 3D view provides the analyst a better overview for certain quantitative values (Andrienko N & Andrienko G, 2007). This kind of visualization can logically make use of the height component of the elements inside a given geographic information data set, which itself were given by their certain 3D position. For some reasons, mainly due to the fact that the height component is often not very interesting for the inspection of the traffic situation, the 3rd dimension is used for the representation of other attributes. A frequently used example is the extrusion of lines or areas based on a number of points or average velocity. The extruded line or area appears as surface or volume with different extrusion height that represents specific values of a chosen attribute.

One example introduced by the Wasatch Front Regional Council (WFRC) (Grant M et al., 2011) shows a 3D map representation of transportation delay/congestion in the Salt Lake City region. Figure 9 shows the visualization results.

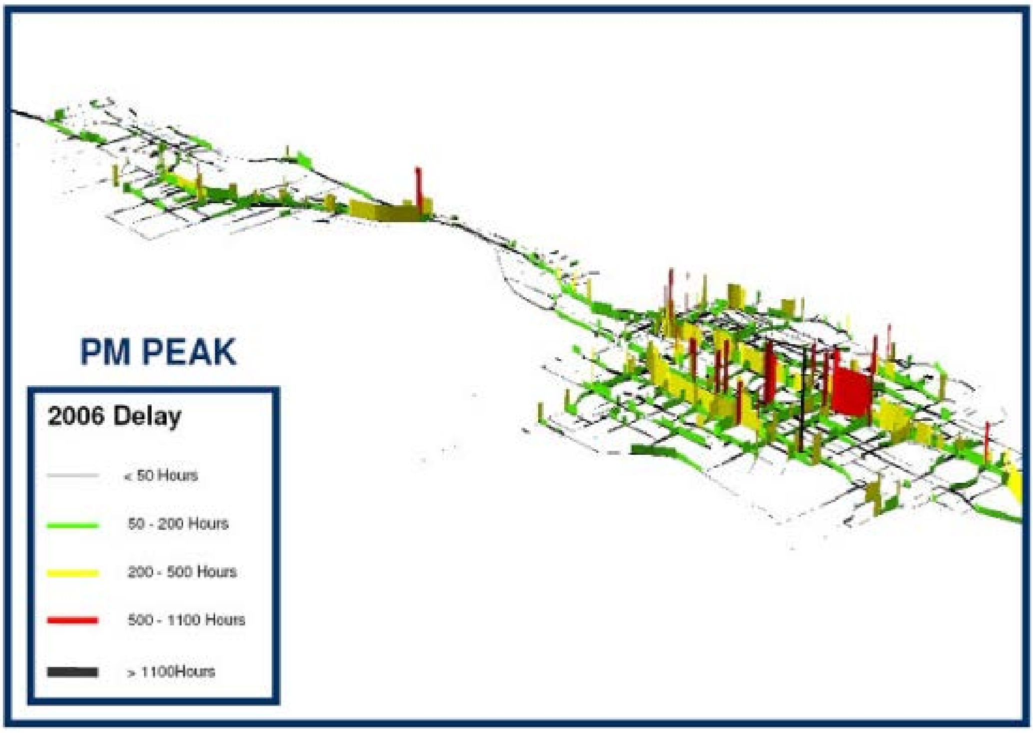


Figure 9 3D visualization of the transportation delay in the year 2006 of the Salt Lake City region, Wasatch Front Regional Council (WFRC) (Grant M et al., 2011)

Here the extrusion is based on line elements, which symbolize the centerline of a certain road. It uses a 3D graphic both in color and vertical height to display regional delay patterns which create an effective visual representation (Grant M. et al, 2011).

The 3D graphs are placed above the road network which symbolizes the level of delay of each road. The height and the different color give a deeper meaning of the information as users could easily distinguish the delay level and pattern based on the height.

C. Proportional Symbol Mapping

Proportional symbol maps scale the size of symbols (usually a circle or square) proportionally to the data values at the corresponding locations. They are a simple concept to grasp: The larger the symbol, the ‘more’ of something exists at a location. The investigation area can be divided into sub-regions. The division of the overall investigation area can be realized for a surface unit of measure or as a unique number of areas. The cells of the discrete grids are the base for the counting of geographical referenced raw data. An example for a movement aggregation is shown in Figure 10 (Andrienko N & Andrienko G, 2013).

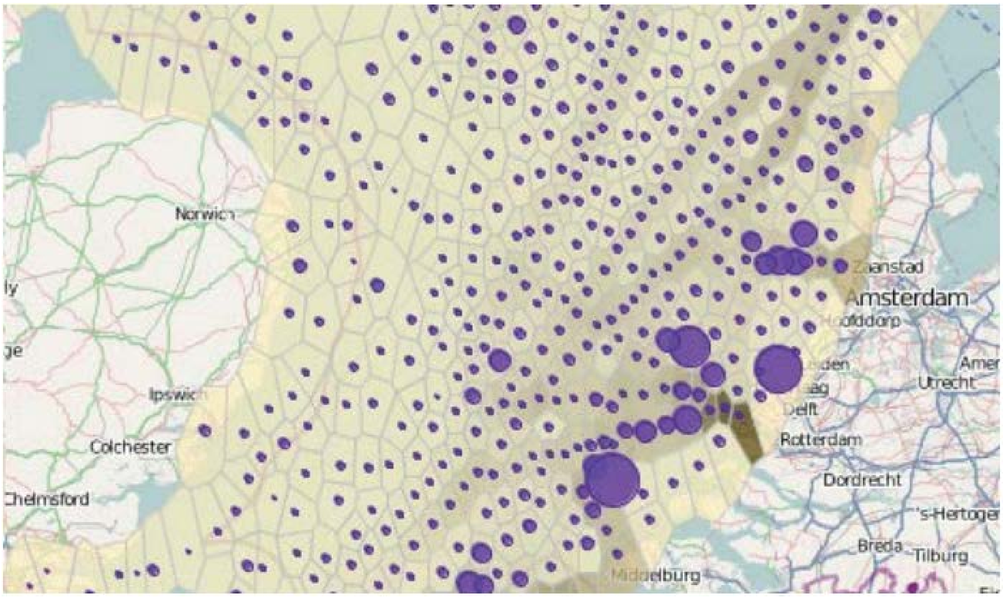


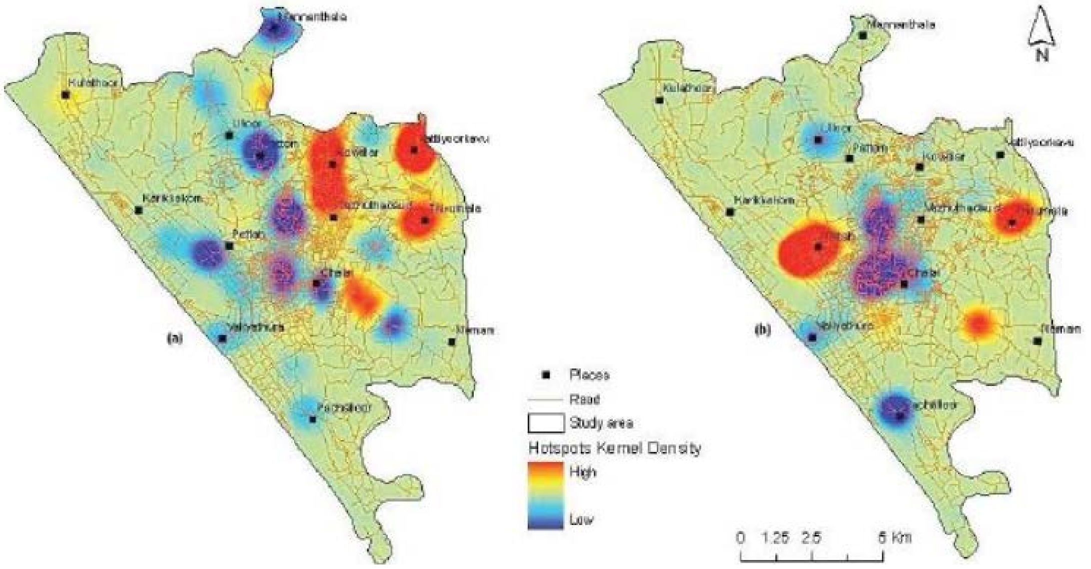
Figure 10 Movement aggregation with the use of irregular grid cells (Andrienko N & Andrienko G, 2013), modified

In this example, the different shading of the sub-regions can be identified, which correlates with the absolute amount of different visiting points of the investigated ships. These visiting points comply with the real time vessels of ships in the North Sea. The second information in this kind of visualization is the displaying of circles with different diameters, which show the mean duration of single ships lingers inside certain grid cells. The overall information, which emerges after the consideration of both elements, the cell shading and the circle inside each grid cell, can be distinguished as a correlation between the high density of ships visits inside the grid cells and the low duration of these visits. Consequently, the longer times of visits are spent in cells that are not intensively visited.

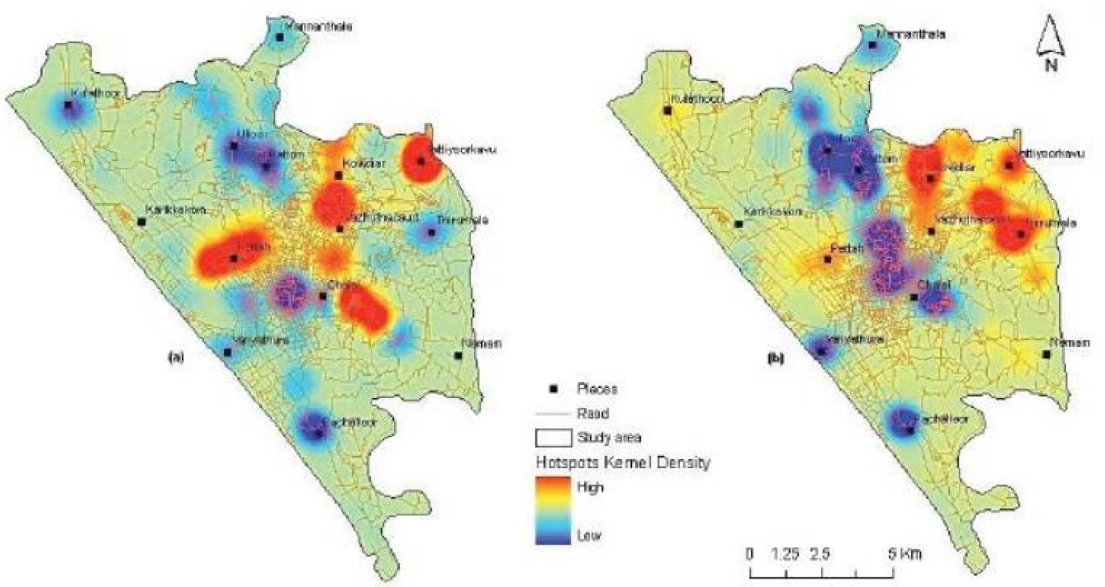
D. Density Mapping

Density mapping (usually known as ‘dot density mapping‘) is a simple yet highly effective way to show density differences in geographic distributions across a landscape. Kernel Density Estimation (KDE) is one of the popular mathematical methods that computes density by smoothing data points (Willems N et al., 2009). The objectives of this method are the reduction of sampling artifacts in the point visualization and additionally the visual determination of hot spots in the point data sets (Willems N et al., 2009). Figure 11 shows an example of density mapping.

In this example, Prasannakumar, et al (2011) evaluated road accident hot spots based on spatiotemporal clustering of road accidents. This research goal is to find patterns of localization and distribution of hotspots to determine whether spatial or temporal factors, such as the proximity to the school or the season, have influences to the road accidents. They are using Moran’s I method for spatial autocorrelation and Getis-Ord function and kernel density for clustering method.



(a) (b)



(c) (d)

Figure 11 Hotspots distribution of road accidents (Prasannakumar, et al., 2011) (a) monsoon period (b) non-monsoon period (c) educational institutions (d) religious places

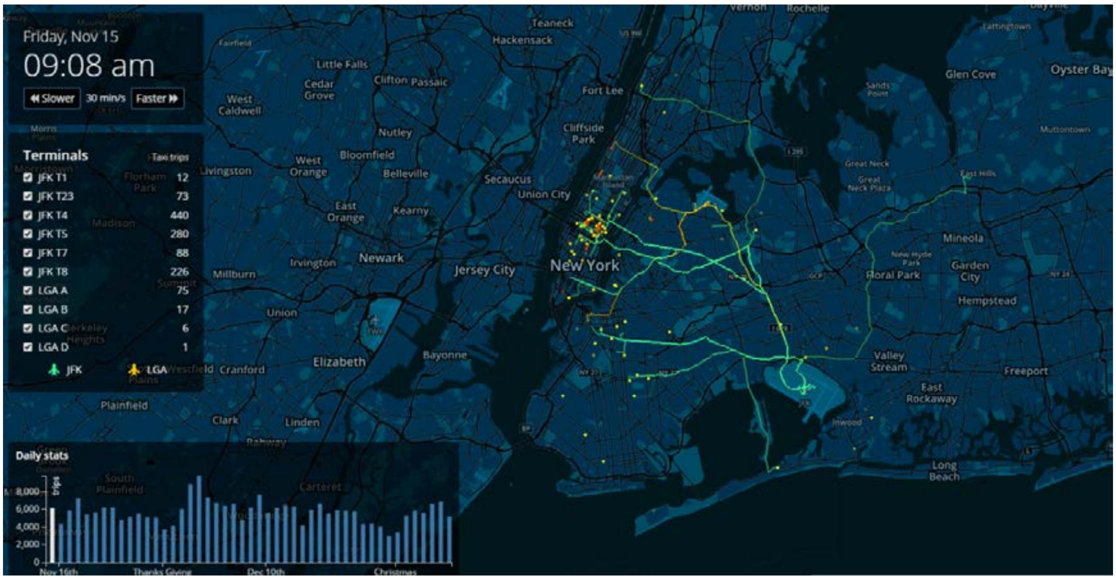
E. Interactive Mapping

Interactive mapping is normally developed to support visual exploration using interactive techniques, for instance, finding the change in some dimension by adding the temporal component and the statistics about the aggregates. Sometimes this kind of map can be described in the instantaneous state. Here we have several examples of interactive mapping.

Figure 12 is an animated ‘NYC Taxi Holiday Visualization System‘ which used 2013 NYC Taxi data to visualize traffic from JFK and LGA airports during the holiday season (Nov 15th to December 31st). Users can observe the traffic patterns and filter the visualization results by individual airlines or terminals. By the combination of the histogram bar on the right bottom, users could clearly see the statistics of the taxi data per day and easily select their interested date.

The map shows the trend that few people travel on the actual holidays, such as Thanksgiving or Christmas. But a few days later trips surge. Take December 2, 2013, the Monday after Thanksgiving as an example, there were nearly 10,000 taxi trips from

JFK and LaGuardia. It also shows that JFK's T4 and T5 have relatively more trips than other terminals.



Friday, Nov 15

09:08 am

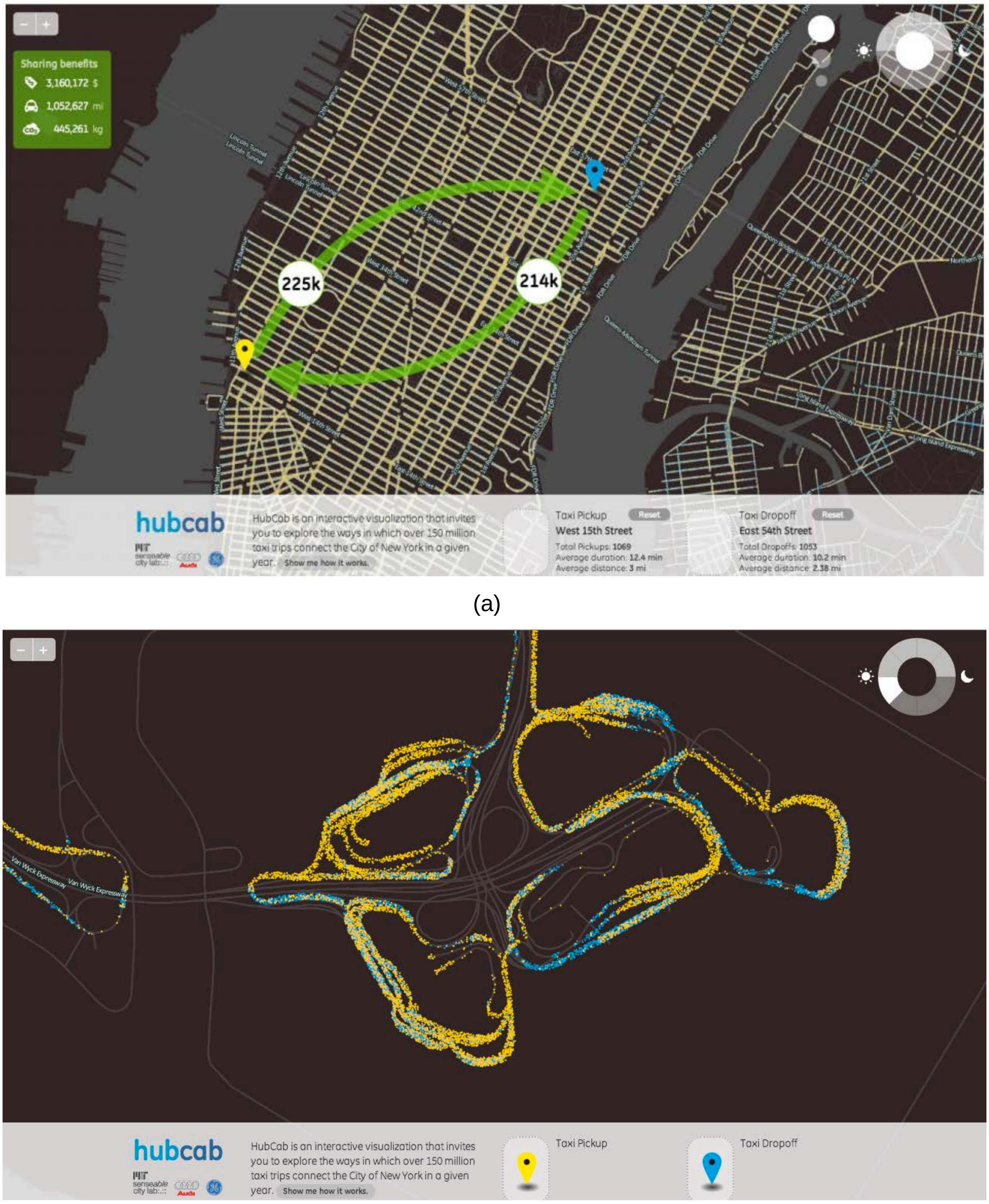
Terminals	Taxi trips
☑ JFK T1	12
☑ JFK T23	73
☑ JFK T4	440
☑ JFK T5	280
☑ JFK T7	88
☑ JFK T8	226
☑ LGA A	75
☑ LGA B	17
☑ LGA C	6
☑ LGA D	1

Figure 12 Screenshot of NYC Taxi Holiday from JFK and LGA airports

Source: *http://taxi.imagework.com*

Another example is about the HubCab system (shown in Figure 13) developed by MIT Senseable city lab. It allows users to get insight into the taxi mobility patterns at a fine granularity and supports future taxi sharing based on a model named 'shareability networks' (Santi et al. 2014, Szell and Groß 2014). HubCab is an interactive visualization that invites you to explore the ways in which over 170 million taxi trips connect the City of New York in a given year. In HubCab, they target taxicab services as a way to understand the linkages between human travel habits and the most visited places. This interface provides a unique insight into the movement mechanism of the city from the previously invisible perspective of the taxi system with a fine granularity.

HubCab allows to investigate exactly how and when taxis pick up or drop off individuals and to identify zones of condensed pick-up and drop-off activities. It allows users to navigate to the places where the taxi trips start and end and to discover how many other people in your area follow the same travel patterns. For example, the user can navigate to the places where his/her taxi trips start and end and to discover how many other people in his/her area follow the same travel patterns.



Sharing benefits
3,160,172 $
1,052,627 mi
445,261 kg

HubCab is an interactive visualization that invites you to explore the ways in which over 150 million taxi trips connect the City of New York in a given year.

Taxi Pickup
West 15th Street
Total Pickups: 1069
Average duration: 12.4 min
Average distance: 3 mi

Taxi Dropoff
East 54th Street
Total Dropoffs: 1053
Average duration: 10.2 min
Average distance: 2.38 mi

(a)

HubCab is an interactive visualization that invites you to explore the ways in which over 150 million taxi trips connect the City of New York in a given year.

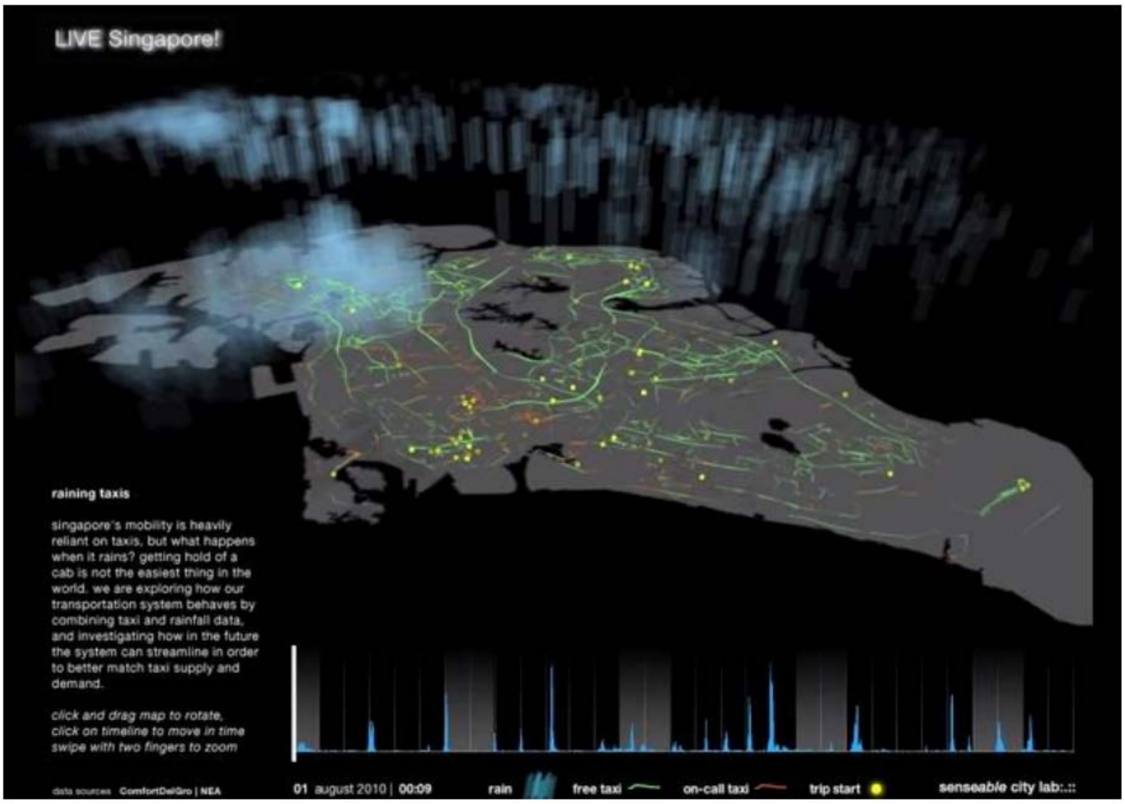
(b)

Figure 13 Screenshot of 'HubCab' (a) shareability networks (b) potential taxi sharing benefits between two locations in Manhattan

Source: *http://hubcab.org/#13.00/40.7219/-73.9484*

Figure 14 shows a more sophisticated animation of taxis distribution on rainy days in Singapore. The visualization aims to provide a deeper understanding of the city's dynamics. Singapore's mobility is heavily reliant on taxis. In this visualization, it is interesting to investigate the patterns when it rains. Researchers are exploring how our transportation system behaves by combining taxi and rainfall data, and investigating how in the future the system can streamline in order to better match taxi supply and demand.

This animation does not only show the taxis distribution on the road network but also presents a visualization of rain above the area of study to give more real feelings to the user. This animation does not only have a time slider function to show a different timeline of events but also a rotate function to make it easier to explore by the users.



LIVE Singapore!

raining taxis

singapore's mobility is heavily reliant on taxis, but what happens when it rains? getting hold of a cab is not the easiest thing in the world. we are exploring how our transportation system behaves by combining taxi and rainfall data, and investigating how in the future the system can streamline in order to better match taxi supply and demand.

click and drag map to rotate,
click on timeline to move in time
swipe with two fingers to zoom

data sources ComfortDelGro | NEA

01 august 2010 | 00:09

Figure 14 Animation of taxis distribution based on time series in raining day in Singapore
Source: *http://senseable.mit.edu/livesingapore/visualizations.html*

3. TEST DATASET

3.1 SHANGHAI FLOATING TAXI DATA PROPERTIES

Recall that in Section 2.1.1 that floating car data of the taxis are first sent to the taxi headquarter and then the data will be sent to the FCD-server of the institute. The cycle times of the positioning are limited by the bandwidth of the communication channel and vary between about 10 and 120 seconds, depending on the status of the individual taxi. The collected GPS positions are sent then with the cycle times of about 10 minutes to the server of the institute. In this thesis, the data was derived from the FTD system.

Floating taxi dataset with the description of attribute information is saved in the form of text file. Each line in text file represents one record of the data and the attribute in each column is separated by a comma. The overview of the general structure of a dataset is described in Table 2, which gives a listing of the available attributes that are contained in the FTD of Shanghai.

Fieldname	Field Value	Details
Date	20100517	8-digit number
Time	082235	6-digit number
Company Code	QS	Initials of the Pinyin of the company name
Car ID	11692	The unique ID of the car, in 5 digits
Longitude	121.61365	In degree; accurate to the 6th decimal place
Latitude	31.201005	In degree; accurate to the 6th decimal place
Instaneous Velocity	34.9	Accurate to 0.1 (km/h)
Instantaneous Altitude	255	Accurate to 1 (m)
Car Status	0	0 for empty; 1 otherwise
GPS Effectiveness	1	1 for effective; 0 otherwise

Table 2 Description of the data format of Shanghai FTD

Here is the detailed description of each attribute:

- **Date** is collected within 47 days from 10th May 2010 to 30th June 2010.

- **Time** records in every 10 seconds. For example ‘082235‘ stands for 08:22:35 am. Theoretically, around 8000 lines of data of each car would be recorded in one day (24 hours).
- **Company Code** is the initials of the company name and it keeps the same for all the data.
- **Car ID** is the identification of each taxi which varied from 10003 to 18991. In total, there are over 2,000 taxis in Shanghai which are recorded.
- **Longitude** and **latitude** values give a precise position of each point with an accuracy to the 6th decimal place so that each point could be visualized geographically by using the WGS84 coordinate system.
- **Car Status** gives an answer about if the vehicle is currently empty or occupied by a passenger.
- **GPS effectiveness** gives the status of an effectively working or not working GPS device inside a vehicle.

In this thesis, since we focus on studying the traffic flows of taxis from or to the certain transport hub. Therefore, the most important information are the ‘Car ID‘, the ‘longitude‘ and ‘latitude‘ position of the taxi, and ‘Car Status‘. The position of the vehicles could be used to calculate the flow rate or volume for each road section, and the vehicle ID is needed to distinguish the vehicles. The status of the car shows if the taxi is occupied with passengers or not.

3.2 REGION OF INTEREST (ROI)

3.2.1 Geographical, Social-Economic Characteristics and Transportation System of Shanghai Hongqiao International Airport

Shanghai Hongqiao International Airport is the main domestic airport with limited international flights. It is located near the town of Hongqiao in Changning District, 13 kilometers (about 8 miles) west of the downtown area and about 60 kilometers (about 37 miles) from Pudong International Airport. It is closer to the city center than Pudong Airport, the Shanghai's main international airport.

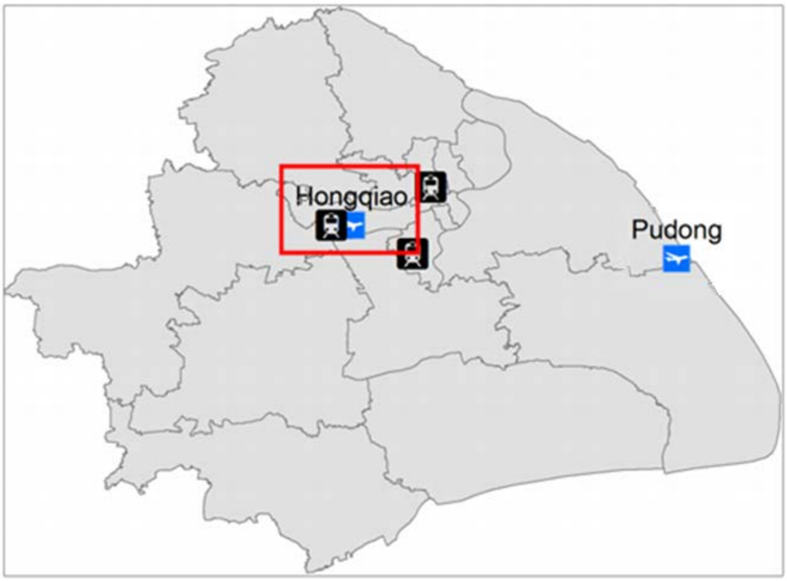


Figure 15 The location of Hongqiao airport in Shanghai

Being the first civilian airport in Shanghai, Hongqiao International Airport is more than eighty years old. After a series of renovations, it has become one of the three international air transit centers in China. 91 airlines currently fly here to both domestic and international cities. In 2014, Hongqiao Airport handled 37,960,200 passengers, a growth of 21.3% over 2010, making it the 4th busiest airport in mainland China and the 21st busiest in the world. The airport was also mainland China's 5th busiest airport in terms of cargo traffic and the 7th busiest by traffic movements. By the end of 2014 Hongqiao Airport hosted 22 airlines serving 82 scheduled passenger destinations.

Hongqiao hub is a mega-world-class integrated transportation hub in the ‘Eleventh Five-Year Plan’ period of China. It has gradually formed a multi-transport hub. When Shanghai Hongqiao comprehensive transport Hub project is completed, it sets railway, Pudong Maglev and Huhang Maglev, airport, bus, subway, long-distance bus terminal and taxi in one. In other words, Hongqiao transport hub will be a center for all modes of transportation.

Figure 16 A picture of Hongqiao international airport

However, during the process of planning, designing, and constructing, the hub ontology and surroundings always face the different transport challenges. Pertinent traffic impact evaluation and transport research have been developed for many times and promoted continuous improvements of designs on the hub planning.

3.2.2 Extraction of Hongqiao Airport

In this work, we chose Shanghai Hongqiao international airport as the main study transport hub. As we can see from Figure 17, we first established the range of our ROI is from 31.174907°N to 31.222126°N and from 121.317655°E to 121.346007°E. Therefore, even if the FTD sets we received is huge, we just need to extract the taxi trajectories which starts from or ends at the Hongqiao Airport area.

In addition, the two possible taxi loading/unloading areas in Hongqiao airport are colored on Figure 17 which are located in Terminal 1 (31°11′50″N, 121°20′32″E) and Terminal 2 (31°11′46″N, 121°19′18″E) respectively. Terminal 2 is four times the size of

Terminal 1 and houses 90 percent of all airlines at the airport. Terminal 1 is now used only for international flights and Spring Airlines.

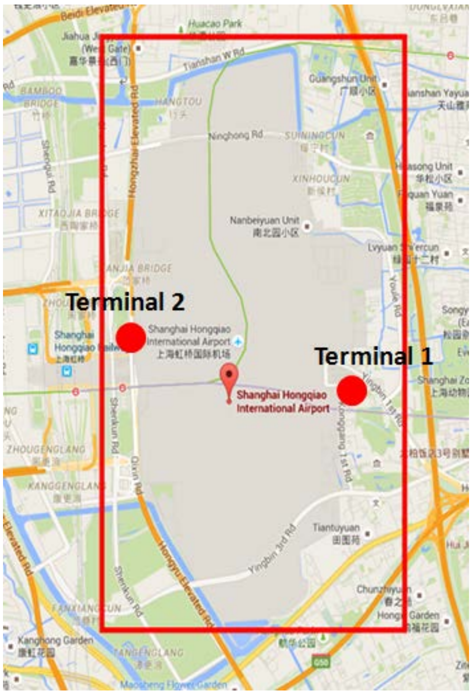


Figure 17 Shanghai Hongqiao airport area as study transport hub

4. METHODOLOGICAL FRAMEWORK

In this chapter, we describe the methodological framework and workflow for visual analysis of floating taxi data related to the Hongqiao airport. In Section 4.1, we propose a visual analysis framework to undertake the data processing, analyzing and visualizing procedures. Section 4.2 introduces the preprocessing procedure of Shanghai floating taxi data. In Section 4.3 we analyze the FTD in temporal dimension and Section 4.4 emphasizes a spatial analysis. In Section 4.5 we aggregate the statistical results from Section 4.4 and Section 4.5 for the spatiotemporal visualization of the FTD.

4.1 THE VISUAL ANALYSIS FRAMEWORK

With the processing of the FTD set of Shanghai, a wide range of useful information can be derived. In the beginning, the data should be generally explored by its correctness and accuracy. Then the given FTD set should be processed in a reasonable way without unnecessary long calculation times. After the preprocessing of the raw data, the following steps aim to convert them into a geographic representation.

In addition to the preprocessing of the raw FTD set, there is also a question to be solved: which kind of spatial clustering methods and visualization techniques described in section 2.2.2 and 2.2.3 are suitable for the visual detection of pick-up and drop-off events from transportation hubs? For the results of the visual representation of ‘hot spot‘, the processing of FTD data generally should be considered with the provided data mining theory in section 2.2.1. According to these approaches, the use of information from FTD can be derived more reasonable.

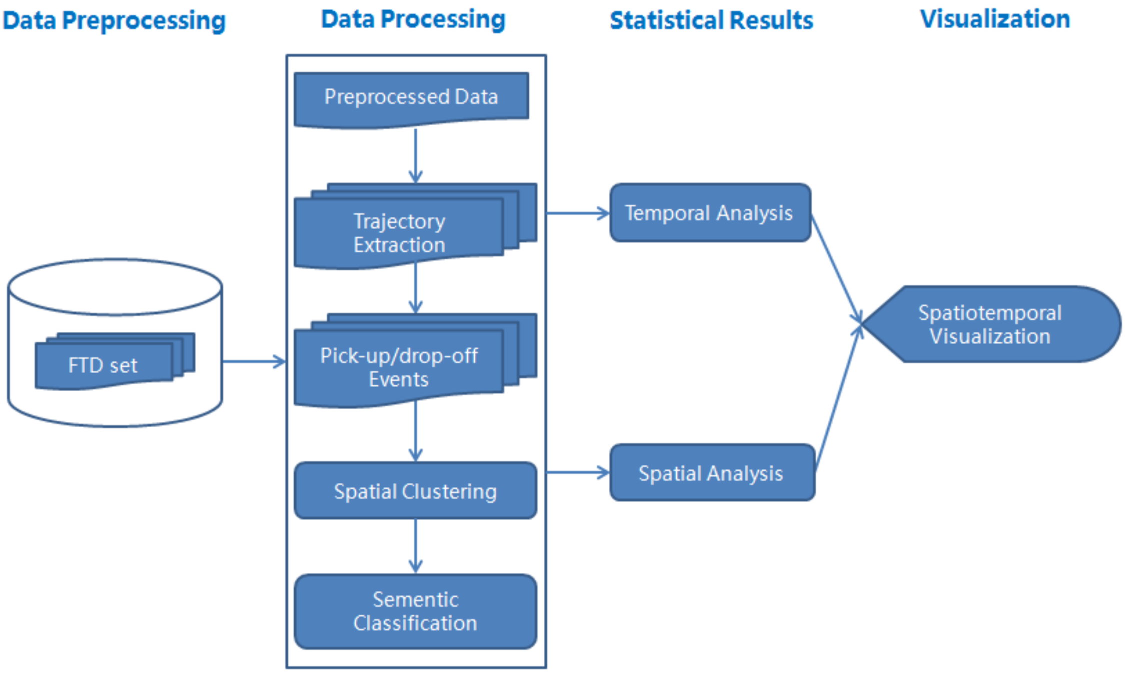


Figure 18 The framework of the visual analysis of traffic flow of transport hubs.

We propose a general framework and workflow as shown in Figure 18 to visually explore the mobility patterns of the flow in and out of the transport hubs. The framework consists of four components

The first component mainly focuses on raw FTD data preprocessing, which includes error elimination, data selection and data partition. We also show the temporal distribution of the data partitions.

In the second component, based on the preprocessed data, we reconstruct occupied trajectories from or to the transport hubs, derive the pick-up or drop-off events, and cluster the pick-up and drop-off events. Spatial clustering methods are applied to identify the dense areas or significant places from the extracted pick-up or drop-off events. These places are represented by the convex hull polygons of the clusters. Furthermore, we classify the significant places into different categories of semantic meanings.

In the third and fourth components, statistical analysis and appropriate visualization methods are proposed to enable the exploration and support the analysis of the spatiotemporal and semantic patterns.

The following sections describe in detail the individual components.

4.2 PREPROCESSING OF SHANGHAI FTD

The size of Floating Taxi Dataset is normally very large due to its high updating frequency (every 10 seconds in our case). Therefore, before the data could be used for further analysis of traffic flow of transport hubs, several steps must be done to simplify the data. As in Figure 19, this thesis undertakes the following data preprocessing steps by using Python source code.

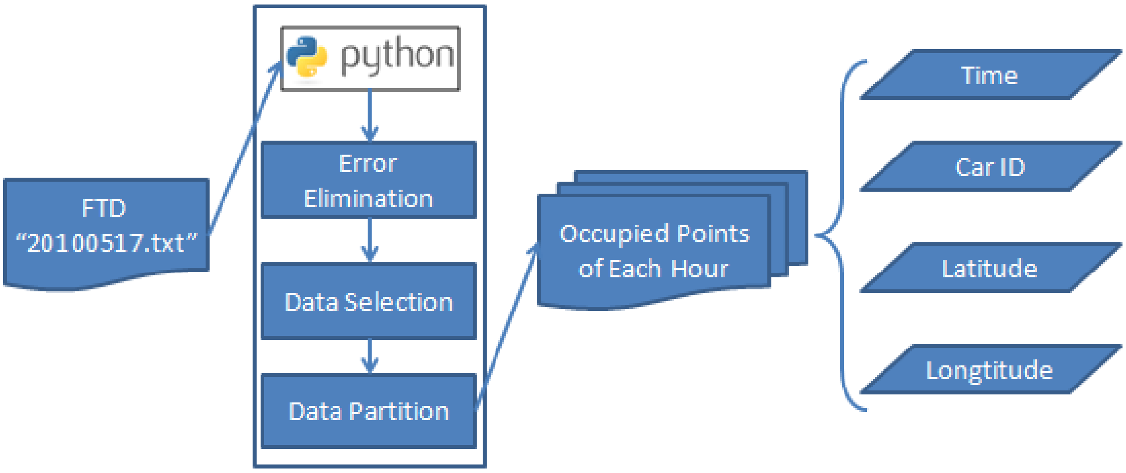


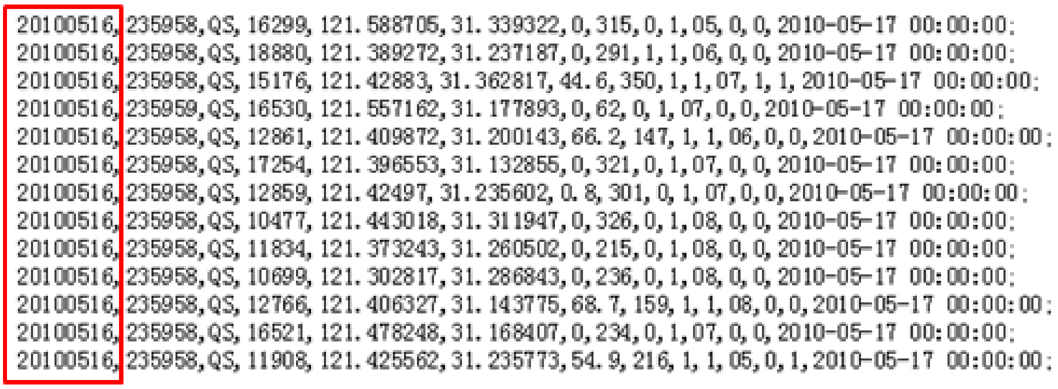
Figure 19 The framework of preprocessing of initial test dataset.

4.2.1 Error Elimination

Due to the slight instability of the sensor device on the floating car – taxi and the disturbance of surrounding environment, the raw data, with no doubt, contains certain errors which need to be eliminated.

First of all, we chose FTD '20100517.txt' (17th of May 2010) as our initial test data. Then we use Python code to filter the data. In general, there are 3 kinds of errors as the following:

- The value of every column must be in the right format. For example, all columns except 'Company Name' must be number.
- The value of the first column 'Date' must be in 8 digits and equal to '20100517'. See error example of Figure 20.



```
20100516,235958,QS,16299,121.588705,31.339322,0,315,0,1,05,0,0,2010-05-17 00:00:00;
20100516,235958,QS,18880,121.389272,31.237187,0,291,1,1,06,0,0,2010-05-17 00:00:00;
20100516,235958,QS,15176,121.42883,31.362817,44.6,350,1,1,07,1,1,2010-05-17 00:00:00;
20100516,235959,QS,16530,121.557162,31.177893,0,62,0,1,07,0,0,2010-05-17 00:00:00;
20100516,235958,QS,12861,121.409872,31.200143,66.2,147,1,1,06,0,0,2010-05-17 00:00:00;
20100516,235958,QS,17254,121.396553,31.132855,0,321,0,1,07,0,0,2010-05-17 00:00:00;
20100516,235958,QS,12859,121.42497,31.235602,0.8,301,0,1,07,0,0,2010-05-17 00:00:00;
20100516,235958,QS,10477,121.443018,31.311947,0,326,0,1,08,0,0,2010-05-17 00:00:00;
20100516,235958,QS,11834,121.373243,31.260502,0,215,0,1,08,0,0,2010-05-17 00:00:00;
20100516,235958,QS,10699,121.302817,31.286843,0,236,0,1,08,0,0,2010-05-17 00:00:00;
20100516,235958,QS,12766,121.406327,31.143775,68.7,159,1,1,08,0,0,2010-05-17 00:00:00;
20100516,235958,QS,16521,121.478248,31.168407,0,234,0,1,07,0,0,2010-05-17 00:00:00;
20100516,235958,QS,11908,121.425562,31.235773,54.9,216,1,1,05,0,1,2010-05-17 00:00:00;
```

Figure 20 Example of data errors of the ‘Date‘ attribute

- The ‘Longitude‘ and ‘Latitude‘ value of each point must be inside of Shanghai bounding box, which we set as 120°E - 122°E and 30°N - 32°N.

4.2.2 Data Selection

After eliminating the errors, the correctness and accuracy of our initial test dataset get higher. In order to extract the pick-up and drop-off points of each taxi trajectory, the next step is to select all the study-related information from the filtered FTD.

Here is our data selection process:

1) As we mentioned in Section 3.1, the most relevant information in FTD is ‘Time‘, ‘Car ID‘, the ‘Longitude‘ and ‘Latitude‘ position of the taxi, and ‘Car Status‘. Thus, we only need to select these 5 columns from our dataset, which will considerably reduce the data volume for further analysis.

2) Since we chose Shanghai Hongqiao airport as our study transport hub and we focus on the traffic flow of taxis, FTD with occupied status starting from or ending to this airport should be extracted. The binary information ‘0‘ and ‘1‘ of ‘Car Status‘ can give a logical separation of attended real time trips of the taxi. That means, the value of ‘Car Status‘ must be 1.

4.2.3 Data Partition

After we acquired all the occupied points, the basic data preprocessing by using Python source code is almost done. However, in order to simplify the work for the further steps and to easily explore the temporal pattern of traffic flow, we divide the data into small parts of one hour acquisition time.

The histogram in Figure 21 indicates the size of data partitions in each hour of a day. We can easily see that the peak hour of the traffic flow of occupied taxis comes in the morning between 6 to 10 am. Nevertheless, from 10 am to 13 pm the taxis are relatively not so occupied.

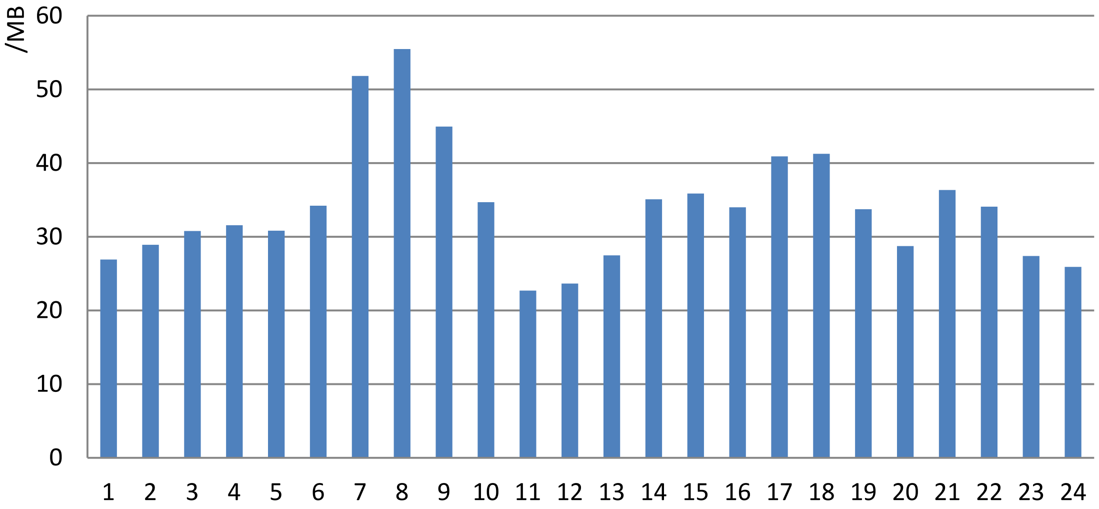


Figure 21 The histogram of the size of filtered data in each hour

For further data process and calculation, we use MATLAB to calculate the taxi statistics of extracted data and their geographical distribution. MATLAB is a high-level technical language and interactive environment which is used by millions of developers and scientists worldwide. In this environment, the user can develop and represent their ideas and also work across disciplines. MATLAB allows matrix manipulations, data visualization, data analysis, implementation of algorithms, plotting of functions and data, the creation of user interfaces, etc. In this thesis, we use MATLAB to further process the filtered data to fulfill our expectation.

4.3 TEMPORAL ANALYSIS OF SHANGHAI FTD

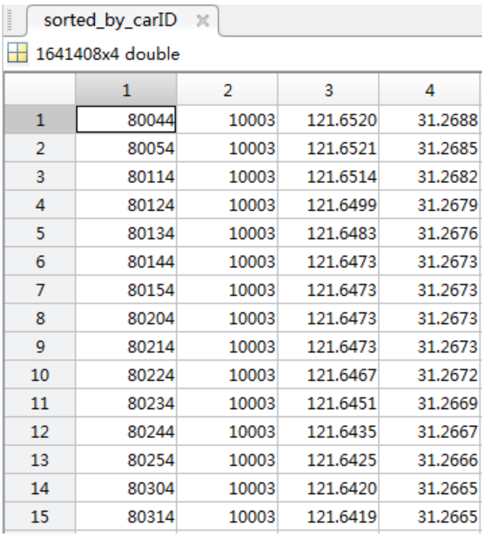
In this section, we focus on the statistical analysis of temporal patterns of the data partitions.

4.3.1 Reconstruction of Taxi’s Trajectories

A first idea of representing movement is to reconstruct the trajectories for each individual taxi. Generation of a trajectory is often based on point position data, which is acquired in a certain time interval that can vary between seconds or hours. The starting and ending point of the trajectories can be defined by introducing a certain investigated time period or by an acquired attribute value, which can specify a state of changing for a special traffic event. Pickup and drop-off points correspond to the starting and ending points of the occupied trips respectively. Here the resulting trajectories are based on the taxi‘s identification number. This kind of trajectories may correspond with the real time taxi rides in the city. A connection in another way can be additionally provided by the temporal and spatial similarity between points and supplemental element types.

Here we take the hourly data from 8 to 9 am as an example. After loading them to MATLAB, the matrix with 1,641,408 rows and 4 columns of data, namely 1,641,408 GPS points of taxis has been stored. In addition, we could see the car identification number varies from 10003 to 18970 and during 8 – 9 am there are 6, 817 occupied taxis in total have been recorded. Therefore in order to generate taxi trajectories from and to Hongqiao airport, the following steps are needed to proceed.

1) Sorting the matrix by each ‘Car ID‘ based on its recording ‘Time‘. See Figure 22.



sorted_by_carID

1641408x4 double

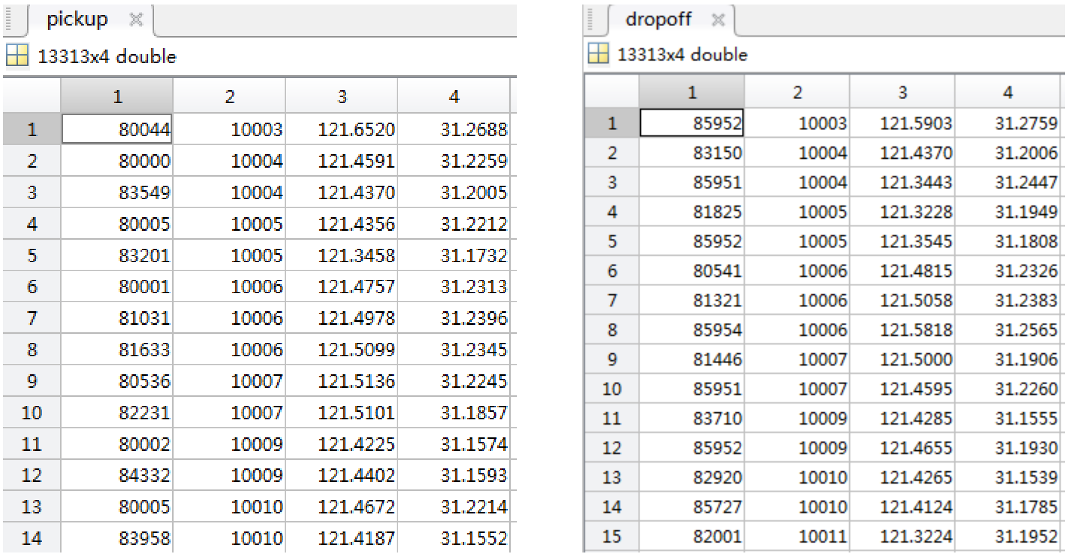
	1	2	3	4
1	80044	10003	121.6520	31.2688
2	80054	10003	121.6521	31.2685
3	80114	10003	121.6514	31.2682
4	80124	10003	121.6499	31.2679
5	80134	10003	121.6483	31.2676
6	80144	10003	121.6473	31.2673
7	80154	10003	121.6473	31.2673
8	80204	10003	121.6473	31.2673
9	80214	10003	121.6473	31.2673
10	80224	10003	121.6467	31.2672
11	80234	10003	121.6451	31.2669
12	80244	10003	121.6435	31.2667
13	80254	10003	121.6425	31.2666
14	80304	10003	121.6420	31.2665
15	80314	10003	121.6419	31.2665

Figure 22 Screenshot of the sorted result in MATLAB

2) Set 3 minutes as the time interval of each two trajectories.

3) Store each trajectory to a new matrix. Altogether 13,313 trajectories have been calculated.

4) Extract all the pick-up points (start point of each trajectory) and drop-off points (end point of each trajectory) from the trajectory matrix to the two new matrices ‘pickup‘ and ‘dropoff‘. See figure 23.



pickup

13313x4 double

	1	2	3	4
1	80044	10003	121.6520	31.2688
2	80000	10004	121.4591	31.2259
3	83549	10004	121.4370	31.2005
4	80005	10005	121.4356	31.2212
5	83201	10005	121.3458	31.1732
6	80001	10006	121.4757	31.2313
7	81031	10006	121.4978	31.2396
8	81633	10006	121.5099	31.2345
9	80536	10007	121.5136	31.2245
10	82231	10007	121.5101	31.1857
11	80002	10009	121.4225	31.1574
12	84332	10009	121.4402	31.1593
13	80005	10010	121.4672	31.2214
14	83958	10010	121.4187	31.1552

dropoff

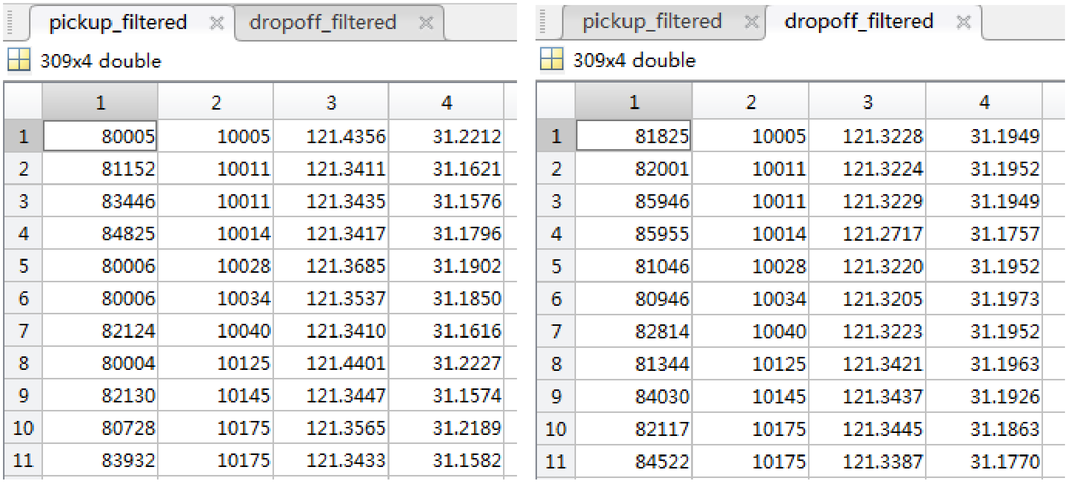
13313x4 double

	1	2	3	4
1	85952	10003	121.5903	31.2759
2	83150	10004	121.4370	31.2006
3	85951	10004	121.3443	31.2447
4	81825	10005	121.3228	31.1949
5	85952	10005	121.3545	31.1808
6	80541	10006	121.4815	31.2326
7	81321	10006	121.5058	31.2383
8	85954	10006	121.5818	31.2565
9	81446	10007	121.5000	31.1906
10	85951	10007	121.4595	31.2260
11	83710	10009	121.4285	31.1555
12	85952	10009	121.4655	31.1930
13	82920	10010	121.4265	31.1539
14	85727	10010	121.4124	31.1785
15	82001	10011	121.3224	31.1952

Figure 23 Screenshots of the ‘pick-up‘ and ‘drop-off‘ matrices in MATLAB

5) Find out all the pick-up points and drop-off points which are inside the bounding box of Hongqiao Airport area: 31.174907°N - 31.222126°N and 121.317655°E - 121.346007°E. Store them to the two new matrices called ‘pickup_filtered‘ and

‘dropoff_filtered‘. Figure 24 indicates that in total there are 309 taxi trajectories out of 13,313 trajectories which head to the airport or leave from the airport.

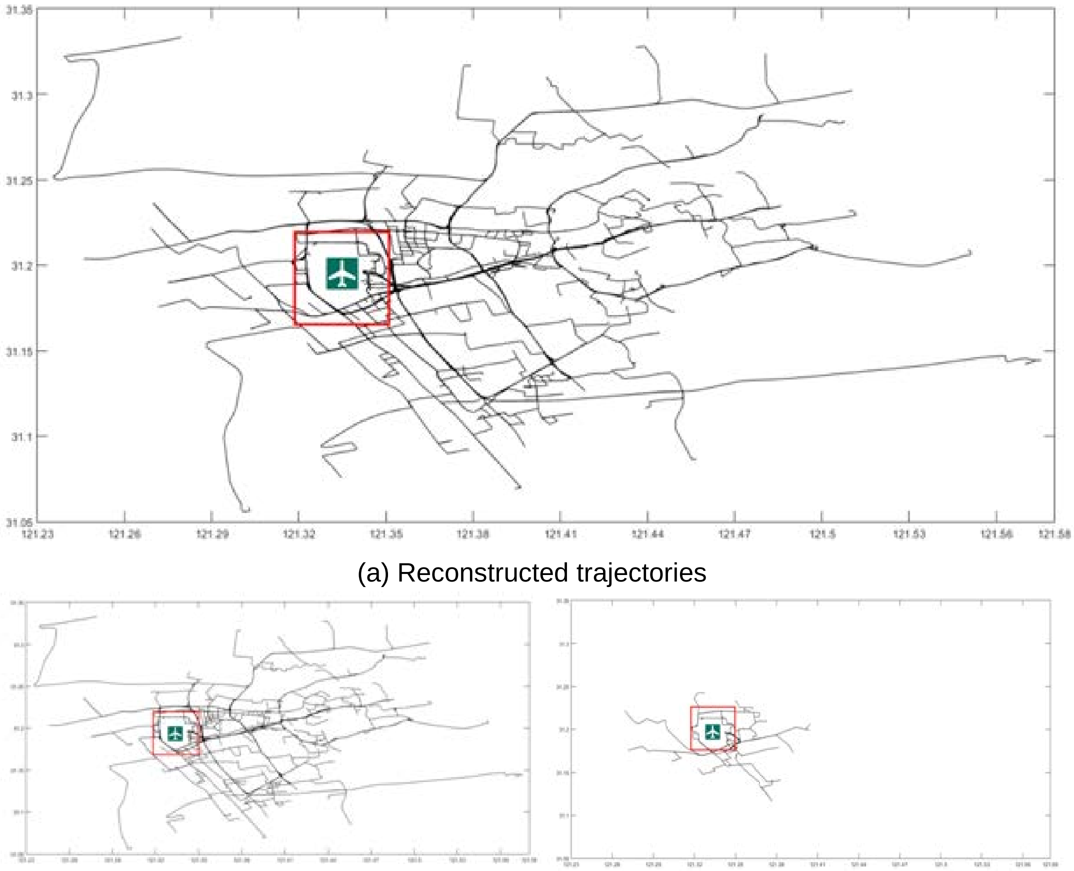


	1	2	3	4
1	80005	10005	121.4356	31.2212
2	81152	10011	121.3411	31.1621
3	83446	10011	121.3435	31.1576
4	84825	10014	121.3417	31.1796
5	80006	10028	121.3685	31.1902
6	80006	10034	121.3537	31.1850
7	82124	10040	121.3410	31.1616
8	80004	10125	121.4401	31.2227
9	82130	10145	121.3447	31.1574
10	80728	10175	121.3565	31.2189
11	83932	10175	121.3433	31.1582

	1	2	3	4
1	81825	10005	121.3228	31.1949
2	82001	10011	121.3224	31.1952
3	85946	10011	121.3229	31.1949
4	85955	10014	121.2717	31.1757
5	81046	10028	121.3220	31.1952
6	80946	10034	121.3205	31.1973
7	82814	10040	121.3223	31.1952
8	81344	10125	121.3421	31.1963
9	84030	10145	121.3437	31.1926
10	82117	10175	121.3445	31.1863
11	84522	10175	121.3387	31.1770

Figure 24 Screenshots of the ‘pickup_filtered‘ and ‘dropoff_filtered‘ matrixes in MATLAB

6) Plot each trajectory. See Figure 25, the red box shows the location of Hongqiao airport.



(a) Reconstructed trajectories

(b) Trajectories to airport

(c) Trajectories from airport

Figure 25 Trajectories of each occupied taxi from 8 – 9 am from/to airport

Each taxi trajectory consists of a series of GPS points which can reflect the road networks of the surrounding area of Hongqiao airport in Shanghai. From the above-generated trajectories, we could see that during the rush hour in the morning, a lot more people take taxis to the airport than those from the airport.

4.3.2 Statistics of Pick-up and Drop-off Events

We use the same data processing method in section 4.3.1 to process data of the other 23 hours. The histogram of the hourly data distribution is shown in Figure 26. In this histogram, the proportion of trajectories from and to the airport in each hour is distinguished by the color of red and blue. We could clearly see that the number of trajectories reaches its peak from 6 am to 8 am in the morning and during this period the number of trajectories to the airport is con

siderably larger than the trajectories from the airport. However, from 18 pm to midnight, trajectories from the airport are dominant.

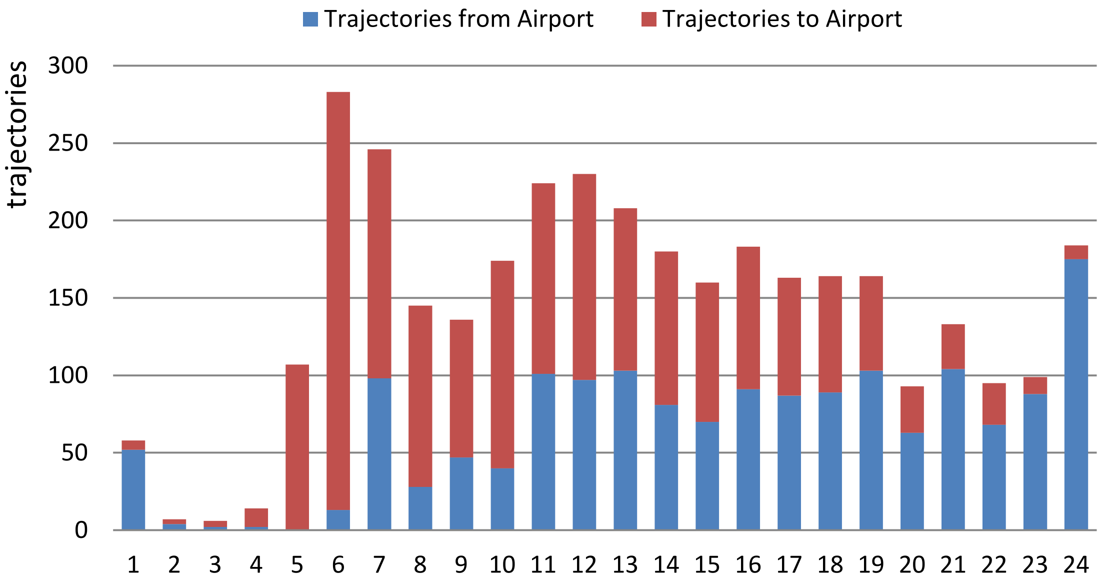


Figure 26 Hourly occupied trajectories from/to the airport

In order to make our data analysis more typical and representative, we divide the 24 hours data into 4 time groups: 00.00 – 06.00, 06.00 – 12.00, 12.00 – 18.00, 18.00 – 24.00. After the calculation, we could see in Figure 27 the temporal distributions of the amount of trajectories from and to the airport. More specifically, on 17th of May 2010 there are 54% (in total 1,850) occupied taxi trajectories to the airport and around 48% of them are in the morning from 6 am to 12 pm. While there are 46% (in total 1,606) occupied taxi trajectories from the airport to the city and 65% of them are from the afternoon to midnight (12 pm to 24 pm).

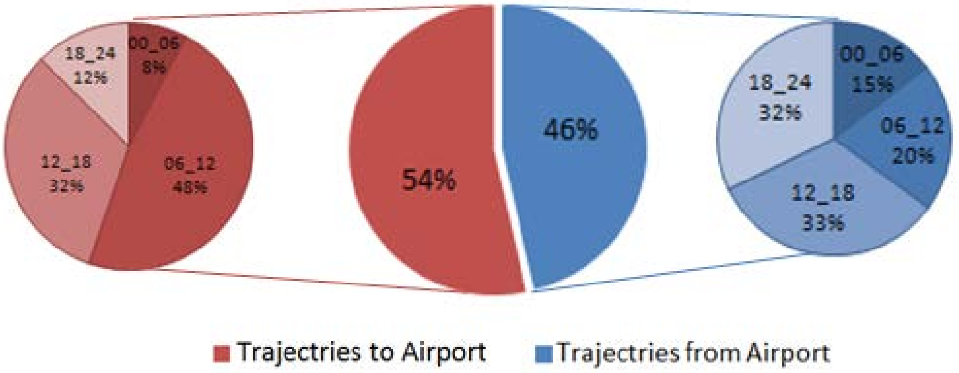


Figure 27 Statistics of the occupied taxi trajectories on 17th May

Moreover, we can ignore the geometric shapes of each trajectory and only focus on the pick-up and drop-off events. Therefore, we filtered the starting and ending point of each trajectory. Furthermore, we removed all the pick-up and drop-off points inside the bounding box of the airport and only keep the pick-up and drop-off points related to the airport. Then we represent those pick-up points in red dots and drop-off points in blue dots. Figure 28 illustrates the distribution of the pick-up and drop-off events of these trajectories.

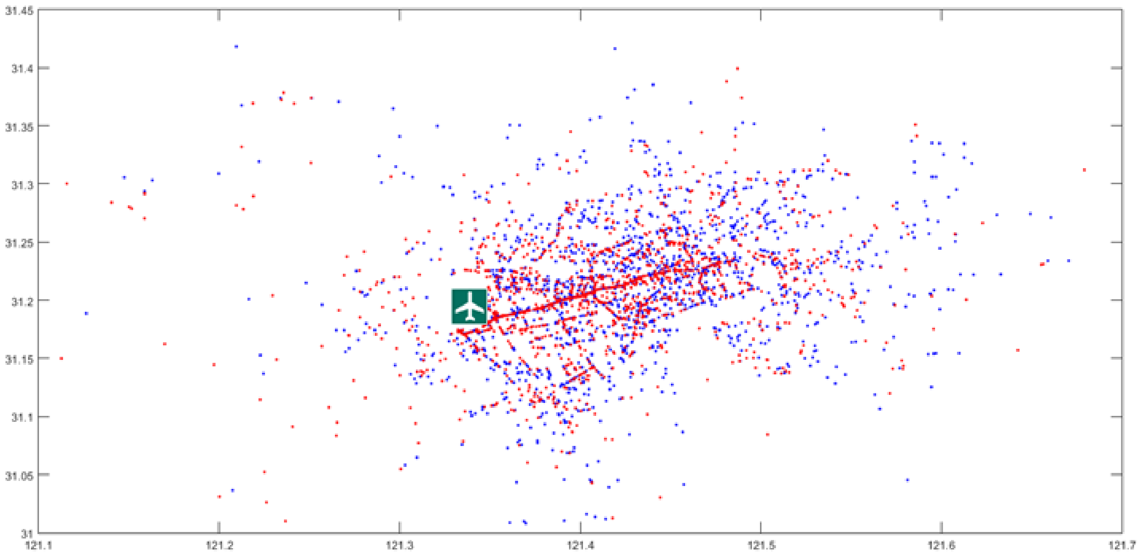
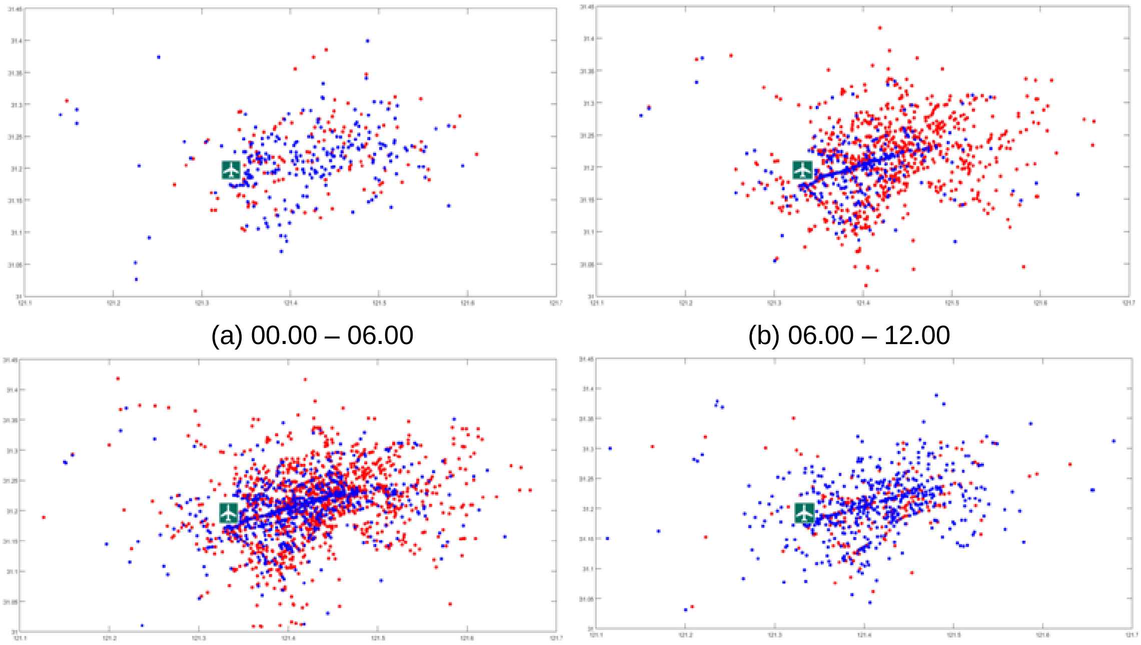


Figure 28 The corresponding pick-up (in red) and drop-off (in blue) points

However, it is really difficult to interpret any effective information from Figure 28 due to the cluttering effect. Therefore, for better analysis of the temporal pattern of these pick-up and drop-off points, we created 4 graphs in 4 different time periods as in Figure 29.



(a) 00.00 – 06.00

(b) 06.00 – 12.00

(c) 12.00 – 18.00

(d) 18.00 – 24.00

Figure 29 Pick-up (in red) and drop-off (in blue) points

Firstly we could see that Figure 29(a) and Figure 29(d), namely, during 18 pm to 6 am, the number of all the pick-up and drop-off events is relatively small. Besides the number of the drop-off events is much more than of the pick-up events. That reflects one typical traffic flow phenomena of the airport that more flight arrives than departs in that time interval. While in Figure 29(b) and Figure 29(c) all the events increased considerably from 6 am to 18 pm. And in Figure 29(b) the pick-up events is more frequent than drop-off events. In figure 29(c), the number of pick-up and drop-off events are almost same. The statistics of the pick-up and drop-off events during the 4 time intervals are presented in Figure 30.

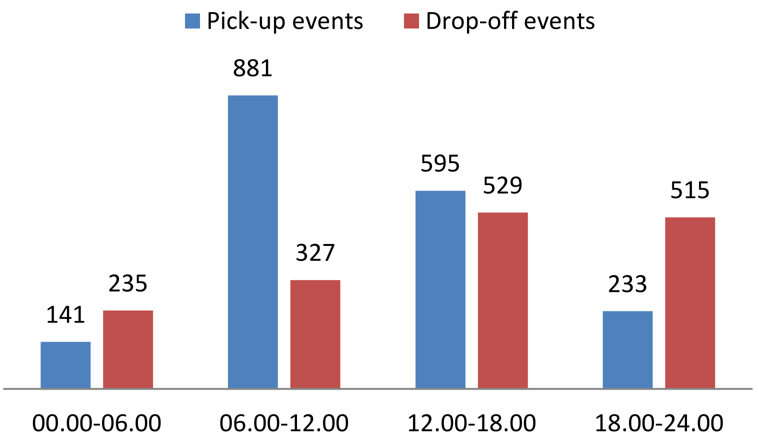


Figure 30 Statistics of pick-up and drop-off events in 4 time groups

4.4 SPATIAL ANALYSIS OF SHANGHAI FTD

In this section, we would like to mine the data in the spatial dimension. We would like to answer questions like: where are the most significant places of the pick-up or drop-off events?

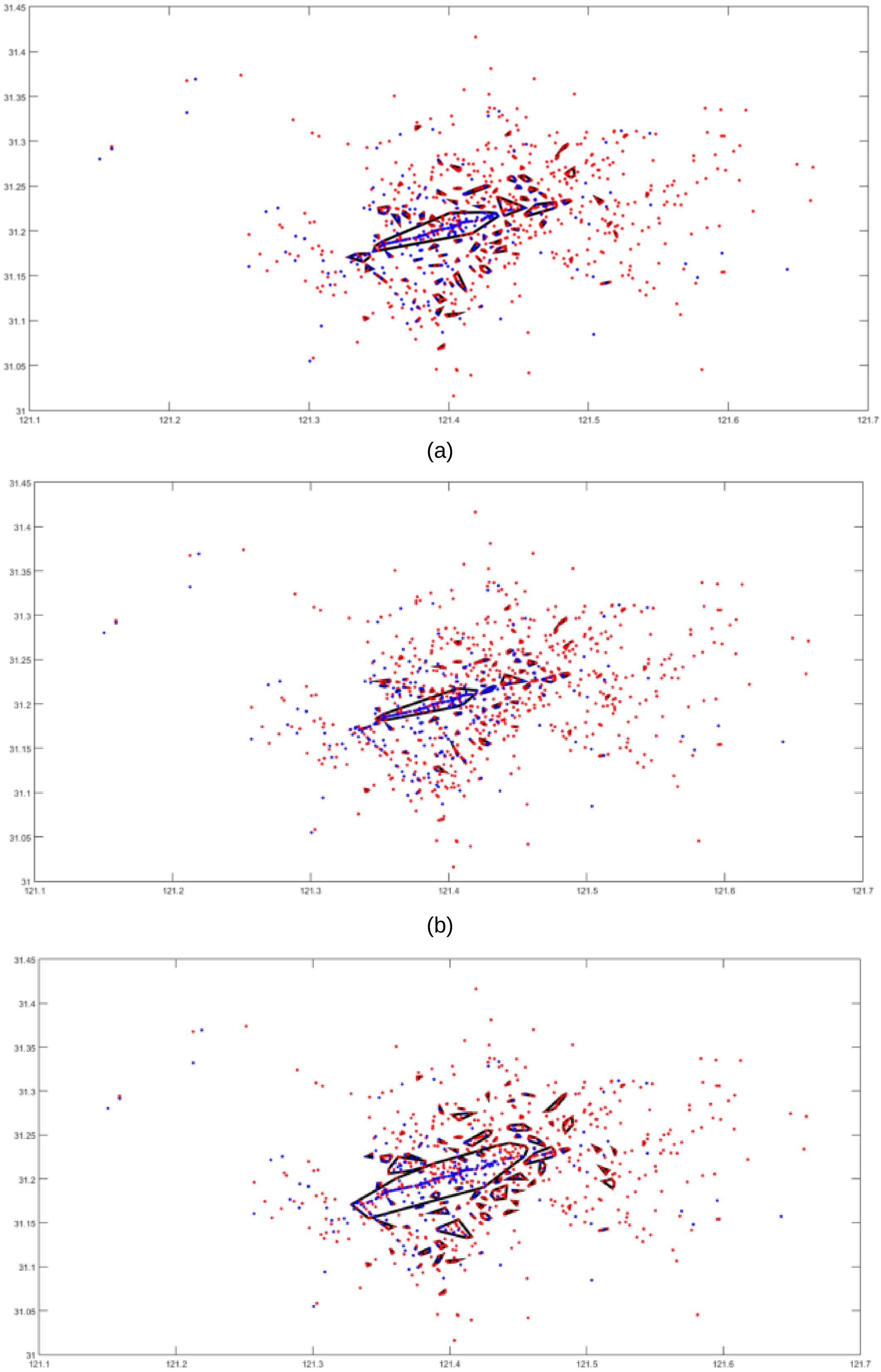
4.4.1 Spatial Clustering of Pick-up and Drop-off Events

Spatial clustering can be used to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis (Han, Kamber and Tung 2001). In section 2.2.2 we already discussed several spatial clustering methods which could apply to FTD set. In this thesis, we use a hierarchical agglomerative clustering method (Kaufman and Rousseeuw 1990) to detect the dense regions where most of the pick-up or drop-off events happen. The reason we choose the hierarchical clustering method is that the clustering results can be easily investigated at multiple scales.

The input of the clustering method is the individual pickup or drop-off events. And the output is the clustered events indicating which cluster the event belongs to. Generally speaking, the input parameters of hierarchical clustering are normally a distance function and a linkage criterion. In our experiment, we use the hierarchical clustering function embedded in MATLAB to cluster the pick-up and drop-off events. There are two options concerned with parameter for us:

1) Let the cluster function create clusters determined by the natural divisions in the data set.
2) Specify the number of clusters 'maxclust' we want to create clusters.

In this thesis, we choose the second option to specify the number of clusters. Since the parameter setting is largely dependent on the applications or tasks, we plan to try a different number of 'maxclust' to find out which clustering result is more empirically appropriate. In order to indicate different cluster on the graph clearly, we use the convex hull to show the spatial extent of each cluster. Here we take the time period '06 - 12 am' as an example, and the hierarchical clustering results are illustrated in Figure 31.



(a)

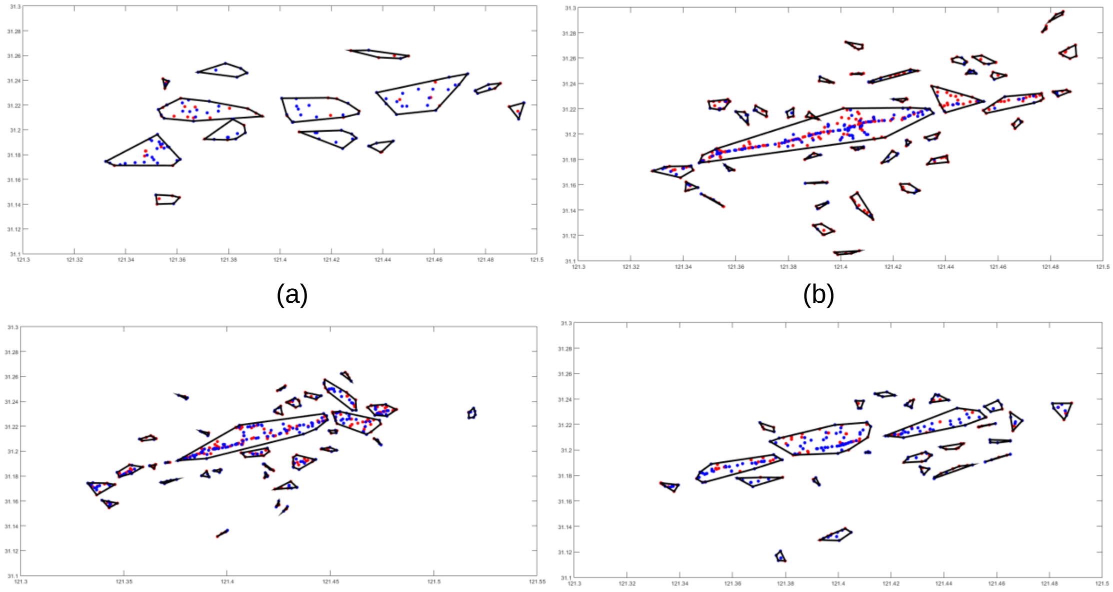
(b)

(c)

Figure 31 Hierarchical clustering with bounding convex hull of pick-up (in red) and drop-off (in blue) events from 6 to 12 am (a) ‘maxclust‘ in 500 (b) ‘maxclust‘ in 600 (c) ‘maxclust‘ in 400

In this time period, there are in total 1,208 events (881 pick-up points and 327 drop-off points) have been recorded. Figure 32 is the clustering results of 3 ‘maxclust’ value as in 400, 500 and 600. From those 3 figures, we could tell that the result of Figure 32(a) is more reasonable than the other two based on the size of clusters. If the ‘maxclust’ value is too large, then the cluster will be too small to be meaningful. On the contrary, the clusters would have some overlap and we cannot extract valuable information from them.

In addition, in order to define the significant places, we set a significant threshold value for the minimum number of the cluster. If a cluster has a number of elements larger than the significant threshold value, then this cluster is a significant cluster. Since the clustering method will assign each point to a cluster and here we are interested only in clusters with a higher density or a larger number of cluster elements, we select clusters with more than 5 points as significant clusters. Therefore, after finding out the best ‘maxclust’ for each time group through the method we mentioned above, the resulted clusters are illustrated in Figure 32.



(a) (b)

(c) (d)

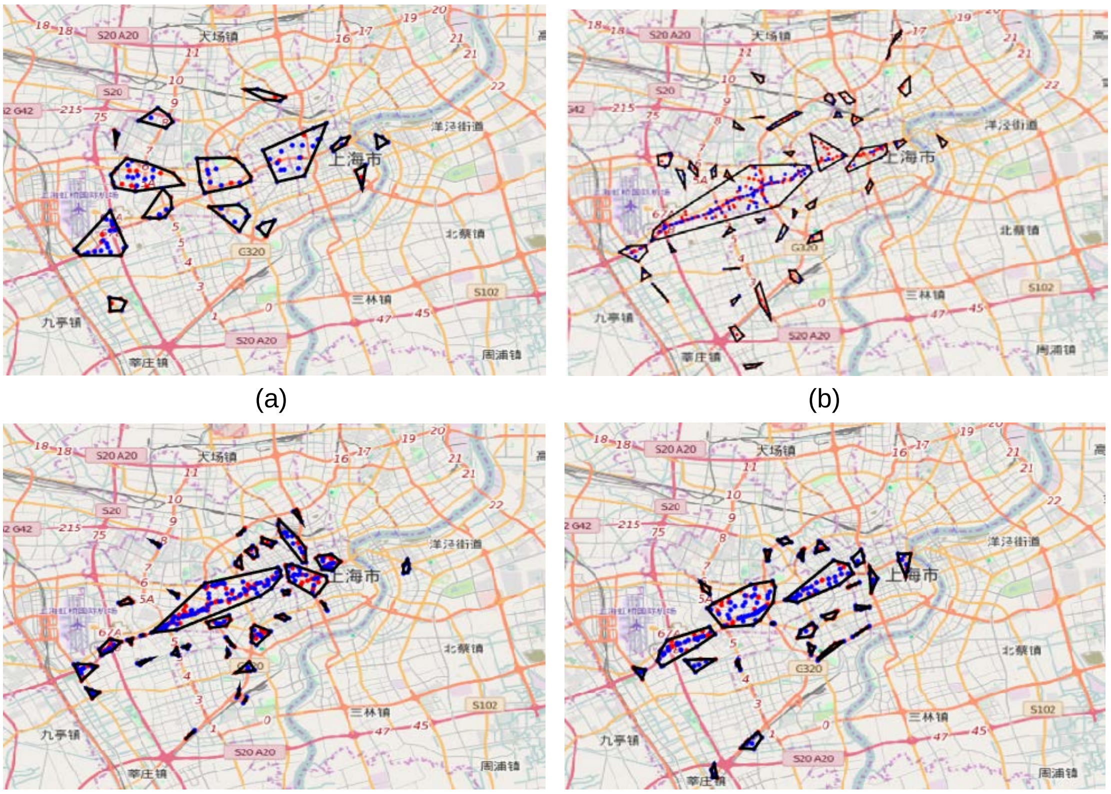
Figure 32 Filtered clusters of pick-up (in red) and drop-off (in blue) events in 4 time periods (a) 0 to 6 am with ‘maxclust‘ in 150 (b) 6 to 12 am with ‘maxclust‘ in 500 (c) 12 to 18 pm with ‘maxclust‘ in 500 (d) 18 to 24 pm with ‘maxclust‘ in 350

From the above-filtered clusters, we could see that some clusters in different time period share the similarity of their location and size. To find out where exactly those clusters in Shanghai are and what kind of semantic category they belong, we need to proceed semantic clustering.

4.4.2 Semantic Labeling of Pick-up and Drop-off Events

With the significant places mined from a different period of taxi pick-up and drop-off data, the semantics of these places can be further identified. In this thesis, we manually check the semantic meanings of the significant clusters by comparing the clusters with the base-map.

We classify all the locations into three commonly used categories, namely 'residential', 'public' and 'commercial'. For example, cluster locations like 'hotel' and 'community' are classified to 'residential', cluster locations like, 'hospital', 'university' and 'school' are classified to 'public', and cluster locations like 'company', 'shopping center' belong to 'commercial'. So we add the OpenStreetMap (OSM) layer to the filtered clusters of each time period as in Figure 33. In this way OSM layer could give us location information of these clusters.



(a) (b)

(c) (d)

Figure 33 Filtered clusters of pick-up (in red) and drop-off (in blue) events in 4 time periods with OSM layer (a) 0 to 6 am with 'maxclust' in 150 (b) 6 to 12 am with 'maxclust' in 500 (c) 12 to 18 pm with 'maxclust' in 500 (d) 18 to 24 pm with 'maxclust' in 350

We zoom in the map layer to identify the location type of each cluster. Here we also take time period '6 am to 12 am' as an example. As we see in Figure 34, the biggest 3 clusters labeled with number 1, 2, 3 stands for the following location types:

- In cluster 1, most pick-up and drop-off events located on 'Yanan elevated road' and its surrounding areas are mostly residential communities. Therefore, we classified this cluster to 'residential'.
- Cluster 2 is also a typical area of residential places in Shanghai.
- Cluster 3 is the area of one big shopping center with several shopping malls. So it goes to 'commercial'.

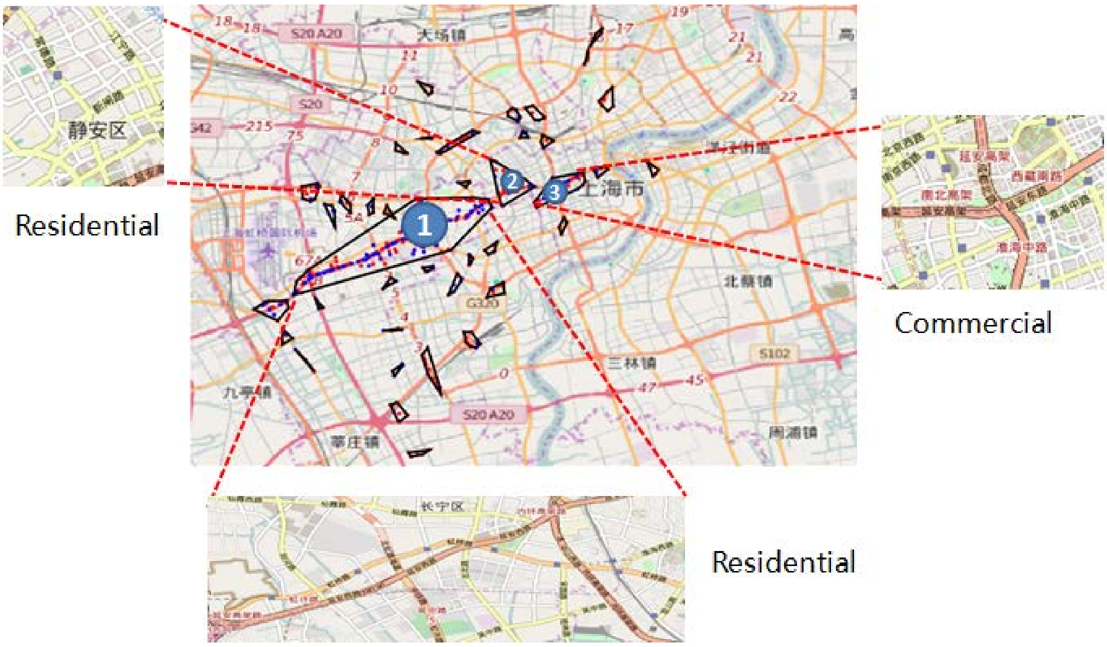


Figure 34 Example of zoomed in map information of different clusters (6 am to 12 am)

In this way, we classified all the filtered clusters of 4 different time periods, and the statistic results are shown in Figure 35. We could see that in general 'residential' type are the mostly significant and the number of 'commercial' and 'public' clusters are relatively less.

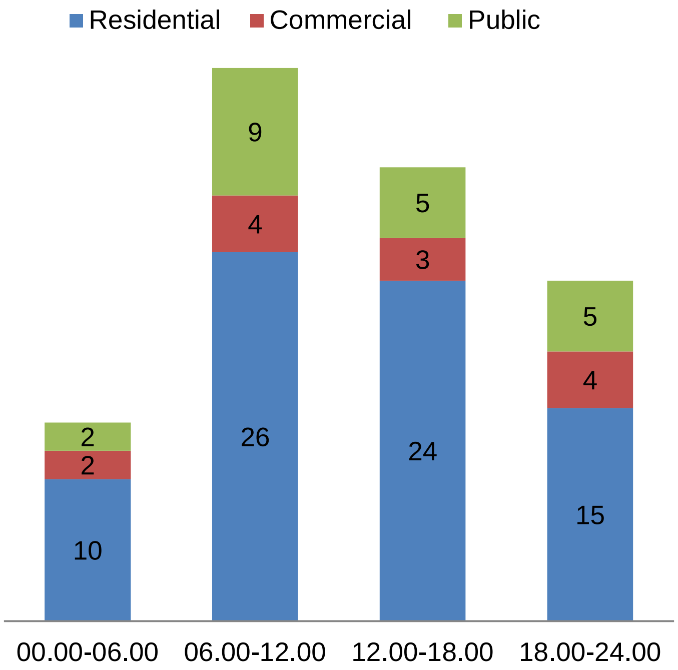


Figure 35 The temporal frequency of four semantic types of significant clusters in 4 different time periods

4.5 SPATIOTEMPORAL VISUALIZATION OF SHANGHAI FTD

In this section, we use Quantum GIS (QGIS) to visualize the analyzed results from Section 4.2 and 4.3. QGIS is a cross-platform free and open source desktop geographic information system (GIS) application that provides data viewing, editing, and analysis. QGIS allows users to create maps which can be assembled in different formats and for different uses. In addition, it also allows maps to be composed of raster or vector layers. The vector data are stored as either point, line, or polygon features. Different kinds of raster images are supported, and the software can geo-reference images.

To begin with, let us recapture the different visualization methods mentioned in section 2.2.3 and then choose appropriate mapping approaches to visualize the analyzed results. So far the analyzed results include 1) the location of each filtered cluster (cluster has over 5 events); 2) the number of pick-up and drop-off events in each cluster; 3) the semantic category of each cluster; 4) the time attributes of each cluster. Therefore, the visualization should, at least, reflect 2 attributes at the same time, like the location of the cluster plus one of the second, third or the fourth results. In this case, 1) 'dot mapping' is rather the illustration of the location of each pick-up or drop-off events instead of cluster with a group of dot events; 2) 'density mapping' could be a choice of visualizing the clusters but compare to 'proportional symbol mapping' which can give us a relatively more accurate and lucid graph which among 'dot mapping', the latter is better. 3) '3D Spatiotemporal Mapping' could present the data in a more intuitive way so it could an option when the two-dimensional illustration is not good enough.

Here we still use the time slot of '6 am to 12 am' as an example and map the spatiaotemporal aggregates from the cluster results in proportional symbols. Since we generated the filtered clusters before, and we aggregate the number of events and calculate the geometrical centroid of each cluster. Proportional bubble symbols are applied to represent the aggregates. The size of the symbol is scaled according to the number of the cluster elements and the location of the symbol is the cluster centroid. Thus, a scaled bubble map is designed to indicate not only the location of the cluster but also the normalized size of the cluster based on the number of events (see Figure 36).

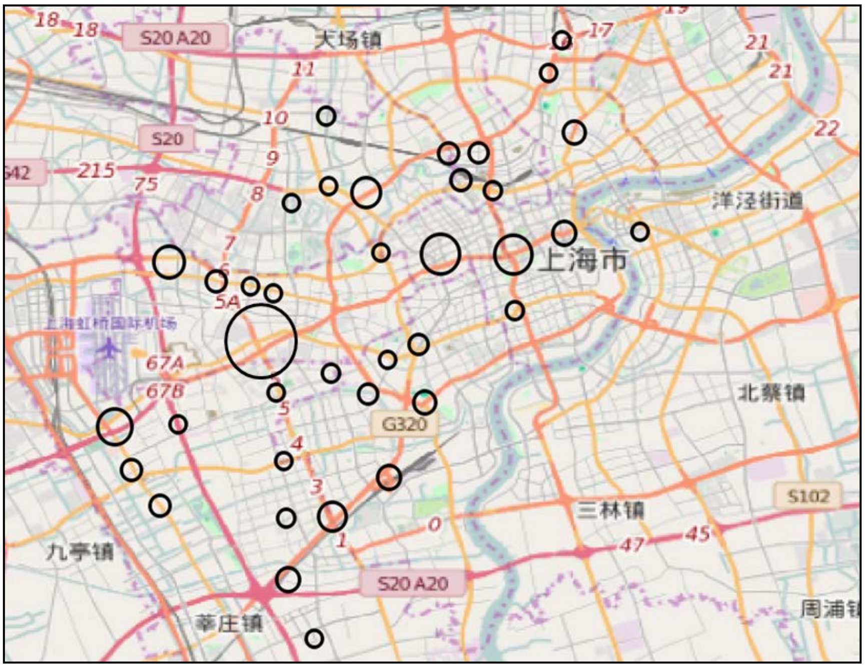


Figure 36 Scaled bubble map in time period ‘6 to 12 am’

Moreover, as shown in Figure 37, we colored each bubble in 3 different colors which represent the semantic categories of the clusters. It is clear to see that in the city center, most pick-up and drop-off events happened in public places while the events happened in residential places scattering all over the city.

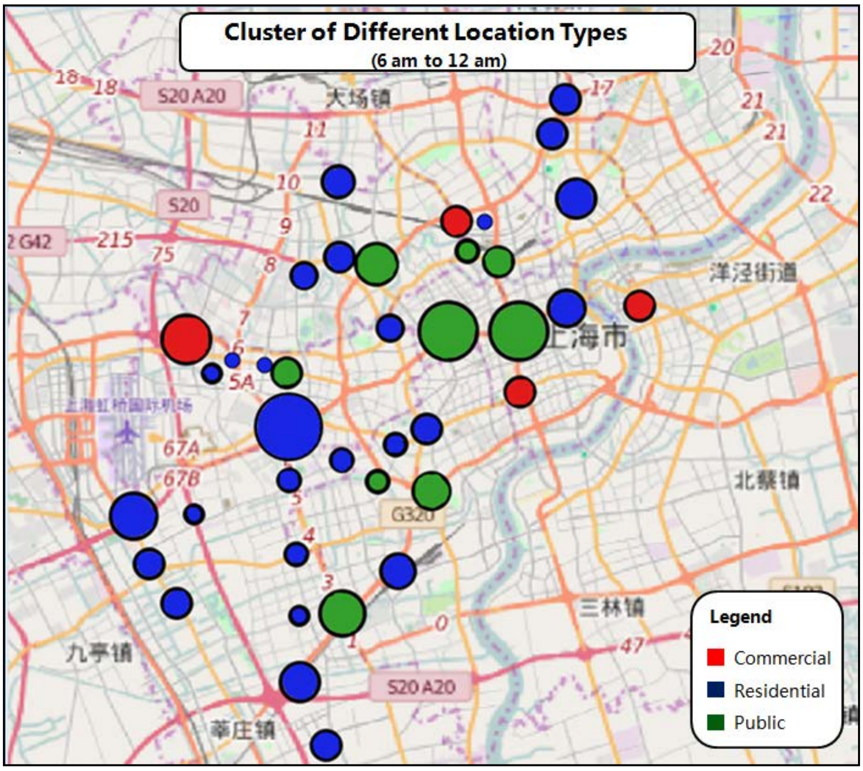
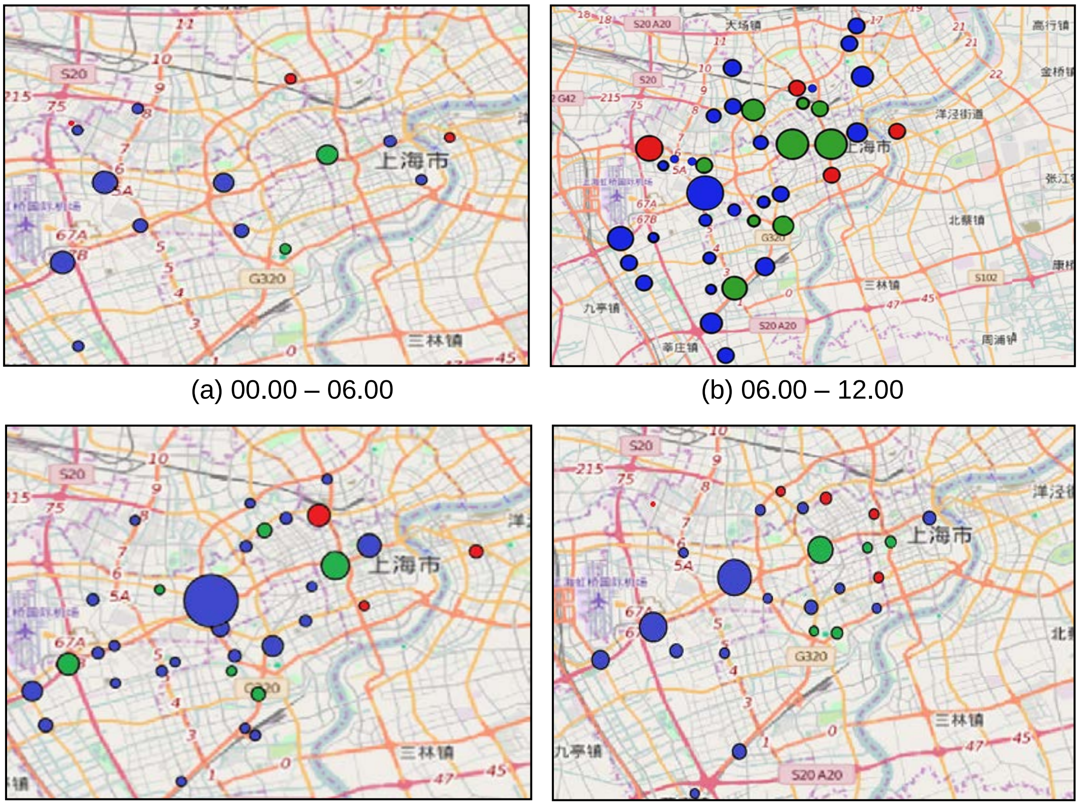


Figure 37 Scaled bubble map of different semantic categories in time period ‘6 to 12 am’

To reveal the temporal differences, we showthe semantic categories in scaled bubbles in 4 different time periods in Figure 38. The spatial clusters or significant places are mainly distributed around the city center with a few scattered far from the center. It is easy to see that in the midnight from 0 am to 6 am, most of the pick-up and drop-off events happened in residential places while in the morning from 6 am to 12 pm, they emerged and increased especially in the city center. From 12 pm to 24 pm the distribution of clusters changes a little bit, it moves from city center back to the residential places and the number of events decreases a little bit.



(a) 00.00 – 06.00

(b) 06.00 – 12.00

(c) 12.00 – 18.00

(d) 18.00 – 24.00

Figure 38 Scaled bubble maps of different semantic categories in different time periods

To sum up, as one of the proportional symbol maps, scaled bubble map demonstrate the location and category information in an easily perceivable way. However in order to observe the number of pick-up and drop-off events in each individual cluster simultaneously, we introduced pie chart to present the proportion of pick-up and drop-off events in each cluster. See an example of time period ‘6 to 12 am‘ in Figure 40.

The pie chart map could give us three-fold information: 1) the location of the pie chart is the centroid of each cluster; 2) the size of each pie indicates the number of events in that cluster; 3) the proportions of red and blue color of each pie represent the ratio of pick-up and drop-off events respectively in that cluster. From Figure 39 we could tell that in general, during 6 am to 12 am in the morning basically in any location of the cluster, the number of pick-up events is considerably bigger than drop-off events.

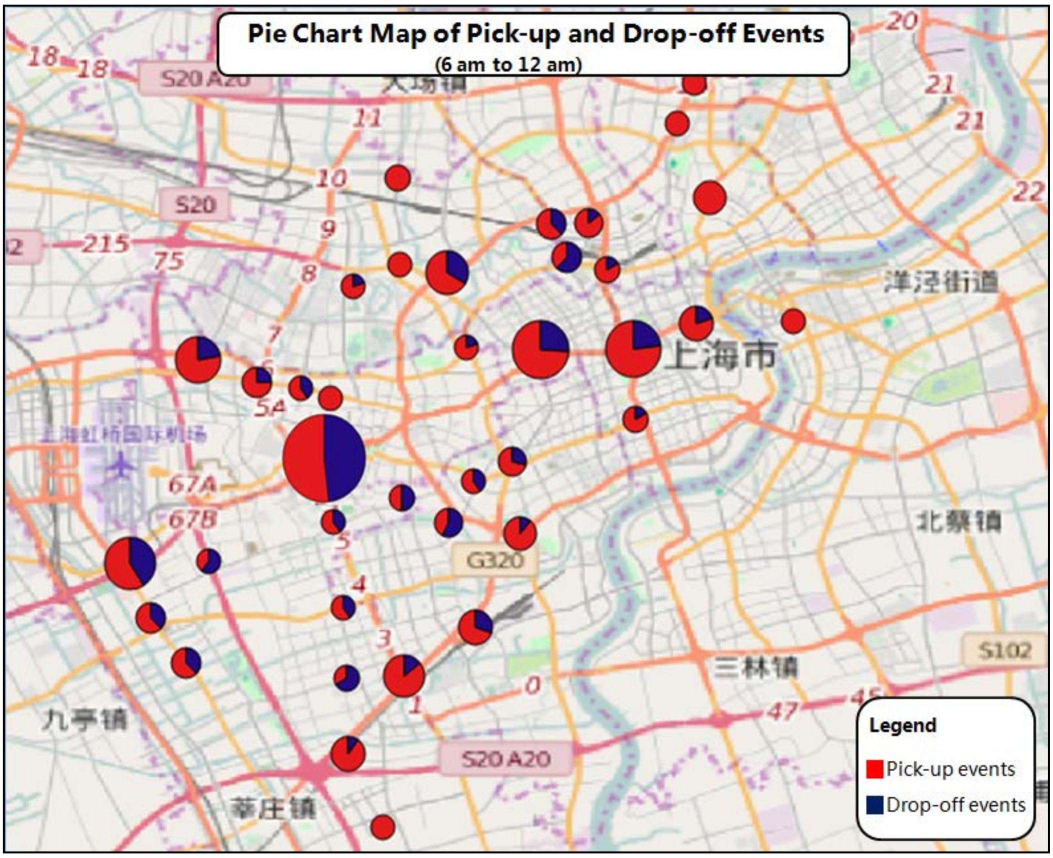
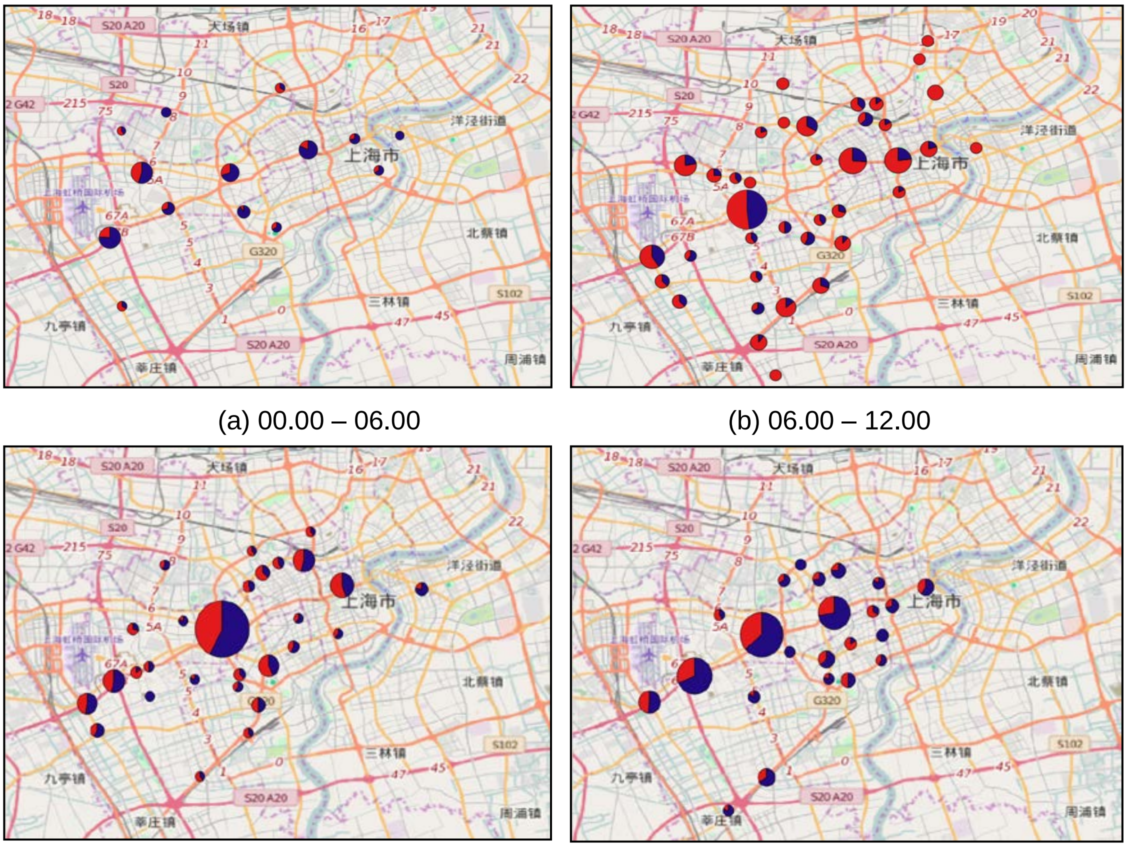


Figure 39 Pie chart map of pick-up and drop-off events in time period ‘6 am to 12 am’

In addition, we create the pie chart map in 4 different time periods as in Figure 41. It is evident that only expect from 6 am to 12 am, the number of drop-off events is quite larger than pick-up events for the rest of the time in a day.



(a) 00.00 – 06.00

(b) 06.00 – 12.00

(c) 12.00 – 18.00

(d) 18.00 – 24.00

Figure 40 Pie chart maps of pick-up and drop-off events

So far we have used 4 different graphs to present 4 different time periods. Sometimes it is difficult to compare the information from the different graph simultaneously. Therefore, in order to present time attributes, the number of events and location of each cluster in one map, we choose pie chart again as our mapping method to present all those information.

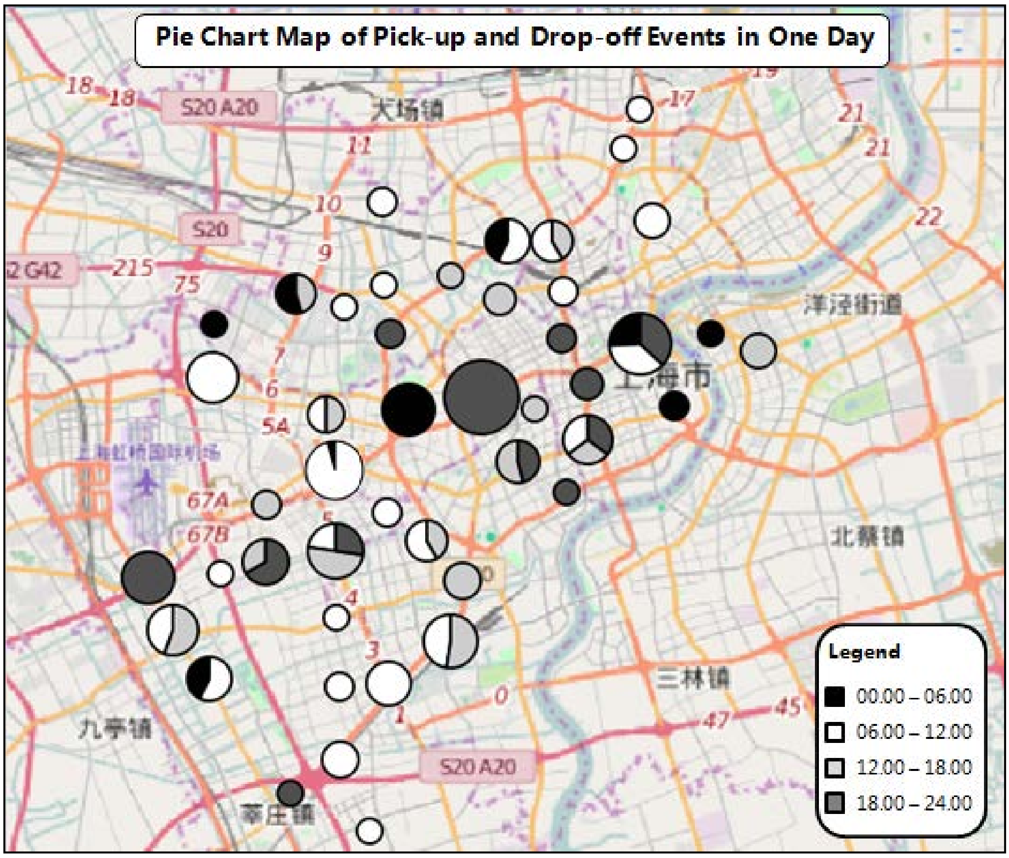


Figure 41 Pie chart map of pick-up and drop-off events in one day

We firstly generate all of the clusters of one day in one map based on its location. Certainly, there are some overlaps among those clusters. Hence, we aggregate the overlapped clusters into one big complex cluster with the shared centroid. Then we demonstrate all the clusters with time attributes on one map. So you can see in Figure 41, for each pie chart, 1) black represent time period ‘0 am to 6 am‘ in the night; 2) white represent time period ‘6 am to 12 am‘ in the morning; 3) light grey represent time period ‘12 am to 18 am‘ in the afternoon; 4) dark grey represent time period ‘18 am to 24 am‘ in the evening.

From Figure 41 it is effortless to tell the proportion of the events in different time zone. For example, the complex cluster in the city center is composed by 3 parts: around a quarter of events is from 0 am to 6 am and more than 1/3 of events is from 6 to 12 am and the rest are from 18 to 24 pm.

5. RESULT AND ANALYSIS

In chapter 4, we proposed a methodological framework and apply it to one-day FTD in Shanghai. The statistical and visualization results turn out to meet our expectations. In this chapter, we apply the proposed workflow to process one-week data (17th May – 23rd May) to understand in depth the spatiotemporal patterns of related to Hongqiao airport.

Firstly we calculate the number of generated taxi trajectories from Monday to Sunday. The statistical result is shown in Figure 42. Red color represents the number of trajectories moving towards the Hongqiao airport and blue represents the opposite moving direction. As we can see, the amount of trajectories of two directions is almost the same. To be more specific, the number of taxi trajectories to the airport is only slightly bigger than from the airport with only one exception on Friday. Furthermore, the number reaches its peak on Monday and the valley comes on Friday and Saturday.

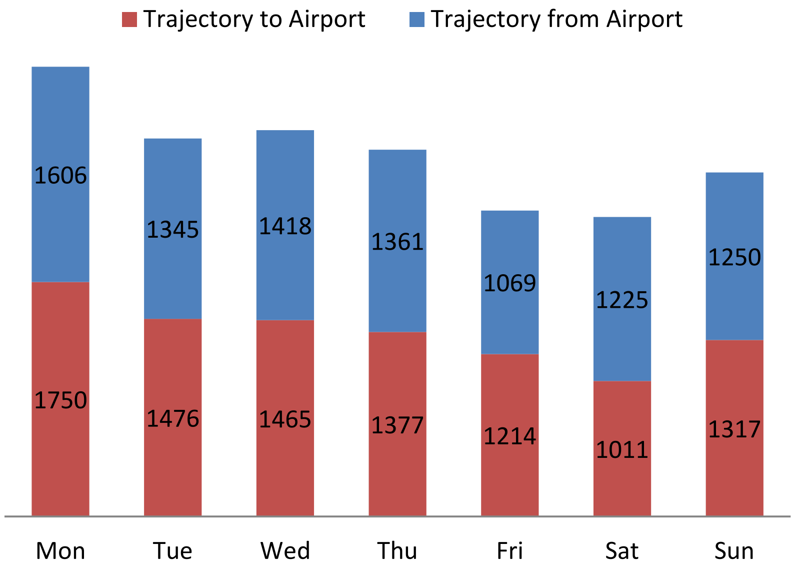


Figure 42 Histogram of the number of generated taxi trajectories in one week

Secondly, in order to investigate the number of trajectories in a detailed temporal interval, we partition the daily data into every two hours. Figure 43 illustrates the statistical distribution of the amount of generated taxi trajectories in every two hours. It is easy to find out that the peak hour of each day is always from 6 am to 8 am. Besides, it is apparent that during this time period, the number of trajectories to the airport is much more than that from the airport. Apart from that, the daily patterns of trajectories number in the rest time periods of each day are almost the same from Monday to Sunday.

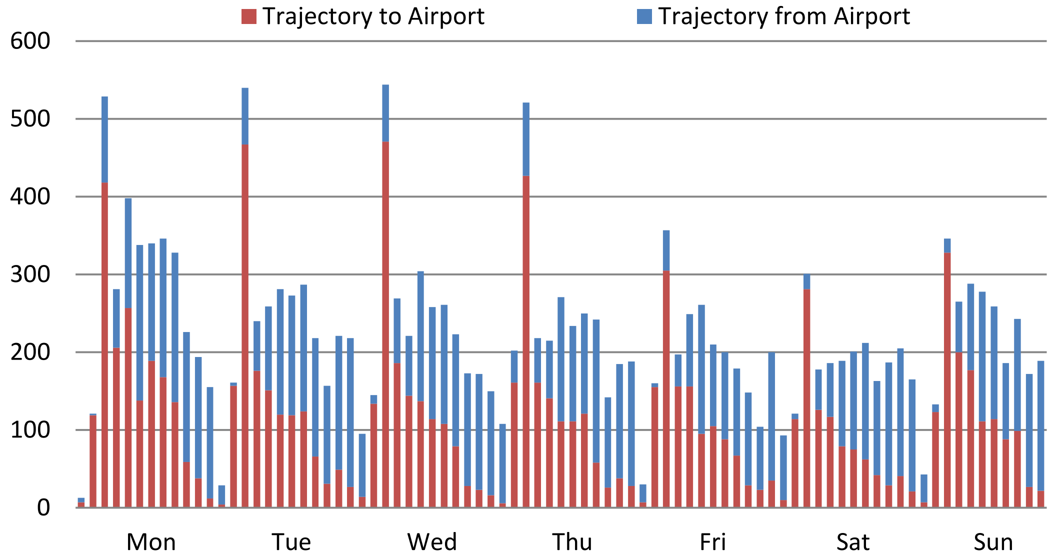
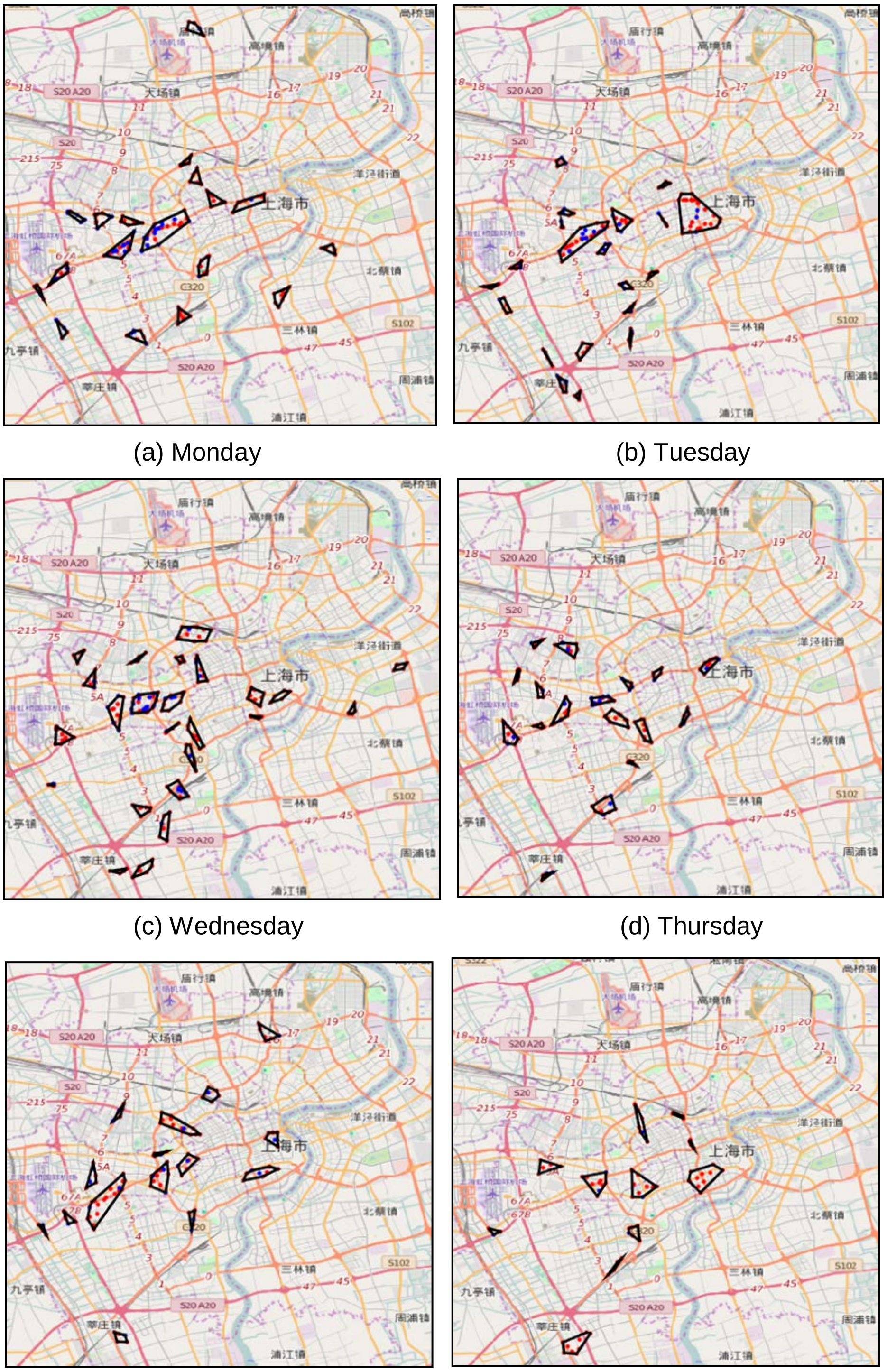


Figure 43 Histogram of the number of taxi trajectories in every 2 hours in one week

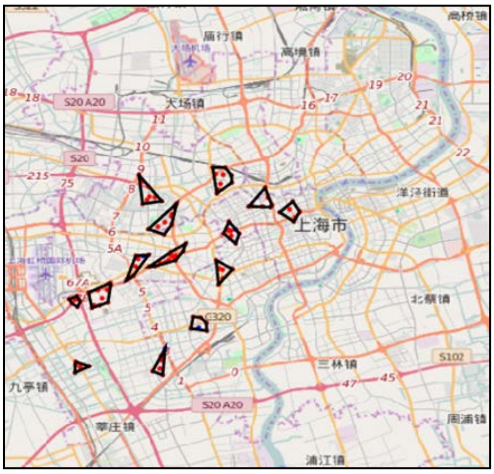
Therefore, we extract the trajectories from 6 am to 8 am to dig more information about the spatial distribution of the airport related significant places. According to Section 4.4, after removing the pick-up and drop-off events of each trajectory inside of Hongqiao airport area, we used the aforementioned hierarchical clustering method to cluster the rest events and only keep clusters with more than five events. To explicitly show the spatial extent of each cluster, we construct its convex hull by connecting its bounding cluster elements. The results of the clustering results are shown in Figure 44. Red points represent the pick-up events and blue points represent the drop-off events.



(a) Monday (b) Tuesday

(c) Wednesday (d) Thursday

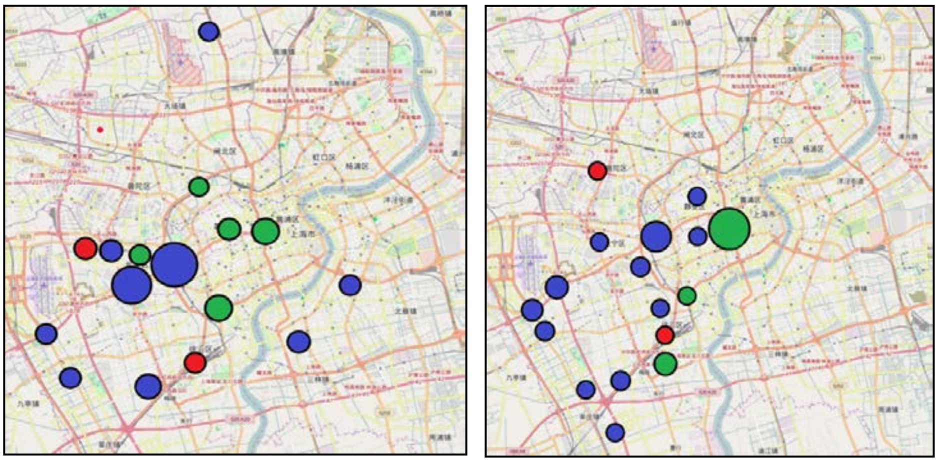
(e) Friday (f) Saturday



(g) Sunday

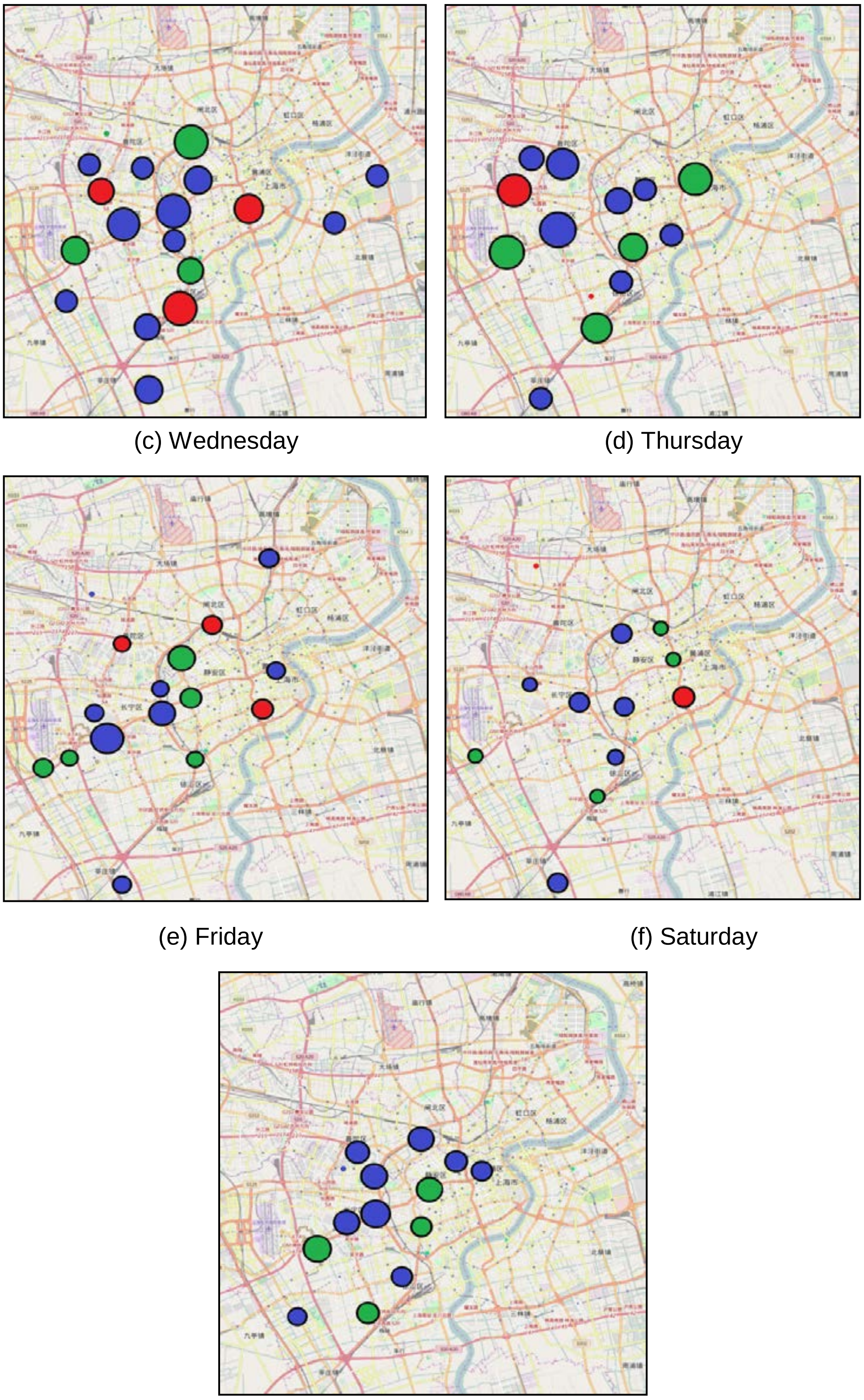
Figure 44 Filtered clusters of pick-up (in red) and drop-off (in blue) events from Monday to Sunday

From Figure 44 we could tell that the filtered clusters scatter mostly in the city center area and the number of clusters starts decreasing from Friday. In order to indicate the semantic categories of significant places, we create a scaled bubble map of each day to present not only the location type of each cluster based on the 3 different colors but also the number of events in each cluster based on the normalized bubble size. As aforementioned, red stands for 'commercial area', yellow stands for 'public area' and blue stands for 'residential area'. See results in Figure 45.



(a) Monday

(b) Tuesday



(c) Wednesday

(d) Thursday

(e) Friday

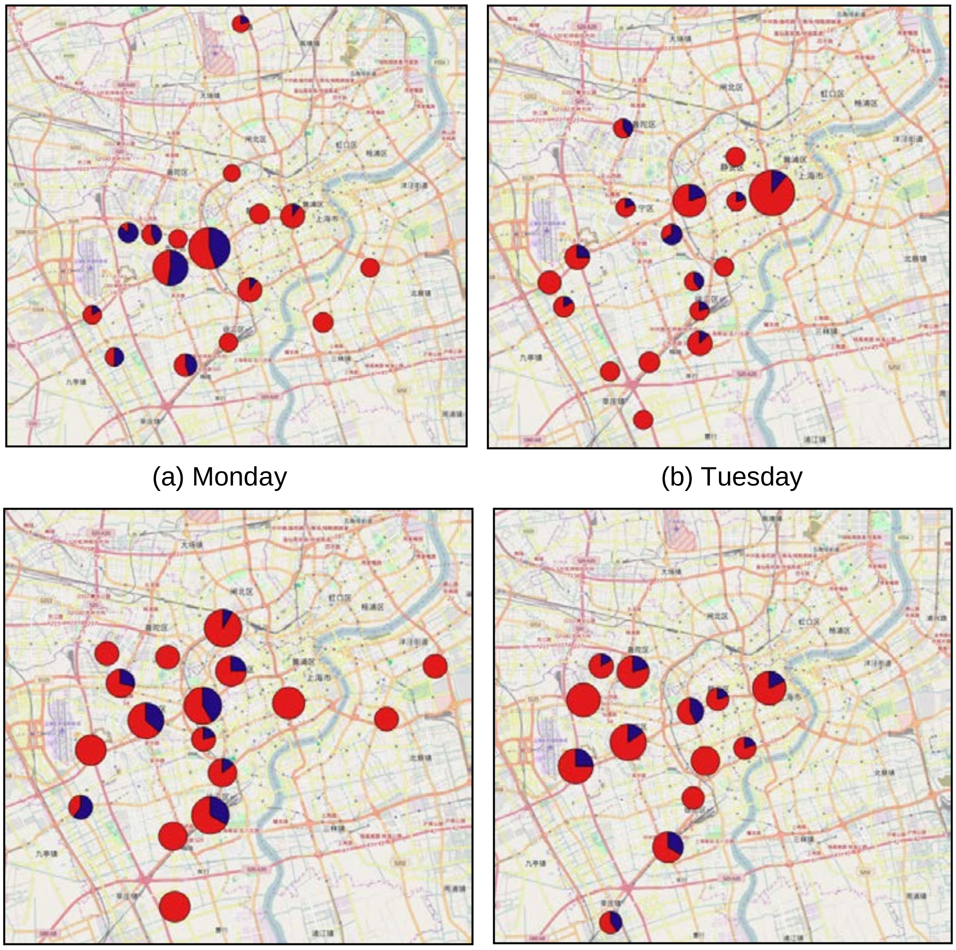
(f) Saturday

(g) Sunday

Figure 45 Scaled bubble map of different semantic categories from Monday to Sunday

The first impression from Figure 45 is that the largest group of bubbles is in blue, meaning that the majority of the significant places is 'residential area'. By comparing Figure 45(a)-(g), we could see that the location type of 'commercial' clusters are insignificant at the weekend (Figure 45(f) and Figure 45(g)) than those during the weekdays. Additionally, we could see that the location of clusters concentrates more in the city center area on Sunday.

Then in each cluster, we use a pie chart to indicate the number of its pick-up and drop-off events. The results are shown in Figure 46. Red and blue represent pick-up and drop-off events respectively.

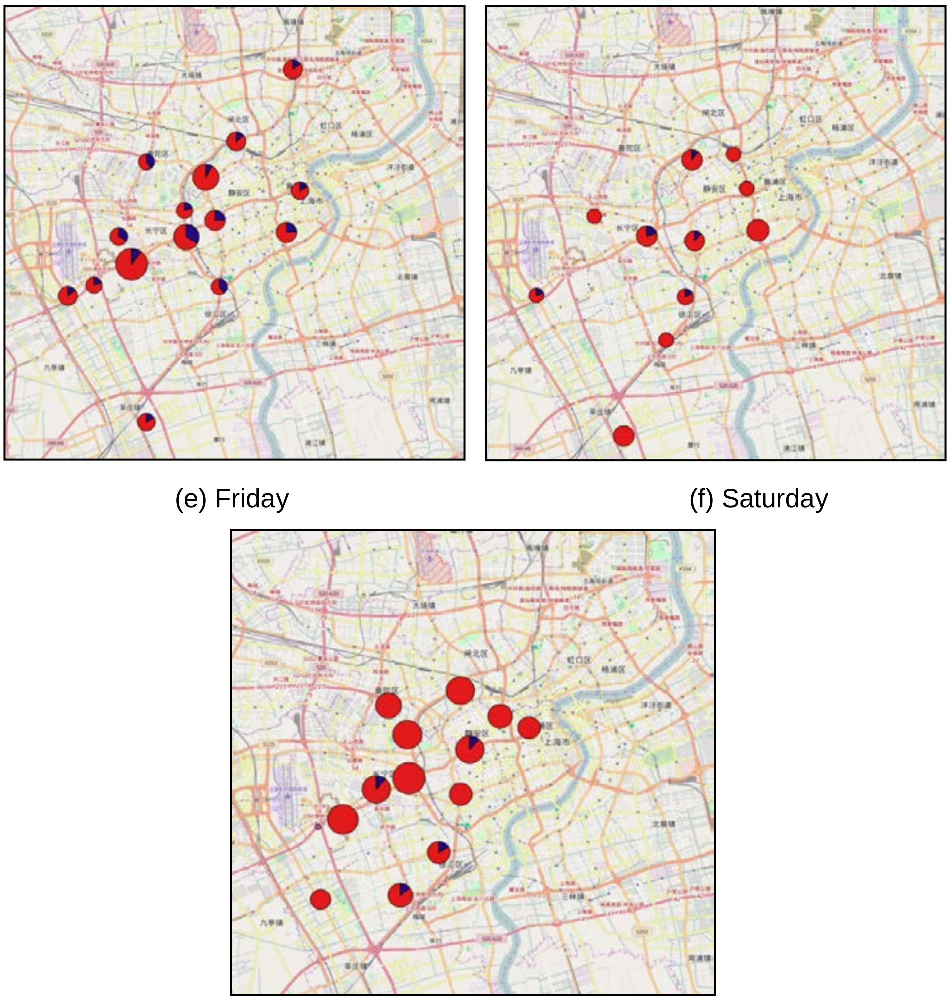


(a) Monday

(b) Tuesday

(c) Wednesday

(d) Thursday



(e) Friday

(f) Saturday

(g) Sunday

Figure 46 Pie chart maps of pick-up and drop-off events of each cluster from Monday to Sunday

Corresponding with the statistical results from Figure 43, Figure 46 presents the data in a more intuitive way that the number of pick-up events to Hongqiao airport is a lot more than the drop-off events from the airport. And combined with Figure 45 we could find out that most of the drop-off events happened in the residential area.

In order to present the weekly data in one map, we first aggregate the overlapped clusters into one big complex cluster with the shared centroid and then create a pie chart over each cluster to demonstrate the distribution of the number of events from Monday to Sunday in each significant place. As shown in Figure 47, red, orange, yellow, green, cyan, blue and purple color represent the sum of the pick-up and drop-off events on Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday respectively.

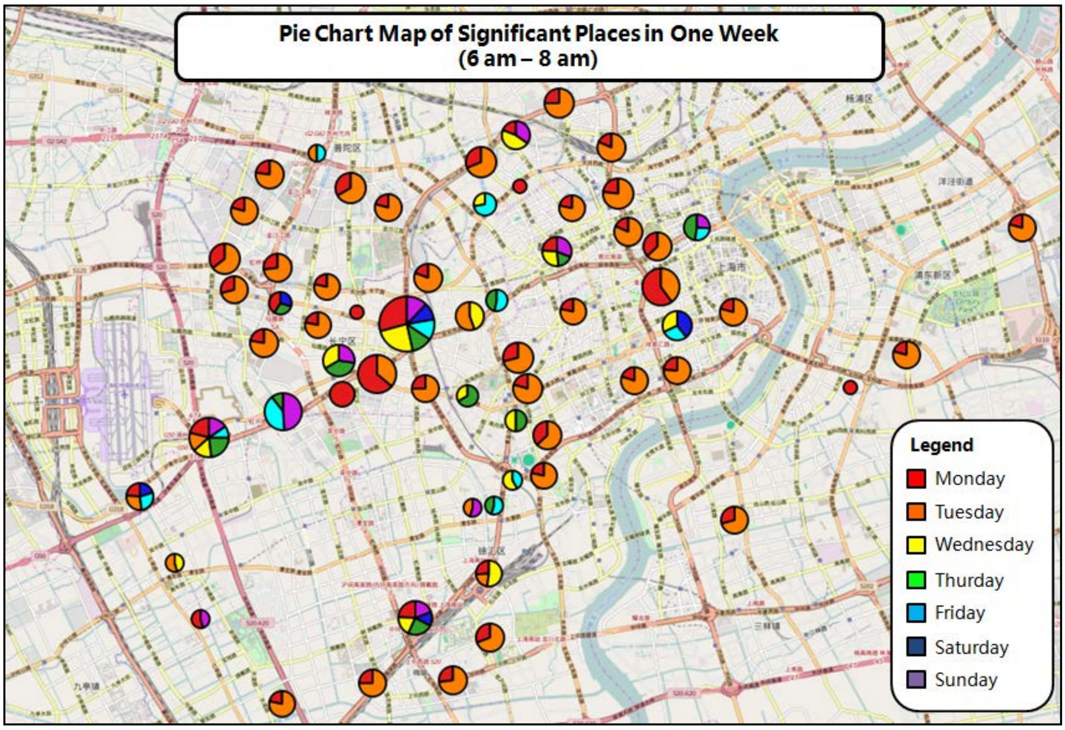


Figure 47 Pie chart map of Hongqiao airport related significant places from Monday to Sunday

It is evident from Figure 47 that the majority of the significant places (mostly belong to location type ‘residential‘) are composed of events from Monday to Tuesday (red and orange), and mostly Tuesday (orange). In addition, we could see that a large amount of significant places located along the main roads.

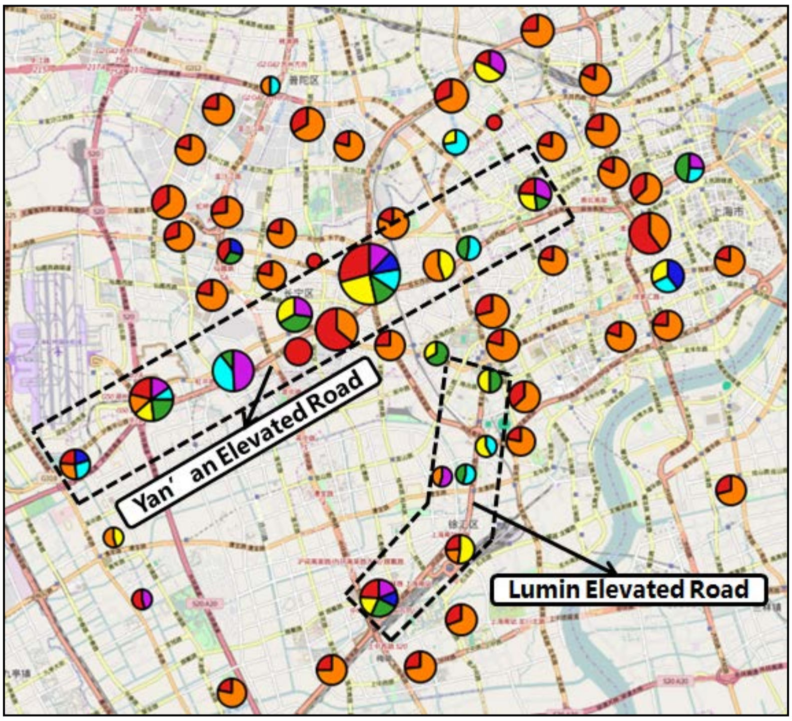


Figure 48 Pie chart map of highlighted roads

As illustrated in Figure 48, along the 'Yan'an Elevated Road' and 'Luming Elevated Road' the time distribution of significant places from Monday to Sunday appears to be more even. That means, from our floating taxi dataset, Yan'an Elevated Road, and Luming Elevated Road should be the most popular road where taxis pick up passengers to Hongqiao airport or drop off passengers from Hongqiao airport.

Moreover, 3 pie charts with the evenest time distribution have been chosen. In order to find out the reason of this phenomenon, we zoom in the base map to explore the surrounding area of them.

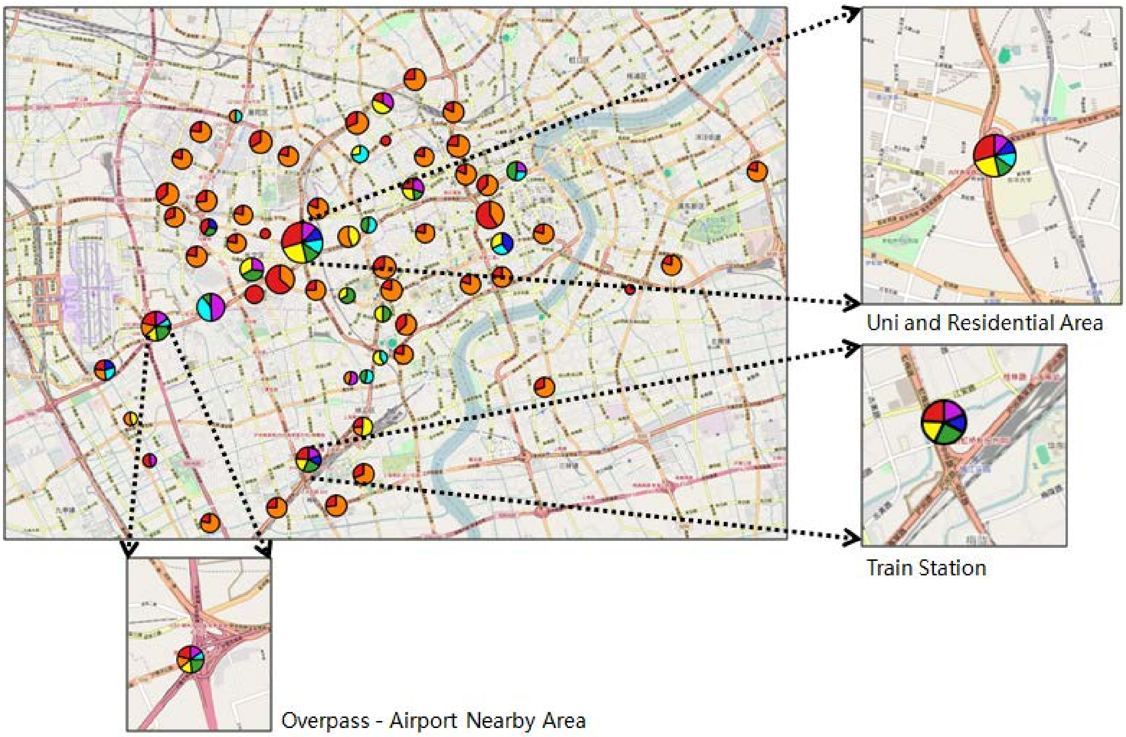


Figure 49 Surrounding areas of highlighted pie charts

As shown in Figure 49, the biggest pie chart is located in one overpass and surrounded by several residential communities and Donghua University. The pie chart on the bottom of the graph is located near Shanghai South Railway Station. And the pie chart close to Hongqiao airport is located on one big overpass.

6. CONCLUSION AND OUTLOOK

This thesis aims to visually analyze the traffic flows and significant places related to transport hubs using movement data. We proposed a visual analysis workflow incorporating computational algorithms, data mining approaches, and visualization techniques for the exploration of the spatiotemporal patterns of taxi flows and visual mining of the significant places. More specifically, we firstly preprocess large amounts of the movement trajectories and extract their starting and ending points (see Section 4.3). Secondly, we identify appropriate spatial and semantic clustering methods to derive and categorize different significant places (see Section 4.4). Finally, we design appropriate spatial and temporal visualization techniques, e.g. dot maps, proportional symbol maps, pie chart maps, to analyze the significant places (see Section 4.5)

In this thesis, a one-week test dataset of floating taxi data in Shanghai city is used to extract the taxi flows information of Hongqiao international airport. We propose a general workflow which mainly consists of data preprocessing, taxi trajectory generation, pick-up/drop-off events extraction, spatial clustering, semantic classification and spatiotemporal visualization.

The experiment results show that our proposed methods are feasible and additionally we answered the questions proposed in the beginning to achieve the goal of this thesis. By using the hierarchical clustering method to cluster the scattered pick-up and drop-off points and by using a semantic clustering method to categorize them to the different significant places, we create a spatial and temporal visualization –proportional symbol maps of the airport related significant places. The maps reveal several phenomenon of taxi flows from and to Hongqiao airport area which are 1) in the daytime (6 am to 18 pm) more people take taxis to Hongqiao airport while at night (18 pm to 6 am) the situation is the opposite; 2) the number of trajectories from/to Hongqiao airport in each day reaches it peak from 6 am to 8 am; 3) the number of trajectories from/to Hongqiao airport in weekday is larger than in the weekend; 4) Yan’an Elevated Road and Luming Elevated

Road are the most popular road where taxis pick up passengers to Hongqiao airport or drop off passengers from Hongqiao airport.

Moreover, by using visual analytics methods, we can conclude that FTD is a very useful data source to derive information about spatiotemporal patterns of traffic flows in the city area. Nevertheless, there are still some improvements could be anticipated in the future.

First of all, semantic clustering of pick-up and drop-off events is done by manually checking the OSM base map. Although the accuracy of the results is pretty satisfying, the process is very time-consuming. For further research, automatic classification approaches should be introduced.

Secondly, interactive techniques and appropriate user interface can be integrated and designed to allow the user to freely explore massive data and their spatiotemporal patterns. Since nowadays it is possible to link and reference the different data in a quite interactive way, further investigation may apply to design one interactive visual interface. For example, the visual interface could consist of three interlinked components: (1) a map view to show the spatial distribution of the individual pick-up and drop-off events (for instance, the Leaflet ‘MarkerCluster Library‘ could be implemented), and the reconstructed occupied trajectories, and the significant places inferred from the spatial clustering as well; (2) a histogram view to allow the visual exploration of different temporal patterns; (3) a pie chart view to show the time periods of semantics and the statistics of the classified significant place. In this case, users can interactively click any of the components to simultaneously inspect the corresponding features highlighted in other components. For example, if the user selects a bar in the histogram view, the corresponding place will be highlighted so that the user knows where the significant place is and how the pickup or drop-off events distribute inside the place. In this way, users can get a general idea of where are the most significant places of the pick-up and drop-off events, what are their semantic meanings, and which routes are frequently taken from the airport to a significant place.

To conclude, this thesis derives useful spatiotemporal patterns related to transport hubs by processing, visualizing and analyzing floating car data. Taxi flows related to transport hubs could reflect the transportation system, traffic condition and urban planning in many ways. This master thesis also serves as preliminary work for future researches to achieve more significant results.

REFERENCES

Andrienko, N. & G. Andrienko. 2011. Spatial generalization and aggregation of massive movement data. Visualization and Computer Graphics, IEEE Transactions on, 17, 205-219.

Andrienko, G., Andrienko, N., Bak, P., Keim, D., Wrobel, S., 2013. Visual Analytics of Movement, Springer, XIV, 397, 178-200.

http://geoanalytics.net/vam/index.html

Birant, D. & Kut, A. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data & Knowledge Engineering, 60(1), 208–221.

Bogorny, V. 2010. Tutorial on Spatial and Spatio-Temporal Data Mining Part II – Trajectory Knowledge Discovery.

www.inf.ufsc.br/~vania/tutorial_icdm.html

Brockfeld, E., Lorkowski, S., Mieth, P. & Wagner, P. 2007. Benefits and limits of recent floating car data technology – an evaluation study. 11th WCTR Conference, Berkeley, USA.

Candra, K. 2015. Visual exploration of spatial-temporal traffic congestion patterns using floating car data. Master's Thesis. Germany: Technische Universitaet Muenchen

Chu, D., Sheets, D. A., Zhao, Y., Wu, Y., Zheng, M. & Chen, G. 2014. Visualizing Hidden Themes of Taxi Movement with Semantic Transformation. Preprint for IEEE PacificVis.

Feifei, X., Xiaohong, C. & Hangfei, L. 2010. Study on Space-Time Distribution Characteristics of Floating Car Data Based on Large Samples. International Conference on Optoelectronics and Image Processing, 2, 449–452.

Gecchele, G., Rossi, R., Gastaldi, M. & Caprini, A. 2011. Data Mining Methods for Traffic Monitoring Data Analysis: A case study. Procedia - Social and Behavioral Sciences, 20, 455–464.

Keler, A. 2013. Visual analysis of traffic congestion based on Shanghai FCD. Master's Thesis. Germany: Technische Universitaet Muenchen.

Kianfar, J., & Edara, P. 2013. A Data Mining Approach to Creating Fundamental Traffic Flow Diagram. Procedia - Social and Behavioral Sciences, 104, 430–439.

Linfang, D., H. Fan & L. Meng. 2015. Understanding taxi driving behaviors from movement data. In AGILE Conference on Geographic Information Science. Lisbon.

Linfang, D., Yang, J. & L. Meng. 2015. Visual analytics for understanding traffic flows of transport hubs from movement data. ICC 2015 - 27th International Cartographic Conference, August 23-28, 2015, Rio de Janeiro, Brazil

Liu, C., Jian, Z.. & Meng, X. 2013. Combining Float Car Data and Multispectral Satellite Images to Extract Road Features and Networks, Progress in Location-Based Services: 29-43.

Liu, X., & Ban, Y. 2013. Uncovering Spatio-Temporal Cluster Patterns Using Massive Floating Car Data. ISPRS International Journal of Geo-Information, 2(2), 371–384.

S. P. Borgatti. How to explain Hierarchical Clustering.
http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html

Santi, P., G. Resta, M. Szell, S. Sobolevsky, S. H. Strogatz & C. Ratti (2014) Quantifying the benefits of vehicle pooling with shareability networks. Proceedings of the National Academy of Sciences, 111, 13290-13294.

Simon, Nick., Gates, Gary. & Burr, Jonathan. Commercial applications arising from a floating vehicle data system in Europe. Proceedings of 9th World Congress on Intelligent Transportation Systems, Vol. 44, 1–8.

Szell, M. & B. Groß. 2014. HubCab – Exploring the Benefits of Shared Taxi Services. In Decoding the City How Big Data Can Change Urbanism.

T. Zhang, R. Ramakrishnan & M. Livny. 1996. BIRCH: An efficient data clustering method for very large databases. 1996 ACM SIGMOD international conference on Management of data, 103-114.

Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu. 2002. An Efficient k-means Clustering Algorithm: Analysis and Implementation. IEEE transactions on pattern analysis and machine intelligence, vol. 24, No. 7.

Tominski, C., Schumann, H., Andrienko, G. & Andrienko, N. 2012. Stacking-Based Visualization of Trajectory Attribute Data, IEEE Transactions on Visualization and Computer Graphics, vol. 18(12), 2565-2574.

A Tutorial on Clustering Algorithms: K-Means Clustering.

http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

Yuan, J., Y. Zheng & X. Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 186-194.

Yuan, N. J., Y. Zheng, L. Zhang & X. Xie (2013) T-finder: A recommender system for finding passengers and vacant taxis. Knowledge and Data Engineering, IEEE Transactions on, 25, 2390-2403.

Yu Zheng, Xiaofang Zhou. 2011. Computing with Spatial Trajectories. Springer.

http://research.microsoft.com/pubs/153826/trajectorycomputing_preview.pdf

Zhu, Y., E. Zhong, Z. Lu & Q. Yang. 2013. Feature engineering for semantic place prediction. Pervasive and Mobile Computing, 9, 772-783.