Technische Universität München
Department of Civil, Geo and Environmental Engineering
Chair of Cartography
Prof. Dr.-Ing. Liqiu Meng

Visual Analysis of Origin/Destination Time Patterns of Travellers

Katre Karja

Master's Thesis

Duration: 01.05.2015 - 15.02.2016

Study Course: Cartography M.Sc.

Supervisor: Dr.-Ing. Mathias Jahnke

2016

Statement of Authorship

I hereby declare that this thesis named „Visual Analysis of Origin/Destination Time Patterns of Travellers“ is my own original work, and has not been submitted elsewhere in any other form for the fulfillment of any other degree of qualification. All direct or indirect sources used are acknowledged as references.

Place and date	(Signature)

Acknowledgements

I would like to express my deep gratitude to my supervisor Dr.-Ing. Mathias Jahnke for his support and guidance during my thesis time and for sharing his knowledge and experience with me. Secondly, I would like to thank Linfang Ding for being very helpful when I had questions related to database queries.

Many thank go to Juliane and Stefan the Cartography Master's program coordinators for everything they have done.

I would like to thank the professors and lectures from the Cartography Department of TUM, of TUD and of TUW for inspiring lectures and guidance during my two years of studies.

My gratitude goes for my friends and colleagues with who we shared together the destiny of living and travelling in different cities during the Master study period, and to those friends who have walked with me in other paths and whose friendship has helped me to stay focused also in my studies.

Most importantly, my special thanks go to my dearest family, my parents and sisters who have been very supportive. My greatest thanks go to Tianbao, who has encouraged me when I encountered difficulties during the writing process of this thesis.

Abstract

Taxis equipped with active Global Positioning System (GPS) devices are as sensors in urban environments. With the Floating Car Data (FCD) method data about traffic is collected. In the metropolis Shanghai, travellers take around 60 000 – 70 000 taxi trips daily. This large number of travelling events leaves a footprint by the means of data that can be used to discover the dynamic of a city. When the historic FCD data is supplemented with data about venues, then an insight into travellers‘ behaviour in areas of interest in the city can be achieved.

The first goal of this thesis is to use taxi FCD data to detect popular places in Shanghai during one week period, 17.05.2010 – 23.05.2010. For this purpose taxi travellers‘ trip origin and destination locations are extracted from the FCD. The popular places (or hot-spots) are defined as areas with high density of taxi pick-up or drop-off points on street-level. For detecting these hot-spots, a density based clustering algorithm DBSCAN (Ester et al., 1996) was implemented with tuning proper parameters. Other different clustering methods were also compared in a literature based evaluation.

The following step is to identify places near the detected hot-spots, in order to give them semantic descriptions (hotel or conference centre). This is done by tagging clusters with nearby point of interests (POI). The POIs are collected from OpenStreetMap (OSM) data layers and the Internet. The OSM is a Volunteered Geographic Information (VGI) project that offers digital data.

The third goal is to explore temporal patterns in semantically labelled hot-spots with the means of visualization techniques. The patterns on different time of a day during one week period is studied with four sample areas. In order to support visual analysis, an interactive web application was built, which is based on Leaflet.js, Crossfilter.js, Dc.js, and MongoDB. The creation of this web application is evolved with the demands of this thesis, and can be used as a platform for general spatio-temporal analysing purpose.

Table of Contents

List of Figures

List of Tables

List of Symbols, Nomenclature or Abbreviations

AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
BSON	Binary JSON (vt JSON)
CSS	Cascade Styling Sheet
D	Densities
DOM	Document Object Model
EDA	Exploratory Data Analysis
ESDA	Exploratory Spatial Data Analysis
FCD	Floating Car Data
GPS	Global Positioning System
GUI	Graphical User Interface
HTML5	Hypertext Markup Language 5
HTTP	HyperText Transfer Protocol
JSON	The JavaScript Object Notation
KDD	Knowledge Discovery from Databases
KDE	Kernel Density Estimation
LISA	Location Indicators of Spatial Association
NNI	Nearest Neighbour Indicator
O	Origins
O/D	Origin/Destination
REST	The Representational State Transfer (REST) architecture
SVG	Scalable Vector Graphics
SOAP	Simple Object Access Protocol
UI	User Interface
UTC	The Current Unix Timestamp
W3C	World Wide Web Consortium
XML	Extensible Markup Language

1 Introduction

1.1 Introduction

The development of location-awareness technologies, e.g., Global Positioning System and tracking devices, have facilitated the possibility to collect a large amount of trajectory data of moving objects. Therefore nowadays travel behaviour studies in which floating car data (FCD), mobile phone data, or social networks data have been used are actual. In addition, in the digital age crowd source mapping is possible and volunteered mapping communities exist, e.g., OpenStreetMap. There are many data mining methods and algorithms which are applicable for large data sets. The domain of cartography of spatio-temporal data for the movement data is evolving (Tsou, 2015) and there exist challenges in employing new analyse methods and in visualising the data. In addition, the visualization and analysis processes in geovisualization support each other, in a way that they evolve simultaneously. Thanks to that of (geo-)visual analytics very wide scale exploration in data is possible with data mining methods and interactive filtering tools. Different available data sources can be combined and used to complement each other in analysis. When supplementing FCD with open-source data of venues, then diverse human behavioural patterns in urban environment can be studied. Human activities are to the interest of for the city planners.

In urban areas FCD of taxis is collected and they can be used for analysing travellers‘ behaviour, so that the popular places of the city can be detected and their activity’s time patterns can be discovered. When the FCD is combined with point of interests alias venues then patterns can be studied for regions of interests. In this thesis, taxi travellers‘ origin and destination points were extracted from one week FCD (17.05.2010 – 23.05.2010) and combined with venues derived from OpenStreetMap data to explore human behaviour with the means of data mining and geovisual analysis methods.

The study area of this thesis is the metropolis Shanghai in China. Shanghai is located between latitudes 31°22‘ N – 31°27‘ N and longitudes 120°52‘ E – 121°45‘ E. Shanghai consists of the city proper and suburban districts, and Chongming county. The proper districts with the state of 2010 were Huangpu, Nanhui, Luwan, Xuhui, Changning, Jing’an, Putuo, Zhabei, Hongkou, Yangpu, Minhang, Baoshan, and suburban districts were Pudong New Area (Nanhui), Jiading, Jinshan, Songjiang, Qingpu, Fengxian (Li et al., 2014).

1.2 Problem Definition

Different people take taxi in many different trip purposes and there exist rhythm of taxi takings. These purposes as well the demand for taxi and the dynamics of trips end and start locations is changing in a day. Taxi takings occurring in different locations and time have interesting patterns. Studying the dynamics of travellers activities can be helpful in digging out knowledge that can enhance the quality of service in urban environments as in smart cities.

There are two main goals in this thesis, firstly, to use exploratory data analysis methods to explore the FCD and find O/D densified clusters. The second goal is to study web visualization methods and web mapping functions to build up a user interface to represent the results of data processing for the visual analysis purposes.

According to Tobler's "first law of geography", "everything is related to everything else, but near things are more related than distant things" (Tobler, 1979). High density of taxi trips’ origin or destination events often refer to a popular place. An EDA method, clustering is grouping

data based on similarity measures. Similarity in space can be measured by the distance. The presence of high incidence rate of taxi travellers origin or destination events around a specific location during a certain time period refers to an occurrence of clustering. It means, a high probability for the existence of a hot spot in that location.

In an urbanised city like Shanghai highly intensive taxi activities take place. It is possible to detect travellers' origin/destination (O/D) activity concentrations in several scales of a city, e.g., clusters as big as the city area; clusters taking place over several city blocks; clusters occurring on street level. In this thesis hot-spots are defined as in some areas at certain time of a day taxi travellers' trip start or end events are highly-concentrating compared to other parts of the city. If some events, like conferences or concerts, with high participation take place in a city, then hot-spots on a very local scale with very high intensity of travellers coming and/or leaving will take place. In this study, the hot-spots are detected by the clustering and geovisualization methods, and the time patterns of the activities there are abstracted.

1.3 Research Objectives and Research Questions

This study has four main objectives:

- To use taxi FCD to detect popular places in the city. This is detecting high density O/D hot-spots on street scale
- Identify places (venues) near the detected hot-spots
- Exploring temporal patterns for these detected small-scale hot-spots
- Creating an interactive visualization application that helps to discover, analyse and answer the previous questions

The objectives will be studied by answering the following questions.

- How to detect O/D hot-spots from floating car data?
- Near which types of venues do the O/D hot-spots occur?
- When and where do these high intensity O/D hot-spots occur?
- What visualization functions should the application have to reveal the information?

1.4 Thesis Structure

This thesis is opened with an introduction, after which a theoretical overview is given in the Chapter 2, where firstly definitions related to data mining, and theoretical overview to hot-spot detection and the clustering techniques are given, with a focus one density-based algorithm. As well, overviews of the semantic and geovisualization aspects are given.

Then, in the Chapter 3, the practical part will be implemented and a typical clustering method will be utilized on the data set, which contains seven days of taxi travelling in Shanghai in order to find hot-spots. In general, three-step data analysis processes are edescribed (pre-processing, clustering, and post-processing), with the following description of how visualization application components are built.

As a result, the clustering results will be analysed, the hot-spots during the recorded time will be visualised within an interactive graphic user interface, which will be used for visual analysis to describe the taxi travellers' temporally changing activities.

After this, in Chapter 5, the results are discussed, and last but not least, the conclusions of this study and possible future works are presented in Chapter 6. The different parts of my thesis are illustrated with the Figure 1-1.

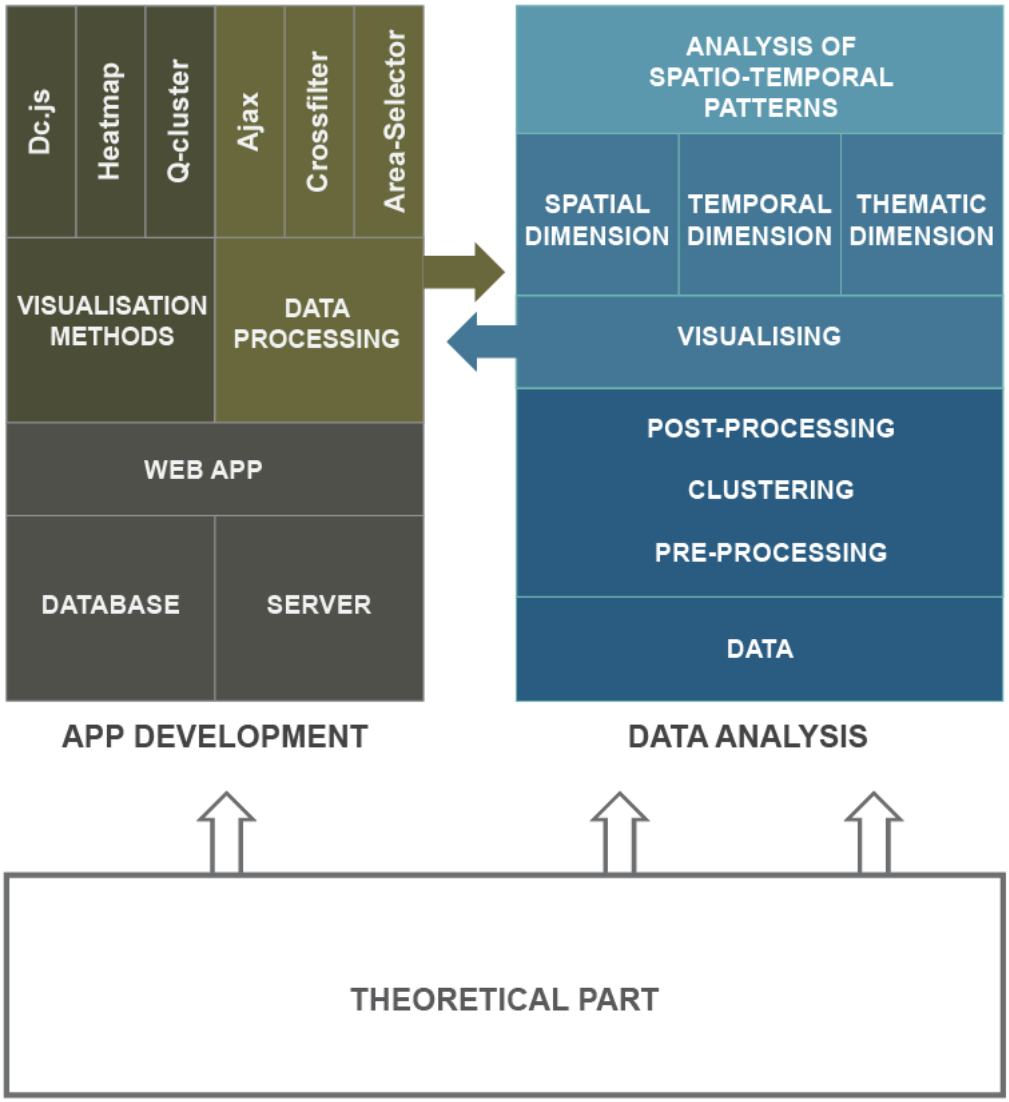


Figure 1-1: The structure of this thesis

2 State of the Art

In this chapter, review of existing research and literature addressing clustering methods and demands for web and cartographic visualization of space-time activities are represented. These topics are being opened in two subchapters respectively. In the first sub-chapter, foremost, the definitions of data mining and geovisualization are opened. Then a short overview to approaches of hot spot detection are given, then later one approach, clustering is further opened. The classification of clustering methods is introduced and then their applicability for spatial, temporal and spatio-temporal analysis are opened/proposed, finally the steps of cluster analysed are suggested. In the second sub-chapter, firstly, the introduction into spatio-temporal visualisation methods is given. That is followed with the description of the concepts of the web visualisation applications.

2.1 Definitions

KDD AND DATA MINING

The knowledge discovery in databases (KDD) is an overall process that is described as the development of methods and techniques for making sense of data. Often KDD is referred as data mining. However, the data mining methods for pattern discovery and extraction are considered to be the core of the process of KDD (Fayyad et al., 1996). The KDD process consists of several steps shown in Figure 2-1 (Fayyad et al., 1996). Firstly, selecting a target data from the data. The second step is preprocessing. The third step is transformations of preprocessed data, which is followed by applying data mining analysis and discovery algorithms for extracting patterns. Finally, the interpretation and evaluation will create knowledge.

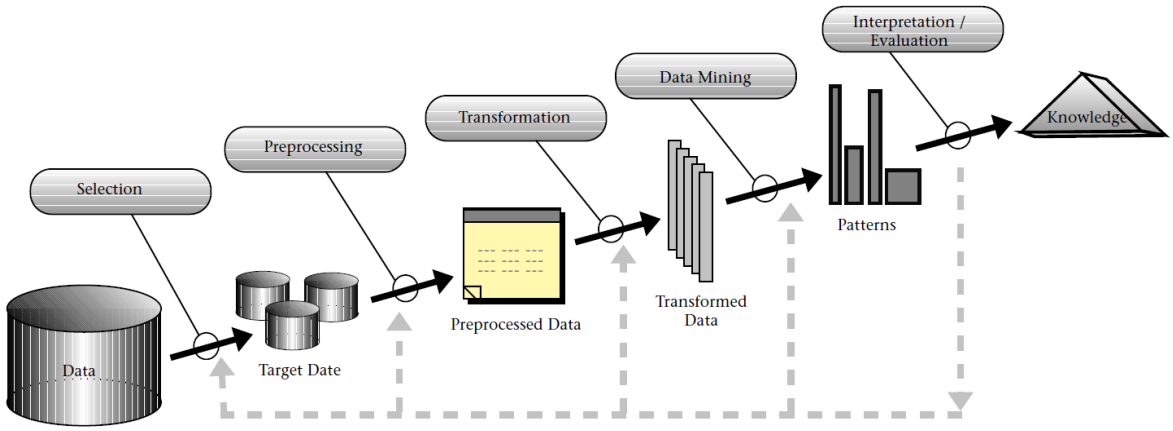


Figure 2-1: Steps of KDD process (Fayyad et al., 1996)

EXPLORATORY DATA ANALYSIS

According to Tukey (Tukey, 1980) the exploratory data analysis (EDA) differs from confirmatory analysis as EDA is about hypotheses generation rather than hypotheses testing. It is an inductive approach of investigating data without preconceptions. In EDA the aspire is towards getting acquainted with data by exploring it. The concept of EDA is associated with the use of graphical representation of data (Andrienko and Andrienko, 2006).

2.2 Approaches for Hot-spot Detection

In crime, traffic, and disease analysis different techniques have been applied for hot-spot detection. From statistical techniques such as location indicators of spatial association (LISA), Moran‘s I, Geary C, NNI, Ripley’s K, kernel density estimation algorithms and other techniques (Hughes et al., 2014), (Kumar, 2009), (Rogerson and Yamada, 2008). On method used for representing hot-spots is Kernel Density Estimation (KDE) to represent hot-spots (Maciejewski et al., 2008). Some of statistical methods rather are used to measure the presence of spatial clustering without determining their locations, others can discover clustering locations. These, more traditional methods are measuring the presence of spatial clustering without determining their locations, including methods of general tests (global statistics), focused tests, and then tests which can detect clustering (Geographical Analysis Machine, Kulldorff’s spatial scan, etc.) (Besag and Newell, cited in (O'Brien, 2009). In many studies the hot-spot analysis is carried out by computing kernel-density-estimation (KDE) based surfaces of the distribution of incidents. Several desktop software, ESRI ArcGIS and QGIS offer tools for generating KDE surfaces. In hot-spot analysis KDE has been useful for detecting overall trends in spatio-temporal datasets and other methods have been used for exploring certain hot-spots (Wolff and Asche, 2009).

In EDA clustering has been frequently used for detecting dense regions in data space (Suh, 2012). Cluster analysis belong to the KDD methods, and it is helpful in finding patterns in large collection of data by grouping data items based on their similarities in data space. Determinations is done based on relationships between data items by finding items that are closer to each other. In the next chapter, cluster analysis techniques are analysed more in detail.

2.3 Clustering

2.3.1 The Classification of Clustering Methods

Clustering or segmentation means assigning a data set into meaningful grouping or classes, that describe the data (Miller and Han, 2009). In general, the assignment is done based on the similarity rule. There exist many clustering algorithms which define the rule for examining relationships between the data items. Clustering belongs to unsupervised learning methods, in which case some influence of subjectivity is added by the analysts‘ choice of parameters. The major categories of clustering techniques include partitioning methods, hierarchical methods, density based methods, and grid-based methods (Miller and Han, 2009).

PARTITIONING METHODS

Partitioning methods are also called representative-based methods. Here a set of representative points are selected and similarity will be calculated based on them. These methods results in partitions of the data, where each partition represents a cluster. They advance by an iterative technique that has the purpose to improve the partitioning as it can move objects from one group to another. The discovery of high-quality clusters in the data is dependent on discovering the high-quality points (Aggarwal, 2015). It is important to notice that this kind of method can only find spherical-shaped clusters (Miller and Han, 2009). For spatial clustering, partitioning methods based on clustering by distance are used. Algorithms of

partitioning methods are *k*-means (Lloyd, 1982), *k*-medoids (Rousseeuw, 2009), CLARANS, and the EM algorithm (Dempster et al., 1977).

***K*-means algorithm** – the sum of the squares of the Euclidean distances of data points to their closest representatives is used to quantity the objective function of the clustering. The weakness of *k*-means algorithm does not detect well when the clusters are of arbitrary shape. On the other hand, Mahalanobis *k*-means algorithm can fit to varying cluster densities (Aggarwal, 2015).

***K*-medoids algorithm –** is similar to *k*-means, but the representatives are selected from the database.This is helpful when handling outliers, and in situations in which it is difficult to define a central representative function because of having many data sets with varying lengths. This method is slower than the *k*-means algorithm, but has a greater applicability (Aggarwal, 2015).

HIERARCHICAL METHODS

Hierarchical methods create a hierarchical decomposition of a given set of data objects. Depending on how hierarchical decomposition is carried out, this method can be divided as agglomerative (bottom-up) or divisive (top-down). Respectively, the examples of them are AGNES and DIANA (Miller and Han, 2009). Another example of hierarchical algorithm is Nearest Neighbor Hierarchical Clustering (NNHC) offered in CrimeStat program (Levine and CrimeStat, 2004). The disadvantage of hierarchical methods is the once made decision of grouping cannot be undone. However, in other algorithms this problem has been solved (BIRCH and Chameleon) (Miller and Han, 2009). For the hierarchical clustering algorithms the level of classification or abstraction have to be chosen.

DENSITY BASED METHODS

The general idea of density based algorithms is that the objects to be clustered can be depicted as points in a matrix. Then, through an iterative process objects will be assigned to clusters in a way that a cluster is formed by area of high density of objects, meanwhile the cluster is separated from other clusters by areas of low density (Han et al., 2009). Typically, density-search algorithms need parameters for defining the threshold for the distance defining whether the pairs of points belong to the cluster region or not. Some examples of algorithms are DBSCAN (Ester et al., 1996), OPTICS (Ankerst et al., 1999), DENCLUE (Hinneburg and Keim, 1998).

GRID-BASED METHODS

Grid-based methods employ the definition of a set of grid-cells for which the objects assinment is done. The density for the each cell is computed. The advantage of grid-based methods is its fast processing time, since data objects are aggregated into the grid values and no distance computations is needed. Clustering is performed on summaries and not individual objects. Finally, it is easy to determine neighbouring clusters. The disadvantage is that shapes of clusters are limited to union of grid-cells (Miller and Han, 2009). Grid-based methods require the specification of density parameter, which is not always intuitive from an analytical perspective (Aggarwal, 2015). Examples of algorithms are STING (Wang W), CLIQUE (Agrawal et al., 1998), and ST-AGRID (Fitrianah et al., 2015) for spatio-temporal data.

2.3.2 Clustering Methods Used for Spatio-Temporal Tasks

According to the literature review, in recent years there exist many approaches using clustering techniques for the different purposes, such as data aggregation in order to reduce the data

complexity, or hot-spot detection. Clustering methods have been applied for spatial, temporal and spatio-temporal analysis, as well for data thematic attributes (place semantics, the trip's speed). For finding hot-spots in space and time, researchers have used following clustering techniques: Chang et. al tested *k*-means, agglomerative hierarchical clustering and DBSCAN for detecting hot-spots of taxi trips origins. They found out that each of them has advantages and disadvantages and there is a need for finding an algorithm having the advantages of all these clustering algorithms. Since clustering methods performances differ depending on the data distributions, they note that it can be challenging to find out an algorithm suitable for the data with varying distribution (Chang, 2008). Some of the interesting studies of efficient clustering, in the context of databases and large data sets, are DBSCAN (Zhao et al.), OPTICS (Zhao et al.), DENCLUE (Zhao et al.), CLIQUE (Zhao et al.) and WaveCluster (Wolff and Asche). Other researchers used clustering methods in mining the semantics of popular origin-destination flows (Zhang et al., 2012). Ehmke used *k*-means clustering for grouping road links of similar speed (Ehmke and Mattfeld, 2010).

2.3.3 A Density Based Spatial Clustering of Application with Noise (DBSCAN)

Density-based clustering algorithms were considered by the author of this work to be suitable in context of this work of finding hot-spots in space and time. From them, an examples is DBSCAN. A DBSCAN algorithm proposed by Ester et. al (1996) has been widely used in spatial data mining. This algorithm detects regions with high density and enables detection of clusters with arbitrary shape, while regions with sparse data will be filtered out as noise. It requires two input parameters, which are epsilon (*Eps* or ε) and the number of minimum points (*minPts*) required to form a dense region. In comparison to some other clustering algorithms, e.g., *k*-means, it does not require the prior knowledge of the number of clusters (Ester et al., 1996).

The density of a data point is defined by a number of points that lie within a radius *Eps* of that point. The data points will be classified into *core*, *border*, or *noise* points (Figure 2-2). The core point, if a data point contains at least *MinPts* data points. Border point, if a data point contains less than *MinPts* points, but it also contains at least one core point within a radius *Eps.* Finally, noise point if a data point is neither a core points nor a border point.

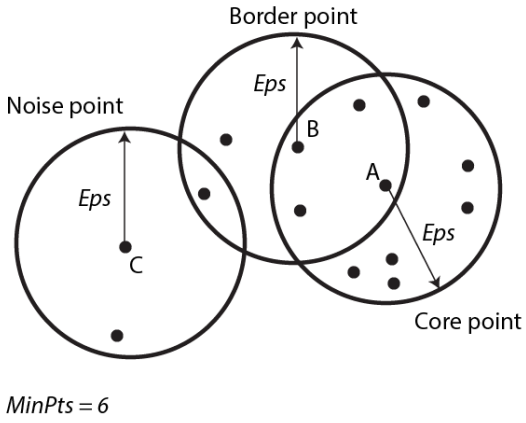


Figure 2-2 Core, border, and noise points (Ester et al. 1996)

DBSCAN detects dense regions from data space, and sparse regions will be assigned as noise. An important note is that DBSCAN algorithm joins regions with different densities into one cluster if they are close to each other. This means, DBSCAN is able to detect clusters in which the density is higher than the threshold defined by its parameters *Eps* and *MinPts*. For separating regions with different densities, recursive call of algorithm with higher value for *MinPts* could be run (Ester et al., 1996). The strength of DBSCAN is its capability to detect places at different granularities and fast performance, but the weaknesses are its vulnerability to altitude variations (Nurmi, 2009). In practice, the choice of parameters for DBSCAN has been found experimentally and it depends on the specific research question and on the distribution of data (Miller and Han, 2009).

DBSCAN clustering method has been used for extracting travellers stops from the trajectories (Vent, 2014). Several authors have developed further adjustments of DBSCAN algorithms for adapting the needs of specific spatial or spatio-temporal clustering problems. For spatial analysis cluster algorithms have been applied to point element, and line element clustering problems. Some examples are the usage of C-DBSCAN (Wang et al., 2014), ST-DBSCAN (Birant and Kut, 2007), P-DBSCAN (Kisilevich et al., 2010). What comes to DBSCAN incapability of performing well in detecting clusters with varying densities, further developments are proposed to address this problem by Pei et. al (Pei et al., 2009).

2.4 Semantics

The place identification is a data analysis task with the aim to analyse measurements and to identify areas that are meaningful to the people (Nurmi, 2009). According to Nurmi (2009), locations derived from devices are represented with coordinates, though, people do not refer to locations using coordinates, but they use semantic descriptions (at home, supermarket), which are denoted as *places.* Relph (Relph, 1976) defines a place as a combination of a physical setting, activities supported by the place and the meanings attributed to the place.

Volunteered geographic information (VGI) stands for the citizens who collect collaboratively geographic information (Goodchild, 2007). One outstanding example of VGI is the OpenStreetMap (OSM) project. Social media platforms collect and offer user-generated data. Some examples of these kind of platforms that collect user-generated geo-data are Foursquare, Flickr, and Twitter. The digital media data can offer Points of Interests (POIs) which could be added as the semantic component to the movement data when analysing the reasons or meanings of human travels.

The place identification with user collected digital data is a challenging task. One aspect to consider is that voluntarily collected data or user-generated data generally has varying level of consistency in different geographic regions. That is dependent on social media users' behaviour or the popularity of VGI activities. Therefore, there may be many places that are not recorded. Further challenges of place identification are related to uncertainty issues when tagging movement locations with semantic labels. In some places the POIs are located densely, then many different venues or places could be eligible.

In literature different methods for determining the semantics of places have been used. Zhang et al. mined the semantics of popular O/D flows by dividing the study area by a grid and applying the semantics of main building to it. Then they created feature vectors of visits frequencies for each cells and did *k*-means clustering for them (Zhang et al., 2012). The social media data has been used for adding semantic information by Andrienko et al. (2013b) and Krueger et al, (2014). Andrienko et al. used GPS, GSM and Twitter data for detecting semantics for places. They clustered human activity areas, created a buffer zone for each cluster and computed

count of visit for each point of interest falling into the cluster zone (Andrienko et al., 2013b). Another example of the user-generated data usage for the context enrichment is the Krueger et al. work in which the FCD and Foursquare data are linked for detecting events automatically (Krueger et al., 2014). In addition, Krueger et al. consider also the uncertainty visualization methods.

2.5 Visualization

At first, the definitions of visual analysis and geovisualization are given. Then, the focus of this chapter is on determining the needs and possibilities of visualization application by describing the characteristics of the spatio-temporal geovisualizations. Then an overview of different interactivity functions and Web mapping concepts are given.

2.5.1 Visual Analysis

Visual analysis is based on the facility of graphics (here maps and charts) to represents information with the means of geometry, relations and structures in a form that is directly perceivable. In visual analysis the analysis involves human activity when examining the data display. The visual analysis is more effective with combination of several views to the data components in a unified display space (Andrienko and Andrienko, 2006). Since the human visual system is effective at recognizing patterns, visualization of data is considered a powerful method in knowledge extraction process (Miller and Han, 2009).

2.5.2 Geovisualization

Geovisualization refers to the visual exploration, analysis, synthesis, and presentation of geospatial data for which it provides theory, methods, and tools (MacEachren and Kraak, 2001). With virtual data presentation tools geovisualization can enable interactive mode of data presentation. MacEachren (MacEachren et al., 2004) has developed a conceptual model for the different functions of geovisualization (Figure 2-3). The model has evlovled through several presentations. According to this geovisualization has different functions depending on the purpose of it. Whether the task of visualization is to support exploration over data in order to construct new knowledge, or on the other scale is the task of information sharing with public, for which case the geovisualization is used for presenting results. For the presentation task, the information has been extracted and it needs to be communicated. During the process of knowledge construction the search is being done for detecting patterns and associations in data. According to this model, during the stage of exploration, the specialist has a need for high level and more sophisticated interaction tools with the visualization application, meanwhile on presentation purpose low level interactivity can be offered.

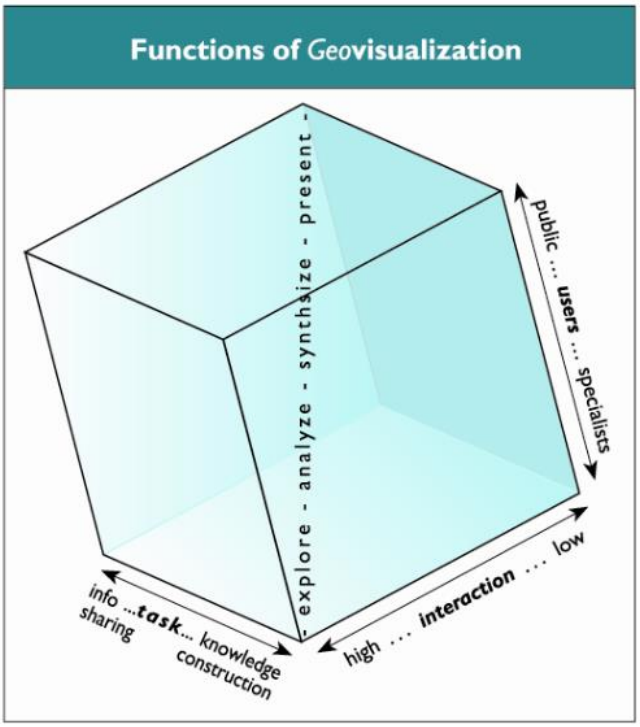


Figure 2-3: The Functions of Geovisualization (MacEachren et al., 2004)

Geovisualization integrates approaches from different fields which are geovisual analytics, cartography, exploratory data analysis, scientific visualization and other related fields described by Schieve in 4 *(*Schieve, cited in Peters 2014*)* (Peters, 2014)

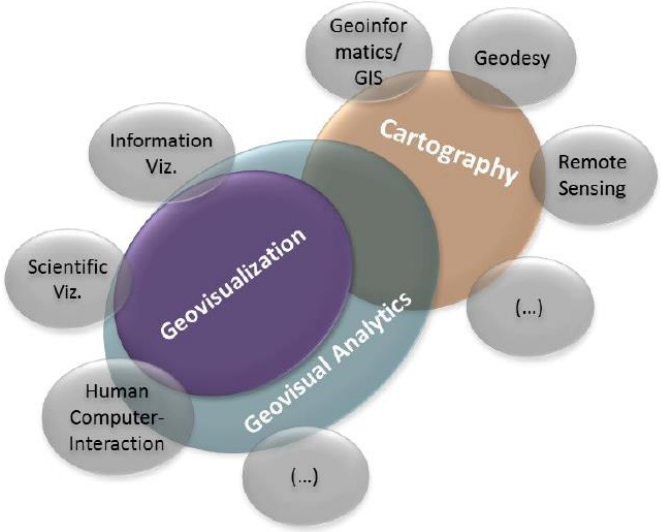


Figure 2-4 How geovisualization interplays with other domains (Schiewe, cited in Peters 2014)

2.5.3 Geovisualization pipeline

The visualization pipeline describes the process of generating visual representation of data (Tominski, 2006). In general, the visualization pipeline consists of four main steps (Figure 2-5), which are data analysis, filtering, mapping and rendering (Dos Santos and Brodlie, 2004). The process begins with data preparation for the analysis. It involves data cleaning from

errouneous measurements. Then, the prepared data will continue further processing which is called filtering. In this stage a selection of data to visualize is made. After this, next stage is data mapping and rendering. Focus data is mapped to geometric primitives (point, line, polygon) and their attributes. These attributes are in the context of cartographic theory analogues with visual variables including color, size, position, shape, etc. (Bertin, 1967), (MacEachren, 1992). The described steps in Dos Santos and Brodlie, 2004 work are considered to be computer-centered involving little user interaction (Tominski, 2006).

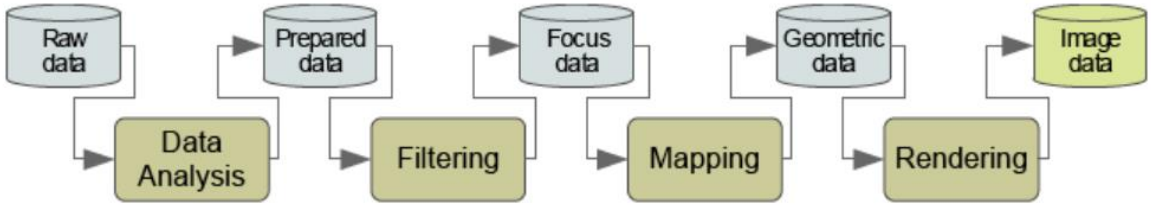


Figure 2-5 The visualization pipeline (Dos Santos and Brodlie, 2004)

According to Thomas Cook and Thomas, 2005 geovisual analytical reasoning process consists of four main steps, which are following: gathering information, creating visual representations, developing insight, and producing results. In those the creation of visual representation means choosing visual forms that aid analysis. The insight is developed through exploration and dynamic visual inquiries. Finally, results will be produced by communicative and story-telling presentations (Figure 2-66).

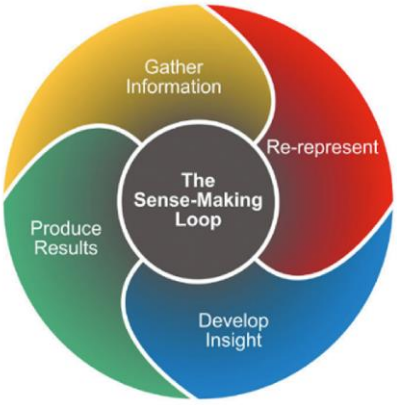


Figure 2-6 The geovisual analytical reasoning process (Cook and Thomas, 2005)

2.5.4 Spatio-Temporal Visualization

There is going on development of spatio-temporal visualization methods and tools for describing temporally changing spatial locations (Tsou, 2015). Spatio-temporal visualisations are used in the context of dynamic transportation (Ding et al., 2015), in studies about migration of people (Kwan and Lee, 2004), (Andrienko and Andrienko, 2008),and in descriptions of natural phenomena, e.g., movement of thunderstorms and lightning (Resch, 2013), (Peters, 2014).

The characteristic of spatio-temporal data is that it is multidimensional, as it can contain a number of attributes. They are two or three spatial coordinates, time coordinate and other many thematic attributes can be added. Andrienko et al. (2003) have given an analytical insight into spatio-temporal data properties and respective tools, as well a classifications and typologies of spatio-temporal data, according to which the characteristics of spatio-temporal phenomena

or objects or events can be described as existential changes, such as appearance or disappearance; as spatial changes, which depict changes in location, size, orientation or altitude; finally, thematic changes are changes in categories, showing changes in qualitative, ordinal or numeric values.

In visual domain the data properties are mapped into graphical elements. The concepts describing spatio-temporal processes can be visualized by „cartographic tools“ of visual variables, and geometric primitives points, lines, polygons (Bertin, 1967). Some examples of visual variables described by Bertin are location, size, shape, orientation, color hue, color value, and texture. Then MacEachren (1992) has extended the list with transparency, crispness, and resolution. An important consideration any visual representation is its reliability and readability. In design-wise a suitable choice of the use of visual variables for showing quantitative and qualitative information need to be made.

For exploring and representing space-time patterns different visualization methods have been used. Representations vary from the scale of static maps to interactive dynamic visualization techniques. Map series can be used for showing changes in time in static representation. In favor of addressing dynamic changes animations can be used. Animation is an intuitive visualization mode for representing changes in time. The drawback of animations in exploratory visualization is that they do not give an overview of the whole situation at a glance, but show the evolvement of a process tools (Brunsdon, 2007). In addition, due to human cognitive learning processes speciality, users can forget earlier frames of animation (Harrower, 2007).

Then, for representing spatio-temporal data, the data attributes can be mapped into dimensional presentation. The examples of different dimensional representations are 2D, 2,5D, 3D, or 4D maps. With the technology development more often high dimensional, both 3D and 4D visualizations have been studied. In 3D mapping the concept of space-time-cube has been adopted in a number of studies and widely used (Hägerstrand 1070 via (Andrienko et al., 2013a)), in which the path of a temporally evolving trajectory is mapped. One development of space-time-cube idea is the 3D wallmap technique that is useful for showing in time (Cheng et al., 2013). Usually the created applications for higher dimensional representation employ the benefits of interactivity or animation. An example of it is 4D visualization in which three dimensions are being used for data mapping and the fourth is for time as an animation (Resch, 2013).

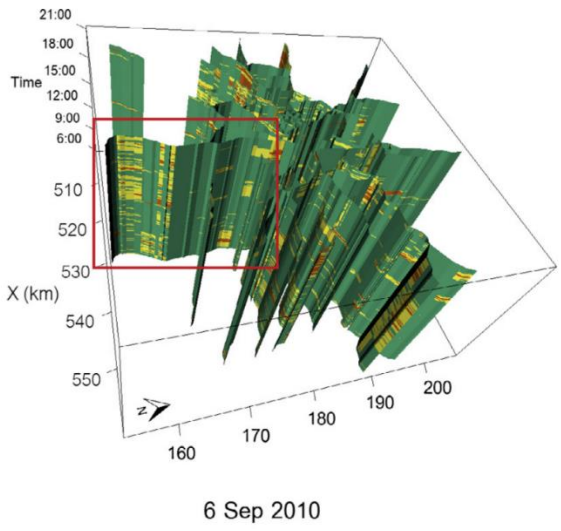


Figure 2-7: A wallmap, a 3D presentation of temporal changes in a network (Cheng et al., 2013)

2.5.5 Interactivity

Nowadays computer-based visualizations enable plotting high amount of data. Cartwright (2001) discuss that in computer technology the interactivity is conducted with interfaces and therefore, in order to create a well usable geospatial representation, the correct design of interface is important. A typical property and limitation of different electronic devices used for conveying information is their relatively small graphical display. On the purpose of offering visualization application users a chance to explore through the „area“ of visual information representation different techniques are used. Interaction between humans and interfaces is kind of communication in which process people can understand how to use and use interfaces as tools and interfaces are designed to respond the need for solving certain tasks. Roth describes as:

> „Cartographic interaction is the dialogue between a human and a map mediated through a computing device. Such a definition acknowledges three components of a cartographic interaction that are necessary for the interaction to be completed digitally: (1) the human (leading to a user-centred perspective on cartographic interaction); (2) the map (leading to an interface-centred perspective on cartographic interaction); and (3) the computing device (leading to a technology-centred perspective on cartographic interaction).“ (Roth, 2012)

According to Cartwright (Cartwright et al., 2001), some of the components of the interaction design of geospatial visualisation application involves are the navigation design (navigation over the information presentation space), the design of how to access data and manipulate it. Depending on the purpose of the visualisation, different interaction techniques can be built for the visualisation applications. There exist several interaction typologies, here the overview for two of them, to Crampton’s and Roth’s taxonomies, are given. Crampton (Crampton, 2002) categorizes different interactivity types into four groups, which are interactivity with the data, interactivity with the data representation, interactivity with the temporal dimension, and contextualising interaction. The methods for each category are shown in Table 2-2.

Table 2-1: Crampton’s Typology of Interactivity Types (2002)

Interaction with the Data Representation	Interaction with the Temporal Dimension	Interaction with the Data	Contextualising Interaction
Lighting	Navigation	Database querying	Multiple views
Viewpoint (“camera”)	Fly-Through	Brushing (Statistical, Geographical, Temporal)	Combining Data Layers
Orientation of Data	Toggling	Filtering (Excluding)	Window Juxtaposition
Zoom-in/Zoom-out	Sorting or Re-expression	Highlighting (Including)	Linking
Rescaling			
Remapping Symbols			

(Roth, 2012) has created a taxonomy of interactivity primitives for map-based visualizations which describe the tasks the user wants to accomplish, the specific actions taken, and the

objects to which actions are being executed. Accordingly, these three types are the objective-based, operator-based, and operand-based approaches. According to him, some interactivity primitives are more often used. From them the more common ones are listed next:

Objective-based approaches:

- Identify
- Compare

Operator-based approaches:

- Brushing – selecting groups of information items for which visual changes will be applied: highlighting, shadow-highlighting, labelling (details-on-demand), or deleting
- Focusing – as providing more detail, as filtering, as being part of brushing in order to emphasise
- Linking – visually relating changes in coordinated multiple views; sometimes considered to be part of brushing.
- Functions for manipulating symbolisation of the cartographic representation
    - Altering map information that is represented (toggling visibility, dynamic projection)
    - Altering the type of cartographic representation
    - Altering the graphic parameters (setting graphical values, altering symbolisation)
- Functions for manipulating the user‘s viewpoint to the cartographic representation (navigation, rotation, distortion, panning)

Operand-based approaches:

- Type specific
    - Temporal
    - Object
    - Three-Dimensional
    - One-Dimensional
    - Two-Dimensional
    - Multi-Dimensional
- State specific
    - Data
    - Visualization abstraction
    - Algorithms/software
    - Representation
    - View
    - Data structure

2.5.6 Web Application Concepts and Tools

CLIENT-SERVER MODEL

Client-server architecture is based on the communication between the client side and the server side. The communication format is determined by the standard Web protocol, Hypertext Transfer Protocol (HTTP). The client initiates the communication by sending a request to server and the server sends a response to the client. On each side different programs are run. In Web communication model, the client runs on a program called browser that sends requests to web server. For that the Uniform Resource Locator (URL) format is used. Similarly works a client/server database access architecture. A database server implements a data manipulation language which functions are to directly access and update the database and present interfaces to client. The simplest client/server database architecture consists of only two components, and it is called two-tier system. However, three-tier architectures are more common. It consists of a Web browser that provides the user interface; the Web server and the application

that requires database access; and the database server and the database itself. (Newton and Villarreal, 2014).

The data access layer of the application is called back-end or server side. The back-end code resides on the server. In the contrary, the front-end code resides in the client. The front-end or client side is any component manipulated by the user. It is an interface between the user and the back-end. In front-end side the code (HTML and JavaScript) are used to create the graphical user interface (GUI) and enable the functionality of the web application via data manipulation and user interactivity functions. (Newton and Villarreal, 2014)

WEB BASICS

Many tools are available for building up web mapping applications. These tools support common web mapping functions, such as interactive visualization, statistical graphs, basic map tools, animation tools. Commonly following World Wide Web Consortium (W3C) standards are used in web mapping: HTML5, CSS3, and SVG. HyperText Markup Language (HTML5) is used to structure content for web browser. It is a tool for specifying semantic structure, or attaching hierarchy, relations and meaning to content (Newton and Villarreal, 2014). Cascading Style Sheets (CSS3) are used to style the visual presentation of DOM (Document Object Model) elements. Scalable Vector Graphics (SVG) is an XML markup language that is used to describe two-dimensional vector graphics for the Web. The HTML, CSS, and SVG integrate with Web standards such as Document Object Model (DOM) that defines the way how documents are accessed and manipulated. (Newton and Villarreal, 2014)

MASHUP

Mashup is a composition technique in which existing reusable parts from different sources are composed together by building up applications. Mashup components typically use technologies as SOAP Web services, RESTful Web services, JavaScript APIs, UI components, CSV and XML files, and W3C gadgets etc. Web applications can be developed by using existing APIs or Web services (Mashup components, Florian Daniel, Matera). In a Web-based mashups the users' browser is used to fetch and manipulate the different components. Mashup applications are created by building up the user interface (UI), then the web services layers and the data layers are included. (Daniel and Matera, 2014)

AJAX

Asynchronous JavaScript and XML enables applications to receive asynchronous data from web servers in the background without interfering with the display of the existing page, e.g., AJAX is using URL to request the data that can have different formats (CSV, JSON).(W3Schools, 2009)

DATA FORMATS

Different data formats exist and can be used for including data into mashups. Typical examples are CSV, JSON, or GeoJSON. The JavaScript Object Notation (JSON) is an alternative to XML that is a syntax for storing and exchanging the data. XML stands for Extensible Markup Language. GeoJSON is a specific JSON format used for describing geographic data structures (W3Schools, 2009).

JAVASCRIPT APIs

API stands for application programming interface. It provides tools for building up software and applications. When using APIs Web Client makes request to the API server in order to call out functions. In recent years, many JavaScript API client libraries have been created, which are suitable tools for building up web map applications. Some typical map APIs are the Google Maps API, the Leaflet API, and the OpenLayers API.

D3.js

D3.js is a Data-Driven-Documents visualization library for creating interactive and dynamic graphics and charts, which utilizes the means of SVG, HTML and CSS. D3.js is able to bind data elements to the Web page's DOM and is therefore compatible with other DOM frameworks. Data formats it uses are, e.g., JSON, GeoJSON, and CSV (Newton and Villarreal, 2014) The advantage of D3.js is that it enables a control over each detail of created graphics, on the other hand, it has a long learning curve. Based on D3.js other libraries are built and many Leaflet plugins take advantage of it.

Dc.js

Dc.js is a Dimensional Charting JavaScript Library that is based on D3.js, and it offers higher level functions to draw more standardized graphics. As well, it has a native support for Crossfilter.

Crossfilter.js

Crossfilter.js usage purpose is data filtering and grouping. It empowers exploring large multivariate datasets in the browser. The advantage of Crossfilter.js is that it enables interactive filtering in client side, however, in some cases of large data sets it is more useful to implement it in back-end.

CartoDB

CartoDB is a Software as a Service (SaaS) cloud computing platform that provides GIS and web mapping tools for display in a web browser.

LEAFLET

Leaflet is an Application Programming Interface (API) that provides sets of classes and functions that can be specifically used for making web maps. Leaflet provides JavaScript library and it is open-source (Agafonkin, 2011). Leaflet is extended with numerous plugins, some of them are q-cluster and leaflet-heat.

Q-cluster.js (Hallahan, 2014) creates visual clusters of points on the map. The goal of this algorithm is to avoid cluttering of visual representations when zooming between different map scales. It also offers D3.js categorization for attributes of data.

Leaflet-heat.js (Agafonkin, 2014) creates *heatmap*-like graphics by drawing each point with defined radius, and blur level. Then, the each point has a certain opacity level. Now, in dense regions the points will overlap and therefore their opacity level will change and more intense colour will be rendered. In addition, it possible to define a colour-scheme for each opacity level. The leaflet-heat is suitable for data-intensive applications since it clusters points into a grid for performance. This presentation adapts to different zoom levels. Leaflet-areaselect.js offers functions for drawing a rectangle for defining the bounding box of the map area which size and shape can be adjusted by the user.

3 Methodology

In this study, detection of street level hot-spots are of interests. The origins and destinations of the trips are not uniformly distributed in the space and time. On the map can be seen that on the street level there occur areas with both high and low concentration of events. In the space and time events take place with different variations in intensity. For detecting hot-spots of taxi travellers origin and destination events, the interest is in detecting space-time subsets in which the concentration of events is higher than in other. During different parts of a day, travellers can have different reasons for making trips. The assumption is that travellers start and end taxi trips close to the places they plan to visit or have visited. After detecting hot-spots location, the nearby venues will be detected for each hot-spot. Venues are buildings or places with desirable or useful features for travellers, e.g., hotel, conference centre, restaurant, shopping mall.

3.1 Data Description

In this study, data sets used are a taxi GPS data (FCD) and semantic data. The semantic data set was used to associate the hot-spot areas with places near them. The semantic data is described further, in Chapter 3.2.3. One week FCD is used in this study over the period of 17.05.2010 – 23.05.2010. The GPS data is collected in Shanghai, China. The FCD set contains GPS location traces of 1690 taxis. The one week data extracted from the database comprises approximately xxx records which are stored as point data. This data was imported to a NoSQL database MongoDB. Data is stored in MongoDB in JSON format. Each data record represents one point that has next attributes: a unique ID, timestamp, company, car id, longitude, latitude, speed, altitude, car status, GPS effectiveness and orientation (Figure 3-1Figure 3-1). From them the attributes interesting for this study are the ones that describe taxi pick-up or drop-off events: spatial attributes latitude and longitude; temporal attribute timestamp; non-spatial attributes car status and car id (Figure 3-2). The car status is in boolean format showing if the taxicab is with a passenger or empty, has values 1 or 0. The timestamp is in UTC format that is presented in seconds from 01.01.1970. In MongoDB timestamp is wrapped with a mongo ISOdate() function which enables calculations with time and date.

.

```
{
  "_id" : ObjectId("53319c06653603217fbdd010"),
  "timestamp" : ISODate("2010-05-22T04:49:16.000Z"),
  "company" : "QS",
  "car_id" : 10003.0000000000000000,
  "lon" : 121.4239120000000000,
  "lat" : 31.3497069999999990,
  "v" : 0.0000000000000000,
  "altitude" : 104.0000000000000000,
  "car_status" : 1.0000000000000000,
  "gps_effectiveness" : 1.0000000000000000
}
```


Figure 3-1: Data description with variables


```
{
  "_id" : ObjectId("53319c06653603217fbdd010"),
  "timestamp" : ISODate("2010-05-22T04:49:16.000Z"),
  "car_id" : 10003.0000000000000000,
  "lon" : 121.4239120000000000,
  "lat" : 31.3497069999999990,
  "car_status" : 0.0000000000000000
}
```


Figure 3-2: The relevant variables for this study

3.2 Data Processing

In this Chapter, it is firstly described how data is prepared for applying the cluster analysis for detecting hot-spots. Secondly, the DBSCAN algorithm is applied to the data. Thirdly, the sub-collections of data is collected together, a semantic label is tagged for hot-spots, and finally data is prepared for the visualization via the web application.

3.2.1 Pre-Processing

In this stage, data is imported to the MongoDB database after which it is sorted and cleaned as a preparation for selecting the target data of O/D points from the data set. After this O/D points are grouped into temporal intervals.

In taxi data set, taxi pick-up or drop-off events are the locations of travellers' origins and destinations. These events are described by the change in taxi status. The pick-up event takes place when taxi status changes from 0 to 1, and the drop-off when 1 to 0.

The preconditions for O/D extraction query was a sorted and cleaned data set. The data set was initially sorted by timestamp, but not by fleets. Therefore the data set was firstly sorted by fleets (field *car_id*). Sorting can be a time demanding operation with MongoDB. One reason for it could be that mongo aggregation pipeline that was used for it requires writing results into a new collection, but while MongoDB is fast in reading, it is much slower in writing procedures. One strategy for enhancing the speed is indexing. Therefore the field *car_*id was indexed before the sorting operation. As a result, the order of data records was sequential to the fleet type (*car_id*) and temporally (*timestamp*). This sequence revealed that from time to time there existed errors that several pick-up or drop-off points occurred very close in time and space, the time difference for these events were less than a minute. These occurrences were caused by technical error and they needed cleaning. Cleaning means eliminating errors such as car status changes taking repeatedly place in a very short time intervals. The FCD precision on temporal scale is 10 seconds. It means that during the duration of a trip after every 10 seconds a value for the *car_status* is recorded. In the data set, there are some incorrect status changes which take place during very short temporal intervals. These cases are described with a very short subsequence of *car_status* changes from values 0 to 1 or opposite. Cleaning the data means smoothing out isolated instances of values 1 or 0, to later avoid extracting O/Ds with as short intervals as e.g., 40 seconds.

The idea behind extracting destinations or origins is based on a comparison of the current and previous documents and looking for time moments when changes appear in the field *status.* This is the moment from which the traveller entered or left the taxi car. The status of the taxi with a passenger is 1 and 0 denotes to empty taxi cub. As for extracting the pick-up points, the following criteriums were used for detecting the pick-up and drop-off events: for the pick-up event the occurring change needs to be from 0 to 1, and for the drop-off event from 1 to 0 (Figure 3-3).

```
if (document_curr.car_status - document_pre.car_status < 0)    -  ( Destination )
if (document_curr.car_status - document_pre.car_status > 0)    -  ( Origin )
```

Figure 3-3: The extraction of origin/destination points from the large data of FCD points

In this way, from one week of FCD set origins and destinations were extracted. The taxi travellers' origin-destination event is represented as a point in space-time continuum. In this work, detecting geographic areas of higher this kind of taxi activities, which are related to travellers trips start or end is solved with cluster analysis. For 7 days there were 86 875 666 records, from them 399844 origins and destinations.

In dealing with time concept the author of this study chose the approach of dividing time period into intervals for which the spatial clustering algorithm will be implemented. So, next step for preparing for the clustering was to extract data points from each temporal interval, for which clustering was later applied. The testing with DBSCAN revealed that during the one hour interval a few O/D concentrations occurred on the street level. Most of the places have just a little O/D events and a few areas stand out. During three hour intervals, more O/D events occurred and the peaks and low areas were more distinct, therefore, for this study the three hour interval was chosen (Table 3-1).

Table 3-1: 3-hour intervals over 24 hours of a day

00-03	**03-06**	**06-09**	**09-12**	**12-15**	**15-18**	**18-21**	**21-24**

Finally, the prepared data for clustering include next information: date and time of the measurement, latitude, and longitude (Figure 3-4Figure 3-4).


```
{
   "_id" : ObjectId("53319c06653603217fbdd010"),
   "timestamp" : ISODate("2010-05-22T04:49:16.000Z"),
   "lon" : 121.4239120000000000,
   "lat" : 31.3497069999999990,
}
```


Figure 3-4 An example of a data record that is prepared for clustering

3.2.2 Clustering

In this step of data processing, the spatial clusters in data are detected during each temporal intervals with clustering algorithm DBSCAN. This is clustering for two dimensional data, for spatial dimension of x and y coordinates.

For event detection the aim was to detect the gatherings of origins or destination events that occur during short time and in concentrated area. Therefore, a smaller range of radius for the buffer zone (*Eps*) for the clustering algorithm and time interval of three hours for clustering. Clustering of taxi events might take place during relatively short time near conference centres if there is any event taking place. Concentration of taxi stops can appear intensively just before the start of the conference or during longer time-range in a day, since people may come and go in between the conference events. Since for finding hot-spots in a city, only the areas with highest densities were of interest. It means points from the low density areas were out of interest and they do not need to belong to a cluster.

There is a minimal previous knowledge about the number of outcome clusters, even more, during different time intervals the number of hot-spot areas is expected to vary. The assumption is that number of clusters could change for each time interval, as well there is a little knowledge of how many clusters could occur.

The summary of conditions for clustering are followings:

- Detecting high intensity gathering of O/D events all over the city
- Low intensity areas of O/D points are not in interest
- Clusters can have varying shapes
- For each interval, there can be a different distribution of data points
- In each interval, there can be a different number of hot-spot areas
- The number of hot-spots for each interval is unknown

The DBSCAN algorithm is considered suitable for the clustering task of Shanghai O/D hot-spots detection, since it does not need an input parameter for the expected number of clusters

and it labels data items with low densities as noise. DBSCAN can detect areas of high densities by defining distance parameter. DBSCAN has also the ability to detect clusters with different shapes, which possibly describes the natural situation of hot-spots shapes in space.

DBSCAN algorithm is available in software packages Weka, ELKI, and R package DCluster. The choice was made for ELKI, since it has a fast calculation speed as it supports creating spatial indexes. ELKI stands for Environment for Developing KDD-Applications Supported by Index-Structures. It is an open-source data mining software written in Java (DatenBanksysteme). ELKI offers data index structures that can offer benefits in performance.

As for the data preparation, from the database the O/D collections for each temporal intervals were queried and exported to Comma Separated Variables file type (CSV*)* files. They became inputs for ELKI DBSCAN algorithm, that was to run for each temporal interval‘ data set. ELKI has a support for geodetic distance functions for using with DBSCAN. The *LatLngDistanceFunction* was used for the distance function with the geomodel of *SphericalVincentyEarthModel*. With these settings the *Eps* radius unit is in meters. ELKI supports spatial indexing that can improve the performance. The *Rstar* index was chosen in this task. Then, DBSCAN requires two input parameters to be defined, *Eps* and *MinPts*. The parameters for DBSCAN were chosen considering the previously listed conditions for cluster detection. According to that, for detecting patterns in local, street level, a small distance for the *Eps* and relatively higher *MinPts* should be chosen. Then, tests of clustering with DBSCAN algorithm with different *Eps* values were done, which were 300 meters, 200 meters, 100 meters, and 25 meters, and *MinPts* with values 10 and 6 (Figure 3-5Figure 3-5).

ELKI writes results of each cluster into separate text files. Therefore, later they had to be collected into one file. A simple Python script was used to do it. Finally, the results of ELKI DBSCAN clustering were collected and written into one CSV file, which was imported into MongoDB. The output of cluster analysis step was a collection with features id, timestamp, cluster_id, interval (Figure 3-6Figure 3-6).

Parameter	Value
rtree.insertionstrategy	Default: CombinedInsertionStrategy
rtree.insert-directory	Default: LeastEnlargementWithAreaInsertionStrategy
rtree.insert-leaf	Default: LeastOverlapInsertionStrategy
rtree.splitstrategy	Default: TopologicalSplitter
rtree.minimum-fill	Default: 0.4
rtree.overflowtreatment	Default: LimitedReinsertOverflowTreatment
rtree.reinsertion-strategy	Default: CloseReinsert
rtree.reinsertion-amount	Default: 0.3
rtree.reinsertion-distancce	geo.LatLngDistanceFunction
geo.model	SphericalVincentyEarthModel v
spatial.bulkstrategy	
time	true
algorithm	clustering.DBSCAN
algorithm.distancefunction	geo.LatLngDistanceFunction
geo.model	Default: SphericalVincentyEarthModel
dbscan.epsilon	25.0
dbscan.minpts	6

Figure 3-5: Parameters in ELKI MiniGUI. *Eps* is 25m, *MinPts* is 6

```
{
   "_id" : ObjectId("53319c06653603217fbdd010"),
   "timestamp" : ISODate("2010-05-22T04:49:16.000Z"),
   "lon" : 121.4239120000000000,
   "lat" : 31.3497069999999990,
   "cluster_id" : "1",
   "interval" : "4"
}
```


Figure 3-6: The output of cluster analysis step

3.2.3 Post-Processing

3.2.3.1 Preparing the Semantic Data Layer

As the goal was to examine to near which places the hot-spots occur, the semantic information was added to discovered clusters. The idea was to detect possible reasons for the hot-spots occurrences based on traveller's behaviour that can be defined by the closeness of places. Therefore the meanings attributed to the hot-spot depend on what is near it. Nearby objects can define the possible activities of humans, e.g., eating in a restaurant or participating an event in the conference centre. The objects, e.g., buildings, parks, in which people can do certain activities were taken as references to travellers' behaviour in this study. They are further called venues.

Semantic data layers for the purpose of finding meanings for taxi trips concentration areas are derived from OpenStreetMap (OSM) data layers *buildings* and *landuse*. In the area of Shanghai province layer *buildings* contains 11506 records and *landuse* 2366 records. Most of them are located in the inner districts of Shanghai. Smaller number of OSM objects are mapped outside inner-districts borders. Not all of the OSM objects were taken into study. These are 6942 unclassified objects and objects which were not meaningful for the goal of this study. Data from two different layers were brought to one. Since in Shanghai the OSM data is mapped quite sparsely, the author completed the list with additional data describing exhibition and conference centres. The list was completed with exhibition centres, convention and exhibition centres, conference centres which were unified into class *exhibition centre.* This data was collected from online portals and maps. After cleaning and mapping the semantic layer contained 3130 objects.

Later some venues of interest were chosen to visualize. The chosen venues from OSM to take in for this study were hostels, hotels, retail, shopping malls, stores, offices, finance, commercial, public, business. OSM offers various types for venues that are similar and can belong to the same category. In addition, in MongoDB the types of records are case sensitive. Therefore, similar types of venues were unified under same category. After unification the resulted classes were hotel, commercial, office, exhibition centre, station, airport (Figure 3-7Figure 3-7).

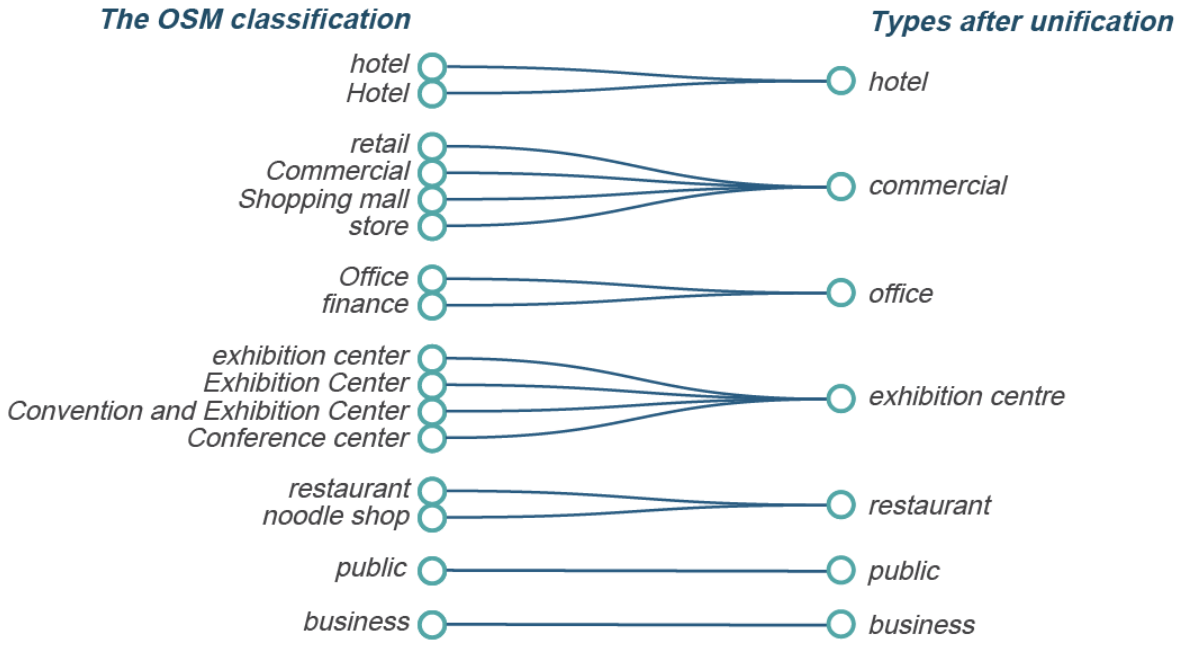


Figure 3-7: The unification of different types of amenities

3.2.3.2 Labelling Clusters

The taxi FCD data contains information about locations of trips‘ origin and destination points that are represented by latitude and longitudes. The place identification of the detected hot-spots was done by discovering the nearest venue around them. For the detected clusters, the semantic labels were added based on the nearby venues. For that the one week collection of destination hot-spot points and one week collection of origin hot-spot points were geo-queried with the venues’ layer. A buffer zone was applied for each *hot-spot* cluster point and venues inside buffer zone detected. Then, from them the nearest venue was selected. Based on this, for each point/record a new field „tag“ was written. The distance of the buffer zone is a short walking distance of 100 meters as the assumption is that pick-up and drop-off areas are very close to the related venues. Since the sample of venues is not covering the city, many points remained unidentified. In the next step, tags with values „unidentified“ and other values different from what is represented in Figure 3-8 were filtered out. As a result, a collection of identified hot-spots was created. Further on, the JSON objects in MongoDB were converted to GeoJSON, which became the input for the web map application. GeoJSON was later rendered on a map in web application.

Variables used for the visualization application are represented in Figure 3-8Figure 3-8. These *are* location variables *longitude* and *latitude*, temporal variable *timestamp*, and semantic variable *tag*. In addition, a variable type was added to describe whether the point describes the *origin* or *destination* point. There are two input data files for the visualization, the *Destinations* and the *Origins.*

```
{
   "_id" : ObjectId("53319c06653603217fbdd010"),
   "timestamp" : ISODate("2010-05-22T04:49:16.000Z"),
   "lon" : 121.4239120000000000,
   "lat" : 31.3497069999999990,
   "tag" : "exhibition center",
   "type" : "origin"
}
```

Figure 3-8 Variables used for the visualization

3.3 The Visualization Application

In this subchapter the description of creation process of an interactive web application for presenting distribution of hot-spots is given. From the related literature in web mapping and cartographic visualization, some important methods and ideas were taken to apply in this application work.

The task is to visualize temporally and spatially changing hot-spot incidents over the city. There are two types of hot-spot, they are traveller's origin and destination places. In addition to this, these hot-spots are semantically labelled according to near which place they do occur. The time range in the study is the seven sequential days on three-hour interval basis.

The interest of conducting the exploration of hot-spots in a city is in discovering patterns and information in different zoom level. With data processing steps the hot-spot areas over the city of Shanghai were revealed. However, these hot-spots were detected on the street scale and therefore they are most interesting to explore on the local scale. Since the the hot-spots could be discovered in small scale and studied further in large scale views, then the visualizations needed to be meaningful when viewing in different zoom levels.

3.3.1 The Structure of the Visualisation Applications

The front-end of the application contains both visualisation and interactivity components. The visualisation approach chosen was to use 2D maps with interactive charts and controllers. The main user-interface (UI) components of this application are the map view which shows the spatial distribution of hot-spots and the map tools panel offering temporal and thematic charts, which are linked together to display information synchronously. The drawer panel for the map tool panel is built with the Cabinet template.

The spatial component of the visualisation is a map view containing two types of thematic layers with the basemap and *building* layer as a background. The basemap is a Leaflet map with OSM tiles that are styled with the CartoDB Positron style that uses grey colours and is a suitable background for thematic maps. The layer *Buildings* is drawn with Leaflet, for which the data in GeoJSON format is called out with D3.js asynchronous request function. The thematic layers are the heat map layer, the donut-chart layer and the pie-chart layer. The heat map layer is built with leaflet-heat.js and it shows different densities for the hot-spots. The donut-chart and pie-chart layers are built with q-cluster.js and they show the point event's categories.

Secondly, on the tool panel the temporal and thematic components are showed as the timegraphs and the categorical charts built with Dc.js. These charts play two roles, they simultaneously present distribution of a specific dimension and enable visual queries by reacting to users' mouse operations in order to render filtered data. The tools panel can be called out by clicking on a button.

Different views are linked together by using associative colours for the chart and map visual elements that are representing either the Origins or Destinations information layers. Secondly, linking has done by using different controllers. Further, different controllers are offered, which are:

The map view and charts are linked via Crossfilter.js, DC.js and Area-selector.js. The linking views means when the selector is brushed in the DC-charts, or the area-selector is changed, the displayed data set will be changed correspondingly both on DC-charts and Leaflet map views. In addition, the cognitive linking between views is achieved by using associative colours for the chart and map visual elements that are representing either the Origins or Destinations information layers.

Other UI components built with Leaflet.js are the map controller enabling to turn on and off the map layers *Origins* and *Destinations;* and the Leaflet *EasyButton* that fires the area-selector, built with the Area-Selector.js, which is a tool for creating a bounding box to select the area of interest from the map view.

The web application is hosted locally by Flask, which will respond to client requests with rendered page. The data contained in the pages are queried from the database (MongoDB) with AJAX codes.

3.3.2 Map I: Hot-spot's Semantics Map

The input for this map are two collections containing data for origin and destination clusters. Each record has features *id, timestamp, coordinates,* and semantic *tag* (Figure 3-9Figure 3-9).

```
"properties":
    {
        "_id" : ObjectId("53319c06653603217fbdd010"),
        "timestamp" : ISODate("2010-05-22T04:49:16.000Z"),
        "tag" : "exhibition centre"
     },
"geometry":
  {
     "type" : „Point"
    „coordinates": [121.45021, 31.249082]
  }
```

Figure 3-9: Excerpt of variables used for the visualization in Map I in geoJSON format

3.3.3 Map II: The O/D Location Map

The input for this map is a collections containing data for origin and destination clusters. Each record has features *id, timestamp, coordinates,* hot-spot type (Figure 3-10).

```
“properties“:
    {
        "_id" : ObjectId("53319c06653603217fbdd010"),
        "timestamp" : ISODate("2010-05-22T04:49:16.000Z"),
        “type“: “destination“
     },
“geometry“:
  {
     "type" : „Point“
    „coordinates“: [121.45021, 31.249082]
  }
```

Figure 3-10 Excerpt of variables used for the visualization in Map II in GeoJSON format

3.3.4 Map III: Temporal Distribution in Popular Places

The input for this map are two collections containing data for origin and destination clusters. Each record has features *id, timestamp, coordinates*, and temporal *interval* (Figure 3-11).

```
“properties“:
    {
        "_id" : ObjectId("53319c06653603217fbdd010"),
        "timestamp" : ISODate("2010-05-22T04:49:16.000Z"),
        "type" : "destination",
       “interval“: “3“
     },
“geometry“:
  {
     "type" : „Point“
    „coordinates“: [121.45021, 31.249082]
  }
```

Figure 3-11 Excerpt of variables used for the visualization in Map III in GeoJSON format

4 Results

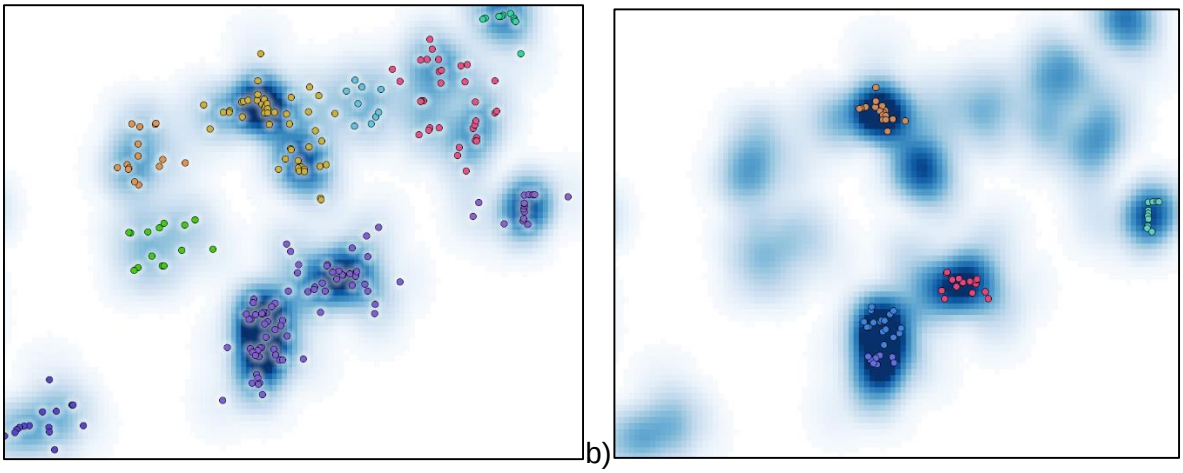
4.1 Clustering Results

4.1.1 The Parameter Choice for DBSCAN Algorithm

In this study, several parameters for *eps* and *MinPts* when clustering with DBSCAN for the taxi floating car data in Shanghai were tested. As shown in Table 4-1 and Figure 4-1, if bigger *eps* was chosen, then the DBSCAN algorithm detected clusters with larger areas and with varying densities. When visualising clusters with parameter settings of *eps*=100 and *MinPts*= 10 with Kernel Density Estimation (KDE) maps in QGIS, then several high density peaks could be revealed inside one cluster. When a smaller value for the *eps* was chosen, then clusters with smaller areas and higher densities were detected and previous high peaks were split into separate clusters, meanwhile with the detected lower density areas detected with first parameter settings were omitted and assigned as noise. Here we can see that with different values for parameters a threshold of density can be created from which higher density areas are detected. The goal was to detect street scale high intensity locations. Since with higher values for *Eps,* the clusters on block level (over several streets) were detected with varying densities, then a small value for the *Eps* was chosen. The *Eps* = 25 m and *MinPts* = 6 were chosen for extracting the hot-spots.

Table 4-1 Comparison among different *Eps* and *MinPts* for DBSCAN with resulting clusters descriptions (The density peaks inside one cluster (DPIC))

Eps	***MinPts***	**Scale**	**Inner Structure**
300 m	10	Blocks level	Several DPIC
200 m	10	Blocks level	Several DPIC
100 m	10	Blocks level	Several DPIC
25 m	6	Street level	Single DPIC



a) b)

Figure 4-1: a) Clustering results with *Eps* = 200m and *MinPts*=10; b) Clustering results with *Eps* = 100 and *MinPts*=10

4.1.2 Spatial Clustering Results for Defined Time Intervals

DBSCAN algorithm was used for detecting high density O/D hot-spots on the street level. The time window for clustering was set to be three hours. The figures and tables illustrate the

distribution of detected clusters for each three hour interval during one week period. For each interval the counts of clusters is given for destination-clusters in Table 4-2 and origin-clusters in Table 4-3; for each interval the comparison between different days is given in Figure 4-3 and Figure 4-4.

During one week (17.05.2010 – 23.05.2010) 4876 clusters of high density O/D areas were detected. In number, there are more destination hot-spots over a week than origin hot-spots, respectively 2726 and 2150 (Table 4 and 5). For origins, the maximum total of a day is on Saturday with 360, then follows Wednesday with 327 clusters, while the average of seven days is 307.14 clusters. The maximum total of a day for destination clusters is on Saturday, 465 hot-spots, with closely following Friday with 432 detected clusters, and the average of seven days is 389.43 clusters.

4.1.2.1 The Comparison of Intervals

Firstly, the description of clusters distribution is given separately for destination-clustering and origin-clustering by comparing different daily intervals. Then, a comparison between origin- and destination-clustering events is made.

The general dynamics of the travellers trips destination-clustering events is illustrated in Table 4-2. Firstly, very low number of clusters were detected during night for the hours of 00:00-03:00 and 03:00-06:00. The numbers of detected clusters is between 0 to 10. For the moring hours of 06:00- 09:00 a sudden rise had taken place and a high number of destination clusters were detected. In average, 81 clusters during this interval occurred during seven days of study. The next interval, from 09:00 to 12:00, was the highest activity period, then the highest numbers of clustered taxi drop-off events took place. In the next interval, 12:00-15:00 high clustering period continued. However, in the afternoon and evening, less clustering events were detected, the average numbers remained close to 44 and 49 for the intervals of 15:00-18:00 and 18:00-21:00. For the late evening it dropped further till 19 clusters.

Table 4-2: The number of detected high intensity hot-spots of destinations on 3-hour interval basis during one week period (blue denotes the low-activity period, red high-activity period)

	Interval	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Avg of all days	Weekday avg
1	00:00 - 03:00	2	1	0	1	1	6	12	3.29	1.00
2	03:00 - 06:00	12	5	5	7	7	10	8	7.71	7.20
3	06:00 - 09:00	78	82	69	82	93	107	60	81.57	80.80
4	09:00 - 12:00	100	97	111	109	108	104	98	103.86	105.00
5	12:00 - 15:00	84	68	78	70	93	73	93	79.86	78.60
6	15:00 - 18:00	52	39	47	39	58	76	32	49.00	47.00
7	18:00 - 21:00	36	42	43	50	50	54	38	44.71	44.20
8	21:00 - 24:00	13	7	15	26	22	35	18	19.43	16.60
	Avg of intervals	48.68								
	Total of a day	377	341	368	384	432	465	359	389.43	380.40
	Total of a week	2726								

The general description of travellers trips origin-clustering events is shown in Table 4-3. In comparison with destination-clusters, there were less variations in the counts of trips origin-clusters. Also, for origins a very low-clustering period occurred during two intervals, when the number of clusters were less or around 10. In comparison with destination clustering, the origins low-clustering period is shifted, taking place on 03:00-06:00 and 06:00-09:00. After the low period in the night and early morning, the number of origin clusters raised during the late morning, 09:00-12:00, till 41. After this during two consecutive intervals the average number of detected clusters was near 50, and in the evening, from 18:00-21:00 and 21:00-24:00 close to 60.

Table 4-3: The number of detected high intensity hot-spots of origins on 3-hour interval basis during one week period (blue denotes the low-activity period, red high-activity period)

	Interval	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Avg of all days	Weekday avg
1	00:00 - 03:00	16	19	16	21	20	42	32	23.71	18.40
2	03:00 - 06:00	11	0	2	4	2	12	8	5.57	3.80
3	06:00 - 09:00	18	12	4	13	12	11	13	11.86	11.80
4	09:00 - 12:00	49	40	45	45	44	34	31	41.14	44.60
5	12:00 - 15:00	49	36	58	58	54	34	65	50.57	51.00
6	15:00 - 18:00	62	50	62	39	46	72	34	52.14	51.80
7	18:00 - 21:00	53	64	70	50	73	69	58	62.43	62.00
8	21:00 - 24:00	13	66	70	70	55	86	58	59.71	54.80
	Avg of intervals	38.39								
	Total of a day	271	287	327	300	306	360	299	307.14	298.2
	Total of a week	2150								

The comparison of the origin-and destination-clustering patterns is illustrated in Figure 4-2. In daily rhythm different patterns for travellers trips origin-clustering and destination-clustering occur. In Figure 4-2, the origins line grows over a day gradually and has the highest peak in the evening at 18.00-21.00. The destinations line grows fast in the morning in hours between 06:00-09:00 and has a very high peak during the interval of 09:00-12:00. Destinations hot-spots appear very intensively during three intervals in daytime, 06:00-09:00, 09:00-12:00, 12:00-15:00 respectively. In the daytime, from 06:00 till 15:00 dominate destination hot-spots over origins. Between 15:00-18:00 both of them are equal, and the origin-hot-spots dominate in the evening and late night, from 18:00-24:00. From 03:00 to 06:00 appearance of each type of hot-spots are almost equally low again.

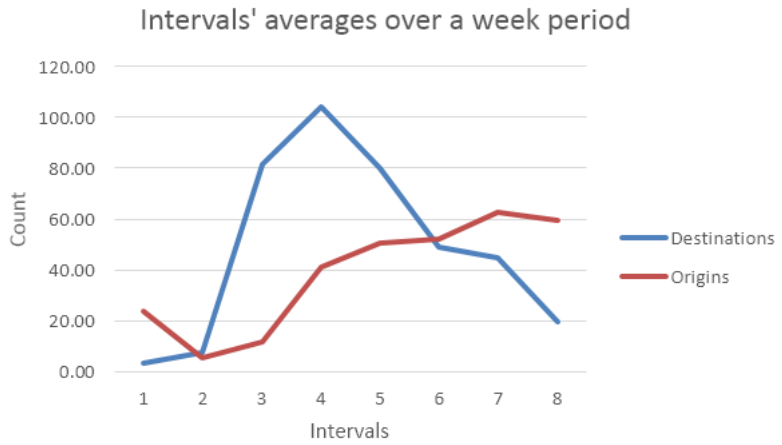


Figure 4-2: Intervals‘ averages over a week period for origin- and destination-clusters

4.1.2.2 The Comparison of Days

The second type of description of clustering events can be done by comparing different days. Saturday differs very significantly from other days by having the highest number of clusterings both for destination and origin events. During the interval of 00:00-03:00 the higher number of clustering events take place on Saturday and Sunday. These events are probably related to the previous days activities. During one week period, there are variations of clustering events in all seven days, but also in between weekdays.

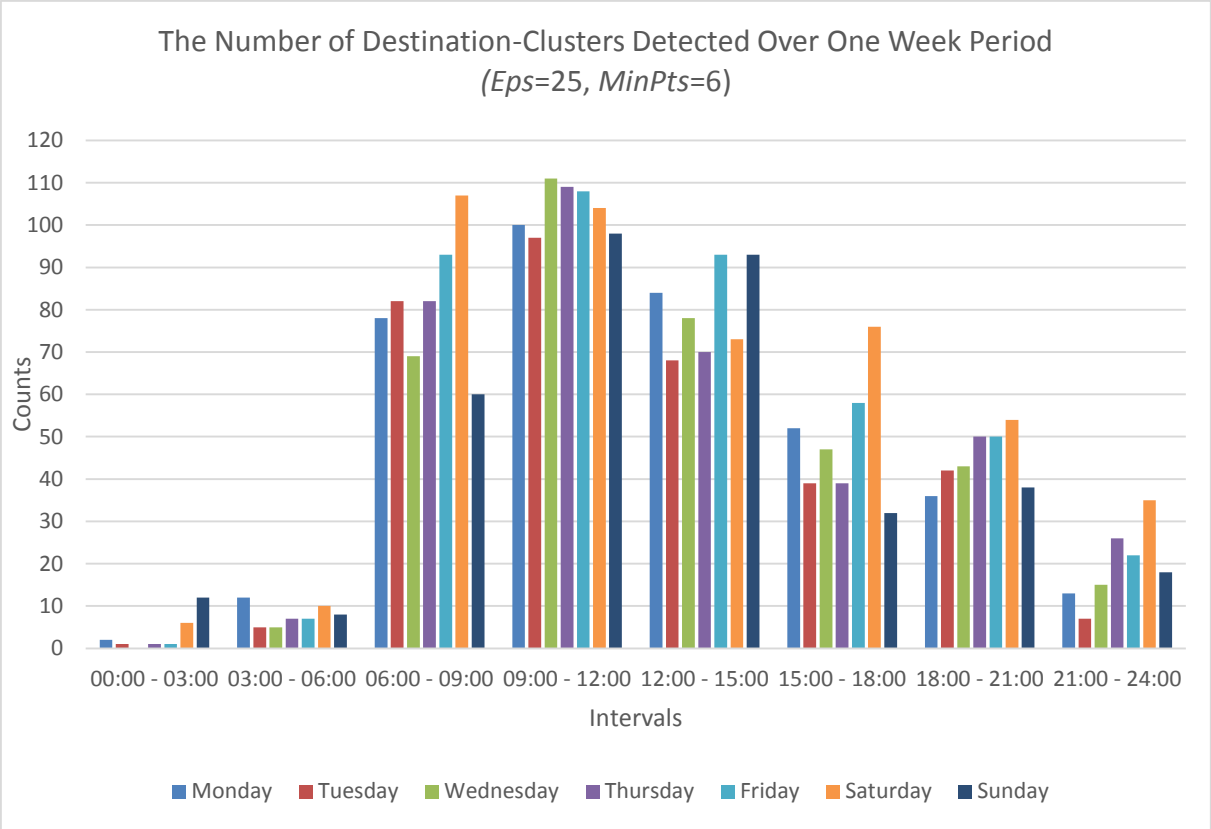


Figure 4-3 The distribution of destination clusters during one week period

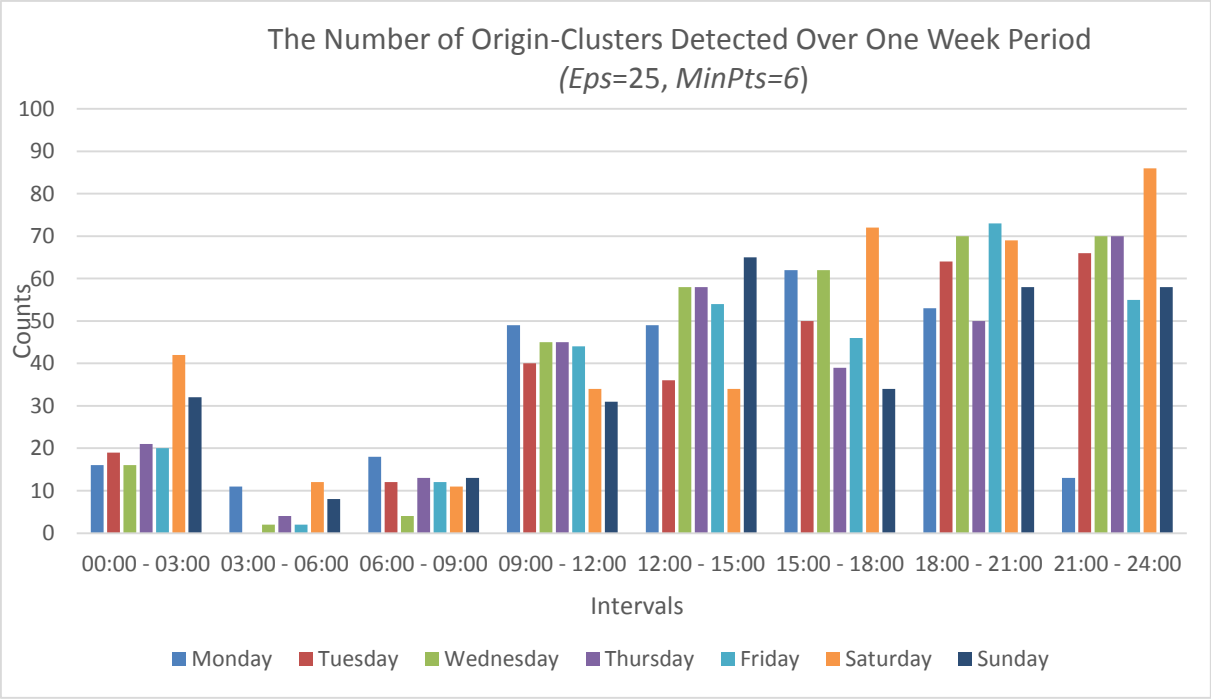


Figure 4-4: The distribution of origins-clusters during one week period

4.2 Visualisation Application

The application has three map views with data filtering tools and statistical charts. The charts show concurrently the statistics of temporal distributions and perform as filters. With them, snapshots of different temporal lengths can be chosen for exploring the data. In time scale choices can be made in daily and hourly levels. In addition to time related information, also semantic context of hot-spots is represented with pie-charts on the map view with the corresponding filtering tool.

The map views are:

- Pie-chart symbol-map with labelled hot-spot locations revealing venue types;
- Heat-map showing O/D hot-spots with temporal filtering tools;
- Pie-chart symbol-map for discovering temporal distribution for hot-spots;

4.2.1 Map I: Hot-spot‘s Semantics Map

Firstly, when starting the exploration in data, the hot-spot’s semantics map gives the overview regarding near which venue the taxi stop events take place. The data layer on the map is represented with pie-charts which denote to the categories of hot-spots according to their semantics. It is possible to search hot-spots according to which type of venues they are taking place at and choose districts of interest. In Figure 4-5 the map view after filtering by venue type is shown. The districts in which these events took place are active in the chart and others are marked inactive. In Figure 4-6 a district Pudong is chosen and in the map view hot-spot’s locations with different categories for venues are shown.

This donut-charts on the map are built with Leaflet q-cluster plugin. In this map each donut shows the categories of venues around it and the amount of points grouped together. On the map tools area, a pie-chart is showing the proportions of events over the all city. When needed, they can be filtered by district.

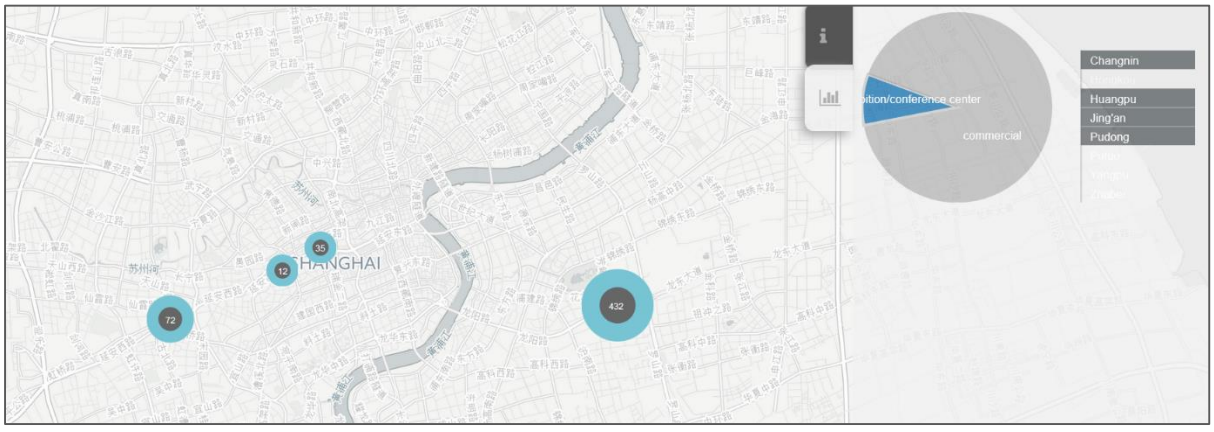


Figure 4-5 Hot-spots near conference and exhibition buildings in city occur in certain districts

Figure 4-6 Hot-spots in Pudong are chosen. Pie-charts describe the semantics of hot-spots

4.2.2 Map II: The O/D Location Map

This visualization combines the two layers of heatmaps, origins and destinations, which are linked with statistical interactive charts that show the distribution of hot-spot on two temporal scales, in days and in hours, respectively (Figure 4-7). These charts enable brushing to change the map view. Then, in this map a user can make a selection of an area and results in charts will be filtered based on this subset. First, the user can use the area-selector for choosing the area of interest and from then on the filtering will be done and results in the charts shown based on the selected area. The user can turn on and off the heatmap layers for origins and destinations. In this map view it is possible to control the two map layers, origin and destination layers, with unified controllers.

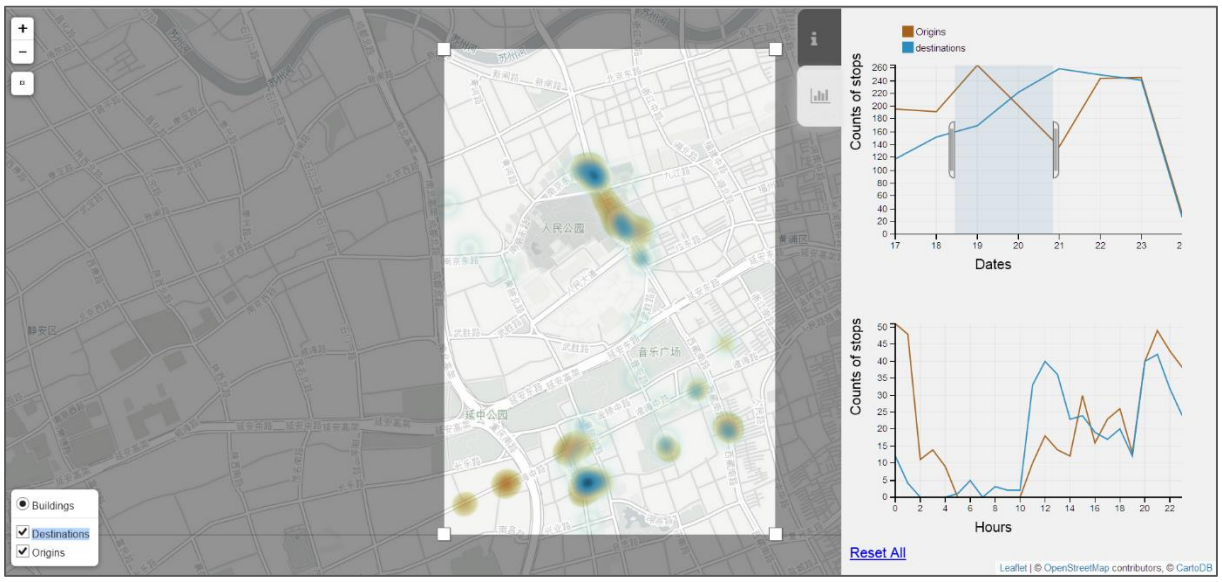


Figure 4-7 Both, the hot-spots of origins and destinations are shown on the map.

4.2.3 Map III: Temporal Distribution in Popular Areas

In this map the temporal distribution of each hot-spot is presented with pie-chart symbols. The pie-charts are built up with q-cluster.js which collects points for different scales and presents different categories with colors. When the mouse cursor is hovered above the cluster, single points that belong to this cluster will appear. Currently this map enables representation of only one of origins or destination layers at a time.

The extracted hot-spots occur in general in small areas with high intensity events and in popular places they occur repeatedly during different time intervals in same locations. Therefore these hot-spots are described with point symbols showing categories for different time intervals. This method can be suitable for comparing the proportions of different intervals for each place (Figure 4-8). The pie-charts are calculated for each map zoom level. However, their symbol sizes do not scale so well for each level and the current configuration is best viewed on a large scale view at zoom level 17. In the tools panel are with pie-charts for days and intervals offering filtering functions. Eight intervals during 24 hours are shown on the interval-chart.

Figure 4-8 Comparison of two intervals 09.00-12.00 and 12.00-15.00 during weekdays. Different colors on pie-charts denote to different time intervals.

4.3 Results of Taxi Travellers‘ Patterns

In this chapter, the results of taxi travellers’ temporal patterns are presented. Here are answered the research questions (2) When and where do these high intensity hot-spots occur; and (3) Near which types of venues they occur. The process of analysis is described with the use of created visualization application. In order to illustrate it, four study areas are chosen. These study areas are taken as typical samples of conference and exhiition centres, shopping, and business areas.

4.3.1 Study Area I

The main venues in the study areas of I and II are conference and exhibition centres. The study area I is near the Shanghai New International Expo Centre (NIEC) in Pudong district.

PROBLEM

1) Different types of hot-spot events take place over the city. We want to know which types of hot-spots do occur in the city level or in the district level. If we detect that an hot-spot of interest, e.g., hot-spots near conference centres, occur, then we want to find out where are their locations and what are their intensities.

2) When we detect hot-spots with high intensity activities, then we want to study them more in detail by choosing suitable temporal granularities for analysing the O/D events.

3) For the same area where the hot-spots occur, we want to compare different temporal periods in order to find out during which times more events took place.

SOLUTION

1) The Hot-spot semantic map shows areas in which destination events take place near mapped venues. By interacting with this map view, it is possible to find answers to the first problem. We can firstly study the different types of hot-spot events taking place in the city area by exploring them in the city level or by the district. In the venue-chart, the proportions of different categories of venues, which occur in the selected district are shown. When we click on the venue-chart to the category of interest, e.g., conference centres, then in the map view only the distribution of hot-spots for which the selection applies to is presented.

As shown in Figure 4-9, during the study week destination events take place near four types of venues in Pudong. Slightly more than half of the detected events occur near commercial venues, while relatively big amount of events near conference and exhibition centres. A few events are detected near public and business areas. In Figure 4-10, the search is done for the events which occur near conference and exhibition centres. The result shows that this week events take place in four districts. A few of them in Jing'an and Huangpu, higher activity is near centre that is in Changnin, and there is very high activity near the centre in Pudong. Now it is a chance to choose an area and study it further. It is described in further chapter.

Figure 4-9 Hot-spots in Pudong are chosen. Pie-charts describe the semantics of hot-spots

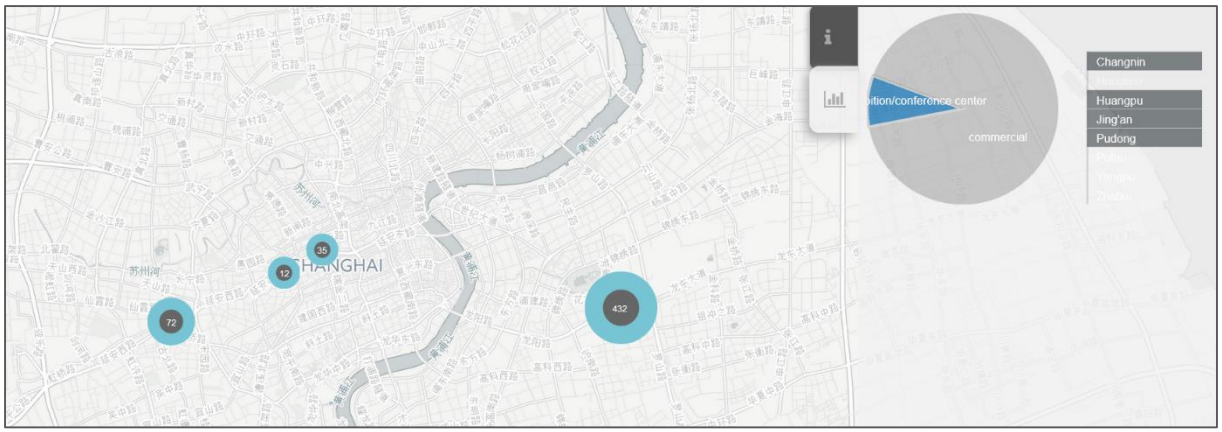


Figure 4-10 Hot-spots near conference and exhibition buildings

2) The heat map showing hot-spots' locations can be used to solve the second problem (Figure 4-11). As we have detected the location of the conference centre, near which high intensity activities take place, we can find the same location in the heat map and continue with more detailed analysis of activities taking place near the Shanghai New International Expo Centre. As high intensity events took place during four days, we can brush the date-chart showing activities by days and choose each day in order to get an insight of daily activities. Then in the map view and the hour-chart the results are displayed accordingly. The hour-chart views for four days are presented in the Figure 4-12.

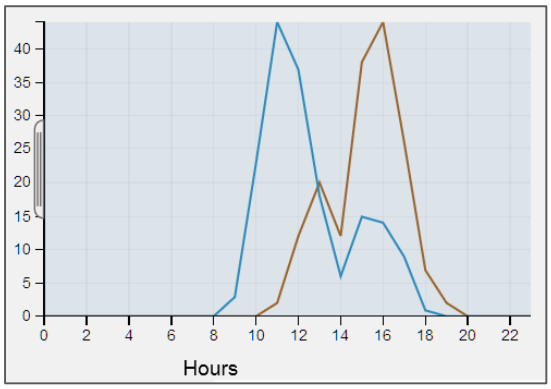
On Wednesday, 19.05.2010 the graphs for destinations and origins mirror each other Figure 4-12-a. First, very high intensity destination events took place in the late morning, the highest peak reached close to 50 destination events around 11.00. The duration of destination events activity period is 9 hours from 9.00 till 18.00. The taxi trips origin events were shifted from destinations, their duration period lasted 6 hours from 11.00 till 19.00 with the highest peak 16.00 which reaches also almost to 50 origin events.

On Thursday, 20.05.2010 only destination events occurred in this area from 10.00 till 17.00 with highest activity times around 11.00 and 12.00 (Figure 4-12-b).

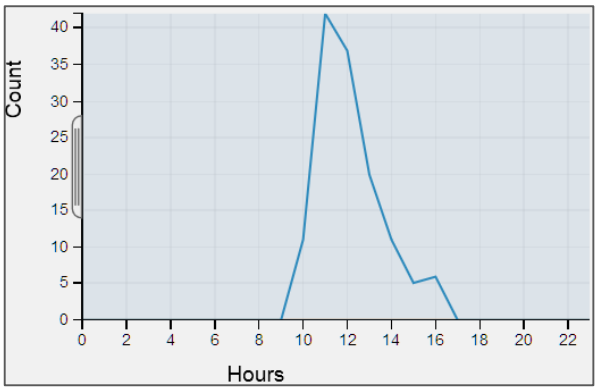
On 21.05.2010, Wednesday destination events took place from 9.00 till 16.00 and origin events from 14.00 till 16.00. Again, the highest activity took place around 11.00 and 12.00 for destination events. On this day the rate for the origins events was small, from 14.00 till 16.00 with 10 events during the highest hours (Figure 4-12-c).

In conclusion, very high intensity events took place in this area during the trade fair. In contrast, on other days no hot-spots occurred with an exhibition of Tuesday. In this area, taxi events can describe well traveller's behaviour during an event. It shows that in certain days the demand for taxi in this area grows a lot, while on other days it can be very low.

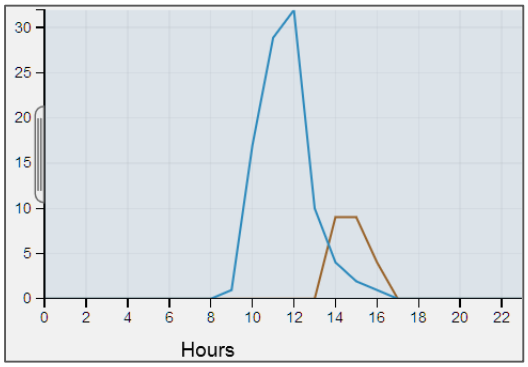
Figure 4-11 Shanghai New International Expo Centre (NIEC) in Pudong district



a) O/D events on 19.05.2010



b) O/D events on 20.05.2010



c) O/D events on 21.05.2010

Figure 4-12 Images a,b,c are representing days from 19.05.10 – 21.05.10.

3) In order to get an overview to the proportions of intervals during which high intensity activities took place, it is possible to use map (Hot-spots' temporal proportions). By clicking on the day-chart, it is possible to reveal the temporal proportions near the NIEC on Wednesday or during three days of high intensity activities, which occurred on Wednesday, Thursday, and Friday.

On Wednesday, on the current zoom two hotspot areas are marked with pie-symbols near the NIEC. In one spot most of the taxi drop-off events took place during the interval of 09:00-12:00. Other active intervals were 06:00-09:00, 12:00-15:00, and 15:00-18:00. During this time the

drop-off events are spread over a wider area as more dots in blue are being shown on a map (Figure 4-13).

In Figure 4-14, the sums over three days are given. Here we can see that in comparison of different intervals the drop-off events during the interval 09:00-12:00 took place very intensively, and the in smaller scale during on the time period of 06:00-09:00 and 12:00-15:00. Only a small part of events took place in between 15:00-18:00 (Figure 4-14).

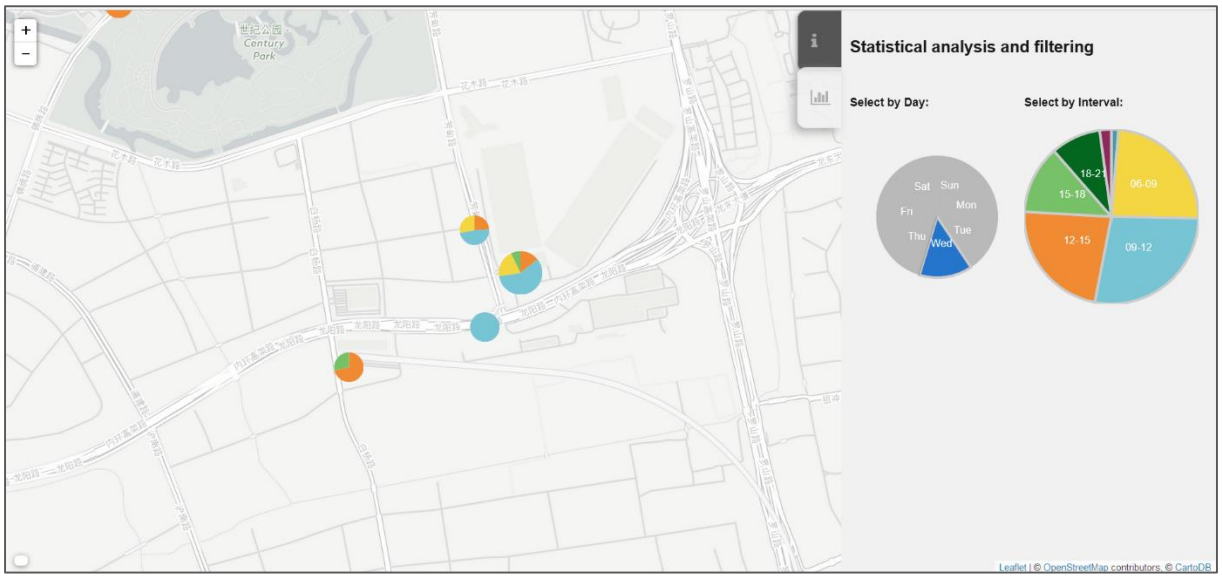


Figure 4-13: Destination spots in front of the Shanghai New International Expo Centre (NIEC) on Wednesday

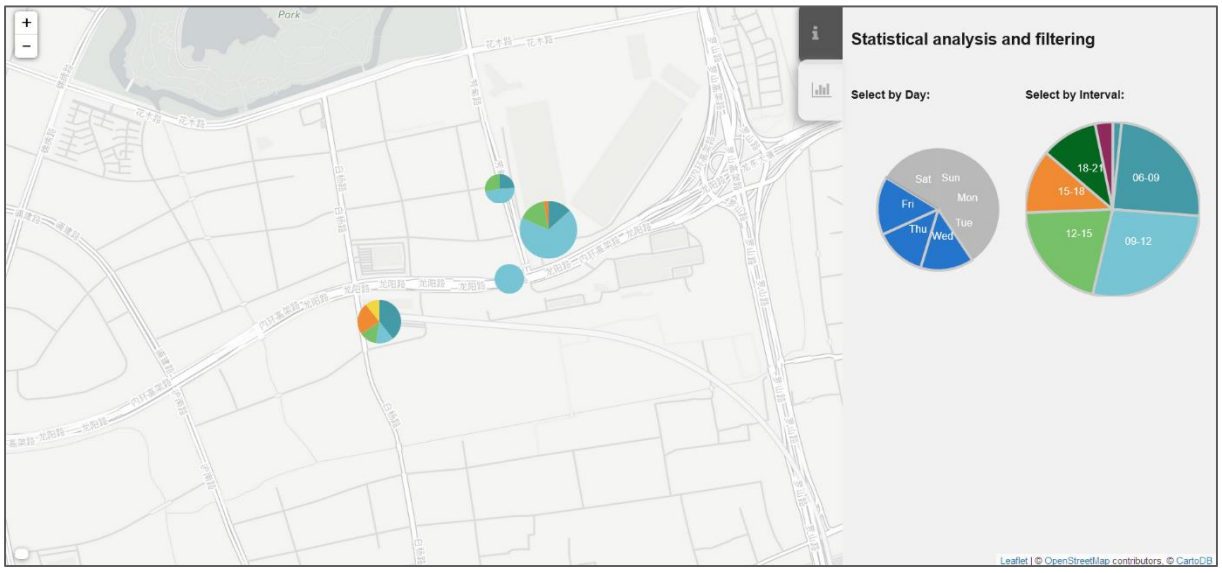


Figure 4-14: Destination spots in front of the Shanghai New International Expo Centre (NIEC) during three days (on Wednesday, Thursday, Friday)

4.3.2 Study Area II

The study area II is another exhibition and conference centre, the Shanghai Exhibition Center in Jing‘an district.

PROBLEM

When do hot-spot events occur near the Study Area II?

SOLUTION

With the heat map the study area II can be described by using the area-selector for focusing this area, so that the temporal charts (the day-chart and the hour-chart) are recalculated (Figure 4-15).

Figure 4-15 Near Shanghai Exhibition Center destination hot-spots occurred on two days of the week, on Friday and Saturday

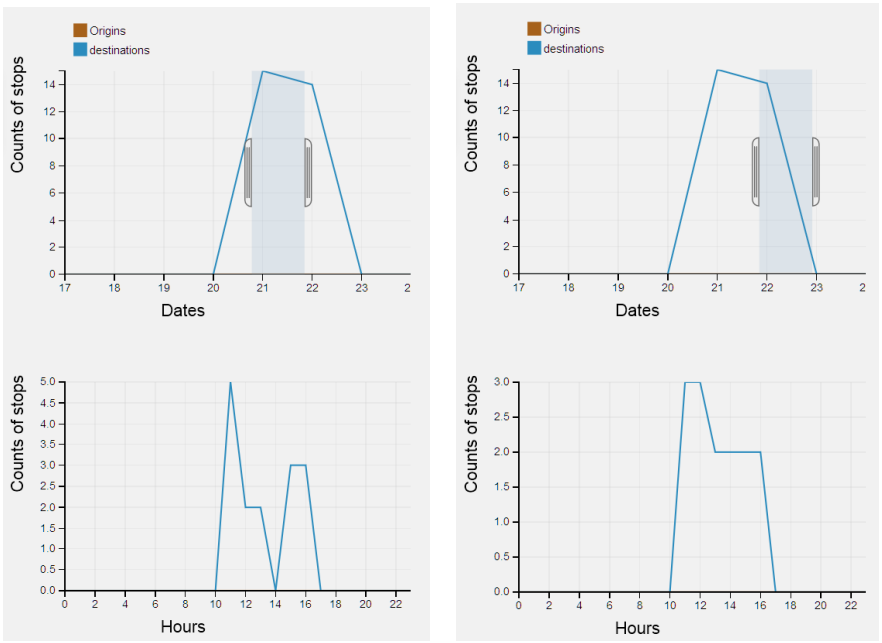


Figure 4-16 Travellers high intensity activities on 21.05.2010 (Friday) and on 21.05.2010 (Saturday)

In this area destination hot-spots occurred on two days of the week, on Friday and Saturday (Figure 4-16). During this week there were no origin hot-spots. During these two days the activity hours were similar, from 11.00 to 17.00 with highest peaks around 11.00-12.00.

During one week period the activity was low in this region with hot-spots occurring only on two days. However, in comparison with study area I, the activity level around this exhibition center was low, but may still be associated with a smaller scale public or entertainment event taking place in this area.

4.3.3 Study Area III

The study area III is the West Nanjing Road that is a famous commercial street in Shanghai. From Middle Xizang Road in Huangpu district the West Nanjing Road extends in the southwest direction to West Yan'an Road near Jing'an Temple in Jing'an district. There are many shopping malls on this street, some examples are Plaza 66, Wagas, and Meilongzhen Plaza.

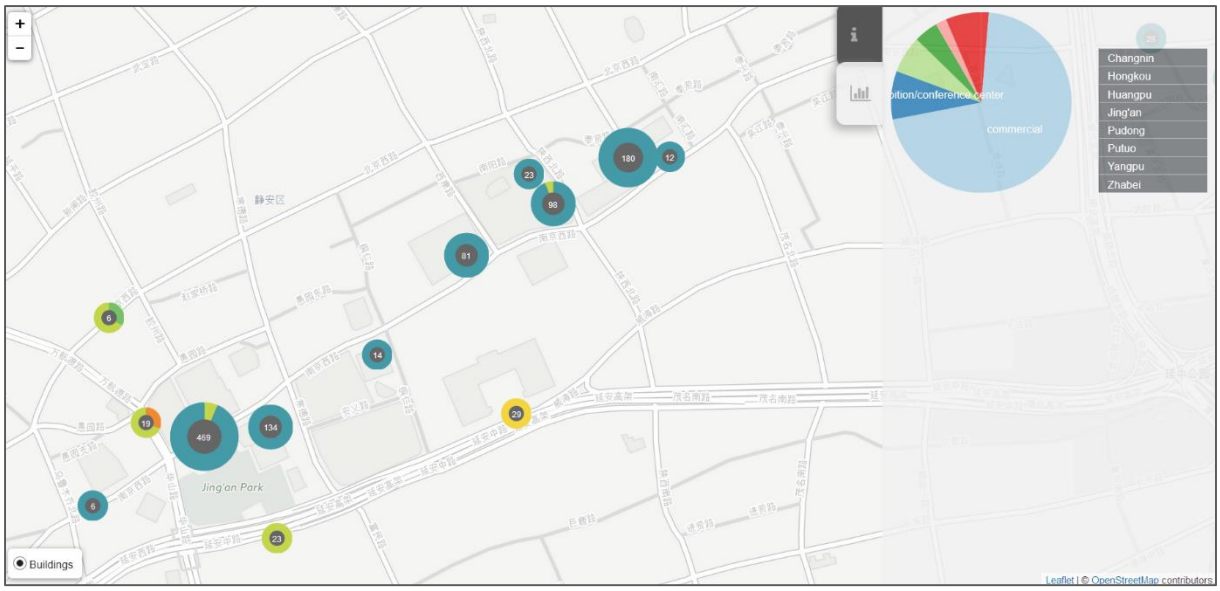


Figure 4-17: Venues labelled with tag "commercial" along the West Nanjing Road

PROBLEM

The Venue map (Figure 4-17) shows that on the West Nanjing Road the hot-spots are mostly labelled with tag "commercial" shown in Figure 32. In comparison, we can see that in nearby streets the hot-spots are marked with other types of venues. We are interested in exploring this shopping area.

1) What are the locations of origins' hot-spots and destinations' hot-spots over one week period and compare them.
2) What is the temporal distribution for hot-spots in different locations along the street.

SOLUTION

1) In the heat map showing hot-spots' locations, it is possible to turn on and off the layers of taxi travellers' destinations and origins.

The taxi travellers' origins hot-spots are shown in Figure 4-18 and destination hot-spots are in Figure 4-19. Along the West Nanjing Road the patterns for the pick-up and drop-off locations overlap in many locations, but there are also some places where only on type of hot-spot occurs. From the heat map it is possible to detect that there are four destination areas with very high intensity peaks, meanwhile the number of very high peak hot-spots for origins is two.

There are more hot-spots for origins with moderate intensities than for destinations.

Figure 4-18: Taxi travellers' origin locations along the West Nanjing Road and in surrounding area

Figure 4-19: Taxi travellers' destination locations along the West Nanjing Road and in surrounding area

2) The second question can be discovered with the pie-chart map describing the hot-spots' temporal proportions.

The map reveals different spots along the road in which activities take place (Figure 4-20). In some areas the pie-charts are bigger and the drop-off events occur with higher intensity. Near these areas locate shopping malls and also the Jing'an Temple and park. In addition, higher temporal variations occur near these areas. In other spots the drop-off events take place mostly during the two intervals of 09.00-12.00 and 12.00-15.00, and there are also hot-spots occurring only during one interval periods.

Figure 4-20: Pie-chart representing hot-spot locations with temporal distributions along West Nanjing Road. West Nanjing Road zone is coloured red.

4.3.4 Study Area IV

The study area IV is a business area. The main building next which the hot-spots occur is an office building called the Century Commercial and Trade Plaza. Next to it is also another office building and across the street is a residential building and a restaurant. There is also one coffee shop in the main office building.

PROBLEM

1) What are the temporal patterns in a typical business area?

SOLUTION

A heat map showing hot-spots' locations is used to describe temporal patterns in the study area IV. The study area is taken into focus with the area-selector after which the temporal patterns for the day-chart and hour-chart are displayed.

In Figure 4-21 the weekly pattern is shown. Near Century Commercial and Trade Plaza the destination hot-spots occurred on weekdays, but none of them on the weekend. Both, destination and origin hot-spots occurred there. In total there were higher rate of destinations than origins. Generally, high intensity destination events' gatherings occurred during office hours. Activities of high intensity periods had peak hours around 11.00-12.00 and 15.00-16.00. Origin events have two activity periods, one is during day, and another in the late evening. During day time, taxi drop-off events follow pick-up with a delay, having peaks around 14.00 and 16.00. In the evening, higher rate of drop-off events take place from 20.00 till 21.00.

However, daily activity charts (Figure 4-22) reveal higher variations for O/D events. In general, during the five days the rhythm of destination events is more stable in comparison with origin events, e.g., on Thursday there are no origin hot-spots in this area and for other days the hot-spot periods are shifted and durations vary a lot.

In conclusion, the taxi demand during one week period in this area was rather steady and the need was moderate during the weekdays, meanwhile it is a quiet area during the weekend.

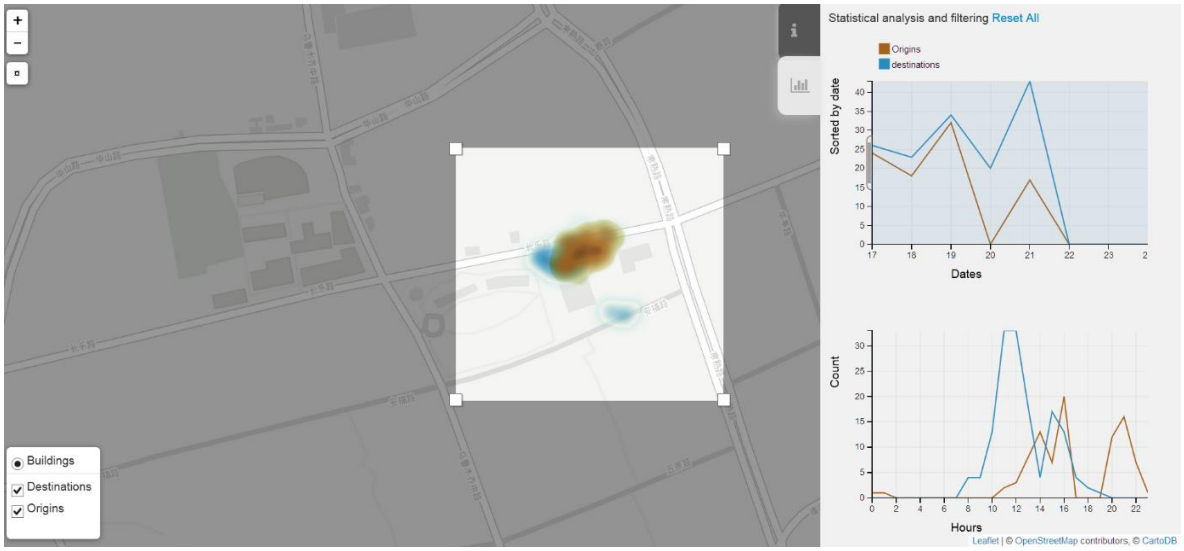
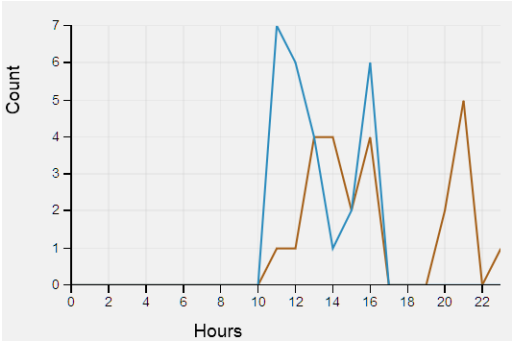
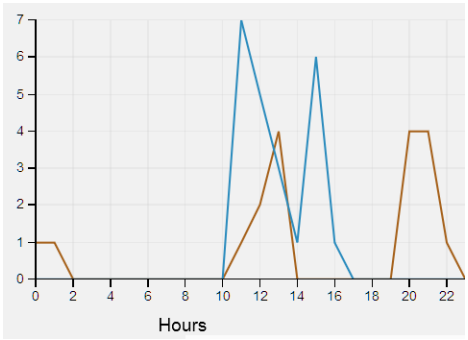


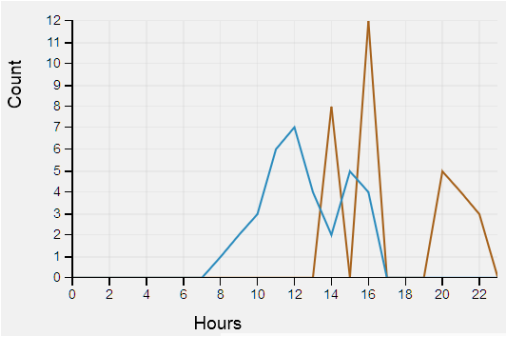
Figure 4-21: Century Commercial and Trade Plaza in Xuhui district



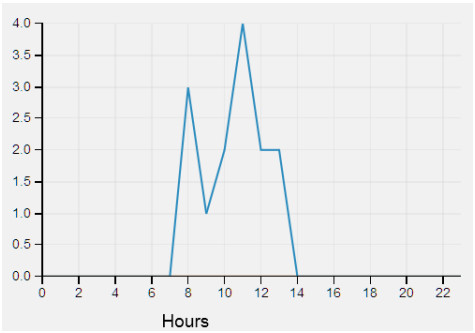
a) O/D events on Monday, 17.05.2010



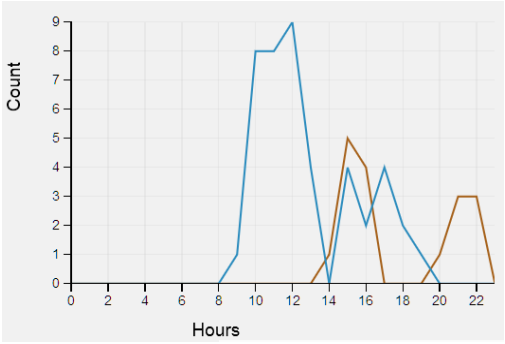
b) O/D events on Tuesday, 18.05.2010



c) O/D events on Wednesday, 19.05.2010



d) O/D events on Thursday, 20.05.2010



e) O/D events on Friday, 21.05.2010

Figure 4-22: Daily activities during five consecutive weekdays of taxi travellers' O/D activities in the study area IV

5 Discussion

5.1 Clustering O/D Hot-spots with DBSCAN Algorithm

Different types of clustering methods were studied based on literature. From them DBSCAN was chosen. The exploration with DBSCAN revealed that, firstly, the choice of *Eps* depends on distribution. Secondly, with different *Eps* and *MinPts* different size of clusters can be detected. Thirdly, the in different level of scales of the city different number of clusters and clusters with different sizes can be detected with DBSCAN. In the city level, there are few visual clusters, but one big covering most of the city. Then, with higher *Eps* value, clusters with higher variability of inner densities could be found. DBSCAN finds automatically clusters, which makes it suitable to apply it if there is no previous knowledge of possible number of clusters, and in situation, if number of clusters is varying a lot for each sequentive interval.

5.1.1 Detecting O/D Hot-spots with DBSCAN Algorithm

DBSCAN is not solving the KDE problem that is to find areas with different densities. DBSCAN is suitable for detecting higher clustering than the set distance threshold. If given a big value for the *Eps* and small for the *MinPts*, then the clustering results in large areas with possibly strongly varying clustering, including both sparse and high concentration areas. The opposite, with a high *MinPts* and small *Eps* detected clusters are small areas with high densities, and lower density areas are omitted.

The tests with different parameters for clustering with DBSCAN algorithm showed that the choice of *Eps* is equivalent to the level of scale (in city) in which clustering can be detected. For detecting the highest peaks of densities a smaller *Eps* with higher *MinPts* can be chosen. The *Eps* of 25 meters and *MinPts* 6 were found suitable for local hot-spot detection for the given data, it revealed the highest clustering that took place on the street near just a few venues. DBSCAN clustering for finding street level hot-spots over the city scale data set seems to work.

The choice of detecting clusters in three hour interval basis was made because in this time period clustering occurred more significantly compared to one hour time. In addition, clusters that could possibly appear because of a certain time, could appear around the full hour (before and after it), and that would have been undiscovered with hourly intervals. Cutting data by three hour intervals allows detecting clustering that occurs in this period. If the concentration event started in between one interval and lasted until the second interval, then their densities might not have been significant enough to be detected by the DBSCAN algorithm. Another method could have been the usage of moving time-windows or clustering 3-dimensional space-time data.

During relatively small temporal interval, three hours, and in a small area, high gathering may indicate an event taking place near a place where hot-spot occurred. However, it is not revealed if this detected cluster is a peak of a bigger generally dense area (but not dense enough to be detected with the current parameters) or the detected cluster is detached from other taxi taking events.

5.1.2 Temporal Patterns of O/D Hot-spots

On interval basis the travellers' trips origin and destination events were detected. More destination hot-spots were detected than origin hot-spots. It shows that taxi drop-off locations occur more often in a similar area, and pick-ups occur more diffusely. Diffusely occurring taxi taking events are not found with DBSCAN algorithm, which detects dense regions.

In different times of the day the different types of hot-spots prevailed. There were more D hot-spot events taking place on daytime, meanwhile the origin's hot-spot events dominated during the hours of evening and night. These phenomena are probably taking place due to taxi travellers' trip taking reasons, which change during the day.

There were variations in the number of detected clustering events for different days. The variations occurred distinguishably for the weekend days during some intervals, but between different weekdays. Since, only one week data set is used, it is not known which variations are common and which uncommon. Still, the general trends of the occurrences of low-clustering times and high-clustering times for the daily intervals were depicted with these seven days.

The results depend on constraints defined by the temporal intervals, since we did not look inside these intervals in smaller granularities, e.g., hourly basis, and some processes (rises or drops) could have started in the middle of the interval.

5.1.3 Hot-spots' Semantics

There are several drawbacks with the application semantic layer which would need further work. Firstly, the scarce data that was available about the study area for the venues. Secondly, the reasoning where the traveller would go from the taxi drop-off location or where they come for taking the taxi remains unclear. There might be various semantical reasons in small scale analysis and there are often several different types of function zones around one taxi stop.

With open-source data it is possible to gain insight in Shanghai only to the certain regions, since compared to some European or American thoroughly mapped regions, Shanghai has a small number of venues mapped. More consistent semantic data layer is needed to get a valid result that would describe the study area. At the moment, it is only possible to study the semantics for a few regions. The author of this work was accessible only for Western open-source data offers, while not to their Chinese equivalents. Another consideration is in China, the unauthorized mapping activities are not applicable.

A small number of venues, 3150 of them, were mapped and used for labelling the detected clusters. Therefore, after labelling the hot-spot clusters the majority of them remained unidentified. In the application on the semantic layer, only hot-spots for which semantics were detected were visualized. Since the semantic layer is thin, it is not possible to compare on the city level where which venues most of the hot-spots occur, but it is possible to study the occurrences of hot-spots near mapped venues and study their daily and weekly behaviour.

The semantics of hot-spots is defined by the idea of Tobler's „first law of geography" that close things in space are related to and the labels are given according to the closest venue detected from the buffer zone of the taxi stop data point. However, the semantic layer for the application currently offers simplified labelling of the hot-spots, since the each point is labelled just with the nearest venue, while there might be more venues in walking distance from the taxi pick up or drop off location. Because of this, the semantics of the hot-spots remains uncertain. It might be reasonable to apply further analysis for all of the venues in the buffer zone for selecting out the main venue. The meaning could be generalised by detecting the main venue from the group of venues in buffer zone. Then, a two level of place identification could be offered, that

is to find the main label for the hot-spot; offer other possibilities on demand. Another way is to add all tags and then offer for the user a chance to filter out the tags of interests.

5.2 Visualization Application

Here, the considerations of interactive visualization application in the point of view of research questions and application requirements are discussed. An overview of visual analysis results discovered with the use of this application is given.

The requirements for the visualization application according to the data were that multidimensional data was being presented. The two spatial coordinates and contextual information of time and semantic labels. In addition, two different types of taxi stop events were analysed. Since two-dimensional visualization methods were chosen in web environment, then the goal was to enable the map user to go through the different levels of representations by interacting with data by the means of filtering. Different aspects of data were revealed with linking map view with charts.

The result of data pre-processing and clustering revealed the high intensity places, which were small in area, but contained high intensity of point events. These detected hot-spots were notably detached on the street level, and on the street block level they were revealed as spots of places. These areas were presented in two maps with point symbols (pie-charts) and in one map with heat map presentation. One goal of the application was to show events in different zoom level. However, in smaller zoom level, when having a small scale view, point symbols start to clutter. Therefore in this applications visualization methods that scale in different zoom level were chosen.

The cluttering problem appeared also when showing point events that took place in the same place repeatedly, but they had different contextual attributes, either timewise or semantically. For dealing with this problem, the categorical pie-charts were drawn on two maps and in another map two charts with temporal filtering functions were offered.

In this application, the area based filtering is offered with Area-Selector which creates an adjustable rectangle on the map view which size and shape can be modified. In the process of analysis, this tool was tested and considered useful. However, as a further adjustment and more flexible area-selector could be offered, by which areas with more customised shapes could be chosen.

In this application some of the data transformations has been done in client with JavaScript, e.g., filtering with Crossfilter. This solution is well suitable for the data amount used in current work. However, when scaling up the application, then the load for the client-side engine would grow and different methods of web mapping need to be used, e.g., filtering methods could be brought to the back-end. Secondly, a further aggregation of clustering results could be done, e.g., by visualising only centroids of clusters instead of all points.

In this application, different goals were achieved with different visualization methods. Each of them has its strengths. The heat-layer scales for different zoom levels so that in larger scale they show small areas of hot-spots, but in smaller scale different areas are united and a continuous representation of an area is given. In the opposite, the q-cluster pie-charts are a representation of point symbols which serve the purpose in large scale when the interest is to reveal different spots in the area, but in smaller scale they do not give that intuitive view of the spatial coverage. However, they are useful for showing different categories.

5.2.1 Map I: Hot-spots' Semantics Map

The pie-charts (*Figure X*) of hot-spot semantics map are showing clearly the distribution of possible meanings of stops in different zoom scales of the map view. The Crossfilter filtering system enables studying the temporal pattern for different venue types. However, for enhancing the map functionalities, in future an additional filtering by selecting out pie symbols on the map for finding out the temporal distribution for each hot-spot could be supplemented.

5.2.2 Map II: The O/D Location Map

The O/D location map offers an insight into general trends of hot-spot locations. It is presented with heat-layer which will react to different zoom level. Origin and destination hot-spots tend to occur in the same areas, therefore on the map view their symbols often overlap. For solving this problem, user can turn on and off O/D layers and in need can switch their visibility order. Then, in this map interactive tooltips that occur during mouse-hover events for the statistical charts are not enable, because of their conflict when using the brushing tool for filtering. However, tooltips are necessary since they offer better experience for reading the chart values.

If to filter over several days, then the chart for hours shows sums of events for each hour over chosen days. As another option, it could be interesting to have a line that shows hourly averages over chosen days.

When considering usability issues, then the Crossfilter filtering brush could adapt more to the chosen units and be less sensitive for choosing areas in between units.

5.2.3 Map III: Temporal Distribution in Popular Areas

Although, pie-charts are calculated for each zoom level, then the representation of a pie-chart may not give the best overview of the spatial coverage of hot-spots events, since the pie-charts are pointwise representations. The pie-charts are calculated with q-cluster plugin, which uses dynamic coloring method for the pie's categories. As a consequence, it was challenging to link pie-chart colours in the map view with crossfilter pie-chart colours in the map tools chart. The same problem occurs in the Hot-spots' semantics map.

5.3 Taxi Traveller's Time Patterns

The concentration of this study is on taxi travellers' activities in popular areas, in which high intensity activities took place. The clustering events are detected on a local scale, so that detected high intensity events can be associated with nearby venues. Several types of venues extracted mainly from OSM data were used in associating hot-spots' points with them. Later some venue types of interests were chosen to visualise on the map to use for further analysis. The chosen venues were commercial areas, venues in which bigger public places can take place, e.g., exhibition and conference centres, hotels which can have conference rooms, public buildings as libraries, and business-buildings as offices.

The map application is designed to enable discovering the patterns in the city level. For presenting the application use and discover traveller's patterns four study areas were chosen which were described with the use of three map views of the Taxi-map application. As expected, in different places different traveller's behavioural patterns came forward.

In the first study a traveller's O/D activities near conference and exhibition halls in a city scale were explored. Then, from the information shown in the map the *Study Area I* and *II* were chosen and studied more in detail. In general, hot-spots around exhibition and conference centres during this week occurred occasionally. Very high intensity activities were revealed for

one centre over three days, but for other days no hot-spots occurred around it with one exception. The *Study Area II* was another exhibition centre, around which only drop-off activities took place during two days of a week, but with a rather low intensity.

The third study area reveals travellers' patterns on a busy shopping street. Many stops occurred along the street during seven days a week. The highest activities took place near bigger shopping malls and near one cultural landmark with a park, which can attract tourists or could be popular among locals. The comparison the temporal distributions of hot-spots along the street shows that in the most popular locations for taxi trips destinations activities occur all over the day from early morning till late evening. There were clusters that occurred during several intervals. For other the high intensity activities lasted for shorter time, e.g., in addition to very popular hot-spots, many hot-spots in other locations along the street occur during the day-time from 09.00 till 15.00.

The *Study area IV* described an area in which the main building is an office building surrounded with a few other types of buildings. In this area both origin and destination hot-spots appeared generally more regularly over the five weekdays, but with some variations when studying them more in detail. In this area, destination patterns occurred during the office hours, but origins had higher variations.

These study areas show different travellers' behavioural patterns driven by different travel purposes. However, the results of these analysis are influenced by the definition of temporal intervals (events detected in three-hour ranges), the number of mapped venues and by the method of defining the semantic labels for the hot-spots.

6 Conclusions

6.1 Conclusion

Two general goals of this study were to take an insight into the movement data and to implement a web visualization representing the results of data processing. To conclude, a clustering algorithm DBSCAN was employed for discovering O/D high density areas of taxi traveller's in Shanghai and a web visualization application that presents the results and helps in discovering and analysis process was created. With this tool the research questions (2), and (3) regarding to traveller's behaviour were answered and a comprehensive insight into the taxi data was taken.

One of the established research question in this study was about how to detect taxi travellers' O/D hot-spots from the movement data. The hot-spots were defined as high intensity areas for trips origins or destinations. A general overview of used statistical techniques solving the hot-spot detection problem were given. In which, different methods have been used for concerning whether the interest is in detecting the existence of clustering in data or to detect the clusters itself (their locations in data space). Further, the data mining technique clustering was considered suitable for detecting hot-spots of FCD and a classification of them was given with choosing a density-based clustering algorithm DBSCAN for implementation. The DBSCAN algorithm requires two parameters as an input and deciding their values are done experimentally. They are application specific since they depend on the data distribution and in which level of scale is the interest for finding the clusters. DBSCAN enables finding clusters that have higher concentration than defined threshold of mininum points in a cluster in a defined area of interest. So, this was taken as a criterium for the hot-spot event definition. The DBSCAN algorithm was applied over 2D-dimensional spatial domain for the data set that was segmented over temporal domain by three-hour intervals. In conclusion, the DBSCAN algorithm can be used to detect high intensity areas, meanwhile extracting them from other data that are assigned as noise.

The second research question of this study was when and where do these high intensity origin/destination hot-spots occur. To answer this question, a web application to visualize the detected hot-spots was created

The third question of this work concerns the semantics of detected hot-spots and tries to reason the meaning of hot-spots by travellers possible activities in surrounding regions, determined by the closeness of amenities, and then study the distribution of the trips purposes. The third map is showing the distribution of amenities for each hot-spot with a pie-chart symbol map. In the application the semantic chart and filtering over the semantic domain help in discovering the semantic distribution.

The last question concerns issues related to technical implementations of the visualization application and the ways how previous two questions can be answered with the means of application. So, this question is partly opened via the two previous questions. The created interactive web-based geovisualization application helps to investigate and understand the patterns of taxi travellers activity (high intensity traxi travellers gathering event detection). The 2D representation of data on a map is supported with linked charts describing the temporal and semantic distribution and filtering functions. So that an interactive mode would enable the discovery to the different aspects of visualization. The last question concerns the issues related

to the creation of the visualization application. The created application enables to study the distribution of hot-spots in spatial and temporal domains. Two different map views were created to describe the temporal and spatial distribution of hot-spots. The first one serves the purpose of showing the distribution of origin and destination hot-spots in space with controllers for modifying the time window. The second map is a pie-chart symbol map showing the temporal distribution as categories in a pie.

6.2 Outlook

Here a discussion is given on what challenges remain on the field of this study with some plans for future directions.

In this work, the main hot-spots in the city were detected, however, another interesting case would be to detect hot-spots which are uncommon and irregular. For that further analysis could be done with the additional data of comparison. Either by comparing the results of hot-spots with common hot-spots that are detected from studying longer period of data in order to find out what is typical through a year or month, or by comparing the hot-spots with additional data layers such as demographic layer. In addition, another data mining method machine learning could be used for studying the data and detecting unusual hot-spots occurrences. If to add more data for the application, not only about amenities, but data about public events taking place in the city, then an interesting question would be the prediction of taxi traveller's hot-spot locations. Then automation of all data processing steps can be useful for levering up the application for the real-time visualization purposes. It would be usable for monitoring high intensity taxi hot-spot creation in real time. In order to create automated surveillance visualizations.

Secondly, these maps and charts describe activities during high intensity periods, however, it would be interesting to combine in the same application the two layers of data – first, the data in higher scale can show the hot-spot areas detected with clustering, when zoomed in and after choosing the area of interest with area selector, then the whole data set describing this area could be requested and rendered.

The current application is a prototype and a tool to represent the analysis results. However, if to target a similar application for a larger audience, then it would need to be adapted to meet the user cognitive understanding needs. Therefore further steps of unifying different maps into one and usability tests of the application could be done for evaluating user experience and possible improvements could be implemented. Currently several single visualizations are built, which in the future could be combined into one complex, but user intuitive user interface.

In the topic of web visualizations, when it comes to scalability issues, then some further methods could be considered. In this work filtering calculations are run in the client, however with larger data sets problems with browser engine and memory could occur. In order to overcome these issues, more server-side data filtering processing could be done, e.g., Crossfilter.js could be implemented in back-end. When the speed of rendering visualizations is important, then an approach of using HTML5 canvas for drawing instead of SVG could be a possible solution, since SVG adds more to the DOM complexity.

References

AGAFONKIN, V. 2011. *Leaflet an open-source JavaScript library for mobile-friendly interactive maps* [Online]. [Accessed 22.01 2016].

AGAFONKIN, V. 2014. *Leaflet.heat* [Online]. [Accessed 10.10 2015].

AGGARWAL, C. C. Mining text data. Data Mining, 2015. Springer, 429-455.

AGRAWAL, R., GEHRKE, J., GUNOPULOS, D. & RAGHAVAN, P. 1998. *Automatic subspace clustering of high dimensional data for data mining applications*, ACM.

ANDRIENKO, G. & ANDRIENKO, N. Spatio-temporal aggregation for visual analysis of movements. Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on, 2008. IEEE, 51-58.

ANDRIENKO, G., ANDRIENKO, N., BAK, P., KEIM, D. & WROBEL, S. 2013a. *Visual analytics of movement*, Springer Science & Business Media.

ANDRIENKO, G., ANDRIENKO, N., FUCHS, G., RAIMOND, A.-M. O., SYMANZIK, J. & ZIEMLICKI, C. Extracting semantics of individual places from movement data by analyzing temporal patterns of visits. Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place (COMP'13), 2013b.

ANDRIENKO, N. & ANDRIENKO, G. 2006. *Exploratory analysis of spatial and temporal data: a systematic approach*, Springer Science & Business Media.

ANDRIENKO, N., ANDRIENKO, G. & GATALSKY, P. 2003. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing,* 14**,** 503-541.

ANKERST, M., BREUNIG, M. M., KRIEGEL, H.-P. & SANDER, J. OPTICS: ordering points to identify the clustering structure. ACM Sigmod Record, 1999. ACM, 49-60.

BERTIN, J. 1967. 1983 Semiology of Graphics: Diagrams, Networks, Maps. Madison, WI, University of Wisconsin Press.

BIRANT, D. & KUT, A. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering,* 60**,** 208-221.

BRUNSDON, C., CORCORAN, J., HIGGS, G., 2007. Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems,* 31**,** 52-75.

CARTWRIGHT, W., CRAMPTON, J., GARTNER, G., MILLER, S., MITCHELL, K., SIEKIERSKA, E. & WOOD, J. 2001. Geospatial information visualization user interface issues. *Cartography and Geographic Information Science,* 28**,** 45-60.

CHANG, H., TAI, Y., CHEN, H. AND HSU, JANE Y. iTaxi: Context-Aware Taxi Demand Hotspots Prediction Using Ontology and Data Mining Approaches. The 13th Conference on Artificial Intelligence and Applications (TAAI 2008), 2008.

References

CHENG, T., TANAKSARANOND, G., BRUNSDON, C. & HAWORTH, J. 2013. Exploratory visualisation of congestion evolutions on urban transport networks. *Transportation Research Part C: Emerging Technologies,* 36**,** 296-306.

COOK, K. A. & THOMAS, J. J. 2005. Illuminating the path: The research and development agenda for visual analytics. Pacific Northwest National Laboratory (PNNL), Richland, WA (US).

CRAMPTON, J. W. 2002. Interactivity types in geographic visualization. *Cartography and geographic information science,* 29**,** 85-98.

DANIEL, F. & MATERA, M. 2014. Mashup Components. *Mashups: Concepts, Models and Architectures.* Berlin, Heidelberg: Springer Berlin Heidelberg.

DATENBANKSYSTEME, L. I. F. I. L.-U. F. F. *ELKI: Environment For Developing KDD Applications Supported By Index Structures* [Online].

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)***,** 1-38.

DING, L., FAN, H. & MENG, L. 2015. Understanding taxi driving behaviors from movement data. *AGILE 2015.* Springer.

DOS SANTOS, S. & BRODLIE, K. 2004. Gaining understanding of multivariate and multidimensional data through visualization. *Computers & Graphics,* 28**,** 311-325.

EHMKE, J. F. & MATTFELD, D. C. 2010. Data allocation and application for time-dependent vehicle routing in city logistics.

ESTER, M., KRIEGEL, H.-P., SANDER, J. & XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd, 1996. 226-231.

FAYYAD, U., PIATETSKY-SHAPIRO, G. & SMYTH, P. 1996. From data mining to knowledge discovery in databases. *AI magazine,* 17**,** 37.

FITRIANAH, D., HIDAYANTO, A., FAHMI, H., GAOL, J. L. & ARYMURTHY, A. 2015. ST-AGRID: A Spatio Temporal Grid Density Based Clustering and Its Application for determining the Potential Fishing Zones. *International Journal of Software Engineering and Its Applications,* 9**,** 13-26.

HALLAHAN, N. 2014. *q-cluster* [Online]. [Accessed 16.11 2015].

HAN, J., LEE, J.-G. & KAMBER, M. 2009. An overview of clustering methods in geographic data analysis. *Geographic data mining and knowledge discovery, 2nd edn. Taylor and Francis, London***,** 149-187.

HINNEBURG, A. & KEIM, D. A. An efficient approach to clustering in large multimedia databases with noise. KDD, 1998. 58-65.

HUGHES, C., NAIK, V. S., SENGUPTA, R. & SAXENA, D. Geovisualization for cluster detection of Hepatitis A & E outbreaks in Ahmedabad, Gujarat, India. Proceedings of the Third ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health, 2014. ACM, 39-44.

KISILEVICH, S., MANSMANN, F. & KEIM, D. P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos.

Proceedings of the 1st international conference and exhibition on computing for geospatial research & application, 2010. ACM, 38.

KRUEGER, R., THOM, D. & ERTL, T. Visual analysis of movement behavior using web data for context enrichment. Pacific Visualization Symposium (PacificVis), 2014 IEEE, 2014. IEEE, 193-200.

KUMAR, V. 2009. *Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, Chapman & Hall/CRC.

KWAN, M.-P. & LEE, J. 2004. Geovisualization of human activity patterns using 3D GIS: a time-geographic approach. *Spatially integrated social science,* 27.

LEVINE, N. & CRIMESTAT, I. 2004. *A spatial statistics program for the analysis of crime incident locations,* Washington, DC, the National Institute of Justice, Washington, DC.

LI, H., WEI, Y. H. D. & HUANG, Z. 2014. Urban land expansion and spatial dynamics in globalizing shanghai. *Sustainability,* 6**,** 8856-8875.

LLOYD, S. P. 1982. Least squares quantization in PCM. *Information Theory, IEEE Transactions on,* 28**,** 129-137.

MACEACHREN, A. M. 1992. Visualizing uncertain information. *Cartographic Perspectives***,** 10-19.

MACEACHREN, A. M., GAHEGAN, M., PIKE, W., BREWER, I., CAI, G., LENGERICH, E. & HARDISTRY, F. 2004. Geovisualization for knowledge construction and decision support. *Computer Graphics and Applications, IEEE,* 24**,** 13-17.

MACEACHREN, A. M. & KRAAK, M.-J. 2001. Research challenges in geovisualization. *Cartography and Geographic Information Science,* 28**,** 3-12.

MACIEJEWSKI, R., RUDOLPH, S., HAFEN, R., ABUSALAH, A., YAKOUT, M., OUZZANI, M., CLEVELAND, W. S., GRANNIS, S. J., WADE, M. & EBERT, D. S. Understanding syndromic hotspots-a visual analytics approach. Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on, 2008. IEEE, 35-42.

MILLER, H. J. & HAN, J. 2009. *Geographic data mining and knowledge discovery*, CRC Press.

NEWTON, T. & VILLARREAL, O. 2014. *Learning D3. js Mapping*, Packt Publishing Ltd.

NURMI, P. 2009. *Identifying meaningful places.* PhD thesis, University of Helsinki.

O'BRIEN, C. M. 2009. Statistical Detection and Surveillance of Geographic Clusters by Peter Rogerson, Ikuho Yamada. *International Statistical Review,* 77**,** 474-474.

PEI, T., JASRA, A., HAND, D. J., ZHU, A.-X. & ZHOU, C. 2009. DECODE: a new method for discovering clusters of different densities in spatial data. *Data Mining and Knowledge Discovery,* 18**,** 337-369.

PETERS, S. 2014. *Dynamics of spatially extended phenomena: Visual analytical approach to movements of lightning clusters.* München, Technische Universität München, Diss., 2014.

RELPH, E. 1976. *Place and placelessness*, Pion London.

References

RESCH, B., HILLEN F., REIMER, A., SPITZER, W., 2013. Towards 4D Cartography - Four-dimensional Dynamic Maps for Understanding Spatio-temporal Correlations in Lightning Events. *The Cartographic Journal,* 50**,** 266-275.

ROGERSON, P. & YAMADA, I. 2008. *Statistical detection and surveillance of geographic clusters*, CRC Press.

ROTH, R. E. 2012. Cartographic interaction primitives: Framework and synthesis. *The Cartographic Journal,* 49**,** 376-395.

ROUSSEEUW, P. J. 2009. *Finding groups in data: An introduction to cluster analysis*, John Wiley & Sons.

SUH, S. 2012. *Practical applications of data mining*, Jones & Bartlett Publishers.

TOMINSKI, C. 2006. *Event based visualization for user centered visual analysis.* Citeseer.

TSOU, M.-H. 2015. Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science,* 42**,** 70-74.

TUKEY, J. W. 1980. We need both exploratory and confirmatory. *The American Statistician,* 34**,** 23-25.

VENT, K. 2014. *Inimese tegevuskohtade leidmine nutitelefonipõhiste käitumisandmestike alusel.* Master degree, Tartu Ülikool.

W3SCHOOLS. 2009. *W3Schools* [Online]. 2009-2016: Refsnes Data. Available: http://www.w3schools.com/ [Accessed 08.02 2016].

WANG, W., TAO, L., GAO, C., WANG, B., YANG, H. & ZHANG, Z. 2014. A C-DBSCAN Algorithm for Determining Bus-Stop Locations Based on Taxi GPS Data. *Advanced Data Mining and Applications.* Springer.

WANG W, Y. J., MUNTZ R. 1997. STING: A Statistical Information Grid Approach to Spatial Data Mining. Athens: Proceedings of the 23rd Conference on VLDB'1997. 186-195.

WOLFF, M. & ASCHE, H. 2009. Geovisualization Approaches for Spatio-temporal Crime Scene Analysis – Towards 4D Crime Mapping. *In:* GERADTS, Z. M. H., FRANKE, K. & VEENMAN, C. (eds.) *Computational Forensics.* Springer Berlin Heidelberg.

ZHANG, W., LI, S. & PAN, G. Mining the semantics of origin-destination flows using taxi traces. UbiComp, 2012. 943-949.

ZHAO, P., QIN, K., ZHOU, Q., LIU, C. & CHEN, Y. 2015. Detecting Hotspots from Taxi Trajectory Data Using Spatial Cluster Analysis. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences,* 1**,** 13.