TECHNISCHE UNIVERSITÄT MÜNCHEN

Ingenieurfakultät Bau Geo Umwelt

Lehrstuhl für Kartographie

Visual Analysis of Large Floating Car Data - A Bridge-Maker between Thematic Mapping and Scientific Visualization

Linfang Ding

Vollständiger Abdruck der von der Ingenieurfakultät Bau Geo Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzende(r): Univ.-Prof. Dr.techn. Roland Pail

Prüfer der Dissertation:
1. Univ.-Prof. Dr.-Ing. Liqiu Meng
2. Hon.-Prof. Dr.-Ing. Gerd Buziek
   Leibniz Universität Hannover
3. Univ.- Prof. Dr. rer. nat. Gennady Andrienko
   City University London, United Kingdom

Die Dissertation wurde am 03.11.2015 bei der Technischen Universität München eingereicht und durch die Ingenieurfakultät Bau Geo Umwelt am 29.02.2016 angenommen.

Acknowledgments

First of all, I would like to give my sincere thanks to my advisor Prof. Liqiu Meng. As a supervisor, she brought me into the field of Geovisualizaiton, gave me an interesting topic and devoted her time to valuable discussions. I still remember at the beginning of my PhD study she spent a whole afternoon teaching me how to write a scientific paper. She also encouraged me to develop my soft skills by involving me in teaching lectures, organizing conferences, and assisting Master programs. I also appreciate the life experiences she shared with us during coffee breaks, and the delicious food she cooked every year in the Christmas parties at her home. Those are all precious memories.

I am grateful to my second supervisor Prof. Gerd Buziek for reviewing my thesis and giving me valuable comments from both scientific and engineering perspective. I greatly appreciate Prof. Gennady Andrienko, who is a pioneer in Geovisual Analytics. My work has been greatly inspired by his innovation and creativity.

Besides my supervisors, Prof. Jukka Krisp guided me doing research during my first two years and has been continuously paying attention to my research work; Prof. Xiaoxiang Zhu involved me into the "4D-City" IGSSE project and I learned a lot from her and other team members; I am highly grateful to Dr. Hongchao Fan for his inspiring and constructive discussions, and also for our happy collaborations. I would like also to thank Prof. Chun Liu for sharing with us the Floating Car Data.

It was a great experience working in the team of Cartography. Luise Fleißer supported me in many aspects during the PhD study from visa application to the final defense. I would like to thank Holger Kumke, Mathias Jahnke, Christian Murphy for discussing research ideas, teaching master courses together, and especially helping with my daily life issues related to German. For Jian Yang who accompanied me along my PhD years, I appreciate his consistent support and encouragement, the funny activities he organized and the delicious dishes he cooked. Thanks also to Hao Lyu for always being there when I need his help, to Juliane Cron who tells me how to arrange the master program, and to Xiao Xie for sharing her research experience in China. Thanks also for Stefan Peters, Nina Polous and Ekaterina Chuprikova for bringing laughter to me in the office.

My warm gratitude also goes to all my friends in Munich for making my life interesting and wonderful.

Finally, I would like to thank my family, especially my parents for their loves and unconditional support; my brother for sharing his life with me; my grandparents for their positive life attitude and encouragement. My special thanks goes to my husband Guohui Xiao. Without his support and love, this thesis would never exist.

The thesis was funded by China Scholarship Council.

Abstract

In an information society, large volumes of data are being captured or produced on a daily basis using various technologies in many domains such as remote sensing, computational simulations and location-based services. However, getting insights into big data is very challenging due to their complexities in the size and structure. These challenges demand joint research efforts from multiple disciplines such as cartography, geovisualization, scientific visualization, information visualization, and the recently coined discipline visual analytics.

The aim of this thesis is to bridge the gaps between thematic mapping and scientific visualization and to achieve their synergetic effects for the visual analysis of big data. To establish a theoretical foundation, the author first conducted a systematical comparative study of thematic cartography and scientific visualization. The comparison covered seven essential aspects along the visualization pipeline: purposes, data sources, georeferencing, data preprocessing, classification, symbolization, and finally perception and cognition. The results showed that these two disciplines reveal different visual analytical levels and are mutually complementary, which encourages the author to seek their synergetic effects.

Based on the theoretical findings, the author conducted extensive experiments of visually analyzing massive and complex real-world taxi floating car data (FCD). Two FCD event categories at different abstraction levels, namely point-based and trajectory-based events, were identified for visual analysis. For point-based events, from the scientific visualization perspective, the author proposed several visualization methods including pie radar glyphs, time graphs and a salience-based method. The resulting compact visualizations revealed the underlying interesting data distributions. From the thematic mapping perspective, choropleth mapping and proportional mapping techniques are applied to communicate easily understandable spatiotemporal patterns. To analyze the trajectory-based events using scientific visualization methods, highly interactive techniques were employed, including interactive clustering methods and a parallel coordinates technique, to let users visually explore potentially interesting clusters, which were further explored on map views using gradient line rendering, direct line rendering and space-time cube approaches. A variety of thematic mapping techniques, including flow mapping, dot mapping, temporal matrix, and pie-chart mapping, are proposed to effectively communicate the spatial interactions and the taxi drivers' behavior patterns.

These visual analysis experiments have demonstrated that the techniques from both disciplines can strongly support users to win insight into the movement data by jointly revealing human mobility patterns, land use types and taxi driving patterns, which otherwise cannot be discovered by employing visualization methods from either discipline. The results have confirmed the synergetic effects of both disciplines.

Zusammenfassung

In einer Informationsgesellschaft werden tagtäglich große Datenmengen durch die verschiedensten Technologien in vielen Bereichen, wie etwa in der Fernerkundung, der Computersimulationen oder der ortsbezogenen Dienste, erhoben oder produziert. Allerdings ist es herausforderd, einen Überblick über die Big Data aufgrund ihrer Komplexität der Größe und Struktur zu verschaffen. Angesichts dieser Herausforderung sind multidisziplinäre Forschungsarbeiten, beispielsweise in der Kartographie, der Geovisualisierung, der Wissenschaftlichen Visualisierung, der Informationsvisualisierung, sowie in der neu entstandenen Disziplin Visual Analytics, erforderlich.

Das Ziel dieser Arbeit ist zwischen der Thematischen Kartographie und der Wissenschaftlichen Visualisierung eine Brücke zu bauen und die Synergieeffekte beider Disziplinen für die visuelle Analyse von Big Data zu gewinnen. Zu Bildung der theoretischen Grundlage findet zuerst eine systematische Vergleichsanalyse statt. Die Reihenfolge des Visualisierungsprozesses folgend, deckt die Vergleichsstudie sieben essentielle Bereiche ab: Verwendungszweck, Datenquellen, Georeferenzierung, Datenvorverarbeitung, Klassifizierung, Symbolisierung und letztlich Wahrnehmung und Kognition. Die Ergebnisse haben unterschiedliche visuell-analytische leistungen der beiden Disziplinen sowie deren komplementäre Rollen gezeigt. Diese Ergebnisse deuten auf die Machbarkeit und Notwendigkeit hin, die Vorteile der komplementären Eigenschaften bzw. Synergieeffekte dieser zwei Disziplinen zu gewinnen.

Von den theoretischen Befunden ausgehend führt die Autorin umfangreiche Untersuchungen der visuellen Analyse von großen und komplexen Taxi-Bewegungsdaten (FCD) durch. Zwei FCD Eventkategorien in verschiedenen Abstraktionsebenen, der punktbasierten und der trajektorienbasierten Events, wurden für die visuelle Analyse identifiziert. Für punktbasierte Events sind mehrere wissenschaftliche Visualisierungsmethod einsatzbar, die Kreis-Radar-Diagramme, Zeitgraphen und eine Aufmerksamkeitsbasierte Methode umfassen. Die resultierende kompakte Visualisierung offenbart grundlegend interessante Datenmuster. Thematischen kartographische Methoden, wie Choroplethenkarten und proportionale Symboltechniken dagegen kommen zum Einsatz, um leicht verständliche raumzeitliche Muster zu kommunizieren. Um die trajektorien-basierten Events mit wissenschaftlichen Visualisierungsverfahren zu analysieren, hat die Autorin interaktive Techniken, wie z.B. die interaktive Clusterbildung und parallele Koordinaten, entwickelt, damit die Nutzer visuell potentiell interessante Cluster erkunden, die genauer auf Kartenansichten mit direkt-Rendering Methoden und im Raum-Zeit-Würfel untersucht werden können. Eine Auswahl an thematischen Visualisierungstechniken, die Dynamische Ablaufkarten, Punktstreuungskarten, temporale Matrizen und proportionale Symbolkarten umfassen, werden für zur effektiven Kommunikation von räumlichen Interaktionen und Taxifahrtverhalten eingesetzt.

Die visuell-analytischen Experimente, die die Techniken beider Disziplinen synthetisieren, sind in der Lage, z.B. Mobilitätsmuster, Landnutzungstypen und Bewe-

gengsmuster der Taxifahren zu entdecken, die mit einer einzigen Disziplin allein nicht möglich wären. Die Ergebnisse haben die Ausgangshypothese bestätigt, dass die beiden Disziplinen aufgrund ihrer komplementären Eigenschaften vorteilhafte Synergieeffekte erzeugen können.

Contents

Acknowledgements iii

Abstract v

Zusammenfassung vii

1 Introduction 1
- 1.1 Background 1
- 1.2 Research Tasks 3
- 1.3 Thesis Structure 3

2 Scientific Visualization vs. Thematic Mapping 5
- 2.1 A Comparative Study Framework 5
- 2.2 Purposes 6
- 2.3 Data sources 9
- 2.4 Georeferencing 10
- 2.5 Data preprocessing 16
- 2.6 Classification 19
- 2.7 Symbolization 22
- 2.8 Perception and Cognition 24
- 2.9 Summary 26

3 Movement Events and Shanghai Taxi Floating Car Data 29
- 3.1 Movement Events 29
- 3.2 Shanghai Taxi Floating Car Data 30
- 3.3 Derivation of Point-based Events 31
  - 3.3.1 Temporal Partition 33
  - 3.3.2 Spatiotemporal Partition 34
  - 3.3.3 Data Normalization 35
- 3.4 Derivation of Trajectory-based Events 39
  - 3.4.1 Categorization of Trajectory-based Events 39
  - 3.4.2 Origin-to-Destination Event and Non-occupancy Event 42
  - 3.4.3 Derivation of the Income of Taxi Drivers 44

4 Visual Analysis of Point-based Events **47**
4.1 Related Work 47
4.2 Visual Exploration by Means of Scientific Visualization 48
4.2.1 Pie Radar Glyph 49
4.2.2 Time Graph 53
4.2.3 Salient Feature Image 55
4.3 Visual Communication based on Thematic Mapping 61
4.3.1 Cluster Analysis 61
4.3.2 Spatial Partition and Aggregation 63
4.3.3 Univariate Choropleth Mapping 64
4.3.4 Thematic Mapping of Temporal Changes 67
4.4 Analysis of the Visualization Results 69
4.5 Comparison and Synthesis of the Visualization Methods 75
4.6 Summary 81

5 Visual Analysis of Trajectory-based Events **83**
5.1 Related Work 83
5.2 A Visual Analytics Framework for Trajectory-based Events 85
5.2.1 Visual Querying of Movement Database 87
5.2.2 Interactive Visual Clustering and Aggregation 88
5.2.3 Visual Representations 89
5.3 Origin-to-Destination Event Visualization 92
5.3.1 Exploration by means of Scientific Visualization 93
5.3.2 Communication via Thematic Mapping 101
5.4 Non-Occupancy Event Visualization 105
5.4.1 Exploration by means of Scientific Visualization 106
5.4.2 Communication using Thematic Mapping 110
5.4.3 Analysis of the Visualization of Non-Occupancy Events 120
5.5 Summary 121

6 Conclusion and Outlook **123**
6.1 Conclusion 123
6.2 Outlook 124

References **125**

Relevant Publications **135**

Curriculum Vitae **137**

List of Figures

2.1 The specific aspects in the visualization pipeline involved in the comparative study. 6
2.2 Illustration of the purposes of thematic mapping and scientific visualization in the map use cube. 8
2.3 The comparison of thematic mapping and scientific visualization in terms of data sources. 14
2.4 Examples of georeferencing in different visualization techniques. 15
2.5 A choropleth map with three ordinal market values. 18
2.6 Exemplary classification approaches applied in choropleth mapping, point map, and tree map visualization. 21
2.7 Four simulated gray scale images. 23
2.8 A square icon with four equal sub-squares. 24
2.9 A compact presentation of the 4-variate color icon. 24
2.10 A compact presentation of the four variables by the parallel coordinate system technique. 25
2.11 Four choropleth maps showing distributions of the mean values of four variables. 25

3.1 GPS points of the taxi with ID 10003 on the 12th May, 2010. 32
3.2 The focus + context hourly temporal variation of the total numbers. 33
3.3 The hourly temporal variation of the total amount of different events. 34
3.4 The frequency distribution of the total numbers of (O, N, P, D) events in each cell. 37
3.5 The spatial distribution of normalized events (O, N, P, D) after the logarithm transformation. 38
3.6 The frequency distribution of the daily number of reconstructed occupied trajectories. 42
3.7 The reconstructed occupied trajectories of 100 cars on 10 May 2010 and their corresponding origin-destination lines. 43
3.8 The distance frequency distribution and the logarithm of the occupied trajectories on 10 May 2010. 44
3.9 Distribution of taxi incomes. 45

4.1 A pie radar glyph. 49

4.2 The pie radar glyph visualization results. 51
4.3 Temporal distributions of (O, N, P, D) events using the pie radar glyph. . . 52
4.4 A schematic example of a time graph for a variable. 53
4.5 The temporal patterns of the four event types. 54
4.6 Salience-based map of four variables. 56
4.7 The salience-based spatial distribution of pick-up and drop-off events at (a) 7-8h and (b) 18-19h. 57
4.8 The color scheme for temporal change from 7-8h to 18-19h. 58
4.9 The salience change map. 59
4.10 Regions of Interests. 60
4.11 Hierarchical clustering results. 63
4.12 The Voronoi tessellation. 64
4.13 The univariate choropleth maps of (O, N, P, D) events. 66
4.14 Bivariate proportional symbol mapping of the spatial distribution of (P, D) events. 67
4.15 Choropleth mapping of salient features of (P, D) events. 68
4.16 Temporal change map of salient features of (P, D) events from 7-8h to 18-19h. 69
4.17 Spatial density distributions of (O, N, P, D) events. 72
4.18 The temporal patterns of the events in Figure 4.5. 73
4.19 The spatiotemporal patterns of (O, N, P, D) events using different methods. 74
4.20 The temporal changes using (a) salience-based method and (b) choropleth mapping. 75
4.21 The previous process and visualization results from high to low level of detail. 77
4.22 The comparison of the temporal change of the salient variable. 80

5.1 The framework for visual analysis of movement trajectory-based events. . . 86
5.2 The histogram of taxi drivers' income. High income bars are selected. 87
5.3 The time line view for brushing specific time window. 88
5.4 Interactive clustering interface for feature selection and distance setting. . . 93
5.5 The parallel coordinates view. 95
5.6 The map view of the selected $(o{-}d)$ lines. 97
5.7 Significant clusters of destinations at 7h-10h on 31 May 2010. 98
5.8 Interactive clustering of origins from 07:00 to 10:00 on 31 May 2010. 99
5.9 Large clusters of origins at 7h-10h on 31 May 2010. 100
5.10 Exploration based on different distance values showing the distance decay effects. 100
5.11 The placement of the line symbols. A1 and A2 are two regions. 102
5.12 Flow maps of origin-to-destination events. 103
5.13 Comparison of a place from Figure 5.12(a) and Figure 5.9(d). 105
5.14 Non-occupancy event clusters based on "starting location" in the parallel coordinates and significant clusters. 107
5.15 Identified transport hubs. 108

5.16 The non-occupancy trajectory-based events related to Pudong international airport on 31 May 2010. 109
5.17 The procedure of detecting the stationary spots. 111
5.18 The average hourly cruising duration for top and bottom income groups. . 113
5.19 The average daily spatial mean centers of the cruising trips and their standard deviations. 114
5.20 Temporal patterns of stationary events. 115
5.21 Comparison of the spatiotemporal distribution of the cluster centroids of the stationary spots. 117
5.22 Comparison of the spatiotemporal distribution of the stationary spots with space-time cube. 118
5.23 Comparison of the spatiotemporal distribution of the parking and traffic congestion places. 119
5.24 The average stationary duration spent on the traffic congestion and parking places (waiting for passengers) for top and bottom taxi groups. 120

List of Tables

2.1 The nature of thematic mapping vs. scientific visualization. 9
2.2 Georeferencing in thematic mapping vs. scientific visualization. 13
2.3 The data preprocessing for thematic mapping vs. scientific visualization. . . 18
2.4 Classification in thematic mapping vs. scientific visualization. 20
2.5 The symbolization in thematic mapping vs. scientific visualization. 23

3.1 Test data properties. 31
3.2 Four types of point-based events (O, N, P, D). 32
3.3 Normalization of numbers of events. 36
3.4 Categorization of trajectory-based events using abstraction levels. 40
3.5 Categorization of trajectory-based events using attributes. 40
3.6 The numbers of the occupancy and non-occupancy trajectory-based events. 42
3.7 The taxi fare calculation system of Shanghai in 2010. 45

4.1 Comparison of the aspects of different visualization methods. 79

5.1 The numbers of the occupied, cruising trips and the stationary spots. 111
5.2 Visualization techniques for trajectory-based events. 122

1 Introduction

1.1 Background

In an information society, large amounts of data are being captured or produced on a daily basis using various technologies in many domains such as remote sensing, computational simulations and location-based services. The exponentially growing data are mostly heterogeneous and multi-dimensional. Getting insights into these data and discovering interesting patterns are demanding and challenging for both data suppliers and data consumers. Visualization that unites the human vision, domain expertise and computational power is an effective means that helps to turn the large data volumes into meaningful information (interpreted data) and, subsequently, into knowledge (understanding derived from integrating information) [MK01].

The idea of using visual representations for humans to interpret and understand datasets has a long history and is well reflected in cartography which is dedicated to the visualization of the earth's surface. Only until recent decades, with the rapid development of computer technologies, the advanced hardware and software have made it popular to render large datasets on screens. System developers can develop complicated interactive tools to visualize as much data as possible, for instance, multimedia cartographic information systems emerged around 2000s benefit from multimedia technologies [Buz01]. To the one end, users, particularly system developers and domain experts, enjoy a lot of freedom to explore the unknowns hidden in the graphics. However, too much information embedded in a graphic does not necessarily mean a concise design with more knowledge. Ill-designed visualization systems with a visual overload would confuse domain specialists or even system developers themselves. To the other end, there is a growing demand on the communication of the known scientific results to general users (e.g., political decision makers, publics), which requires that visualization results should be more easily understandable and synthesized.

Facing these challenges, joint efforts from multiple disciplines, such as cartography, geovisualization, scientific visualization, and information visualization, are demanded. In particular, thematic mapping and scientific visualization are two dedicated areas for visualizing a wide range of domain-specific datasets. Since its emergence in mid-17th century, thematic mapping methods have been widely applied to show the spatial patterns of one or more geographic attributes or variates. The term "mapping" in cartography mainly refers to depicting geographic space, although thematic maps can also be used as metaphors to represent non-geographic space. Thematic mapping methods are mostly computerized as a consequence of the technological (r)evolutions. Outside cartography, the term "visualization" is widely adopted in computer sciences, especially in scientific visualization that emerged in late 1980s and addresses all kinds of data. By taking the advantage of human-computer interactions, scientific visualization is designed for the exploration of the unknowns in large multivariate datasets.

From a general perspective, thematic mapping and scientific visualization reveal large overlaps in terms of visualization methods. However, due to diverse historical developments and disciplinary focuses, the concerns involved in the thematic mapping workflow that ranges from data sources, data processing to map design principles may be quite different from those in scientific visualization, and vice versa. With the increasing share of the rapidly developing technologies and the necessity to handle extremely large datasets, the two disciplines now tend to increasingly converge. Current cartography-related systems, such as GeoVISTA Studio software [TG02] and V-Analytics (formally called CommonGIS) software [AA04], have integrated visualization techniques from scientific visualization and thematic mapping to explore large geospatial datasets and allow both overview and detailed views. Some interdisciplinary researches between general mapping and scientific visualization have been conducted. The widely used concept spatialization in scientific visualization [Che06] as well as in cartography [SF03] are an example. However, little knowledge about the differences and commonalities between thematic mapping and scientific visualization is available. To avoid reinventing techniques and raise awareness of the overlapping areas, an in-depth comparative study is necessary.

This thesis aims to bridge the gaps between thematic mapping and scientific visualization by identifying their methodological strengths and achieve their synergetic effects for the visual analysis of big data. To establish a theoretical foundation, a systematical comparative study of those two disciplines is essential.

Moreover, the theoretical findings of the comparative study motivate us to carry out extensive experiments for visual analysis of massive and complex data. Ideally, the expected test data for the experiments should be real world data of large volume and complex structures. The Floating Car Data (FCD) satisfies all the requirements and will be served as test case in this thesis. Indeed, FCD analysis, in general movement data analysis, is currently a hot research topic with numerous studies focusing on modeling human mobility patterns [GHB08; Zhe+08; DYM15], uncovering taxi driving behaviors [LAR10; DFM15], mining interesting locations or places [Zhe+09; And+11; And+13b], and inferring urban land uses and city structures [Liu+15a], etc. We believe that insights can be derived from the movement data by taking advantage of the complementary characteristics from scientific visualization and thematic mapping.

1.2 Research Tasks

This thesis strives for an interdisciplinary research that bridges the gap between thematic mapping and scientific visualization. It involves the following research tasks:

- Identify the complementary characteristics of thematic mapping and scientific visualization by investigating their commonalities and differences.
- Demonstrate the necessity and feasibility of achieving synergetic effects for visual analysis of large geospatial database by taking advantage of the two disciplines.
- Identify use cases of real-world movement data for experimenting the synergetic effects.
- Propose concrete scientific visualization and thematic mapping techniques for visual exploration and communication of movement events.
- Implement and evaluate the proposed visualization methods.

1.3 Thesis Structure

The thesis is structured as follows.

- Chapter 2 is dedicated to the theoretical foundation for bridging the gap between scientific visualization and thematic mapping. The author first conducts a systematical comparative study of these two discipline. Following the order in the visualization pipeline, the comparison covered seven essential aspects: purposes, data sources, geo-referencing, data preprocessing, classification, symbolization, and finally perception and cognition.
- Chapter 3 introduces the test datasets and data preprocessing steps. Massive and complex real-world taxi floating car data (FCD) collected in Shanghai in the year 2010 are used for conducting extensive experiments. According to an event-based view, the author first categorizes the movement events and then identifies two types of events: point- and trajectory-based events, for visual analysis.
- Chapter 4 deals with the visual exploration and communication of point-based event patterns. From the scientific visualization perspective, we propose pie radar glyphs, time graphs and a salience-based method for effective visual exploration and analysis of four types of point-based events, referred as *(occupancy, non-occupancy, pick-up, drop-off)*. From the thematic mapping perspective, choropleth mapping and proportional mapping techniques are applied to communicate spatiotemporal patterns.
- The trajectory-based events are analyzed in Chapter 5. Using scientific visualization methods, highly interactive techniques are employed, including interactive cluster-

ing methods, the parallel coordinates approach, gradient and direct line rendering techniques, and space-time cube, to let users visually explore the spatiotemporal patterns of potentially interesting clusters. A variety of thematic mapping techniques, including flow mapping, dot mapping, temporal matrix, and proportional mapping, are proposed for the communication of the spatial interactions and taxi drivers' behavior patterns.

- Chapter 6 concludes the thesis and discusses the future work.

2 Scientific Visualization vs. Thematic Mapping

This chapter is dedicated to the theoretical foundation for bridging the gaps between scientific visualization and thematic mapping. We conduct a systematical comparative study of these two disciplines following the order in the visualization pipeline, which covers seven essential aspects: purposes (Section 2.2), data sources (Section 2.3), georeferencing (Section 2.4), data preprocessing (Section 2.5), classification (Section 2.6), symbolization (Section 2.7), and finally perception and cognition (Section 2.8). In Section 2.9, we conclude that the comparison results show that these two disciplines reveal different visual analytical levels and play a complementary role with each other. These results suggest the necessity and feasibility of achieving synergetic effects from two disciplines by taking advantage of their complementary characteristics.

2.1 A Comparative Study Framework

Thematic mapping and scientific visualization are two dedicated areas for visualizing a wide range of domain-specific data sets. From a general perspective, thematic mapping and scientific visualization reveal large overlaps in terms of visualization methods. However, these two disciplines are shaped by different driving forces and usage purposes. Accordingly, the concerns involved in the thematic mapping workflow that ranges from data source, data processing to map design principles may be quite different from those for scientific visualization and vice versa. Although some interdisciplinary researches between general mapping and scientific visualization have been conducted [SF03; Che06], little knowledge about the differences and commonalities between thematic mapping and scientific visualization is available. This chapter is devoted to an in-depth comparative study with the aim at bridging these two disciplines by identifying their methodological strengths and synergetic effects.

Thematic mapping and scientific visualization both cover a wide variety of aspects and issues such as those related to static vs. dynamic visualization and human-computer interaction. It is difficult to list and compare their characteristics exhaustively. In this study, we will tackle some essential aspects that could demonstrate their emphases, commonalities and distinctions. This section introduces a framework that confines the overall spectrum of this comparative study.

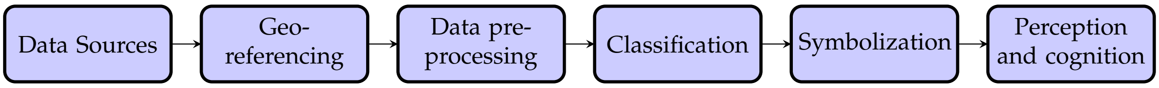


Figure 2.1: The specific aspects in the visualization pipeline involved in the comparative study.

Section 2.2 firstly gives an overview about the two disciplines by examining their purposes, disciplinary natures and different visualization guidelines. Detailed comparisons on specific aspects of these two fields will be unfolded mainly according to the visualization pipeline. The term "visualization pipeline" used in scientific visualization describes the (step-wise) process of creating visual representations of data, which mainly consists of data filtering, data mapping and data rendering. In thematic mapping, the corresponding term is "map making", which is a creative procedure from raw data to cartographic representations. The thematic mapmaking procedure normally contains steps such as data preprocessing, classification and symbolization. Generally speaking, "data filtering" is a sub-step of "data preprocessing"; "data mapping" is synonymous to "symbolization". To consistently compare different aspects, this work adopts terms and processes involved in thematic mapping. More specifically, we compare the following aspects in the visualization pipeline: data sources (Section 2.3), georeferencing (Section 2.4), data preprocessing (Section 2.5), classification (Section 2.6), and symbolization (Section 2.7). Figure 2.1 illustrates these involved aspects in the comparative study framework.

Beyond the visualization pipeline, the user interaction with the visual representations or cartographic representations is also an important aspect and a rather complicated creative process. In Section 2.8, we address the generic perception and cognition issues that demonstrate users' understanding of different visual displays and possible influences on their behavior. At this moment, extensive experiments on (map) reading of the visualization results generated from thematic mapping and scientific visualization using the same dataset are not discussed here because of the theoretical nature of this section.

2.2 Purposes

Visualization serves for a variety of purposes and functions. For example, DiBiase [DiB90] models map-based scientific visualization as a four-stage process that starts with exploration, moves to confirmation, transitions to synthesis, and ends with presentation. MacEachren and Kraak [MK97] developed the space-cube representation for map-use ranging from private-high interaction-exploration of the unknown corner to public-low interaction-presentation of the known corner. However, distinguishing those visualization stages and use goals does not mean there are crisp boundaries among them. In fact, exploratory visualization is also capable of communicating useful results, while visualization for presentation may promote new insights as well. Rather, the stress is to apply different visualization strategies according to different purposes.

Thematic mapping emerged as a response to the growing knowledge of natural sciences and the intensified exploration of our living planet in the 17th century. In its early stage of development, it was focused on the communication of known information from the map designer to the map user. Effortful cognitive processes are involved in each step of the map design procedure. For example, to effectively portray the phenomenon of interest, cartographers must decide upon the appropriate spatial level of detail. In most cases, cartographers would aggregate data to certain geographic units, which involves both spatial and semantic aggregations that largely reduce the data amounts and reveal meaningful information. Besides, cartographers choose appropriate classification and symbolization methods so that the map information can be conveyed to users without their considerable cognitive effort. For example, most thematic maps show data with no more than seven classes. When reading a thematic map, map users are supposed to be able to understand the meaning of each symbol immediately. With the digital (r)evolutions, thematic mapping tends to be increasingly democratized, allowing the users to takes over more analytical and explorative responsibilities. This has added to the thematic mapping a flavor of scientific visualization. In particular, with the embedded analytical functions in an interactive mapping environment, users may not only efficiently receive the known information, but also explore the unknown information. For example, in modern GIS systems with various specialized functions, map overlay is dedicated to exploring and detecting correlations of multiple layers. Nevertheless, the individual layers are still thematic maps conveying the known qualities or quantities of the underlying themes. In a similar way, most elaborately designed thematic maps in Atlas information systems serve as information sources and their designer bares the responsibilities of communicating the processed information to the viewer who but enjoys the freedom of interactively sifting through the maps in order to detect information beyond what is known to the system developer. In this scenario, the communication process in thematic mapping can be regarded as a push-based approach that cartographers take more responsibilities, hence the potential of making mistakes as well.

The purpose of scientific visualization is to provide a thinking instrument that helps domain specialists to form and verify hypotheses, to explore the unknown information hidden in the graphics and understand scientific datasets. Scientific visualization systems provide the possibilities to cope with massive data and facilitate exploration through human-computer interactions. Driven by the explorative nature, scientific visualization does not usually generalize data as thematic mapping does. Its data preprocessing mainly deals with the normalization of data instead of replacing the original data with synthesized values. Semantic aggregation is often applied to reveal the distributional patterns or relationships of the data items, whereas spatial aggregation is rarely applied so that individual-level data could be further explored. Since there is no much spatial aggregation, symbols (e.g., glyphs, icons) of individual data items are rendered at their representa-

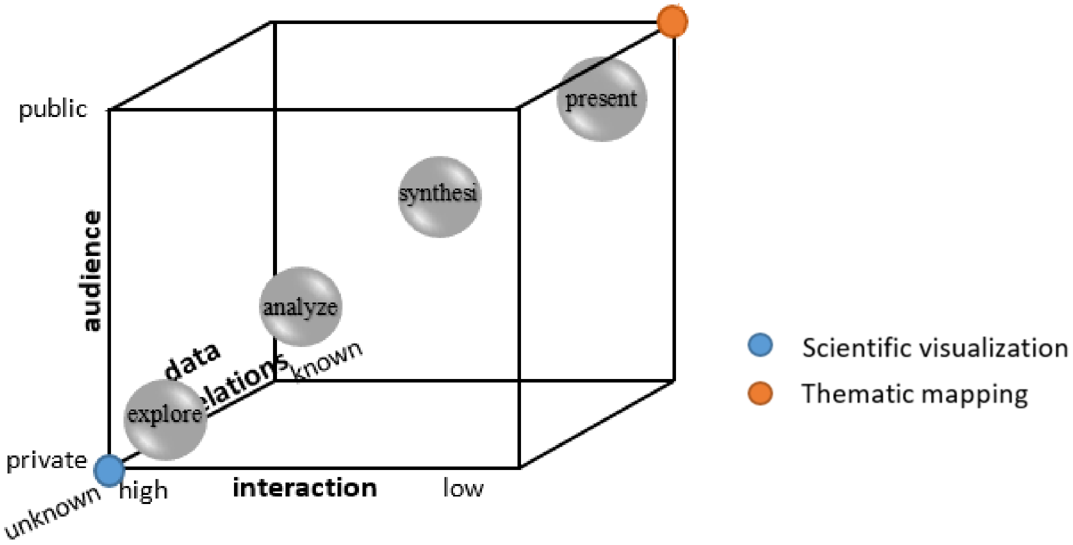


Figure 2.2: Illustration of the purposes of thematic mapping and scientific visualization in the map use cube.

tive locations. Consequently, scientific visualization often results in a visual representation of compact symbols. Its whole structure is obvious while the individual symbols are indistinguishable. Interaction techniques are often used for information retrieval and display, and most of the information is just temporarily visualized on demand by the user [Buz99]. Being characterized by its high interactivity, scientific visualization systems then allow users the maximum freedom to manipulate the visualization, such as zooming into interesting hotspots, querying exact data values by selection, extracting cluster boundaries or the distribution structures. Since the final analytical step is left to users, users rather than system developers must spend significant cognitive efforts to analyze the visualization results. Instead, system developers pay much attention to develop advanced computational algorithms to cope with mass data and enhance the rendering performance. This kind of pull-based interaction style also indicates that users rather than system developers or designers would run the risk of making mistakes and getting lost in the information space.

From the aforementioned paragraphs, we can learn that thematic mapping is based on the communication model and emphasizes the communication of known information from the map maker to the map user via information abstraction. In contrast, scientific visualization is based on an exploration model and emphasizes the exploration of unknown information. Their distinctive purposes and natures reveal that the two disciplines can be regarded as locating at the two corners of the well-known map-use-cube as illustrated in Figure 2.2. We summarize the detailed comparisons of the nature of thematic mapping and scientific visualization in Table 2.1.

	Thematic mapping	Scientific visualization
Theory	Communication model	Exploration model
Design style	Push-based	Pull-based
Information abstraction degree	High	Low
Cognitive efforts of the developer	High	Low
Cognitive efforts of the user	Low	High
Main responsibility	Map makers	Users

Table 2.1: The nature of thematic mapping vs. scientific visualization.

2.3 Data sources

Thematic mapping and scientific visualization both cover a variety of raw data sources collected or created by advanced sensors, computational simulations, financial transactions, etc. In thematic mapping, the data source could be individual data items but more likely spatially aggregated data, such as national census data. There are a variety of reasons using aggregated data. For example, national agencies mostly provide aggregated data to ensure personal confidentiality. For thematic mapping, even if massive raw data scattering over the space are available, it is necessary to partition the space into smaller units and aggregate the individual data items within each unit to one single value in order to summarize information and reach a more effective communication.

However, it is essential for scientific visualization starting with individual raw data items, because besides a global picture of the spatial and temporal behavior of the data, scientific visualization should allow interactive exploration of the details. Raw data are normally associated with relatively accurate positions. Data processing in scientific visualization should also keep the data as accurate as possible. That means, analytical algorithms can help reduce data noise and enrich data by means of semantic aggregation, but the spatial resolution or the level of detail remains unchanged.

For example, remote sensing imagery serves as an abundant data source for both disciplines. In most cases, scientific visualization starts from the raw remote sensing images, and processes the imagery data, e.g., by classification, segmentation, based on the semantic meaning of each pixel. The spatial aggregation is not applied and the image resolution does not change. The well-classified images reveal the underlying data distribution and the global structure (sometimes with extracted boundaries based on pixel semantics). In thematic mapping, those well-classified images are often taken as its raw data or background rather than the final products. Classified images can be further aggregated or (re)sampled to lower resolutions or to geographic areas (e.g., states, districts, blocks). Classification and symbolization then could be applied to show patterns of these areas.

In this study, we use Financial Times Global 500 2012 as a test dataset to illustrate the characteristics of thematic mapping and scientific visualization. This dataset provides an annual snapshot of 500 individual companies with several indicators, e.g., company name, sector, market value, and the number of employees. Fictional areas are designed to represent geographic areas where the companies are located. Each company is assigned to a position inside one of these areas. While a thematic map usually displays the spatially aggregated or classified data as shown in Figure 2.3(a) with one area being selected, the scientific visualization may tend to choose a direct rendering of the raw dataset containing the individual attribute values of 500 individual companies shown in Figure 2.3(b). One single company is highlighted. Figure 2.3(b) can also be regarded as a pre-stage for Figure 2.3(a).

We can conclude from the aforementioned examples that the significant distinction of the two fields in terms of data sources is that scientific visualization mostly starts from individual raw data items while thematic mapping could start from individual raw data items but mostly from spatially aggregated data.

2.4 Georeferencing

Georeferencing explicitly or implicitly relates data such as world events and placenames to geographic locations. Georeferencing theories and techniques are essential for cartography, geography and related disciplines especially when data from different sources need to be combined. In this section, we discuss the differences of georeferencing in thematic mapping and scientific visualization.

Georeferencing in thematic mapping

Thematic data in thematic mapping are inherently geo-referenced to locations (e.g., enumeration units) in a 3-D geographic space (i.e. geospace). Geospace is a continuous physical space, in which location, distance, orientation, area, size and so forth can be measured based on physical or functional measurements. For instance, distance on thematic mapping can be physical Euclidean distance or time distance; regions can be homogeneous zones based on physical objects (e.g., forest, lake) or functional regions (e.g., administrative units).

In thematic mapping, geospatial features are mostly represented by geometrical primitives, points, lines, polygons (areas) and volumes. Vector data model is used to represent those points, lines and polygons as a series of coordinates (e.g., longitude and latitude), and may also store topological information that describes the relative positions of objects to each other. Irregularly shaped administrative and statistical base areas, which are mainly used for the aggregation of the attribute or thematic data, are easily represented by polygons. For example, U.S. census bureau offers TIGER/LINE shapefiles as a vector data format to represent thematic base data, such as coastlines, country borders, and census tract boundaries. Besides vector data model, raster data modal is another way to represent spatial objects, discrete or continuous, static or dynamic. Vector data can be overlaid on top of a raster image.

To represent the 3-D geographic space on physical planes (e.g., paper maps, computer screens), thematic mapping should appropriately choose map projections depending on the purposes. In addition to orthographic projections, central and parallel perspective projections widely used in map-related representations could also be applied. Geometric distortions are no longer the unwanted side effect [Men03]. Instead, map distortion or anamorphosis with functions such as value-by-area, poly-focal projection or dynamic lens is increasingly designed as an effective means to make thematic maps more attractive. During the distortions, invariance (e.g., topological relationships) would be preserved to make sure that no confusion occurs in the communication process. For example, in the choropleth map in Figure 2.4(a) the aggregated market values are geo-referenced to the fictional areas. The larger the quantity is, the darker the color tone. In the value-by-area map in Figure 2.4(b) the distorted fictional geographic areas are proportional to the corresponding market values, which may evoke the user's attention and curiosity.

Georeferencing in scientific visualization

In scientific visualization, two distinctive visualization categories can be identified in terms of spatial properties of the data. One category concerns on data with reference to the earth or a specific object in the physical space. The other category emphasizes on visualizing aspatial data.

Spatial data

This category concerns data with reference to the earth or a spatial object (e.g., human bodies in medical visualization) in a physical space. Here, we focus on measurable spatial data that are geo-referenced to the earth's surface with location, distance, orientation, etc. Remote sensing images, geo-located social media data (e.g., Geo-Twitter), floating car data generated from GPS-enabled taxis, etc. belong to this kind of geo-referenced data that are huge in amount and with a full coverage of a certain geographic region. These geo-referenced data could represent discrete entities or continuous fields and are encoded either in vector or raster format. Geospatial locations and attribute values occur in raster format in regular pixels and color tones. In vector format, they are typically expressed using primitive features such as points, lines and polygons which are not generalized as required in thematic mapping. For example, Figure 2.4(c) shows the individual market values of individual companies. The size of each green circle is scaled to reflect the relative market value of a company. Each circle is geo-referenced to a geographic location inside one polygon. Depending on the visualization purpose, some additional efforts of converting pixels or points (e.g., discrete measurements of continuous surfaces) to irregular vector shapes are necessary. Typical methods are triangulations, re-sampling on uniform grids and mesh grids. However, during these procedures, individual values should be kept as accurate as possible.

To represent the physical space on a screen, similar projection methods as in thematic mapping are also employed. While for the accuracy purpose, distortions should be minimized.

Aspatial data

Aspatial data do not have a spatial reference. For example, a set of publications is an aspatial dataset, in which each document contains information such as title, author, and publishing date. To allow a better comprehension, aspatial data with multiple attributes could be associated with a high-dimensional virtual space. Mapping an aspatial dataset on a spatial reference is termed as spatialization [SB97]. The imposed spatial metaphors, such as location, distance, size and orientation, make the structure and relations of aspatial attributes straightforward to see and to compute. For example, multidimensional scaling (MDS), latent semantic indexing (LSI), or statistical correlations can be used to locate the numerical data. Generally speaking, the relative locations on the base map are more important than the absolute locations and the distance indicates the strength of semantic relationships such as similarity.

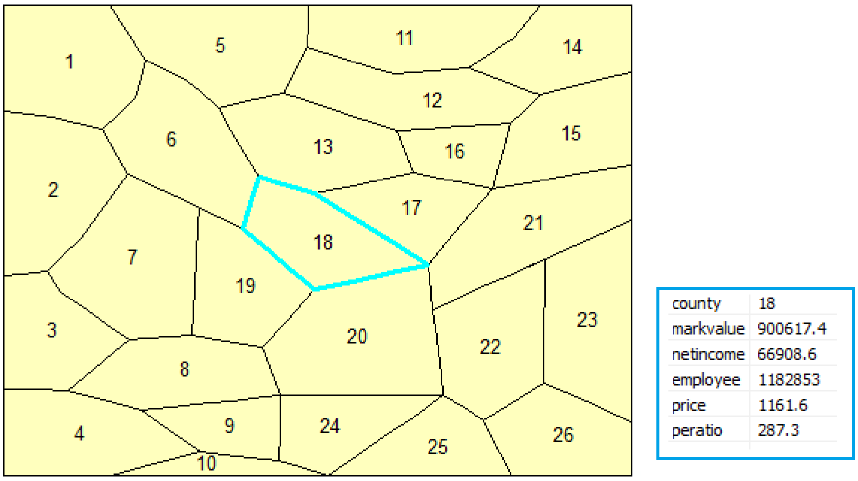
The spatialized virtual space is normally modeled by a directional graph composed of a set of nodes (vertices) and their attributes, as well as a set of links (edges). Nodes represent individual abstract objects, and attributes represent the properties of those objects. Links indicate object relationships that can be derived by means of statistical or mathematical data analysis methods. For instance, Figure 2.4(d) is a treemap visualization of the companies. Each rectangle is a node and represents a company. The rectangle sizes are proportional to the market values. The absolute locations of the rectangles have no geographic meaning while their relative locations shows the relationships (e.g., hierarchical relationship indicated by different color tones) among the rectangles. Fabrikant and Skupin [FS05] have theoretically investigated a series of spatial metaphors and empirically examined those metaphors in a number of experiments. Software programs utilizing spatial metaphors are also common, such as Citespace [Che06], Galaxy and ThemeView visualization [INS05]. Besides spatialization, distortion techniques, such as hyperbolic graph and perspective wall [MRC91], are deliberately applied to make visual representation more attractive by means of focusing on details while preserving an overview of data.

Concluding this section, we can say that in thematic mapping aggregated data in geospace

	Thematic mapping	Scientific visualization	
Space	3D geographic space	3D physical space	Hyper- or high dimensional virtual space
Spatial concepts (Location, distance, region, etc.)	Physical or functional measures	Physical measures	Spatial metaphors
Projections	Physical to physical	Physical to physical	Virtual to physical
Data model	Mostly vector data model	Both raster and vector data model	Directed graphs
Geometry	Irregularly or regularly shaped polygons	Grids (e.g., pixels, triangles) or primitive geometries representing individual objects (e.g., scatter points)	Any kind
Spatial aggregation	Common	Uncommon	
Pitfalls	Statistical pitfalls	No pitfalls	
Distortion	Geometries and metric relationships can be intentionally distorted in order to emphasize or preserve important thematic information	Absolute locational accuracy is preferred	Geometric distortion without physical meaning

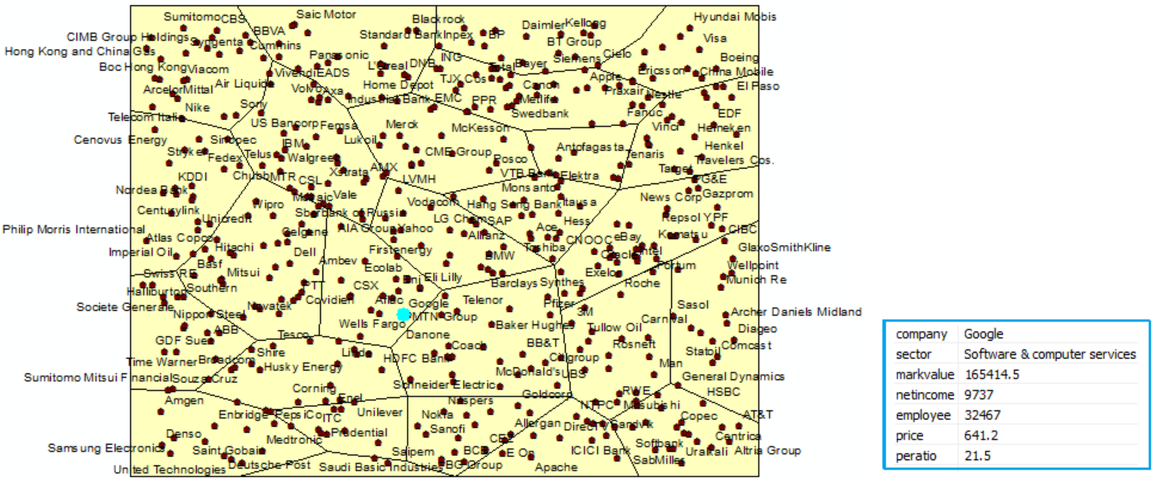
Table 2.2: Georeferencing in thematic mapping vs. scientific visualization.

are normally geo-referenced to irregularly or regularly shaped polygons, while in scientific visualization individual data items with spatial characteristics can be geo-referenced to accurate locations of individual objects or measurements in a 3-D physical space, and aspatial data in a hyper-dimensional space can be spatialized to a 2-D or 3-D physical space usually based on a graph model. More detailed comparisons between thematic mapping and scientific visualization in terms of georeferencing are summarized in Table 2.2.



county	18
markvalue	900617.4
netincome	66908.6
employee	1182853
price	1161.6
peratio	287.3

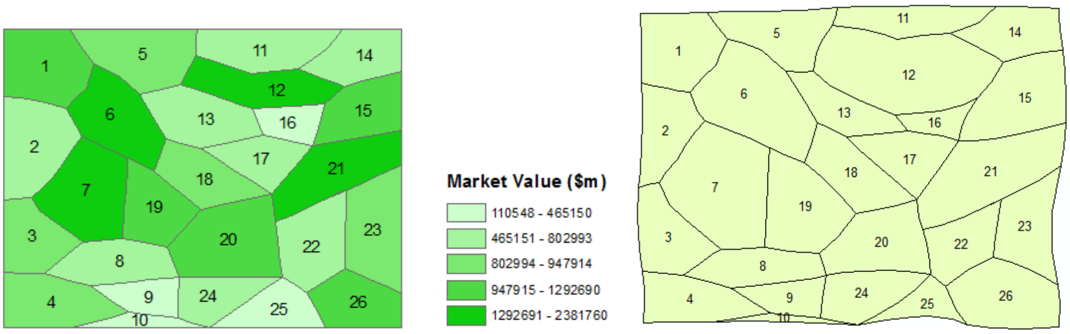
(a) Aggregated data of fictional areas in thematic mapping



company	Google
sector	Software & computer services
markvalue	165414.5
netincome	9737
employee	32467
price	641.2
peratio	21.5

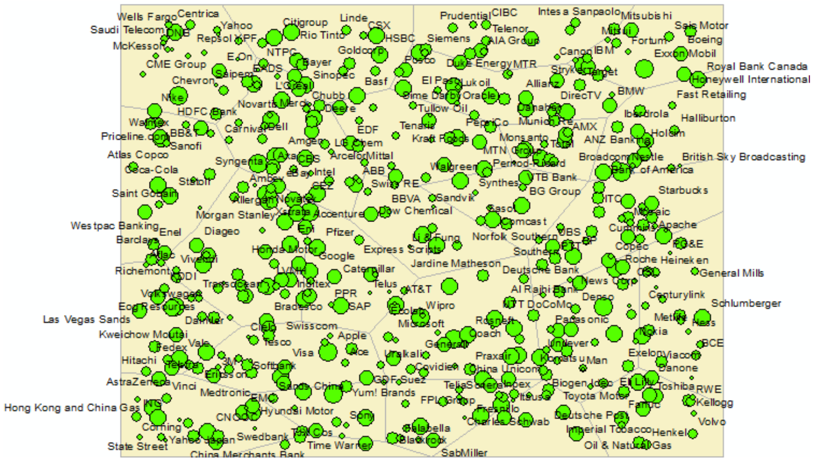
(b) Individual companies with their attribute values used in scientific visualization.

Figure 2.3: The comparison of thematic mapping and scientific visualization in terms of data sources.

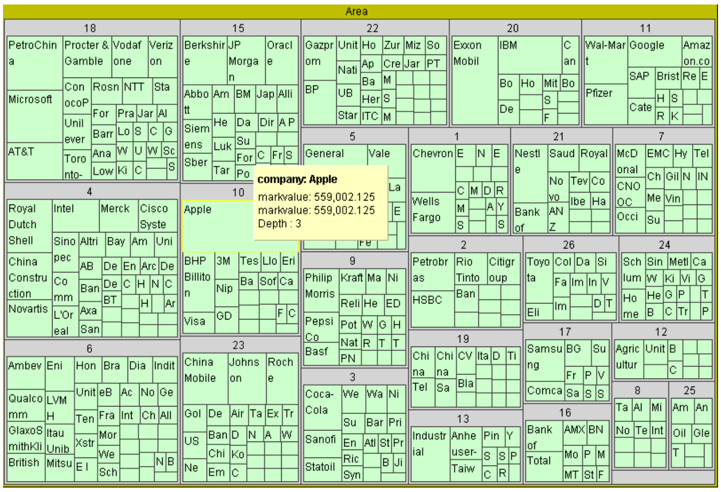


(a) A choropleth map with the aggregated market values geo-referenced to the fictional geographic areas

(b) A value-by-area map with the fictional areas proportional to the corresponding market values



(c) A point map of the geo-referenced individually scaled market values



(d) A treemap visualization with locations of rectangles bearing no spatial meaning, but sizes reflecting the relative market values

Figure 2.4: Examples of georeferencing in different visualization techniques.

2.5 Data preprocessing

In thematic mapping, data preprocessing aims to reduce the amount of data and reveal certain spatial patterns of the themes. Besides the regular data preprocessing steps like noisy data filtering, spatial and semantic aggregation are highly involved. Generally speaking, if the raw data source contains numerous individual data items, the very first step is to aggregate the data to an appropriate spatial level of detail. Then, the data could be scaled to certain level of measurements using spatial statistical techniques.

Spatial aggregation. Spatial aggregation is a rather subjective procedure. Cartographers may choose a specific spatial level that they believe fits the purpose of the map or the goal of the problem. The data can be spatially aggregated to a certain level of geographic areas (such as blocks, states, and countries) when an appropriate base map is accessible, or they can be aggregated to area units that are calculated based on their spatial distributions. For instance, when creating a point density map of incident data, the incident points are firstly clustered and conceptual points are generated as the geometric centers of the clusters. Then Voronoi diagram techniques can be applied to partition the space into polygons that contain a conceptual point each. Incident points are then spatially aggregated into these polygons. Finally, a choropleth map can be derived based on the Voronoi meshes. Sometimes, regular grids of relatively large cells are superimposed on the mapping area and data inside each grid are then aggregated.

Data scaling. A wide spectrum of techniques could be applied to scale geospatial data to nominal, ordinal, interval or ratio level of measurements. Descriptive statistical methods, such as the mode or chi-squared technique, are appropriate to scale count values to nominal level of measurement (e.g., race, ethnicity, marital status). Ordinal scaling assigns observations, usually collected from social surveys of preferences and perceptions (e.g., English-speaking ability, education attainment), to discretely ranked categories. Statistical techniques, such as median and percentile, could be useful in dealing with ordinal data. For interval and ratio scaling, more descriptive as well as inferential statistical methods could be applied, such as mean, standard deviation, correlation, regression, and analysis of variance. Besides, geostatistical techniques and space-related data analysis methods are also well developed. Depending on the level of measurements, different analytical operations, such as grouping, isolation, cross-tabulation, differentiation, and classification, could be applied [Chr02]. Data scaling often leads to a reduction of data amount and the improved insight into the datasets. For example, in Figure 2.5, the aggregated market values of the fictional areas can be further projected onto an ordinal level, which leads to the loss of the numerical values of the fictional areas but helps reveal the relative quantities in a more abstract and legible manner.

Pitfalls induced by spatial aggregation. Since thematic data are mostly aggregated to the geo-referenced areas, the sizes and shapes of the areas may exert pronounced influences on the accuracy of derived values, thus the accuracy of corresponding thematic maps. In

[Mac85] and [Sch55; HXR70], size and shape effects have been investigated on the accuracy of choropleth and isopleth mapping respectively. More seriously, aggregated data could also induce spatial data analysis pitfalls. Typical pitfalls are modifiable unit area problems, edge effects, ecological fallacy, etc. For example, edge effects may arise where an artificial boundary is imposed on a study area, and are particularly prominent in studies of spatial point patterns [FR93]. Therefore, map makers should be aware of these limitations and try to maximize the thematic map accuracy when designing a thematic map, such as by choosing an appropriate form of symbolization. Meanwhile, map users could be alerted of the possible pitfalls (such as provided with the measure of the thematic accuracy), and take caution in interpreting the insufficiently accurate thematic map.

In scientific visualization, data preprocessing emphasizes on normalizing attribute data and data enrichment via various data analysis methods. Data normalization (or standardization) is widely applied due to the fact that measurement scales are not uniform. For example, temperature could be measured in both Fahrenheit and Centigrade. Both are valid, but they produce different numbers. The measurements need to be scaled to the same "neutral" or "standard" level to allow the comparison. Normalization techniques are different, depending on whether the variables are scaled at nominal, ordinal, interval, or ratio levels. For example, nominal (e.g., different naming systems for flowers in different areas) and ordinal variables (e.g., clothes sizes in different countries) could be normalized by looking up standardization tables; linear or some non-linear (e.g., logarithm) transformation methods are often used for normalizing interval and ratio data. In general, normalized data are at the same level of the original data and the accuracy of the original data is kept as much as possible. Besides data normalization, other data analysis methods could be used to derive data values for data enrichment. For example, in Citespace [Che06], besides the original data (e.g., author, title, citations), derived values, such as frequency and betweenness centrality, are also calculated using statistical methods, text mining algorithms or graph analysis methods. Note that the derived data are intended to enrich the original data rather than to replace them. Therefore, the data preprocessing in scientific visualization results in more accurate and enriched data. Since data are not spatially aggregated as required in thematic mapping, statistical pitfalls are also avoided. However, interpreting the less generalized dataset demands for more cognitive endeavors.

In short, through data preprocessing, data items in thematic mapping are normally scaled to more abstract levels of measurement based on semantic aggregation and the data amounts can be largely reduced through spatial aggregation which may induce some statistical pitfalls; whereas in scientific visualization raw individual data items are normalized and filtered for the accuracy purpose based on semantic aggregation. Table 2.3 lists the compared results of data preprocessing between thematic mapping and scientific visualization.

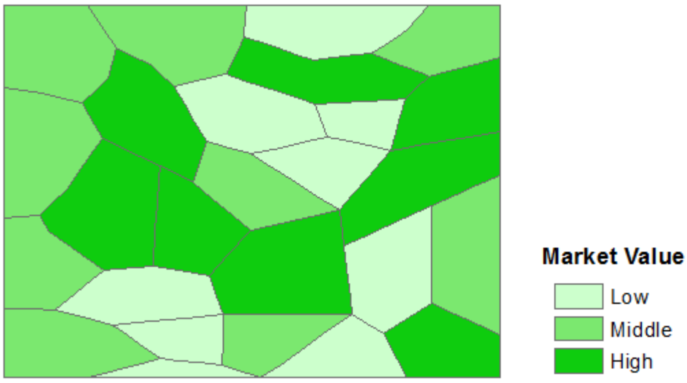


Figure 2.5: A choropleth map with three ordinal market values.

	Thematic mapping	Scientific visualization
Preprocessing techniques	Mostly geostatistical techniques and space-related data analysis methods	Mostly data normalization (e.g., statistics, image processing, graph analysis)
The scaling of measurement	The scaling at the same level or more aggregated level	Mostly the same level of scaling
Data amount	Largely reduced	Rarely reduced
Derived data	May replace the original data	As additional data
Original data accuracy	Decreased	Kept
Information abstraction	High	Low
Cognitive effort	Low	High

Table 2.3: The data preprocessing for thematic mapping vs. scientific visualization.

2.6 Classification

In thematic mapping, classification allows the designer to structure the message of the thematic communication [Den90]. The quantitative information, being directly measured (e.g., total population) or derived (e.g., population density), is usually classified before its symbolization in a thematic map. Theoretically, accurate classes that best reflect the distributional character of the dataset can be calculated. However, thematic mapping should also take the perceptual constraints into account by restricting the number of classes, thus allows a quick and immediate overview and interpretation of spatial patterns. In addition, a thematic map is also assumed to permit its readers to identify exactly the class to which each individual symbol belongs [Dob73]. Therefore, thematic maps are normally with a limited number of classes that fits for perception. Note that the perception-driven classification is different from the data-driven scaling process that aims to reflect the data distribution. The widely adopted principle in thematic mapping is that no more than seven classes are used. Under this constraint, optimal classifiers, which strongly affect the visual impression of a thematic map, are applied to determine the class intervals and class boundaries. Extensive empirical case studies on classification issues have been conducted, such as the determination of an appropriate number of classes for human perception [JK03], the selection of class intervals and boundaries [Eva77; Har03; GB09; Fis10].

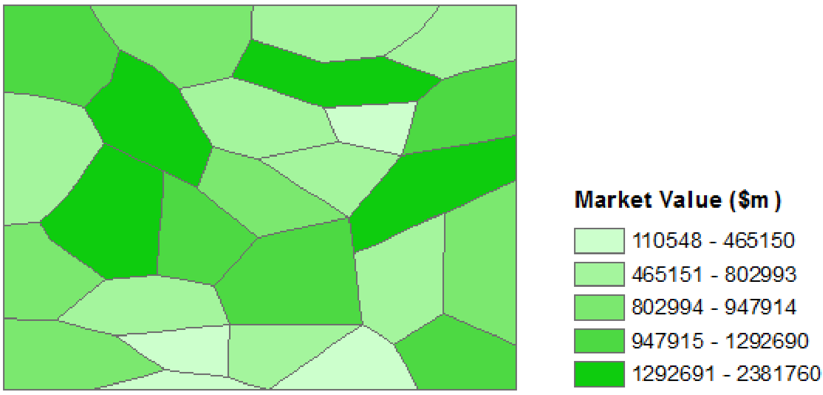
In scientific visualization, accurate information should be preserved for the purpose of future exploration. According to the goodness measures, a variety of mathematical and statistical techniques are employed to find optimal classifiers, which assure that each individual data item precisely falls into a certain group. It is often the case that more than 10 categories are classified. For example, satellite images are segmented and classified to many land use types. Sometimes, data values are merely normalized and mapped to graphic variables without classification. For example, normalized data values in stick figure icon visualization are directly mapped to corresponding stick angles [PG88]. Those accurately classified groups or unclassified individual data can be helpful for the further exploration but may call for longtime interpretation or understanding.

Figure 2.6 illustrates the classification emphasis of the two disciplines by using three visualization techniques. Figure 2.6(a) shows a choropleth map with 5 classes. In Figure 2.6(b), the color tones of the circles are scaled to reflect the relative market values with more than 20 classes along the color bar. The treemap visualization in Figure 2.6(c) uses unclassified market values to calculate the sizes of the individual rectangles. The changing darkness of the color tones reflects the quantity of each company's net income value (the larger the quantity, the darker the color tone). Obviously, the more generalized thematic map with 5 classes leads to the loss of more detailed information, and reduced data accuracy. However, users can comprehend the classified results more easily.

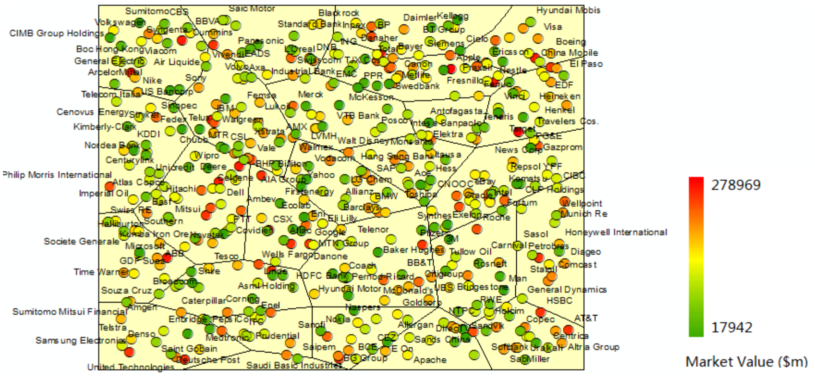
We can summarize from the aforementioned paragraphs that classification in thematic mapping normally has more constraints from perception consideration that can largely affect the visual impression of a thematic map; while classification in scientific visualization try to make sure that each of the individual data items can be optimally classified to a group. The comparison of classification between thematic mapping and scientific visualization is illustrated in Table 2.4. For better exploration as well as user-centered effective perception in the visualization systems, a trade-off of classification is required between the fitness for the natural distributions and the fitness for human perception.

	Thematic mapping	Scientific visualization
Necessity	Mostly classified	Classified or unclassified
Driving force	Fitness for perception	Goodness measure reflecting the distributions
Number of classes	Usually <7	Can be very large (e.g., >10)
Evaluation	Readability and legibility	Accurate classification of individual data items

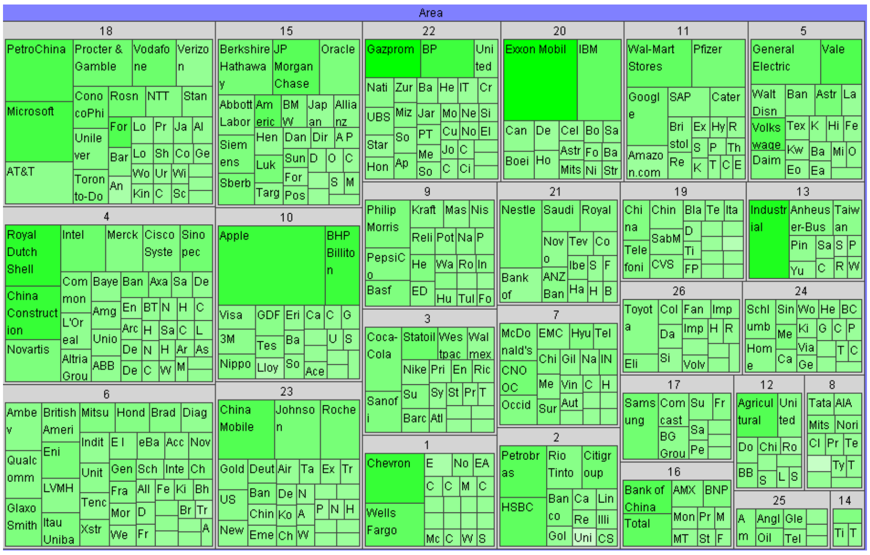
Table 2.4: Classification in thematic mapping vs. scientific visualization.



(a) A choropleth map of the company market values with five classes



(b) A point map of the company market values with more than 20 classes



(c) Treemap visualization of unclassified market values represented proportionally by the sizes of the corresponding rectangles.

Figure 2.6: Exemplary classification approaches applied in choropleth mapping, point map, and tree map visualization.

2.7 Symbolization

Semiotics or semiology, in particular Bertin's semiology of graphics [Ber83] and extended graphic variables [Mac95; Buz01], has profoundly influenced both thematic mapping and scientific visualization. A wide spectrum of symbols together with their graphic variables, as well as various design styles could be found in both disciplines. However, in terms of symbolization results and methodologies, there are many differences.

For the purpose of communication, each symbol representing certain classified data on a thematic map should be clearly seen. Users should quickly and intuitively relate a symbol with its referent so that they can understand the meaning of the symbol. Although there are no firm rules for symbol construction, cartographers have carried out many research works to enhance thematic map legibility and reliability [Men01]. How to do properly with symbolization has been a long concern. MacEachren [Mac94] developed a useful set of models of geographic phenomena arranged along discrete-continuous and abrupt-smooth continua and a set of symbolization methods appropriate for these models. Numerous usability studies on symbolization issues and map effectiveness have been conducted as well. For example, user perception issues related to the different symbol styles on thematic maps were investigated [Slo+08]. However, more studies should be done on relatively new symbolization methods and graphic variables, such as animation, transparency, shading, and viewing angles.

In scientific visualization, symbolizing large amounts of unclassified data or classified data (usually more than 10 classes) often results in a clustered visual pattern with indiscernible individual symbols. "How-to-do"-symbolization has been well stated in this discipline. For example, a transfer function is often specified in volume visualization so that scalar data values can be mapped to optical quantities such as colors and opacities [WS11]. However, "how-to-do-properly"-symbolization is largely missing. Although some symbolization issues are partly solved, questions concerned with the impacts of design solutions on users are seldom addressed. Researchers rarely study or apply what is known about the visual system when designing visualization techniques [JH04], and the growth of usability studies and empirical evaluations has been relatively slow [Che05].

To demonstrate the symbolization characteristics of thematic mapping and scientific visualization, we have designed a scenario of multivariate visualization. Figure 2.7 shows four equal-sized gray scale images that are simulated in a random process, each representing one variable. For the purpose of scientific visualization, we designed a square icon with four equally divided sub-squares (Figure 2.8). Each sub-square corresponds to one variable. We firstly normalize all of these four variable values to the range [0,1] and then map the four variables to the sub-squares according to a specific order. Figure 2.9 shows the 4-variate scientific visualization result. The whole structure of data distribution with several hot spots could be easily perceived, while the individual color icons in the whole structure are indiscernible. However, each variate value can still be acquired. We also

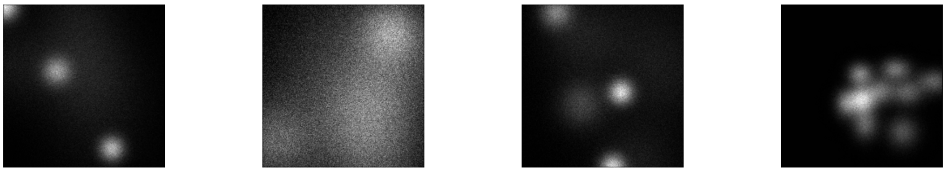


Figure 2.7: Four simulated gray scale images.

apply the parallel coordinate system technique on the same dataset to show the similar compact effect of scientific visualization in Figure 2.10. For the purpose of thematic mapping, we predefine several irregular areal units and impose the boundaries on each gray scale image. Image pixels inside each polygon are aggregated and a mean value is generated. Figure 2.11 shows multiple choropleth maps of the aggregated mean values of each variable. Due to the spatial aggregation, individual symbols within the referenced area and the whole pattern could be easily perceived and interpreted. Note that, superimposing many themes over each other or integrating them with each other in thematic mapping may require too much cognitive efforts on the user's part.

The aforementioned paragraphs and examples illustrate that sophisticated symbolization design in thematic mapping makes the whole pattern of data distribution easily comprehensible and each individual symbol easily discernible; while the symbolization results of scientific visualization usually show a whole pattern of data distribution with compact and indiscernible individual symbols. The compared results in terms of symbolization in thematic mapping vs. scientific visualization are presented in Table 2.5.

	Thematic mapping	Scientific visualization
Symbolization results	Often with known patterns	Often naturally clustered symbols
Individual symbols	Easily discernible	Hardly discernible
Usability studies	More	Less
State of the art	Both how-to-do and why-to-do	How-to-do instead of why-to-do

Table 2.5: The symbolization in thematic mapping vs. scientific visualization.

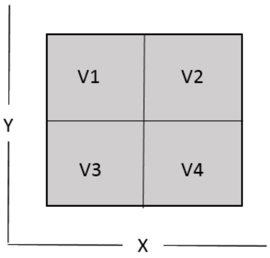


Figure 2.8: A square icon with four equal sub-squares. In this example, four variables are mapped to the sub-squares according to a specific order; x and y axes indicate the position of the icon on the image.

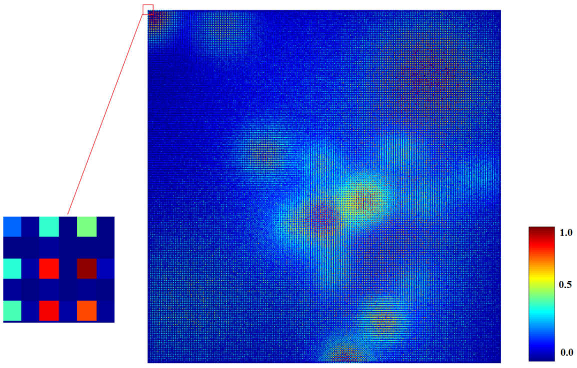


Figure 2.9: A compact presentation of the 4-variate color icon. The left sub figure shows the enlarged pixels.

2.8 Perception and Cognition

For both thematic mapping and scientific visualization, users construct or adjust their personal mental images based on their perception and cognition of the visual representations.

Properly designed thematic mapping results in easily discernible individual symbols. Individual symbols with high pre-attentive distinctions can immediately draw user's attention and be easily memorized by users. This is particularly important when users surf on Internet and look for interesting contents. Graphics (e.g., Internet maps) that do not possess the ability of pre-attention will be possibly ignored during the short-time usage of Internet. In addition, most individual symbols are self-explaining or can be quickly understood by general users referring to the legend. Similarly, just by an overview of a thematic map, users can perceive and understand the whole spatial pattern without much cognitive effort. Moreover, multiple representation methods are often simultaneously applied to help users get a comprehensive understanding of the phenomena.

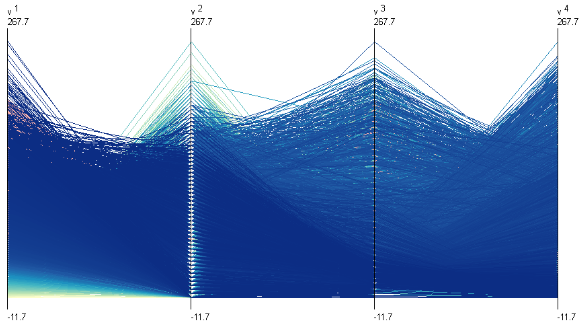


Figure 2.10: A compact presentation of the four variables by the parallel coordinate system technique, generated by XmdvTool (`http://davis.wpi.edu/xmdv/`).

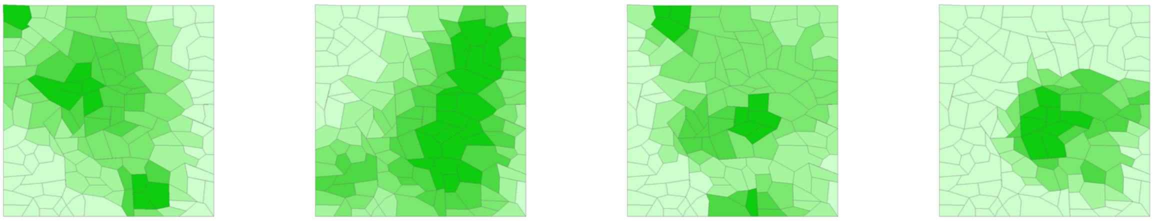


Figure 2.11: Four choropleth maps showing distributions of the mean values of four variables.

From the scientific visualization result, domain experts may get a rough idea about the data distribution, whereas general users without domain knowledge often find it difficult to interpret the naturally clustered symbols. An in-depth understanding of the interesting hotspots or outliers often demands more cognitive efforts. Users need to drill down the information space, often via complicated interactive operations (e.g., zoom in/out, brushing and liking), in order to figure out the meaning of each symbol. When visualizing aspatial data, familiar spatial metaphors may be helpful to enhance quick understanding of the representations.

To conclude, the whole pattern of the data distribution on a well-designed thematic map should be easily comprehensible and each individual symbol be easily discernible, while in scientific visualization the whole pattern of the data distribution is easy to perceive but the compact and indiscernible individual symbols may require more cognitive effort and interactions to interpret.

2.9 Summary

In this chapter, we systematically compared some essential aspects in thematic mapping and scientific visualization. They include the purposes, data sources, georeferencing, data preprocessing, classification, symbolization, and perception and cognition.

From the aforementioned comparisons, we can learn that thematic mapping is based on the communication model and emphasizes the communication of known information from the map maker to the map user via information abstraction. This information abstraction is reflected in every step of the visualization pipeline. Thematic mapping starts with raw data sources, which can be individual data items but more likely aggregated data geo-referenced to irregularly or regularly shaped polygons. Through data preprocessing, thematic data are normally scaled to more abstract levels of measurement based on semantic aggregation; while at the same time, the data amounts can be largely reduced through spatial aggregation. Classification in thematic mapping further abstracts information by imposing classification constraints such as limiting the number of classes up to seven. These constraints largely reduce the data complexity and allow an immediate understanding of the whole classification results. Furthermore, sophisticated symbolization design principles in thematic mapping make the whole pattern of data distribution easily comprehensible and each individual symbol easily discernible. However, information abstraction in each step will lead to the loss of detailed information and may also be misleading and result in statistical pitfalls.

In contrast, scientific visualization is based on the exploration model and emphasizes on the exploration of unknown information from large amounts of dataset. With the aim of exploring detailed information, scientific visualization starts normally with raw individual data items and tries to keep the data accuracy during every step of the visualization pipeline. Individual data items with spatial characteristics can be geo-referenced to accurate locations in 3-D physical space based on either a vector or raster data model, while aspatial in hyper-dimensional space can be spatialized to the 2-D or 3-D physical space based on a graph model. Through data preprocessing steps, raw individual data items may be normalized and filtered based on semantic aggregation. In the data classification step, optimal classifiers are calculated to accurately group each of the individual data items to a class without strict constraints of the number of classes. The symbolization in scientific visualization usually results in a whole pattern of data distributions with compact and indiscernible individual symbols. Thus, the visualization results may require more cognitive effort and interactions to interpret. During the visualization procedure, accurate individual data items can be acquired at each step.

We can conclude from the comparative study that thematic mapping and scientific visualization reveal different visual analytical levels and play a complementary role with each other. Thematic mapping emphasizes the communication of known information to general users via easily perceivable graphics; whereas scientific visualization emphasizes the exploration of unknown information and usually needs more efforts from the user. In terms of data preprocessing, thematic mapping goes steps further than scientific visualization while scientific visualization relies more on the capability of users to visually process the data by themselves. Based on their complementary characteristics, we can say that scientific visualization can very well profit from thematic mapping methods to achieve some communication effects while thematic mapping methods can be very well extended to embrace some explorative scientific visualization methods for the hotspot investigations.

This systematic study also provides a theoretic background to achieve synergetic effects of the two disciplines in future visualization systems, particularly providing meaningful ways to tackle with significant challenges posed by big data. From a technical point of view, facing big data, such as the flourishing crowdsourcing data, both thematic mapping and scientific visualization need intermediate steps to process the big data and make it manageable. In thematic mapping, the raw data sets would be largely reduced through spatial aggregation. The much smaller amount of aggregated data are then visualized on maps, resulting in an overview of data distribution. In scientific visualization, the big data should be divided into smaller and computationally manageable subsets. The scientific visualization systems then may progressively load data and allow users to interactively explore the information space. For example, the recent imMens system utilizing multivariate data tiling and parallel query processing techniques are developed to support real-time visual analysis of big data [LJH13]. According to Meng [Men13], the big data problem will enforce the shift of the map-making procedure from coverage-orientation to feature-orientation. We envision that the synergetic effects of thematic mapping and scientific visualization may visualize big data in a more controlled way. For example, based on a well-designed thematic map with an overview of density distribution, users can first identify some hotspots, and then progressively analyze and explore the subsequent data utilizing techniques from scientific visualization.

3 Movement Events and Shanghai Taxi Floating Car Data

Mobility has been one of the top priorities in the industrial regions and has profound impacts on economic growth. With the advances in location positioning and wireless communication technologies, collecting spatial trajectories that represent the mobility of a variety of moving objects (e.g. people, vehicles) becomes prevalent in the digital society. The resulting complex and large movement data give rise to unprecedented opportunities for understanding the mobility patterns and their social interactions of moving objects. In our thesis, we pick floating car data (FCD) collected from moving taxis as our test dataset to understand mobility patterns of taxis, and as the basis for the empirical proof of concept.

This chapter aims at identifying use cases of real-world movement data for experimenting the synergetic effects of scientific visualization and thematic mapping. In Section 3.1, we introduce movement data and investigate the concepts of movement events. More specifically, we distinguish movement events at two different levels, which can provide two different views for the visual exploration of movement data. At the basic level, each movement record naturally represents an individual event; beyond the basic level, for a series of continuous records, one could group them into trajectories and derive a variety of trajectory-based events. Section 3.2 describes the test movement data set, i.e. taxi floating car data, which are collected from about 2000 taxicabs in the year 2010 in Shanghai. In Section 3.3 and Section 3.4, we investigate the point-based and trajectory-based events in-depth, derive and pre-process the corresponding event dataset for the subsequent visual analysis.

3.1 Movement Events

Recent years have witnessed pervasive usage of location positioning and communication technologies (e.g. GPS devices, mobile phones) and dramatic increases of movement data produced from these technologies (e.g. floating car data, mobile phone data). Those massive movement data contain rich information and bring new opportunities for us to understand urban dynamics, which are crucial for decision making in environmental and transportation planning. Movement data analysis is currently a hot research topic with numerous studies focusing on modeling human mobility patterns [GHB08; Zhe+08; DYM15], uncovering taxi driving behaviors [LAR10; DFM15], mining interesting locations or places [Zhe+09; And+11; And+13b], and inferring urban land uses and city structures [Liu+15a], etc. In addition, movement data have also been combinatorially explored with other data sources, e.g. POI data [YZX12], check-in data [Liu+14; Liu+15b], and Flickr data [MT15], for understanding place semantics. In spite of the numerous work, due to the inherent spatiotemporal complexity and uncertainty, analysis of movement data is still challenging.

Basically, movement data consist of position records generated by moving objects. Each record can be represented by a point, e.g. of $p = (x, y, t)$. A series of chronologically ordered points form a spatial trajectory $(p_1, p_2, ..., p_n)$. There are different views of movement data. For instance, Andrienko, Andrienko, and Heurich [AAH11] stated that movement can be viewed as consisting of continuous paths in space and time, referred as trajectories, or as a composition of various spatial events. Hence, movement data can be analyzed both as trajectories and as spatial events [And+13a].

In this work, we analyze taxi floating car data from an event-based view, although we still use “trajectory” as a general reference to the traces in the geographic space. In spite of extensive definitions of event and movement event [BDP08; And+13a], we distinguish two abstract levels of movement events, i.e. point-based and trajectory-based events. The point-based event view considers each discrete point (e.g. a GPS entry) as a point-based spatial event, while the trajectory-based event view considers a sequence of temporally ordered points (e.g. GPS records) as a trajectory-based event. We will introduce in detail the two types of events and their derivation and preprocessing steps in Section 3.3 and Section 3.4.

3.2 Shanghai Taxi Floating Car Data

The test fcd dataset is temporally ordered position records collected from about 2000 GPS-enabled taxis within 52 days from 10th May to 30th June 2010 in Shanghai with a temporal resolution of 10 seconds, resulting in more than half billion GPS entries. Each GPS entry is associated with a certain amount of fields, i.e. date, time, car identifier, location, instantaneous velocity, car status and GPS effectiveness. Table 3.1 lists the fields for each GPS record along with sample values and the description.

As an illustration of the spatial distribution of the raw GPS points, we extract GPS points of a taxi with an identification number of 10003 on 12th May 2010. We distinguish two states of the extracted GPD points based on the value of “car status”, i.e. occupied (O) with a “car status” value of 1, and non-occupied (N) of 0. Figure 3.1 illustrates the raw GPS points of a taxi with an identification number of 10003 on 12th May 2010 with the occupied and unoccupied states coded in blue and red. In Figure 3.1, the street networks and occupied and non-occupied trajectories can be clearly identified.

The raw data are provided as compressed CSV files. For each day, the size of CSV file

Field	Example value	Field description
Date	20100517	8-digit number, yyyymmdd
Time	235903	6-digit number, HHMMSS
Company name	QS	2-digit letter
Car identifier	10003	5-digit number
Longitude	121.472038	Accurate to 6 decimal places, in degrees
Latitude	31.236135	Accurate to 6 decimal places, in degrees
Velocity	16.1	In km/h
Car status	1/0	1-occupied; 0-unoccupied
GPS effectiveness	1/0	1-GPS effective; 0-ineffective

Table 3.1: Test data properties.

is about 4.5G after decompression. We load the data into MongoDB[1] because of its flexibility in data modeling and its native support for geo-spatial operators and indexes. In MongoDB, the geometries can be stored naturally as GeoJSON[2] objects. The data processing is mostly implemented in the Mongo Aggregation Framework and some ad-hoc JavaScriptfunctions. The screenshots in the following sections are captured from interactive HTML pages, where several JavaScript libraries are employed, including notably the OpenLayers[3] library for displaying maps, and d3.js[4] and c3.js[5] for plotting figures.

3.3 Derivation of Point-based Events

We consider each raw GPS point as an event of some status from an event-based view. For instance, a GPS point with the car status of 1 defines an occupancy event and 0 a non-occupancy event. Beside the occupancy and non-occupancy events, we derive from a time series of GPS points two additional special types of event, namely pick-up and drop-off event. A pick-up event is an event when the current one is non-occupancy and the next is an occupancy. Similarly, a drop-off event is an event when the current status is occupied and the next is non-occupied. Table 3.2 summarizes four types of events, i.e. occupancy, non-occupancy, pick-up and drop-off.

According to the definitions in Table 3.2, we firstly differentiate in the database the oc-

---

[1] `https://www.mongodb.org/`

[2] `http://geojson.org/`

[3] `http://openlayers.org/`

[4] `http://d3js.org/`

[5] `http://c3js.org/`

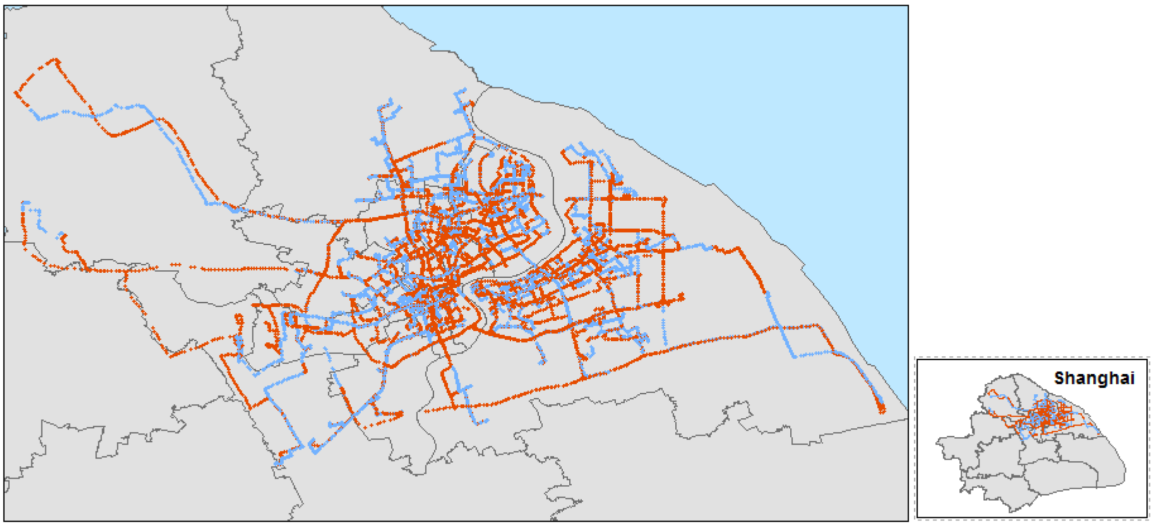


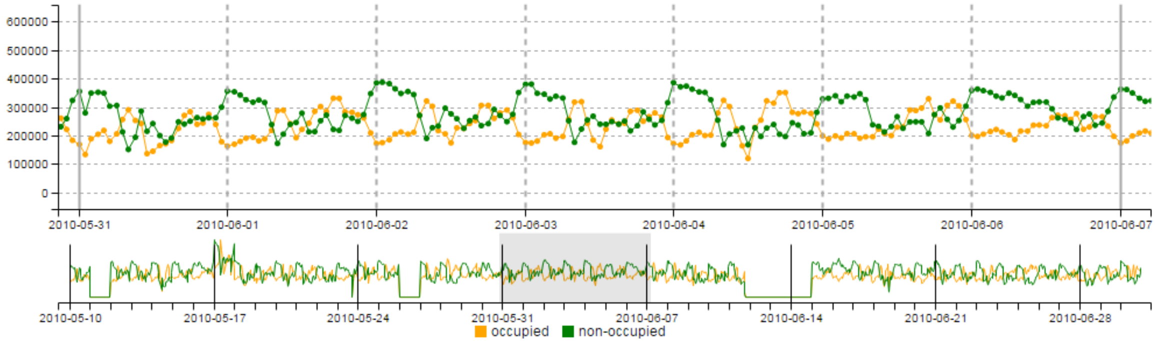
Figure 3.1: GPS points of the taxi with ID 10003 on the 12th May, 2010. Red and blue dots respectively indicate GPS points of occupied and unoccupied states.

Event type	Description
Occupancy (O)	A taxi is occupied by a passenger
Non-occupancy (N)	A taxi is without a passenger
Pick-up (P)	A taxi is picking up a passenger
Drop-off (D)	A taxi is dropping off a passenger

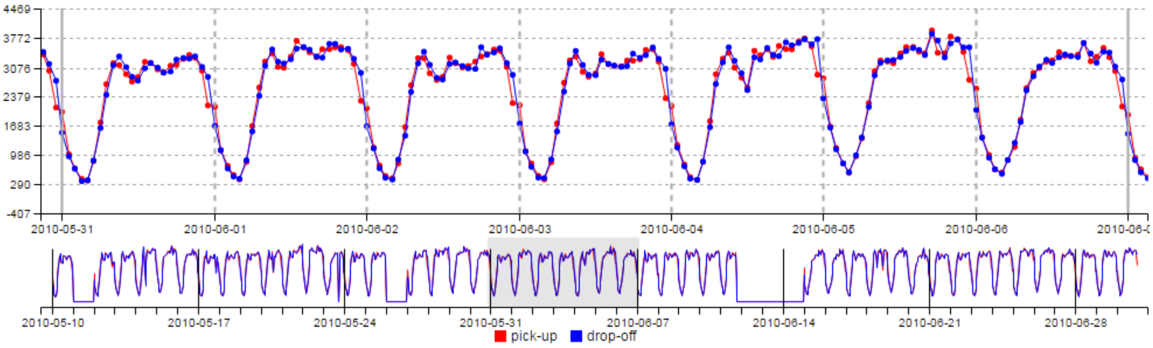
Table 3.2: Four types of point-based events (O, N, P, D).

cupancy and non-occupancy events, and then derive the pick-up and drop-off event accordingly. The extracted and derived data of (O, N, P, D) events will be used for analysis in the following chapters.

Due to the huge amount of (O, N, P, D) events in the time span in the whole study area, we partition them into certain spatial (e.g. $100m \times 100m$) and temporal chunks (e.g. 1 hour). For each spatial and temporal partition, we calculate for each event type its summarized value (e.g. the total number of Os) and assign it to the partition. The following subsections firstly introduce how we compute and investigate the temporal patterns of the dataset. Then we look into the spatiotemporal patterns and show the necessity of data normalization.



(a) occupancy and non-occupancy events



(b) pick-up and drop-off events

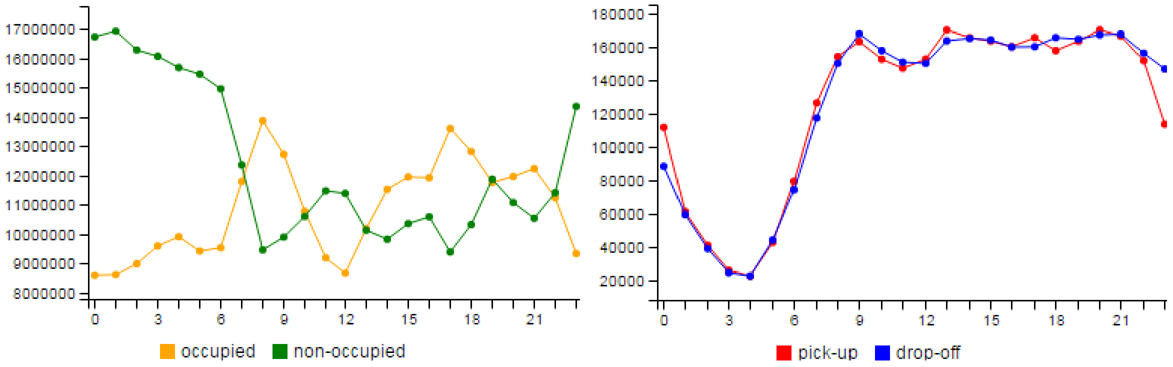
Figure 3.2: The focus + context hourly temporal variation of the total numbers. The context views show the temporal pattern from 10 May – 30 June, 2010, while the focus views show in detail one week temporal pattern from 31 May – 6 June, 2010.

3.3.1 Temporal Partition

For each hour in the 52 days, we compute the total numbers of (O, N, P, D) events respectively in the study area. Since the O and N events are of the same orders of magnitude, and P and D are at the same orders of magnitude, we plotted in the pair of (O, N) and (P, D) separately in Figure 3.2(a) and Figure 3.2(b).

From Figure 3.2, the temporal variation of the four variables can be clearly identified:

1. The frequency distributions of the four event types exhibit a strong daily rhythm.
2. The occupancy and non-occupancy events have a negative correlation. At midnight and early morning from about 10 p.m. to 6 a.m., the number of non-occupancy events is much larger. There are three peaks of occupancy events at about 8 a.m., 17 a.m. and 21 p.m. when the number of occupancy events exceeds that of the non-occupancy.
3. On the contrary, the pick-up and drop-off events have a positive correlation. The



(a) occupancy and non-occupancy events

(b) pick-up and drop-off events

Figure 3.3: The hourly temporal variation of the total amount of different events.

overall temporal pattern of the two variables has three peaks (at about 8 a.m.-9 a.m., 12 p.m. and 18 a.m.) and one deep valley (at about 4 a.m.).

Furthermore, we can calculate the general temporal patterns during each hour. We use $O[d,h]$ for the number of occupancy events in the hour h of on the day d, and $O[h]$ for the number of occupancy events in the hour of h on all days. Formally, the total numbers of events of O, N, P and D in each hour h of total 52 days are

$$\begin{aligned} O[h] &= \sum_{d=1}^{D} O[d,h] & N[h] &= \sum_{d=1}^{D} N[d,h] \\ P[h] &= \sum_{d=1}^{D} P[d,h] & D[h] &= \sum_{d=1}^{D} D[d,h] \end{aligned}$$

where $0 \leq h \leq H(=23)$.

The results are plotted in Figure 3.3 showing similar temporal patterns as Figure 3.2.

3.3.2 Spatiotemporal Partition

Besides temporal patterns, we also investigate furthermore the spatiotemporal patterns of the dataset. We firstly decompose the study area based on the simple regular $R \times C$ grids or cells. Each cell has a fixed size (e.g. $500m \times 500m$) and is denoted as a pair of (i,j) where $1 \leq i \leq R, 1 \leq j \leq C$. This method is relatively simple and easy to implement, and one can directly determine the size of the resulting decomposition by changing the size of the grids (Castro et al. 2013).

Based on the area decomposition, for each cell during the studied time span of 52 days, we compute the total number of events of O, N, P and D in a spatial partition. Let

$O[i, j, d, h]$ indicate the number of occupancy events in the cell (i, j) in the hour of h on the day d, and $O[i, j, h]$ for the number in the cell (i, j) in hour of h on all days. Formally, the total number of O, N, P and D in each cell (i, j) in each hour h of total 52 days are

$$O[i,j,h] = \sum_{d=1}^{D} O[i,j,d,h] \qquad N[i,j,h] = \sum_{d=1}^{D} N[i,j,d,h]$$

$$P[i,j,h] = \sum_{d=1}^{D} P[i,j,d,h] \qquad D[i,j,h] = \sum_{d=1}^{D} D[i,j,d,h]$$

where $1 \leq i \leq R, 1 \leq j \leq C, 0 \leq h \leq H(= 23)$.

Furthermore, for each cell during the studied time span of 52 days, we compute the total numbers of events of O, N, P and D.

$$O[i,j] = \sum_{h=0}^{H} O[i,j,h] \qquad N[i,j] = \sum_{h=0}^{H} N[i,j,h]$$

$$P[i,j] = \sum_{h=0}^{H} P[i,j,h] \qquad D[i,j] = \sum_{h=0}^{H} D[i,j,h]$$

where $1 \leq i \leq R, 1 \leq j \leq C$.

We plot the results in Figure 3.4. To reveal the underlying distribution, an equal interval classification method is used to group the values to seven classes and a bipolar color scheme from blue to red is applied to show the distribution of the data.

From Figure 3.4, we can see the spatial patterns of the four event types. Most of the cells are of small values reflected by the dominant blue color, especially the occupancy and non-occupancy events. There are a few light blue dots scattered in Figure 3.4(a), and notably two red dots around the two airports in Figure 3.4(b). Compared with the spatial distribution of the occupancy and non-occupancy events, the pick-up (Figure 3.4(c)) and drop off (Figure 3.4(d)) events are more concentrated in the city center of Shanghai, which means the center region are associated with very strong pick-up and drop-off activities than the suburban area.

3.3.3 Data Normalization

Although the visualization results of the original data in Figure 3.4 can reveal the underlying data distribution, it is hard to detect obvious spatial patterns. This motivates us to inspect the frequency distribution of the total amount of events for each variable. In Table 3.3, the 2nd and 3rd columns show the data frequency and its normalized distribution using the range scaling method. The two histograms with heavy tails reflect that the majority of grids have a very small amount of events and only a very few portion of grids

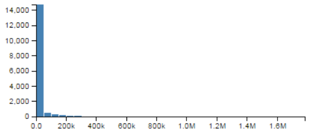
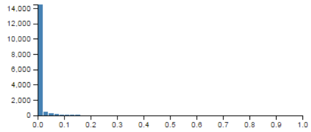
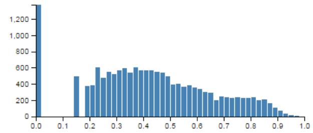
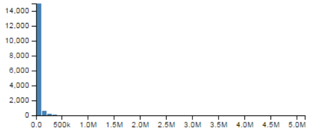
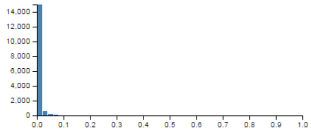
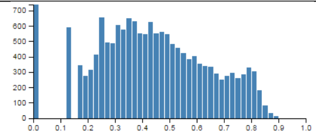
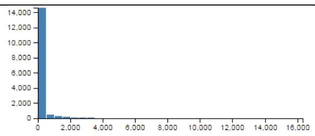
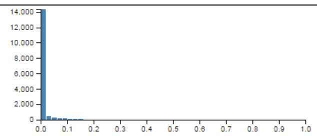
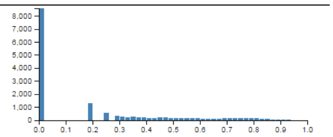
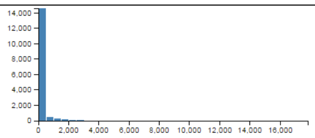
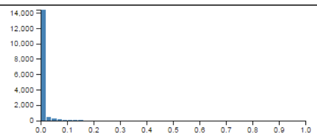
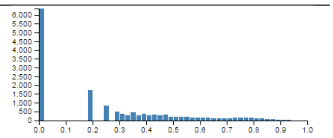
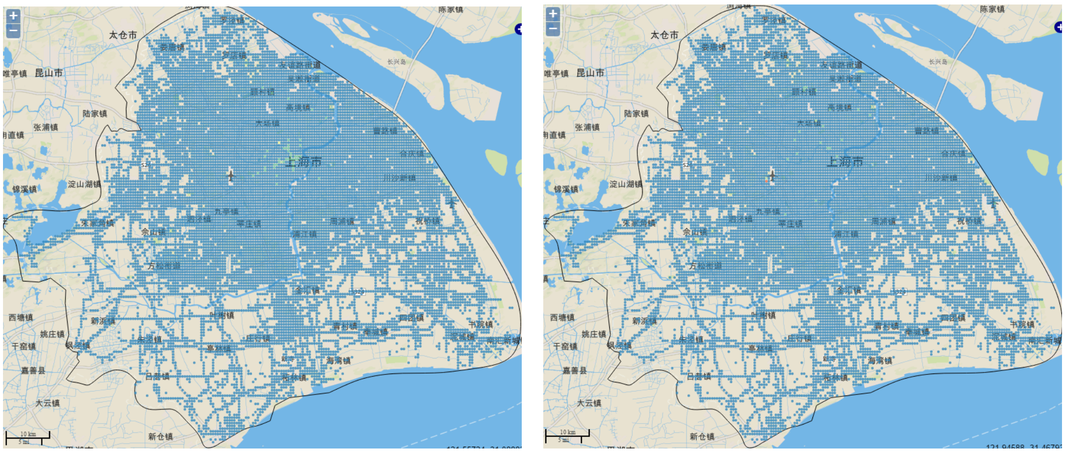
Event type	Number of events A_{ij}	Normalization $A_{ij}/max(A_{ij})$	Nomalization $\frac{log(A_{ij}+1)}{max(log(A_{ij}+1))}$
Occupancy			
Non-occupancy			
Pick-up			
Drop-off			

Table 3.3: Normalization of the number of events: the number of events in each cell (2nd column); the number of events divided by the max value (3rd column); The logarithm of the number of events divided by the max log value (4th column).

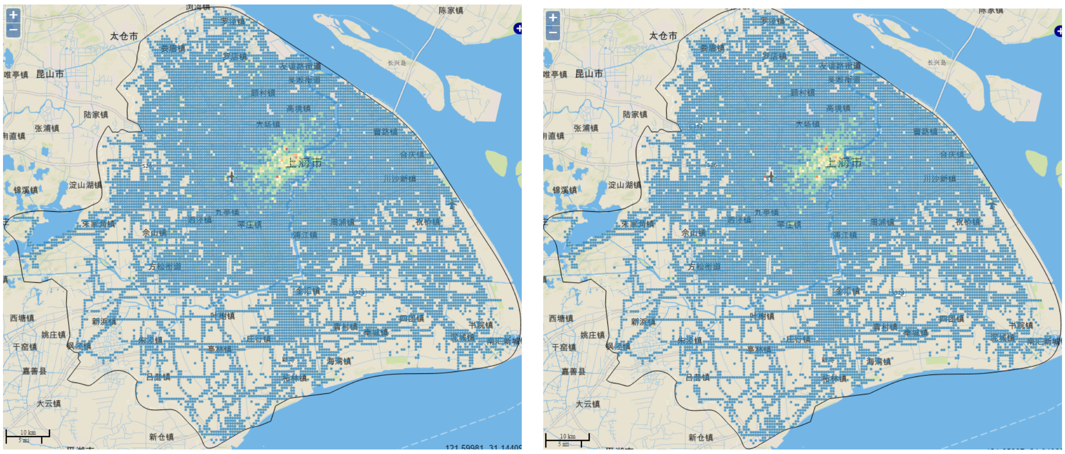
has extreme large number of events. For further pattern recognition and more meaningful visualization results, an appropriate normalization technique should be selected.

Due to the long tail characteristic of the variable frequency distribution, a logarithm transformation is a good choice. The 4th column in Table 3.3 shows the frequency distribution of (O, N, P, D) based on the range scaling normalization after the logarithm transformation.

We apply the same equal interval classification method with seven classes and the same color scheme to visualize the normalized data of logarithm transformation. The visualization result is shown in Figure 3.5.

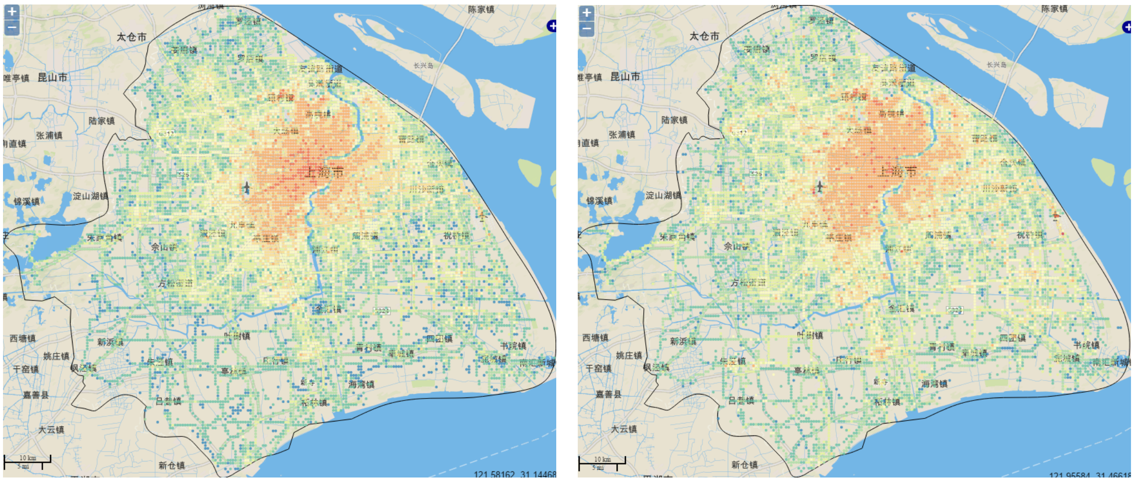


(a) occupancy (b) non-occupancy

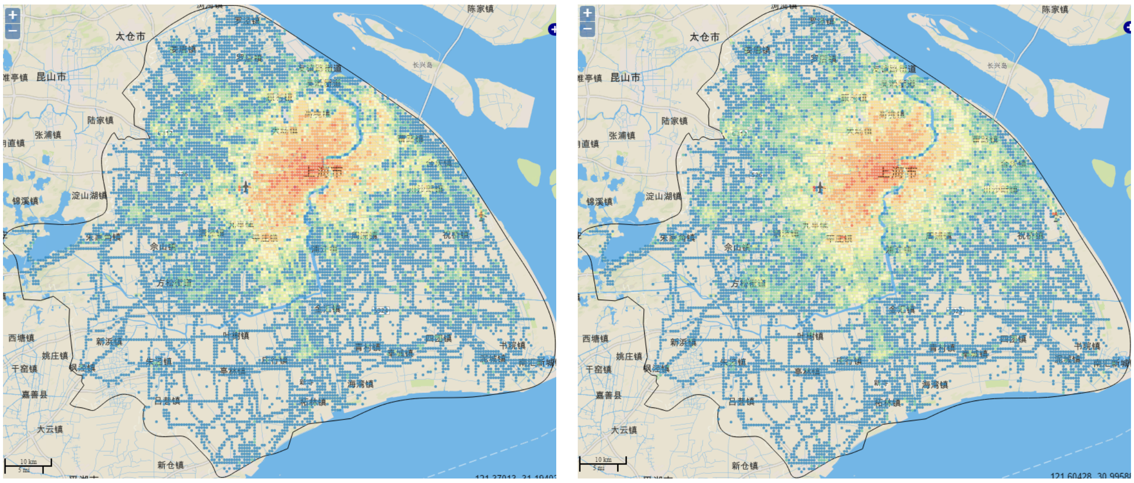


(c) pick-up (d) drop-off

Figure 3.4: The frequency distribution of the total numbers of (O, N, P, D) events in each cell.



(a) occupancy (b) non-occupancy



(c) pick-up (d) drop-off

Figure 3.5: The spatial distribution of normalized events (O, N, P, D) after the logarithm transformation.

Compared with Figure 3.4, the visualizations of the normalized (O, N, P, D) events after the logarithm transformation in Figure 3.5 reveal visually enhanced spatial patterns, showing the densities of the four events decrease from city center to suburban area. In Figure 3.5(a) the heavily occupied major roads can be identified in the city center. In Figure 3.5(b), unlike the distribution of the occupancy events, there are no obvious significant clusters in the city center or along with the road network. Instead a few dark cells associated with very high non-occupancy events are scattered around the city. According to the non-occupied characteristics that taxis can cruise on the road or parking at some places, we may infer that the scattered dark red cells may correspond to possible long parking places (e.g. airport waiting pools, parking lots, taxi drivers' homes). Compared with the distribution of the occupancy and non-occupancy events, the pick-up and drop-off events in Figure 3.5(c) and Figure 3.5(d) are much denser in the city center than suburb regions. Some peaks (in dark red) can also be easily detected.

3.4 Derivation of Trajectory-based Events

A trajectory is a complex object consisting of many GPS records with several associated attributes. Numerous types of events can be extracted from a collection of trajectories. In Section 3.4.1, we first categorize trajectory-based events into a variety of different types along two dimensions: attributes and abstraction levels. Then, we select two types of trajectory-based events, i.e. origin-to-destination event and non-occupancy event, for further investigation.

3.4.1 Categorization of Trajectory-based Events

In our study, we categorize the trajectory-based events that can be derived along two dimensions, namely the abstraction levels and the attributes. The abstraction levels refer to the levels of detail resulting from a variety of generalization methods that can be applied to a trajectory. The attributes mainly refer to spatial, temporal and thematic attributes, which impose constraints to a trajectory for the derivation of new types of trajectory-based events. According to the two dimensions, we summarize the categorization matrix in Table 3.4 and 3.5 respectively. In addition, we give specific examples for each category. The types of trajectory-based event that will be analyzed in the following chapters are highlighted.

Abstraction level	Examples
non-generalized	**reconstructed trajectories**
generalized	**point-to-point**
	simplified trajectories (e.g. after Douglas-Peucker algorithm)

Table 3.4: Categorization of trajectory-based events using abstraction levels.

	Attributes	Constraints	Example values
spatial	lon & lat	**starts from**	starts from an airport
		ends to	ends to a station
		inside	inside a county
		crosses	crosses a bridge
temporal	timestamp	starts	starts at 9 a.m.
		ends	ends at 12 p.m.
		before	before May
		during	during daytime
thematic	**occupation**	occupancy	1
		non-occupancy	0
	speed	high	>80
		low-speed	<20
	direction	eastwards	E, NE, SE

Table 3.5: Categorization of trajectory-based events using attributes.

Table 3.4 illustrates the categorization of the trajectory-based events using their *abstraction levels.* According to different mapping purposes, a trajectory can be represented at different levels of detail. The most detailed trajectory-based events are reflected by the non-generalized trajectories that consist of almost all the raw GPS records without sacrificing much information. This kind of events reflects the detailed dynamics of the moving object during the time period. In our study, the *non-generalized trajectory-based event* refers to the event extracted from the reconstructed trajectory formed by connecting a time series of recorded GPS points. On the other hand, trajectories can be generalized to different levels of detail according to a variety of cartographic line generalization methods. The trajectory-based events derived from generalized trajectories show the distinguished characteristics of the reduced spatial traces of the moving object. One extreme case of the simplified events is the *point-to-point (P2P) event* that keeps only the first and the last points of a trajectory disregarding all the intermediate details.

Table 3.5 shows the categorization of trajectory-based events based on the *attributes* of a trajectory. A trajectory is associated with a variety of spatial, temporal and thematic attributes. The three basic attributes impose a variety of constraints on the trajectory based on which a variety of different types of trajectory-based events can be derived.

Spatial and temporal attributes. Due to the inherent spatial and temporal characteristics of spatial trajectories, different types of trajectory-based events can be derived based on their locations and timestamps. Spatial constraints are numerous, for example, “starts from”, “ends to”, “inside”, and “crosses” etc. More specifically, for instance, by imposing the “starts from” constraint to an airport on a trajectory, the “starts from the airport” trajectory-based event can be extracted. Similar with the spatial attributes, a variety of temporal constraints, e.g. “starts”, “ends”, “before”, “during”, can also be naturally applied. For instance, by imposing “during” to lunch (e.g. 12h-13h), “during lunch” events can be derived.

Thematic attributes. A trajectory is associated with different thematic attributes . In this study, the reconstructed trajectories have a variety of derived attributes including occupation status, speed, direction, distance. For instance, occupancy and non-occupancy trajectory-based events can be derived based on their occupation status. More categories, e.g. high-speed, eastwards, can be found in Table 3.5.

Note that countless new types of events can be derived by imposing constraints to the listed types of events. For instance, the occupancy trajectory-based events derived based on the thematic attribute of “occupation status” can be further abstracted to occupancy point-to-point events, which reflects the passenger transfer event from the origin to destination and will be referred as *origin-to-destination events* in this thesis.

	number
Occupancy events	2906716
Non-occupancy events	2919597

Table 3.6: The numbers of the occupancy and non-occupancy trajectory-based events.

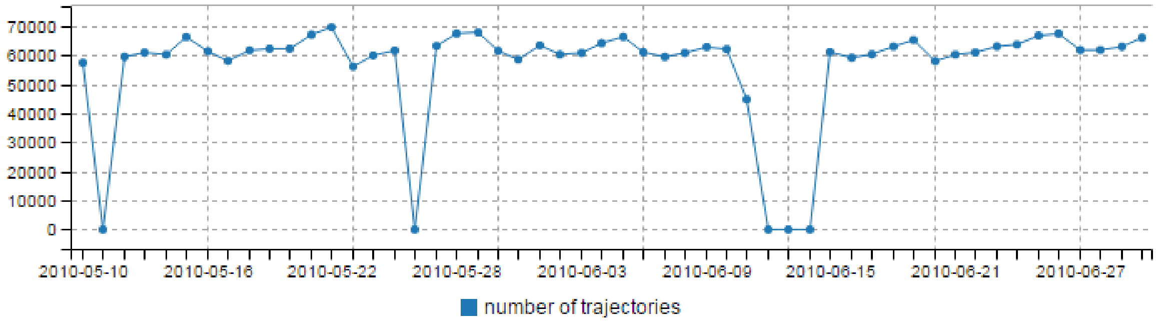


Figure 3.6: The frequency distribution of the daily number of reconstructed occupied trajectories.

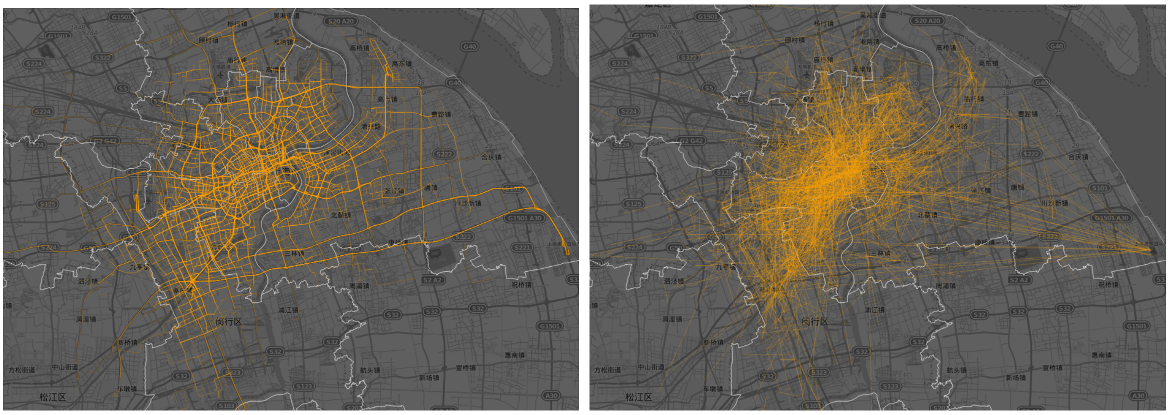
3.4.2 Origin-to-Destination Event and Non-occupancy Event

Based on the categorization, we can easily distinguish two types of trajectory-based events, namely occupancy and non-occupancy events. The derivation is based on the thematic attribute constraint "car status". Occupancy events are of "car status" value of 1 and non-occupancy events are of "car status" value of 0.

We connect the temporal sequences of GPS points of each event type and reconstruct the event trajectories and save them to the movement database. Table 3.6 shows the number of the respective types of trajectory-based events.

As an illustration, we plot the number of occupied trajectories of each day in Figure 3.6. From Figure 3.6, we can observe the temporal distribution of the daily total number of occupied trajectories. The overall trend exhibits a slow increase in every week with a peak usually shown on Saturday, which reflects that there might be more activities using taxicabs on Saturday.

Figure 3.7 illustrates the reconstructed occupied trajectories of about 100 cars on 10 May, 2010, and their corresponding origin-destination lines. In Figure 3.7(a), each orange line represents a reconstructed occupied trajectory and is drawn with an opacity of 0.3. The trace footprints show clearly the road network structure. Darker orange lines indicate denser trajectories along the roads. Figure 3.7(b) shows the origin-destination lines of the corresponding trajectories. We can see that there are some very popular origins or des-



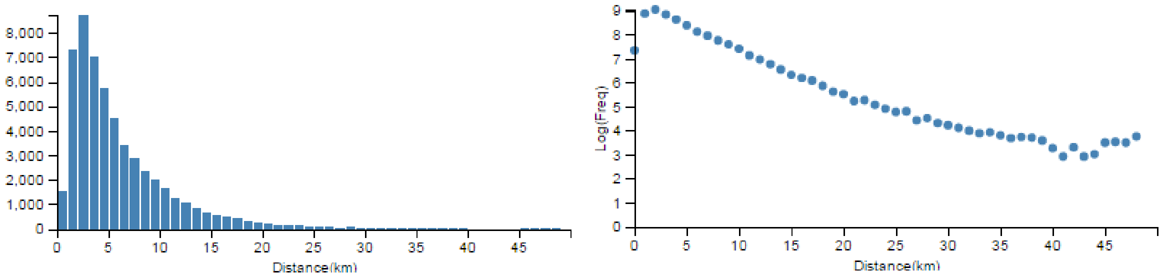
(a) Occupied trajectories

(b) Origin-Destination lines

Figure 3.7: The reconstructed occupied trajectories of 100 cars on 10 May 2010 and their corresponding origin-destination lines.

tinations. For instance, there is clearly a hotspot on the right of the screenshot, which corresponds to the Shanghai Pudong international airport.

Based on the reconstructed trajectories, a number of associated trajectory statistics can be derived e.g. trajectory distance, duration. For instance, Figure 3.8(a) shows the frequency distribution of distance of the derived occupied trajectories on 10 May, 2010. Similar to the frequency distribution of the total number of events for spatial cells, the histogram of the traveling displacement and elapsed time show long tails as well. The logarithm of the frequency represented by the dot graphs in Figure 3.8(b) show that the frequency follows roughly the exponential law.



(a) The distance frequency distribution of the occupied trajectories

(b) The logarithm of the frequency

Figure 3.8: The distance frequency distribution and the logarithm of the occupied trajectories on 10 May 2010.

When the taxi is occupied, the taxi driver usually finds out the fastest route to send passengers to a destination based on his knowledge [Yua+12]. On the contrary, when the taxi is vacant, the taxi driver has a large freedom to plan his/her routing to minimize his/her waiting time of the next trip. Generally speaking, experienced taxi drivers are more likely to pick up their next passengers quickly while novice drivers may cruise on the roads for a longer time. Therefore, different taxi driving behaviors may exhibit significant influence on the income of taxi drivers.

Based on these assumptions, in our work, for occupancy events, we do not care about their real physical traces and pay attention to the origin-to-destination events, which can be extracted by connecting the starting and ending point of the occupancy event trajectories. For non-occupancy events, we would focus on the analysis of two groups of taxi drivers, i.e. high-income and low-income groups, and investigate their different driving behaviors when they are free during the non-occupied time periods.

3.4.3 Derivation of the Income of Taxi Drivers

Occupied trips and non-occupied trips can be reconstructed by connecting the temporal sequences of GPS points with the 'car status' attribute value of 1 and 0 respectively. A number of associated trip statistics, e.g. trip distance and duration, are also derived. Based on the occupied trip distance and the taxi fare system of Shanghai in 2010 shown in Table 3.7, the average daily income of each taxi driver can be calculated. The following formula is used to calculate the fare of an occupied trip.

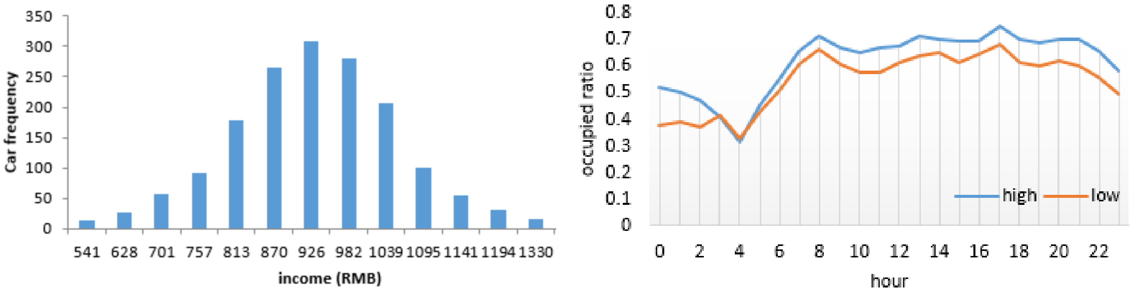
$$f(d) = P_0 + P_3 * \min(\max(d-3,0),7) + P_{10} * \max(d-10,0)$$

where d is the distance of the occupied trip.

In Figure 3.9(a), the histogram represents the distribution of the average daily income

	ITEM	Day timing	Night timing
	(05:00 - 23:00)	(23:00 - 05:00)	
P_0	Minimum fare for the first 3km	12 CNY	16 CNY
P_3	Fare above minimum fare until 10km	2.4 CNY/km	3.1 CNY/km
P_{10}	Fare above 10km	3.6 CNY/km	4.7 CNY/km

Table 3.7: The taxi fare calculation system of Shanghai in 2010.



(a) Average daily income distribution (b) Occupied ratio of high- and low-income taxis

Figure 3.9: Distribution of taxi incomes.

of the taxi dataset (2000 taxis in 47 days). The normal distribution of the daily income indicates the income discrepancy among the taxis. This paper categorizes the 100 highest-income taxis as the top group and the 100 lowest-income taxis as the bottom group.

Moreover, we estimate the occupied ratios by utilizing the distance of the occupied and unoccupied trip with the formula

$$dist(trips_{occupied})/(dist(trips_{occupied}) + dist(trips_{unoccupied}).$$

The line chart in Figure 3.9(b) illustrates the occupied ratios of the top and the bottom taxis. Basically, the occupied ratio of both groups has several peaks (e.g. at 8am) and valleys (e.g. at 4am) and the high-income taxi group generally has a much higher occupied ratio. However, in the early morning between 2am and 7am, the differences between the top and bottom taxi groups are much smaller than that in other time slots.

4 Visual Analysis of Point-based Events

This chapter is dedicated to analyzing the point-based (O, N, P, D) events extracted from floating car data in Section 3.3. A variety of visualization techniques by means of scientific visualization and thematic mapping are proposed for effective and comprehensive visual exploration and communication of the four types of events. Section 4.1 investigates the state-of-the-art of the theories, methodologies and applications related to point-based events. In Section 4.2 we propose three visualization techniques for visual exploration of the movement events: a multivariate visualization method based on pie radar glyph, a univariate visualization method based on time graph, and a salience-based visualization method. Section 4.3 aims to communicate the spatiotemporal patterns by means of thematic mapping techniques. Based on the aggregates resulting from spatial clustering, Voronoi tessellation and aggregation, easily legible thematic maps are then designed to convey the spatiotemporal patterns. Section 4.4 analyzes and summarizes the visualization results. Section 4.5 compares and synthesizes the proposed visualization methods. Section 4.6 summarizes this chapter.

4.1 Related Work

Point-based movement data can be viewed at multiple granularities from multiple aspects, e.g. characteristics of location, their dynamics over time, as well as the relations between events and the properties of the context. For a comprehensive exploration and understanding of massive point-based event data, a wide range of visual techniques, being direct depictions of original/raw data or representations of computationally extracted patterns [Zhe+14], are necessary and should be integrated.

Numerous researches and applications have been conducted on exploration of movement events. At a conceptual level, Beard, Deese, and Pettigrew [BDP08] proposed a general framework for visual exploration of events. Andrienko, Andrienko, and Heurich [AAH11] suggested an event-based conceptual model for context-aware movement analysis. Research projects like ViAMoD (Visual Spatiotemporal Pattern Analysis of Movement and Event Data) [1] aim to develop theoretical foundations as well as appropriate analysis methods that combine visual, interactive and algorithmic techniques for the scalable analysis of movement and event data. In addition, researchers developed new methods

---

[1] `http://geoanalytics.net/and/projects.html`

to detect and track dynamic clusters of spatial events [And+14], and discover and analyze temporal patterns in complex event data [Peu+15]. Novel visualization methods in the interactive environment are addressed for visually exploring multidimensional and multivariate movement events using space-time cube [GAA04], interactive difference views [LKH10] and non-overlapping aggregated multivariate glyphs [SVV14].

At the individual level, a point-based event can be represented by a dot and a straightforward way to show event patterns is by dot plotting. Interactive techniques, such as filtering, brushing, linking, can be applied to the dot plots to query the interesting part of the data set, explore the spatial distributions of each variable and their relationships. For instance, the interactive visualization system HubCab [2] developed by MIT Senseable city lab plots individual pick-up and drop-off events as blue and yellow dots. The system allows users to get insight into the taxi mobility patterns at a very fine granularity and supports future taxi sharing based on a model named "shareability networks" [San+14]. In some scenarios, multiple types of events may occur at the same location during a time interval. Those co-located events can be either simultaneously shown in one view by mapping them to multiple visual variables, or presented separately in small multiples. Advanced multidimensional multivariate visualization techniques are developed and widely used, e.g. scatter plot, color-icon, stick figure icon, and hyperbox [WB97].

At the collective level, visual analysis of massive events incorporates data transformation and data aggregation, which aims to represent groups of objects and reduce visual cluttering and overlapping effects. A typical procedure for visual analysis of events involves event clustering, space partition, spatiotemporal aggregation and the analysis of aggregated data. Andrienko and Andrienko [AA10] systematize the possible approaches for aggregation of movement data in a framework that clearly defines what kinds of exploratory tasks each approach is suitable for. The aggregates can be mapped to proportional symbols, diagrams, etc. The analysis of event aggregates at the collective level has many applications, for instance, aggregating the events to meaningful places and analyzing the relevant places [And+13c] and detecting dynamic activity patterns [SL14].

This chapter develops visualization methods for visual exploration and communication of the spatiotemporal patterns of multivariate point-based events, starting at a fine resolution and gradually moves to more abstract levels. The test data are the (O, N, P, D) events extracted from floating car data in Shanghai (see Section 3.3) with the spatial resolution of $100m \times 100m$ and the temporal resolution of 1 hour.

4.2 Visual Exploration by Means of Scientific Visualization

This section inspects the spatiotemporal patterns of the movement events at a detailed level. The normalized data based on the range scaling of maximum values from Sec-

---

[2] `http://hubcab.org/`

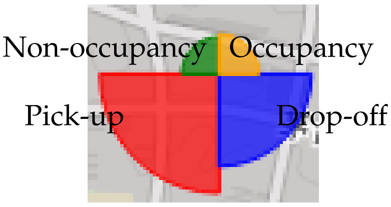


Figure 4.1: A pie radar glyph.

tion 3.3.3 are applied as input data set to reflect the underlying data distribution. We firstly introduce a multivariate visualization method based on a pie radar glyph to show the spatial density distributions of the co-located (O, N, P, D) events simultaneously. Secondly, a time graph mapping method is designed focusing on the investigation of uni-variate temporal patterns. Furthermore, we propose a salience-based method to highlight only the most salient variable and its quantity rather than all the variables at a specific location. Finally, the salience-based method is applied to visualize the temporal changes of pick-up and drop-off events, which leads to the exploration and inference of urban land use types.

4.2.1 Pie Radar Glyph

An appropriate visual representation is necessary to map the co-located spatial events of multiple variates. Glyph-based visualization methods like star/radar plot are widely used to represent multivariate scientific data. Rose diagram (also known as polar area diagrams) and its variant flower diagram is also applied to visualize groups of co-located spatial events. For instance, Andrienko et al. [And+13a] adopted a rose diagram to represent cyclic phenomena.

In this work, we propose a pie radar glyph to represent the co-located events. The glyph is comprised of four filled sectors assigned with orange, green, red and blue anticlockwise to represent the variates of O, N, P, D respectively. The radius of each sector is proportional to the value of the mapping variate. Figure 4.1 depicts the mapping of a pie radar glyph from the co-located four variate data.

To avoid clutter effects, we design the map by carefully choosing the following elements:

Filtering To reduce the computational complexity and the visual cluttering effects, we firstly filter out the very small data values of the four variables that are visually insignificant. The filtered multivariate data are then mapped to pie radar glyphs accordingly.

Ordering In spite of the filtering mechanism, visual cluttering effects still cannot be totally avoided due to the fine spatial resolution. A proper order of glyph rendering is necessary to enhance the visual appearance. This is achieved by assigning a larger z-index value to small glyphs so that small glyphs are on top of larger glyphs rather than occluded.

Opacity setting To make sure that the underlying big glyph would not be overlapped, semi-transparency rendering technique is applied. The radius of the glyph sectors are also carefully chosen.

The extracted (O, N, P, D) events in 52 days at each $100m \times 100m$ is mapped to the designed pie radar glyph. The visualization result is illustrated in Figure 4.2. We also zoom in several interesting places with attractive patterns.

The overall distribution in Figure 4.2 shows that the events in the city center are generally denser than in the suburban area. There are several significantly dense places. For the occupancy (orange) and non-occupancy (green) events, there are only a few dense places, which means that their spatial distributions are very uneven. On the contrary, pick-up (red) and drop-off (blue) events scatter more evenly.

We zoom in some of the dense areas which might be interesting to be explored in detail. Area (1) and (2) carry several extremely large glyphs, of which there are two big sectors of the non-occupancy events (green) and several relatively big sectors of pick-up and drop-off events. The background map reveals these two areas as two international airports, which makes the phenomena easily understandable. The two big green sectors of non-occupancy events are exactly located in the waiting pools of the two airports, where most of the taxis wait for a long time for the passengers resulting in a large number of non-occupancy events. There are also several pick-up and drop-off hotspots in the terminals, which correspond to the taxi stands where passengers arrive or depart the airports. Area (3) has a unusually large sector in orange. The place is a cross-way with a very high density of occupancy events resulting from long traffic jams. Area (4) and (5) with very high densities of pick-up and drop-off events are two railway stations. Area (6) and (7) also demonstrate very dense pick-up and drop-off activities. But unlike (4) and (5), they correspond to two famous commercial centers.

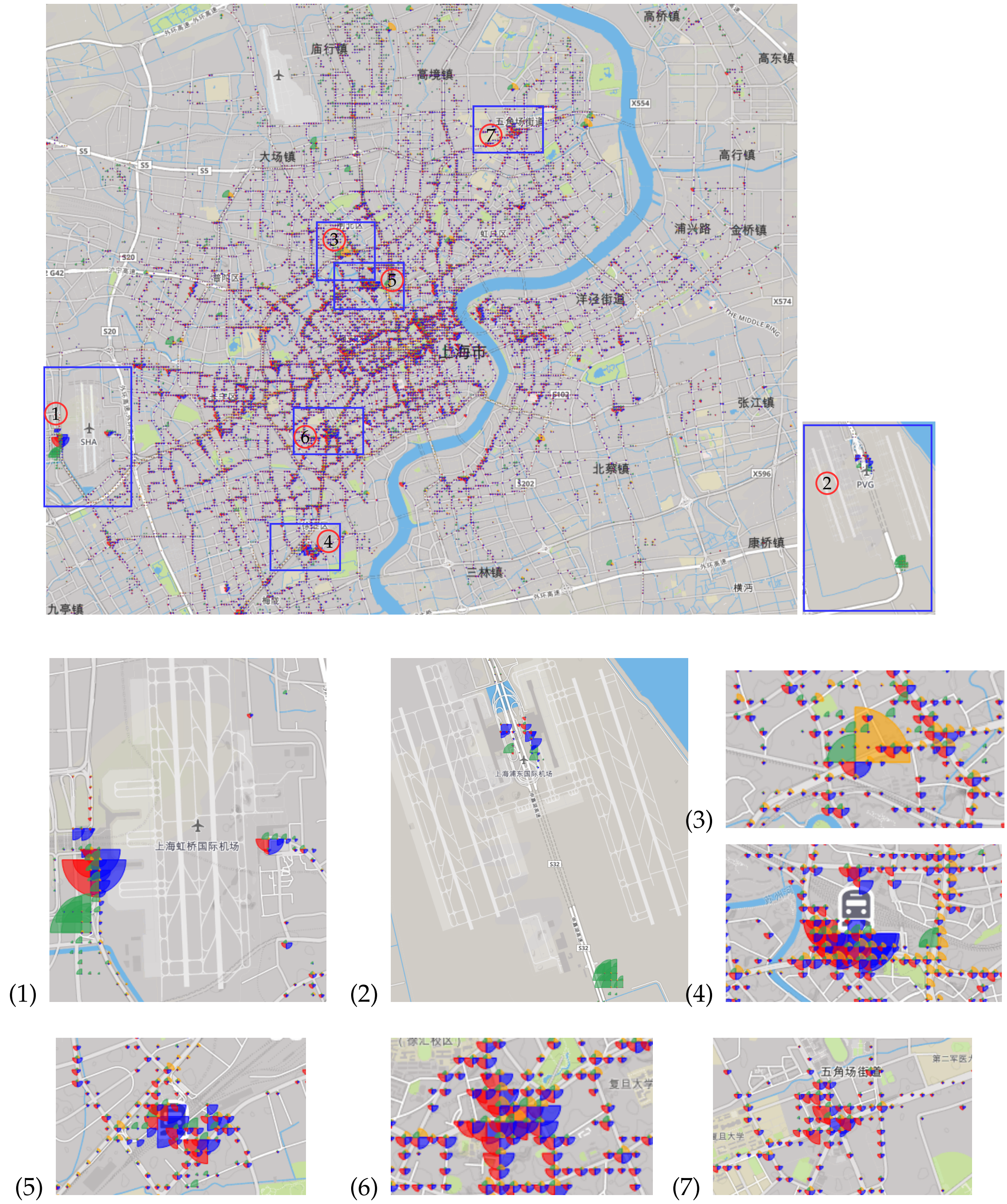
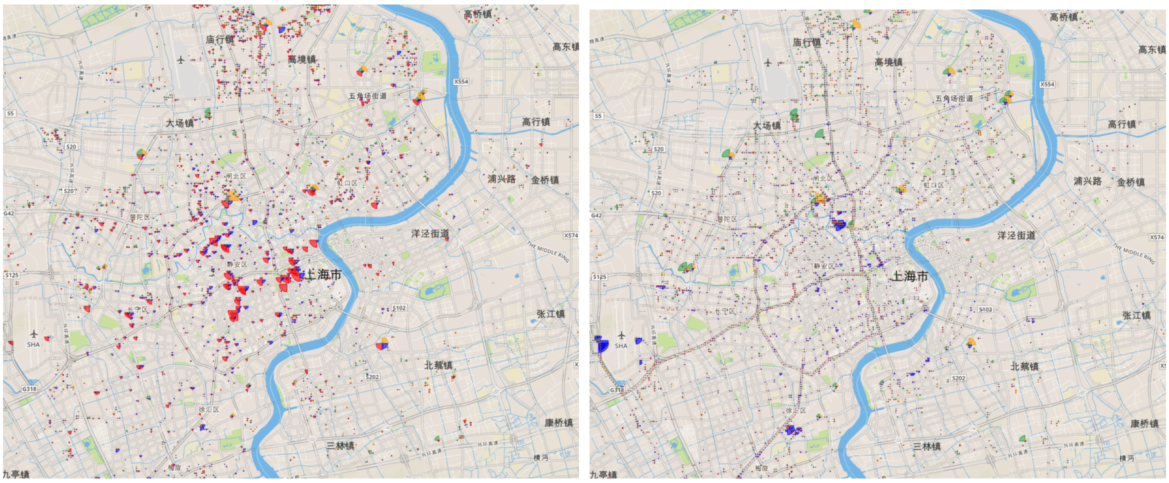
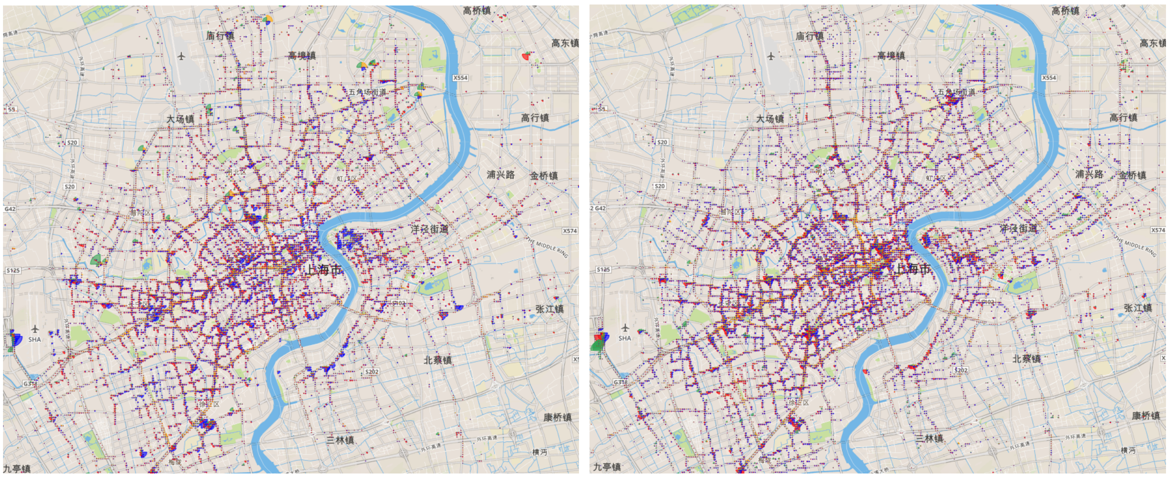


Figure 4.2: The pie radar glyph visualization results. The rectangles represent areas of interest (AOIs): (1) Hongqiao international airport; (2) Pudong international airport; (3) A cross-way; (4) Shanghai railway station; (5) Shanghai south railway station; (6) Xujiahui commercial zone; (7) Wujiaochang commercial zone.

To explore the temporal patterns, we may apply the glyph in different time slots. Figure 4.3 shows the spatiotemporal variances of the event distributions in four time slots.



(a) 3-4h (b) 6-7h



(c) 8-9h (d) 18-19h

Figure 4.3: Temporal distributions of (O, N, P, D) events using the pie radar glyph.

From Figure 4.3, we can observe very distinct temporal patterns from the compact visualization results. Figure 4.3(a) at 3-4h shows dominant red sectors indicating relatively active pick-ups, while Figure 4.3(b) at 6-7h has a few places of significantly dense drop-offs. The three biggest blue sectors on the west, south, and middle (from left to right) are located in the Hongqiao international airport (Terminal 2 and 1), Shanghai south railway station and Shanghai railway station. Figure 4.3(c) and Figure 4.3(d) are daily rush hours at 8-9h and 18-19h which have relatively larger number of events and the main roads in the city center are dominantly related to occupancy events. During the rush hour at 8-9h

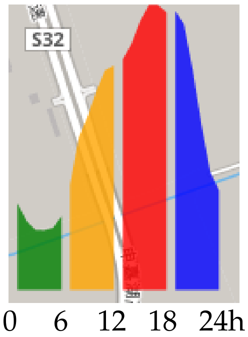


Figure 4.4: A schematic example of a time graph for a variable.

there are obvious places of blue sectors of drop-off events and the red sectors of pick-up events scattering around. On the contrary, during the rush hour at 18-19h, there are obvious places of red sectors of pick-ups. Besides, there is a big green sector of non-occupancy events in the airport.

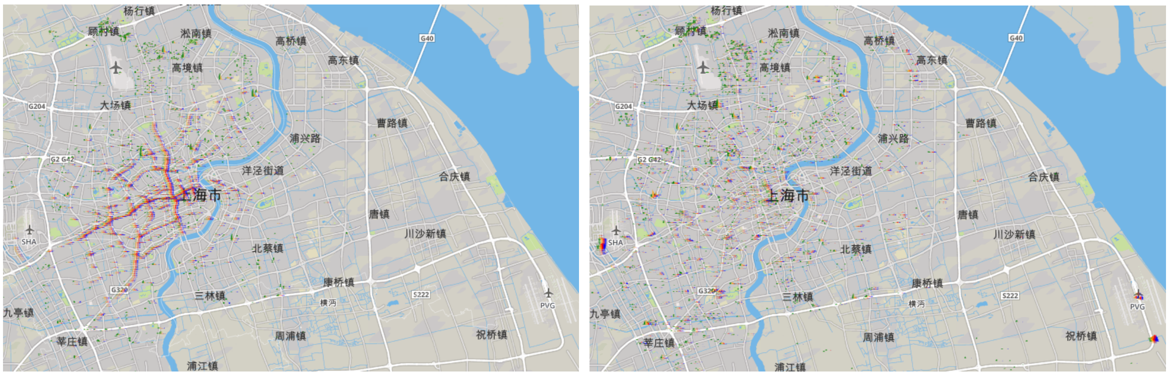
4.2.2 Time Graph

Besides using small multiples in Figure 4.3 to visualize the temporal patterns of the multivariate data, we want to simultaneously observe the temporal patterns of each individual variate at different time intervals. A time graph is designed to show the uni-variate temporal variations. It consists of four components as four area graphs, representing four time intervals, i.e., early morning (00:00 – 06:00), morning (06:00 – 12:00), afternoon (12:00 – 18:00) and evening (18:00 – 24:00). Colors of green, orange, red, and blue are assigned to the four area graphs to represent the four time intervals respectively. Each area graph is plotted using its 6 hourly event quantities with the value mapped to the height of the area. Figure 4.4 is a schematic example of a time graph.

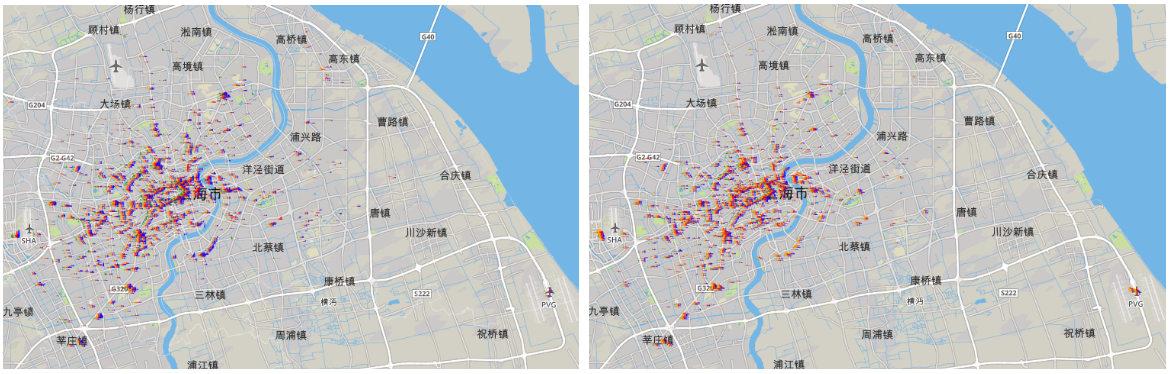
Similar to the pie radar glyph introduced in Section 4.2.1, each time graph is located in the center of each cell. Filtering, ordering and opacity setting approaches are applied to reduce the computational complexity and visual clutters. The time graph visualization of the individual types of events is shown in Figure 4.5.

From the compact symbols shown in Figure 4.5, we can observe the following patterns:

(1) The distribution of the occupancy events in Figure 4.5(a) exhibits quite obvious spatiotemporal patterns. In the time slots (06:00 – 24:00), the occupancy activities are highly concentrated in the city center continuously along the main roads (shown in orange, red and blue). This temporal phenomena is quite easy to understand since it is normal during the day time that main roads are mostly taken. However, we can also observe quite counter-intuitive phenomena that in the early morning (00:00 – 06:00)



(a) Occupancy (b) Non-Occupancy



(c) Pick-up (d) Drop-off

Figure 4.5: The temporal patterns of the four event types.

there are also intense occupancy activities (in green) in the north suburban areas in Shanghai. The reason might be that there are recording mistakes, for example, when the driver arrives at home he/she forgets to turn-off the GPS devices or switch the car status to non-occupied.

(2) The density of the non-occupancy events is unevenly distributed. Generally speaking, there are strong non-occupancy activities in the early morning (00:00 – 06:00) in the whole Shanghai areas, especially in the north part. In the other three time intervals, some hotspots are scattered around the Shanghai area. There are two prominent hotspots at the westernmost and easternmost corresponding to the Hongqiao and Pudong international airports.

(3) The spatial distributions of the pick-up and drop-off events are very similar to each other that events are concentrated around the city center with a few exceptions far from the city center. With regards to the temporal patterns, generally the pick-up and drop-off events are very active from 06:00 to 24:00. However, pick-up events happen more frequently from the afternoon till midnight (12:00 – 24:00) reflected by the dominant

red and blue colors in Figure 4.5(c), while drop-off events happen more frequently from morning till evening (06:00 – 18:00) reflected by the dominant orange and red colors in Figure 4.5(d).

4.2.3 Salient Feature Image

Section 4.2.1 and Section 4.2.2 can reveal the spatial and temporal distribution of movement events at a detailed level. In spite of the easily perceivable overall compact visualization results after data filtering, ordering and opacity setting, cognitive efforts and interactive operations are still highly demanded to explore and understand the cluttered symbols.

To simplify the visual complexity, we propose a salience-based method to represent only the most salient feature in each cell. More specifically, the co-located four types of events in each cell are regarded as four features and the feature of the largest value is selected as the salient feature. Formally, the selection of the salient feature is as follows:

$$salient\_feature_{c_i} = \mathrm{argmax}(occupancy_{c_i}, non\_occupancy_{c_i}, pick\_up_{c_i}, drop\_off_{c_i})$$

We apply the salience-based method to the normalized multivariate data. Figure 4.6 shows the visualization results. The same color scheme as previous sections is chosen to represent the four types of events. The quantities of the events are mapped to the opacity values. The smaller the value is, the more transparent the cell.

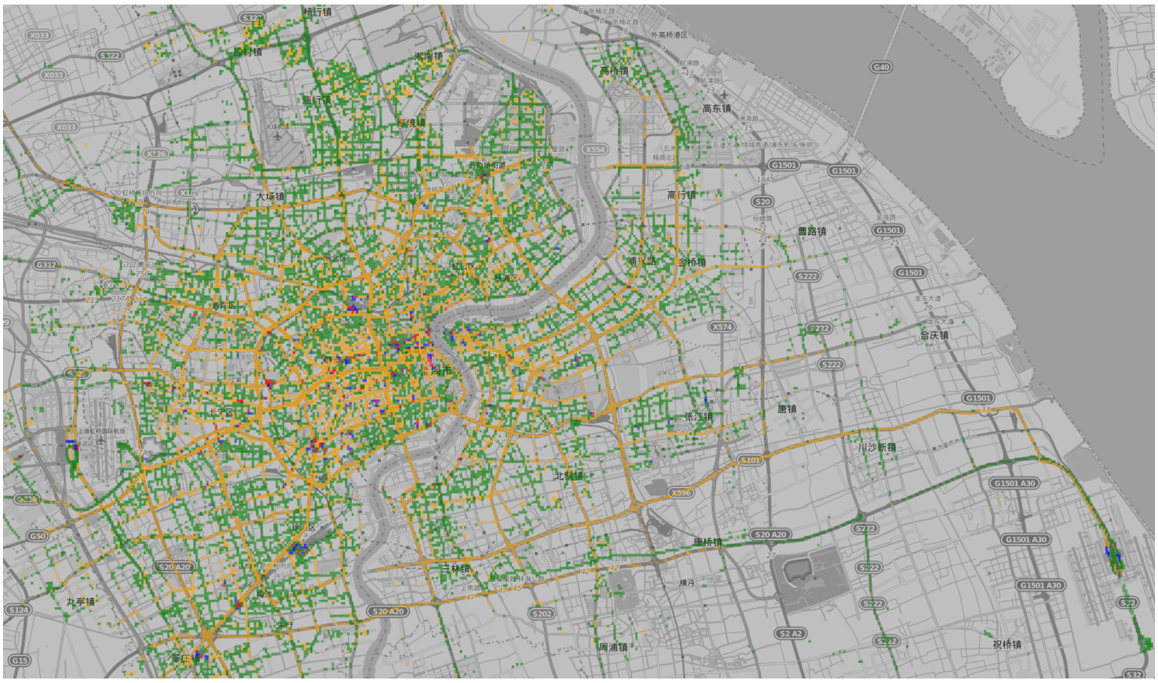
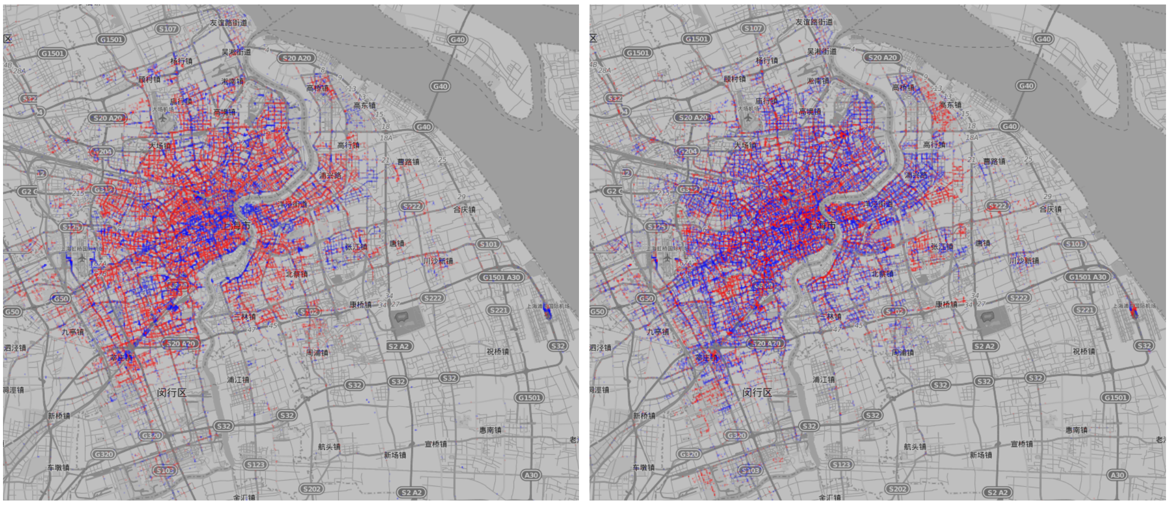


Figure 4.6: Salience-based map of four variables.

In Figure 4.6, occupancy (in orange) and non-occupancy (in green) events are the most salient features. Occupancy events are mostly distributed on the main streets and non-occupancy events on small streets as well as in off-street areas. Pick-up (in red) and drop-off (in blue) events are highly concentrated in only a few small places.

Besides the resulting spatial distributions, we can also investigate the temporal characteristics of the events using the salience-based method. Taking two hourly time intervals 7-8h and 18-19h of two event types - pick-up and drop-off - as exemplary features, the spatial distribution of two features in the two time intervals are demonstrated in Figure 4.7(a) and Figure 4.7(b).



(a) 7-8h

(b) 18-19h

Figure 4.7: The salience-based spatial distribution of pick-up and drop-off events at (a) 7-8h and (b) 18-19h.

From the two images, we can easily observe dense pick-up (red) and drop-off (blue) areas at the two time intervals as well as roughly complementary spatial distribution patterns of the salient features. For instance, the dense drop-off areas at 7-8h become dense pick-up areas at 18-19h.

In spite of the distinctive patterns of Figure 4.7(a) and Figure 4.7(b), comparing the two individual images side by side is still tedious and the temporal change patterns are not explicitly presented. An appropriate solution is to synthesize the two images into one temporal change map and do further spatial temporal analysis. More specifically, we view each cell as a pixel in the image and do algebraic addition operation on the corresponding pixels of two images. The result is a synthesized temporal change image.

With regard to the salience-based images of pick-up and drop-off events at 7-8h and 18-19h, we define three categories of temporal changes. The first category is the change of the salient feature from drop-off at 7-8h to pick-up at 18-19h. The second one is the change of the salient feature from pick-up at 7-8h to drop-off at 18-19h. In the last category, the salient feature is unchanged. More specifically, the salient features are either “pick-up” at 7-8h and 18-19h or “drop-off” at 7-8h and 18-19h, but the values of the salient feature might be changed at the two time intervals.

For the visual mapping, we use a color scheme of magenta, steel blue and yellow to represent these three categories. Figure 4.8 shows the three categories of the temporal changes.

Furthermore, we calculate the amount of change in each cell as follows. For the first

	7h–8h		18h–19h
🟧	Drop-off	→	Pick-up
🟦	Pick-up	→	Drop-off
🟨	Pick-up	→	Pick-up
	Drop-off	→	Drop-off

Figure 4.8: The color scheme for temporal change from 7-8h to 18-19h.

category, the amount of the change in a cell is calculated by adding up the quantity values of drop-off events at 7-8h and of pick-up events at 18-19h. Similarly, the amount of temporal change of the second category is the addition of the quantity value of pick-up events at 7-8h and the value of drop-off events at 18-19h. With regard to the third category, the amount of change is the difference between the quantities at 7-8h and 18-19h. The resulting value of change in each cell is then proportionally mapped to an opacity value.

Figure 4.9 shows the temporal change image of the pick-up and drop-off events at 7-8h and 18-19h, which allows a straightforward detection of change patterns during the corresponding time intervals. With regard to the spatial extent, the dominant temporal change is the change of the salient variable from pick-up to drop-off events (in steel blue), occupying almost the whole Shanghai area. The next spatially large temporal change is the change of salient variable from drop-off to pick-up events, which can be identified by several magenta areas of irregular size. Finally, we can detect quite a few relatively small but very shiny yellow areas scattered in Shanghai, which indicates these areas are of either significant pick-up or drop-off changes during this time intervals.

Stimulated from the spatiotemporal pattern in Figure 4.9 and based on the knowledge of the regular human activity patterns, we naturally associate the temporal change patterns with distinct land use types. For instance, areas with intense drop-off activities at 7-8h and intense pick-ups at 18-19h probably correspond to the working places. There is a class of land use/cover change detection problems in a variety of fields. An example in remote sensing field is understanding landscape dynamics by applying image change detection techniques to multi-temporal satellite imagery. With the increasing usage of data collected from mobile devices and social medias, some hot research topics have emerged in recent years with the aim to infer land use types from mobile phone data [Gra+15], GPS trajectories [YZX12], social network [Jia+15], and the combination of a variety of data sources, like GPS and social media data [Liu+15b; MT15].

To deeply understand the temporal change patterns in Figure 4.9, we enclose areas of interest in orange and yellow with irregular polygons. These areas of interest are then manually checked with the Shanghai base map and labeled with their functions or names. Figure 4.10 shows the distinct areas of interest.

In Figure 4.10, the irregular polygons with orange frames and labels are regions espe-

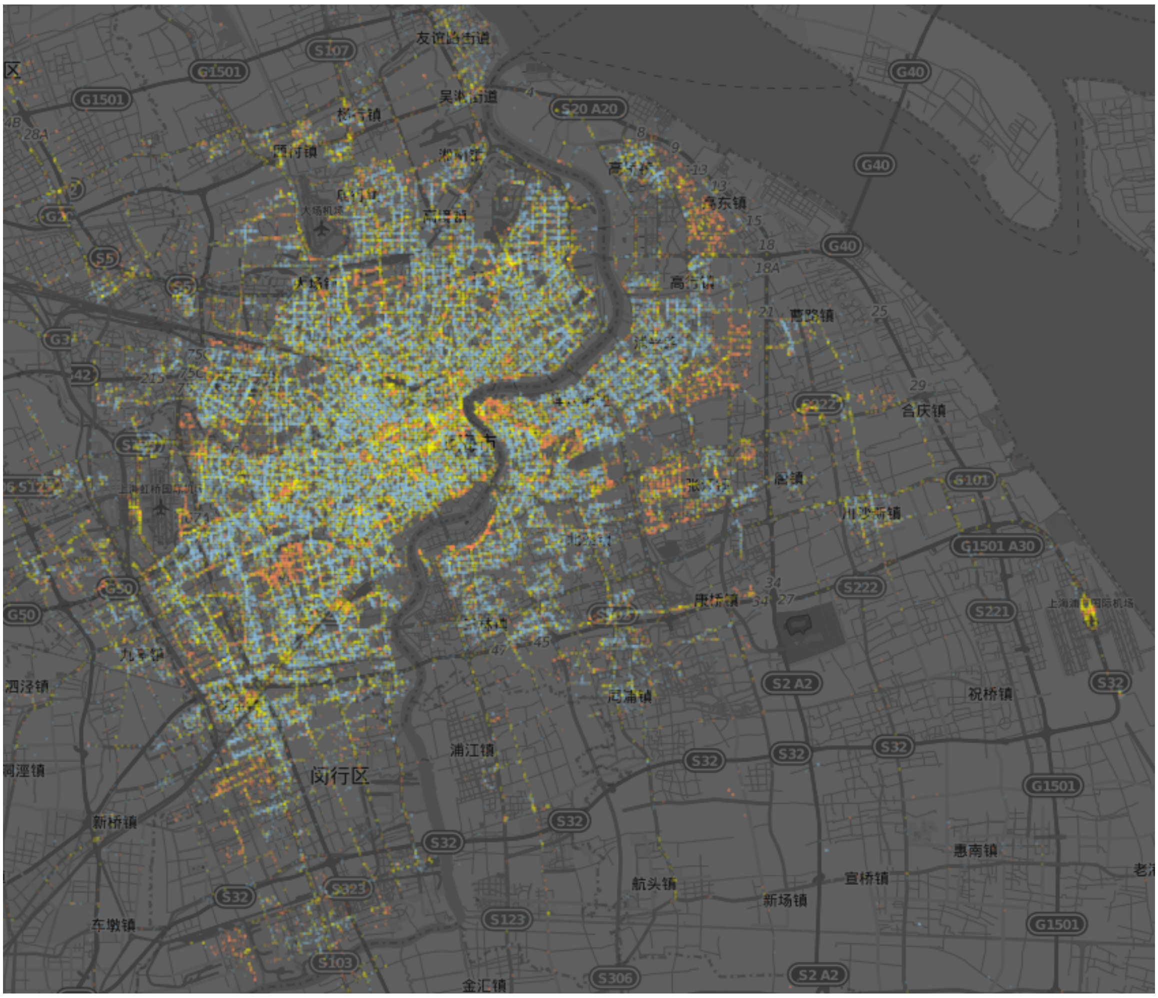


Figure 4.9: The salience change map.

cially active with drop-off events at 7-8h and pick-up events at 18-19h. As Castro et al. [Cas+13] explained, people use vehicles for commuting to and from work, for regular and 'irregular' chores, and for leisure activities. We would infer that these regions might correspond to the working places where people take taxis to the place in the morning and back home in the evening. After manually checking the functions of these regions, we find that most of them are free trade zones, high-tech zones, industrial and development zones, and financial centers, which match our assumptions fairly well. Interestingly, we find that the area of Shanghai Expo 2010, located on both banks of the Huangpu River, is also of this kind of temporal change pattern. Since Expo 2010 was held from 1 May to 31 October 2010 and covered the total time span of our data set, it is reasonable that the data during this time period can reveal the human mobility patterns related to this international events.

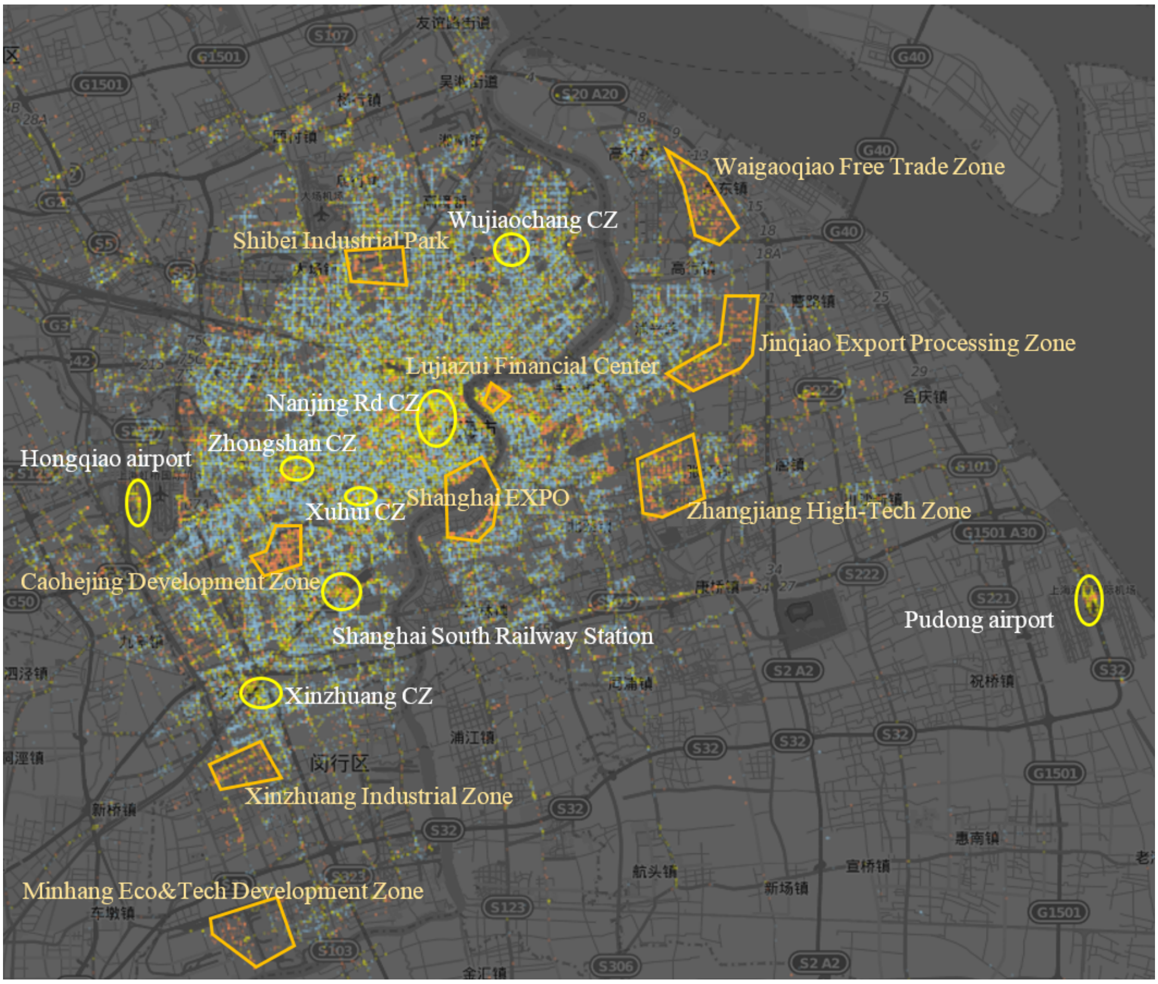

Waigaoqiao Free Trade Zone
Wujiaochang CZ
Shibei Industrial Park
Jinqiao Export Processing Zone
Lujiazui Financial Center
Nanjing Rd CZ
Zhongshan CZ
Hongqiao airport
Xuhui CZ
Shanghai EXPO
Zhangjiang High-Tech Zone
Caohejing Development Zone
Shanghai South Railway Station
Pudong airport
Xinzhuang CZ
Xinzhuang Industrial Zone
Minhang Eco&Tech Development Zone

Figure 4.10: Regions of Interests. (CZ stands for commercial zone.)

This kind of event-driven land use type usually happens during limited time intervals and is a sort of come and go type, which is termed “Ad hocs” POI/AOI. Regretfully, we do not have a data set of longer time span to test whether the mobility pattern in this area disappears or changes prior or post the Expo event.

Similarly, the areas enclosed by ellipses with yellow frames and white labels are associated with significant changes of either pick-up or drop-off events from 7-8h to 18-19h. By reading the corresponding regions on the base map, we find most of the areas are transport hubs and commercial zones (CZs). For example, the easternmost and westernmost ellipses correspond to Pudong and Hongqiao international airports. The other ellipses are mostly commercial zones, e.g., the commercial zones of Wujiaochang in the north, Xinzhuang in the south, and Nanjing road, Zhongshan and Xuhui in the city center.

Since the steel blue pixels cover almost the whole Shanghai area, it is not so interesting to explore them in detail at the moment. Nevertheless, we would infer from the temporal pattern of dense pick-ups at 7-8h and drop-offs at 18-19h that these areas are probably of residential function.

4.3 Visual Communication based on Thematic Mapping

In spite of the advantage of examining multivariate events at a very fine granularity, highly cognitive efforts are required to understand the compact symbols in the visualization results. To reduce the visual complexity and let users comprehensively explore the events, we can aggregate events to coarser but appropriate levels. Depending on various analysis tasks, the data can be aggregated to diverse spatial partitions. For instance, if the task is related to the administrative divisions, the aggregation is necessarily conducted on the predefined administrative regions. Otherwise, the spatial partition can be proceeded based on the underlying data distribution.

In this section, we focus on communicating spatiotemporal distributions at high levels by virtue of thematic mapping techniques. We firstly group event data through hierarchical clustering methods, based on which the study area are divided into irregular subregions by means of Voronoi tessellations. The event aggregates calculated through the Voronoi partitions can reveal the underlying data distribution quite well and are used for choropleth and proportional symbol mapping.

4.3.1 Cluster Analysis

Spatial clustering, which groups similar spatial objects into classes, can be used as a standalone tool to gain insight into the data distribution, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. It may also serve as preprocessing step for other algorithms, such as classification and characterization, which will operate on the detected clusters [HKT01]. There are several popular clustering methods, including partitioning method (e.g. k-means), hierarchical method, density based methods (OPTICS, DBSCAN).

In this work, a hierarchical agglomerative clustering method [KR90] is applied to group similar events and serves as a pre-processing step for future classification. The principle of the algorithm is that for a given set of data objects it is hierarchically decomposed, forming a dendrogram - a tree that splits the database recursively into small subsets. The dendrogram can be formed either ‘bottom-up’ or ‘top-down’. The agglomerative approach adopts the ‘bottom-up’ way. It starts with each object forming a separate group and successively merges the objects or groups according to some measures like the distance between the two group centers, which is done until a termination condition holds.

We choose the hierarchical clustering method for two reasons. Firstly, it allows an easy investigation of the clustering results at multiple scales. Secondly, since each cell is already associated with a density value, we use this value as the third dimension with the spatial location of the input data. The algorithm has two parameters, namely a distance function and a linkage criterion. The parameter setting is largely dependent on the applications or tasks. The output is the assignment of events to its cluster.

There are two possibilities in applying the method to clusters of the four types of events. One is to do the tessellation of the territory once using a summarized variate, e.g. the total number of the four types of events in each cell. Then, the aggregation is done separately for each variate. The other possibility is to do the tessellation separately for each type of the events. The latter one can reflect the data distribution of each type of events quite well but is hard for the comparison and analysis of the relationship among the variables. In this work, we adopt the former one using the same spatial tessellation for the four types of events. The variable for generating the tessellation is the total number of events inside the cells, which is the sum of the (O, N, P, D) events. This tessellation can roughly reflect the data distribution of the four variables.

The parameters for the clustering are chosen as follows. Taking into account of the spatial scale of the analysis (the whole territory of Shanghai) and the sparseness of the relevant events in suburban areas, the maximal neighborhood radius has been set to a Euclidean distance of 700m and the linkage criteria is set to the average. The resulting number of the representative clusters is 93. Figure 4.11 illustrates the clustering results. Each cluster is represented as a gray polygon in which the small gray dots represent cluster elements. The gray polygons are the convex hulls of the corresponding cluster elements. The black circles represent the cluster centroids. The resulting cluster pattern in Figure 4.11 shows that far more clusters are distributed in the city center than the surrounding areas.

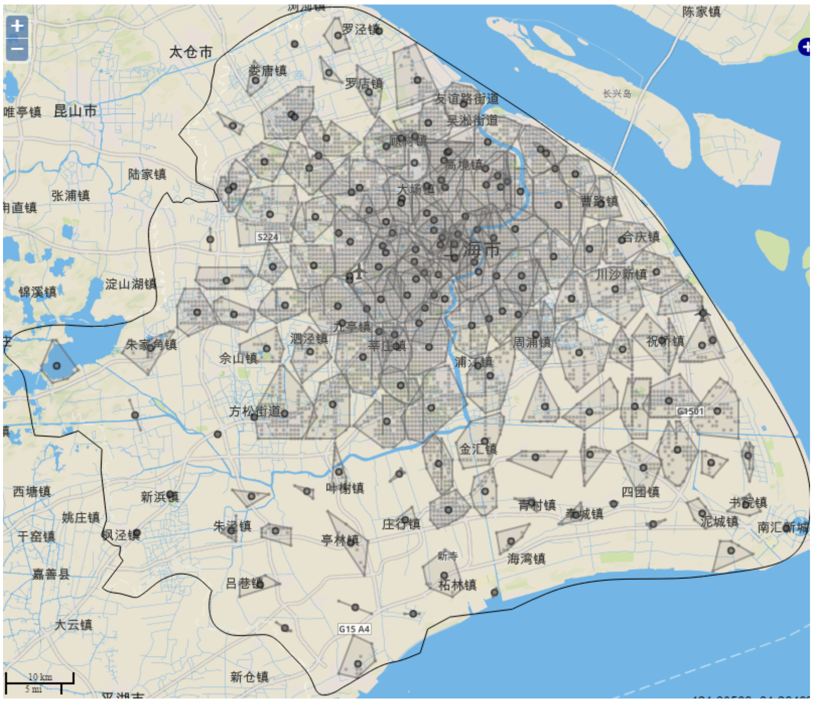
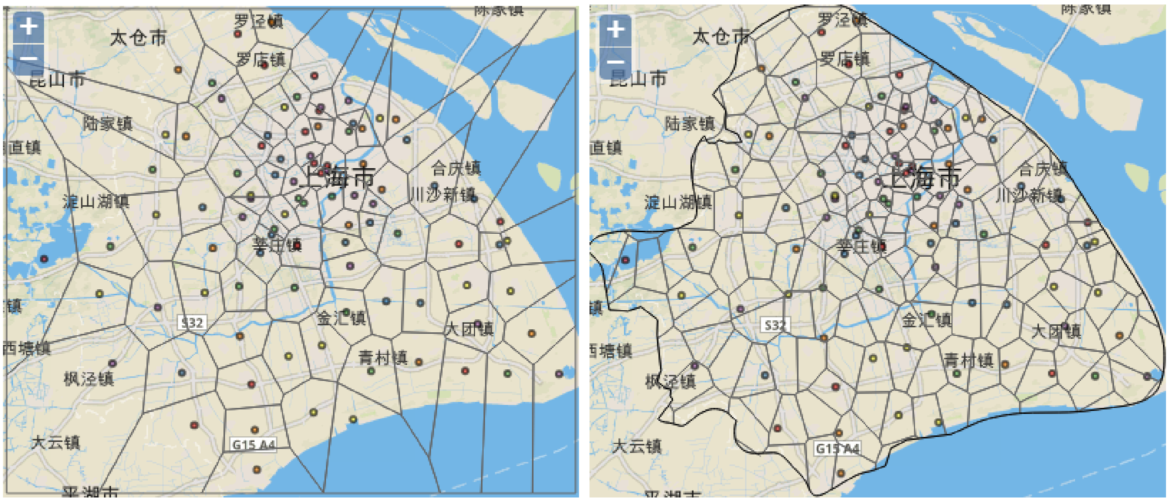


Figure 4.11: Hierarchical clustering results.

4.3.2 Spatial Partition and Aggregation

The cluster results in Figure 4.11 reflect the natural spatial divisions of underlying event similarities. Based on the clustering results, we partition the study area into corresponding sub-regions as follows.

- Voronoi tessellation. To divide the territory into appropriate compartments, we apply the Voronoi tessellation method. The centroids of the 93 clusters are the generating points for Voronoi cells. The resulting Voronoi cells are shown in Figure 4.12(a). Due to the uneven distribution and the sparseness of data in the suburban areas, the resulting Voronoi cells are of arbitrary sizes, some of which are especially large.
- Adding additional generating points. To obtain cells of more even sizes and shapes, we adopt the similar method described in [AA11] to generate additional generating points in the areas where there are no characteristic points from the original data, i.e. the clustering centroids. The principle of the method is that we preset the maximum size of the Voronoi cells. If the size of a Voronoi cell exceeds the predefined maximum size (e.g. 500m), then the Voronoi cell is split into smaller cells.
- Clipping the area. From the accuracy and aesthetic point of view, the Voronoi polygons inside the Shanghai border are clipped. The number of the Voronoi cells after



(a) The original Voronoi tessellation

(b) The tessellation with additional generating points and boundary clipping.

Figure 4.12: The Voronoi tessellation.

adding the additional generating centroids and the clipping of the Shanghai boundary is 197. Figure 4.12(b) shows the final spatial partitions.

Based on the spatial partitions in Figure 4.12(b), for each type of (O, N, P, D) events we aggregate the total number of its events to the corresponding Voronoi cell. Additionally, we compute their statistics by the hourly time intervals. The aggregates are then applied in the following sections for choropleth and proportional symbol mapping.

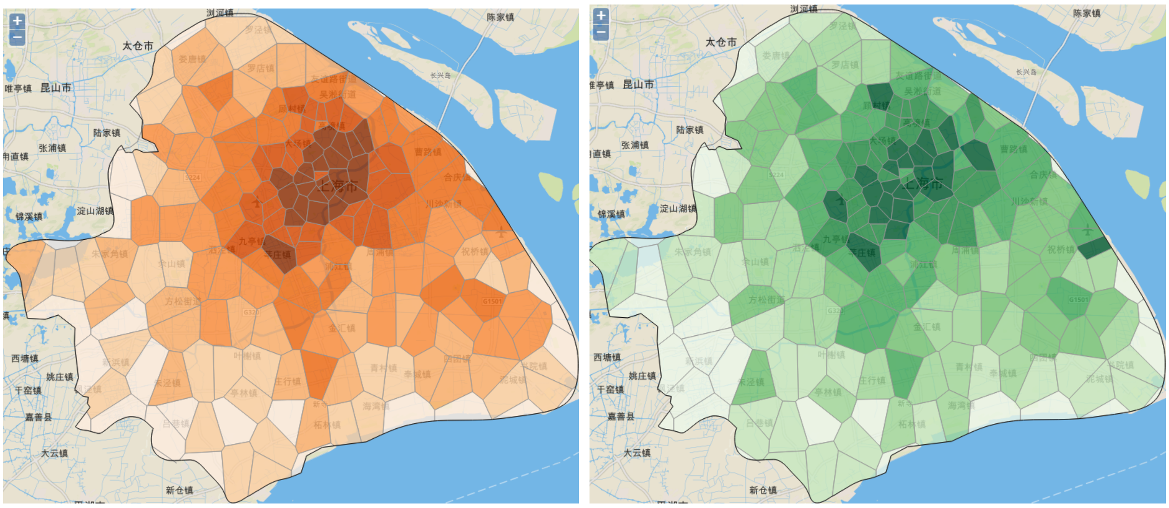
4.3.3 Univariate Choropleth Mapping

A choropleth map is a thematic map in which areas are shaded or graded in proportion to the measurement of the statistical variable being displayed on the map. It provides an easy way to visualize how a measurement varies across a geographic area. In this section, we adopt the choropleth mapping technique for visualizing the (O, N, P, D) events.

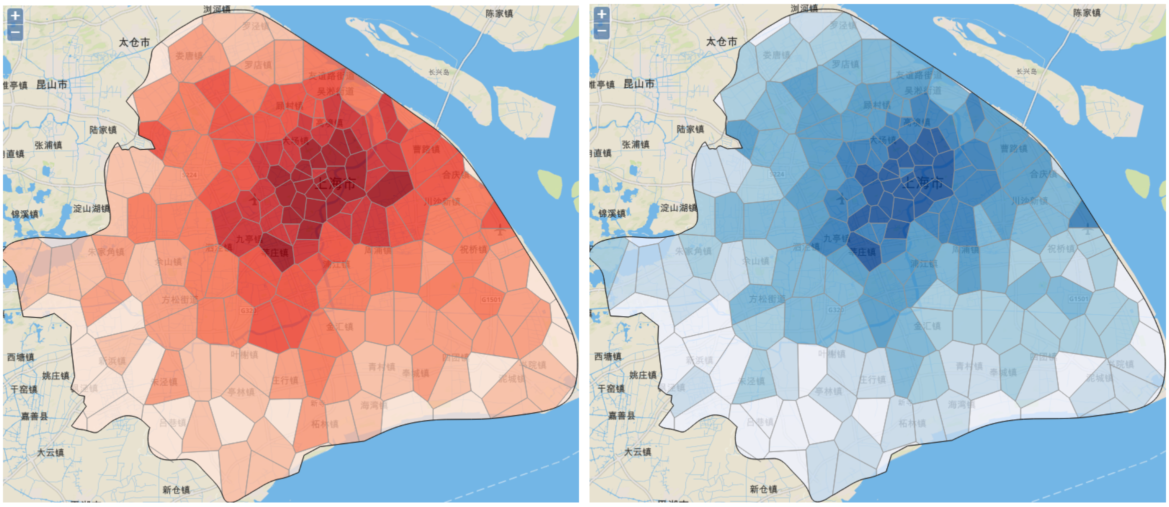
According to the cartographic design principles, the aggregates generated from Section 4.3.2 are classified into 7 classes for an easy perception of the spatial distribution. Equal interval classification methods are chosen to generate a balanced visual appearance with each class of the same number of class members. The same color scheme used in Section 4.2.1 is adopted for the four variates. An opacity value of 0.7 is applied to each spatial division for orientation purpose.

Figure 4.13 shows the univariate choropleth maps of the four variables respectively. The general distributions of the four types of events reveal that the density gradually decreases from the city center to the suburban areas. In the southwest of the city center, there is an obvious outlier area of high dense events, which is distinct with its neighboring areas. This outlier corresponds to a town called Xinzhuang, where a commercial zone and an industrial zone are located.

Compared to the other types of events, the non-occupancy events are more disperse and have several dense areas outside the city center. For instance, there is one dense event area in the north part of the city and two other dense areas located at the easternmost and westernmost of the city. The pick-up and drop-off events have similar spatial distribution patterns except that the drop-off patterns are more compact.



(a) Occupancy (b) Non-occupancy



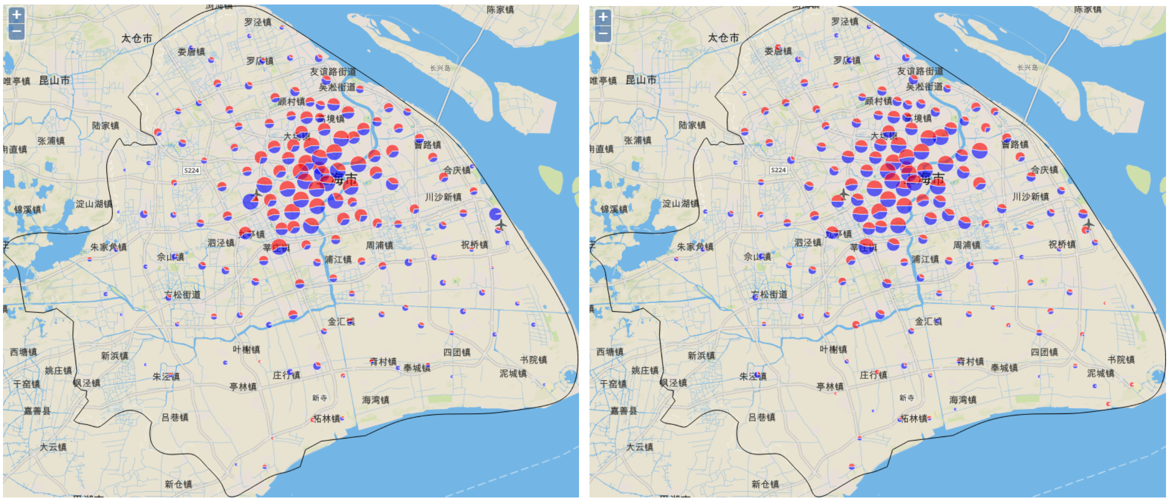
(c) Pick up (d) Drop off

Figure 4.13: The univariate choropleth maps of (O, N, P, D) events.

4.3.4 Thematic Mapping of Temporal Changes

This section investigates the spatiotemporal patterns of the events and the temporal change patterns. To be consistent with Section 4.2.3, the pick-up and drop-off events at 7-8h and 18-19h are extracted and mapped using thematic maps for the illustration of their effectiveness.

We adopt mainly two thematic mapping techniques to visually represent the spatiotemporal patterns as well as the temporal change patterns. Firstly, proportional symbol mapping techniques are applied to show the spatiotemporal distributions of the total number of the pick-up and drop-off events at the two time intervals. Pie charts are selected to represent the two variables. The size of the pie chart is proportional to the sum of the total number of each variable. Each pie chart has two sectors. Red and blue are assigned to the sectors representing the two variables and the size of each sector is proportional to the quantity of the aggregate of each variable. The total number of pick-up and drop-off events are classified as usual into 7 classes. An opacity value of 0.7 is applied for the orientation purpose. Figure 4.14(a) and Figure 4.14(b) illustrate the bivariate proportional symbol mapping results. The spatial distribution of each variable and the temporal differences can be clearly identified.



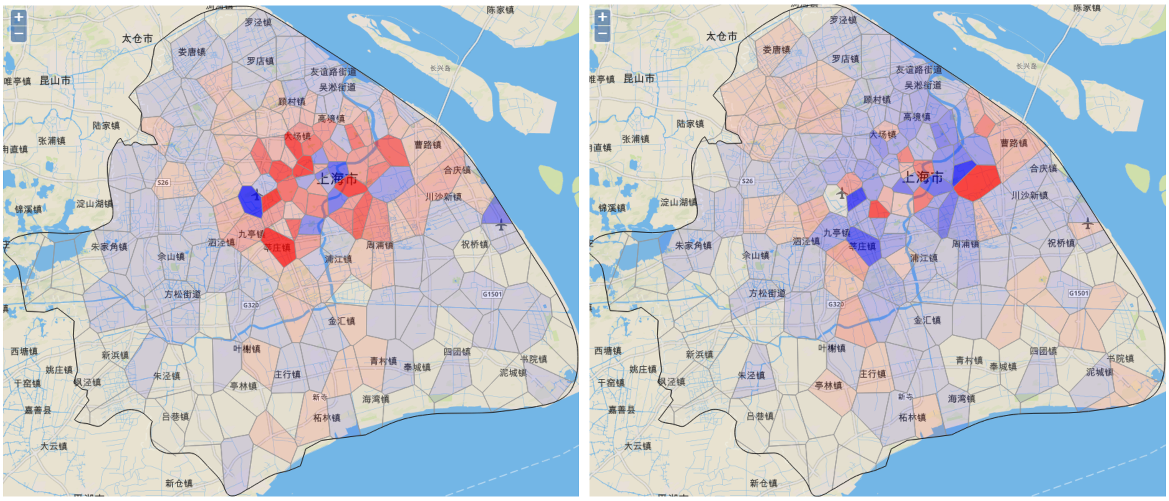
(a) 7-8h

(b) 18-19h

Figure 4.14: Bivariate proportional symbol mapping of the spatial distribution of (P, D) events.

Figure 4.14(a) and Figure 4.14(b) show the spatiotemporal patterns of the pick-up and drop-off events at 7-8h and 18-19h. At 7-8h, there are several places with high proportion of drop-off events, for instance, in the city center as well as two extreme drop-off dense areas in the two international airports. Compared with the spatial patterns at 7-8h, the proportions of pick-up and drop-off events are equal at 18-19h except some small unequal quantities in the suburban areas.

Furthermore, the choropleth mapping technique is chosen to map the salient variables of larger aggregated values inside the Voronoi cells. For each Voronoi cell, we calculate the salient variable and its quantity value. The principle of calculating the salient variable and its value is the same as the calculation of the salient feature in Section 4.2.3. Two color hues red and blue represent the two variables. The values are classified into 7 classes, each being represented by a color value. An opacity of 0.7 is applied for orientation purpose. Figure 4.15(a) and Figure 4.15(b) show the spatiotemporal distribution of the salient variables at 7-8h and 18-19h.



(a) 7-8h

(b) 18-19h

Figure 4.15: Choropleth mapping of salient features of (P, D) events.

Figure 4.15(a) and Figure 4.15(b) show respectively the distributions of the salient variables and their quantities at 7-8h and 18-19h. These two maps reveal a trend of opposite patterns. Interestingly, the region around the Hongqiao international airport has nearly no values at 18-19h. If we check the proportional symbol map at 18-19h, we can see that the quantities of pick-up and drop-off events are both large but quite close to each other, resulting in small change values.

Finally, a synthesized choropleth map is designed to show the temporal change pattern. We calculate the temporal difference of the distribution at the two time intervals and categorize them into three categories. The categories and the calculation of the quantity of the temporal changes as well as the color assignment are the same as Section 4.2.3. The difference is that the change quantities of each category are classified into 4 classes and mapped to 4 color values of each color hue.

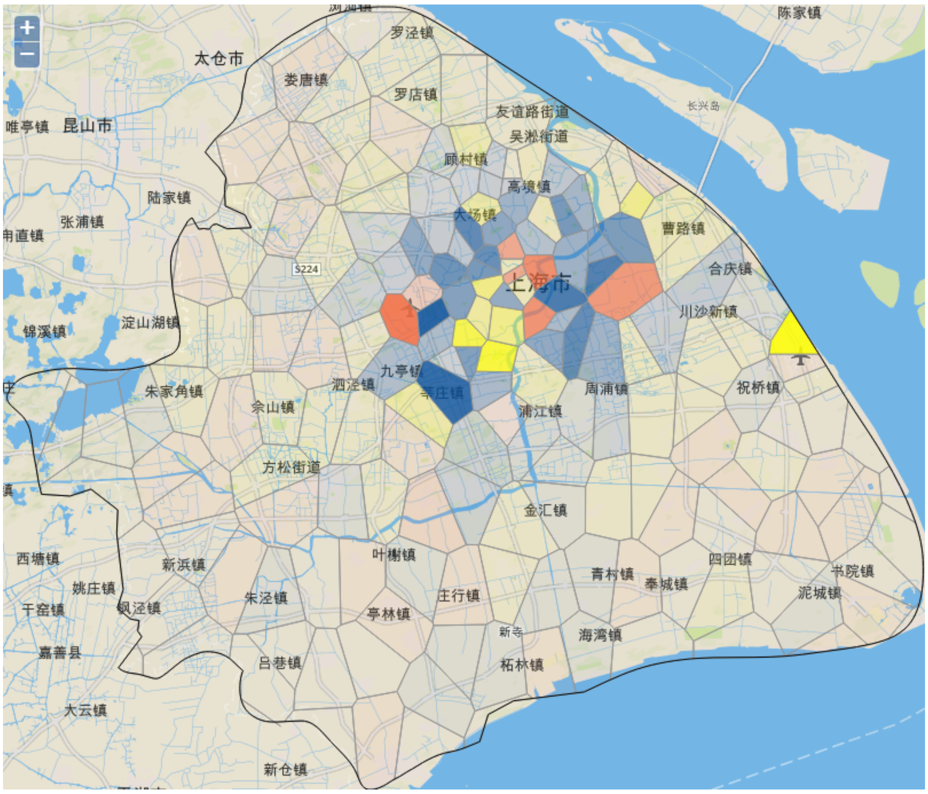


Figure 4.16: Temporal change map of salient features of (P, D) events from 7-8h to 18-19h.

The choropleth map in Figure 4.16 illustrates the temporal change patterns. The first category with significantly active pick-ups at 7-8h and drop-offs at 18-19h (Blue) is areas of Xinzhuang, around the Hongqiao airport, and of the Lujiazui financial center. The second category of active drop-offs at 7-8h and pick-ups at 18-19h (red) is Hongqiao international airport and the Lujiazui financial center etc. The third category of no salient feature changes but of large quantity change is the Pudong international airport and Shanghai south railway station etc.

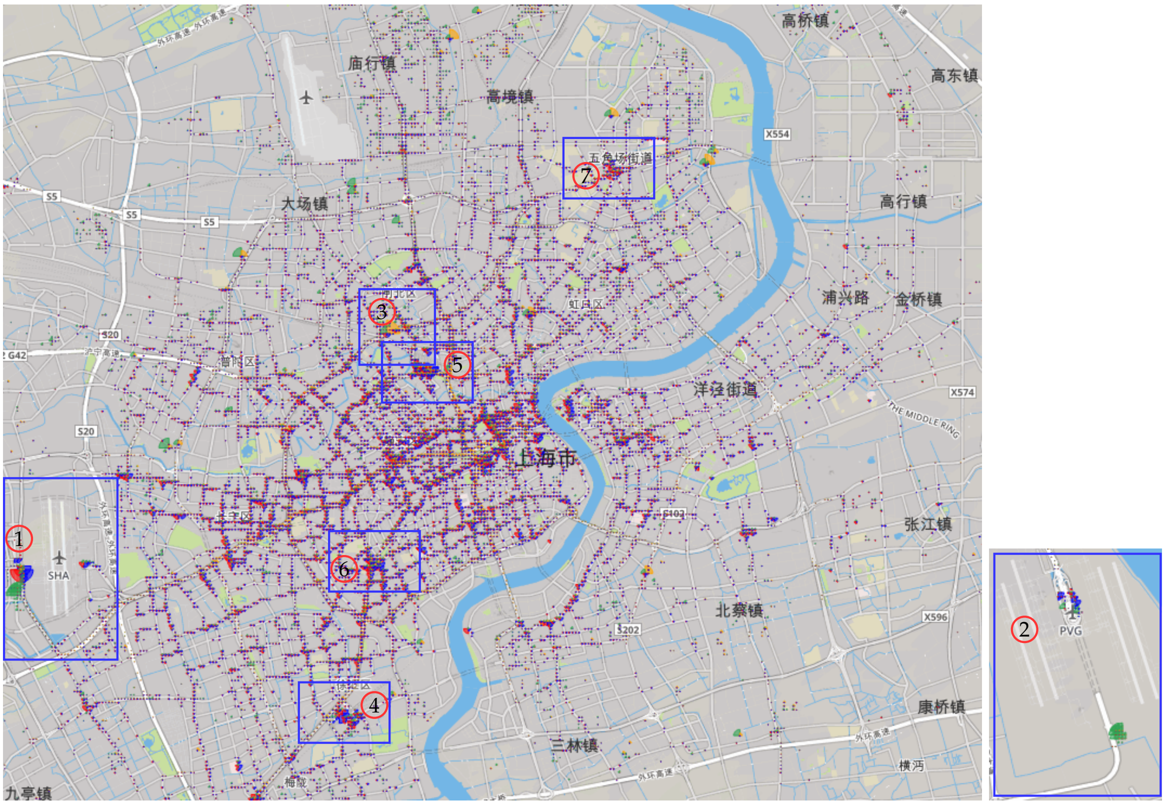
4.4 Analysis of the Visualization Results

This section summarizes and systematically analyzes the visualization results from the previous sections. The analysis and the findings are as follows:

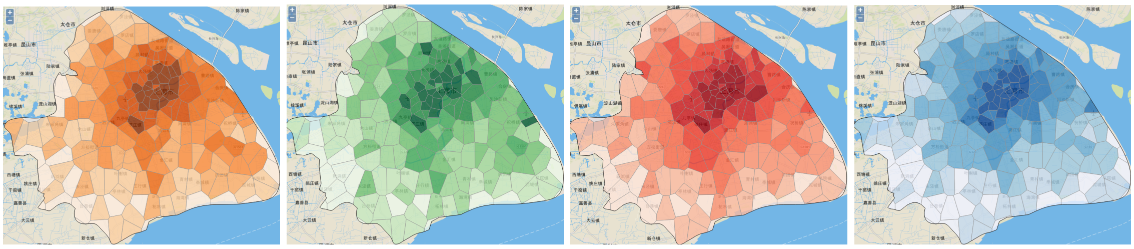
The spatial density distribution of the four types of events. Generally, the overall density of each type of events decreases gradually from the city center to the suburban areas as shown in Figure 4.2 (see Figure 4.17(a)) and Figure 4.13 (see Figure 4.17(b)), except that in Figure 4.13(b) the non-occupancy events are more disperse and have several dense areas outside the city center. Besides, in Figure 4.13 we can detect an obvious outlier in the southwest of the city center, which is located in a town called Xinzhuang and is very likely related to a commercial and an industrial zone. Furthermore, in Figure 4.2, we can observe several significantly dense places. For the occupancy (orange) and non-occupancy (green) events, there are only a few dense places, which means that the occupancy and non-occupancy events are unevenly distributed. On the contrary, pick-up (red) and drop-off (blue) events are relatively more evenly distributed. We zoomed in those hotspots and examined them carefully with the base map. The analysis results show that Areas (1) and (2) with dense non-occupancy events (in green) and relatively dense pick-up and drop-off events correspond to two international airports. Area (3) of dense occupancy events is a road intersection. Areas (4) and (5) with very high densities of pick-up and drop-off events are two railway stations. Areas (6) and (7) of dense pick-up and drop-off activities are two very famous commercial centers.

The temporal patterns of the events. In general, the overall temporal distribution of the four types of events reveals large differences. Figure 4.5 (see Figure 4.18) illustrates large temporal variations. The occupancy events in the early morning (00:00 – 06:00) are mostly distributed in the northern part of the Shanghai area and in the other time slots are highly concentrated in the city center along the main roads. The density of the non-occupancy events are unevenly distributed with more non-occupancy activities in the early morning (00:00 – 06:00) in the whole Shanghai area, especially in the northern part, and in the other three time slots, there are two hotspots with very large values in the two international airports. The pick-up and drop-off events have similar spatial patterns and temporal distributions from 06:00 to 24:00. However, there are temporal lags of pick-up and drop-off events with more pick-up events in the afternoon till midnight (12:00 – 24:00) and more drop-off events in the morning till evening (06:00 – 18:00).

The spatiotemporal patterns of the events. Previous visualization techniques show significant spatiotemporal patterns of the four variates. Figure 4.3 (see Figure 4.19(a)) presents very distinct spatiotemporal patterns of the multivariate events at four hourly time slots, i.e. 3-4h, 6-7h, 8-9h and 18-19h. At 3-4h, pick-up events are relatively active mostly in the city center. At 6-7h, there are a few significantly dense drop-off places concentrated in the big transport hubs. During rush hours at 8-9h and 18-19h, the events become dense in the whole area with the main roads in the city center mostly of occupancy events (orange). At 8-9h there are quite a few places with dense drop-off events while at 18-19h these places are of more pick-up events. In terms of the non-occupancy events, they are highly concentrated in the Hongqiao airport reflected by the big green pie radar sector. When examining the spatiotemporal pattern using the salience-based method, we also found their significant spatiotemporal distributions. Figure 4.7 (see Figure 4.19(b)) uses pick-up and drop-off events as exemplary variables to show roughly complementary spatial distributions at the two time slots, i.e. 7-8h and 18-19h. This stimulates the interpretation of the land use types related to the usage of the pick-up and drop-off events. Figure 4.9 (see Figure 4.19(c)) synthesizes the temporal changes of Figure 4.7 at a disaggregated level. At an aggregated level, the proportional symbol maps in Figure 4.14 (see Figure 4.19(d)) illustrate the different distributions of the pick-ups and drop-offs at 7-8h and 18-19h. At 7-8h there are several places with high proportion of drop-off events in the city center and two extreme dense areas of drop-offs in the two international airports. At 18-19h the quantities of pick-up and drop-off events are nearly the same except a few regions in the suburban areas. Figure 4.15 (see Figure 4.19(e)) further abstracts salient variable of pick-up and drop-off events and shows a complementary temporal patterns at 7-8h and 18-19h. Figure 4.16 (see Figure 4.19(f)) synthesizes the two figures in Figure 4.15 and explicitly shows significant temporal changes in one choropleth map.

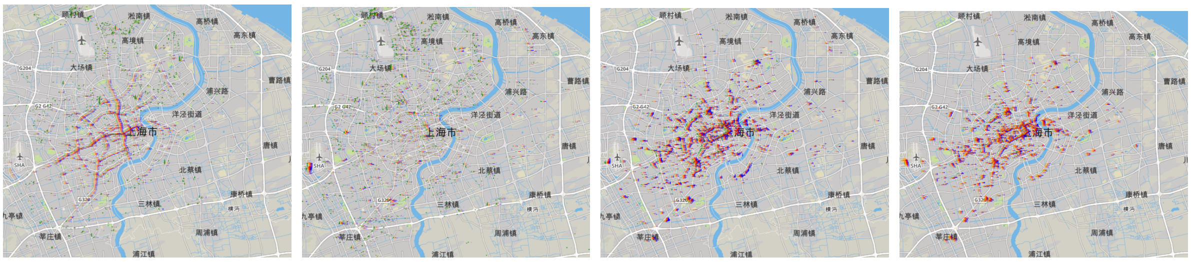


(a) Figure 4.2



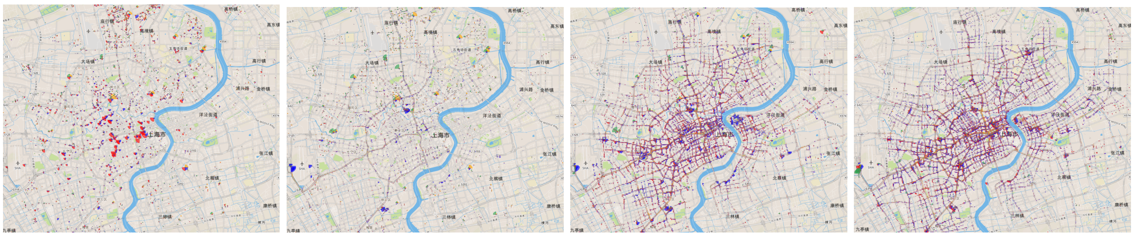
(b) Figure 4.13. From left to right (O, N, P, D).

Figure 4.17: Spatial density distributions of (O, N, P, D) events using (a) pie radar glyph and (b) choropleth mapping.

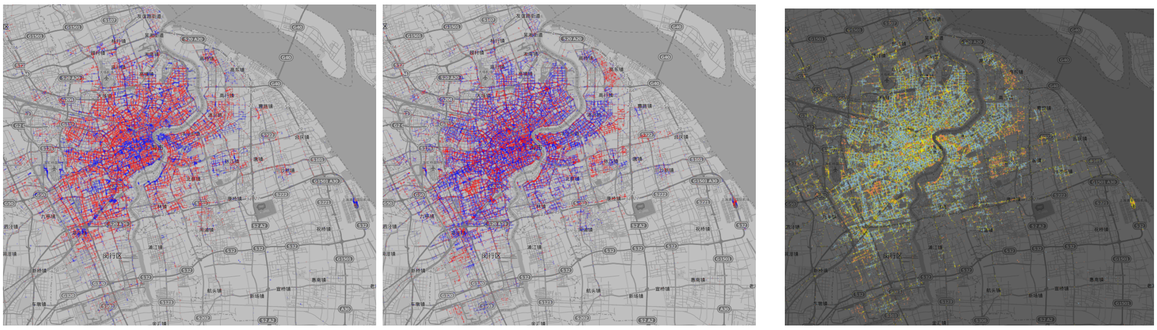


(a) Occupancy (b) Non-occupancy (c) Pick-up (d) Drop-off

Figure 4.18: The temporal patterns of the events in Figure 4.5.

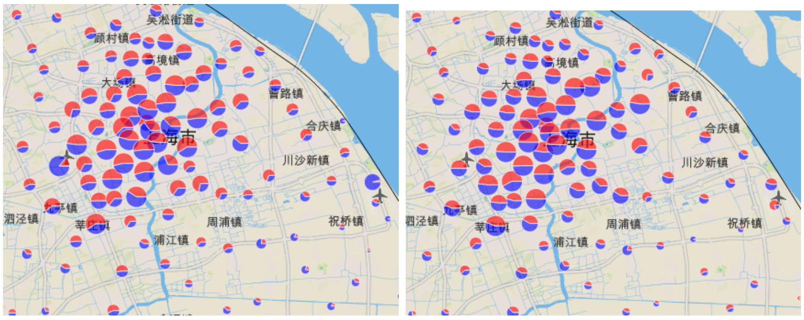


(a) Figure 4.3

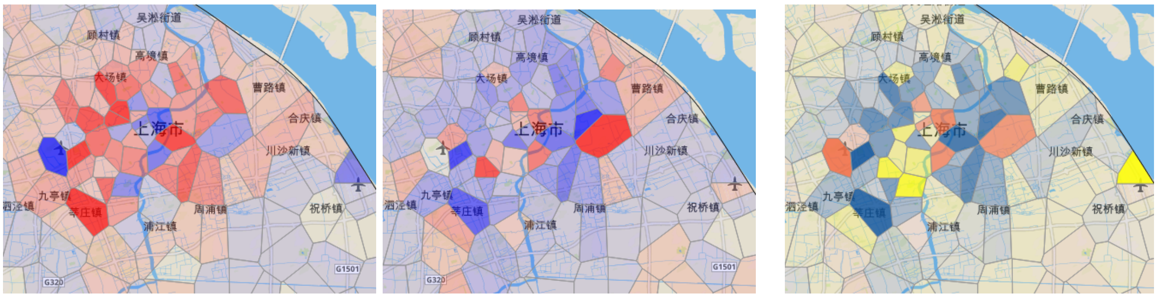


(b) Figure 4.7

(c) Figure 4.9



(d) Figure 4.14

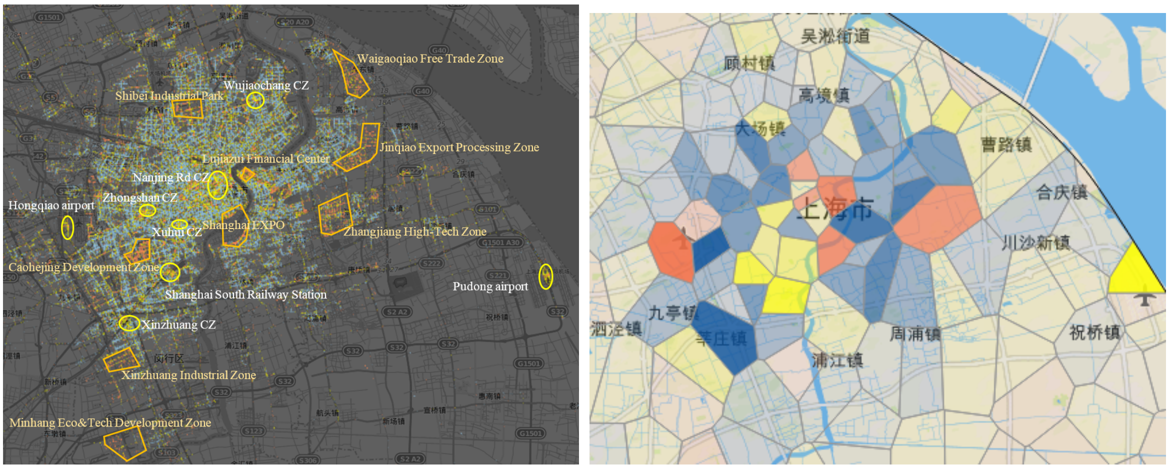


(e) Figure 4.15

(f) Figure 4.16

Figure 4.19: The spatiotemporal patterns of (O, N, P, D) events using: (a) pie radar glyph; (b, c) salience-based method; (d) proportional symbol mapping; and (e, f) choropleth mapping.

Temporal changes and land use. Figure 4.10 (see Figure 4.20(a)) and Figure 4.16 (see Figure 4.20(b)) illustrate the temporal changes at a disaggregated and an aggregated level respectively. The analysis of these temporal changes obviously reveals distinct land use types. In Figure 4.10, the significant areas with more drop-off events at 7-8h and more pick-up events at 18-19h are mostly free trade zones, high-tech zones, industrial and development zones, and financial centers. The areas with significant changes of drop-off or pick-up events mostly correspond to transport hubs and commercial zones. We also suppose that the areas with more pick-ups at 7-8h and more drop-offs at 18-19h are probably residential areas. At the aggregated level, the choropleth map in Figure 4.16 illustrates the final temporal changes in three categories. The regions (blue) with more active pick-up at 7-8h and drop-off at 18-19h include Xinzhuang, the Hongqiao airport and the areas around the Lujiazui financial center. The regions (red) which are more active drop-off at 7-8h and pick-up at 18-19h include Hongqiao international airport and the Lujiazui financial center etc. The regions (yellow) which are always active of drop-off at 7-8h and 18-19h or always active of pick-up correspond to the Pudong international airport and Shanghai south railway station etc.


Waigaoqiao Free Trade Zone
Wujiaochang CZ
Shibei Industrial Park
Jinqiao Export Processing Zone
Lujiazui Financial Center
Nanjing Rd CZ
Zhongshan CZ
Hongqiao airport
Shanghai EXPO
Xuhui CZ
Zhangjiang High-Tech Zone
Caohejing Development Zone
Shanghai South Railway Station
Pudong airport
Xinzhuang CZ
Xinzhuang Industrial Zone
Minhang Eco&Tech Development Zone
吴淞街道
顾村镇
高境镇
大场镇
曹路镇
上海市
合庆镇
川沙新镇
九亭镇
泗泾镇
莘庄镇
周浦镇
浦江镇
祝桥镇

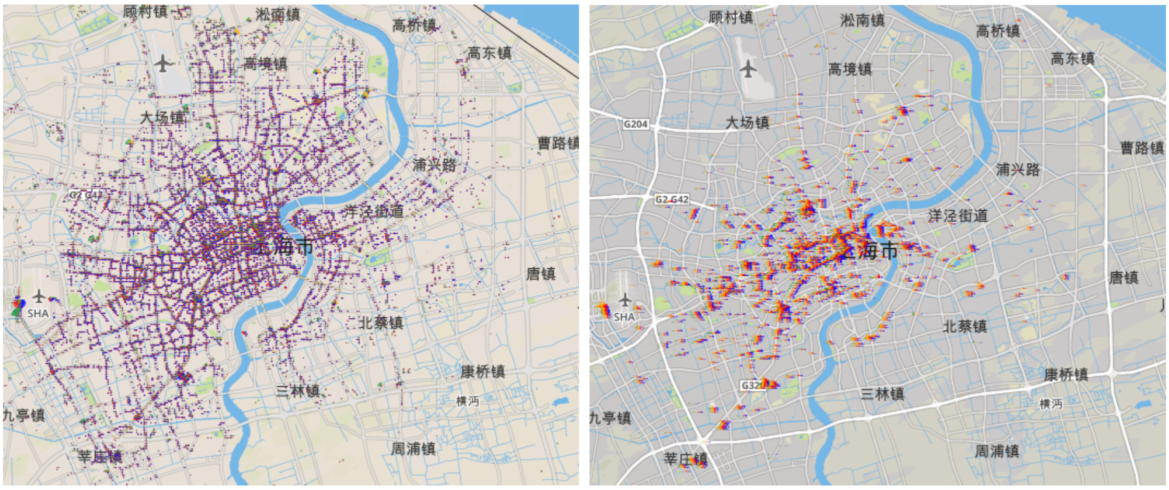
(a) Figure 4.10

(b) Figure 4.16

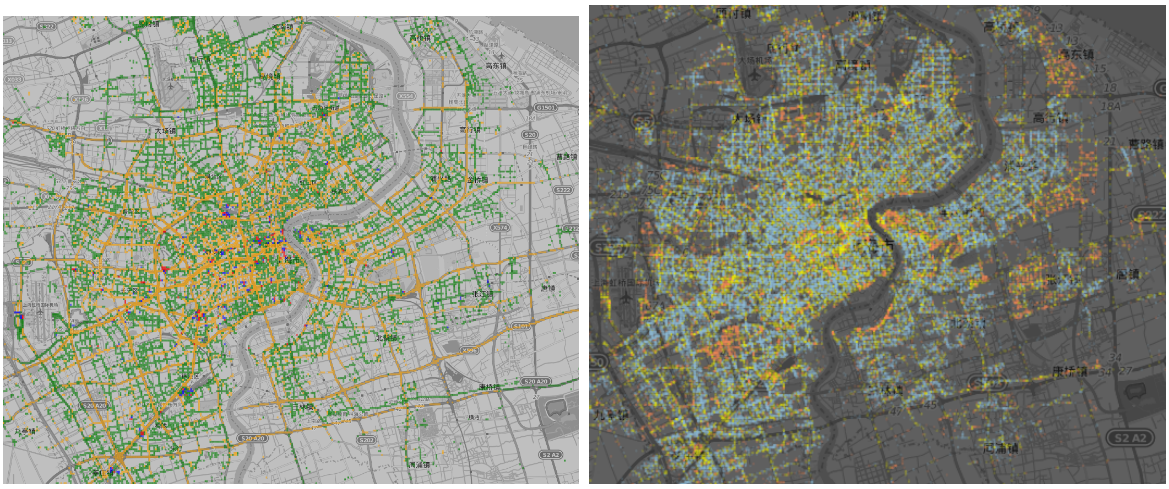
Figure 4.20: The temporal changes using (a) salience-based method and (b) choropleth mapping.

4.5 Comparison and Synthesis of the Visualization Methods

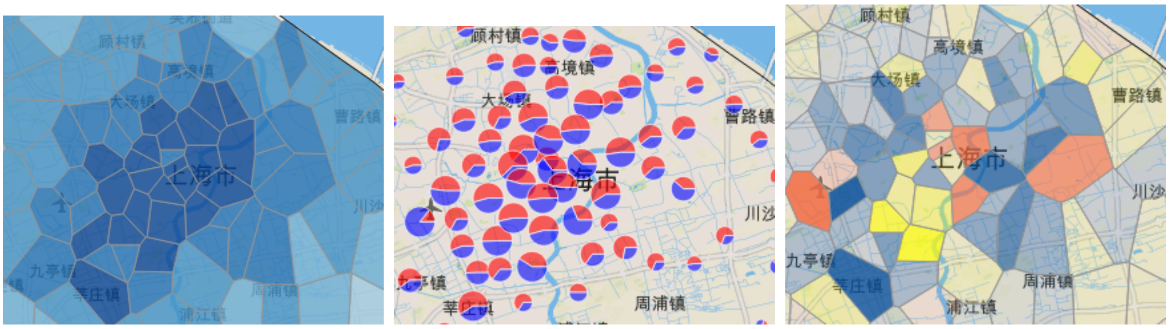
The analysis results in Section 4.4 show that the visualization methods by means of scientific visualization and thematic mapping reveal distinctive spatiotemporal patterns at multiple scales from different perspectives. From both data and visual abstraction point of view, the results in previous sections follow an order of increasing abstraction and construct a consistent example for comprehensive visual exploration of the event data. To explicitly compare the visualization results, we select the representative visualization results of “drop-off” at a same region and sructure them in Figure 4.21 by the sequence of the visualization techniques proposed in the previous sections.



(a) Figure 4.2 (b) Figure 4.5(d)



(c) Figure 4.6 (d) Figure 4.9



(e) Figure 4.13(d) (f) Figure 4.14(a) (g) Figure 4.16

Figure 4.21: The previous process and visualization results from high to low level of detail.

Figure 4.21(a) shows the multivariate visualization resulting from Section 4.2.1. Utilizing the pie radar glyph, this method shows the spatial distributions of the four types of (O, N, P, D) events simultaneously. The pre-attentive visual variables, i.e. color and size, are mixed to allow an immediate perception and detection of significant big pie radar sectors from the compact visualization. The spatial patterns of drop-off events (in blue) can also be observed with especially dense locations in city center as well as very dense locations in some areas. By zooming into interesting places, users can examine their accurate quantities. In addition, by using temporal screenshots in Figure 4.3, spatiotemporal patterns of the multivariate at different time slots can be inspected.

To further explore and analyze the temporal patterns, Section 4.2.2 applied a time graph (see Figure 4.21(b)) to show the univariate event quantities at an hourly time interval. From the visualization result, the user can effortlessly get a general idea about the spatiotemporal distribution of “drop-off” event, which are dense around the city center with a few exceptions far from the city center. With regards to the temporal patterns, drop-off events are especially active from 06:00 to 18:00. Similar to the multivariate visualization based on pie radar glyph, a hotspot or outlier can be examined in detail by interactive operations.

Instead of representing the co-located multivariate events simultaneously, Section 4.2.3 simplified the visual representation utilizing the salience-based method to just show the most significant feature at a specific location (see Figure 4.21(c)). Several significant “drop-off” hotspots scattering around the city center can be detected. This method benefits in further steps from a variety of image processing methods. For instance, two images resulting from this method at different time slots can be arithmetically added to detect temporal changes. Temporal changes of “drop-off” events and the relationships to “pick-up” events can be detected and are used to infer different land use types (see Figure 4.21(d)).

Rather than only inspecting the spatial pattern at the pixel level, Section 4.3.1 and Section 4.3.2 used sophisticated clustering methods to group similar objects, partition the space into proper spatial compartments, and calculate the aggregates that can reflect the underlying data distributions. Based on the aggregates, choropleth maps (see Figure 4.21(e)) in Section 4.3.3 are designed to allow the users to interpret the spatial patterns of individual types of events at a coarser level of detail. The general distributions of “drop-off” events reveals a very compact dense patterns in the city center and the density gradually decreases to the suburban areas.

Along with the generalized information, the visual complexity is largely reduced. Furthermore, in Section 4.3.4, proportional symbol mapping (see Figure 4.21(f)) and choropleth mapping techniques (see Figure 4.21(g)) are applied to illustrate the temporal changes.

To systematically compare these methods in detail, we summarize several of their aspects ranging from data processing to the perception of the visualization results in Table 4.1.

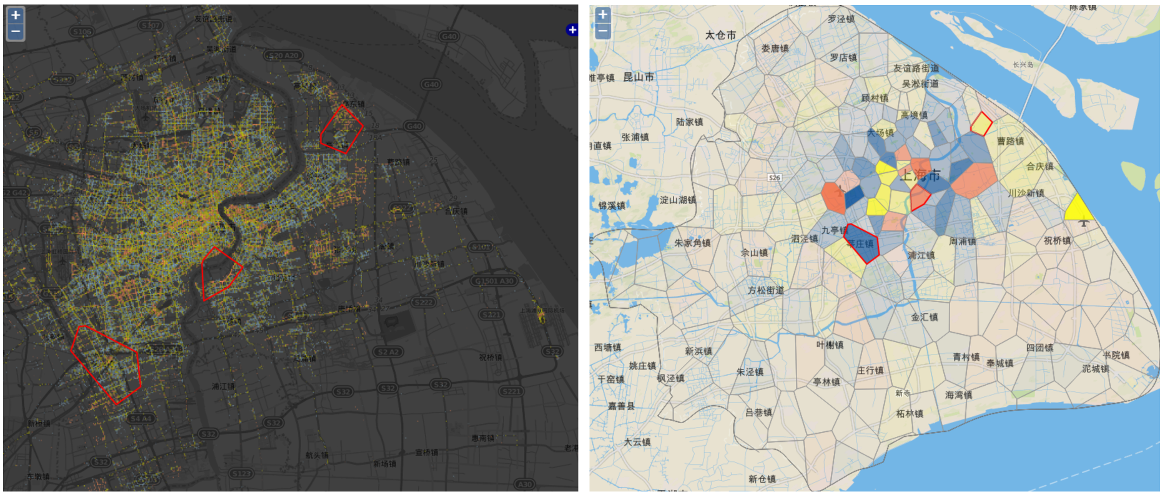
	Scientific Visualization			Thematic Mapping	
	Pie radar glyph	Time graph	Salient feature	Proportional symbol mapping	Choropleth mapping
Filtering	✓	✓	✓		
Ordering	✓	✓			
Clustering				✓	✓
Aggregation				✓	✓
Classification				✓	✓
Symbol overlaps	✓	✓			
Visual variables	color, size	color, size	color	color, size	color
Legibility	low	low	medium	high	high
Cognitive efforts	high	high	medium	low	low

Table 4.1: Comparison of the aspects of different visualization methods.

From Table 4.1, we can see that with the increasing abstraction the characteristics of the methods on data processing, clustering and classification, design principles of visual variables tend to diverge. At the low abstraction level, with the pie radar glyph, there is no need to aggregate the data and the four types of events and their quantities can be simultaneously visualized. The overall spatiotemporal distribution and the hotspots or outliers can be perceived immediately with the compact visual representations. However, at such a fine resolution, individual symbols can be hardly distinguished and understood because of the symbol overlapping and the complex meaning of the multivariate symbol. The users need to inspect very carefully by zooming into the interesting areas. On the contrary, at the high abstraction level, e.g. the choropleth maps, the legibility of each symbol is very good due to the data aggregation but it is impossible to inspect the detailed data items.

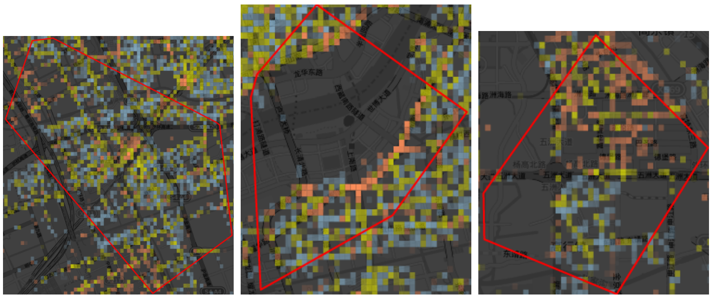
The comparison results of this example suggest that for a comprehensive visual exploration of massive point-based events, it is necessary to synthesize a series of effective visual representations, ideally, a series of representations of continuous abstraction levels. Furthermore, with more advanced interactive techniques, e.g. filtering, linking and brushing, the different visualization results may validate each other and their synergetic effects can also be explored.

As an demonstration, we link the visualization results of Figure 4.9 (see Figure 4.22(a)) and Figure 4.16 (see Figure 4.22(b)) for the purpose of comparison and synthesis. Highlighting and linking techniques are applied for interactive selection of interesting spatial



(a) Figure 4.9

(b) Figure 4.16



(c) Three polygon areas that are highlighted in (a) and (b) and their zoomed-in images

Figure 4.22: The comparison of the temporal change of the salient variable between (a) the salience-based visualization method and (b) the choropleth mapping method.

divisions in the choropleth map and investigation in detail the specific types of events inside them on the left image. Figure 4.22 show the three zoomed in homologe polygons.

As explained in Section 4.2.3, Figure 4.22(a) is the temporal change map of (P, D) at 7-8h and 18-19h resulting from the salience-based method and the algebra addition operation. Quite a few regions in certain specific colors can be distinguished from the compact displays of the image, for instance, the regions in orange. However, when examined closely, each uni-color region is a mixture of different types of temporal changes. Then it is difficult for human to associate the region with certain specific functions. On the contrary, in Figure 4.22(b) each region is assigned with a deterministic function that can be easily understandable.

On the other hand, scientific visualization results can be used to validate and compensate the thematic mapping result. Take the three highlighted polygons as examples. The polygon in the middle in Figure 4.22(a) is in orange and corresponds to the Shanghai Expo area. The counterpart in Figure 4.22(b) consists of mostly the orange pixels as well, which means the aggregate reflects the underlying data distribution quite well. However, overgeneralization phenomena might appear in the aggregation procedure. For instance, the left polygon in Figure 4.22(b) is in blue. However, when examining the corresponding area in Figure 4.22(a), it includes a small shiny yellow area which is overgeneralized by the abstraction process. Furthermore, the aggregation may also lead to pitfalls. For instance, the right polygon in yellow in Figure 4.22(b) indicates that there is no temporal change of the salient variable in the region. While examining the counterpart in Figure 4.22(a), we find that it consists of pixels both in blue and red. The contradictory patterns in the two images show that the scientific visualization result is a necessity to validate and compensate the thematic mapping result.

4.6 Summary

In this chapter, we propose three scientific visualization techniques for visual exploration of multivariate events and adopt two thematic mapping techniques for effective communication of event patterns. More specifically, a pie radar glyph, a time graph, a salience-based method, proportional symbol and choropleth mapping techniques are applied in combination with spatial data mining and image processing techniques. The proposed methods are tested using the (O, N, P, D) events derived from the real-world floating car data. The systematic visual analysis results demonstrate that our proposed methods can reveal the spatial and temporal patterns of the underlying data at multiple levels of details from different perspectives. We also suggest that visual analysis of event data integrating scientific visualization and thematic mapping methods can achieve complementary effects that support a deeper understanding of event patterns.

5 Visual Analysis of Trajectory-based Events

In Chapter 4, we have analyzed the spatial distribution of point-based events using scientific visualization and thematic mapping techniques. This chapter focuses on the analysis of trajectory-based events by synthesizing the scientific visualization and thematic mapping techniques. More specifically, we investigate two types of trajectory-based events based on different abstraction levels of trajectories. The first one is the generalized trajectory-based events, namely the origin-to-destination events. The second one is the non-generalized non-occupancy trajectory-based events.

Section 5.1 reviews the related work of the methods and applications for analysis of trajectories. In Section 5.2, we introduce a general visual analytic framework for analyzing trajectory-based events. Consequently, this framework is applied to the two above-mentioned concrete types of trajectory-based events in the following two sections. Section 5.3 is dedicated to the visual exploration and communication of origin-to-destination event patterns. To explore the overall individual events, we propose and incorporate interactive clustering techniques, parallel coordinates, and gradient line rendering techniques in the framework. For the communication of summarized origin-to-destination event patterns, flow maps are designed building on Voronoi spatial divisions. In Section 5.4, we emphasize on visual analysis of non-occupancy events of two income taxi groups, i.e. high- and low-income taxi drivers. For exploration purpose, direct line rendering and space-time cube techniques are investigated showing different dynamic processes of the two groups. At a high abstraction level, dot mapping, time matrix and pie-chart mapping techniques are applied to reveal the distinctive mobility patterns between these two taxi groups. Finally, Section 5.5 summarizes this chapter.

5.1 Related Work

Massive movement data can be explored and analyzed from a trajectory-based view for understanding the spatial interaction mechanism and human mobility patterns in urban areas. There have been extensive researches and applications in understanding the spatial interactions through visual analytics, which is helpful for understanding human mobility patterns and revealing the city structures at an intracity scale [Liu+12; Liu+15a; KLW14] or intercity scale [Kan+13; Liu+14].

At the individual level, a straightforward way to visualize trajectories is to draw a line connecting the adjacent trajectory segments. Recall the individual reconstructed occupied trajectories and their origin-destination counterparts shown in Figure 3.7. However, overplotting or too many line intersections make the users hardly discern any meaningful patterns. To deal with the visual cluttering issues, numerous techniques have been proposed. For instance, the NYC Taxi holiday visualization system[1] uses animation technique to show the dynamic behaviors of the trajectories. Triple perspective visual trajectory analytics (TripVista) [Guo+11] and layered visual analytics approach [Elz+13] adopt interaction techniques to inspect a subset of massive data in a controlled way. Timeline [WY14] and Space-Time cube [And+13d] show movement patterns in a compact way via data transformation techniques. Edge bundling [Hol06; HW09; Zho+13] is another popular technique in information visualization field to reduce edge intersection and visual clutter.

At the aggregated level, spatial trajectories can be grouped to visualize the movement flows. In cartography, flow mapping techniques have been long used to show the movement of objects, such as people in a migration from one location to another. Andrienko and Andrienko [AA08] presented a generic spatial generalization and aggregation approach for visual analysis of trajectory-based movement events by transforming trajectories into aggregate flows between areas. In addition, Andrienko et al. [And+09] introduced an interactive visual clustering approach by combining clustering and classification for visual analysis of large collections of trajectories at individual and aggregated levels. Scheepens and his colleagues focused on using density maps to explore the multivariate trajectories [Sch+11; Sch+12; SWW14].

Thanks to the advancement of computational and graphical techniques, multivariate trajectory-based events can be explored in an synthesized environment. Buziek [Buz01] designed 2-D and 3-D multimedia interactive environments using a variety of visual and acoustic variables to analyze traffic congestion related information, for instance, by using vertical bands in perspective views. Tominski et al. [Tom+12] used a hybrid 2D/3D display to visualize trajectories as stacked 3D trajectory bands along which attribute values are encoded by color. Elzen and Wijk [EW14] enabled users to gain insights through selected interests (manually or automatically), and producing high-level, infographic overviews simultaneously. For mapping origin(O)-destination(D) flows, Guo et al. [Guo07; Guo09; Guo+12; GZ14] developed and implemented new approaches and systems like VIS-STAMP for flow mapping and multivariate visualization of large spatial interaction data. Wood, Dykes, and Slingsby [WDS10] proposed a technique, by mapping OD vectors as cells while preserving the spatial layout of all O and D locations, for the visual exploration of origins and destinations arranged in geographic space. Boyandin et al. [Boy+11] presented Flowstrates technique to allows the users to perform spatial visual queries, focusing on different regions of interest for the origins and destinations, and to analyze the changes over time.

---

[1] http://taxi.imagework.com/

5.2 A Visual Analytics Framework for Trajectory-based Events

In this section, we propose a visual analytics framework for exploration and communication of trajectory-based event patterns. As illustrated in Figure 5.1, this framework consists of three components ranging from visual querying of the movement database, interactive clustering and aggregation, to visual representations. The interactions in the framework are annotated with the icon . We implemented the framework in a web-based interactive environment incorporating database querying techniques, spatial clustering and aggregation methods, scientific visualization, and thematic mapping techniques.

(1) *Movement Database Visual Querying*. This component allows visual query of the movement database for interesting trajectory subset. Since the income is an essential factor for taxi drivers and might be of interest for the analysis, we design a histogram view to show their average incomes as well as allow the interactive selection of data at certain income levels. Besides, a time graph view is designed for the selection of data in specific time windows. Through visual queries, the amount of data is largely reduced and only the interesting sub-dataset is selected.

(2) *Interactive Visual Clustering and Statistical Calculation and Aggregation.* In this component, there are three aspects that process the queried trajectories.

Firstly, according to the trajectory-based event categorization in Section 3.4.1, we can derive a variety of types of events. In this study, we focus on "origin-to-destination" and "non-generalized non-occupancy" events. Secondly, the two types of events are dynamically clustered through the adjustment of the clustering parameters, which optimizes and steers the complicated processes. Thirdly, we can derive the statistics about the two types of events as well as their spatiotemporal aggregates.

(3) *Visual Representations*. The resulting clusters and aggregates are displayed in a variety of visual representations that are helpful for the investigation of the movement data at multiple scales from different perspectives.

In the following part of this section, we will introduce the details of each component.

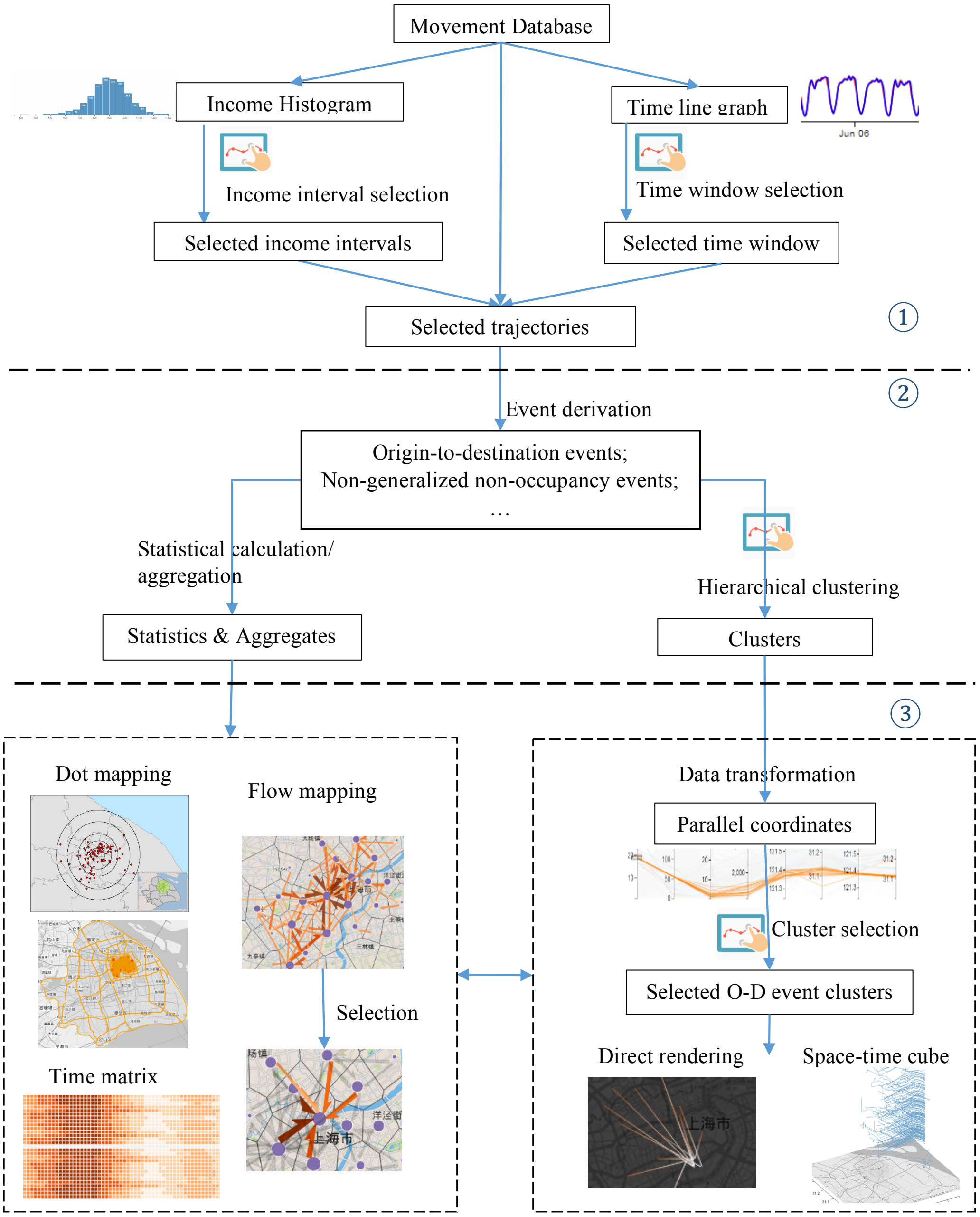


Figure 5.1: The framework for visual analysis of movement trajectory-based events.

5.2.1 Visual Querying of Movement Database

For the sake of the computational efficiency, a reasonable way to inspect massive data is to retrieve only relevant interesting data partitions from the database. In our framework, we retrieve relevant data sets according to taxi drivers income in specific time intervals by interactively brushing an income histogram and a time line graph. As illustrated in Figure 5.1, the first component for querying movement database has the following four parts.

- *Movement database*. The database contains raw GPS data, and derived occupied and non-occupied trajectories among others.
- *A histogram* of taxi drivers' average income derived from the movement database for querying taxis at a certain income level (See Section 3.4 for the derivation of incomes). In our experiment, the daily income of the 2007 taxis in 47 days are classified into 20 groups. As shown in Figure 5.2, the daily income follows a normal distribution, which indicates the income discrepancy among the taxis. Interactive selection of the histogram bars allows users to freely access objects at certain income levels for further inspection, for instance, to inspect behavior differences between high- and low-income taxi driver groups.

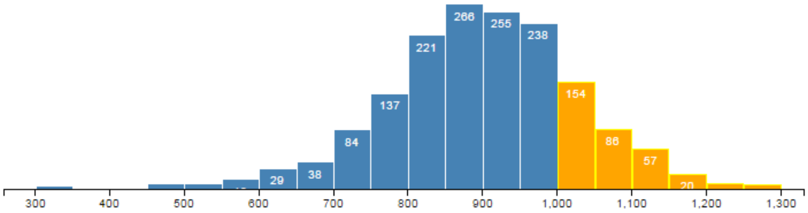


Figure 5.2: The histogram of taxi drivers' income. High income bars are selected.

- *A time line graph* for querying data in a time window from the database. The temporal view with line charts shows the temporal variation of certain attributes. A focus + context technique is applied to show the detailed and overview information. Users can select a time window from the context view and inspect the details in the focus view. Figure 5.3 illustrates the time line view. The statistics in the time line graph is about the hourly numbers of pick-up and drop-off events of occupied trajectories.

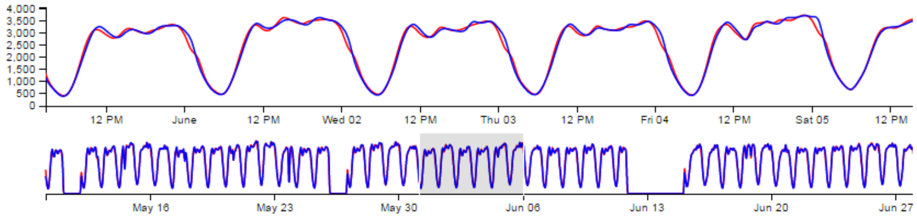


Figure 5.3: The time line view for brushing specific time window.

- *Selected trajectories.* By brushing the income histogram and time line graph, a subset of trajectories are selected according to the range of incomes and time periods for future analysis.

5.2.2 Interactive Visual Clustering and Aggregation

The second component of our framework (shown in Figure 5.1) aims to derive and analyze interesting types of trajectory-based events. According to the categorization discussed in Section 3.4, we can derive a variety of trajectory-based events from the selected trajectories using spatial, temporal and thematic constraints. In this chapter, we focus on two specific types of events: “origin-to-destination” and “non-generalized non-occupancy” events. Two parallel streams are designed to analyze the two derived events at low- and high-abstract levels using interactive clustering and aggregation techniques respectively. To analyze the trajectory-based events at a detail level, the interactive visual clustering method is designed to allow users to select clustering features and adjust clustering parameters. To understand the trajectories at a collective level, we calculate relevant statistics and aggregate the events to coarser spatial and temporal partitions.

Interactive clustering via feature selection and parameter setting

Clustering is a generic data mining approach for exploring massive data. For multivariate data, multivariate clustering techniques can be used for the identification of potentially interesting multidimensional subspaces from a high-dimensional data space. We stress that meaningful clustering results, requiring appropriate clustering features and optimal parameters as inputs for a clustering algorithm, are hardly possible to achieve with fixed features and parameters.

Take the derived trajectory-based events as an example. The events have multiple attributes including timestamps and locations of the origins and destinations, distance, and duration. Analysts sometimes might be interested in inspecting event patterns of trajectories starting from or ending to the same locations, which requires the clustering algorithm to group them according to the spatial locations of the same origins or destinations. Likewise, analysts may be interested in events of certain semantic properties, e.g. distance, then it is necessary to cluster the events based on the distance attribute. Furthermore, appropriate parameter setting is a key factor in the clustering method. Different clustering parameters will lead to diverse results and therefore have strong influences on further analysis. For instance, trajectory-based events of same origins may show distinctive event patterns with different distance parameters. Thus, for exploration purpose, it is crucial to allow interactive selection of clustering variables/features and adjustment of parameters. In our study, we design an interactive clustering interface to serve this purpose. The agglomerative hierarchical clustering method is applied for the multivariate clustering of the trajectory-based events.

Statistics calculation and data aggregation

In parallel with the visual clustering process, we apply data statistics and aggregation techniques to summarize the characteristics of the derived events. Unlike the visual clustering of individual trajectory-based events, data statistics and aggregation serve as effective ways to reveal the general tendency of the dataset and the overall patterns of the events.

To reveal different event characteristics, a number of statistic values can be calculated, for instance, the number of the events, the geometric center of the trajectory-based events, the distribution of events in specific time intervals. Furthermore, besides the basic statistics, we aggregate the trajectory-based events into irregular spatial partitions which can reflect the underlying data distribution quite well. Here we adopt similar techniques used in Section 4.3.1 for spatial aggregation, which includes steps of spatial clustering, Voronoi tessellation and spatial aggregation. The statistical and aggregated values are then visually communicated by using a variety of mapping techniques, e.g. direct rendering and flow mapping, as described in the next section.

5.2.3 Visual Representations

The third component of our framework (in Figure 5.1) focuses on visually representing the event clusters and the summarized and aggregated event data derived from previous steps by means of scientific visualization and thematic mapping techniques. More specifically, we apply the following scientific visualization methods: parallel coordinates for visual exploration of trajectory clusters, gradient line rendering for inspection of the origin-to-destination event clusters, and direct line rendering and space-time cube techniques for the investigation of the dynamic process of the non-occupancy events. For visual communication of spatial interaction patterns of origin-to-destination flows, flow maps of the aggregated origin-to-destination events are designed; while for communicating the spatiotemporal non-occupancy event patterns, dot maps, time matrix graphs and pie-chart maps are proposed.

Visual representation using scientific visualization methods

For the exploration of individual trajectory-based events, parallel coordinates are proposed and well designed to show the resulting clusters from the interactive clustering process. Users can further select interesting clusters by inspecting the patterns in the parallel coordinates. Depending on the specific event types, the individual events in the selected clusters will be shown on a map view, either using gradient rendering technique for origin-to-destination events or direct rendering and space-time cube for non-occupancy events.

- *Parallel coordinates* for visualizing and analyzing the high-dimensional geometry of the precomputed clusters, and as an interactive medium for interactive selection of interesting clusters.

  Parallel coordinates [Ins97] have been widely applied to visualize and explore high-dimensional and multivariate data. The compact parallel lines can easily reveal natural clusters of data and is helpful for discerning cluster patterns within the data. As Heinrich and Weiskopf [HW12] summarized in their survey that parallel coordinates can be used to address a set of high-level tasks, e.g. clustering, classification, regression, summarization, dependency modeling, and change and deviation detection.

  In our work, the parallel coordinates approach is designed to primarily fulfill two tasks: to visualize the precomputed clusters and their characteristics; and to serve as an intermediate tool for visually detecting and selecting potentially interesting clusters, which is beneficial for future analysis.

- *Gradient line rendering* for showing the spatial and directional distribution of origin-to-destination clusters.

  As the parallel coordinates show the high-dimensional origin-to-destination events in transformed way, we use a 2-D map to show the spatial distribution of the events in the geographical space. In our work, this 2-D map is linked with the parallel coordinates. Only the visually salient clusters in the parallel coordinates selected by the analysts will be rendered on the map for further investigation. Through this linking technique, the number of individual lines is significantly reduced, resulting in decreased line intersections.

  Inherently, a generalized origin-to-destination trajectory contains a dynamic process and directional behavior starting from its origin and ending at its destination. To represent the origin-to-destination clusters in a straightforward way, we design a gradient rendering technique by applying gradients to the lines, which can effectively convey the dynamic and directional information and achieve preattentive visual effects.

- *Direct line rendering and space-time cube* for visualizing the dynamic process of the non-generalized trajectory-based events.

The non-generalized trajectories reconstructed from the GPS points approximately represent the continuous process of the moving objects in the geographical space. A straightforward way to present the events on a 2-D map is direct line drawing. Here we apply color and opacity to render lines in an easily understandable way. Furthermore, space-time cube is used to show the lines in a 3-D space. Space-time cube is a commonly used visualization technique to show the dynamics of objects. In its basic appearance these images consist of a cube with on its base a representation of geography (along the x- and y-axis), while the cubes height represents time (z-axis). A typical space-time cube could contain the space-time-paths of for instance individuals or bus routes [Kra88]. In this work, we extract non-occupancy events of two income-level taxi groups related to an exemplary airport and use space-time cube to explore and understand their dynamic processes.

Visual representation using thematic mapping methods

To effectively communicate spatiotemporal patterns of trajectory-based events, thematic mapping techniques are employed representing highly abstract data derived from the origin-to-destination and non-occupancy events. More specifically, we design flow maps to allow an easy perception of origin-to-destination flow patterns. To understand the taxi drivers' behaviors from non-occupancy events, in particular two sub-event types: cruising and waiting, we propose dot mapping, time matrix graph and pie-chart mapping techniques to let users inspect the spatial distribution of the cruising events and the spatiotemporal patterns of long-waiting spots.

- *Flow mapping* to show the origin-to-destination movement patterns.

  Based on the spatial distribution of origins and destinations of the origin-to-destination events, the study area is divided into irregular spatial regions and events are aggregated to the regions. Based on the aggregated data, the origin-destination matrix of each pair of spatial regions can be calculated and used for designing flow maps. Proportional line symbols are designed to convey the flows between different spatial regions, and proportional circle symbols to represent the inner flows of a spatial region.

- *Dot mapping* to represent the spatial distribution of the mean centers of non-occupied cruising events.

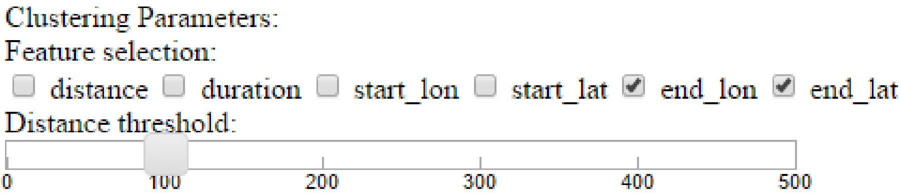
  For individual trajectory-based events, one possible and reasonable way is to get a summarized indicator that reveals their high-level distribution patterns. In this work, we abstract for each driver in the two income-level groups their spatial daily mean centroids of corresponding trajectory events and investigate their general spatial distribution patterns. Dot mapping techniques are applied to show the daily mean centroid patterns with each dot representing a centroid of a group of individual lines. In addition, standard deviation is derived to represent the spread trends of the daily mean centroids. Ellipse symbols are applied to represent the standard deviations.

- *Time matrix graph* to get a temporal overview of the stationary spots extracted from non-occupancy events. We calculate temporal aggregates and design a time matrix graph to visualize daily and quarterly density of the stationary events extracted from the non-occupancy events.
- *Pie-chart mapping and proportional symbol mapping* to reveal the spatiotemporal distribution of non-occupied long-stationary event clusters and stationary spots in off-peak hours.

  Long-stationary events can be interesting for analysis since they reflect unusual behaviors during the moving process, especially when the events are too long to affect a taxi driver's income. For long-stationary event clusters, we use pie-chart maps to visualize their spatial patterns at four time intervals, namely midnight, morning, afternoon and evening. Furthermore, we derive the parking and traffic congestion places from stationary events and investigate on a proportional symbol map the different spatiotemporal distributions between the two taxi groups.

5.3 Origin-to-Destination Event Visualization

In this section we apply the proposed visual analytics framework to the visualization of origin-to-destination events. Recall that in Section 3.4.1, an occupied trajectory can be generalized to an origin-to-destination event, denoted as a pair $(o{-}d)$, which discards the details of the trajectory and emphasizes on the simplified passenger transfer event starting from a location o and ending at a location d. In this section we focus on the visual analysis of the origin-to-destination event patterns at multiple levels by means of scientific visualization and thematic mapping. In Section 5.3.1, we apply a clustering method to group the events and use a parallel coordinates view to show the clustering results as well as allow the exploration of individual potential interesting clusters. The selected clusters with individual origin-destination lines will be then inspected on a map view by using the gradient line rendering technique. In Section 5.3.2, we design a flow map to communicate the spatial interaction patterns of the origin-to-destination trajectory events at the aggregated level.



Clustering Parameters:

Feature selection:

☐ distance ☐ duration ☐ start_lon ☐ start_lat ☑ end_lon ☑ end_lat

Distance threshold:

Figure 5.4: Interactive clustering interface for feature selection and distance setting.

5.3.1 Exploration by means of Scientific Visualization

In this section, we visually explore and analyze massive and complex $(o{-}d)$ trajectory-based events by employing a hierarchical clustering method, the parallel coordinates and a gradient rendering technique. Correspondingly, we build an interactive clustering interface for feature selection and parameter setting, a parallel coordinate view for cluster visualization and brushing, and a map view for spatial representation of pairs of individual $(o{-}d)$s.

Interactive clustering interface

For a comprehensive exploration of massive and complex $(o{-}d)$ events, it is necessary to first group similar events to different categories and then investigate potentially interesting groups. In Section 5.2.2, we have discussed the importance of integrating analysts knowledge for visual exploration of $(o{-}d)$ events by allowing them to interactively choose clustering features and adjust parameters.

As a widely applied tool for analyzing multivariate data at multiple scales, the agglomerative hierarchical clustering method (refer to Section 4.3.1 for algorithm descriptions) is appropriate for our analytic tasks. In our case, the $(o-d)$ events can be abstracted as multi-dimensional multivariate points. Here, the location of the origin and destination, the duration and distance are used as exemplary attributes. Besides, the algorithm has two parameters, namely a distance function and a linkage criterion. The parameter setting is largely dependent on the applications or tasks. The output is an assignment of events to its cluster.

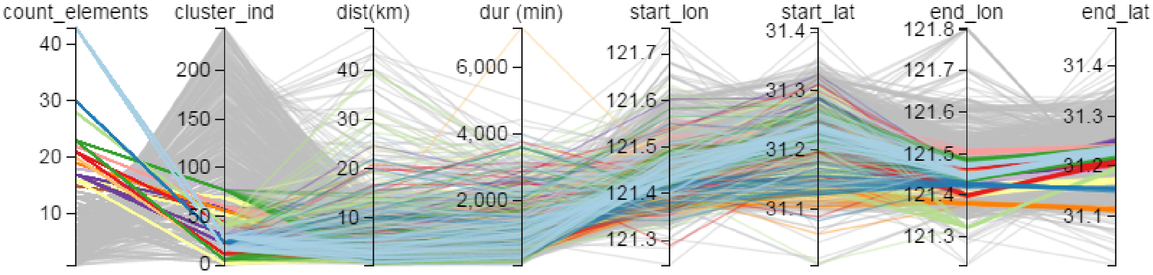
For interactive clustering, we design an interface (shown in Figure 5.4) for feature selection and parameter setting. In this interface, a checklist of related features is used for feature selection and a slider bar is employed for distance parameter setting. As shown in Figure 5.4, the selected features are the longitude and latitude of the destination and the clustering distance is 100m.

A parallel coordinates view

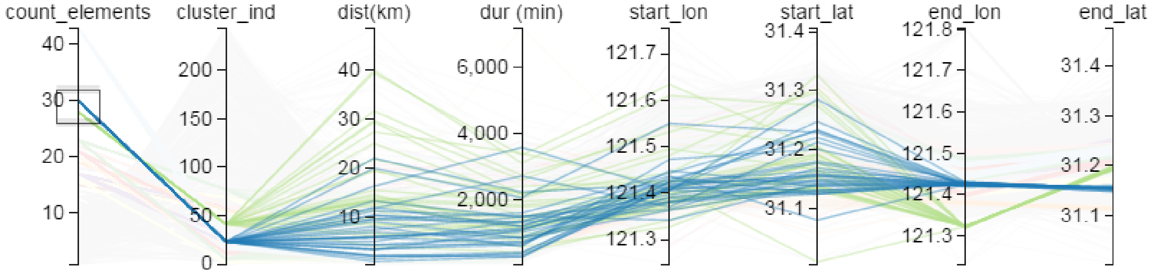
To explore potentially interesting patterns, we would expect that the resulting clusters can be presented in an intuitive way that the natural and significant clusters can be distinguished easily. To achieve this, we adopt the interactive parallel coordinates technique and design the visual representation of the clustering results as follows.

Firstly, besides the aforementioned multiple attributes, we add to the parallel coordinates two more features, i.e. the individual cluster identification numbers and the number of the elements of each cluster. The aim is to explicitly show the cluster results as two axes and allow an easy inspection of individual clusters. Secondly, we design the parallel coordinates in an easily understandable way by ordering the objects using Z-index and semi-transparency to clusters. We order the objects using Z-index according to the number of elements in each cluster so that clusters with large number of elements would be displayed on top of the small numbers. The semi-transparency can on the other hand reveal the hidden objects. Finally, for immediate perception of individual clusters, we assign distinctive colors to clusters with distinct number of cluster elements to reveal the natural clusters. For the clusters with relatively more elements, which normally are potential hotspots, we assign distinctive colors to them. Here, we adopt the qualitative color scheme designed from colorbrewer. For the other clusters with few elements, we use the same color.

Take the trajectory data from 07:00 to 10:00 on 31 May 2015 as an example. The origin-destination pairs can be selected from the time line graph. We cluster the $(o{-}d)$ events by using their destinations. The distance parameter for the clustering method is set to 100m, which results in about 250 clusters. These 250 clusters are then represented by the parallel coordinates shown in Figure 5.5(a). The first axis represents the number of elements in each cluster. The second axis shows the identity number of each specific cluster. The third and fourth axes represent the distance and duration of the $(o{-}d)$ events. The last four axes show the locations of origin and destination.



(a) Clusters of events based on "destination" from 07:00 to 10:00 on 31 May 2010.



(b) Selection of interesting individual clusters.

Figure 5.5: The parallel coordinates view.

As shown in Figure 5.5(a) the parallel coordinates reveal the natural data distribution and the clusters in an intuitive manner. For an easy perception of prominent groups and multivariate relationships, large clusters with more than 15 elements of individual $(o - d)$ events are colored according to the chosen categorical color scheme. Since we cluster the events using their destinations, it is natural that the large clusters converge at the last two axes. One more interesting pattern is that larger clusters converge at the middle range of the last two axes, which indicates that large clusters of $(o - d)$ events happen in the middle of the study area.

Looking at the first axis, we can get an overview about the distribution of the number of elements in each large cluster. For instance, the largest clusters have approximately 40 elements, and the second largest clusters around 30 elements. For further exploration and analysis, the users can inspect interesting individual clusters by brushing any of the axis or multiple axes at the same time. For instance, Figure 5.5(b) shows two elements in the first axis are brushed, which correspond to two clusters of about 30 elements. Clearly, the events in blue cluster are of relatively shorter distances than those in the green cluster and also there are differences in terms of the start and end locations. The un-brushed events in the parallel coordinates are shown in light gray as context information. Users can also brush other axes to examine events with specific characteristics. For instance, to investigate small or local scale spatial interactions, the users can brush the distance axis with the range from $0km$ to $3km$.

A map view of $(o{-}d)$ events by gradient line rendering

The parallel coordinates can reveal the general multivariate relationships of the events. Users can also infer spatial distributions of the clusters from the last four axes. For instance, large event clusters of destination converging at the middle of the last two axes means that a lot of $(o{-}d)$ events end in the middle of the study area. However, high cognitive efforts are needed to interpret and understand the spatial relationships without a spatial reference. In this section, we use a map view to explicitly show the spatial distribution and interaction of the interesting $(o - d)$ events. Since it can get easily cluttered with line objects on a map, the map view is linked with the parallel coordinates and only render the selected clusters of events for detailed examination.

To allow an intuitive interpretation of the $(o{-}d)$ lines, we apply the following techniques to render them on the map.

1. Round the coordinates of origins/destinations. To reduce the line intersections and make the rendering results more appealingly, we simplify the spatial locations of the origin or destination to round values.
2. Order the origin-destination lines according to their distances. Lines with long distances are pushed into the background (using a small z-index) so that short line would not be hidden.
3. Assign the lines from origin to destination with gradient colors. Firstly a line is segmented to a certain number of line segments and the intermediate nodes are interpolated. Secondly the line segments from the origin to destination are assigned color from dark to light color values to give the user an intuitive feeling of line direction.

Take the selected clusters in Figure 5.5(b) as an example, Figure 5.6 shows the individual $(o{-}d)$ events on the map view after applying the aforementioned gradient rendering techniques. The detailed spatial interaction can be clearly identified. The two clusters correspond to two transport hubs with the blue cluster of the Shanghai south railway center and the green one of the Hongqiao international airport.

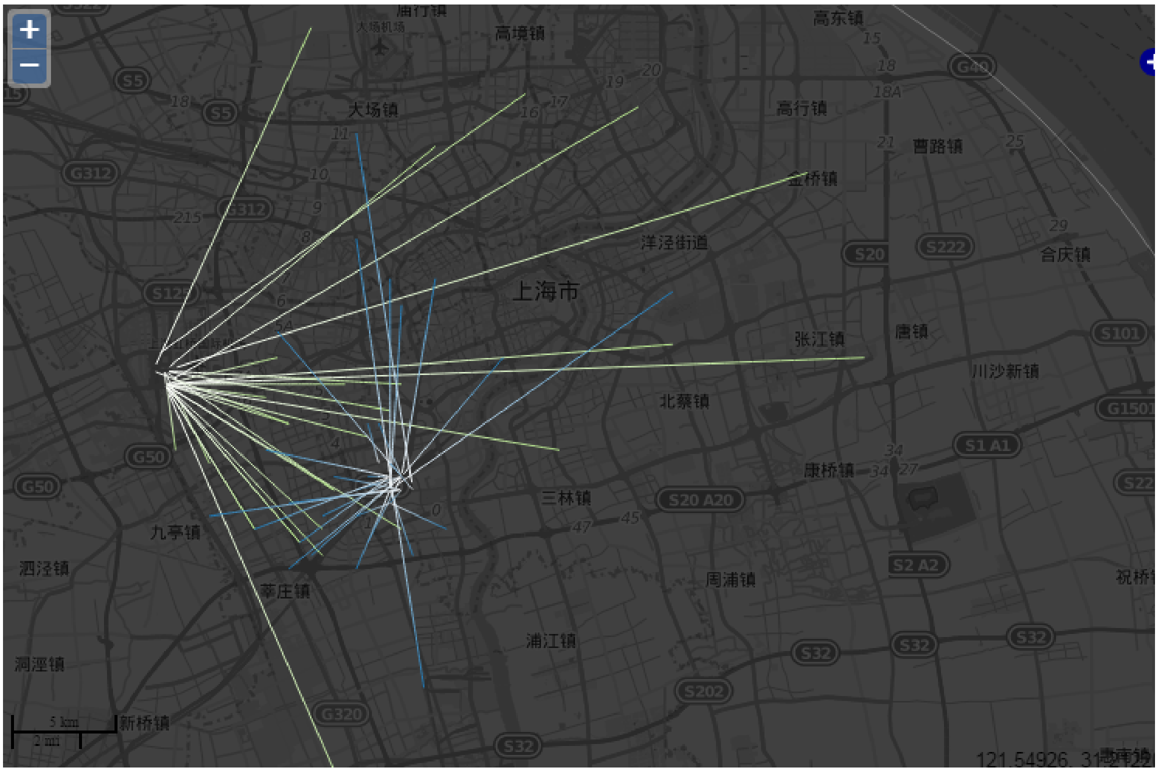
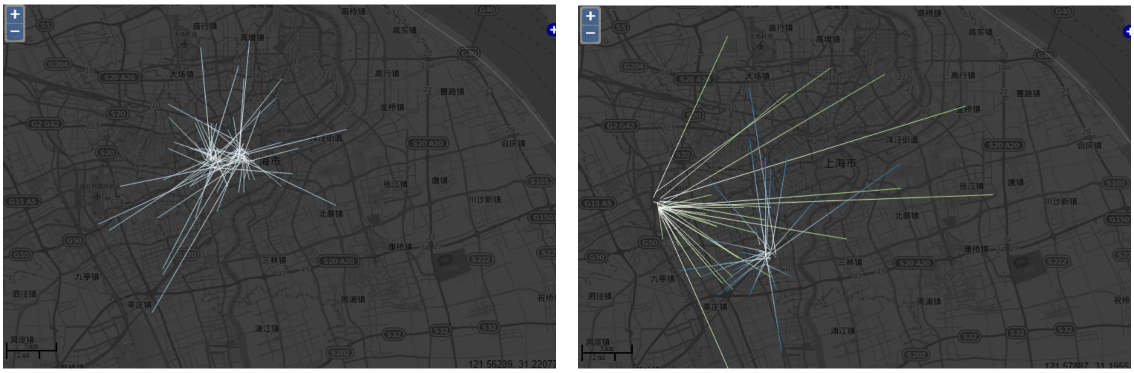


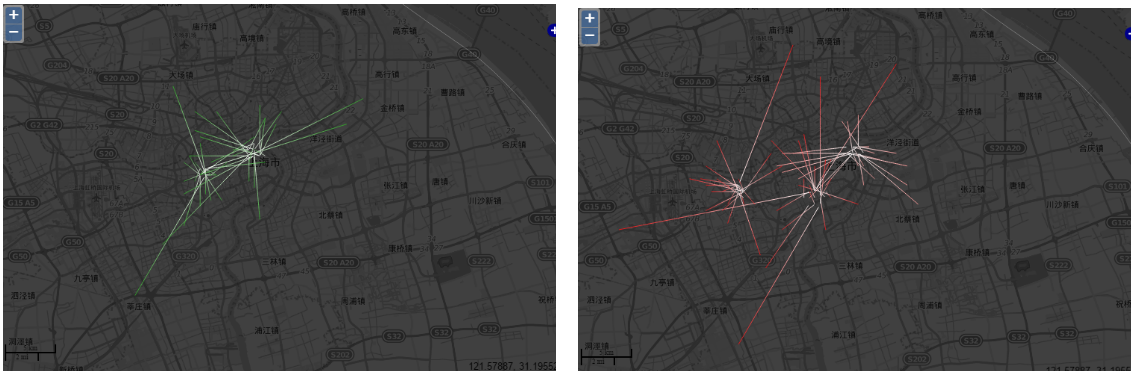
Figure 5.6: The map view of the selected $(o{-}d)$ lines.

So far, we have introduced the parallel coordinates visualization technique and the gradient rendering technique. We use the $(o{-}d)$ events from 7h to 10h on 31 May 2010 to demonstrate the effectiveness of these methods. The events are clustered using "destinations" and the distance parameter is set to 100 meter. The clusters can be visually detected in the parallel coordinates in Figure 5.5(a). We brush large clusters and show the individual events in these clusters in Figure 5.7. The largest two clusters of about 42 individual event elements (Figure 5.7(a)) are with their destinations concentrated in the city center. This phenomena is reasonable since there should be more taxis going to the city center during rush hours. The third and fourth largest clusters of about 30 individual events (Figure 5.7(b)) correspond to Hongqiao airport and Shanghai south railway station. Due to the spatial locations of the two transport hubs, especially Hongqiao airport, there are some long distance $(o{-}d)$ events. The fourth to ninth largest clusters (Figure 5.7(c) and Figure 5.7(d)) are of around 25 to 30 event elements and are primarily with destinations near the city center.

The gradient line rendering results show not only the spatiotemporal distribution of the lines but also the spatial interaction of the events between different areas. Mostly, there are more local interactions. The radiation shape of the $(o{-}d)$ lines are different and difficult to



(a) cluster 1 and 2 (b) cluster 3 and 4

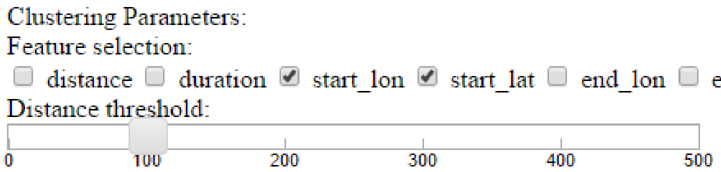


(c) cluster 5 and 6 (d) cluster 7,8 and 9

Figure 5.7: Significant clusters of destinations at 7h-10h on 31 May 2010.

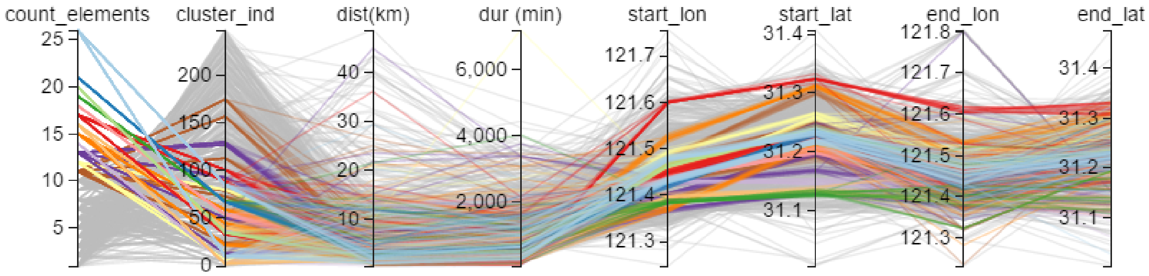
foresee, which relies on the spatial location of the cluster. Likely, users can explore the spatial patterns of other clusters by brushing their correspondences in the parallel coordinates view.

Similarly, we can cluster the same dataset by using "origins". We also set the distance parameter of 100m. The clustering results are shown in Figure 5.8(b). Compared with cluster results using "destination" in Figure 5.5(a), Figure 5.8(b) has relatively less clusters with large number of elements. For instance, the largest cluster in Figure 5.5(a) has about 42 event elements, however, the largest cluster in Figure 5.8(b) only has about 25 elements. This means that during 7h-10h on 31 May 2010, the destinations are more concentrated than the origins.



Clustering Parameters:
Feature selection:
☐ distance ☐ duration ☑ start_lon ☑ start_lat ☐ end_lon ☐ end_lat
Distance threshold:

(a) Clustering features (origin) and parameters (distance).



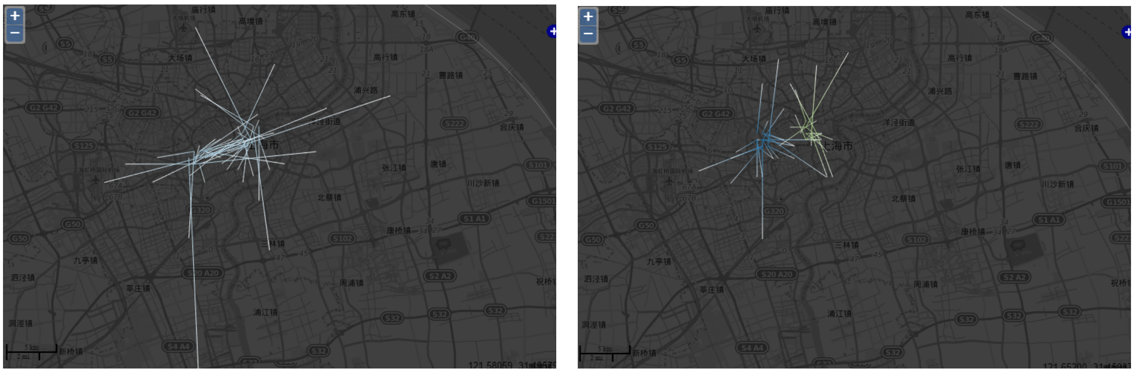
(b) Clusters of events based on "origin" from 07:00 to 10:00 on 31 May 2010.

Figure 5.8: Interactive clustering of origins from 07:00 to 10:00 on 31 May 2010.

Similarly, we brush several large clusters of origin and inspect their spatial distribution in Figure 5.9. The first and second clusters (Figure 5.9(a)) are of about 27 individual events with the destination centers in the city center. However, the two clusters are relatively interconnected in comparison with the two largest clusters of destination in Figure 5.7(a). The third and fourth clusters (Figure 5.9(b)) are of about 20 to 23 individual events. The fourth to ninth clusters (Figure 5.9(c) and Figure 5.9(d)) are of around 17 to 20 elements. We can roughly judge from the geographical location that the green cluster in Figure 5.9(c) origins from the Xinzhuang town and the red cluster on the left in Figure 5.9(d) origins near the Waigaoqiao Free Trade Zone.

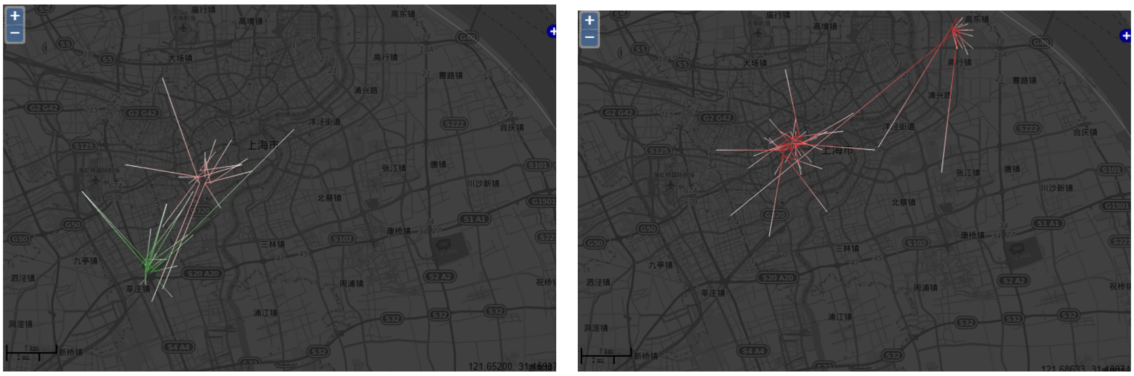
By comparing Figure 5.7 and Figure 5.9, we can observe asymmetric patterns of starting and ending behaviors. There are fewer large origin clusters than destination clusters. In addition, the spatial interactions of the large origin clusters are more local with their near neighboring areas compared with the destination clusters.

By taking the advantage of the interactive functions, the users can also explore different event patterns by imposing a variety of attribute constraints. For instance Figure 5.10 shows the spatial distribution of events with distance less than $3km$ and greater than $30km$ respectively. Several large clusters can be found near the city center and Xinzhuang Town in Figure 5.10(a). While in Figure 5.10(b) one prominent cluster is related to the location of easternmost Shanghai, which corresponds to the Pudong airport. Besides, the two figures show far more event clusters of short distance than clusters of long distance. The phenomenon conforms to the geographic first law and reflects the distance effect, and thus can be easily understood.



(a) cluster 1 and 2

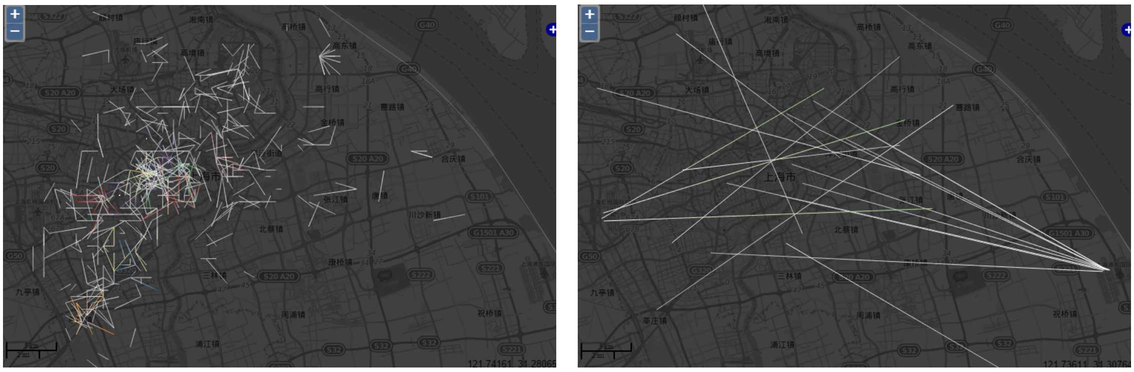
(b) cluster 3 and 4



(c) cluster 5 and 6

(d) cluster 7,8 and 9

Figure 5.9: Large clusters of origins at 7h-10h on 31 May 2010.



(a) Distance < 3km

(b) Distance > 30km

Figure 5.10: Exploration based on different distance values showing the distance decay effects.

5.3.2 Communication via Thematic Mapping

In the previous Section 5.3.1, we have introduced the visual exploration of the $(o{-}d)$ events by means of scientific visualization. Highly interactive techniques, e.g. selection, slider, brushing and linking, incorporated in the hierarchical clustering and parallel coordinates steps allow users to freely query and explore fractions of event dataset at rather detail levels. For instance, the parallel coordinates visualization reveal the multivariate relationships in a compact way by transforming the high-dimensional data into two dimensions. Complementary to the aspatial patterns in the parallel coordinates, the gradient line rendering of significant clusters of $(o{-}d)$ events on a map reveals their spatiotemporal movement patterns.

In spite of fully explorative characteristics of the aforementioned visualizations, effortful cognitive endeavors are required to control the interactive process and interpret the complicated visualization results. For example, users need to select interesting features and set parameters during hierarchical clustering to generate appropriate clusters in the parallel coordinates. For the interpretation of the cluttered lines in the parallel coordinates, users need to understand the meaning of each axis and mentally associate the visual representation to the underlying data distribution and relationship. Besides, the explorative steps are more effective and computationally efficient when relatively small fractions of dataset are applied so that line clutters are reduced and the whole exploration procedure is accelerated.

To relieve the cognition burden and show overall pattern of large amount of $(o{-}d)$ events in an easily understandable manner, an alternative visual representation is possible by data summarization and representation of their movement patterns at a high abstraction level. In our study, we propose a flow mapping technique to communicate the movement patterns of the $(o{-}d)$ events.

Flow mapping is a widely applied technique that can effectively visualize the collective movement of objects from one location to another. The algorithm and design of the flow mapping technique is described as follows.

Firstly, we partition the study area to meaningful sub-regions that can well reflect the $(o{-}d)$ distributions. We adopt the hierarchical clustering techniques used in Section 4.3.1 for the spatial division. The spatial locations of the origins and destinations are used as the clustering features. The centroids of the resulting clusters are applied as generating points for Voronoi tessellations. To generate more evenly sized sub-regions, additional generating points and boundary clipping techniques are applied (refer to Section 4.3.1). Secondly, we calculate the $(o{-}d)$ flow matrix of each pair of the sub-regions. The Voronoi cells resulting from the first step are used to aggregate the events. To generate the aggregates, we take each Voronoi cell as a node and calculate the number of $(o{-}d)$ events that moves from one cell and to another. Then the number of the flows between each pair of nodes is assigned as the edge weight. Note that we also calculate the flow values inside each cell, i.e., the

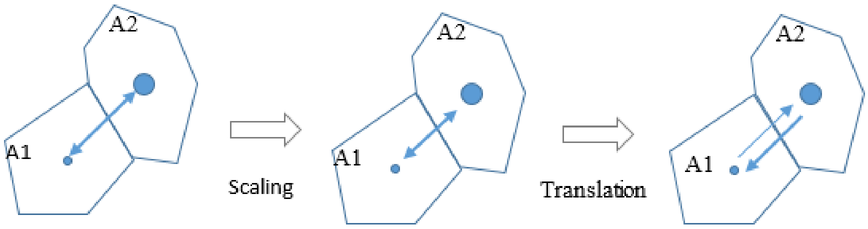
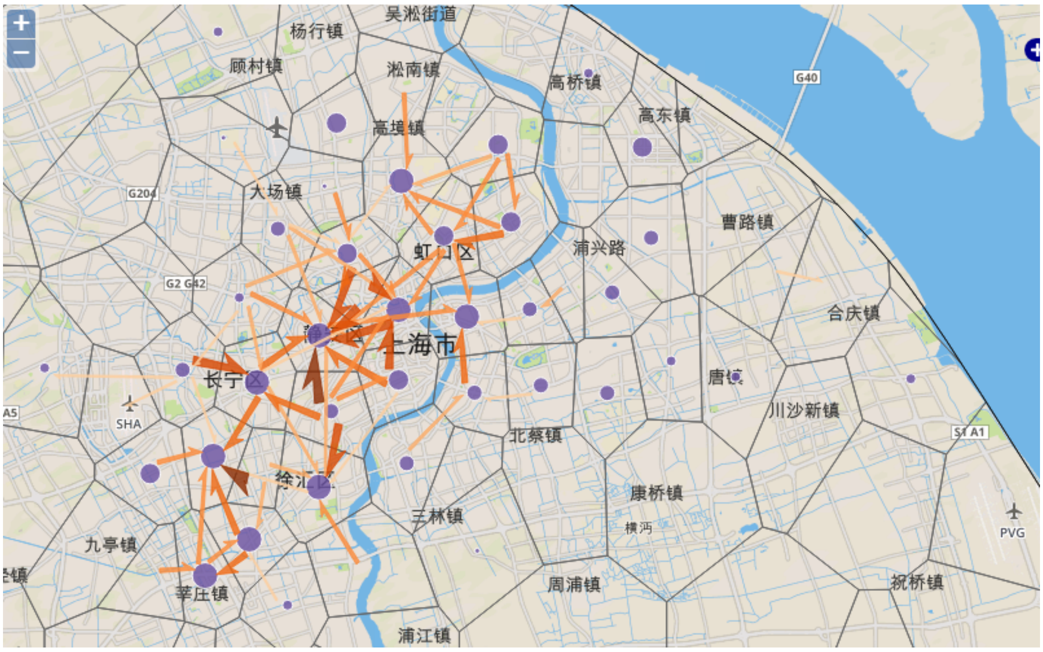


Figure 5.11: The placement of the line symbols. A1 and A2 are two regions.

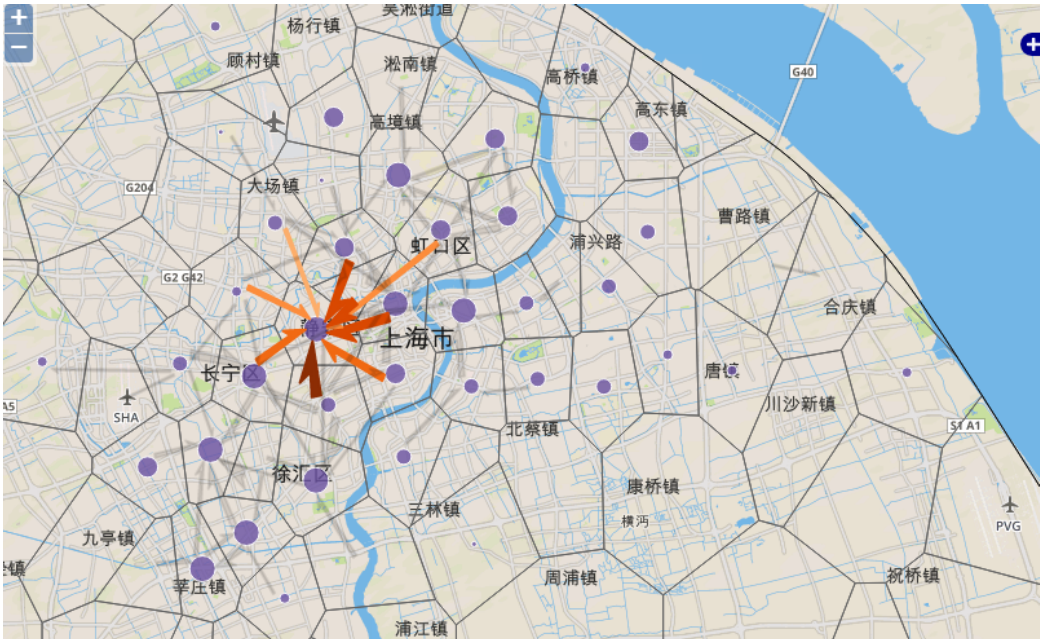
number of trajectories from the cell to itself. Finally, we design a flow map to show the calculated flow matrix. Basically, we use proportional line symbols to represent the flows between pairs of sub-regions (inter-regional flows) and proportional circle symbols for the movements inside each sub-region (intra-regional flows). For the line symbols, two visual variables, i.e., color and line width, are applied to represent the calculated flow aggregates. According to cartographic design principles, the aggregates are grouped to five classes and proportionally mapped to a sequential color scheme and also an array of line sizes. The symbol of half arrow is added to the end of the line to indicate the flow directions. The size of the circle symbols are proportional to the aggregate values calculated for each sub-region.

To reduce line intersections and enhance the legibility of the flow map, we design the flow symbols and their placements carefully. Firstly, we scale each line connecting each pair of the regions using a factor of 0.9 so that the lines do not directly connect the geographic centroids of each region, which will reduce line symbol overlapping effects especially around the centroids. Secondly, we translate the lines along their normal directions so that the line symbols in opposite directions can be separated. The design schema of the placement of the line symbols are shown in Figure 5.11.

Following the design principles, flow maps of $(o-d)$ event data from 7h to 10h on 31May are designed as an illustration to show the high-level spatial interaction patterns. The origin and destination points are as input for the hierarchical clustering method with the distance parameter of 300 meter. Figure 5.12(a) shows the resulted flow map. Furthermore, we also utilize basic interaction techniques to allow the users investigating the flows of specific regions. For instance, when the users are interested in one region, they can select the region which will highlight the line symbols connecting to the cluster centroid and gray other line symbols as context. Figure 5.12(b) shows the highlighted flows related to a selected region.



(a) A flow map at 7 a.m. - 10 a.m. on 31 May 2010.



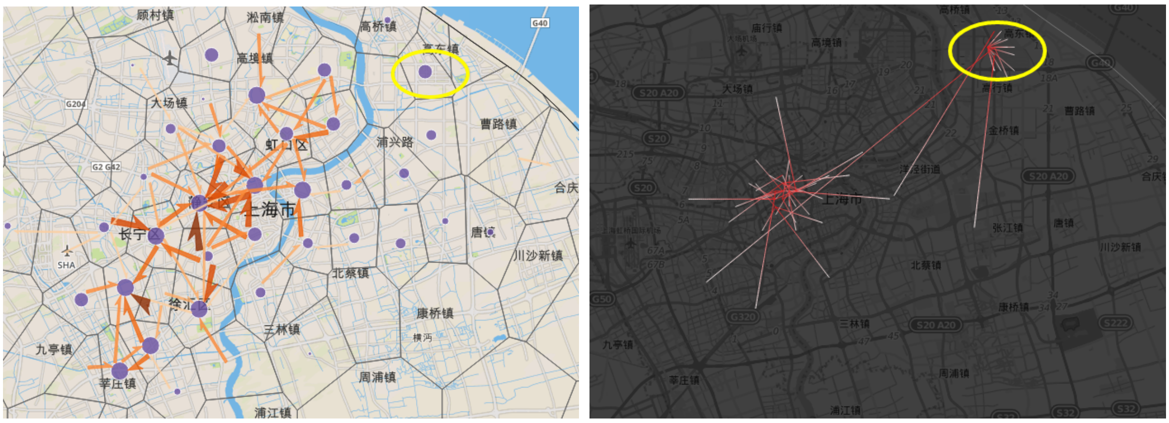
(b) The flow map after selecting a spatial partition.

Figure 5.12: Flow maps of origin-to-destination events.

In Figure 5.12(a), we can see that the flow map clearly reveal the overall flow patterns. Both the volume and the direction information of the flows are legible and can be easily interpreted with little effort. The overall distribution of the inter-regional $(o{-}d)$ flows at this time slot exhibits a northeast-southwest shape and shows that neighboring areas has stronger spatial interactions than distant areas. In addition, through the unevenly sized directional lines, the users can easily discern asymmetrical flow patterns between pairs of regions, indicating that at this time period some regions have more flow-ins than flow-outs, and vice versa. Keeping in mind that the purple circles in the flow map reveal the intra-regional flow volumes in respective regions, we can observe that normally regions of strong inter-regional connections also have strong intra-regional movements. Several peripheral regions are of high intra-regional movements but without obvious inter-regional movements.

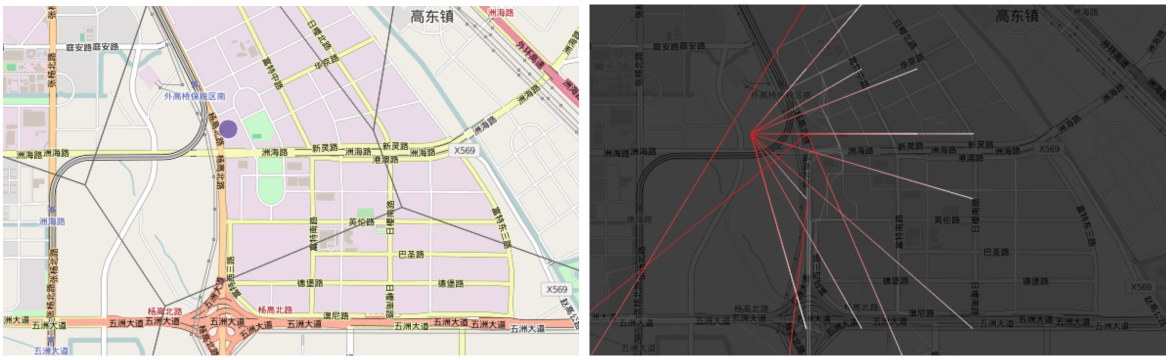
Figure 5.12(b) shows a selected region in the city center with related flows highlighted. The users can notice that the flow directions show the movements are flowing into this region.

We also compare the visualization results from flow mapping and gradient line rendering results. Figure 5.12(a) and Figure 5.9(d) are taken as examples. An area on the top right corner of Shanghai in both figures is framed in orange to illustrate their differences (Figure 5.13(a) and Figure 5.13(b)). Their enlarged counterparts are shown in Figure 5.13(c) and Figure 5.13(d). From the size of the circle symbol in Figure 5.13(c) we can estimate the flow volume in this region. In Figure 5.13(d), instead of the summarized value, we can investigate in detail each individual $(o{-}d)$ events and their destinations and directions inside this area. We can also see three long-distance lines from this region to other remote regions. However, due to the generalization effect in Figure 5.13(c), the inter-regional flows of this region are invisible.



(a) Figure 5.12(a)

(b) Figure 5.9(d)



(c) Enlarged part of Figure 5.12(a)

(d) Enlarged part of Figure 5.9(d)

Figure 5.13: Comparison of a place from Figure 5.12(a) and Figure 5.9(d).

5.4 Non-Occupancy Event Visualization

In this section we apply the proposed visual analytics framework to the visualization of non-occupancy events by means of scientific visualization and thematic mapping. As we have already stated in Section 3.4.2, when a taxi is occupied, the taxi driver usually takes the fastest route based on his/her knowledge to bring passengers to a destination; on the contrary, when the taxi is vacant, the taxi driver has a large freedom to plan his/her routing to minimize his/her waiting time for the next trip, which may result in large income differences for experienced taxi drivers and novices. This motivates us to visually analyze the non-occupancy events with the emphasis of two groups of taxi drivers, i.e. high-income and low-income groups.

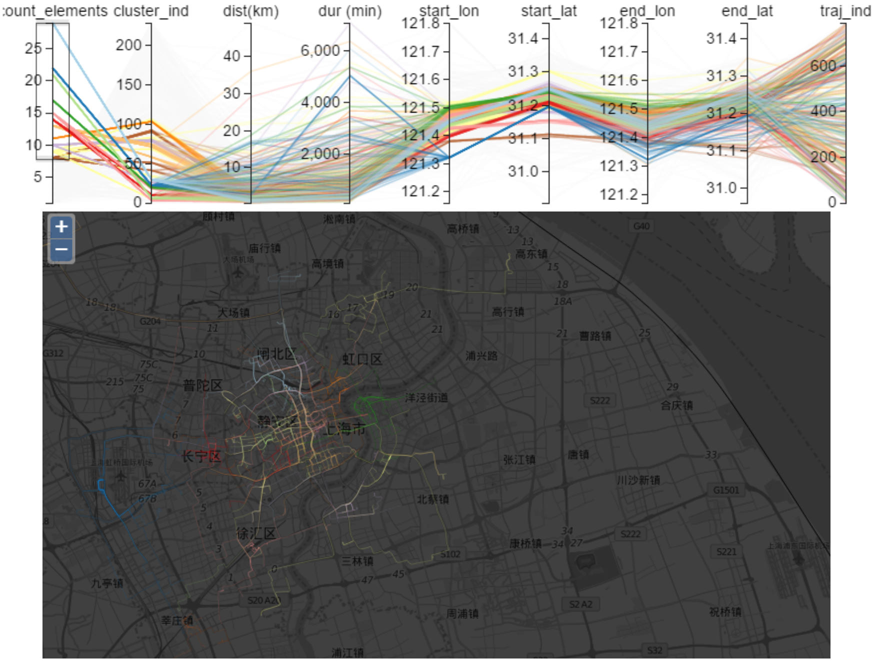
In Section 5.4.1, similar to the procedure in Section 5.3.1, we interactively cluster the retrieved non-occupancy events and visualize the clusters in the parallel coordinates. The users then can select significant clusters in the parallel coordinates and render the individual cluster elements on a 2-D map. Furthermore, a space time cube is applied to show the non-occupancy event in 3-D to reveal the space-time dynamic profiles of the moving taxis. In Section 5.4.2, we focus on the visual communication of the driving behavior patterns between the two specific taxi groups by dot mapping, time matrix and pie-chart mapping.

5.4.1 Exploration by means of Scientific Visualization

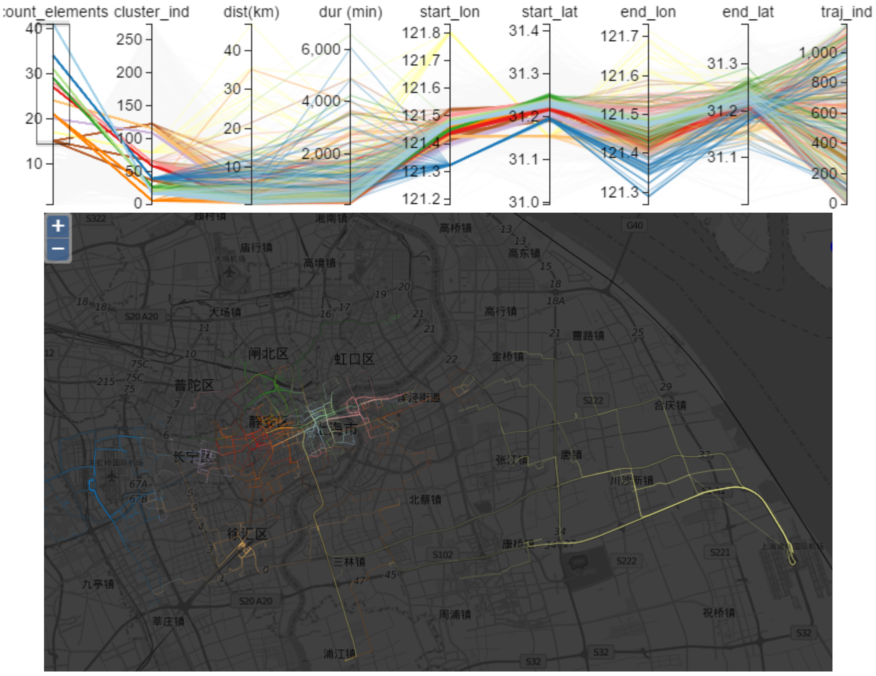
In this subsection, we focus on the visual analysis of the non-generalized non-occupancy events at detailed level. The following principles are analogous to those in the visual exploration of origin-to-destination by means of scientific visualization. *(a)* Using interactive clustering methods, the analysts can cluster a chunk of non-occupied events by choosing clustering features and setting appropriate parameters via the interactive graphic user interface. *(b)* To reveal the underlying data distribution patterns of the non-occupancy events and the multivariate relationships, we adopt the parallel coordinates technique, which allows easy visual detection of interesting clusters. *(c)* Moreover, analysts can interactively inspect interesting clusters by brushing individual clusters and render them accordingly on a map to support the investigation of their spatial distribution.

To demonstrate, we retrieve the non-occupancy events of the bottom and top taxi groups from 06:00 to 12:00 on 31 May 2010. Based on their "starting" points, we cluster these events and show the clustering results in the parallel coordinates in Figure 5.14. Then we select the significant clusters in the respective parallel coordinates, which are rendered on the 2-D map with an opacity of 0.2 and with the same colors of the clusters in the parallel coordinates.

In Figure 5.14, we can see from the parallel coordinates that there are more large clusters of non-occupancy events in the high-income taxi group (Figure 5.14(b)) than those in the low-income group (Figure 5.14(a)). In the map views, the overall spatial distribution of the selected clusters and the driving routes starting from the cluster centers can be easily detected with the help of the different colors and opacity values of individual spatial trajectories. The denser line areas (i.e. rendered with lower values of opacity) are with more frequent non-occupancy events. We also observe that in each cluster there is a distance decay effect of the event frequency from its center to its border. Moreover, in terms of the spatial distribution, the cluster centers of both taxi groups largely overlap. There are also some notable differences. For instance, in the top taxi group there is a significant cluster related to the Pudong airport (on the easternmost), which does not appear in the bottom group.



(a) bottom income group



(b) top income group

Figure 5.14: Non-occupancy event clusters based on "starting location" in the parallel coordinates and significant clusters on the map from 06:00 to 12:00 on 31 May 2010 for two groups.

By comparing cluster centers with the base map, we can also infer their possible semantics. For instance, we can identify the most important transport hubs in the test area, including Shanghai railway station, Shanghai south railway station, Hongqiao airport, and Pudong airport. Figure 5.15 annotates these transport hubs by their names.

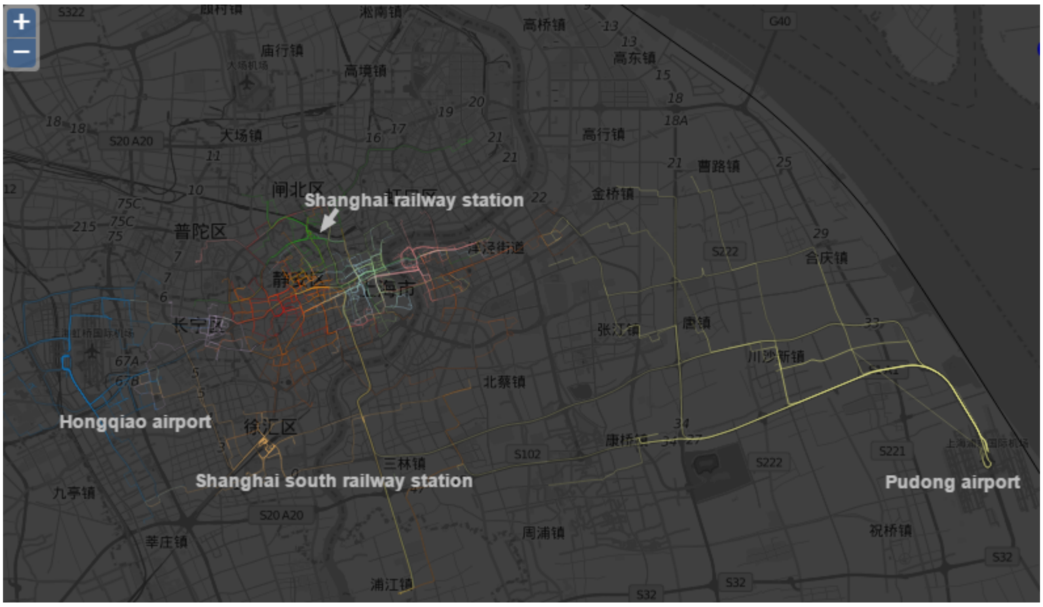
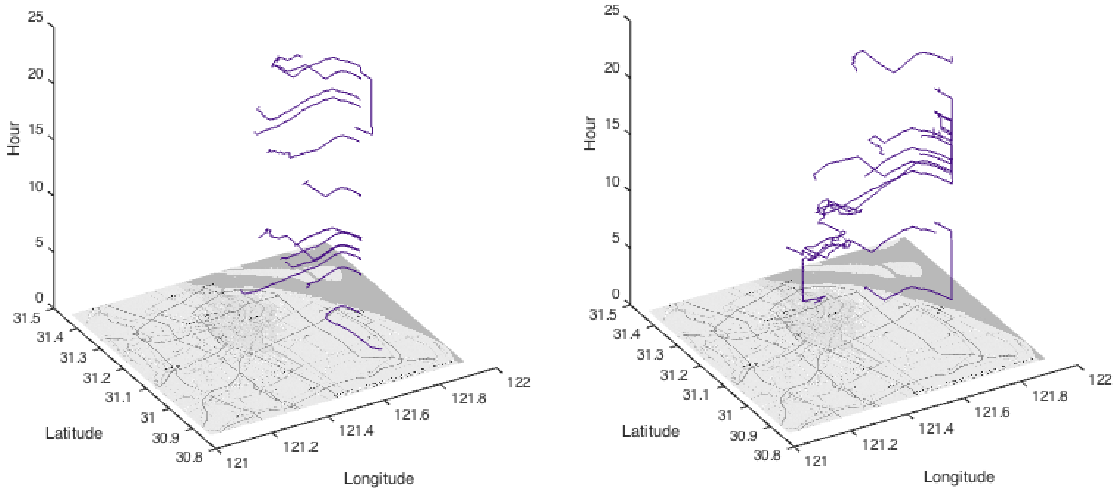


Figure 5.15: Identified transport hubs.

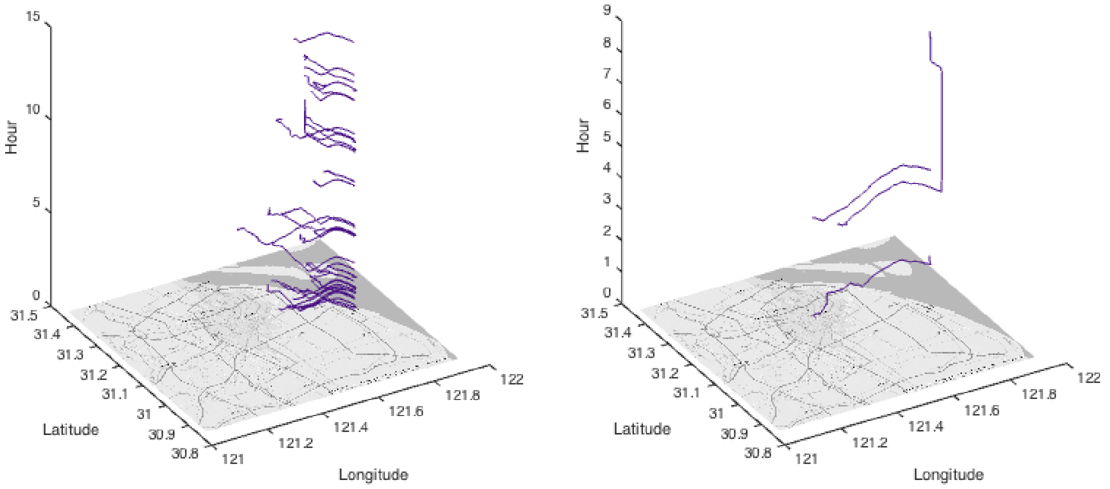
The map view visualization by rendering the individual trajectories is suitable for the investigation of the overall spatial distribution of non-occupancy events. However, one cannot observe the spatial-temporal behavior of individual trajectories. To address this, we use the space-time cube to reveal their dynamic processes.

Since we have already detected different spatial patterns related to Pudong airport for high-income and low-income taxi groups, we continue to use Pudong international airport as a test case. For each taxi group, we extract the non-occupancy trajectory-based events starting from or ending to the airport from the movement database. The dynamics of the non-occupancy events related to the airport on 31 May 2010 are visualized in the space-time cubes shown in Figure 5.16.



(a) bottom income group from Pudong

(b) bottom income group to Pudong



(c) top income group from Pudong

(d) top income group to Pudong

Figure 5.16: The non-occupancy trajectory-based events related to Pudong international airport on 31 May 2010.

From Figure 5.16, we can get an overview of the spatiotemporal profiles of the four types of events related to the airport. Regarding the amount of events, obviously for the top taxi group, there are far more events of “non-occupancy from airport” (Figure 5.16(c)) than those of “non-occupancy to airport” (Figure 5.16(d)). The large number of “non-occupancy from airport” events indicates that most of the high-income taxi drivers directly cruise back to the city center after dropping off passengers in the airport rather than waiting there for the next passengers. While the small amount of “non-occupancy to airport” indicates that only few high-income taxi drivers cruise to the airport without passengers. On the contrary, for the bottom income taxi group, we cannot find such difference in “non-occupancy from airport” events (Figure 5.16(a)) and “non-occupancy to airport” events (Figure 5.16(b)).

Looking at the time dimension, we can also observe the frequency distribution patterns. For instance, in the bottom taxi group, there are more “non-occupancy to Pudong” events (Figure 5.16(b)) happening in the afternoon around 15:00. While in the top income taxi group there are more “non-occupancy from airport” events (Figure 5.16(c)) in the early morning (00:00-05:00).

Furthermore, we can identify from the “non-occupancy to airport” events in both taxi groups (Figure 5.16(b) and Figure 5.16(d)) that there are many vertical lines. These vertical lines reflect that the taxis are stationary at particular locations for rather a long period. Indeed we checked these places on the base map and found that they correspond to airport taxi waiting pools where taxi drivers can rest or wait for picking up their next passengers.

5.4.2 Communication using Thematic Mapping

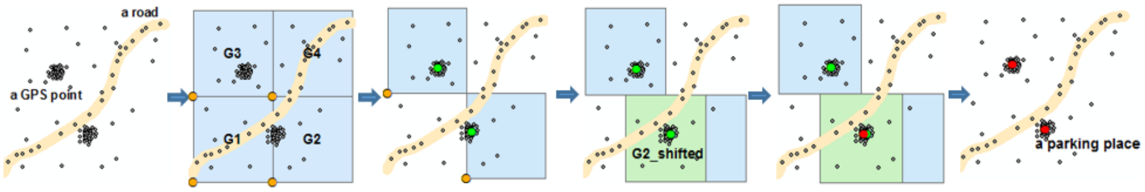
In this section, we emphasize the investigation of the overall spatiotemporal patterns by using thematic mapping techniques for visualizing the statistics and aggregates of the non-occupancy events introduced in Section 5.2.2. We firstly differentiate two sub event types of stationary spots and cruising trips in Section 5.4.2. The differentiation of the two subtypes is stimulated by the vertical lines observed from space-time cube visualization results in Figure 5.16(b) and Figure 5.16(d). In Section 5.4.2, for both taxi groups we investigate the spatial distribution of the cruising events using dot mapping. In Section 5.4.2, we focus on the spatiotemporal patterns of the stationary events by proposing a time matrix and a pie-chart mapping technique.

Detection of stationary spots and cruising trips

From the vertical lines shown in space-time cube visualization results (Figure 5.16(b) and Figure 5.16(d)), we observe that there are different driving behaviors when the taxi is non-occupied, either stationary for a long time or keep moving without passengers. This stimulates us to extract two sub event types from the non-occupancy events, namely stationary

	Top taxi group	Bottom taxi group
Cruising trips	249600	141698
Stationary spots	30623	24263

Table 5.1: The numbers of the occupied, cruising trips and the stationary spots.



(1a) (1b) (1c) (2a) (2b) (2c)

Figure 5.17: The procedure of detecting the stationary spots. The black dots are stop events.

event and cruising event.

This section details the process for detecting stationary spots from the non-occupied events. A stationary spot is the location where a sequence of stop events occur and last for more than five minutes. A two-step approach is designed to detect the stationary spots. The specific work flow is illustrated in Figure 5.17 and described as follows:

First, we aggregate a sequence of static GPS points ($v = 0$) of a taxi to predefined grids (grid size is about 100m$\times$100m) and calculate the centroids of the points inside each grid.

- (1a) Extract raw GPS points from the dataset with cars of unoccupied state and with the instantaneous velocity of zero, and partition the study area into grids with the side length of l_g.
- (1b) Assigning the extracted GPS points into the grids. In Figure 5.17 (1b), G1, G2, G3 and G4 are the grid divisions.
- (1c) Calculate the duration and the count of the time-series GPS points inside each grid for each taxi. If a sequence of stop events in a grid lasts a long enough time period (e.g. more than 5 minutes) and are geographically close enough (e.g. $l_g/4$), we treat such a sequence as a cluster. In Figure 5.17(1c), the green dots in the grid G2 and G3 are two clusters of stop events.

Next, a refinement step is applied to the clusters, whose centroids are close to the boundary of the grid by adjusting the surrounding grids to recalculate the clusters inside the new grid.

(2a) Move the grid by $l_g/2$ if the mean center inside a grid is close to its grid boundary. For instance, in Figure 5.17 the green rectangle named $G2_{shifted}$ is a new grid.

(2b) Repeat step (1c) and get the cluster in the new grid. The red dot in Figure 5.17 (2b) is the new mean center of $G2_{shifted}$.

(2c) Finally, the mean centers are identified as the stationary spots of the GPS points. In Figure 5.17 (2c), there are two stationary spots denoted by red dots.

The detection of stationary spots by the above procedure largely reduces the amount of GPS points. Note that the detected stationary spots are associated with temporal and thematic attributes (e.g. stationary duration, starting and ending timestamp) and they may refer to locations with different meanings such as parking lots, railway stations or hotels, or traffic congestion locations at road segments or intersections.

After the detection, the stationary event can be extracted from the non-occupancy event and the rest of the GPS points are of the cruising state and can be used for reconstructing the cruising trips. Table 5.1 shows the numbers of the reconstructed occupied, cruising trips and the stationary spots analyzed in this thesis.

In the following part of this section, we focus on the visual analysis of spatiotemporal patterns of the cruising and stationary events. To investigate the general spatial patterns of the cruising events, we use dot mapping technique and standard deviational ellipses for visualizing their average daily spatial mean centers and the corresponding standard deviations. To inspect the temporal patterns of the stationary events, we use time matrix graph representing the temporal aggregates. For the spatiotemporal patterns of the stationary events, we design pie-chart maps.

Dot mapping

When a taxi driver is cruising, he or she plans to minimize his or her cruising time and intends to quickly pick up the next passenger. To investigate the temporal patterns of the cruising trips between top and bottom income taxis, we calculate the average hourly cruising duration (shown in Figure 5.18) between the two groups. The top income taxis are normally cruising longer than the bottom group, especially in the time slots of midnight and early morning from 22:00 to 06:00. During the day, typically at 09:00 and 17:00, the differences of their cruising durations are relatively small.

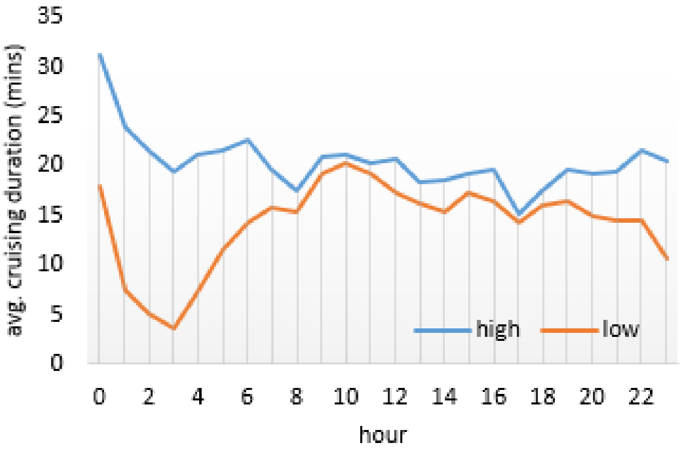


Figure 5.18: The average hourly cruising duration for top and bottom income groups.

To get an overview of the spatial distribution of the cruising events, we examine the distribution of their mean centers.

Given a taxi trip $t = ((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, its spatial mean center is

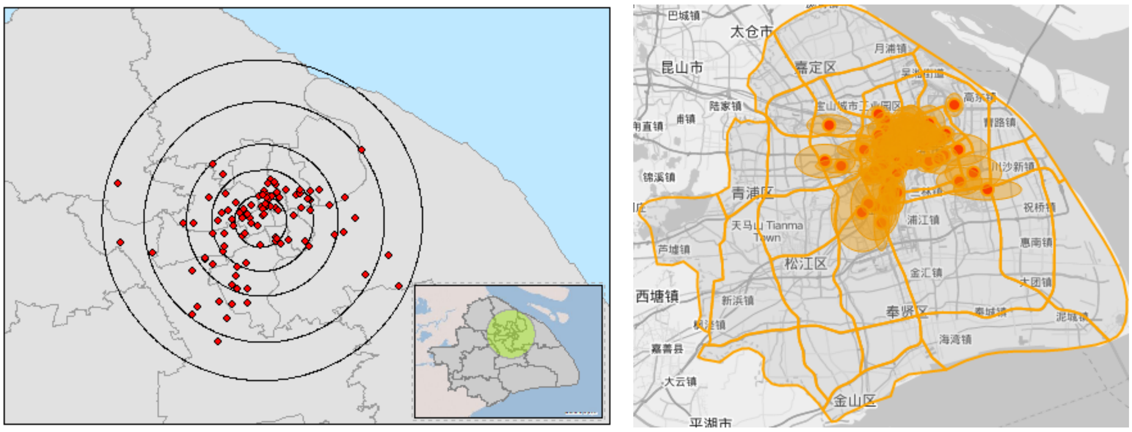
$$mc_{trip}(t) = (1/n \sum_{i=1,...,n} x_i, 1/n \sum_{i=1,...,n} y_i)$$

Consequently, given a set of taxi trips $T = (t_1, t_2, ..., t_m)$ in one day, the spatial mean center of the trips is

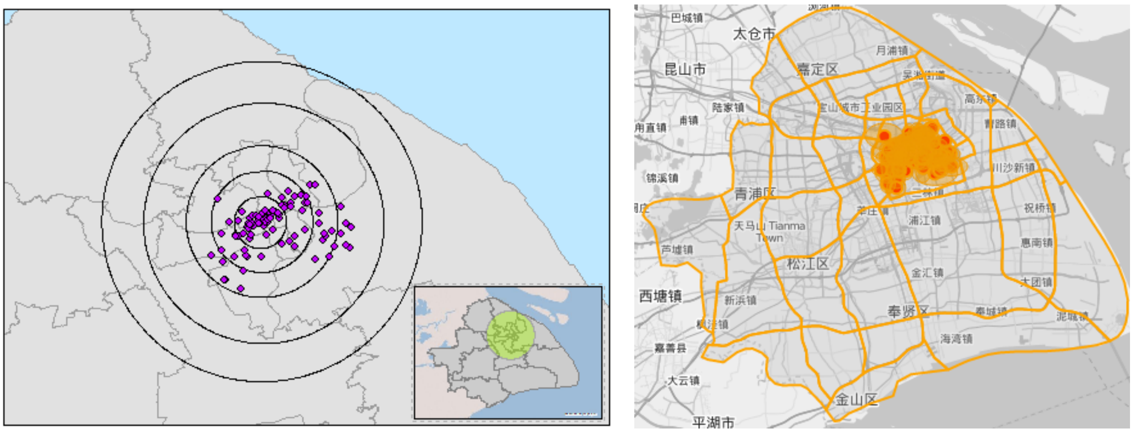
$$mc_{trip_{day}} = 1/m \sum_{t \in T} mc_{trip}(t)$$

Based on the above definitions, the average daily spatial mean centers of each car in the 47 days are calculated. Furthermore, to investigate the amount of variation or dispersion, we also calculate the standard deviation of the daily spatial mean centers of each taxi.

Dot maps are designed to show the average daily spatial mean centers of selected trajectories and their standard deviations patterns or uncertainties. We use filled circles to represent the daily spatial mean centers, and semi-transparent ellipse to represent the corresponding standard deviations. Figure 5.19 shows the exemplary results of the dot maps.



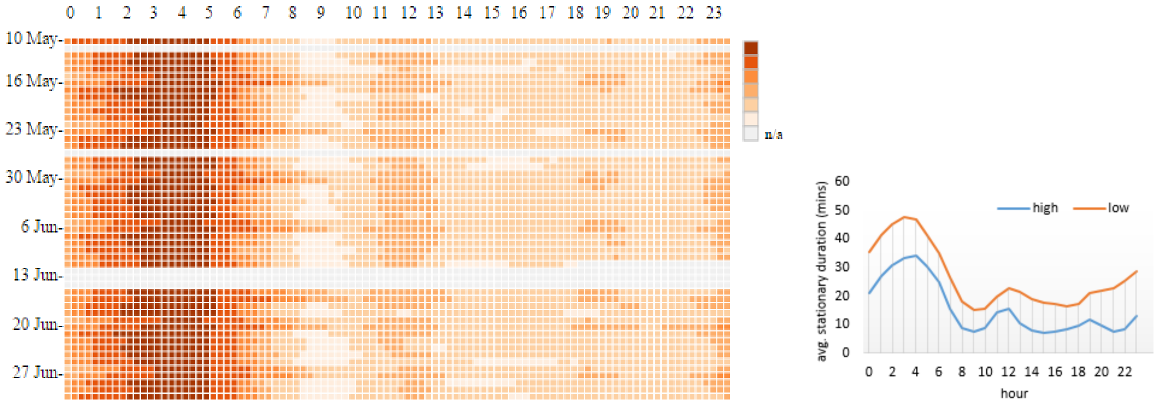
(a) bottom income group (b) bottom income group



(c) top income group (d) top income group

Figure 5.19: The average daily spatial mean centers of the cruising trips (a, b) and their standard deviations (c, d). In (a, b), dots represent the average daily spatial mean centers of each taxi and ellipses in (c, d) their standard deviations.

Figure 5.19 shows the distribution of the average daily spatial mean centers (Figure 5.19(a) and 5.19(b)) of the cruising trips and their standard deviations (Figure 5.19(c) and 5.19(d)) of the bottom group (Figure 5.19(a) and 5.19(c)) and top group (Figure 5.19(b) and 5.19(d)). The rings in Figure 5.19(a) and 5.19(b) are spatial partitions indicating the areas from the center of Shanghai to the peripheral. One can easily observe that the spatial distribution of the average daily cruise centers of the top group is more compact and concentrated in the city center, while the spatial distribution of the daily cruise mean centers of the bottom group is more dispersed. The orange ellipses in Figure 5.19(c) and 5.19(d) indicate that the daily spatial mean center for the top group is more centralized than those of the bottom group. For the bottom group, the further they cruise from the city center, the more they



(a) The time graph of the stationary durations aggregated into 15-minute intervals.

(b) Average hourly durations at stationary spots.

Figure 5.20: Temporal patterns of stationary events.

spread from their daily cruising spatial mean centers.

Time matrix graph and pie-chart mapping

To get a temporal overview of the stationary spots, we calculated the average stationary durations of all the 2000 taxis, divided into 15-minute time intervals. The average stationary duration is visualized in a time graph (see Figure 5.20(a)). The rows represent the dates from 10 May to 30 June, and the columns represent every 15 minute in 24 hours. The color scheme is chosen from the colorbrewer system. The darker the individual block is, the longer the taxis remain static. The white color means missing data (11 May, 4 hours on 11 June, and 12 - 14 June).

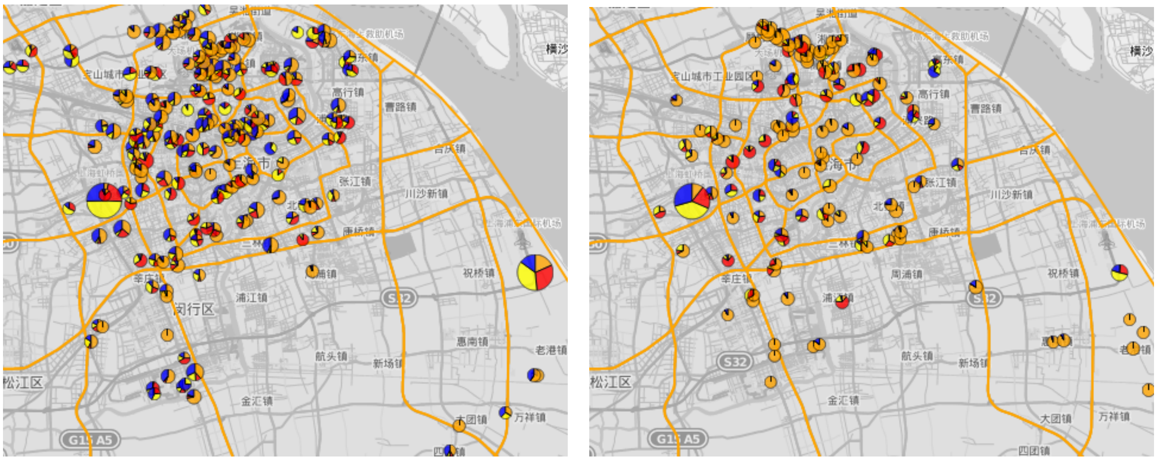
The time graph in Figure 5.20(a) shows significant daily and weekly rhythm. The daily pattern is illustrated in each row where the dark red colors occur in the early morning (01:00 - 06:00), at noon (especially from 11:30 - 13:00), and in a short period of evening (around 19:00). These time slots can be interpreted as the sleeping, lunch and dinner periods respectively.

The weekly pattern is that for every seven rows the dark red color occurrences shifted about two or three hour's right, which means that stationary durations at weekends are about 1 - 2 hours longer and later than that at weekdays. Interestingly, one exceptional weekend pattern can be found on 15 and 16 June (Tuesday and Wednesday). By looking up the Chinese calendar of 2010, we found that 14th to 16th June is the three-day national holiday for the Duanwu Festival (unfortunately the data on 14 June are missing). To study the difference between the top and bottom groups, we also inspected the time graphs of the stationary duration for them and found similar daily and weekly patterns. Next, we created the line graph of Figure 5.20(b) which shows the average hourly stationary duration of the top and bottom drivers. Unsurprisingly, the bottom taxi drivers stay at one places longer than the top taxi drivers.

We investigate the spatiotemporal distribution of stationary spots with relatively long durations, which are normally breaks or long waiting periods. For instance, during the night, such stationary spots are mostly related to sleeping, and during the day, they may imply long waiting in a queue. It is important to know the spatiotemporal distribution of such events since they may have significant impacts on the taxis' overall income.

Here, we extract and examine the stationary spots with the duration more than 30 minutes. A density-based spatial clustering approach is applied to cluster these stationary spots. Specifically, DBSCAN (density-based spatial clustering of applications with noise) algorithm [Est+96] is applied to detect clusters of arbitrary shapes, with the distance parameter set to 20m and the minimum point parameter to 10. The number of the extracted spatial clusters of the bottom and top groups are 216 and 137 respectively.

Furthermore, we divided one day into four time intervals, namely midnight (00:00 - 06:00), morning (06:00 - 12:00), afternoon (12:00 - 18:00) and evening (18:00 - 24:00) and calculate in each cluster the number of cluster elements (stationary spots) in each interval. Pie-chart maps are used to visualize the spatial and temporal distribution of the clusters. The location of each pie chart represents a cluster centroid; the size of the pie chart is proportional to the total number of elements in each cluster; and the colors (i.e. orange, red, yellow and blue) of the pie chart sectors correspond to the four time intervals and these sectors are proportionally sized to the respective number of the cluster elements in each time interval.



(a) Bottom income group

(b) Top income group

● Midnight ● Morning ● Afternoon ● Evening

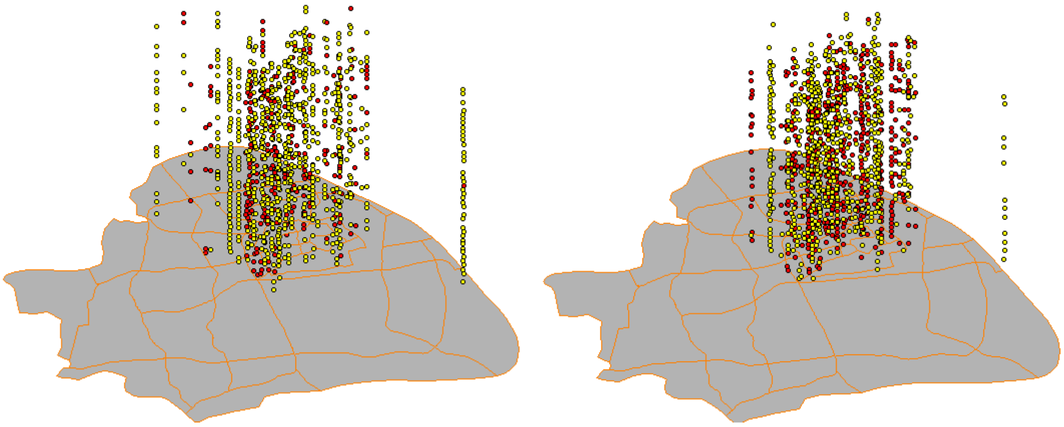
Figure 5.21: Comparison of the spatiotemporal distribution of the cluster centroids of the stationary spots. The predefined four time slots are color coded; the number of stationary spots in the each cluster is proportionally sized to the size of each pie.

The two screeshots in Figure 5.21 show the spatiotemporal distribution of the cluster centroids of stationary spots with the duration more than 30 minutes. There are two relatively large pie charts in both graphs, which means that both groups often stay at these two places. These two pie charts have large proportions of afternoon and evening. Actually, they are located at the two transport hubs of Hongqiao (Figure 5.21(a)) and Pudong international airport (Figure 5.21(b)). We also notice that the cluster at Pudong for the top group is much smaller than that for the bottom group. For both taxi groups, the relative small pie charts are mostly in orange and correspond to the midnight while for the bottom taxi groups there are more blue portions (Figure 5.21(a)) and refers to more stationary spots during the evening interval (18:00 - 24:00). We can interpret that the small circles with a large midnight and evening portions may correspond to their long-break and familiar places, e.g. the locations of the taxi companies or taxi drivers' home.

Proportional symbol mapping of stationary spots in off-peak hours

We noticed in (Figure 3.9(b)) there is a large discrepcany of the occupied ratios between the top and bottom driver groups from 10am to 12pm. This off -peak-hour occupied ratio difference might result from distinct driving behaviors and is interesting to get an insight.

To understand the spatiotemporal patterns of the stationary spots from 10:00 to 12:00,



(a) Bottom income group (b) Top income group

Figure 5.22: Comparison of the spatiotemporal distribution of the stationary spots with space-time cube between the bottom (a) and top (b) taxi groups from 10am to12pm in the 47 days. Red dots represent the traffic congestion places and yellow dots the parking places.

we extract from the detected stationary spots about 1100 and 1500 stationary spots during this time interval for the bottom and top taxi groups respectively.

Until now, we did not differentiate the meanings of the stationary spots. Here, we classify the stationary spots from 10:00 to 12:00 to traffic congestion and parking places. To identify the traffic congestion clusters, we utilize the road networks to calculate the distance from the clustered stationary spots to their nearest roads. A threshold (e.g. 20 m) is adopted between a traffic congestion place (< 20m) and a parking place (≥ 20m). Figure 5.22 shows the space-time cube of the stationary spots of the bottom and top taxi groups from 10:00 to 12:00 in the 47 days. Red dots represent the traffic congestion places and yellow dots the parking places. From Figure 5.22, we can see that there are obviously more parking events indicated by the yellow dots from the bottom taxi group (Figure 5.22(a)), while there are more congestion events denoted by the red dots from the top taxi group (Figure 5.22(a)).

To compensate the occlusion effect the space-time cube, a pie chart map of the cluster centroids are inspected (shown in Figure 5.23). Similarly, we apply the DBSCAN clustering method and get 95 and 102 clusters after the clustering method. Each pie chart corresponds to a cluster centroid. If more than half of the elements in a cluster are traffic congestion, the cluster centroid is regarded as a traffic congestion place and coded in red; otherwise it is a parking place and coded in yellow. The color values correspond to the average hourly stationary duration in the cluster. The size of the pie charts is proportional to the number of elements in clusters. There are about 33 cluster centroids representing traffic congestion for the bottom taxis and 47 for the top taxis. We can see from Figure 5.23 that bottom taxi

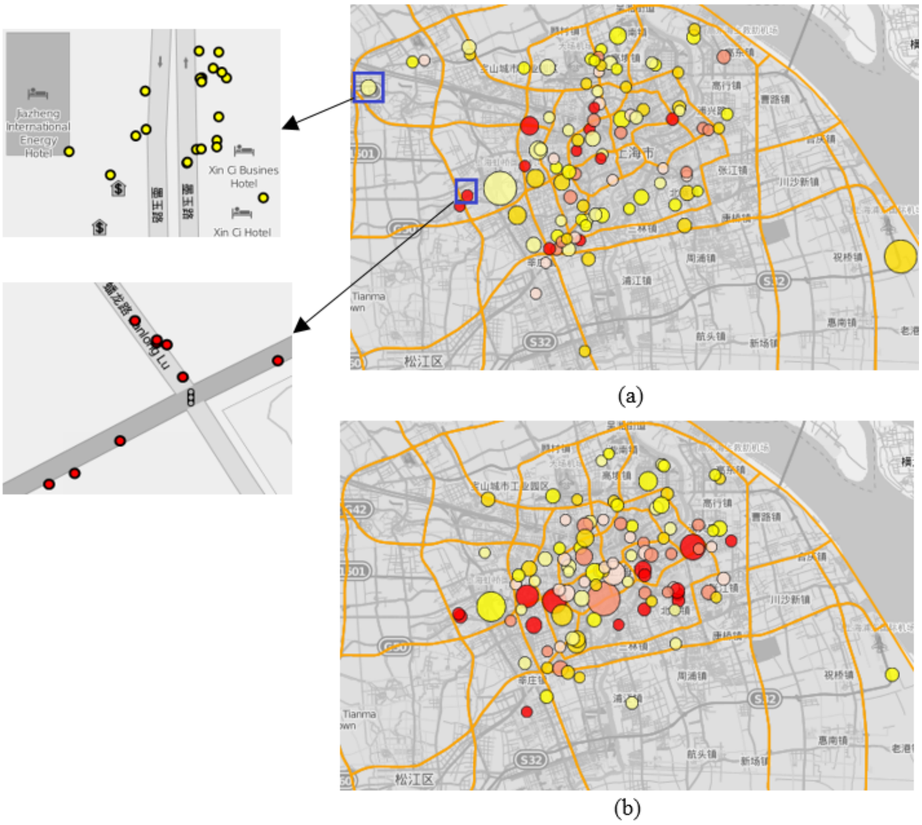


Figure 5.23: Comparison of the spatiotemporal distribution of the parking and traffic congestion places between the bottom (a) and top (b) taxi groups. The parking places are in yellow and the congestion places in red. The color values correspond to the average duration at the places. The size of the circle is proportional to the count of the cluster elements of parking or traffic congestion.

group are more likely go eastward to the airport "Pudong" at this time interval with a longer parking time.

Figure 5.24 shows the average stationary duration (in minutes) spent on the traffic congestion and the parking places (waiting for passengers) for top and bottom taxi groups. The average time spend on traffic congestion is more or less the same, while the parking time from 10:00 to 12:00 for the bottom taxi group are much longer than the top taxi group.

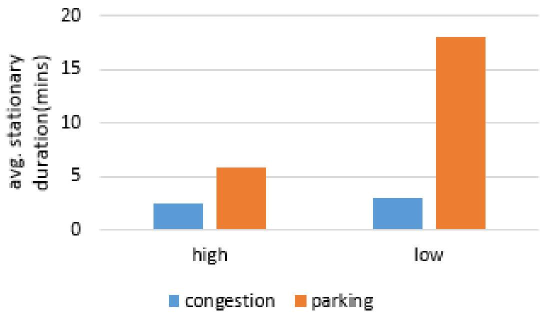


Figure 5.24: The average stationary duration spent on the traffic congestion and parking places (waiting for passengers) for top and bottom taxi groups.

5.4.3 Analysis of the Visualization of Non-Occupancy Events

The above sections studied in detail the spatial and temporal distribution of the trajectory traces of top and bottom taxi groups, which reflects their distinctive driving behaviors.

Differences of the overall temporal patterns when they are cruising or stationary The trend of the occupied ratios between the top and bottom taxi groups (Figure 3.9(b)) reveals that the top taxi group drives longer distance with passengers than the bottom group. In spite of their longer occupied trips, in terms of cruising durations, the top taxi group normally cruises longer than the bottom taxi group especially in the evening and during the midnight (Figure 5.18), which might reduce the profit via the cost of gas consumptions. However, in terms of stationary durations, top taxis have constantly shorter average durations (Figure 5.20(b)), which might in turn compensate the loss of longer cruising. In addition, we also found the daily and weekly routines of all the taxis by investigating their stationary durations (Figure 5.20(a)). The daily patterns show that long breaks often occur in the midnight and early morning, lunch or dinnertime. The weekly pattern shows obvious differences between weekdays and weekends.

Differences of their spatial cruising distributions In terms of spatial driving behaviors, the cruising patterns of the top taxi group are more compact and concentrated in the city center, as shown in Figure 5.19, while the spatial distribution of the bottom taxi group is more dispersed and the further they cruise from the city center, the more they spread. Moreover, comparing the cruising patterns with the spatial distribution of the long break stationary spots, we can observe that there are no obvious relationships of the cruising mean centers and the long-parking places around midnight, and thus can interpret that taxis might not cruise around their long-parking or long-break places.

Differences of their stationary spatiotemporal characteristics We observe from the occupied ratio plot (see Figure 3.9(b)) that there are relatively large differences during off-peak hours, which might be the main reason that result in the income differences. Thus we study in detail the spatial and temporal distributions during the off-peak time intervals, especially from 10:00 to 12:00, and found that the top and bottom taxi group spend similar quantity of time on the road congestion but significantly distinct quantity of time on the parking places (see Figure 5.24). One reason might be that bottom drivers wait much longer in some places, for instance, in Figure 5.23, we can easily see that there are a large number of bottom taxis in the Pudong airport waiting for a relative long time period.

5.5 Summary

In this chapter, we have extensively investigated the spatio-temporal patterns of two types of trajectory-based events, i.e. origin-to-destination and non-occupancy events, by using specific scientific visualization and thematic mapping techniques.

We have first considered the origin-to-destination events. For the extensive exploration, we have employed highly interactive techniques which can serve for diverse purposes. Move specifically, we have proposed an interactive visual clustering method, parallel coordinates, and gradient line rendering techniques. For the communication of overall patterns, we have design easily understandable flow maps. Meanwhile, simple interaction techniques have been employed to highlight the flows of particularly interesting spatial regions.

We have then moved to visual analysis of non-occupancy events focusing on the two taxi groups of different incomes. For exploration, we have applied the interactive clustering technique, parallel coordinates, direct line rendering, and space-time cube to show the dynamic process of the events. For communication, we have further identified two sub-type events, i.e., cruising and stationary events, and focus on mapping their statistics and aggregates by thematic cartography techniques. More specifically, we have used dot mapping techniques to reveal the spatial patterns of cruising events, and time matrix and pie-chart mapping of stationary events.

To summarize, we list in Table 5.2 the techniques of visualizing the trajectory-based events in this chapter. We conclude from the aforementioned examples that:

- Scientific visualization techniques are highly interactive and suitable for event exploration for diverse purposes. Thematic maps are especially designed to communicate specific aspects of the summarized data and easily understandable patterns.
- The exploration stage in scientific visualization can be helpful to identify interesting patterns that can be communicated through thematic mapping. For instance, the space-time cube visualization results (Figure 5.16) stimulate us to differentiate two sub types of non-occupancy events, which are further mapped using a diverse thematic mapping techniques to show their respective spatiotemporal patterns.
- Thematic mapping results with some hotspots can be helpful for orienting the next

<table>
<tr><th>Techniques</th><th>Origin-to-destination</th><th>Non-occupancy</th></tr>
<tr><td rowspan="3">Scientific Visualization</td><td colspan="2">Interactive visual clustering</td></tr>
<tr><td colspan="2">Parallel coordinates</td></tr>
<tr><td>Gradient line rendering</td><td>Direct line rendering</td></tr>
<tr><td>Thematic mapping</td><td>Flow mapping</td><td>Dot mapping
Time matrix graph
Pie-chart mapping</td></tr>
</table>

Table 5.2: Visualization techniques for trajectory-based events.

step exploration. For instance, from the flow map (Figure 5.12), users can perceive the flow volume between pairs of the spatial divisions and may easily find interesting flow patterns in a specific area. The users then can select the area and check the individual events by rendering the lines on a map.

6 Conclusion and Outlook

6.1 Conclusion

This thesis is devoted to bridging the gaps between two disciplines thematic mapping and scientific visualization to achieve synergetic effects for the visual analysis of big data. To demonstrate the necessity and feasibility, the author first conducted an in-depth comparative study of these two disciplines, and proposed a variety of thematic mapping and scientific visualization techniques for the visual exploration and communication of spatiotemporal patterns of massive movement events derived from taxi floating car data.

This work makes the following main contributions:

- It established the theoretical framework to achieve synergetic effects for the visual analysis of big data by taking advantage of the two disciplines. Chapter 2 conducted a systematical comparative study of thematic cartography and scientific visualization. Following the order in the visualization pipeline, the comparison covered seven essential aspects: purposes, data sources, georeferencing, data preprocessing, classification, symbolization, and finally perception and cognition. The comparison results showed that these two disciplines reveal different visual analytical levels and play a complementary role with each other. These findings suggest the necessity and feasibility of achieving synergetic effects from two disciplines by taking advantage of their complementary characteristics.
- Based on the theoretical findings, we identified specific use cases and conducted extensive experiments of visually analyzing massive and complex real-world taxi floating car data (FCD). Two FCD event categories at different abstraction levels, namely point-based and trajectory-based events, were identified for visual analysis.
- For the visual analysis of multivariate point-based events (O, N, P, D), we proposed a variety of scientific visualization and thematic mapping techniques, which reveal spatiotemporal patterns of the events at multi-scales from multiple perspectives. The visual exploration by means of scientific visualization methods is based on the proposed pie radar glyphs, time graphs and a salience-based method, which reveal the spatiotemporal patterns of multivariate events, the temporal patterns of each individual variate and the most salient variates. Their compact visualizations revealed the underlying interesting data distributions, which can be used for further mining, e.g. inferring urban land use types. To effectively communicate the patterns of the

four types of point-based events, choropleth and proportional mapping techniques, are applied, which allow an easy perception and interpretation of high-level spatiotemporal event patterns.

- For trajectory-based events, we firstly identified two specific trajectory-based events, i.e. origin-to-destination and non-occupancy events, and then proposed a visual analysis framework integrating multiple scientific visualization and thematic mapping techniques. From the perspective of scientific visualization, highly interactive techniques were employed, including interactive clustering methods, a parallel coordinates approach, gradient and direct line rendering techniques, and space-time cube, to allow users freely explore potentially interesting clusters. A variety of thematic mapping techniques, including flow mapping, dot mapping, time matrix graph, and proportional mapping, are proposed for the effective communication of the spatial interaction patterns and taxi driving behavior patterns.

6.2 Outlook

The diverse historical developments of thematic mapping and scientific visualization led to the different disciplinary focuses. However, facing the challenges proposed by big data, future geovisual analytics needs to integrate exploration functions by means of scientific visualization and the communication functions based on thematic mapping to allow the analysts to effectively get insights into massive geospatial data.

On the basis of this work a number of future improvements are possible:

- Based on the theoretical framework and empirical findings of this thesis, we will consolidate a methodological design system that can effectively couple thematic mapping with scientific visualization techniques.
- Following the experiments on the FCD data, we plan to carry on user evaluations on fundamental questions of what and how much information the user can effectively and efficiently derive from various combinatorial functionalities from scientific visualization and thematic mapping.
- From the visualization technical point of view, we will continue to extend the existing visualization techniques for better exploration and communication. From the practical usage perspective, we will identify target user groups that can benefit from using the developed tools, e.g. taxi companies and city/traffic/transport planners. For instance, understanding the significant different driving behaviors between high- and low-income taxi groups of the non-occupancy trajectory events (Section 5.4) may help both taxi drivers and taxi companies to make more profits.
- We also plan to experiment our techniques on other large datasets, e.g., social media data, point clouds, and mobile phone data.

References

[AA11] Natalia Adrienko and Gennady Adrienko. “Spatial Generalization and Aggregation of Massive Movement Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.2 (Feb. 2011), pp. 205–219 (page 63).

[AA10] G. Andrienko and N. Andrienko. “A General Framework for Using Aggregation in Visual Exploration of Movement Data”. In: *The Cartographic Journal* (2010), pp. 22–40 (page 48).

[And+13a] Gennady L. Andrienko, Natalia V. Andrienko, Peter Bak, Daniel A. Keim, and Stefan Wrobel. *Visual Analytics of Movement*. Springer, 2013 (pages 30, 49).

[And+13b] Gennady L. Andrienko, Natalia V. Andrienko, Georg Fuchs, Ana-Maria Olteanu Raimond, Jürgen Symanzik, and Cezary Ziemlicki. “Extracting Semantics of Individual Places from Movement Data by Analyzing Temporal Patterns of Visits”. In: *ACM SIGSPATIAL International Workshop on Computational Models of Place, COMP 2013, November 5, 2013, Orlando, Florida, USA*. 2013, pp. 9–15 (pages 2, 30).

[And+11] Gennady L. Andrienko, Natalia V. Andrienko, Christophe Hurter, Salvatore Rinzivillo, and Stefan Wrobel. “From movement tracks through events to places: Extracting and characterizing significant places from mobility data.” In: *IEEE VAST*. IEEE, 2011, pp. 161–170 (pages 2, 30).

[AA08] Gennady Andrienko and Natalia Andrienko. “Spatio-temporal aggregation for visual analysis of movements”. In: *In Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST 2008), IEEE Computer*. Society Press, 2008, pp. 51–58 (page 84).

[AAH11] Gennady Andrienko, Natalia Andrienko, and Marco Heurich. “An Event-based Conceptual Model for Context-aware Movement Analysis”. In: *Int. J. Geogr. Inf. Sci.* 25.9 (Sept. 2011), pp. 1347–1370 (pages 30, 47).

[And+13c] Gennady Andrienko, Natalia Andrienko, Christophe Hurter, Salvatore Rinzivillo, and Sophie Wrobel. “Scalable analysis of movement data for extracting and exploring significant places”. In: *Visualization and Computer Graphics, IEEE Transactions on* 19.7 (2013), pp. 1078–1094 (page 48).

[And+09] Gennady Andrienko, Natalia Andrienko, Salvatore Rinzivillo, Mirco Nanni, Dino Pedreschi, and Fosca Giannotti. “Interactive Visual Clustering of Large Collections of Trajectories”. In: *VAST* (2009) (page 84).

[And+14] N. Andrienko, G. Andrienko, G. Fuchs, and H. Stange. "Detecting and tracking dynamic clusters of spatial events". In: *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. Oct. 2014, pp. 219–220 (page 48).

[And+13d] Natalia V. Andrienko, Gennady L. Andrienko, Louise Barrett, Marcus Dostie, and S. Peter Henzi. "Space Transformation for Understanding Group Movement". In: *IEEE Trans. Vis. Comput. Graph.* 19.12 (2013), pp. 2169–2178 (page 84).

[AA04] Natalia Andrienko and Gennady Andrienko. "Interactive visual tools to explore spatio-temporal variation". In: *Proceedings of the working conference on Advanced visual interfaces*. AVI '04. Gallipoli, Italy: ACM, 2004, pp. 417–420 (page 2).

[BDP08] Kate Beard, Heather Deese, and Neal R. Pettigrew. "A Framework for Visualization and Exploration of Events". In: *Information Visualization* 7.2 (Apr. 2008), pp. 133–151 (pages 30, 47).

[Ber83] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps, Madison*. Translated by W.J. Berg. WI: University of Visconsin Press, 1983 (page 22).

[Boy+11] Ilya Boyandin, Enrico Bertini, Peter Bak, and Denis Lalanne. "Flowstrates: An Approach for Visual Exploration of Temporal Origin-destination Data". In: *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*. EuroVis'11. Bergen, Norway: The Eurographs Association &#38; John Wiley &#38; Sons, Ltd., 2011, pp. 971–980 (page 84).

[Buz99] Gerd Buziek. "Dynamic Elements of Multimedia Cartography". English. In: *Multimedia Cartography*. Ed. by William Cartwright, MichaelP. Peterson, and Georg Gartner. Springer Berlin Heidelberg, 1999, pp. 231–244 (page 8).

[Buz01] Gerd Buziek. "Eine Konzeption der kartographischen Visualisierung". Habil.-Schr. Hannover University, 2001 (pages 1, 22, 84).

[Cas+13] Pablo Samuel Castro, Daqing Zhang, Chao Chen, Shijian Li, and Gang Pan. "From Taxi GPS Traces to Social and Community Dynamics: A Survey". In: *ACM Comput. Surv.* 46.2 (Dec. 2013), 17:1–17:34 (page 59).

[Che05] Chaomei Chen. "Top 10 Unsolved Information Visualization Problems". In: *IEEE Comput. Graph. Appl.* 25.4 (July 2005), pp. 12–16 (page 22).

[Che06] Chaomei Chen. *Information Visualization: Beyond the Horizon*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006 (pages 2, 5, 12, 17).

[Chr02] Nicholas Chrisman. *Exploring geographic information systems (2nd Edition)*. New York: John Wiley & Sons, Inc., 2002 (page 16).

[Den90] Borden D. Dent. *Cartography: Thematic Map Design 3rd Ed.* Dubuque, Iowa: Wm. C. Brown Publishers, 1990 (page 19).

[DiB90] D. DiBiase. "Visualization in the earth sciences". In: *College of Earth and Mineral Sciences, Pensylvania State Univ.* 59.2 (1990), pp. 13–18 (page 6).

[Dob73] Michael W. Dobson. "Choropleth maps without class intervals? A comment". In: *Geographical Analysis* 5.4 (Oct. 1973), pp. 358–360 (page 19).

[Elz+13] S. van den Elzen, J. Blaas, D. Holten, J.-K. Buenen, J.J. van Wijk, R. Spousta, A. Miao, S. Sala, and S. Chan. "Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics Approach". In: *In V. Blondel et al., Mobile Phone Data for Development, Selected contributions to the D4D challenge 3rd International Conference on the Analysis of Mobile Phone Datasets (NetMob 2013), Cambridge, MA, May 1-3, 2013*. 2013, pp. 85–94 (page 84).

[EW14] S. van den Elzen and J.J. van Wijk. "Multivariate Network Exploration and Presentation: From Detail to Overview via Selections and Aggregations". In: *Visualization and Computer Graphics, IEEE Transactions on* 20.12 (Dec. 2014), pp. 2310–2319 (page 84).

[Est+96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*. Ed. by Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad. AAAI Press, 1996, pp. 226–231 (page 116).

[Eva77] Ian S. Evans. "The selection of class intervals". In: *Transactions of the Institute of British Geographers* 2.1 (1977), pp. 98–124 (page 19).

[FS05] S. I. Fabrikant and A. Skupin. "Exploring GeoVisualization". In: Amsterdam: Elsevier, 2005. Chap. Cognitively plausible information visualization, pp. 667–690 (page 12).

[Fis10] C. Fish. "Change Ditection in Animated Choropleth Maps". In: (2010). Master thesis. Michigan State University (page 19).

[FR93] A. Stewart Fotheringham and Peter A. Rogerson. "GIS and Spatial Analytical Problems". In: *International Journal of Geographical Information Systems* 7.1 (1993), pp. 3–19 (page 17).

[GAA04] P. Gatalsky, N. Andrienko, and G. Andrienko. "Interactive analysis of event data using space-time cube". In: *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*. July 2004, pp. 145–152 (page 48).

[GB09] Kirk Goldsberry and Sarah Battersby. "Issues of Change Detection in Animated Choropleth Maps." In: *Cartographica* 44.3 (2009), pp. 201–215 (page 19).

[GHB08] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. "Understanding individual human mobility patterns". In: *Nature* 453.7196 (2008), pp. 779–782 (pages 2, 29).

[Gra+15] Sebastian Grauwin, Stanislav Sobolevsky, Simon Moritz, Istvan Godor, and Carlo Ratti. "Towards a Comparative Science of Cities: Using Mobile Traffic Records in New York, London, and Hong Kong". In: *Computational Approaches for Urban Environments*. Ed. by Marco Helbich, Jamal Jokar Arsanjani, and Michael Leitner. Vol. 13. Geotechnologies and the Environment. Springer International Publishing, 2015, pp. 363–387 (page 58).

[Guo07] D. Guo. "Visual analytics of spatial interaction patterns for pandemic decision support". In: *International Journal of Geographical Information Science* 21.8 (2007), pp. 859–877 (page 84).

[Guo09] Diansheng Guo. "Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data". In: *Visualization and Computer Graphics, IEEE Transactions on* 15.6 (Nov. 2009), pp. 1041–1048 (page 84).

[GZ14] Diansheng Guo and Xi Zhu. "Origin-Destination Flow Data Smoothing and Mapping". In: *IEEE Trans. Vis. Comput. Graph.* 20.12 (2014), pp. 2043–2052 (page 84).

[Guo+12] Diansheng Guo, Xi Zhu, Hai Jin, Peng Gao, and Clio Andris. "Discovering Spatial Patterns in Origin-Destination Mobility Data". In: *T. GIS* 16.3 (2012), pp. 411–429 (page 84).

[Guo+11] Hanqi Guo, Zuchao Wang, Bowen Yu, Huijing Zhao, and Xiaoru Yuan. "TripVista: Triple Perspective Visual Trajectory Analytics and its application on microscopic traffic data at a road intersection". In: *IEEE Pacific Visualization Symposium, PacificVis 2011, Hong Kong, China, 1-4 March, 2011*. 2011, pp. 163–170 (page 84).

[HKT01] Jiawei Han, Micheline Kamber, and Anthony K. H. Tung. "Spatial Clustering Methods in Data Mining: A Survey". In: *Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS*. Ed. by Harvey J. Miller and Jiawei Han. Taylor and Francis, 2001 (page 61).

[Har03] Mark Harrower. "Tips for Designing Effective Animated Maps". In: *Cartographic Perspectives* (2003), pp. 63–65, 82–83 (page 19).

[HW12] Julian Heinrich and Daniel Weiskopf. "State of the Art of Parallel Coordinates". In: *Eurographics 2013 - State of the Art Reports*. Ed. by M. Sbert and L. Szirmay-Kalos. The Eurographics Association, 2012 (page 90).

[Hol06] D. Holten. "Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data". In: *Visualization and Computer Graphics, IEEE Transactions on* 12.5 (Sept. 2006), pp. 741–748 (page 84).

[HW09] Danny Holten and Jarke J. van Wijk. "Force-directed Edge Bundling for Graph Visualization". In: *Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization*. EuroVis'09. Berlin, Germany: The Eurographs Association &#38; John Wiley &#38; Sons, Ltd., 2009, pp. 983–998 (page 84).

[HXR70] M.L. Hsu, M. Xu, and A.H. Robinson. *The fidelity of isopleth maps: an experimental study*. University of Minnesota Press, 1970 (page 17).

[Ins97] A. Inselberg. "Multidimensional Detective". In: *Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97)*. INFOVIS '97. Washington, DC, USA: IEEE Computer Society, 1997, pp. 100– (page 90).

[JK03] G. F. Jenks and D. S. Knos. "The use of shading patterns in graded series". In: *Annals of the Association of American Geographers* 20 (2003), pp. 177–95 (page 19).

[Jia+15] Shan Jiang, Ana Alves, Filipe Rodrigues, Joseph Ferreira Jr., and Francisco C. Pereira. "Mining point-of-interest data from social networks for urban land use classification and disaggregation". In: *Computers, Environment and Urban Systems* (2015) (page 58).

[JH04] C. Johnson and C. Hansen. *Visualization Handbook*. Orlando, FL, USA: Academic Press, Inc., 2004 (page 22).

[KLW14] Chaogui Kang, Yu Liu, and Lun Wu. "Delineating intra-urban spatial connectivity patterns by travel-activities: a case study of Beijing, China". In: *The 23th International Conference on Geoinformatics*. 2014, pp. 1–6 (page 83).

[Kan+13] Chaogui Kang, Yi Zhang, Xiujun Ma, and Yu Liu. "Inferring properties and revealing geographical impacts of intercity mobile communication network of China using a subnet data set". In: *International Journal of Geographical Information Science* 27.3 (2013), pp. 431–448 (page 83).

[KR90] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. A Wiley-Interscience publication. New York: Wiley, 1990 (page 61).

[Kra88] M. J. Kraak. "The space-time cube revisited from a geovisualization perspective". In: *Proceedings of the 21st International Cartographic Conference* 1995 (1988) (page 91).

[LKH10] Ove Daae Lampe, Johannes Kehrer, and Helwig Hauser. "Visual Analysis of Multivariate Movement Data Using Interactive Difference Views". In: *Proceedings of Vision, Modeling, and Visualization (VMV 2010)*. Siegen, Germany, 2010, pp. 315–322 (page 48).

[LAR10] Liang Liu, Clio Andris, and Carlo Ratti. "Uncovering cabdrivers' behavior patterns from their digital traces". In: *Computers, Environment and Urban Systems* 34.6 (2010). GeoVisualization and the Digital CitySpecial issue of the International Cartographic Association Commission on GeoVisualization, pp. 541–548 (pages 2, 29).

[Liu+15a] Xi Liu, Li Gong, Yongxi Gong, and Yu Liu. "Revealing travel patterns and city structure with taxi trip data". In: *Journal of Transport Geography* 43 (2015), pp. 78–90 (pages 2, 30, 83).

[Liu+12] Yu Liu, Chaogui Kang, Song Gao, Yu Xiao, and Yuan Tian. "Understanding intra-urban trip patterns from taxi trajectory data". In: *Journal of Geographical Systems* 14.4 (2012), pp. 463–483 (page 83).

[Liu+15b] Yu Liu, Xi Liu, Song Gao, Li Gong, Chaogui Kang, Ye Zhi, Guanghua Chi, and Li Shi. "Social sensing: a new approach to understanding our socio-economic environments". In: *Annals of the Association of American Geographers* (2015) (pages 30, 58).

[Liu+14] Yu Liu, Zhengwei Sui, Chaogui Kang, and Yong Gao. "Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data". In: *PloS one* 9.1 (2014), e86026 (pages 30, 83).

[LJH13] Zhicheng Liu, Biye Jiang, and Jeffrey Heer. "imMens: Real-time Visual Querying of Big Data". In: *Computer Graphics Forum (Proc. EuroVis)* 32 (3 2013) (page 27).

[Mac85] A. M. MacEachren. "Accuracy of thematic maps: Implications of choropleth symbolization". In: *Cartographica* 22.1 (1985), pp. 38–58 (page 17).

[Mac94] A. M. MacEachren. "Some Truth with Maps: A Primer on Design and Symbolization". In: 1994. Chap. Cartographic Language (page 22).

[MK01] A. M. MacEachren and M. J. Kraak. "Research challenges in geovisualization". In: *Cartography and Geographic Information Science (CaGIS)* 28 (2001), pp. 3–12 (page 1).

[MK97] Alan M. MacEachren and M. J. Kraak. "Exploratory cartographic visualization: Advancing the agenda". In: *Computers & GeosciencesComputers & Geosciences* 23.4 (May 1997), pp. 335–343 (page 6).

[Mac95] AM MacEachren. *How maps work: representation, visualization, and design*. New York: Guilford Press, 1995 (page 22).

[MRC91] Jock D. Mackinlay, George G. Robertson, and Stuart K. Card. "The perspective wall: detail and context smoothly integrated". In: *Proceedings of CHI*. 1991, pp. 173–176 (page 12).

[MT15] Jean Damascene Mazimpaka and Sabine Timpf. "Exploring the Potential of Combining Taxi GPS and Flickr Data for Discovering Functional Regions". English. In: *AGILE 2015*. Ed. by Fernando Bacao, Maribel Yasmina Santos, and Marco Painho. Lecture Notes in Geoinformation and Cartography. Springer International Publishing, 2015, pp. 3–18 (pages 30, 58).

[Men01] Liqiu Meng. "Scroll the Space and Drill-Down the Information". In: *Proceedings of the 20th International Cartographic Conference Beijing 2001*. Vol. 4. 2001, pp. 2436–2443 (page 22).

[Men03] Liqiu Meng. "Missing Theories and Methods in Digital Cartography (English)". In: *CD-Proceedings of the 21st International Cartographic Conference Durban 2003*. 2003 (page 11).

[Men13] Liqiu Meng. "Cartography and Maps Beyond Disciplines". In: *Special issue of Kartographische Nachrichten* (2013), pp. 115–122 (page 27).

[Peu+15] Donna J. Peuquet, Anthony C. Robinson, Samuel Stehle, Franklin A. Hardisty, and Wei Luo. "A method for discovery and analysis of temporal patterns in complex event data". In: *International Journal of Geographical Information Science* (2015), pp. 1–24 (page 48).

[PG88] R. M. Pickett and G. Grinstein. "Iconographic Displays for Visualizing Multidimensional Data". In: *Proceedings of IEEE Conference on Systems, Man, and Cybernetics*. 1988, pp. 514–519 (page 19).

[San+14] Paolo Santi, Giovanni Resta, Michael Szell, Stanislav Sobolevsky, Steven H Strogatz, and Carlo Ratti. "Quantifying the benefits of vehicle pooling with shareability networks". In: *Proceedings of the National Academy of Sciences* 111.37 (2014), pp. 13290–13294 (page 48).

[SVV14] R. Scheepens, H. Van De Wetering, and J.J. Van Wijk. "Non-overlapping Aggregated Multivariate Glyphs for Moving Objects". In: *Visualization Symposium (PacificVis), 2014 IEEE Pacific*. Mar. 2014, pp. 17–24 (page 48).

[SWW14] Roeland Scheepens, Huub van de Wetering, and Jarke J. van Wijk. "Contour based visualization of vessel movement predictions". In: *International Journal of Geographical Information Science* 28.5 (2014), pp. 891–909 (page 84).

[Sch+11] Roeland Scheepens, Niels Willems, Huub van de Wetering, Gennady L. Andrienko, Natalia V. Andrienko, and Jarke J. van Wijk. "Composite Density Maps for Multivariate Trajectories". In: *IEEE Trans. Vis. Comput. Graph.* 17.12 (2011), pp. 2518–2527 (page 84).

[Sch+12] Roeland Scheepens, Niels Willems, Huub van de Wetering, and Jarke J. van Wijk. "Interactive Density Maps for Moving Objects". In: *IEEE Computer Graphics and Applications* 32.1 (2012), pp. 56–66 (page 84).

[Sch55] C. E. Schmid. "Basic Problems, Techniques, and Theory of Isopleth Mapping". In: *Journal of the American Statistical Association* 50 (1955) (page 17).

[SL14] Ruojing W. Scholz and Yongmei Lu. "Detection of Dynamic Activity Patterns at a Collective Level from Large-volume Trajectory Data". In: *Int. J. Geogr. Inf. Sci.* 28.5 (May 2014), pp. 946–963 (page 48).

[SB97] Andre Skupin and Barbara P. Buttenfield. "Spatial Metaphors for Visualizing Information Spaces". In: *Proceedings of ACSM/ASPRS Annual Convention and Exhibition*. 1997, pp. 116–125 (page 12).

[SF03] André Skupin and Sara Irina Fabrikant. "Spatialization methods: a cartographic research agenda for non-geographic information visualization". In: *Cartography and Geographic Information Science* 30 (2003), pp. 95–119 (pages 2, 5).

[Slo+08] Terry A. Slocum, Robert B. Mcmaster, Fritz C. Kessler, and Hugh H. Howard. *Thematic Cartography and Geovisualization (3rd Edition) (Prentice Hall Series in Geographic Information Science)*. Prentice Hall, Apr. 2008 (page 22).

[INS05] IN-SPIR. *IN-SPIRE Tips and Techniques "At a Glance"*. Tech. rep. IN-SPIR, Nov. 2005 (page 12).

[TG02] Masahiro Takatsuka and Mark Gahegan. "GeoVISTA Studio: A Codeless Visual Programming Environment For Geoscientific Data Analysis and Visualization". In: *Computational Geoscience* 28 (2002), pp. 1131–1144 (page 2).

[Tom+12] Christian Tominski, Heidrun Schumann, Gennady L. Andrienko, and Natalia V. Andrienko. "Stacking-Based Visualization of Trajectory Attribute Data". In: *IEEE Trans. Vis. Comput. Graph.* 18.12 (2012), pp. 2565–2574 (page 84).

[WS11] Chaoli Wang and Han-Wei Shen. "Information Theory in Scientific Visualization." In: *Entropy* 13.1 (2011), pp. 254–273 (page 22).

[WY14] Zuchao Wang and Xiaoru Yuan. "Urban trajectory timeline visualization". In: *International Conference on Big Data and Smart Computing, BIGCOMP 2014, Bangkok, Thailand, January 15-17, 2014*. 2014, pp. 13–18 (page 84).

[WB97] Pak Chung Wong and R. Daniel Bergeron. "30 Years of Multidimensional Multivariate Visualization". In: *Proceedings of Scientific Visualization*. IEEE Computer Society Press, 1997, pp. 3–33 (page 48).

[WDS10] Jo Wood, Jason Dykes, and Aidan Slingsby. "Visualisation of Origins, Destinations and Flows with OD Maps". In: *The Cartographic Journal* 47.2 (2010), pp. 117–129 (page 84).

[YZX12] Jing Yuan, Yu Zheng, and Xing Xie. "Discovering Regions of Different Functions in a City Using Human Mobility and POIs". In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '12. Beijing, China: ACM, 2012, pp. 186–194 (pages 30, 58).

[Yua+12] Jing Yuan, Yu Zheng, Xing Xie, and Guangzhong Sun. "T-Drive: Enhancing Driving Directions with Taxi Drivers' Intelligence". In: *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (Jan. 2012) (page 44).

[Zhe+14] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. "Urban Computing: Concepts, Methodologies, and Applications". In: *ACM Transaction on Intelligent Systems and Technology* (2014) (page 47).

[Zhe+08] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. "Understanding mobility based on GPS data". In: *Proceedings of the 10th international conference on Ubiquitous computing*. ACM. 2008, pp. 312–321 (pages 2, 29).

[Zhe+09] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. “Mining interesting locations and travel sequences from GPS trajectories”. In: *Proceedings of the 18th international conference on World wide web*. ACM. 2009, pp. 791–800 (pages 2, 30).

[Zho+13] Hong Zhou, Panpan Xu, Xiaoru Yuan, and Huamin Qu. “Edge bundling in information visualization”. In: *Tsinghua Science and Technology* 18.2 (Apr. 2013), pp. 145–156 (page 84).

Relevant Publications

[DFM15] Linfang Ding, Hongchao Fan, and Liqiu Meng. “Understanding taxi driving behaviors from movement data”. In: *18th AGILE Conference on Geographic Information Science, 9-12 June 2015, Lisboa, Portugal*. June 2015 (pages 2, 29).

[DM14] Linfang Ding and Liqiu Meng. “A comparative study of thematic mapping and scientific visualization”. In: *Annals of GIS* 20.1 (2014), pp. 23–37.

[DXM12] Linfang Ding, Guohui Xiao, and Liqiu Meng. “Derivation and Visual Exploration of 4-D Building Deformation from High-resolution SAR Data”. In: *Proceedings of the Symposium on Service-Oriented Mapping 2012, pages 359368, Vienna, Nov 2012. Jobstmedia Management Verlag*. Ed. by M. Jobst. Jobstmedia Management Verlag, Nov. 2012.

[DYM15] Linfang Ding, Jian Yang, and Liqiu Meng. “Visual Analytics for Understanding Traffic Flows of Transportation Hub from Movement Data”. In: *ICC 2015 - 27th International Cartographic Conference, August 23-28, 2015, Rio de Janeiro, Brazil*. Aug. 2015 (pages 2, 29).

[DZM13] Linfang Ding, Xiaoxiang Zhu, and Liqiu Meng. “Visual Analysis of Large Amounts of 4-D Building Deformation Data”. In: *26th International Cartographic Conference 2013 Dresden, Germany, 25-30*. Aug. 2013.

[FD15] Hongchao Fan and Linfang Ding. “Can FCD data indicate problems in urban planning? A case study in Shanghai”. In: *LBS 2015 - International Conference on Location Based Services 2015, September 16-18, 2015, Augsburg, Germany*. Sept. 2015.

[Kri+12] Jukka M. Krisp, Linfang Ding, Yanmin Jin, and Patrick Peer. “Indoor Routing - Is a centrality measure for an indoor routing network useful?” In: *Mobile Tartu 2012, August, Tartu, Estonia, 22-25*. Aug. 2012.

[KDW13] Jukka M. Krisp, Linfang Ding, and Lianhuan Wei. “Visual clustering of spatio-temporal hotspots for taxi activity in Shanghai”. In: *GeoViz2013, Hamburg, Germany, 6-8*. Mar. 2013.

[KPD12] Jukka M. Krisp, Patrick Peer, and Linfang Ding. “Classification of an indoor routing network based on graph theory”. In: *Geoinformatics, Hongkong, China, 15-17*. June 2012.

[Wei+13] Lianhuan Wei, Jukka M. Krisp, Timo Balz, Mingsheng Liao, and Linfang Ding. *Urban Subsidence Surveillance combining PS-InSAR and Visual Analytics*. Publication. 26th International Cartographic Conference 2013, Dresden, Germany, 25-30, Aug. 2013.

Curriculum Vitae

Personal Data

Name	Linfang Ding
Date of Birth	April 29, 1985
Place of Birth	Henan, China
Nationality	P. R. China
Gender	Female

Education

Oct 2010 –	PhD Candidate Department of Cartography, Technische Universität Müchen, Munich, Germany
Sept 2007 – July 2010	Master of Science School of Earth and Space Science, Peking University, Beijing, China
Sept 2003 – July 2007	Bachelor of Science School of Earth and Space Science, Peking University, Beijing, China
Sept 2004 – July 2007	Bachelor in Economics (Double Degree) China Center for Economic Research, Peking University, Beijing, China

Experience

Oct 2015 –	Scientific Researcher Universität Augsburg, Augsburg, Germany
Oct 2010 –	Research Assistant Technische Universität Müchen, Munich, Germany