Geospatial Knowledge Discovery using Volunteered Geographic Information: a Complex System Perspective

Tao Jia

Doctoral Dissertation
© Tao Jia 2012

Doctoral Dissertation
Division of Geodesy and Geoinformatics
Department of Urban Planning and Environment
Royal Institute of Technology (KTH)
SE-100 44 STOCKHOLM, Sweden

**Trita SoM 2012-16**

ISSN 1653-6126
ISRN KTH/SoM/10-12/SE
ISBN 978-91-7501-531-6

Printed by e-print, Sweden 2012
Abstract

The continuous progression of urbanization has resulted in an increasing number of people living in cities or towns. In parallel, advancements in technologies, such as the Internet, telecommunications, and transportation, have allowed for better connectivity among people. This has engendered drastic changes in urban systems during the recent decades. From a social geographic perspective, the changes in urban systems are primarily characterized by intensive contacts among people and their interactions with the surrounding urban environment, which further leads to subsequent challenging problems such as traffic jams, environmental pollution, urban sprawl, etc. These problems have been reported to be heterogeneous and non-deterministic. Hence, to cope with them, massive amounts of geographic data are required to create new knowledge on urban systems.

Due to the thriving of Volunteer Geographic Information (VGI) in recent years, this thesis presents knowledge on urban systems based on extensive VGI datasets from three sources: highway dataset from the OpenStreetMap (OSM) project, photo location dataset from the Flickr website, and GPS tracking datasets from volunteers, taxicabs, and air flights. The knowledge primarily relates to two issues of urban systems: the urban space and the corresponding human dynamics. In accordance, on one hand, urban space acts as a carrier for associated geographic activities and knowledge of it benefits our understanding of current social and economic problems in urban systems. On the other hand, human dynamics reflect human behavior in urban space, which leads to complex mobility or activity patterns. Its investigation allows a derivation of the underlying driving force that is very instructive to urban planning, traffic management, and infectious disease control. Therefore, to fully understand the two issues, this thesis conducts a thorough investigation from multiple aspects.

The first issue is investigated from four aspects. First, at the city level, the controversial topic of city size regularity is investigated in terms of natural cities, and the conclusion is that Zipf’s law holds stably for all US cities. Second, at the sub-city level, the size distribution of spatial units within different cities in terms of the clusters formed by street nodes, photo locations, and taxi static points are explored, and the result shows a remarkable scaling property of these spatial units. Third, enlightened by the scaling property of the urban space at the city or sub-city level, this thesis devises a novel tool that can demarcate the cities into three categories: compact cities, normal cities, and sprawling cities. The tool is then applied to cities in both the US and three European countries. In the last, another representation of urban space is taken into account, namely the transportation network. The findings report that the US airport
network displays the properties of scale-free, small-world, and disassortative mixing and that the individual natural airports show heterogeneous patterns that are probably subject to geographic constraints and socioeconomic factors.

The second issue is examined from four perspectives. First, at the city level, the movement flow contributed by agents using two types of behavior is investigated through an agent-based simulation, and the result conjectures that the human mobility behavior is mainly shaped by the underlying street network. Second, at the country level, this thesis reports that the human travel length by air can be approximated well by an exponential distribution, and subsequent simulations indicate that human mobility behavior is largely constrained by the underlying airport network. Third, at the regional level, the length that humans travel by car is demonstrated to agree well with a power law with exponential cutoff distribution, and subsequent simulation further reproduces this levy flight characteristic. Based on the simulation, human mobility behavior is again revealed to be primarily shaped by the underlying hierarchical spatial structure. Finally, taxicab static points are adopted to explore human activity patterns, which can be characterized as the regularities in space and time, the heterogeneity and predictability in space.

From a complex system perspective, this thesis presents the knowledge discovered in urban systems using massive volumes of geographic data. Together with new knowledge from empirical findings, the development of methods, and the design of theoretic models, this thesis also shares the research community with geographic data generated from extensive VGI datasets and the corresponding source codes. Moreover, this study is aligned with a paradigm shift in that it analyzes large-size datasets using high processing power as opposed to analyzing small-size datasets with low processing power.

**Keywords:** knowledge discovery, urban systems, complex system, VGI, OSM, GPS tracking dataset, scaling, heavy-tailed distribution detection, urban sprawl, Zipf’s law, human activity/mobility patterns, agent-based modeling, complex network.
List of papers


Table of contents

Abstract .......................................................................................................................... i
List of papers ............................................................................................................... iii
Table of contents .......................................................................................................... iv
List of abbreviations ..................................................................................................... vii
List of figures ............................................................................................................... viii
List of tables ............................................................................................................... xi
Acknowledgements .................................................................................................... xii
1. Introduction ........................................................................................................... 1
  1.1. Background ......................................................................................................... 1
  1.1.1. Urban systems ................................................................................................ 2
  1.1.2. Complex system ............................................................................................ 4
  1.1.3. GIS and geographic data .............................................................................. 6
  1.1.4. VGI and data intensive research ................................................................... 8
  1.2. Thesis aims ........................................................................................................ 9
  1.3. Thesis structure ................................................................................................. 11
  1.4. Thesis declaration ............................................................................................. 13
2. Literature review .................................................................................................... 15
  2.1. Knowledge discovery ....................................................................................... 15
  2.2. Geographic knowledge discovery ..................................................................... 15
  2.2.1. Challenges and strategies .......................................................................... 16
  2.2.2. Spatial point clustering .............................................................................. 17
  2.3. Theories of complex system for geographic knowledge discovery .............. 18
  2.3.1. Complex system ........................................................................................ 19
  2.3.2. Space syntax .............................................................................................. 20
  2.3.3. Complex network analysis ......................................................................... 21
  2.3.4. Agent-based modeling (ABM) .................................................................. 23
  2.3.5. Scaling analysis .......................................................................................... 24
  2.3.6. Information theory ..................................................................................... 26
3. The geographic data and its preprocessing .............................................................. 27
  3.1. OpenStreetMap (OSM) ...................................................................................... 27
    3.1.1. Development and components of OSM .................................................... 27
    3.1.2. Format and usage of OSM data ............................................................... 30
    3.1.3. Street nodes extraction ............................................................................ 33
  3.2. GPS tracking datasets ..................................................................................... 37
    3.2.1. Taxi floating dataset ................................................................................ 38
    3.2.2. Flight tracking dataset ............................................................................. 39
    3.2.3. Volunteer movement dataset .................................................................... 41
4. Methodologies ........................................................................................................ 45
  4.1. Overall structure .............................................................................................. 45
  4.2. Heavy-tailed distribution detection ................................................................. 47
  4.3. Head/tail division rule ..................................................................................... 53
  4.4. Spatial point clustering method ....................................................................... 55
  4.5. Urban sprawl detection ................................................................................... 58
  4.6. Complex network analysis ............................................................................. 60
  4.7. Agent-based modeling (ABM) ....................................................................... 62
5. Results and discussions .......................................................................................... 65
  5.1. Overview .......................................................................................................... 65
  5.2. Paper VII: Validating Zipf’s Law for all the US cities ..................................... 65
  5.3. Paper II: Uncovering scaling property of urban systems ................................ 67
  5.4. Paper I: Measuring urban sprawl using massive street nodes ....................... 68
  5.5. Paper III: Analyzing the US airport network .................................................. 70
  5.6. Paper VI: Exploring human mobility patterns at the city level ....................... 73
  5.7. Paper VIII: Exploring human mobility patterns at the country level .............. 76
  5.8. Paper V: Exploring human mobility patterns at the regional level ................. 79
  5.9. Paper IV: Exploring human activity patterns .................................................. 80
6. Conclusions and future work .................................................................................. 83
  6.1. Conclusions ....................................................................................................... 83
  6.2. Future work ....................................................................................................... 85
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABM</td>
<td>Agent-based Modeling</td>
</tr>
<tr>
<td>ASP</td>
<td>Average Shortest Path</td>
</tr>
<tr>
<td>CC</td>
<td>Clustering Coefficient</td>
</tr>
<tr>
<td>CCA</td>
<td>City Clustering Algorithm</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Distribution Function</td>
</tr>
<tr>
<td>CGIS</td>
<td>Canadian Geographic Information System</td>
</tr>
<tr>
<td>DBSCA</td>
<td>Density-Based Spatial Clustering of Applications</td>
</tr>
<tr>
<td>EHCA</td>
<td>Entropy-based Hierarchical Clustering Algorithm</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td>FAA</td>
<td>Federal Aviation Administration</td>
</tr>
<tr>
<td>GAM</td>
<td>Geographic Analysis Machine</td>
</tr>
<tr>
<td>GoF</td>
<td>Goodness of Fit</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>JOSM</td>
<td>Java OSM Editor</td>
</tr>
<tr>
<td>KSS</td>
<td>Kolmogorov-Smirnov Statistic</td>
</tr>
<tr>
<td>LiDAR</td>
<td>Light Detect and Ranging</td>
</tr>
<tr>
<td>LLR</td>
<td>Log Likelihood Ratio</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimation</td>
</tr>
<tr>
<td>OSM</td>
<td>OpenStreetMap</td>
</tr>
<tr>
<td>POI</td>
<td>Point of Interest</td>
</tr>
<tr>
<td>REST</td>
<td>Representational State Transfer</td>
</tr>
<tr>
<td>RS</td>
<td>Remote Sensing</td>
</tr>
<tr>
<td>SAP</td>
<td>Selective Availability Policy</td>
</tr>
<tr>
<td>SPs</td>
<td>Static Points</td>
</tr>
<tr>
<td>STING</td>
<td>Statistical Information Grid</td>
</tr>
<tr>
<td>TCA</td>
<td>Triangular Clustering Algorithm</td>
</tr>
<tr>
<td>TIGER</td>
<td>Topologically Integrated Geographic Encoding and Referencing</td>
</tr>
<tr>
<td>TIN</td>
<td>Triangular Irregular Network</td>
</tr>
<tr>
<td>VGI</td>
<td>Volunteered Geographic Information</td>
</tr>
<tr>
<td>VTS</td>
<td>Vuong’s Test Statistic</td>
</tr>
<tr>
<td>WGS 84</td>
<td>World Geodetic System 1984</td>
</tr>
<tr>
<td>WPR</td>
<td>Weighted PageRank</td>
</tr>
</tbody>
</table>
List of figures

Figure 1.1: A model for the pre-industrial cities (Source: Sjoberg 1960) .......... 3
Figure 1.2: A model for the industrial cities (Source: Pacione 2009) .............. 3
Figure 1.3: Demonstration of the rank size distribution ..................................... 4
Figure 1.4: Organization map of a complex system adapted from Hiroki Sayama via Wikimedia Commons ........................................................................ 5
Figure 1.5: Illustration for the geographic entity representation .......................... 8
Figure 1.6: Overview of the structure of this thesis .............................................. 11
Figure 2.1: A fictive urban space, its (a) axial map and (b) connectivity graph (Source: Jiang 2009) ........................................................................ 20
Figure 2.2: (a) Small-world and (b) scale-free properties of complex network . 22
Figure 2.3: The structure of ABM ..................................................................... 23
Figure 3.1: The trend of OSM users and uploaded points (source: Openstreetmap.org) ................................................................................. 28
Figure 3.2: Overview of OSM workflow (source: Openstreetmap.org) .......... 29
Figure 3.3: Demonstration for the node, way, and relation ............................... 31
Figure 3.4: Demonstration of OSM database scheme in the node case .......... 32
Figure 3.5: Illustration of street nodes extraction ............................................. 34
Figure 3.6: Illustration of an OSM data snippet ............................................... 35
Figure 3.7: Overview of the street node extraction from the class perspective .. 35
Figure 3.8: Work flow of extracting street node ID information ....................... 36
Figure 3.9: Work flow of pinpointing street node location information .......... 37
Figure 3.10: Map of taxi GPS locations (blue dot) overlaid on street network (grey line) ..................................................................................... 38
Figure 3.11: Space-time view of the trajectory in terms of MPs and SPs ......... 39
Figure 3.12: Map of the domestic flight locations (blue dot) overlaid on the US boundaries (white line)..................................................................................................................40

Figure 3.13: The procedure of extracting flight dataset from the raw flight tracking dataset..........................................................................................................................41

Figure 3.14: Map of volunteer movement locations (blue dot) overlaid on the Sweden national highway network (white line).................................................................42

Figure 3.15: The procedure of extracting purposive locations dataset from the raw volunteer movement dataset..................................................................................................43

Figure 4.1: Overall structure of the methodologies ............................................45

Figure 4.2: Heavy-tailed distribution (red) and Gaussian distribution (blue)....47

Figure 4.3: Illustration of the process to calculate normalization constant $c$ of the lognormal distribution............................................................................................................49

Figure 4.4: Work flow for selecting the best model from the five heavy-tailed distributions ..........................................................................................................................53

Figure 4.5: The hierarchical structure of the US Census 2000 urban area population obtained by the head/tail division rule...............................................................54

Figure 4.6: Map of the hierarchical structure of the US Census 2000 urban area population obtained by the head/tail division rule...........................................................55

Figure 4.7: Demonstration of generating clusters with (a) TCA and (b) CCA .. 57

Figure 4.8: Illustration of the linear and power sprawl ruler to determine the sprawling status of cities .................................................................................................................60

Figure 4.9: Work flow to illustrate complex network analysis.........................61

Figure 4.10: Illustration of the movement of (a) random agent and (b) purposive agent on an axial map..................................................................................................................63

Figure 4.11: Illustration of agent mobility in graph............................................64

Figure 5.1: Work flow for validating Zipf’s law for all the US cities ............66

Figure 5.2: Work flow for uncovering scaling property of spatial units in urban systems .............................................................................................................................67

Figure 5.3: Work flow for measuring urban sprawl using street nodes.........69
Figure 5.4: Map of natural cities of (a) the US and (b) three European countries: sprawling (red), normal (yellow), and compact (green) ........................................... 70

Figure 5.5: Work flow for building and analyzing the US airport network ...... 71

Figure 5.6: Map of the categorization of the airports .................................. 72

Figure 5.7: Work flow for exploring the mobility patterns at the city level....... 74

Figure 5.8: Illustration of different moving behaviors in ABM ...................... 76

Figure 5.9: The route of the US airport network rendered by its geometric length .................................................................................................................... 77

Figure 5.10: Work flow for exploring human mobility patterns at the regional level ........................................................................................................... 79

Figure 5.11: Work flow for exploring human activity patterns ..................... 81
List of tables

Table 4.1: The five heavy-tailed distributions .......................................................... 48
Table 4.2: The normalized five heavy-tailed distributions ................................. 48
Table 4.3: Estimated values for the parameters of heavy-tailed distributions.... 50
Table 5.1: Power law scaling exponent for both natural cities and the US Census urban area ............................................................................................................ 66
Table 5.2: Results of scaling analysis of the three datasets .............................. 68
Table 5.3: R square values between the movement flows calculated by the footprint counters and the seven morphological metrics ............................................. 74
Table 5.4: R square values between the movement flows calculated by the gate counters and the seven morphological metrics .................................................... 75
Table 5.5: Model selection results for both observed flight length and simulated flight length ................................................................................................................. 78
Table 5.6: KS distance between observed flight lengths and simulated ones .... 78
Acknowledgements

I would like to say thanks to my primary supervisor Professor Bin Jiang. He gives me a chance to be employed as a research assistant in Hong Kong Polytechnic University for a half year, a position of research engineer in Future Position X (FPX) with one and a half years, and a position of research assistant in University of Gävle with one year. Meanwhile, he helps me to be registered as a PhD student in Royal Institute of Technology (KTH). His enthusiasm and concentration on research makes me comprehend the fundamental spirit in doing research. His encouragement and motivation on pursuit of cutting edge research in combination of Geographic Information System (GIS) and complex system helps me to begin the journey that leads to this thesis. Thanks are also given to his valuable comments on shaping this thesis.

I would like say thanks to my assistant supervisor Professor Yifang Ban who also gives me very important suggestions and comments during my study at KTH. I appreciate very much the comments and efforts from Dr. Lars Harrie who takes quality control of this thesis, which improves this thesis substantially. In particular, I should say thanks to Professor Itzhak Benenson who is willing to act as the opponent in the thesis defense and gives constructive comments on shaping the thesis. Besides, I am also grateful to the financial support from FPX and University of Gävle.

My thanks are also due to all the colleagues from both University of Gävle and KTH for their support during my study in Sweden. I would like to thank Dr. Martin Sjöström and Mr. James Morrison for their kindly help on polishing the language. And especially, I should offer my gratitude to my master’s supervisor Professor Zhongtang Fu who gave me valuable advice during my study in Sweden.

Last but not least, I dedicate this dissertation to my parents, my wife and my daughter. Without their understanding and support, it is impossible for me to accomplish the study. Particularly, I would like to express my deep sense of gratitude to my wife Meixue Ji, whose patience and encouragement give me great power to conquer the difficulties encountered in the PhD study.

Tao Jia
1. Introduction

1.1. Background
Given the rapid progression of urbanization, an increasing number of people are residing in cities or towns. In 2007, according to The Millennium Development Goals Report (United Nations 2007), approximately 50 percent of the world’s population lives in urban areas. Urban systems have undergone drastic changes during recent decades due to global (e.g. economical, technological, political) and local factors (e.g. regional tax regulation) (Pacione 2009). Cities have experienced unprecedented expansion in terms of their extent, which consequently engendered urban agglomerations. In practice, expansion blurs city boundaries and challenges the conventional administrative definition of urban extent. On the other hand, cities are more connected with one another through transportation networks and communication facilities as carriers of, for example, the flow of migration, goods, and information. These changes not only boost economic growth (Henderson 2003) but also lead to many challenging problems, such as the spread of infectious disease, traffic jams, environmental pollution, and urban sprawl (Knox and McCarthy 2005).

In essence, such changes indicate the complexity of urban systems because (1) the geographic space itself bears heterogeneous characteristic in terms of its individual size distribution (Paper VII, Jiang and Liu 2012, Paper II) or its topological relationship with others (Jiang 2007), which possibly implies the process of hierarchical spatial organization. Furthermore, (2) intensive contacts exist among people and their interactions with the surrounding geographic space, such as street network (Paper VI) or spatial points of interest (POI) (Paper V), which may represent a process of mutual reinforcement and adaptability. Lastly, (3) external economic factors and inner social development policies (Albeverio et al. 2008) may have a significant influence, which serves as incentives to shape urban systems. Therefore, to better understand and model the processes underlying urban systems, it is advisable to investigate them from a complex system perspective. In this respect, urban systems are regarded as self-organizing and display the properties of adaptability and transformability (Chen and Zhou 2008). This view is in line with others in the literature (Alberti et al. 2003, Riccardo et al. 2006, Kumar et al. 2007).

Along with the complexity of urban systems, the advancement in geographic information science in general and the revolution of strategies and methods of geographic data collection in particular have ensured the possibility of conducting empirical studies on urban systems. Similar to famous historical
experiments in natural science, such as the law of photoelectric effect by Albert Einstein, the geographic data not only play an important role in verifying the proposed theoretic viewpoint but also help to discover the novel hidden knowledge. However, the geographic knowledge may be biased if only a small sample of geographic data in terms of kilobytes is employed, which is partly because of the dynamic and complex properties of the geographic space. From this point, the revealed geographic knowledge would be valid and reliable if a significant amount of geographic data in terms of gigabytes is involved. In recent years, a paradigm shift has occurred in examining the issues related to urban systems from using a small-size dataset with a low computing power computer to using a large-size dataset with a high processing capacity workstation.

This new paradigm is called data intensive science, which is the fourth scientific paradigm following the first three paradigms in science, namely empirical description a thousand years ago, theoretical proposition a hundred years ago, and computational simulation a few decades ago (Gray and Szalay 2007, Bell et al. 2009). Initially, data intensive research was inspired by the availability of a flood of scientific data from sensor networks, satellites, telescopes, and detective instruments (Bell et al. 2009). In a sensor network, in which nodes are individual volunteers carrying GPS devices and edges are formed using Web 2.0 technology, the generated geographic data are called Volunteered Geographic Information (VGI, Goodchild 2007a). The VGI data are new in society, and their massive volume challenges scientists from the fields of geography, regional science, and ecology, although some innovations are already available, such as cloud computing and distributed digital archiving technologies. However, the opportunities are obvious because VGI permits taking action and being responsive to the growing complex urban problems.

The following parts introduce this study from four aspects. First, they describe the overall context or frame within which the study is carried out. Second, the strategy or the view adopted in this thesis is presented. Third, the development of geographic information science in general and the revolution of data collection method in particular are illustrated. Fourth, the emergence of VGI and data intensive research is described.

1.1.1. Urban systems
The earliest city can be traced back to 3500 BC in Mesopotamia along the rivers of Tigris and Euphrates (Pacione 2009). Although numerous debates exist around the origin of cities, such as hydraulic theory (Wittfogel 1957) and economic theory (Lynch 1960), an indubitable fact is that cities have emerged as a series of factors, including social, economic, and environmental, through
reinforcing interactions over a long period of time (Pacione 2009). In particular, the evolution of cities throughout the world can be roughly classified into three stages: pre-industrial cities, industrial cities, and post-industrial cities (Pacione 2009). The pre-industrial cities had a simple spatial pattern to reflect the urban society (see Figure 1.1), whereas the industrial cities were affected by the large gap between the wealthy and the excluded (see Figure 1.2). In this respect, the pre-industrial and industrial cities were primarily managed in a centralized manner, and a clear boundary existed between cities and rural areas.

Figure 1.1: A model for the pre-industrial cities (Source: Sjoberg 1960)

Figure 1.2: A model for the industrial cities (Source: Pacione 2009)

The post-industrial cities experienced the process of deconcentration through the restructuring of the economy from a manufacturing base to a service orientation, and they can be roughly characterized as fragmentation in terms of urban form (Berry and Garrison 1958). The resulting fragmented urban form parallels with segregation of the population, expansion of the urban infrastructure, and other side effects related to, such as, traffic jams, environmental pollution, and urban
sprawl. Moreover, these phenomena are interwoven to form a complex urban image with a heterogeneous population distribution, a complex traffic network, and a dynamic flow of human movement. To a certain extent, cities are more connected with one another than ever before, and a significant change in one city may probably affect the other cities within urban systems. In this respect, methods or strategies from a complex system perspective should be adopted to handle these issues.

Among the previously noted studies, an important topic on urban systems focuses on city size distribution, which is also known as the rank size regularity problem (Zipf 1949). This problem states that, in systems of cities (Berry 1964), the size of the r-ranked city should be expected to be the same as the size of the top-ranked city multiplied by 1/r (see Figure 1.3, where x-axis is rank and y-axis is size). This regularity not only indicates the heterogeneous hierarchy of city size during a period but also holds stably from one period to another, although a few cities can be expected to change positions (Berry 1967, Yeates 1990). Nonetheless, explanations and models around this regularity have also attracted extensive attention from both economists and statistical physicists, such as the general systems explanation based on innovation diffusion theory (1962) and Christaller’s central place theory (1933), the regional-historical explanation based on Innes’ staples theory (1920s) and economic base theory (1928), the autocatalytic processes (Malcai et al. 1999), the self-organized criticality (Bak et al. 1987), and even the stochastic process (Murtra and Sole 2010). Obviously, this research direction requires knowledge from multiple disciplines.

![Figure 1.3: Demonstration of the rank size distribution](image)

1.1.2. Complex system

Although still no consensus exists regarding the definition of a complex system, it has been used broadly in many fields in both the natural sciences and the social sciences. Typically, it is regarded as a system composed of interconnected individual parts that not only displays the emergent collective behavior not observed from individuals but also adapts to its environment (Miller and Page...
Different from a complicated system consisting of a large number of small components like a machine, a complex system is primarily characterized by the properties of decentralization, non-linear dynamics, emergence, and heterogeneity. As shown in Figure 1.4, we can observe several domains of a complex system, which provides many effective tools to examine the phenomena undertaken. These tools exist in the form of mathematics or computer simulations because no direct way exists to derive the emergent properties (Bossomaier and Green 2007). For example, the relationship between the spread of fire and the density of a forest can be reported using cellular automata with simple rules on each cell (Wilensky 1997b).

In general, two main approaches exist to cope with the complex system (Newman 2011). The first one is to construct an abstract mathematical model that is a simplified version of the complex reality. The approach mimics the most important character of the real system, and the solution through mathematic derivation can shed insight into the real system. For example, the logistic map (May 1976) is adopted as an effect equation to model the discrete time demographic dynamic. The other approach is to construct a simulation model from a bottom up perspective, which addresses the detailed interaction process among the individual parts. The approach captures and models the complex behavior of the real system, and hence the final observed emergence is
more realistic and comprehensive. A typical example of this approach is the agent-based modeling (ABM). Moreover, the two approaches can be applied to many real systems characterized by complex behavior, and classic examples include the financial and market system, the human brain or immune system, the ecosystem, and the urban systems composed of several components such as road traffic, population dynamics, and human mobility.

In particular, with respect to the issues in urban systems, the conventional thinking of reductionism and determinism (Fuchs 2003) is replaced by the complexity in terms of emergence and chaos. From the viewpoint of ecodynamics, Scandurra (1994) thought of a city as “an artificial ecosystem where there is a continuous exchange between living organisms and the physical man-made environment where they live”. In these respects, the properties or patterns resulting from the interaction between the individuals and the urban environment cannot be explained or predicted by the simple summation of the individuals’ properties. Instead, they should be treated as the emergence of collective behavior (Pulselli et al. 2006). Empirical studies aligned with this vision have focused on the exploitation of the locations of human movement to derive the hidden patterns of urban dynamics (Pulselli et al. 2006, Reades et al. 2009, Liu et al. 2009, Ahas et al. 2010).

Moreover, investigations into computer based simulation models have been accelerated and improved through the development and release of agent-based simulation softwares, such as NetLogo and SWARM. These models not only help to handle the issues related to the dynamics of urban land use change (Almeida et al. 2003), but are also targeted to simulate the urban growth involved with diverse ingredients, such as biophysical factors (Mundia and Murayama 2010). As important tools for exploring complex system, these software packages have eased the development of urban simulation models with the integration of human wisdom. Hence, they should have much wider applications in the domain of urban studies.

1.1.3. GIS and geographic data
In the beginning, GIS, or geographic information system, was defined from a toolbox-based perspective as an information system for capturing, storing, checking, managing, analyzing, and visualizing data that are spatially referenced to the earth (US Department of Environment 1987). One of the earliest prototypes of GIS can be traced back to the mid-1960s, when Canadian geographer Roger Tomlinson organized and developed the Canadian Geographic Information System (CGIS) to assist in regular procedures of land use management and resource monitoring. The main contribution of CGIS is that it achieved digital management of theme or layer based geographic entities,
which paved the way for advanced spatial analysis and modeling. Following CGIS, several GIS laboratories were established throughout the world, such as the Harvard Laboratory for Computer Graphics and Spatial Analysis which developed several general-purpose mapping softwares including SYMAP and GRID (Robertson, 1967), and the Geographic Information System Laboratory at the SUNY Buffalo. During the 1980s, GIS development flourished with the release of the professional GIS software ARC/INFO by ESRI and the establishment of several international GIS journals and conferences. This development continued into the 1990s, when Goodchild (1992) suggested the discipline of Geographic Information Science, which deals with fundamental issues in the development and use of GI-technology. During the recent decade, given the popularity of the personal computer and the advent of the Internet and mobile devices, the tendency is moving towards service-oriented GIS and volunteered GIS, as coined by Goodchild (2007a and 2007b).

An important issue in GIS is geographic data representation. That is, how geographic entities are to be modeled or represented. Geographic entities are three-dimensional in the real world, and they can be conceptualized as two types in GIS: field-based and object-based. The field-based entities represent a continuous surface of the underlying phenomenon like elevation or precipitation, and they can be represented as a grid model or a Triangular Irregular Network (TIN) model. On the other hand, the object-based entities refer to individual features on the ground like a house or a road, and they can be represented as a raster model or a vector-based point, line, and polygon model (see Figure 1.5). The above representation can better capture the geometric properties of each entity, but it lacks the ability to reveal the relationship among the entities. Here the street network is used as an example: the conventional line-based representation has an advantage in determining the length of each street, but it cannot tell which street is most connected with others. However, from a topological aspect, the street network can be modeled as an extremely useful graph in urban studies, in which the street is a node and the connection between two streets is an edge (Jiang and Claramunt 2004).
Moreover, geographic data play an important role in the application of GIS to other fields. Fortunately, the collection methods or strategies have changed drastically, enabling the acquisition of massive amounts of data in short periods of time. First, the traditional ground surveying employing the equipment of transits and theodolites is now complemented by the measurements using Global Positioning System (GPS), digital photogrammetry, and Light Detect and Ranging (LiDAR). The vector data acquired in this way not only have high accuracy but also have a short production period. Second, using the technique of Remote Sensing (RS), large amounts of image data can be used to acquire properties on geographical objects. The acquired raster data not only have a high spatial and temporal resolution but also can cover large areas, which is useful in land change detections. Although the data acquired with the above methods can be used for the investigation of some issues in urban systems, they are so expensive that only a few large institutions or groups can afford them.

1.1.4. VGI and data intensive research

With the advent of neogeography (Turner 2006) accompanied by a sharp decline in the price of mobile device with a GPS unit and the corresponding improvement on accuracy because of the removal of the SA policy on GPS signal by the US government in 2000, user generated geographic data about their movement trajectory are emerging and exploding. These kinds of data, known as VGI (Goodchild 2007a and 2007b), are voluntarily driven geographic information and are regarded as a rival for commercial giants, such as Google Map and Tele Atlas. Particularly, VGI is a wiki-based and voluntary-driven geographic information activity organized in a bottom up manner in which every volunteer collects geographic data, uploads the data to the central server, edits
the data collected or collected by others, and downloads the data for her or his own use, regardless of whether she or he is a GIS professional or amateur.

As one successful example of VGI, the OpenStreetMap (OSM, Haklay and Weber 2008) project is a wiki-like collaboration to create a free editable map of the world, using data from portable GPS devices, aerial photography, and other free sources. OSM data may have massive volume owing to the increasing number of contributors and the collective behavior among them. Another example of VGI, the GPS tracking data have also increased rapidly during the recent years given the increasing number of taxis, airplanes, or even people equipped with a GPS receiver. The rich and complex characteristics of VGI data collected at an individual level make it a very valuable data source for uncovering something hidden and unconventional within geographic systems, especially urban systems.

This notion is in line with the data intensive research, which is considered as the fourth paradigm in science and is characterized as being collaborative, networked and data driven (Gray and Szalay 2007, Bell et al. 2009). It is based on the availability of massive scientific data from satellites, sensor networks, and telescopes, and with the assumption that clues to new science can be distilled from them using specific tools (Frankel and Reid 2008). This trend has also stimulated a research interest in the geographic field, which is termed as the data intensive geospatial analysis (Jiang 2010). Compared to geographic analysis with small data sample, phenomena revealed by data intensive geospatial analysis may appear different and results may prove more reliable. For example, space seems to bear enormous heterogeneity when investigated with a massive volume of geographic data (Jiang 2010).

1.2. Thesis aims
Urban studies conventionally rely on geographic data which are produced and disseminated by national mapping agencies or gathered from questionnaires designed to acquire information on human activity. These data are not only insufficient for data intensive research on urban domain, but are also expensive and time-consuming. As Batty (2003) stated, “The city has become more complicated thanks to these new innovations, rather than less, and our ability to make sense of these changes in theoretical and scientific terms have not kept up…”. Hence, in order to enhance our understanding of urban systems, massive geographic data are required and needed to be thoroughly mined to gain sufficient insight.

The insights on urban systems are primarily related to two aspects: urban space
and the corresponding human dynamics. In this context, urban space is investigated from its scaling property and the subsequent implications, whereas human dynamics are examined by the emergent property from the interaction between individuals and their urban environment. However, to understand the two aspects empirically, it is advisable to tackle the following four problems: (1) how to construct the urban entity or its structure to represent the urban space requires deep investigation; (2) there is still an absence of a systematic way to detect the scaling property of the underlying phenomenon; (3) what is the implication of scaling property of urban systems, or can we exploit it to solve some urban problems, such as urban sprawl and transportation management; (4) what is the relationship between human dynamics and urban spatial structure, or can human movement behavior be determined by urban spatial structure.

To handle these problems, this thesis aims to: (A) propose a systematic strategy to detect the scaling property of the underlying phenomenon; devise new spatial point clustering method without parameter input; suggest a method to construct the hierarchical structure of urban space; devise simulation models to mimic the human dynamics in urban space; (B) obtain empirical knowledge on the two issues of urban systems through the application of these methods to extensive VGI datasets. Moreover, the aims for the entire work examined in this thesis are generalized as the following.

**AIM A:** Developing methods or tools in complex system for investigating the issues in urban systems.

**AIM B:** Applying the methods or tools to extensive VGI datasets for uncovering the knowledge related to the issues.

More specifically, the purpose of AIM A is to develop methods or tools in a complex system for investigating two aspects of urban systems, namely the problems related to the urban space, including scaling property, city size regularity, urban sprawl, and network structure; and the issues related to human dynamics in urban space, including human mobility patterns and human activity patterns. On another hand, the purpose of AIM B is to apply these methods or tools to the massive OSM dataset and the GPS tracking datasets to uncover knowledge about these issues in urban systems, which may be valuable for the research community as well as decision or policy makers.
1.3. Thesis structure

This thesis consists of six chapters which are based on the papers previously noted (list of papers). The first chapter presents a brief description of the background of this study, aims of this thesis, its structure and declaration. The next chapter conducts a literature review on the related theories and previous studies. Chapter 3 describes the fundamental concepts of OpenSteetMap (OSM, Haklay and Weber 2008) and GPS tracking datasets, and introduces the related procedure of data preprocessing. In the following two chapters, chapter 4 and chapter 5, we present the methodologies adopted in this study which relate to several disciplines and the subsequent findings or results around the topics in urban systems. Finally, chapter 6 presents a conclusion and points out areas for future work. Additional details about the structure of this thesis are shown in Figure 1.6.

![Figure 1.6: Overview of the structure of this thesis](image)

In chapter 2, this thesis conducts a literature review on related theories, methodologies, and previous studies. Specifically, the review begins with the methodologies in the field of data mining and knowledge discovery, and then focuses on spatial point clustering algorithms. Second, the theories and applications from complex system to the studies in urban systems are introduced. These theories and applications include the application of space syntax principle in street network, the model selection method in statistical physics, the method of agent-based modeling (ABM), the complex network analysis, and the information theory.
Chapter 3 describes the geographic data adopted in this thesis and introduces the data preprocessing procedures, respectively. The development of OSM is introduced in the beginning, and its components are then elaborated from a technical perspective. Subsequently, this chapter discusses the data organization in OSM and its applications, which state that using OSM data to study urban systems is one of the most flexible usages. Then, we present a procedure to extract the street nodes using the program developed by this study, which is adopted as one dataset in this thesis. The second part starts with a description of GPS tracking data in general. And then it demonstrates three specific types of datasets adopted in this thesis, namely taxi floating dataset, flight tracking dataset, and human movement dataset, and further elaborates on their preprocessing procedures.

Chapter 4 concentrates on the methodologies adopted in this thesis. Firstly, we present an overall view of the methodologies. Secondly, we propose the heavy-tailed distribution detection method, which automatically suggests the best fitted model from the five potential heavy-tailed models for the underlying dataset based on the knowledge of Maximum Likelihood Estimation (MLE, Shanbhag and Rao 2001), improved Kolmogorov-Smirnov Statistic (KSS, Clauset et al. 2009), and Vuong’s Test Statistic (VTS, Vuong 1989). Thirdly, we present the head/tail division rule which is used to reveal the hierarchical structure of the heavy-tailed distributed data. Fourthly, we put forward two spatial point clustering methods, namely the Triangle Clustering Algorithm (TCA) and the Entropy-based Hierarchical Clustering Algorithm (EHCA). After that, a strategy is shown that detects the phenomenon of urban sprawl, which is mainly based on the application of head/tail division rule. Sixthly, the chapter reports on the method of complex network analysis adopted in this thesis. Finally, we present three agent-based schemes to model human mobility in the underlying spatial structure.

Chapter 5 presents the results obtained from the applications of these methods. Firstly, the conclusion is made that Zipf’s law holds stably for all US cities. Secondly, the scaling property of spatial units is presented in terms of the clusters formed by street nodes, taxi static points, and photo locations. Thirdly, the US cities are classified according to their degree of sprawling: compact, normal, and sprawling. Fourthly, the chapter reports the findings on both the structure and the traffic patterns of the US airport network. Fifthly, at the city level, human mobility behavior is conjectured as shaped mainly by the underlying street network irrespective of random walk or purposive walk. Sixthly, at the country level, the exponential distribution of human flight length is presented and the same finding as before is suggested: that human mobility behavior is largely constrained by the underlying airport network. Seventhly, at
the regional level, the levy flight characteristic of human travel length is reported and the underlying hierarchical spatial structure is conjectured to determine human mobility behavior. Finally, the patterns of regularity, heterogeneity, and predictability of human activities in urban space are shown.

Chapter 6 summarizes the entire work of this thesis, presents the major findings and further highlights challenging problems for future work.

1.4. Thesis declaration

This thesis is primarily based on eight peer-reviewed papers around two issues in urban systems: urban space and the corresponding human dynamics. Professor Bin Jiang conceived of and proposed the main idea in Paper VI, whereas I carried out the work including data processing, data analysis, programming, and data statistics. Professor Bin Jiang also devised the research idea in Paper VII, and I am responsible for the rest of the work, including data processing, data analysis, programming, and data statistics. As for Paper VIII, I am responsible for the research idea, data processing, data analysis, programming, and data statistics, whereas Professor Bin Jiang provided constructive suggestions on shaping this work. Besides, I am responsible for the rest of the first author papers: Paper I, Paper II, Paper III, Paper IV and Paper V. Any problem found in this thesis is declared the sole responsibility of the thesis author.
2. Literature review

2.1. Knowledge discovery
Knowledge discovery is defined as the process of extracting unknown patterns from large volumes of data with novel methods or tools. A simple example to define knowledge discovery process can be shown as: given a dataset $D$, a language collection $L$, and a set of certainty measurements $M$, the knowledge or pattern can be defined as a statement $S$ in $L$ that represents relationships among data in $D$ with a certainty measurement, such that $S$ is simpler than the enumeration of all data in $D$ (Frawley et al. 1992). Pioneering applications of knowledge discovery are diverse and have emerged from many different fields. Retail giants have exploited it to identify potential products or market areas which can contribute to maximum profit from daily transaction database; the American Airlines has expanded its market by identifying potential travelers via its flyer database; and sanitation agencies have relied on patterns revealed in patient record database to identify periods of high incidences of epidemic disease in order to suggest corresponding strategies.

Knowledge discovery has also been applied to urban studies with the availability of volumes of geographic data, and the so-called geographic knowledge discovery. This trend has attracted the attention of scientists from multiple disciplines, including geography, urban planning, computer science, physics, and even mathematics. Therefore, for a better understanding of this thesis, this chapter is aimed to thoroughly review the literature on methods and applications of geographic knowledge discovery in urban studies.

The basic methods or tools in geographic knowledge discovery are first reviewed. These methods or tools are insufficient for gaining deep insights from complex urban systems. Hence, we subsequently concentrate on the literature regarding the methods or tools used for complex system, particularly their applications in mining geographic knowledge.

2.2. Geographic knowledge discovery
In this section, previous studies concerning geographic knowledge discovery in general are firstly reviewed. Secondly, one specific domain in this field, spatial point clustering, is presented in detail. Spatial point clustering has been demonstrated to have potential applications in urban studies.
2.2.1. Challenges and strategies

In recent decades, investigations around geographic knowledge discovery have experienced a revival with the availability of rich geographic data and high processing capacity of computers. Geographic data not only expand drastically in terms of their volume, but also extend broadly in terms of their sources. For example, apart from conventional geographic data, such as satellite imagery or topographic map, the human trajectory data can be acquired with the use of devices such as GPS logger or cell phone. However, conventional methods or tools in geographic information system targeting initially for data-poor and computation-poor analysis may not be adapted well to this new situation, just as Miller and Han (2001) stated that “the traditional spatial analytical techniques cannot easily discover new and unexpected patterns, trends and relationships that can be hidden deep within very large and diverse geographic datasets.”

In this respect, new methods or techniques have to be devised or developed to meet the requirement of data-rich and computation-rich era. These new methods or techniques embedded in the geographic knowledge discovery process aim to distill useful information from massive geographic dataset and further derive the hidden knowledge or facts about underlying phenomenon which is instructive to both researchers and decision makers. Furthermore, the uncovered knowledge or facts should be valid, novel, useful, and finally interpretable by humans (Fayyad et al. 1996). Valid means that the knowledge should be general enough so that it can be reproduced from other datasets. Novel means the knowledge should be interesting and non-trivial. Useful means that the knowledge should benefit researchers and decision makers alike. Whereas interpretable means that the knowledge should be simple and understandable by humans (Miller and Han 2001). Procedurally and basically, four steps are involved in the process of geographic knowledge discovery, including geographic data selection to focus on the study area, data processing to remove noise or outliers, data mining to employ spatial or statistical method for uncovering hidden information, and data interpretation to extract knowledge or fact from the reported information by humans.

Commonly used strategies or methods in the process of geographic knowledge discovery include but are not limited to the following: (1) Spatial clustering, which deals with the classification of geographic objects according to their spatial proximity and similarity of attributes. Useful methods include k-means clustering (Lloyd 1982) and support vector machine clustering (Abe 2010). (2) Spatial autocorrelation analysis, which concerns the tendency or relationship among the properties of geographic objects in space. Typical methods include Moran’s I (Moran 1950), Getis and Ord G statistic (Getis and Ord 1992), and network mixing pattern (Newman 2003). (3) Spatial outlier analysis, which concentrates on the finding of anomalous geographic objects bearing distinct
properties from the others. For example, Ng (2001) studied and compared several outlier detection methods including distance-based, distribution-based, noise-based, and depth-based, whereas recently Alvera-Azcárate et al. (2012) examined outliers in satellite data using spatial coherence. (4) Spatial trends detection, which intends to “find the patterns of change with respect to the neighborhood of some spatial objects” (Miller and Han 2001). Applications of this method can be found in areas of urban temperature (Fujibe 2009) and rainfall prediction (Caloiiero et al. 2011). (5) Spatio-temporal patterns detection, which employs several data mining tools or methods to detect the novel patterns of geographic objects in space and time. For example, Openshaw (1994) proposed the Geographic Analysis Machine (GAM) to explore the space-time-attribute patterns in geographic data.

2.2.2. Spatial point clustering

Although the methods adopted in the process of geographic knowledge discovery have been briefly described in the previous part, it is imperative to shed light into the spatial point clustering method which not only has immense applications but also constitutes the main methodology of this study. Note that points are two-dimensional unless other is stated. Generally speaking, the algorithms around spatial point clustering can be categorized into four classes: partition-based, density-based, grid-based, and hierarchy-based (Han et al. 2001). In the following, we further elaborate on previous studies with respect to the four general categories.

Partition-based algorithm agglomerates spatial points into a specified k clusters according to similarity of attributes or proximity in space. One of the most popular ones is the k-means clustering (Lloyd 1982) method, which iteratively assigns each point to the nearest cluster based on the criterion that the total sum of the variance within the clusters reaches the minimum whereas the total sum of the variance among the clusters comes to the maximum. Recently, an improvement on this algorithm was proposed by Arthur and Vassilvitskii (2007), which is known as k-means++ targeting as an approximation algorithm for the NP-hard k-means problem. The improvement in k-means++ is to perform a procedure with the intuition of spreading k initial cluster centers away from each other before the application of k-means method. On another hand, from the perspective of statistical probability, the Expectation Maximization (EM, Dempster et al. 1977) method also belongs to this type. The idea of EM method is to assume that each cluster follows a predefined probability distribution and then to maximize the log likelihood of the mixture probability model (weighted sum of each cluster model). Traditionally, the Gaussian model is adopted as the cluster model since the Gaussian mixture model can be applied to approximate any density distributions (Scott 1992).
Density-based algorithm relies on the idea that clusters contain dense regions separated by sparse regions in data space (Han et al. 2001). A classic method of this type is the Density-Based Spatial Clustering of Applications with noise method (DBSCA) proposed by Ester et al. (1996). This method generates clusters with the gradual combination of high density regions. A total of two parameters are involved in this algorithm, namely radius and minimum number of points. Recently, a variant of this algorithm is the method proposed by Rozenfeld et al. (2008), which is called City Clustering Algorithm (CCA) applied to population sites in the US (Rozenfeld et al. 2008). This algorithm agglomerates the points with spatial distance to each other less than a specified radius as a cluster, and hence distance effect is involved whereas local density effect is removed in this simple method. Note that an important step in the above methods is to search the neighborhood of each spatial point, and it is time-consuming if massive geographic data are involved. In this respect, basic spatial data indexing technique, such as grid or R-Tree indexing (Guttman 1984), should be employed before application of the clustering method.

Next, grid-based algorithm aims at decomposing the data space into several layers composed of increasing grid size or resolution. From this point, all of the operations for clustering are performed on the grid structure (Han et al. 2001). Wang et al. (1997) proposed the Statistical Information Grid (STING) approach to cluster spatial data objects. This method divides the space area into rectangular cells with respect to a particular level, and the cell in a high level is further partitioned into four children cells in a low level. According to the statistic information precomputed for each cell in each level, a pruning process starting from the highest level to the lowest level is gradually applied to the cells in each level to reserve the cells satisfying clustering requirement (Wang et al. 1997). Similar to the above method, hierarchy-based algorithm also decomposes the spatial objects into smaller subsets in an increasing level to form a dendrogram (Han et al. 2001). An important representative of this method, coined CHAMELEON, was proposed by Karypis et al. (1999). This technique firstly constructs a k-nearest neighbor graph of spatial points, then performs a partition process recursively, and finally carries out a merging process on the subgraph based on certain criterions (Karypis et al.1999).

2.3. Theories of complex system for geographic knowledge discovery
This section begins with an overall review in the state-of-art of complex system. Followed by a general review of five specific theoretical tools or methods and their applications in uncovering the pattern or knowledge in complex system. These methods include space syntax, complex network analysis, agent-based modeling (ABM), scaling analysis, and information theory. The first two
methods put an emphasis on deriving the structural measurements of spatial environment. The third method focuses on simulating the dynamic measurements due to the interactions with the spatial environment and the last two methods are intended to reveal the overall patterns of complex system.

2.3.1. Complex system
Complex system is a relatively new field associated with many disciplines (Newman 2011). It can be characterized by interconnected individual parts communicating with their environment to display the emergent collective behavior not observed from the individuals (Miller and Page 2007). Typical examples of complex system can be found in areas of financial markets (May et al. 2008), ecological system (Levin 1998, May et al. 2008), biological system including the cells or the human brain (Edelman and Gally 2001), swarm behavior like the flocking of birds or schooling of fish (Reynolds 1987), transportation system (Chowdhury et al. 2000), and urban systems (Batty et al. 1999, Barabási et al. 2002). Specifically, the phenomena in urban systems have exhibited particular interest to researchers in many disciplines including geography, economy, and even physics. For example, in the early 1960s, Jacobs (1961) published a book entitled “The Death and Life of Great American Cities”, which was pioneering in that it suggested the study of cities from a complex system perspective. Moreover, Bettencourt et al. (2007) addressed the scaling analysis in urban environment, and Batty (2008a) strengthened the bottom-up evolution of cities with scaling property.

However, in practice, it is difficult to identify the characteristics of complex system using the conventional tools or techniques. Hence, a lot of theoretical tools or methods have been developed to mimic and uncover the patterns underlying complex system, such as complex network analysis, information theory, cellular automata, agent-based modeling (ABM), and even the techniques from statistical physics and space syntax. Generally, the aforementioned tools or methods can be further categorized as three classes: computer simulation method, structure or system analysis method, and mathematical modeling method. Note that we do not intend to make an exact categorization of these methods, but instead to make an approximate classification so that it becomes much clearer to the methods adopted in this thesis. Therefore, computer simulation methods include ABM; structure or system analysis methods contain space syntax theory, complex network analysis, and information theory; and mathematical modeling methods include the technique of scaling analysis. The following parts will give a detailed review on the applications of these methods in the literature.
2.3.2. Space syntax

Space syntax is a set of principles and techniques for studying the spatial configurations of geographic entities which can range from a single room in a building to a whole city. Spatial configuration means “the relations which take account of other relations” (Hiller 2007) in the space syntax theory. Spatial configuration is also an effective method to examine problems related to urban morphology, which is an interdisciplinary research on urban form, its formation, and its linkage to social-economic forces over time. In this respect, to assess the spatial configuration, two steps are typically involved, namely spatial decomposition and morphological measurement. The first step refers to the process of representing the underlying space with small spatial units. Conventional ways of spatial decomposition include convex space popularized by Peponis and his colleagues (1997), axial map devised by Hillier and Hanson (1984), and isovist fields proposed by Batty (2001). Among these spatial decompositions, the axial map seems to be the most flexible and effective way of representing both small-scale space and large-scale space (see Figure 2.1 for an axial map of a small-scale urban space). On the other hand, morphological measurement can be derived from the connectivity graph shown in Figure 2.1c. Suggested measurements are connectivity, local integration, global integration, betweenness, and closeness.

![Figure 2.1: A fictive urban space, its (a) axial map and (b) connectivity graph (Source: Jiang 2009)](image)

Applications of space syntax into the urban studies have flourished in recent decades. For example, Jiang and Claramunt (2002) have integrated the concepts and functions of space syntax into geographic information system (GIS), which is extended as a GIS package coined Axwoman (Jiang et al. 2000). Jiang and Claramunt (2004) carried out a topological analysis of the street networks in three cities, and they found that the connectivity graph exhibited the small-world property but displayed no scale-free property. Moreover, Jiang (2008) proposed the 20/80 rule of street network, namely 20% of streets are highly connected whereas 80% of them are less connected. Apart from the structural analysis of
urban environment, some other studies have focused on predicting traffic flow based on the morphological measurements, for instance, which measurement can better capture traffic flow. Specifically, some empirical studies conducted by researchers in space syntax community claimed that traffic flow can be better captured by local integration (Hillier et al. 1993, Penn et al. 1998), whereas Jiang (2009) suggested that Weighted PageRank (WPR, Page and Brin 1998) with an optimal damping factor value is better at predicting traffic flow.

2.3.3. Complex network analysis
Networks have been the focus of mathematicians for a long history since the foundation of graph theory because of the solution of the seven bridges of Königsberg by Swiss mathematician Leonhard Euler in 1735. Graph theory has been successfully applied in many fields and solved a lot of practical problems, such as the routing problem with shortest path (Sniedovich 2006) and the map coloring problem using minimum number of colors but with different colors in neighbors (Boccaletti et al. 2006). Recently, with the advancement in several fields, such as telecommunications, the Internet, and biology, the “systems” seem to be more complex than ever before. For example, in urban systems, the urban entities are more connected with each other due to urbanization in terms of increasing population migration and their intensive interaction with the urban environment. On the other hand, society has also witnessed an increase in computing capacity and the availability of large diverse network datasets. In these respects, to solve the complex problems in society, there is a revival of interest on the studies of complex network and its applications (Boccaletti et al. 2006). This trend is further stimulated by the surprising findings on the small-world (Watts and Strogatz 1998) and scale-free (Barabási and Albert 1999) properties of complex network (see Figure 2.2).
Figure 2.2: (a) Small-world and (b) scale-free properties of complex network

As another powerful tool for investigating spatial structure, complex network analysis has been applied to a various fields in urban systems. In the transportation domain, many kinds of transportation networks have been examined. For example, Jiang (2007) examined the topological structure of street networks of 40 cities in the US and found the scale-free and small-world properties of these networks. Chan et al. (2011) investigated the geometrical properties of the road networks of the 20 largest cities in Germany, and they reported a large degree of similarity of the small-scale geometric features among these road networks. The maritime transportation network was also explored from a structural perspective with the small-world and scale-free properties (Hu and Zhu 2009) and from a traffic perspective to verify the gravity models of ship movement for a better understanding of global trade and bioinvasion (Kaluza et al. 2010).

Besides, Soh et al. (2010) studied the public transportation network in Singapore, and they stressed that dynamic analysis enriched the knowledge obtained from conventional topological analysis. As a crucial transportation infrastructure, the airport network plays an important role in urban systems in terms of economic development, infectious disease control, and human
migration management. Therefore, it has attracted extensive attention from a wide range of studies (Li and Cai 2004, Guimerà et al. 2005, Guida and Funaro 2007, Bagler 2008, Wang et al. 2010). These studies are based on the datasets provided by air service departments, which is composed of flight records with origin and destination airports. Generally speaking, in their airport network, a flight record corresponds to an edge and the individual airport or aggregated airport within a city is considered as a node.

2.3.4. Agent-based modeling (ABM)
Apart from the aforementioned tools used for structural analysis of spatial environment, ABM is an effective tool to derive the simulated measurements from the interaction between the individual and the environment in complex system. Avoiding the rules in a top-down centralized way, it assigns simple rules in a bottom-up decentralized way to each agent in terms of its communication with others and the environment (Macy and Willer 2002). Although the rules followed by each agent are simple, the collective emergent pattern can be expected through the interactions at the individual level. Structurally, ABM is composed of three components (Macal and North 2010): the agent, the underlying environment, and the rule set. Here, the agent can be summarized as two types (Jiang 1999), namely reactive agent and cognitive agent. Reactive agent only receives information from the underlying environment or other agents and performs the corresponding action, whereas cognitive agent can sense its environment, react to it, learn this experience, and tend towards a purposely new destination. The underlying environment can be of any type, such as a land use map or a street network. The rule set should be as simple as possible. Moreover, a complete view of this structure can be seen in Figure 2.3.

![Figure 2.3: The structure of ABM](image)
In addition, there have been several software packages developed and released: (1) Swarm (available at: swarm.org) developed by the Santa Fe Institute aims to model complex system, and the modeling language is objective C or Java; (2) StarLogo developed by MIT (available at: education.mit.edu/projects/starlogo-tng) is relatively easy to manage since it adopts the simple logo language; and (3) NetLogo software package (available at: ccl.northwestern.edu/netlogo/) released by the University of Northwestern is a totally cross platform system written in Java and easy to learn. The releases of these softwares have stimulated the applications of ABM to many fields.

In ecology, based on the idea of Boids simulation (Reynolds 1987), Wilensky (1998) explored collective flocking behavior emerged from the simple rules followed by individual birds, namely alignment, separation, and cohesion. In human society, using the concept of racial segregation proposed by the Laureate Schelling (1971), Wilensky (1997a) implemented a model and obtained the pattern that individuals prefer to live near others like themselves. In urban systems, Jiang (1999) examined human movement behavior in a virtual urban environment, and he concluded that random pedestrian movement rates had a significant correlation with local integration measurement in space syntax. Fontaine and Rounsevell (2009) presented a framework to model the dynamics of future housing demand in a polycentric region. They reported a heterogeneous spatial pattern that some locations would grow faster than the others in terms of urban development.

2.3.5. Scaling analysis

To further reveal the hidden pattern in complex system, the derived variables from the aforementioned tools (such as the structural measurements from space syntax and complex network analysis, the simulated measurements from ABM, etc.) are examined through a scaling analysis. It is one of the most significant methods of complex system theory, and it is typically used to identify the power law distribution of the size of the underlying variables in complex system (Newman 2011). Strictly speaking, scaling is the property of power law distribution and it means that the distribution maintains the shape whenever the measured size is multiplied by a constant (rescaled). Identification of power law distribution of the measured unit is not an easy job because of the diversity of the underlying phenomena and the limitations of the available statistical techniques (Clauset et al. 2009).

Substantial efforts have been made to explore new methods. The most common fitting technique can be traced back to the Italian economist Pareto’s work at the end of 19th century (Arnold 1983), when he simply made a histogram of wealth
data and further took a logarithm on the two axes. If a straight line was observed, then the power law distribution of the data might be concluded and the scaling exponent could be calculated by measuring the slope of the straight line or using the method of least square linear regression. This method is straightforward but suffers from several problems as elaborated by Clauset et al. (2009). Following the above method is the adoption of Maximum Likelihood Estimation (MLE, Shanbhag and Rao 2001) for calculating the model parameters and Goodness of Fit (GoF) for testing the model significance (Wasserman 2003, Resnick 2006, Clauset et al. 2009). The GoF test can be accomplished by different techniques, such as the Anderson-Darling test (Anderson and Darling 1952), the Shapiro-Wilk test (Shapiro and Wilk 1965), or the Kolmogorov-Smirnov (KS) test (Shanbhag and Rao 2001). Importantly, conventional usage of the significance test is based on comparison of the test statistic with the critical value obtained from a statistical table, but the strategy devised by Clauset et al. (2009) is based on simulations and coined as improved KS test without a statistical table.

Observations of scaling property of phenomena in complex system have exploded in recent decades with the availability of massive data (Newman 2005). For example, Zipf found that the frequency of words used in a piece of English text followed a power law distribution which is also known as Zipf’s law; the number of citations received by a scientific paper was reported to obey the power law distribution (Price 1965, Redner 1988); the number of hits received by a website was found to exhibit the scaling property (Adamic and Huberman 2000); the magnitude of earthquakes that occurred in California between January 1910 and May 1992 have also been reported to follow a power law distribution (Newman 2005); and the scaling property was also found in many other fields, such as the intensity of wars (Roberts and Turcotte 1998), the wealth of society (Levy and Solomon 1997), and the population of cities (Rosen and Resnick 1980, Gabaix 1999, Soo 2005).

Along with the empirical findings, the literature has also welcomed the explanations or origins on the scaling property. In the beginning, scientists tried to postulate that a single mathematical mechanism might be responsible for all these scaling behaviors. However, this naïve thinking failed when faced the diverse phenomena (Newman 2005). Thus, there is still a long way to go to form a united theory, and consequently different mechanisms have been explored. Simon (1955) was probably the first man who proposed the mechanism of “rich get richer” or “preferential attachment”, which was later elaborated by Barabási and Albert (1999) in detail. Bak et al. (1987) examined this regularity from a perspective of self-organized criticality, which adopted the cellular automata model of “self-organizing sand pile”. Furthermore, Krugman (1996) attributed it to the positive feedback mechanism in the economic models, Malcai et al.
explained it using the autocatalytic processes, and Murtra and Sole (2010) tried to give a solution based on stochastic process.

2.3.6. Information theory

The information theory was initially developed by the American mathematician Shannon (1948) in his work around signal processing operations and reliably storing communication data. Since its inception, it has been applied in a wide variety of areas and has benefited many sub-fields including data compression, channel coding, algorithmic complexity theory, and information measurement (Reza 1994). The last sub-field is also called entropy and is an effective tool to measure the information contained in a system. As Newman (2011) stated, “a pattern is precisely recognizable as a pattern because its information content is low. For instance, there is little information in a periodically repeating sequence of symbols, numbers, colors, etc. If we can accurately predict the next symbol in a sequence then that symbol contains little information since we knew what it was going to be before we saw it”. In this respect, the entropy not only provides us with a quantitative way to characterize the patterns exhibited in complex system, but also benefits us by being able to predict patterns changing with time.

Applications of information entropy in Geographic Information Science (GIS) as well as in urban systems have been a hot topic in recent decades. Bjorke and Myklebust (1997) proposed an entropy-based algorithm to eliminate the point features in the process of cartographic generalization. Li and Huang (2002) examined the information content of a map and devised three new information measurements, namely metric information, topological information, and thematic information, based on the Voronoi diagram of map symbols. In addition to applications in cartography, information entropy has also been adopted in urban studies. For example, Batty (1977) devised a new measurement called spatial entropy mathematically and elaborated on its applications on one- and-two dimensional aggregation problems. Yeh and Li (2001) investigated the measurement of monitoring urban sprawl in developing countries using the concept of entropy. Recently, Batty (2008b) treated cities as evolving systems with structure emerging in a bottom-up way and further investigated the derivation of more realistic urban models based on the framework of entropy maximization. Lastly, Song et al. (2010) examined to what degree human behavior is predictable. They concluded a 93% of potential predictability of user mobility through an investigation on the entropy of trajectories of individual mobile phone users.
3. The geographic data and its preprocessing

3.1. OpenStreetMap (OSM)
This section starts by introducing the development of OSM and elaborates on its components from a technical perspective. This section further discusses the data organization and management in OSM and presents some successful applications based on OSM source followed by stressing that using OSM data in urban studies is beneficiary. Finally, based on the organization of OSM data, the procedure to extract the street nodes is demonstrated.

3.1.1. Development and components of OSM
OSM, started at the University College London in 2004 by Steve Coast (Chilton 2009), is a project aimed specifically at creating and providing free geographic data such as street maps to anyone. The project started because (1) most maps you think of as free actually have legal or technical restrictions on their use, holding back people from using them in creative, productive or unexpected ways, and (2) the high availability of GPS receivers and the improved accuracy of GPS signals motivates people to participate.

For many years, mapping products were solely produced by professional surveyors and cartographers in national mapping agencies. These geographical data are usually collected by national mapping agencies in a number of ways, such as geodetic survey, aerial photogrammetry, remote sensing (RS), global positioning system (GPS) or other advanced techniques such as light detection and ranging (LiDAR), etc. Because of the vast territory and the relatively long period of data production, the currency of fundamental geographic database is far lagging the market needs. It is even worse that people have to pay the map agency to obtain geographic data. This situation is pretty true in the US and European countries. In the US, you could obtain the basic geographic data through the US Census Bureau’s TIGER (Topologically Integrated Geographic Encoding and Referencing) Line program, but the details are relatively coarse and the currency is comparatively low because of the high cost of mapping. For example, these data do not include green space, landmarks, and the like (Haklay and Weber 2008), and road features do not contain some attribute information such as if a road is one-lane or two-lane. Among European countries, accurate geographic data is expensive for the public, small businesses, or community organizations (Haklay and Weber 2008). Besides, it is the same case in commercial companies which rely on the charge to earn a living, such as
NAVTEQ and Tele Atlas. Therefore, motivated by the belief to provide detailed geographic data free of charge, volunteers participate in the OSM project.

Along with the demand for free data among users, there is another important factor in promoting this project, namely the technical improvement in GPS receivers and the improvement in accuracy of GPS measurements. The technical improvement in GPS receivers and the advance in telecommunications have witnessed a sharp decline in cost of mobile phones with GPS functionality. It is said that a GPS receiver only cost about 100 US dollars in 2001 (Hightower and Borriello 2001), with price still going down. Apart from the decline in price of GPS receivers, accuracy have also been greatly improved from 100 meters to about 6 meters in normal conditions with the removal of the Selective Availability (SA) policy by the US government (Haklay and Weber 2008). Thus, the technological change over the past years has made it possible for anyone to engage in the work of mapping their surrounding environment.

With the motivations from the two aspects noted previously and by adopting a wiki-based (Web 2.0) technology, OSM is evolving rapidly in terms of both the number of registered users and the volume of geographic information. It is reported that OSM had about 33,000 registered users in 2008 (Haklay and Weber 2008). It is estimated that this number is now almost 20 times more than that. Figure 3.1 shows that the number of registered users has increased steeply.
from 2009 to 2012 together with the volume of geographic data contributed. Online statistics show that the total number of uploaded GPS points is almost 2,861,039,307, the total number of nodes is almost 1,453,960,761, the total number of ways is almost 135,554,648, and the total number of relations is almost 1,403,807 (available at: openstreetmap.org/stats/data_stats.html).

![Figure 3.2: Overview of OSM workflow (source: Openstreetmap.org)](image)

From a technological perspective, OSM is mainly composed of four parts, namely backend database, map editing software, map rendering software, and the main website (source: wiki.openstreetmap.org/wiki/Component_Overview). (1) The backend database used in OSM is PostgreSQL, which is primarily composed of two types of tables, master table and current table. The master table holds all the previous edit versions, whereas the current table only stores the latest version for drawing the map. (2) Map editing software includes Java OSM Editor (JOSM), Potlatch, Merkaator, etc. JOSM is targeted for experienced contributors, and it resembles a traditional GIS package and provides advanced functions, such as linking OSM features to photos or audios, supporting data conflicting solution, and supporting function extension by third party plug-ins (Haklay and Weber 2008). Compared with JOSM, Potlatch is a light-weight flash based online editor and could be used for beginners. (3) Map rendering software is Mapnik or Tiles@Home. Mapnik is an open source library for generating high-quality map tiles, and it uses a weekly database dump as the data source for the rendering of map tiles of all zoom levels (Haklay and Weber 2008). However, it is a time consuming job if considering the weekly update of OSM data of the entire world. On the contrary, the Tiles@Home rendering system is developed as a distributed rendering system, which comprises a central coordinating server and over 100 software clients. This distributed rendering
system renders number of map tiles in near real time. (4) The last component is the front website which is based on open source AJAX library OpenLayers to display the “slippy map”. This website not only provides interactivity with the map like zoom and pan functionality, but also enables users to search for places based on external web service. Besides, the communication between the database and the software is based on the Representational State Transfer (REST) web service interface, which accepts or outputs data in the format of OSM XML.

As shown in Figure 3.2, the entire workflow for the OSM project is that the users firstly collect the data in a GPX format or similar. They then edit these geographic data either using the online Potlatch editing software or the advanced offline JOSM editing software. After the editing, they upload the data to the database via the REST web service interface. The database will then send the geographic data every one week to the rendering system either Mapnik or Tiles@Home. Once the rendering system receives the OSM data, it will output the geographic data into map tiles with the specified map style and store them on the tile server. Finally the “slippy map” will be displayed on the website using the OpenLayers library.

3.1.2. Format and usage of OSM data
To understand OSM data structure, this part examines the data organization and management in OSM, and presents some applications based on OSM source. OSM adopts a tag based or key-value pair based data model to manage the spatial data and attribute data seamlessly. This novel data model is based on taxonomy of entities of the complex real world, and it allows the abstract or virtual decomposition of the real-world. Interestingly, this virtual decomposition among all the real-world entities mimics the advantages of the object-oriented model without its complicated implementation. Basically, a geographic entity is composed of two elements in OSM dataset, namely location and tag. Location is the geometric information of a real-world entity, whereas tag is its property information. Based on the tag scheme, the entities of the real-world are virtually decomposed to form a hierarchical structure, such as relationship of the super-class and the sub-class in the object-oriented model. This can be explored from two perspectives. (1) If we look at it from the key perspective, then the set of entities with the same key belong to the same class and the hierarchy of the key leads to the hierarchy of super-class and the sub-class of entities. For example, the entities with a key equal to highway form a highway class and the entities with a key equal to way form the way class. Obviously, the latter class is the super-class of the former one. (2) If we observe it from the value perspective, then the set of entities with the same key but different values constitute different sub-classes of the same super-class. For example, the entities with a key equal to
way but value equal to highway constitute a highway sub-class of way class, whereas the entities with a key equal to way but value equal to railway constitute a railway sub-class of the way class. In other words, OSM stores the atomic type of the real-world entity as its basic element, and this basic element cannot be further decomposed.

In practice, OSM data structure is composed of three fundamental elements, namely node, way, and relation (see Figure 3.3). Node is the dot that is used to mark a specific location (such as a post box) or for comprising way. For example, a node maybe the spatial point of interest (POI) such as a city mall. Way is an ordered list of nodes which are referenced by IDs and displayed as connecting line segments. Way includes two sub types: non-closed way and closed way. Non-closed way is used to describe a line feature like a road, whereas closed way is used to describe an area feature like a park or a lake. The last element is relation which is an ordered list of ways associated with each other. Relation references the ways by their IDs, and it is used to describe the role of each way played in complex physical feature, such as cycling route, turn restriction, and area with holes (source: wiki.openstreetmap.org/wiki/Beginners_Guide_1.3).

![Figure 3.3: Demonstration for the node, way, and relation](image)

Based on the OSM data structure, the entities of the real-world are stored in geodata tables of OSM’s central database which is implemented by MySQL. The database scheme is designed to support wiki behavior, such as versioning and rollbacks, and keeps copies of modified or deleted features indefinitely (Haklay and Weber 2008). As elaborated previously, the geodata table is composed of master table and current table (see Figure 3.4). The master table stores the entire historical edition (versions) of the features, whereas the current table only stores the latest data for drawing purpose. Besides, the geometry of all geographical features are organized as points with latitude and longitude coordinates combined with user name and timestamp information, and are stored in the nodes table (see Figure 3.4b or 3.4d). The linear feature and simple polygon feature are represented as the way element by reference to a list of
ordered nodes, and complex polygon features are represented as relation element by reference to a list of ordered ways. Apart from the shape information of real world features stored as tables in OSM database, the attributes of these features are stored in the k and v field of the tag tables, such as node-tag table, way-tag-table and relation-tag table (see Figure 3.4a or 3.4c). Thus, this organization only stores location information in nodes table, which avoids the storage redundancy of location data and keeps the topology of entities consistent at updates or modifications.

Figure 3.4: Demonstration of OSM database scheme in the node case

Moreover, the OSM database does not impose any limitation on the tags contributed by users. This benefits users by allowing them to implement their own tagging scheme freely, which leads to diversity of the database content. However, it is suggested that keeping the conventional key-value pairs is a good habit to contribute, because inappropriate free tagging may result in a burden on data management and usage. Note that beginners could get some help on the tagging skill from the editing software. After all, this tagging scheme, which is increasingly being incorporated into a complex taxonomy of real world feature classes and objects, is a core part of the OSM initiative and is community-driven (Haklay and Weber 2008).

After discussing the inner structure and organization of OSM database, it is time now to move to the increasing usages or applications based on OSM source. The applications based on OSM source could be divided into the following categories according to the specific task. (1) The easiest option to use OSM map is to download the map image specified in a certain region for the purpose of
making static print media. (2) A relative easy way to use OSM source is to embed the map on the website using iFrame-based solution provided by Google Map, but it requires a little technical knowledge to accomplish your task. (3) Deploying the map to the website using JavaScript is a bit difficult as it requires some web programming knowledge. However, this is a relatively flexible usage of OSM source because it allows one to implement some interactive functionality with the base map. For example the frequently used OpenLayers which is an open source JavaScript library. This library is used to display the “slippy map” on the OSM website, because it implements a JavaScript API for building rich web-based geographic information applications. Based on OpenLayers, one can display map tiles and markers loaded from any resource, and finally you could mash up a website providing the location related service, such as a routing service. (4) The most flexible way to use OSM source perhaps is to create a complete new map to serve your needs. This is the most difficult application compared with the above ones, because it requires more specialized technical knowledge and additional software or hardware to support the map. A typical example employing this usage is the cycling map created from OSM data (source: opencyclemap.org), which shows the international cycling networks. In this respect, the whole map has to be rendered to exaggerate some important geographic features or to eliminate some unimportant geographic features, because different geographic features have different effects on the cycling.

Finally, (5) extracting the raw geographic data from OSM database for knowledge discovery is also one of the most difficult applications. This kind of geographic data are appropriate for studies in urban systems, because they contain a rich set of attribute information and a huge amount of geographic data. However, because of the complexity and multifaceted of OSM source, emphasis should be focused on only these geographic entities related to studies. For example, a street network researcher would like to study the highway features in OSM source, an urban planning researcher would be more likely to extract the urban extent or similar. Therefore, to meet our special needs, we extract street nodes as one of our datasets to study the urban related issues in this thesis. The process to extract the street nodes is presented in the following section.

### 3.1.3. Street nodes extraction

Street junctions are the intersection points in the street networks, whereas street nodes defined in this thesis not only include street junctions, but also include the start and end of the streets. However, street nodes used here should avoid the following cases. Firstly, the intersection point of a highway bridge and the underlying highway should be excluded, which is shown as the red dot in Figure 3.5. Secondly, the link way which links a lower level street (secondary highway) to a higher level street (primary highway) should be avoided, as you can see the
green dashed line shown in Figure 3.5. Finally, the intersection points (traffic circle) of the roundabout that directs the road traffic to one direction around a central island and the connected highways are dropped, which are shown as blue rectangle in Figure 3.5. Thus, the street nodes in this thesis are a bit different from the street junctions, and they are usually the turning points for residents accessing their houses (as shown the yellow triangle in Figure 3.5). And hence, we believe that street nodes reflect a true image of human activities.

Figure 3.5: Illustration of street nodes extraction

OSM data downloaded with the REST web service interface is an xml file (see Figure 3.6) which is sequentially organized. Basically, the node information is placed at the beginning of the file, and then follows the way information, and then the relation information. Thus, reading the file sequentially is a necessity to extract the street nodes. This job is accomplished in our program by taking two steps, namely extracting street node ID information and pinpointing location information, respectively.

The program extracts the node information of the way satisfying our needs (highway), and writes the node information (node ID, node degree and end-point-flag) to several files in an ascending sequence according to node ID. This procedure could be demonstrated in both Figure 3.7 and Figure 3.8. The ExtractStreetnodeID method of StreetNodeExtractor class in Figure 3.7 is implemented to execute this task in the program, whereas the detailed work flow of this process can be observed from the chart in Figure 3.8. Note that we only write the node information of a maximum number of ways (known as threshold in our program) to a file. In other words, a new file will be created once the number of ways exceed the maximum number. This is a method of batch processing. It delineates the entire data file into several pieces, and each piece could be treated the same way. Therefore, this batch processing allows the
program to handle the data file regardless of its size, and it is especially useful for the data file with extremely large size, say OSM source of the entire world.

Figure 3.6: Illustration of an OSM data snippet

Figure 3.7: Overview of the street node extraction from the class perspective

The second step is to pinpoint the location information based on the node files generated in the first step. As a batch procession, our program reads each data file sequentially and the entire data files synchronously. In other words, once a node is read from one data file as the current node, the program searches the candidates in other data files with the same ID as the one of the current node. This is done by moving the pointers of each data file to the record with the node ID equal to or greater than the one of current node synchronously. After that, the
degree value of the current node can be obtained by summing the degree values of all the candidates (the same way for obtaining end-points-flag value). In this way, the three cases aforementioned are avoided by combining the node information with the highway filter process. This step is demonstrated in both Figure 3.7 and Figure 3.9.

To this point, a batch procedure to extract street nodes as one of our datasets in this thesis has been demonstrated. Importantly, compared with traditional line-line intersection based method to extract street nodes, our method demonstrates two advantages: (1) the size of data file can be extremely large; (2) the time efficiency is drastically improved.

Figure 3.8: Work flow of extracting street node ID information
3.2. GPS tracking datasets

This section intends to introduce the other three GPS tracking datasets employed in this thesis: taxi floating dataset, flight tracking dataset, and volunteer movement dataset. The similarity among these datasets is that they are all GPS-based trajectory reflecting the human mobility behavior. The difference is that they are collected in different ways, namely taxi-based, flight-carried and car-held, and with different geographic extents, namely city-level, country-level and region-level. The similarities and differences in the three datasets thereby provide us with a new perspective to explore issues on human dynamics in urban space.
3.2.1. Taxi floating dataset

Taxi floating dataset was obtained by GPS receivers installed on 54 taxicabs of a local company within the period of October 2007 and from the period of January 2008 to May 2008. It covers the entire study area including four cities or towns in the middle of Sweden: Gävle, Sandviken, Storvik, and Hofors. Each taxicab records its location every 10 seconds, and hence totally it is a massive spatial dataset containing around 59,983,958 records. Apart from the spatiotemporal data in terms of longitude, latitude, and time (where latitude and longitude are WGS 84 geo-referenced and time is when the location is captured), there are customer data covering the same periods, which record the usage information of the current taxi, such as car ID, the time to pick up a customer, the time to drop off a customer, and the status of the current taxi. Totally, there are 166,679 records stored in the customer data table during these periods. Moreover, the map in Figure 3.10 shows a one month taxi floating dataset together with the local street network. However, for the issue of business sensitivity, all other information is skipped except the longitude, latitude, time, and car ID (which is replaced by an arbitrary number specified from 1 to 54).

![Figure 3.10: Map of taxi GPS locations (blue dot) overlaid on street network (grey line)](image)

The raw taxi floating data is further preprocessed to derive the static points (SPs) dataset to study one particular topic of this thesis, namely human activity patterns. By SPs, we mean the locations with zero speed along the tracking trajectory. A trajectory is a path that a moving object follows in space as a function of time. Following this definition, a taxicab trajectory is defined as a path that the taxicab moves in space as a function of time, and it is
mathematically denoted as, $\text{Traj}_{T_1}^{T_2}(i) = \{\text{Loc}(i,t) \mid T_1 \leq t \leq T_2\}$. Based on the taxicab trajectory definition, we split a trajectory into two parts: moving points (MPs) and static points (SPs), denoted by green small dots and red large dots respectively in Figure 3.11. In this context, we assume that SPs are more meaningful than MPs. This is because on one hand they are highly associated with many human events, for instance, traffic jams, customer dropping off (in), and parking. On the other hand, SPs of different trajectories are more likely to be clustered together. These clusters are also associated with spatial points of interest (POI). Therefore, from these perspectives, SPs are considered as a better proxy to investigate the patterns of human activities within urban domain. Generally speaking, SPs are extracted according to the following formula,

$$\text{SPs}_{T_1}^{T_2}(i) = \{\text{Loc}(i,t) \mid T_1 \leq t \leq T_2 \text{ and } \text{Loc}(i,t) = \text{Loc}(i,t+r)\} \quad (3.1)$$

where $r$ is the GPS sampling interval in seconds. By applying the above rule to the raw taxi floating dataset, we obtain a total number of 10,067,674 SPs for one month (from October 1st 2007 to October 28th 2007).

3.2.2. Flight tracking dataset

Flight tracking dataset adopted in this study contains a total number of 7,685,948 records in ASCII files gathered during a period from the 8th August to 18th August in 2010. Each record indicates the current status of a particular en-route flight within and to/from US, and it is updated once every five minutes. The updated information includes eight data elements: current longitude, current latitude, current height, current timestamp, current speed, airline code and flight number, aircraft type code, and domestic or international flag. Note that the
location information with three-dimensions: longitude, latitude, and altitude, is referenced to the World Geodetic System 1984 (WGS 84). Moreover, this massive dataset covers a vast geographic space including the US mainland, Alaska, Hawaii Island and Puerto Rico Island and contains the positional information of all commercial flights. To get an impression of this massive geographic dataset, we present the visualization of domestic flight locations in Figure 3.12. From these points, it offers us a good dataset to explore two issues: the patterns of human mobility at the country level and the relationship between air traffic and the underlying airport network structure.

![Figure 3.12: Map of the domestic flight locations (blue dot) overlaid on the US boundaries (white line)](image)

To cope with the two topics proposed in this study, flight dataset is obtained from the raw dataset using two sub-procedures. The first sub-procedure is data cleaning and data pruning to remove the records that do not meet the study requirement. After a detailed check on the en-route flight records, noisy records are identified as the ones with information loss (which indicates one of the three dimensional location is lost) or information duplication (which indicates two consecutive records have the same time stamp). The noisy records are the invalid records that need to be eliminated. For the first case, we simply drop the records, and we merge the duplicate records into one in the second case. With the two steps of records cleaning, we are left with a series of valid records. The valid records are further filtered to obtain the domestic valid records based on the flag information of domestic or international to meet the requirement of this study. Finally, after the process of data pruning, there are a total number of 4,823,658 domestic valid records left for further investigation.
The second sub-procedure is data extraction to acquire flight dataset from the valid domestic records dataset. The valid domestic records are categorized into individual flight record set which is composed of a series of valid domestic records with the same airline code and flight number. Each flight record set is labeled as regular or irregular based on the statistical information of the time interval of two consecutive flight records, namely $\tau_i = t_i - t_{i-1}$. The regular flight record set is chopped sequentially into several parts, and the flight record in each part is connected sequentially to form a regular flight. Thus, the regular flight repeats several times during the observation period, which reflects the flight routine between the origin and destination airports. On the other hand, the irregular flight record set does not need to be broken apart, and each flight record is connected consecutively to form one single irregular flight, which usually reflects the private or official usage. Figure 3.13 illustrates the procedure of extracting flight dataset from the raw flight tracking dataset. Finally, after the process of data extraction, a total number of 205,662 flights are obtained including both regular and irregular ones which constitute the building blocks of the two research topics.

3.2.3. Volunteer movement dataset
Volunteer movement dataset was obtained by 89 BT-338X devices recording the daily movement of volunteers living in three sites of Borlänge, Sweden (Domnarvet, Kvarnsveden, and StoraTuna) during four periods. BT-338X is a combined GPS receiver and data logger with a Bluetooth interface. Each volunteer attached it to her/his car for around one week. The whole dataset is collected from the period of March 29 to May 15 in 2011. Totally, the volunteer movement dataset includes 258 GPS logger files corresponding to 262,021 movement recordings of all volunteers. Besides, it should be noted that each GPS logger file contains the movement information of one volunteer, and also that each record in the GPS logger file includes the information when the GPS signal is received every 30 seconds. This information is composed of the location in terms of longitude (x) and latitude (y), the time (t) and the speed (s). The longitude and latitude are referenced to the World Geodetic System 1984.
WGS 84) and have an accuracy of 5 meters according to the BT-338X user manual. Although most of volunteers are residents of Borlänge County, the spatial extent of their movement covers more than half of the entire Sweden territory (see Figure 3.14). In this respect, the volunteer movement dataset reflects the picture of human mobility with accurate positions in a regional-level geographical space.

Nonetheless, we cannot directly rely on the raw volunteer movement dataset because of noisy records. Similar to the previous two datasets, data cleaning and data extraction procedures are developed. The first procedure is to remove the noisy records. These are defined as the ones with incorrect time format and are simply removed. The second procedure is to extract the purposive locations characterized by drastic change in time, distance, or angle. This change can be further identified from two criteria, namely a large time interval and a tortuous behavior. The former occurs when both time interval (t) between two consecutive locations exceeding the time threshold (ΔT) and their distance (d) being less than the expected distance (t*s, where s is the estimated speed). The latter denotes that the angle (φ) formed by three successive locations is less than the angle threshold (Δφ). For a better understanding of the two rules, the procedure of extracting purposive locations from the raw volunteer movement data...
dataset is illustrated in Figure 3.15. By applying this procedure to volunteer movement dataset, a total number of 15,423 purposive locations are obtained. Furthermore, the purposive locations are most likely to occur when people go to the office or shopping mall where no GPS signal can be received or when they are outdoors with GPS signal but wander around something interesting.

Figure 3.15: The procedure of extracting purposive locations dataset from the raw volunteer movement dataset
4. Methodologies

4.1. Overall structure

This chapter presents the methods adopted in this thesis to explore urban space and the corresponding human dynamics from the perspective of complex system (see Figure 4.1). Methods or tools in complex system provide us with a brand new way to look at these issues, which is different from the conventional way of a centralized top-down analysis with a small sample of geographic data. On one hand, the two issues investigated in this thesis cover a wide range of topics, including city size distribution, scaling property of spatial units, measurement of urban sprawl, structure of transportation network, patterns of human mobility, and patterns of human activity. On the other hand, methods or tools to cope with these topics can be classified into six categories according to their roles. However, in practice, one or several methods are needed to investigate one topic.

First, the mathematical method is introduced on how to detect the heavy-tailed distribution of the size or property of geographic entities in urban space. The heavy-tailed distribution detection method requires the knowledge of both

![Figure 4.1: Overall structure of the methodologies](image-url)
mathematical statistics and probability, and it offers us a quantitative way to understand the heterogeneous or scaling structure of the underlying systems. In this respect, it serves as the backbone in the investigations of all the topics covered in this thesis.

Second, the classification strategy, namely head/tail division rule, is demonstrated. This method is directly based on the prerequisite of a heavy-tailed distribution of the examined geographic entities as its name suggests. This method was firstly proposed by Jiang and Liu (2012) based on the work of the mathematical detection of the heavy-tailed distribution by Jia and Jiang (Paper III). It clearly demarcates the entire dataset into two parts, namely a minority important head and a majority trivial tail. In this respect, it recommends a new way to look at the hierarchy of the urban space.

Third, a necessity in examining the issues of urban space is to establish the underlying spatial structure, and here the spatial structure may refer to natural cities in the systems of cities, natural street network in the city systems, airport network in the transportation systems, etc. However, a further necessity before constructing the spatial structure is to derive the spatial units, and they can be the cities, the streets, the airports, or the city districts. In this respect, two methods are demonstrated, including Triangle Clustering Algorithm (TCA, Paper III) and Entropy-based Hierarchical Clustering Algorithm (EHCA, Paper II), to cluster individual spatial locations to form the spatial units for the further examination of urban space.

Fourth, based on constructed spatial units, the method of sprawl ruler is shown, which employs the techniques of regression curve and head/tail division rule. This method demarcates the entire cities into three categories: compact, normal, and sprawl, which sets a uniform standard for cross comparison of sprawling levels across an entire country. From this point, it tackles the urban sprawl from a new perspective which is different from the traditional time series comparison way. Besides, it witnesses the successful application of the head/tail division rule on the topic of urban sprawl.

Fifth, based on constructed spatial structure, the method of complex network analysis is illustrated to explore its hidden structural properties. This method has been successfully applied to many kinds of complex networks, such as neural network, social network, etc. Specifically, here it is employed to uncover the structural properties of airport network which might play a non-trivial role on local urban development and national transportation systems. In the following, this method will be explained explicitly in terms of its network measurements and its capacity in predicting traffic flow.
In the last, to explore human dynamics in urban space, the method of agent-based modeling (ABM) is elaborated on uncovering the collective emergent patterns from the interaction between the agents and the underlying spatial structure. This method is composed of three components, namely the agents, the rules, and the environment. To examine the different patterns emerging from the agents bearing different behaviors and interacting with different underlying spatial structures, three types of ABM are proposed, including random/purposive agents in the street network, agents with five different behaviors in the airport network, and random/preferential agents in the hierarchical graph of spatial regions.

### 4.2. Heavy-tailed distribution detection

Scaling property of the observed phenomenon, or power law distribution of its size, is the most typical characteristic of complex system. On one hand, this characteristic indicates the heterogeneity of the underlying system, and specifically it means that there is a relatively high probability of the occurrence of an extremely large event compared with the normal or Gaussian distribution (see Figure 4.2). In this context, it is meaningless to use a typical value to describe the entire system which is completely different from the case of normal distribution. For example, we can say that the average male height in Sweden is around 180 centimeter, but it may be nonsense if we say that the average population of cities in Sweden is around 20,000 because this average value does not capture the most important information of the system. On the other hand, the underlying mechanism to generate this distribution also displays extensive implications, and it can be explained from several aspects, such as stochastic process, self organization criticality, economic positive feedback, etc. From these points, it sounds meaningful and valuable if we can empirically verify the power law distribution of the observed phenomenon.

![Figure 4.2: Heavy-tailed distribution (red) and Gaussian distribution (blue)](image)

However, in practice, a majority of phenomena observed in nature or in the social world do not strictly follow a power law distribution, and instead they can
be approximated by power law with cutoff distribution, exponential distribution, stretched exponential distribution, or lognormal distribution because of the finite size effect. Moreover, the latter distributions also indicate the heterogeneous characteristic of the underlying phenomenon, and they belong to the family of heavy-tailed distribution which is again the obvious characteristic of complex system. In this respect, to embrace the diversity of the phenomena, the scaling analysis is extended to identify the heavy-tailed distribution of the measured unit in this thesis. In other words, the aim is at selecting the best fitted model from the underlying five heavy-tailed distributions. To facilitate our analysis, detailed information on the five models are listed below in Table 4.1 (Note that the model is considered as the continuous case in this study if without specific notation).

Table 4.1: The five heavy-tailed distributions

<table>
<thead>
<tr>
<th>Model</th>
<th>Probability density function $P(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power law</td>
<td>$P(x) = cx^{-\alpha}, (\alpha &gt; 1)$</td>
</tr>
<tr>
<td>Lognormal</td>
<td>$P(x) = (c/(x\sigma \sqrt{2\pi})) \exp(-(\ln x - \mu)^2 / 2\sigma^2)$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$P(x) = ce^{-\lambda x}, (\lambda &gt; 0)$</td>
</tr>
<tr>
<td>Power law with exponential cutoff</td>
<td>$P(x) = cx^{-\alpha} e^{-\lambda x}$</td>
</tr>
<tr>
<td>Stretched exponential</td>
<td>$P(x) = cx^{\beta-1} e^{-\lambda x}, (\lambda &gt; 0, \beta &gt; 0)$</td>
</tr>
</tbody>
</table>

The entire procedure to select the best model can be summarized in a few steps. The first step is the normalization process, which is meant to calculate normalization constant $c$ of the probability density function $P(x)$. It is well known that normalization constant $c$ is used to ensure the whole probability density function $P(x)$ can be integrated to 1 with respect to the $x$ domain. Thus, to calculate normalization constant $c$, a lower bound $x_{\text{min}}$ is imposed for all the five distributions. The integration calculation to obtain normalization constant $c$ seems a bit complicated, and Table 4.2 shows the results for the normalized five heavy-tailed distributions. However, for a better understanding on this process, Figure 4.3 gives a detailed process to calculate normalization constant $c$ of the lognormal distribution.

Table 4.2: The normalized five heavy-tailed distributions

<table>
<thead>
<tr>
<th>Model</th>
<th>Normalized $P(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power law</td>
<td>$P(x) = ((\alpha-1)/x_{\text{min}})(x/x_{\text{min}})^{-\alpha}$</td>
</tr>
<tr>
<td>Lognormal</td>
<td>$P(x) = \sqrt{\frac{2}{\pi\sigma}} \text{erfc}(\frac{\ln x_{\text{min}} - \mu}{\sqrt{2\sigma^2}})^{\frac{1}{2}} x \exp(-(\ln x - \mu)^2 / 2\sigma^2)$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$P(x) = \lambda e^{\lambda x_{\text{min}}} e^{-\lambda x}$</td>
</tr>
<tr>
<td>Power law with exponential cutoff</td>
<td>$P(x) = \gamma^{1-\alpha} \Gamma(1-\alpha)\Gamma(1-\alpha, x_{\text{min}}) x^{-\alpha} e^{-\lambda x}$</td>
</tr>
<tr>
<td>Stretched exponential</td>
<td>$P(x) = \beta \lambda x^{\beta-1} e^{\lambda (x_{\text{min}} - x)^\theta}$</td>
</tr>
</tbody>
</table>
The second step is to calculate model parameters using the method of Maximum Likelihood Estimation (MLE, Shanbhag and Rao 2001). The basic thinking in MLE is to make the probability that the dataset is drawn from the underlying model reach the maximum. Suppose we have an empirical dataset containing n observations $x_i (i = 1, 2, \ldots, n)$, the probability or the likelihood that this dataset is drawn from the model $f(parameters, x)$ is noted as,

$$L(parameters) = \prod_{i=1}^{n} f(parameters, x_i)$$

And then, we take the logarithm of the two sides of this equation,

$$Log(L(parameters)) = \sum_{i=1}^{n} Log(f(parameters, x_i))$$

We now wish to find the values of model parameters so that the likelihood reaches the maximum, and it can be achieved by solving equations of the derivative of $Log(L(parameters))$ with respect to each parameter,

$$dLog(L(parameters))/dparameter_i = 0, i = 1, 2, \ldots$$
In some cases, it is simple to obtain the solution of these equations, for example the power law distribution and the exponential distribution. However, in other cases, it is hard to directly solve these equations because of the difficulty in deriving their inverse functions. And hence, we adopt the technique provided by R to obtain the estimated values of model parameters. R is a collection of powerful packages that is used for data manipulation, calculation, and graphical display. It can be used as a statistical system which implements a rich set of classic and modern statistical techniques (Venables and Ripley 2010). The problem here to obtain the estimated values of model parameters can be treated as the problem of nonlinear optimization. In R, many functions can be used to carry out a minimization of the objective function with the constraints on a range of parameters, such as the function of \( \text{nlm} \) or \( \text{ConstrOptim} \). Therefore, the estimated values for model parameters of the five heavy-tailed distributions can be solved directly or indirectly, and they are listed in Table 4.3.

Table 4.3: Estimated values for the parameters of heavy-tailed distributions

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Estimated values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power law</td>
<td>direct</td>
<td>( \hat{\alpha} = 1 + n [\sum_{i=1}^{n} (\log(x_i/x_{\text{min}}))]^{-1} )</td>
</tr>
<tr>
<td>Lognormal</td>
<td>indirect</td>
<td>( \mu, \sigma = \text{nlm}(f = -\log(L(\mu, \sigma, x_i)), p = (\mu, \sigma)) )</td>
</tr>
<tr>
<td>Exponential</td>
<td>direct</td>
<td>( \hat{\lambda} = n [\sum_{i=1}^{n} x_i - nx_{\text{min}}]^{-1} )</td>
</tr>
<tr>
<td>Power law with exponential</td>
<td>indirect</td>
<td>( \hat{\alpha}, \hat{\gamma} = \text{ConstrOptim}(f = -\log(L(\alpha, \gamma, x_i)), p = (\alpha, \gamma), \text{const}) )</td>
</tr>
<tr>
<td>Stretched exponential</td>
<td>indirect</td>
<td>( \hat{\lambda}, \hat{\beta} = \text{nlm}(f = -\log(L(\lambda, \beta, x_i)), p = (\lambda, \beta)) )</td>
</tr>
</tbody>
</table>

A bit to note is that in practice the lowest bound \( x_{\text{min}} \) is typically unknown, and hence we need to choose it as accurate as possible, because model parameter values are very sensitive to the choosing \( x_{\text{min}} \) according to the research conducted by Clauset et al. (2009). Many methods of choosing \( x_{\text{min}} \) can be exploited, and here this study adopts an objective method based on minimizing the distance between the heavy-tailed model and the empirical distribution model (Clauset et al. 2009). By minimizing the distance, the fitted model will be more similar to the empirical data distribution, and thus obtain a more reliable \( x_{\text{min}} \). Generally speaking, this is an iterative process based on Kolmogorov-Smirnov Statistic (KSS, Shanbhag and Rao 2001) which calculates the maximum distance between the cumulative distribution function (CDF) of the data and the fitted heavy-tailed model of the data for each given \( x_{\text{min}} \). After iterating every possible \( x_{\text{min}} \), the \( x_{\text{min}} \) giving the minimum distance is the most reliable one. This process could be described by the following pseudo-code,
The third step is to verify the plausibility of the heavy-tailed model, and the idea is to perform a Goodness-of-Fit (GoF) test based on improved KSS. This method was proposed by Clauset et al. (2009) to verify the power law distribution of the empirical data, but here we extend this method to the other four members of the heavy-tailed distributions. The fundamental idea in GoF test is to calculate the \( P \) value that quantifies the plausibility of the fitted model. Specifically, the whole procedure is described as following: (1) obtaining the estimated \( x_{\text{min}} \) and parameter values of the fitted heavy-tailed model for the empirical data using the above function Estimating_\( x_{\text{min}} \), and then conducting a GoF test to get the maximum distance \( max_d \); (2) generating \( n \) synthetic datasets with a large sampling number using the empirical fitted heavy-tailed model; (3) employing the function Estimating_\( x_{\text{min}} \) again to each synthetic dataset and obtaining the corresponding synthetic fitted heavy-tailed model for its own; (4) a GoF test is employed on each synthetic dataset to get the corresponding \( max_{di} \); (5) \( P \) value can be calculated as the fraction of the number of \( max_{di} \) greater than \( max_d \) as shown in the following equation,

\[
P = \frac{\text{the number of } max_{di} \text{ greater than } max_d}{n}
\]

(4.1)

where \( n \) denotes the total number of synthetic datasets, and \( max_d \) is KSS for the empirical dataset, and \( max_{di} \) is KSS for each synthetic dataset. Basically, the larger the \( P \) value, the more reliable the fitted model is. However, different \( P \) values should be adopted in different studies. Here, this thesis adopts 0.05 as the \( P \) value to judge if the fitted model is a plausible one. It is important to note that the method adopted in this study to detect the heavy-tailed distribution with KS statistic and the corresponding GoF test is a recent advance in the literature. Among the five potential models, our strategy is to select the one with the largest \( P \) value as the best model, but the selection procedure may not stop here.
if one or more models have the same high $P$ value. With respect to the latter case, we have to move to the fourth step to conduct a Vuong’s test (1989).

The fourth step is to conduct a Vuong’s test (1989) on the potential models in an everyone-to-everyone manner. Vuong’s test is based on the Log Likelihood Ratio (LLR), and its aim is to provide a more convincing statistical way to tell the better fitted model from two competing models. LLR is the logarithm ratio value of the probability that the data is drawn from one model to the probability that the data is drawn from the other one. It tells how the data is more likely under one model than the other. It is typically denoted as,

$$LLR(1,2) = \sum_{i=1}^{n} \log(f_1(x_i)) - \log(f_2(x_i))$$  \hspace{1cm} (4.2)

Where $f_1$ and $f_2$ are the probability density functions of the two competing models respectively. Obviously, if the value of LLR is greater than zero then we can say that $f_1$ is better fitted than $f_2$, and vice versa. However, this is not a statistical significant test, and the result cannot be trusted. According to the central limit theorem, the test static LLR should follow a Gaussian distribution. If the Gaussian distribution has a mean value of zero then the value of LLR is just the result of a statistical fluctuation, and hence both of the two distributions are equally far from the true distribution of empirical data. On another hand, if the Gaussian distribution has a mean value not equal to zero then we can safely rely on the sign of LLR to make judgment. Vuong’s test is solely based on the above thinking and it proposes the hypothesis as following:

$H_0$: LLR follows a Gaussian distribution with mean equal to zero  
$H_1$: LLR follows a Gaussian distribution with mean not equal to zero

In this respect, Vuong’s test static (VTS) is constructed as,

$$VTS = \sqrt{n} \cdot \frac{\text{mean}(LLR_i)}{\text{std}(LLR_i)}$$  \hspace{1cm} (4.3)

Therefore, if the $p$ value obtained from Gaussian table (VTS) is very small, let’s say less than 0.1, then we can trust the value of LLR to make the selection, and vice versa. By applying Vuong’s test to every pair from the five models, we can get a five-by-five matrix with each cell indicating the value of LLR and the corresponding $p$ value. Thereafter we can select the best model from the matrix with the largest value of LLR and small value of $p$. Note that this matrix may fail if every cell displays a large value of $p$, in which case we cannot select the best model from the underlying five models although this may happen rarely.
summary, we present the four steps to select the best heavy-tailed model in Figure 4.4.

![Flowchart showing the four steps to select the best model](image)

Figure 4.4: Work flow for selecting the best model from the five heavy-tailed distributions

### 4.3. Head/tail division rule

Most of the phenomena in urban space or in geographic space tend to exhibit a pattern that can be described as: there are far more trivial small things than the important large ones. For example, there are much less connected streets than the well-connected ones; there are far more small cities than the large ones, etc. This kind of description is totally different from the phenomenon that can be approximated by a Gaussian distribution, where the entire phenomenon can be characterized by the mean value or the average. For example, it is reasonable if we say the average male height in Sweden is around 180 cm, whereas it is nonsense if we say the average population of cities in Sweden is about 20,000.

Besides, this kind of pattern gives us a deep impression of the underlying phenomenon and this impression is not the common things but the few important ones. From the perspective of psychology, we are more likely to talk about the richest man in the world, like Bill Gates, when the topic is around personal wealth; and we are more attracted by the landmark of a city, like the Forbidden City in Beijing, when we are traveling around the world. Why are we
not intoxicated by the common things when we are confronting the phenomenon in urban space or in geographic space? This is probably because of their heterogeneous pattern or scaling property in terms of heavy-tailed distribution. The heavy-tailed distribution does not reflect its intrinsic nature from the typical average but from the hierarchical structure, namely the important ones dominating the head whereas the trivial ones staying at the tail.

Therefore, an important question is raised: How can we build the hierarchical structure of the phenomenon with a heavy-tailed distribution? After extensive empirical experiments on the size or the property of geographic entities, Jiang and Liu (2012) proposed a rule to demarcate the heavy-tailed distributed entities into a minority head part and a majority tail part using the mean value of the size or the property. They named this regularity as head/tail division rule and formally stated it as “Given a variable X, if its values x follow a heavy-tailed distribution, then the mean (m) of the values can divide all the values into two parts: a high percentage in the tail, and a low percentage in the head” (Jiang and Liu 2012). Note that the mean value of the heavy-tailed distributed variable is nonsense in itself, for example, it contains little information to explain the whole, but it is used as a measure point to partition the whole into two parts: less than or equal to 20% of the head and greater than or equal to 80% of the tail. Furthermore, this rule can be employed recursively to build the hierarchical structure of the geographical phenomenon. This procedure is shown in the following Figure 4.5, which is based on the population of the US Census 2000 urban area (available at: census.gov/geo/www/cob/ua2000.html).

Figure 4.5: The hierarchical structure of the US Census 2000 urban area population obtained by the head/tail division rule
In addition, the hierarchical structure of the US Census 2000 urban area is visualized in terms of their population in Figure 4.6, which adopts the same coloring scheme as shown in Figure 4.5. From this map, it can be observed that there are around 90% of urban areas with the population less than the mean value 59,611, and this majority tail part categorized as level one (colored as blue) is scattered around the entire US territory which probably tells little information of US cities. Then, repeated application of the head/tail division rule is employed to the minority head part, and it consequently classifies the urban areas into level two, level three, level four, and level five. Importantly, urban areas in level five (colored as red) are the top five largest cities in the US, namely New York, Philadelphia, Chicago, Los Angeles, and Miami, and they may be the most popular cities when we chat about cities in the US. In this respect, a new classification method has been proposed based on the recursive application of head/tail division rule (Jiang 2012), but a prerequisite is that the underlying phenomenon should follow a heavy-tailed distribution.

Figure 4.6: Map of the hierarchical structure of the US Census 2000 urban area population obtained by the head/tail division rule

4.4. Spatial point clustering method

The methods or tools on how to detect the heavy-tailed distribution have been elaborated in terms of the size or property of spatial units in urban space or in geographic space. However, another important issue is how to define the spatial units, which is associated with the problem of the representation of the geographic space. Generally speaking, there are two types of spatial units: administrative spatial unit and inferred spatial unit. The administrative spatial
unit includes the US Census 2000 urban area or the named street (Jiang and Claramunt 2004), whereas the inferred spatial unit includes the axial line (Hillier 1996), the city block (Lämmer et al. 2006), or the natural cities (Paper I). The inferred spatial units seem more flexible and objective than the administrative ones, although most of the empirical studies have found the scaling property of the geographic space irrespective of the definition. In this respect, coupled by the massive point dataset investigated in this thesis, two point clustering methods are proposed to derive the inferred spatial units.

The first method is adapted from the City Clustering Algorithm (CCA) proposed by Rozenfeld et al. (2008), and it is named as Triangle Clustering Algorithm (TCA, Paper III). As its name suggests, this method derives the clusters through the agglomeration of neighboring triangles, and it may have wide applications on the heterogeneously spatial distributed points. Compared with CCA, TCA should have two advantages: one is that it constructs the Triangular Irregular Network (TIN) to model the relationship of spatial points (see Figure 4.7a), which is different from the way that CCA builds the neighborhood by laying a circle on each point (see Figure 4.7b); and the other one is that TCA adopts the head/tail division rule to obtain the threshold value used in the agglomeration process while it avoids the arbitrary radius selection in the traditional CCA process. Moreover, a spatial autocorrelation factor is considered to remove noisy disruption in forming the clusters. This spatial autocorrelation factor specifies that a triangle belongs to the current cluster if and only if its size as well as the sizes of all its neighbors satisfies the rule. To better understand this process, it is described in the pseudo-code as follows:

Select any triangle with size less than mean value as current triangle;
Define a geometry collection as GC;

**Recursive Function TCA (current triangle)**

  - Retrieve its three neighbor triangles;
  - If (sizes of the three neighbor triangles < mean value) Then
    - Add the three neighbor triangles into the triangle set;
  - End If
  - If (the triangle set = empty) Then
    - Return;
  - Else
    - Remove the current triangle from the triangle set;
    - Add the current triangle into the GC;
    - Pick up any triangle from the triangle set as the current triangle;
  - End If
  - Call TCA (current triangle);

End Function

Merge the triangles in GC as a cluster;
The second method proposed in this thesis is coined as the Entropy-based Hierarchical Clustering Algorithm (EHCA, Paper II). This method firstly constructs a TIN model of the spatial points, and then it adopts a strategy of divide and conquer. Generally speaking, it is primarily composed of two parts with a recursive character: (1) decomposition of TIN based on head/tail division rule to obtain a level of clustering; and (2) selection of the best clustering level with the aid of maximum entropy change. The first part aims to remove the long edges with length larger than the threshold value which is determined by the head/tail division rule and reserves the others. However, in practice, a TIN is reserved if it is still connected (which means you can navigate from any node to any other node) even if by removing the long edges. Otherwise, it is divided into several sub TINs. To this end, both the reserved TINs and the sub TINs constitute the current level of clustering.

The second part is designed to calculate the entropy change between successive levels of clustering. The entropy for the clustering of level \( k \) can be defined as,

\[
H_k = -\sum_{i \in N_i} p_i^k \log p_i^k, \quad p_i^k = n_i^k / \sum_{i \in N_i} n_i^k
\]

(4.4)

Where \( n_i^k \) is the number of locations belonging to cluster \( i \) in level \( k \), \( p_i^k \) is the probability of a location belonging to cluster \( i \) in level \( k \), and \( N_i \) is the number of clusters in level \( k \). Then, the entropy change between level \( k \) and \( k-1 \) is defined as, \( \Delta H_{k-1} = H_k - H_{k-1} \), where \( H_k \) is the entropy value for the \( k^{th} \) level of clustering and \( H_{k-1} \) is the one for the \( (k-1)^{th} \) level. To this point, we conjecture that the clustering with the maximum entropy change should be adopted, and this best level of clustering \( k \) can be derived as:
\[ k = \max(\Delta H_{k-1}^l) | k = 1,2,\ldots \]  \hspace{1cm} (4.5)\\

To better understand this algorithm, it is described in the pseudo-code as follows:

cur_level = 1;
final_TINs = null;

**Function EHCA** (TIN, k levels)

- entropy_old = 0;
- max_diff = Negative Infinite;
- Append TIN into TIN_array;

**Repeat** k level

- **Set** TIN_array = **Call function** Decomposing (TIN_array);
- total_num = 0;

  **For each** tin in TIN_array

  - total_num = total_num + number of points in tin;
  - entropy = 0;

  **For each** tin in TIN_array

  - p = number of points in tin / total_num;
  - entropy = entropy + p \times \log (1 / p);
  - entropy_diff = entropy - entropy_old;
  - entropy_old = entropy;

  **If** entropy_diff > max_diff **then**

  - max_diff = entropy_diff;

  **If** cur_level = cur_level + 1;

**End Function**

Convert every tin in final_TINs into patches or polygons as cluster;

**Function Decomposing** (TIN_array)

**For each** tin in TIN_array

- Calculate the mean edge length of tin;
- Obtain the median edge length of tin;

  **If** mean edge length \leq median edge length **then**

  - Reserve this tin;

  **Else**

  - Remove the edges with length greater than the mean edge length in this tin;

    **If** tin is segmented into several components **then**

    - Remove tin from TIN_array;

    **Else**

    - Append every component as new tin into TIN_array;

  **Else**

  - Reserve this tin;

**End Function**

**4.5. Urban sprawl detection**

Urban sprawl is generally considered as the low density, auto-dependent land development, and considerable efforts have been made to the study of its
measurement. The most popular measurement is to compare the urban land consumption rate with the population growth rate (Fulton et al. 2001, Ewing et al. 2002, Torrens 2008), e.g., a city is said to be sprawling if the urban land consumption rate is faster than the one of population growth. In essence, the above method is based on the comparison of historical urban land development and population growth. On another hand, the boundary of urban area has a non-trivial impact on the final result, although satellite imagery has been applied to extract the land use patterns (Sudhira et al. 2004, Ji et al. 2006). Thanks to the spatial units derived from spatial point clustering methods, we can look at urban sprawling problem with this new data using a uniform standard measurement for cross comparing sprawling levels across an entire country.

The method in this thesis adopts a simple criterion to determine whether or not a city is sprawling, i.e., a city is considered to be sprawling if urban expansion is faster than the growth of a certain property like population. To the contrary, a city is considered to be compact or normal if urban expansion is slower than or equal to the growth of the property. In the literature, there are two different perspectives regarding the relationship between urban area \((x)\) and its property \((y)\) in the systems of cities: The first view assumes a linear relationship, i.e., \(y = kx\), whereas the second view assumes a nonlinear relationship or a power relationship, i.e., \(y = cx^\alpha\), where \(\alpha < 1\). The second view comes from the theory of allometry initially developed from biology on the study of the growth of part of an organism in relation to that of the entire organism (West et al. 1999). When it comes to urban systems, the second view is also known as the economy of scale, implying that the larger the cities, the less infrastructure to be built, such as street networks, gas stations, and water pipelines (Bettencourt et al. 2007).

Inspired from the above two views, the sprawl ruler is proposed for measuring urban sprawl in a uniform manner across an entire country, and this method is built on the regression curve between urban area \((x)\) as the x-axis and its property \((y)\) as the y-axis. In this context, two forms of regression curves can be expected: the linear form and the power form. Specifically, this method demarcates the cities into three kinds: compact cities, sprawling cities and normal cities. The normal cities are closely along the regression curve or they are inside the buffer zone around the regression curve with the width defined as the mean value of all the distances from the city point to the regression curve according to the head/tail division rule. The compact cities should lie above the regression curve or buffer zone, whereas the sprawling cities should lie within the region below the regression curve or buffer zone. To explain this method more vividly, we illustrate it in Figure 4.8 where two sprawling cities, two compact cities and several normal cities are identified.
4.6. Complex network analysis

Complex network analysis is an interdisciplinary research which involves the studies from mathematicians, physicists, biologists, and geographers. Geographers are involved because the spatial networks have been reported recently to exhibit the same structural properties as the previous ones such as the internet networks or the biological networks. In this thesis, a systematic analysis of complex network is suggested, although most of the measurements are borrowed from the previous studies. Generally speaking, a complex network can be modeled as a binary network or a weighted network: the former one can be represented by its adjacent matrix as $\text{Adj}(G) = [a_{ij}, i \neq j]$, where $a_{ij}$ is 1 when there exists a link between node $i$ and node $j$ and 0 when there is no link; and the latter one is simply generated by appending a weight value to the corresponding edge, and its form can be described as $\text{WAdj}(G) = [a_{ij} \cdot w_{ij}, i \neq j]$. To this point, Figure 4.9 presents the work flow of the complex network analysis, and the corresponding measurements are illustrated afterwards.
Degree ($D$)
$D$ is defined as the number of nodes connected with the current node from the aspect of graph theory, $D_i = \sum_j a_{ij}$, and it reflects the importance of the current node (Freeman 1979).

Betweenness ($B$)
$B$ is defined as the number of shortest paths between any two nodes that pass through the current node, $B_i = \sum_{m,n \neq i} \text{Path}(m, i, n) / \text{Path}(m, n)$, and it reflects the extent to which the current node lies in the network (Freeman 1979).

Attraction ($A$)
$A$ is defined as the gravitational force of the edge connecting node $i$ and $j$, $A_{ij} = \sqrt{(D_i \cdot D_j) / \text{dist}(i, j)}$. It is proportional to the product of the degrees of the connecting nodes, and is inversely proportional to their geometric distance.

Capacity ($C$)
$C$ is defined as the total weights of the current node’s incident edges, $C_i = \sum_j a_{ij} \cdot w_{ij}$. It is also called the node strength (Barrat et al. 2004) and is a generalized form of the degree measurement in the topological network.

Clustering Coefficient ($CC$)
From a local perspective, $CC$ is defined as the fraction of the actual number of pairs of connected neighbors of the current node over the maximum number of possible pairs (Watts and Strogatz 1998), $CC_i = R_i / ((D_i \cdot (D_i - 1) / 2)$. It is used to
characterize the local cohesiveness of the current node or the extent to which the
nodes in the network are clustered together.

**Average Shortest Path (ASP)**

ASP is defined as the average shortest number of steps among all pairs of nodes
(Watts and Strogatz 1998), \( ASP = 2 \sum_{i>j} \frac{d_{ij}}{(N*(N-1))} \), and it is used to
classify the efficiency of information circulating on the network.

### 4.7. Agent-based modeling (ABM)

ABM can be used as an effective tool to capture the emergent patterns from the
interactions among individual agents and their interactions with the underlying
environment. In this thesis, three types of ABM are proposed to explore the
patterns emerging from the interaction between individual agents and the
underlying spatial structure. In other words, we are not only focused on the
movement patterns emerged from the simulation, but also we want to explore
the role of spatial structure played in the movement patterns. The three types of
ABM include: random/purposive agents in street network, agents with five
different behaviors in airport network, and random/preferential agents in
hierarchical graph of spatial regions.

However, the three types of ABM can be further generalized into two categories,
namely the geometric constraint and the topologic constraint in terms of the
underlying spatial structure. The geometric constraint refers to the situation that
the movement of agents, like the behavior of turtles crawling along the road, is
constrained to the streets, whereas the topologic constraint means the case that
the movement of agents, like the behavior of frogs jumping from one place to
another, is constrained to the graph with nodes representing the spatial units and
edges representing the connections between spatial units. In the following,
details on the two models are elaborated.

The geometric constraint ABM is used to investigate the relationship between
traffic flow of human mobility and structure of street network. In this model: the
agents are bestowed with two behaviors, namely random walk and purposive
walk; and the environment is set to be the urban street network. The random
agents walk continuously along the street and make a random turn at the street
intersection to reach the next street, whereas the purposive agents walk
continuously along the street from one destination to another with the help of
shortest path yet least turns. These behaviors are further illustrated in the
following Figure 4.10 where the blue car represents the random agent and the
green car is the symbol of purposive agent. Besides, all the footprints left by the
agents are aggregated to the corresponding street, and they are regarded as the
traffic flow of that street. In this respect, this model provides us a fresh look at
the question of how spatial configuration of street network affects traffic flow of human mobility. We implement this model using the Netlogo package which provides a friendly environment for ABM developing, and interested readers may refer to http://fromto.hig.se/~bjg/movingbehavior/ for a demo showing the random or purposive moving behavior.

![Figure 4.10: Illustration of the movement of (a) random agent and (b) purposive agent on an axial map](image)

The topologic constraint ABM is used to examine the levy flight characteristic of human mobility and its relationship with the underlying spatial structure. In this model: the agents have a variety of abilities to jump to the next node; and the environment is set to be the graph which is a representation of the underlying spatial structure like the airport network or the hierarchical spatial regions. Generally speaking, agents have two behaviors to choose the next node: one is to visit each adjacent node with equal probability, and the other one is to visit each neighboring node with different probabilities. For example, a common implementation in the second behavior is to visit the neighboring node with the probability proportional to its degree value. In fact, the mobility of agents on this graph can be regarded as a stochastic process in which an agent continuously visits the neighboring node based on the current node uniformly or preferentially. Besides, we introduce the jumping factor, which resembles the damping factor in the PageRank algorithm (Page and Brin 1998) or the revisiting and exploring behavior (Song et al. 2010), to disrupt the stochastic process. For a better understanding, these mobility behaviors are illustrated in Figure 4.11.
Figure 4.11: Illustration of agent mobility in graph

Note: Node 1, node 2 and node 3 are neighboring nodes of node A.

(a): With equal probability to visit neighboring nodes.
(b): With unequal probability to visit neighboring nodes.
(c): With a probability p to visit the non-neighboring nodes.
5. Results and discussions

5.1. Overview
This chapter presents results and discussions obtained from the applications of these methods to VGI datasets for exploring several topics in urban systems. These topics are mainly around two issues: urban space and the corresponding human dynamics. Moreover, the outlets of these applications are eight peer-reviewed papers, and each paper is focused on one particular topic in urban systems. Therefore, this chapter is organized according to the topics of papers, and the result of each paper is subsequently presented and discussed.

5.2. Paper VII: Validating Zipf’s Law for all the US cities
Zipf’s law has been a hot topic among geographers, economists, or mathematicians for nearly a century. It states that if the cities are ranked by their size (e.g. population), then the size of the first rank city is almost two times of the second rank city, three times of the third rank city, and so on. In other words, the size \( s \) of a city is inversely proportional to its rank \( r \), namely \( s \sim r^{-b}, b = 1 \). Mathematically, the phrase “the r\textsuperscript{th} largest city with a size of \( s \)” has the same meaning as the expression like “the number of cities with size equal to or greater than \( s \) is \( r \)” (Adamic 2002), and thereby the relationship between the power law scaling exponent \( \alpha \) and the Zipf’s law exponent \( b \) can be expressed as, \( \alpha = (1 / b) + 1 \). In this respect, the primary aim of this paper is to verify whether the size of all the US cities can be held stably by a power law distribution with exponent value roughly equal to 2.

A major problem involved in this investigation is how to define or demarcate the city boundary, which will have a non-trivial impact on the size distribution in terms of its extent or population. However, the previous studies typically took the city boundary data defined by the US Census or the administrative department, and consequently their results were suspect to be biased because not all of the population was involved. This paper firstly investigates this issue based on all the US cities derived from a massive street node dataset. Cities defined in this way are believed to include all human settlements or at least the majority, which is based on the fact that human activities are constrained to streets: no street node no human activity. Practically, the methods employed by this paper contain city clustering algorithm (CCA, Rozenfeld et al. 2008), city boundary extraction, and heavy-tailed distribution detection. With the street node dataset as input, the work flow of this study is illustrated in Figure 5.1.
The findings in this paper come from two aspects. At the country level, it suggests that Zipf’s law holds remarkably well for the size of all natural cities within the entire US, which implies that cities are power law distributed with exponent value equal to 2.0 (see Table 5.1). To strengthen this conclusion, we take the same experiment on the US Census 2000 urban area, where it finds that the size follows the power law distribution with the exponent value deviating a little from 2.0 (see Table 5.1). Furthermore, we investigate the natural cities within individual US states. The results suggest that Zipf’s law is not universal as far as the US state is concerned. The scaling exponent values are different from one state to another, although it has reported that in general 33 states out of 51 follow the Zipf’s law.

<table>
<thead>
<tr>
<th></th>
<th>700</th>
<th>600</th>
<th>500</th>
<th>400</th>
<th>Urban Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toppest Cities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nodes (a)</td>
<td>1.98</td>
<td>2.09</td>
<td>2.00</td>
<td>2.10</td>
<td>2.06</td>
</tr>
<tr>
<td>Extent (a)</td>
<td>2.06</td>
<td>2.14</td>
<td>2.14</td>
<td>2.22</td>
<td>1.91</td>
</tr>
<tr>
<td>Top150000</td>
<td>2.01</td>
<td>2.09</td>
<td>2.04</td>
<td>2.11</td>
<td>2.06</td>
</tr>
<tr>
<td>Top15000000</td>
<td>2.01</td>
<td>2.09</td>
<td>2.04</td>
<td>2.11</td>
<td>2.06</td>
</tr>
<tr>
<td>ALL</td>
<td>2.01</td>
<td>2.09</td>
<td>2.04</td>
<td>2.11</td>
<td>2.06</td>
</tr>
</tbody>
</table>

The insights of this paper can be described from two aspects: (1) this study extracts a massive street node dataset with the size of almost 25 million from a 120 gigabytes of OpenStreetMap (OSM) data, and further it is believed that this massive dataset is valuable for the data intensive research community (available at: http://fromto.hig.se/~bjg/ijgis/Zipf); and (2) from the perspective of massive dataset and large geographic scope, this study might hint a fundamental shift in the characteristic of the geographic space from the conventional Gaussian distributed to the heterogeneous power law distributed. Nonetheless, we still
need to explore the implications behind the different states with different scaling exponent values as well as the mechanism driving the regularity of the entire country, which constitute our future work.

5.3. Paper II: Uncovering scaling property of urban systems

Having reported the scaling property of natural cities in above section, this section continues to present the scaling property of urban space with respect to the spatial units within cities. Here the spatial units include the clusters derived from street nodes of Stockholm, photo locations of London, and taxi trajectory static points in Gävle. The three different datasets can represent different aspects of urban systems. For example, street node dataset from the OSM project can reflect a general image of the underlying spatial structure, photo location dataset from the Flickr website is believed to be highly associated with spatial points of interest (POI), and static point dataset from taxi trajectories may be considered as a proxy of human activities.

Different from the investigation on natural cities at a city level, this paper aims to explore the distribution of spatial units at a district or sub-city level. In this respect, the knowledge acquired at this level will definitely enhance our understanding on the inner characteristic of city systems. Another difference from the natural cities is the cluster generating method, and the spatial units are derived using a new method called Entropy-based Hierarchical Clustering Algorithm (ECHA). To this point, methods employed in this study include EHCA and heavy-tailed distribution detection. The work flow can be clearly observed in Figure 5.2.

![Figure 5.2: Work flow for uncovering scaling property of spatial units in urban systems](image-url)
The findings of this paper are relatively simple but important. First, the best clustering level is determined as two after ten iterative clustering processes, and this level is kept unchanged with respect to the three different datasets. Second, the cluster size can be approximated well by a power law distribution, and this scaling property is reported as the common characteristic for three different aspects of urban systems, namely urban structure, spatial points of interest (POI) and human activity. Besides, it is worth to note that the power law model fitted here is statistically significant, because the Goodness-of-Fit (GoF) test reports the \( P \) value greater than the threshold 0.05. For clarity, the overall results can be found in Table 5.2.

<table>
<thead>
<tr>
<th>Street node</th>
<th>Amount</th>
<th>Best clustering level ( k )</th>
<th>( \alpha )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photo location</td>
<td>124,738</td>
<td>2</td>
<td>2.4</td>
<td>0.7</td>
</tr>
<tr>
<td>Taxi static point</td>
<td>316,209</td>
<td>2</td>
<td>2.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

The contribution of this paper can be summarized from two aspects. On one hand, this paper has devised a new clustering method, namely ECHA, to aggregate the individual spatial points to form the components or spatial units in urban space. This method ensures the derived clusters not only contain enough information but also avoid over splitting. Both the robustness and effectiveness of this algorithm are examined afterwards, and the results display remarkable consistency. Furthermore, this paper reports the power law distribution of spatial units derived from three datasets reflecting different aspects of urban systems, which further strengthens our understanding on scaling characteristic of the urban space: the small components or spatial units are much more common than the large ones, whereas the large components or spatial units are far more important than the small ones.

5.4. Paper I: Measuring urban sprawl using massive street nodes

Bearing in mind the scaling property of urban space at the city or sub-city level, this paper investigates the problem of urban sprawl for the cities in both the US and European countries. The contribution of this study can be attributed to two points: (1) this paper proposes a novel method to demarcate the cities into sprawling, normal, and compact, which suggests a uniform standard way for the cross comparison of all cities across an entire country; (2) this paper abandons the conventional way of using administrative city boundaries and population data, and instead adopts natural cities and street nodes to measure urban sprawl.
The reasons behind such a change come from two aspects: (1) administrative city boundaries have been criticized for being subjective and expensive, whereas it is efficient and effective to derive natural cities with OpenStreetMap (OSM) data; (2) street nodes within a city are highly correlated with the city population with R square value equal to 0.89, which indicates street nodes defined at the individual level can be a proxy of population for measuring urban sprawl. This study exploits spatial point clustering method, heavy-tailed distribution detection method, and urban sprawl ruler. The entire work flow is shown in Figure 5.3.

Figure 5.3: Work flow for measuring urban sprawl using street nodes

The results on the sprawling situation of the US cities roughly agree with the literature (see Figure 5.4a), whereas the results for cities of three European countries (France, Germany and the UK) are believed to match well with our general perception (see Figure 5.4b). In the US, this paper finds that a majority of coastal cities suffer less from sprawl and most of the southeastern inland cities suffer more from sprawl, and that it is a bit difficult to identify a general pattern for cities in European countries. At the individual city level, this study reports that only three cities out of the top 25 are inconsistent with the cities in terms of US Census urban area: New York region, Richmond region, and Roanok-Lynchburg region. On the other hand, for European cities, Paris is reported as a compact city, Berlin is classified as a sprawling city, and London is also considered as a sprawling city.
However, how these results about sprawling or compact cities are related to a situation in reality warrants further comparison or verification due to a simple fact that the boundaries of natural cities and real cities could be different. Besides, this paper gave neither explanations around the underlying mechanism nor the driving force hidden within a sprawling or compact city, such as the geographic constraints, the economic structure, etc. Therefore, explanations as well as city growth models can be assisted with the knowledge from the fields of geography, economy, and demography, which points out our future studies.

5.5. Paper III: Analyzing the US airport network
This paper investigates the structure and traffic patterns of airport network which as an important transportation infrastructure plays a vital role within both
the systems of cities and the city systems. Here the airport network is built from
the en-route location data of all the US flights, and it is totally different from the
conventional way of constructing an airport network: with the flight dataset
provided by air service department, a flight record corresponds to an edge and
the individual airport or aggregated airport within a city is considered as a node.
In this respect, our airport network exhibits high flexibility over the
administrative based ones. For example, it considers the effect of the first law of
geography: nearer airports belonging to different cities are more connected with
each other and hence should be merged as one airport.

In this paper, two aims are set to explore the US airport network: one is the
structural and traffic analysis of the network and the other one is the analysis of
individual airport patterns. To assist the two aims, several methods introduced
previously are employed, including flight extraction which has been described in
chapter three, Triangle Clustering Algorithm (TCA) which aims at deriving
clusters from the spatial points, complex network analysis which has been
widely applied to many other kinds of transportation networks, and heavy-tailed
distribution detection which has penetrated into every corner of this thesis. A
general glance of the work flow in this study is shown in Figure 5.5.

Figure 5.5: Work flow for building and analyzing the US airport network

The findings of this paper are mainly concentrated on two points. On one hand,
results from the structural and traffic analysis of airport network suggest that (1)
the US airport network exhibits scale-free, small-world, and disassortative
mixing properties; and (2) that traffic on each route can be approximated by a power law distribution which indicates that a majority of flights are operated on a minority of busy routes; and (3) that a remarkable power relationship between the structural measurements in the binary graph and the traffic measurements in the weighed counterpart, namely degree versus capacity and attraction versus volume, can be clearly observed.

On the other hand, results from the analysis of individual airport patterns may show that (1) traffic handled by each airport can be also approximated by a power law distribution, which indicates a small number of airports handling a huge amount of air traffic and characterizes the heterogeneity of air traffic among the airports. The results further suggest (2) a typological map (see Figure 5.6) of the entire airports according to the similarity of five structural measurements, among which the largest 25 airports belonging to the first category constitute roughly the major US airports reported by Federal Aviation Administration (FAA, available at: fly.faa.gov/flyfaa/usmap.jsp). Lastly, this study reminds us that (3) individual airport traffic patterns might be subject to the underlying socio-economic factor, geographical constraints, and so on.

The contributions of this paper can be stressed as follows. (1) This study constructs the US airport network from the en-route location data of domestic US flights, which shows an information mining process from the underlying massive geographic data that is different from the previous studies relying on flight records. (2) This paper explores the relationship between structural measurements and traffic dynamics in the US airport network, which will benefit future studies on the evolution of air transportation system. Lastly, (3) this paper contributes to the research community with the airport similarity...
analysis in terms of both structural measurement and flight distance distribution, which not only strengthens our understanding of the airport typology but also helps future studies in other related fields, such as airport traffic design and management.

5.6. Paper VI: Exploring human mobility patterns at the city level

At the city level, this paper explores the relationship between human mobility pattern and the underlying street structure. Human mobility pattern refers to the movement flow captured by the street and street structure refers to the topological relationship of streets as to what street is connected with what other streets. Particularly, the movement flow on the street is calculated with two methods: the footprint counter which measures the number of footprints left on that street by the human movements and the gate counter which measures the number of times that movement trajectories passing through the gate set on the street segment. The latter way of movement flow is further obtained by the summation of all the gates’ flow along that street. On the other hand, seven metrics from either space syntax or complex network have been adopted in this paper to quantitatively measure the street structure, and they include Connectivity (Cnt), Control (Ctr), Betweeness (Btw), Local Integration (Ltg), Global Integration (Gtg), PageRank (PR), and Weighted PageRank (WPR).

From this point, two questions are proposed in this paper around the relationship between movement flow and street structure: (1) if movement behavior has a non-trivial influence on mobility pattern, or whether mobility pattern is mainly shaped by the underlying street network; and (2) which street morphological metric is the best indicator of movement flow. To answer the above two questions, an agent-based model is firstly devised to simulate the movements of agents with two different behaviors: random walk and purposive walk. Then, a correlation analysis is carried out between the simulated movement flows and the metrics of street network. Particularly, purposive agents are subdivided into two types: the ones (purposive I) choosing the next destinations globally, and the others (purposive II) with a high probability (80%) of choosing the near destination within two topological steps and a low probability (20%) of choosing the far destination beyond two topological steps. In addition, the techniques adopted in this study include space syntax principle and agent-based modeling (ABM). The entire work flow can be observed in Figure 5.7.
Findings in this paper can be reported from the following points: (1) the observed movement flows captured by the gate counters in the four sites of London are well correlated with the seven morphological metrics, particularly the metrics of WPR and Ltg have the best performance in terms of the correlation whereas Ctr and Gtg have the worst performance; (2) the simulated movement flows captured by the footprint counters in the three areas of London and the Gävle region have a high correlation with all the metrics except Gtg, and importantly WPR can be considered as the best metric to predict the flows irrespective of the movement behavior in Gävle. This is different from the London counterpart where Btw seems to be the best for purposive agents and WPR the best for random ones (see Table 5.3).

Table 5.3: R square values between the movement flows calculated by the footprint counters and the seven morphological metrics

<table>
<thead>
<tr>
<th></th>
<th>Barnsbury</th>
<th>Clerkenwell</th>
<th>S.Kensington</th>
<th>Mean</th>
<th>Gävle</th>
</tr>
</thead>
<tbody>
<tr>
<td>WPR</td>
<td>0.87</td>
<td>0.92</td>
<td>0.91</td>
<td>0.90</td>
<td>0.89</td>
</tr>
<tr>
<td>PR</td>
<td>0.86</td>
<td>0.87</td>
<td>0.84</td>
<td>0.86</td>
<td>0.75</td>
</tr>
<tr>
<td>Cnt</td>
<td>0.86</td>
<td>0.87</td>
<td>0.84</td>
<td>0.86</td>
<td>0.75</td>
</tr>
<tr>
<td>Ctr</td>
<td>0.54</td>
<td>0.48</td>
<td>0.44</td>
<td>0.48</td>
<td>0.41</td>
</tr>
<tr>
<td>Btw</td>
<td>0.67</td>
<td>0.64</td>
<td>0.57</td>
<td>0.63</td>
<td>0.56</td>
</tr>
<tr>
<td>Ltg</td>
<td>0.82</td>
<td>0.83</td>
<td>0.80</td>
<td>0.82</td>
<td>0.36</td>
</tr>
<tr>
<td>Gtg</td>
<td>0.23</td>
<td>0.32</td>
<td>0.27</td>
<td>0.27</td>
<td>0.26</td>
</tr>
<tr>
<td>Purposive walkers (I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WPR</td>
<td>0.63</td>
<td>0.63</td>
<td>0.57</td>
<td>0.61</td>
<td>0.67</td>
</tr>
<tr>
<td>PR</td>
<td>0.59</td>
<td>0.58</td>
<td>0.53</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>Cnt</td>
<td>0.59</td>
<td>0.58</td>
<td>0.53</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>Ctr</td>
<td>0.42</td>
<td>0.42</td>
<td>0.38</td>
<td>0.41</td>
<td>0.42</td>
</tr>
<tr>
<td>Btw</td>
<td>0.81</td>
<td>0.77</td>
<td>0.75</td>
<td>0.77</td>
<td>0.57</td>
</tr>
<tr>
<td>Ltg</td>
<td>0.54</td>
<td>0.51</td>
<td>0.46</td>
<td>0.50</td>
<td>0.16</td>
</tr>
<tr>
<td>Gtg</td>
<td>0.20</td>
<td>0.19</td>
<td>0.16</td>
<td>0.19</td>
<td>0.13</td>
</tr>
<tr>
<td>Purposive walkers (II)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WPR</td>
<td>0.73</td>
<td>0.73</td>
<td>0.68</td>
<td>0.71</td>
<td>0.75</td>
</tr>
<tr>
<td>PR</td>
<td>0.69</td>
<td>0.69</td>
<td>0.64</td>
<td>0.68</td>
<td>0.60</td>
</tr>
<tr>
<td>Cnt</td>
<td>0.69</td>
<td>0.69</td>
<td>0.64</td>
<td>0.68</td>
<td>0.60</td>
</tr>
<tr>
<td>Ctr</td>
<td>0.47</td>
<td>0.46</td>
<td>0.41</td>
<td>0.45</td>
<td>0.33</td>
</tr>
<tr>
<td>Btw</td>
<td>0.80</td>
<td>0.77</td>
<td>0.74</td>
<td>0.77</td>
<td>0.53</td>
</tr>
<tr>
<td>Ltg</td>
<td>0.67</td>
<td>0.63</td>
<td>0.59</td>
<td>0.63</td>
<td>0.33</td>
</tr>
<tr>
<td>Gtg</td>
<td>0.20</td>
<td>0.20</td>
<td>0.17</td>
<td>0.19</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Figure 5.7: Work flow for exploring the mobility patterns at the city level
Besides, (3) the simulated movement flows captured by the gate counters also exhibit a high correlation with all the metrics except Gtg. Specifically, in London area, WPR performs the best for random agents and Btw is the best choice for purposive agents, which is the same as the case of the flows captured by the footprint counters. On another hand, in Gävle region, WPR performs the best for random agents and Btw as well as Cnt appear to be the best for purposive agents, whereas Ltg seems to perform poorly compared with the London counterpart. In addition, in London area, (4) a high correlation can be observed between the observed flows and the simulated flows with respect to three behaviors. Note that random agents appear to have the best correlation followed by purposive II agents, then followed by purposive I agents.

Table 5.4: R square values between the movement flows calculated by the gate counters and the seven morphological metrics

<table>
<thead>
<tr>
<th></th>
<th>Barnsby</th>
<th>Clerkenwell</th>
<th>S.Kensington</th>
<th>Mean</th>
<th>Gävle</th>
</tr>
</thead>
<tbody>
<tr>
<td>WPR</td>
<td>0.90</td>
<td>0.92</td>
<td>0.90</td>
<td>0.91</td>
<td>0.71</td>
</tr>
<tr>
<td>PR</td>
<td>0.91</td>
<td>0.89</td>
<td>0.87</td>
<td>0.89</td>
<td>0.68</td>
</tr>
<tr>
<td>Cnt</td>
<td>0.91</td>
<td>0.89</td>
<td>0.87</td>
<td>0.89</td>
<td>0.68</td>
</tr>
<tr>
<td>Ctr</td>
<td>0.58</td>
<td>0.47</td>
<td>0.45</td>
<td>0.50</td>
<td>0.42</td>
</tr>
<tr>
<td>Btw</td>
<td>0.64</td>
<td>0.59</td>
<td>0.51</td>
<td>0.58</td>
<td>0.47</td>
</tr>
<tr>
<td>Ltg</td>
<td>0.83</td>
<td>0.83</td>
<td>0.80</td>
<td>0.82</td>
<td>0.34</td>
</tr>
<tr>
<td>Gtg</td>
<td>0.24</td>
<td>0.32</td>
<td>0.24</td>
<td>0.26</td>
<td>0.16</td>
</tr>
</tbody>
</table>

This paper sheds some important insights into the relationship between human mobility pattern and the underlying street network. Firstly, human movement behavior has little influence on movement flows left on the street network. In other words, given the street network, human mobility pattern is almost the same irrespective of movement behavior. Secondly, the metric Gtg is found to be the
worst indicator in terms of explaining the movement flow. On the contrary, the metrics of WPR, PR, Cnt, and Btw are suggested to be the good indicators. Nonetheless, this paper puts forward two questions needing further investigations: (1) how to explain different metric performances with respect to different geographic regions; and (2) what about the result if interactions between agents are adopted.

5.7. Paper VIII: Exploring human mobility patterns at the country level
At the country level, this paper explores human mobility patterns using the en-route location information of the US flights and further tries to explain these patterns through an agent-based simulation with different moving behaviors. The first aim means to investigate the human travel length distribution at the country level as well as other geometric or topological properties of the airport network, such as traffic volume between airports, route length distribution, etc. The second aim is supposed to mimic the observed mobility patterns with agent-based simulations including five moving behaviors. Generally, these moving behaviors can be categorized into three types: geometric based (Figure 5.8 a), topological based (Figure 5.8 b, c, d), and preferential return based (Song et al. 2010, Figure 5.8e). Obviously, ABM with the first two moving behaviors can be fully treated as the Markov chain process (Stroock 2005) which is a random process characterized by the memoryless property indicating that the next state of the system solely depends on the current state. The ABM with the last moving behavior is not fully a Markov chain process with the introduction of a preferential return to the previous states.

Figure 5.8: Illustration of different moving behaviors in ABM
Bearing the above aims in mind, findings from the observed en-route flight data suggest that (1) both in-degree and out-degree of the airport network exhibit a striking power law distribution, which indicates the heterogeneous structure of the US airport network. Furthermore, (2) traffic volume between airports can be approximated well by a power law distribution, which shows the same property as the topological structure. Besides, from a geometric perspective, they demonstrate that (3) route length between two connected airports follows stably an exponential distribution, which is visualized in Figure 5.9. Lastly but importantly, (4) human travel length approximated by flight length is also found to obey an exponential distribution. This finding can be seen clearly from the Table 5.5, where the exponential distribution outperforms significantly other competing distributions with high P value passing the GoF test. It is inconsistent with other findings in the literature (Brockmann et al. 2006, Gonzalez et al. 2008, Jiang et al. 2009) but is consistent with recent studies (Bazzani et al. 2010 and Liang et al. 2012).

Figure 5.9: The route of the US airport network rendered by its geometric length

On the other hand, results from ABM report that (1) simulated flight length can be well fitted by the exponential distribution compared with the other competing distributions, and it shows striking consistency irrespective of the five moving
behaviors which can be seen from Table 5.5. In addition, (2) the random moving behavior (T1) seems to be the best one among the five proposed ones to capture the observed mobility pattern, although T2 behavior may be a potential rival which is ranked as second. Note that the similarity is measured according to the average KS distance between observed flight lengths and simulated ones through 100 experiments (Table 5.6).

Table 5.5: Model selection results for both observed flight length and simulated flight length

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Power law</th>
<th>Exponential</th>
<th>Lognormal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$\lambda$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>T1</td>
<td>3.03</td>
<td>9.44E-07</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>1368872</td>
<td>771877</td>
<td>0.69</td>
</tr>
<tr>
<td>T2</td>
<td>3.00</td>
<td>9.79E-07</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>1291135</td>
<td>784208</td>
<td>0.71</td>
</tr>
<tr>
<td>T3</td>
<td>9.35</td>
<td>7.70E-07</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>3602005</td>
<td>802353</td>
<td>0.69</td>
</tr>
<tr>
<td>G</td>
<td>3.47</td>
<td>1.44E-06</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>1368872</td>
<td>457352</td>
<td>0.74</td>
</tr>
<tr>
<td>PR</td>
<td>6.51</td>
<td>9.11E-07</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>3665280</td>
<td>471625</td>
<td>0.71</td>
</tr>
</tbody>
</table>

| Observed     | 5.71      | 8.14E-07    | 0.70      | 13.96    |
|              | 4648157   | 471625      | 0.77      | 409698   |

Table 5.6: KS distance between observed flight lengths and simulated ones

<table>
<thead>
<tr>
<th>Models</th>
<th>G</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>0.3866</td>
<td>0.0295</td>
<td>0.0470</td>
<td>0.1985</td>
<td>0.1264</td>
</tr>
</tbody>
</table>

The contributions of this paper can primarily be concentrated to the following two aspects. (1) This paper supports the argument that human travel length at the country level can be approximated by the exponential distribution, which can be better replicated by the ABM with the underlying airport network and random behavior. Moreover, (2) this paper contributes to the research community with the method of heavy-tailed distribution detection, which suggests a systematic way to select the best model from the underlying five models. However, the main argument proposed in this paper is based on the observed US flight data within one week, which may be criticized for being non-representative of human.
travel at the country level. Therefore, data from other sources should be involved in future studies to verify our argument at the country level.

5.8. Paper V: Exploring human mobility patterns at the regional level

Followed by the above investigations on human mobility at the city and country level, this paper further explores human mobility patterns at the regional level and intends to interpret the patterns via the ABM. Here, human mobility patterns refer to the human travel length distribution which is approximated by the GPS traces of 258 volunteers and other properties, such as home distance pattern and gyration radius. On another hand, the ABM is based on the following two assumptions: (1) the underlying spatial point of interest (POI) would have a strong influence on the patterns of human mobility and (2) the human preferential selection of available resource also might impose a non-negligible influence on the mobility patterns. From a computational perspective, the two assumptions constitute two basic components of the ABM, namely the environment and the moving behavior. The environment is set as a three-level hierarchical regional graph constructed from mobility data, and the moving behavior is the preferential selection. To achieve the above goals, several methods are employed, including spatial point clustering, ABM, and heavy-tailed distribution detection. For a clear illustration, the entire work flow is shown in Figure 5.10.

Figure 5.10: Work flow for exploring human mobility patterns at the regional level
This paper reports both the regular and scaling properties of human mobility. The regularity of human mobility is found in the daily movement in terms of home distance pattern and the number of individual daily purposes. On the other hand, the scaling property of human mobility is reported with respect to human travel length, gyration radius, and time duration at the purposive location. Importantly, scaling property of human travel length is also known as the levy flight characteristic of human mobility, which is consistent with most of the previous studies (Brockmann et al. 2006, Gonzalez et al. 2008, Jiang et al. 2009) but is inconsistent with other studies (Bazzani et al. 2010, paper VIII, Liang et al. 2012). In addition, a three-level purpose cluster is presented, and the size of the clusters in each level exhibits a remarkable power law distribution.

Furthermore, results from the ABM are reported. Firstly, uniform agent cannot mimic the observed levy flight characteristic, whereas preferential agent does capture this observed levy flight characteristic with scaling exponent value equal to 2.02. Secondly, the ABM is extended by adding an ingredient called jumping factor (JF), which refers to the situation in reality that one person might have a probability to cancel the regular mobility schedule and immediately make a decision to move to another place. By tuning the value of JF from 0 to 1, our model covers a relatively wide range of human mobility patterns with different scaling exponent values from 1.55 to 2.05.

The contributions of this study are mainly focused on the following three points. Firstly, this paper contributes to the society with a large amount of interesting purposive locations which may be useful in many fields, such as transportation management and urban planning. Secondly, based on the three-level purpose clusters, this paper constructs a hierarchical regional graph corresponding to different levels of geographical entities: high level city or town, middle level city districts, and low level city blocks, which serve as the spatial structure of this study. Lastly but importantly, this paper supports the argument that human mobility pattern at the regional level is mainly shaped by the underlying spatial structure and the individual preferential behavior. However, two issues are to be dealt with in future work: (1) the implications of the purposive locations, and (2) the improvement on the model to have the ability to mimic the exponential cutoff effect observed in real world.

5.9. Paper IV: Exploring human activity patterns

Apart from the investigations on human mobility patterns, this paper aims to explore human activity patterns which not only strengthen our understanding of the activities in place, but also benefit many other fields, such as urban planning and traffic management. The taxicab static points (SPs) dataset is adopted in this
study as a proxy of human activities, and consequently the human activity is considered as a general activity with no specific type. With this substitution and generality, the patterns of human activity are examined at both the aggregated and individual level. At the aggregated level, this paper conducts a temporal, spatial, and scaling analysis: temporal analysis allows us to understand the pulse of human activities in place; spatial analysis enables us to detect the spatial drifting of human activities from one region to another; and scaling analysis examines the diffusion pattern of human activities. At the individual level, this paper firstly explores the temporal pattern of each taxicab, and then it examines the predictability of taxicabs in space.

To achieve these tasks, several methods or tools are employed in this paper, and they mainly include time segmentation technique, heavy-tailed distribution detection method, and information theory. The time segmentation technique is used to split the continuous observation period into discrete time slots, and it contributes to the aggregated temporal analysis, the aggregated spatial analysis, and the individual temporal analysis. Also, the heavy-tailed distribution detection method is adopted to analyze the scaling property of aggregated human activities, whereas the information theory is applied to estimate the information for activity location prediction. These methodologies constitute the backbone of the subsequent analysis and the final results, and the entire work flow of this study is illustrated in Figure 5.11.

![Figure 5.11: Work flow for exploring human activity patterns](image-url)
The results are reported from five points with respect to the specific analysis. (1) Human activities exhibit an obvious regularity in time, which is in consistence with our common sense. For example, human activities explode in the weekend night whereas they dissipate during the weekday night. (2) They also show a remarkable spatial drifting pattern. For example, in weekday, it is observed that there is an expansion from the central districts towards the surrounding districts during the daytime whereas a shrink from the surrounding districts to the central districts during the nighttime. Besides, (3) they follow a striking power law distribution during the rest periods, whereas they agree well with the lognormal distribution during other periods. Apart from the aggregated results, this paper suggests that (4) individual taxicabs have different temporal regularities with each other, which is also reasonable since every taxi driver has her/his own working schedule. Lastly, this paper indicates that (5) about 1 bit information is enough to estimate the next location of individual activity.

The insights of this paper are twofold. Firstly, this paper suggests a convenient way to explore human activities using SPs, which avoids the conventional examination requiring the diary of human activities in space and time. Secondly, human activities are heterogeneous in space irrespective of their spatial drifting with time, namely a majority of activities are concentrated on a minority of places, whereas a minority of activities are spread over a majority of places. This pattern is in consistence with many other findings in social science (Barabási 2005, Radicchi 2009) and in the geographic space (Lämmer et al. 2006, Jiang 2007, Paper II, Paper VII).
6. Conclusions and future work

6.1. Conclusions
With the continuous progression of urbanization, the percentage of population living in cities had reached almost 75% for more developed regions and 44% for less developed regions in 2007. This trend has drastically increased the contacts among people and their interactions with the surrounding urban environment, which further leads to subsequent urban problems such as traffic congestion, urban sprawl, and environmental pollution. These challenging problems have been reported to be chaotic or non-deterministic, and they cannot be simply tackled from the administrative perspective with road pricing or smart growth. In this respect, urban systems should be undoubtedly considered as a complex system, and consequently the methods or techniques used to examine the complex system should be adopted to explore the complex issues in urban systems.

This thesis is initially inspired from the above facts, and it mainly focuses on two major issues of urban systems, although urban systems can be divided into many subsystems with different issues. Specifically, the two issues are concentrated on the aspects of urban space and the corresponding human dynamics. Urban space can be looked at from different aspects, such as the city boundary, the axial map, or the transportation network. The related data can be obtained from mapping agencies or derived with certain rules. Relying on the naturally derived urban space, this thesis has investigated its scaling as well as other structural properties from different aspects and further applied this scaling property to resolve the problem of urban sprawl. On the other hand, the human dynamics is not only affected by the urban space but also shapes the urban space. This is examined from two perspectives: human mobility patterns and human activity patterns. Both of the two patterns have suggested the heterogeneous characteristic of human behavior and showed its subtle relationship with the underlying urban space or spatial structure, which will definitely benefit many other fields, such as traffic management and infectious disease control.

The above two issues are further associated with several specific topics examined in this thesis: the first issue is around urban space and is to examine the scaling property of cities and spatial units within cities, the problem of urban sprawl, and the structure of transportation network; whereas the second issue is around human dynamics and is to explore the human activity patterns at the city level and the human mobility patterns at three levels: city, region, and country.
Instead of using administrative data such as city boundaries which have high cost, or human activity data gathered from questionnaires which are time consuming, this thesis adopts the Volunteered Geographic Information (VGI) acquired from several sources: highway dataset from the OpenStreetMap (OSM) project, photo location dataset from the Flickr website, GPS tracking datasets from volunteers, taxicabs, and air flights. These VGI datasets are collected at the individual level, are free to use and offers a new way to understand issues in urban systems. In this respect, the procedure to explore the two issues is also a knowledge discovery process based on massive amounts of geographic data using the methods or tools of complex system.

Results from the first issue are the outlets of four peer-reviewed papers. Particularly, paper VII investigated the controversial topic of city size regularity in terms of all the US natural cities, and the conclusion is that Zipf’s law holds stably for all the US cities. Apart from the city level scaling, paper II delved into the spatial units within cities in terms of clusters formed by street nodes, photo locations, and taxi static points, and it further reported the scaling property of spatial units at the sub-city level. Enlightened by the scaling property of the urban space at the city and sub-city level, paper I proposed a sprawl ruler to classify the cities into three categories: compact cities, normal cities, and sprawling cities, which suggested a uniform standard for the cross comparison of the sprawling extent of all cities across an entire country. Lastly, to understand the structure property of transportation network, paper III took a look at the airport network built from en-route location information of US flights, and it reported the scale-free, small-world, and disassortative mixing properties of the USAN as well as the heterogeneous patterns of individual natural airports.

On the other hand, results from the second issue are the outlets of four peer-reviewed papers. First, at the city level, paper VI investigated the correlation between human movement flow and the morphological metrics of street network via agent-based modeling (ABM), and it conjectured that human mobility behavior is primarily shaped by the underlying street network irrespective of random walk or purposive walk. Second, at the country level, paper VIII reported the exponential distribution of observed human travel length by flight, and it suggested the same finding that human mobility behavior is largely constrained by the underlying airport network. Third, at the regional level, paper V presented the heterogeneous and particularly the levy flight characteristic of the observed human travel length, and it again conjectured that human mobility behavior is mainly determined by the three-level hierarchical regional graph, namely city block, city district, and city. The findings from the above three papers are non-trivial, and they together support the argument that human mobility exhibits the scaling property if not limited to the power law and this scaling pattern is highly related to the scaling property of the underlying spatial
structure. Lastly, apart from the investigations on human mobility, paper IV explored human activities at the city level using the taxicab static points, and it further demonstrated that human activities are regular in space and time, heterogeneous in space irrespective of time, and predictable in space for the individual taxicab.

To this point, the aims proposed in this thesis have been achieved and accomplished in terms of devising novel methods or tools and applying them to the VGI datasets for uncovering knowledge on the issues of urban systems. Furthermore, this thesis shares the research community with geographic data generated from extensive VGI datasets and the corresponding source codes (available at: http://fromto.hig.se/~bjg/VGI/VGIDataset.rar) as well as the new knowledge in terms of empirical findings, development of methods, and design of theoretic models. Nonetheless, the knowledge is believed to be useful in many fields of urban studies, such as city planning, transportation management, infectious disease control, and regional economic research. However, the issues investigated only constitute “the tip of the iceberg” of urban systems, and we are motivated to explore more in future scientific endeavors.

6.2. Future work

With the availability of large amounts of Volunteer Geographic Information (VGI), we are granted a chance to carry out the data intensive research related to many complex issues in urban systems and to further uncover the knowledge for urban management. In spite of the findings reported in this thesis around the issues of urban space and human dynamics, more efforts are needed to extend or consolidate the findings reported in this thesis as well as to explore new issues in urban systems. These efforts are brought forward in the following section.

Firstly, it is necessary to extend and consolidate the findings reported in this thesis, and future work will be mainly focused on four aspects. (1) Time series analysis will be applied to study the dynamic character of urban sprawl for individual natural cities, which is different from the way this thesis examined, namely with comparisons from one natural city to another. Apart from the time series analysis, explanations should be given around the underlying mechanism or the driving force hidden within a sprawling or compact city, such as geographic constraints and economic structure. (2) Another issue which deserves future study is to understand the relationship between the scaling exponent value of a US state and its inner structure represented as a combination of economic index and political factors.

Secondly, (3) the human mobility model will be extended by the implementation of interactions among agents. Moreover, the three different agent-based models
should be integrated into one model so that simulations can be performed using a uniform interface. (4) Regarding the distribution of human travel length at the country level, the conclusion drawn in this thesis should be double-checked in future studies using large volume of datasets from other sources.

Lastly, new issues in urban systems will be explored with the employment of our state-of-art methods and extensive VGI datasets. (1) The purposive locations extracted from human movement data will allow us to discover meaningful knowledge about the most probable activities that one has been involved in space and time. In this respect, a complete profile of individual person can be obtained, which will contribute to studies in human geography. (2) Another issue in urban systems which deserves further investigation is the site or landmark appraisal from an ecological perspective. Using the human movement data, it is convenient to estimate the amount of Green House Gas emission related to a specific landmark. Hence, the site or landmark can be evaluated quantitatively in terms of its environmental influence, which further provides strategies for the issue of relocation planning.
References

Ahas R., Aasa A., Silm S. and Tiru M. (2010), Daily rhythms of suburban commuters’ movements in the Tallinn metropolitan area: Case study with mobile positioning data, Transportation Research Part C: Emerging Technologies, 1, 45-54
Batty M. (2003), Unwired cities, *Environment and Planning B: Planning and Design*, 30(6), 797-798
Batty M. (2008a), Cities as complex systems: Scaling, interactions, networks, dynamics, and morphologies, *UCL Working Paper Series*, ISSN 1467-1298
Bell G., Hey T. and Szalay A. (2009), Beyond the data deluge, *Science*, 323(5919), 1297-1298
Berry B. (1964), Cities as systems within systems of cities, *Papers of the regional science association*, 13, 147-163


Chilton S. (2009), Crowdsourcing is radically changing the geodata landscape: Case study of OpenStreetMap, in: *Proceedings of the 24th international cartographic conference*, Santiago, Chile.


Frankel F. and Reid R. (2008), Big data: distilling meaning from data, *Nature*, 455(30)
Fuchs C. (2003), Structuration theory and self-organization, *Systemic Practice and Action Research*, 16(2), 133-167
Gray J. and Szalay A. (2007), eScience—A transformed scientific method, Presentation to the Computer Science and Technology Board of the National Research Council, Mountain View, CA.


Hillier B., Penn A., Hanson J., Grajewski T. and Xu J. (1993), Natural movement: configuration and attraction in urban pedestrian movement, *Environment and Planning B: planning and design*, 20, 29-66


May R.M. (1976), Simple mathematical models with very complicated dynamics, *Nature*, 261(5560), 459-467


93


Price D.J. (1965), Networks of scientific papers, *Science*, 149, 510-515


complex network approach, *Journal of Transport Geography*, 0966-6923