博士論文

Urban human mobility analysis for urban system design based on public transportation smart card data

(交通ICカードデータに基づく都市システムデザインのための人間の都市内移動分析)

前田 高志ニコラス

Urban human mobility analysis for urban system design based on public transportation smart card data

(交通ICカードデータに基づく都市システム
デザインのための人間の都市内移動分析)

Takashi Nicholas Maeda

前田 高志ニコラス

Graduate School of Engineering
The University of Tokyo

This dissertation is submitted for the degree of
Doctor of Philosophy
博士 (工学) 学位請求論文

Abstract

A city is a complex system where areas of a city interact with each other through the medium of people, things, and information. Cities evolve by the feedback loop of these interactions. Such mechanisms make it difficult to capture and predict the current state and future state of each area of a city.

In recent years, the availability of human mobility data is rapidly increasing, and these include smart card data of public transportation. Such data are expected to help contribute to the planning and improvement of cities in developed and developing countries alike.

This research investigates methods to analyze and predict urban systems based on smart card data of public transportation. First, this research attempts to develop a method to estimate the break-down of the number of passengers' arrivals at each station for each activity type. Second, this research attempts to develop a method to predict the changes of future residential mobility based on human mobility on non-working days. Third, this research attempts the comparative examination of network clustering methods for extracting the community structure of a city.

The above methods can be used for the analysis of urban systems without relying on travel survey data that are time-consuming to collect. In addition, the proposed methods are designed for coping with many types of problems in urban space, such as the overpopulation of city centers and the shrinkage of population in commuter towns located far from city centers.

Contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background and motivation

1.1.1 Cities as complex systems

A city is a sytem where areas are interacting with each other in a complex way through the medium of entities such as people, things, and information. The evolution of cities is characterized as the feedback loop of those interactions. It is necessary to think of cities not just as artifacts but living organisms. However, such complex urban dynamism makes it difficult to understand and predict the states of each area of a city, and it causes many kinds of urban problems such as urban sprawl, traffic congestion, rapid increase in land price, crime, air pollution, and lack of jobs. In recent years, many cities in the world are faced with rapid changes due to the increase of population in emerging countries and the aging population in developed countries. It is getting more and more difficult to understand cities or plan effective urban policies.

The World Bank [127] and the United Nations [112, 113] recently emphasize the scope to utilize mobility data for the sustainable development of cities. The availability of urban data is growing rapidly, that ranges from telecommunication data [95] and night light data [78] to social media data [99]. In addition, smart card data of public transportation has been used for identifying deprived areas within London, UK [104], and mobile phone data in Rwanda has been utilized for understanding patterns of local migration within the country [18].

Among those urban data, human mobility data are the most important data to understand cities as complex systems. Human mobility data are recently becoming more available due to the automatic data collection from mobile phones, smart cards of public transportation, and various kinds of sensors. Michael Batty, who has analyzed cities based on the systems theory, notes in his book titled “The New Science of Cities” [9], that it is necessary to understand flows or relations between places to understand a place in a city. In the case of biology, we need to know the flows of blood or nerve signals to understand a part of a body. Similarly, it is necessary to understand a place in a city by analyzing flows in a city. Human mobility is the most important flow to understand urban systems, as blood circulation is the most important to understand the system of a body.

We consider that the biggest challenges for urban planners are as follows: (1) The difuculty in understanding the current states of areas of a city; (2) The difficulty in predicting the future states of areas of a city; and (3) The difficulty in capturing the community structure of a city. It is necessary to understand the current situation of areas to find problems in urban spaces. Also, it is necessary to predict the outcome of urban planning when investing in urban development. Finally, it is necessary to understand the functional clusters of areas of a city for sustainable development.

We need to create models of urban systems in relation to human flows, and we need to develop methods to understand the above facors of cities based on human mobiity data.

1.1.2 Transportation data for the analysis of urban systems

There are various types of human mobility data. GPS data collected from mobile phones have been used for the analysis of human mobility. Gonzalez et al. [42] show that human moblity has high regularity and predictability. GPS data has detailed information about the coordinates and time. Such information is useful for the analysis of individuals’ movements. Smart card data of public transportation have also been used in previous studies. Smart card data has information about boarding time, alighting time, boarding station, and alighting station.

Transportation data are more suitable for the analysis of spatial interaction in urban spaces than GPS data for the following two reasons: First, it is likely that a trip recorded in smart card data of public transportation is caused by a specific trip purpose. People move by train to participate in a specific activity at a destination such as shopping, meeting with other people, working, and schooling. However, GPS data collected from mobile phones include physical movements that are not caused by spcific trip purposes. Second, it is necessary to discretize the space when using GPS data to analyze spatial interaction in urban spaces. GPS data have detailed information about coordinates. It is necessary to divide the space into grid cells or Voronoi polygons to discretize locational information. On the other hand, the locational information of the smart card data of public transportation is already discretized. Furthermore, commercial areas and residential areas develop around staions in a society where people move mostly by train. The locational information of transportation data is more suitable for the analysis of spatial interaction than the discretized locational information of GPS data.

For the above reasons, transportation data are the most suitable human mobility data to analyze cities as systems.

1.2 Problem statements, hypotheses, and research aims

To create models of urban systems, it is necessary to make hypotheses about the relationships between human mobility and other factors of urban systems. After developing the models, the models should be tested using human mobility data. This research aims at developing three models of urban factors in relation to human mobility, (1) urban activities in each area of a city, (2) population growth of each area, and (3) sub-systems of urban space and their boundaries.

1.2.1 Urban activities as the characteristics of places

Urban activities in a place (e.g. shopping, working) are related to trip purposes of visitors. Therefore, analyzing human mobility can help us to better understand urban activities in each area.

This research hypothesizes that urban activities can be understood by analyzing the temporal distributions of visitors arrivals. For example, the number of commuters' arrivals at destinations is likely high around 8 AM. This research attempts to develop a method to infer urban activities in each place from the temporal distribution of visitors' arrivals.

1.2.2 Population growth as the evolution of cities

Human mobility reflects the attractiveness of destinations and the amenities and convenience of origins. We consider that an area with rich amenities is likely to be more populated in the future. In addition, the amenities of residential neighborhoods likely decrease the need to move to distant places on non-working days. Therefore, the population growth of areas can be estimated based on human moblity.

This research hypothesizes that it is possible to predict the number of future in-migrants based on human mobility on non-working days. We create a model to predict future residential mobility based on non-working human mobility.

1.2.3 Sub-systems of urban systems and their boundaries

Areas of a city are functionally linked to each other. For example, a residential area is likely linked to commercial areas or business areas located near by. Such functional linkages form a cluster of places. We call it a sub-system of an urban system. Discovering sub-systems of urban systems is important to plan effective urban policies. We assume that the functional linkage can be estimated by human mobility. We attempt to develop an effective method to find sub-systems of urban systems based on human mobility data.

1.3 Chapters and contributions

This dissertation is organized into a review of related studies (Chapter 2), the framework of this study (Chapter 3), the explanation of data used in this dissertation (Chapter 4), three main research results (Chapters 5, 6, and 7), the possible application of the proposed methods to real problems (Chapter 8), and conclusions (Chapter 9). Chapter 5 attempts to develop a method to estimate urban activities of each area of a city. Chapter 6 explores the causal relationships between human mobility and population growth in each area of a city. Chapter 7 examines network clustering methods for extracting the community structure of a city.

Chapter 2

Related studies

This chapter reviews previous studies on human mobility patterns and urban systems. Many studies on urban systems are based on the assumption that human mobility has high regularity. Such an assumption has been confirmed by recent studies using a large amount of locational data collected from mobile phones (Section 2.1). Spatial interaction models are another important strand of studies. Interactions between areas of a city such as residential movements and human movements have been modeled by extending theories of physics (Section 2.2). Another strand of studies has focused on spatial proximity between areas of a city, which has been considered as one of the most important factors for the analysis of the growth of each area of a city (Section 2.3). We consider that smart card data of public transportation enable us to analyze the areas of a city not relying on physical distance, and they allow us to analyze a city with topological perspectives. We consider such an aproach can fill in gaps in previous studies (Section 2.4).

2.1 Human mobility patterns

Studies on human mobility patterns date back to the analysis of albatrosses' wandering pattern. Viswanathan et al. [116] show that wandering albatrosses' flight-time intervals follow the power-law distribution. Lévy flights are a special class of random walks whose step lengths follow the power-law. Their interpretation is that Lévy flights are suitable for searching for food, and animals' movements can be regarded as Lévy flights. The work is followed by the studies on marine predators search behaviour by Sims et al. [103]. They also find that marine predators' search behaviour follows Lévy flights. It is extended to human mobility by Brockmann et al. [21]. Their study analyzes the movements of bank notes. They show that the movements of bank notes can also be regarded as Lévy flights. Therefore, actual human mobility was inferred to be the same as Lévy flights.

However, the supposition has been denied by González et al [42]. Their study analyzes GPS data collected from mobile phones of 100,000 individuals for six months. The mobility pattern of each person shows both spatial and temporal regularity, mainly between home and work/school. The direction of the mobility pattern of a person is limited, and the length of movement does not change. In regard to the difference of mobility patterns between different people, the lengths of different people's movements follow the power-law distribution. Therefore, the fact that the movements of bank notes follow Lévy flights is because of the difference of mobility patterns between different people. The regularity of human mobility patterns is now considered as a stylized fact though the influence of social ties and geographical factors on human mobility has been noted [27, 88].

2.2 Spatial interaction models

Human flows have been modeled by analogy with Newton's law of universal gravitation. The gravitation $F_{i,j}$ between two masses m_i and m_j separated by a distance $d_{i,j}$ is described as

$$F_{i,j} = \gamma \frac{m_i m_j}{d_{i,j}^2} \tag{2.1}$$

where γ is a constant. The transportation gravity model [125] is then described as

$$T_{i,j} = k \frac{O_i D_j}{d_{i,j}^2} \tag{2.2}$$

where k is a constant. Assume that $T_{i,j}$ is the number of trips from place i to place j, that satisfies the following equations:

$$\sum_j T_{i,j} = O_i \tag{2.3}$$

$$\sum_i T_{i,j} = D_j \tag{2.4}$$

In these equations, O_i denotes the total number of trips from place i, and D_j denotes the total number of trips to place j. However, these equations do not meet Equation 2.2. In addition, the reductionism behind Equation 2.2 has been doubted by sociologists. For example, migration from place i to place j is affected by the opportunities in place k that is located between place i and place j. A sociologist, Samuel Stouffer has proposed the theory of intervening opportunities that is stated as follows:

> *The number of persons going a given distance is directly proportional to the number of opportunities at that distance and inversely proportional to the number of intervening opportunities.* (Samuel Stouffer, Theory of Intervening Opportunities [106])

Wilson [125] addresses the above criticism by extending the transportation gravity model based on an analogy with a different branch of physics, statistical mechanics. The newly defined gravity model is given as

$$T_{i,j} = A_i B_j O_i D_j f(d_{i,j}) \tag{2.5}$$

where $f(d_{i,j})$ is a function of distance, and A_i and B_j are determined by repeating a calculation using the following equations:

$$A_i = \left[\sum_j B_j D_j f(d_{i,j}) \right]^{-1} \tag{2.6}$$

$$B_j = \left[\sum_i A_i O_i f(d_{i,j}) \right]^{-1} \tag{2.7}$$

Wilson [125] shows not only that the model satisfies Equation 2.2 but also that the number of intervening opportunities can be defined by the model.

It is possible to infer the number of human flows between two places using information about total in-flows and out-flows, based on the model.

2.3 Models of urban growth

Cellular automaton models and agent-based models have been used for analyzing urban growth [8]. Given the initial state of each entity and the rules of interaction between entities, the models enable us to simulate the growth of cities.

Cellular automaton models [126] discretize a space by dividing it into grid cells, and also discretize the time. Each cell has a discretized state, which is changed by the states of the neighboring cells. The typical definitions of neighborhood are shown in Fig. 2.1, where red cells denote the neighborhood of the blue cells.

The changes of land use (e.g. residential area, commercial area) have been modeled using cellular automaton models with GIS data [124, 29, 10].

Agent-based models place numerous agents and assume the behavioral rules of the agents. Agents react to the environment, and the environment is also affected by agents' behaviors. The model can simulate the feedback loop between agents and environments.

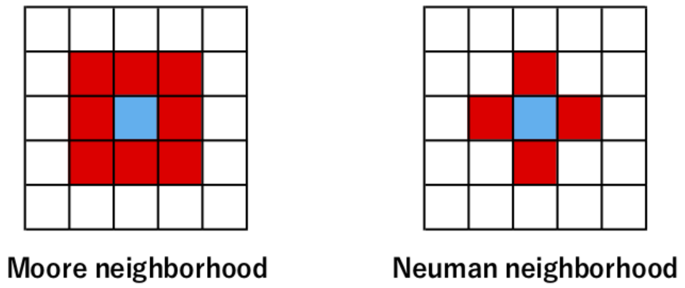


Fig. 2.1 Definitions of neighborhood in two-dimensional cellular automata.

For example, Hasse et al. [45] model the residential choices of individuals, the land use of neighborhoods, and the interaction between the environment and agents' behaviors. Finally, their study simulates residential mobility in a shrinking city. In this manner, land uses have been analyzed by using mutlti-agent based simulation [90].

2.4 Gap in the literature

The exisiting models of urban growth assume that information about the internal state of each area of a city (e.g. land use) is given. The number of human flows can be estimated by spatial interaction models, and the future state of each area can be predicted by cellular automaton models and agent-based models. However, such information about the state of each area is hardly available. In turn, the availability of human mobility data is rapidly growing because of automatic data collection from smart cards and mobile phones.

Little has been explored about the following points by using only human mobility data (e.g. smart card data) without information about the state of each area of a city:

- Inferring the current state of each area of a city.
- Predicting the future state of of each area of a city.
- Discovering the community structure of a city.

This research is dedicated to addressing the above points.

Chapter 3

Framework and definitions

3.1 Urban systems modeling

The science of “sytems” was originally developed in Ludwig von Bertalanffy’s General System Theory [117] in biology and Norbert Weiner’s Cybernetics [123] in engineering. The science of systems is employed in many other domains of research. Rather than aiming to develop a theory that explains everything from the perspecive of reductionism, it focuses on the relationships between elements to understand the dynamism of a system. Many studies of cities have also employed simulation to understand a city as a complex system.

The most important key concepts in the systems theory are elements, interactions, evolution, sub-systems, and boundaries. In urban spaces, those factors are defined as follows:

- **Elements (Places) :** Each place (each area of a city) is regarded as an element (unit) of urban systems. The characteristics of each place can be understood in relation to human mobility. For example, a business area is an area commuters move in the early morning on working days.

- **Interactions (Human mobility) :** Human flow is the interaction between areas of a city. It is the most important factor for understanding cities as complex systems.

- **Evolution (Population growth) :** The evolution of cities is characterized by factors such as the population of residents, the number of visitors, and amenities in an area of a city. We consider that population growth is an especially important factor among the factors.

- **Sub-systems (Clusters of places) :** A system is constituted of a set of sub-systems. Recent studies in network theory assume that a sub-network in a network is one in

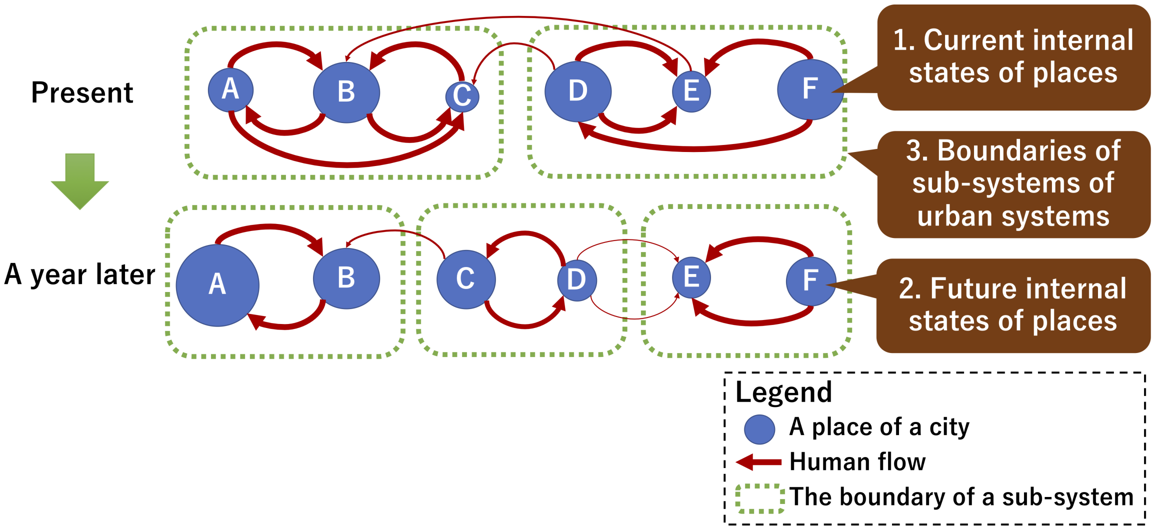

Present
A year later
A
B
C
D
E
F
1. Current internal states of places
3. Boundaries of sub-systems of urban systems
2. Future internal states of places
Legend
A place of a city
Human flow
The boundary of a sub-system

Fig. 3.1 Factors of urban systems and the study targets of this research.

which internal interactions are dense. The sub-systems of urban spaces are defined as the set of places whose spatial and functional roles are similar.

- **Boundaries (Urban boundaries) :** The places of a sub-system of urban spaces are located near each other, and the sub-system of urban spaces shows a geographical boundary.

3.2 Study targets of this research

Fig. 3.1 shows the factors of urban systems and the following study targets of this research.

1. Current internal states of places in urban spaces (Chapter 5)
2. Future internal states of places in urban spaces (Chapter 6)
3. Sub-systems of urban systems and their boundaries (Chapter 7)

In this dissertation, current states of places indicate urban activities. Future states indicate future population growth.

Chapter 4

Data

4.1 Data description

We use the smart card data of public transportation in the Kansai Area of Japan (Fig. 4.1) in Chapters 5, 6, and 7. The Kansai Area is the second-most populated region in Japan after the Greater Tokyo Area. The data were collected at the auto fare collection barriers in each station. The original data have trip records with information on trip ID, passenger ID, boarding time, boarding station, alighting time, and alighting station.

The study period and stations differ according to chapters. The study in Chapter 5 was conducted in the early stage of this research project, and we have used the data that were collected at 723 stations across 6 railway companies shown in Fig. 4.2. The studies in Chapters 6 and 7 use data collected at 1,024 stations run by 14 railway companies and 3 agencies of city governments, shown in Fig. 4.2. The average number of passengers in the data is 1,087,351 per day, and the average number of trips is 2,477,966 per day.

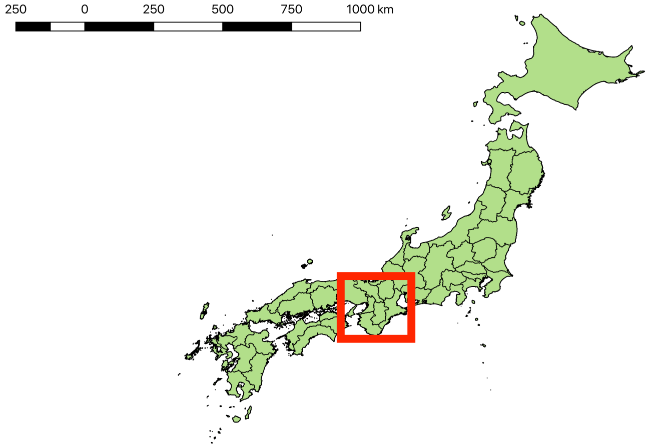


Fig. 4.1 The study area of this dissertation.

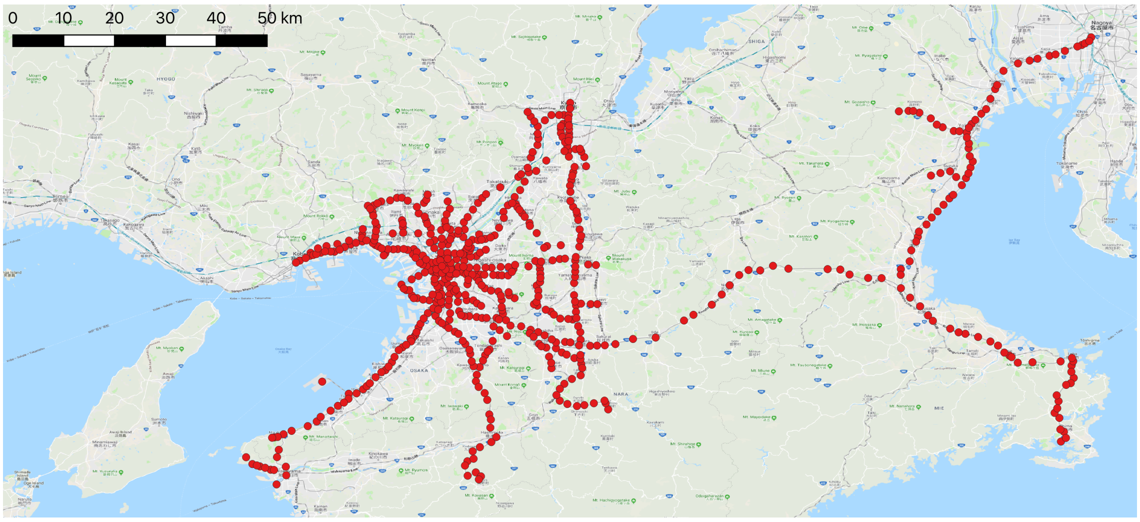


Fig. 4.2 The locations of the 723 railway stations in the data used in Chapters 5.

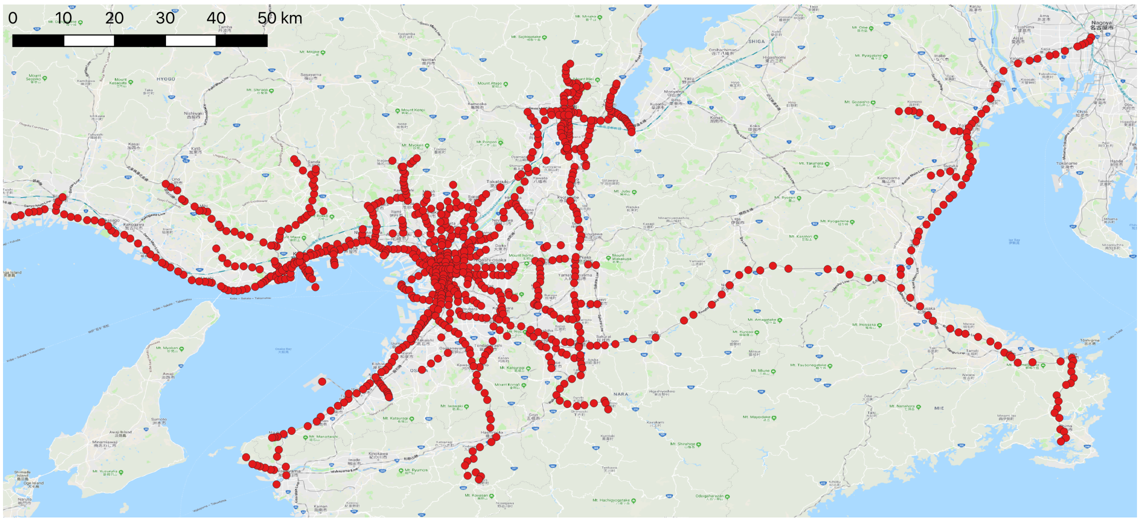


Fig. 4.3 The locations of the 1024 railway stations in the data used in Chapters 6 and 7.

4.2 The advantages of the smart card data of public trasportation

Our study focuses on interaction between areas of a city. Smart card data of public transportation have advantages compared to GPS data collected from mobile phones.

First, trips recorded in smart card data of public transportation are caused by specific trip purposes. We consider that people move by train in order to participate in activities such as shopping or working. However, GPS data include every physical movement. Therefore, we consider GPS data are not suitable for the analysis of urban activities.

Second, the information about locations in the smart card data of public transportation is discretized whereas geographical information in GPS data is continuous. In order to discretize geographical information in GPS data, it is necessary to divide the space into grid cells or polygons. Commercial areas and residential areas develop around staions in a society where people move mostly by train. The locational information of transportation data is more suitable for the analysis of spatial interaction than the discretized locational information of GPS data.

For these reasons, smart card data of public transportation are suitable for the analysis of urban systems.

Chapter 5

Understanding Urban Activities and Changes

5.1 Introduction

Trips within and between cities are manifested through the need to access places and participate in activities [57]. They range from the daily commute to workplaces to ad hoc excursions. Analyzing the patterns of such human mobility marks an important step toward understanding human activities in a city and, thereby, helps planning and managing public transport and road traffic. To this end, urban planners, social scientists and real-estate developers have often relied on travel survey data (e.g. Axhausen et al. [4]). While survey data reflects human mobility and sheds light on the trends of human activities in a city, they are costly and time-consuming to collect.

In recent years, mobility data from smart cards, mobile phones and sensors have become increasingly available [132]. The process of collecting such data is largely automatic and requires minimum effort, yet they reflect real-time human movements within urban space. Given the extensive and comprehensive nature of mobility data, they are expected to help contribute to the planning and improvement of cities in developed and developing countries alike. Indeed, the World Bank [127] and the United Nations [112, 113] emphasize the need to utilize such mobility data for covering population in developing countries. These include cases of the TfL (Transport for London) Oyster Card data being used for identifying deprived areas within London, UK [104], and mobile phone data in Rwanda being utilized for understanding patterns of migration within the country [18].

However, mobility data often lack some of the key information, and these include the purposes of trips for each individual user. Information on trip purposes is crucial for estimating the number of visitors for each type of activity (e.g. commuting, leisure), and their changes over time would help identify the characteristics of each area and how it is changing.

Previous studies have estimated trip purposes, land uses, and activities using individual travel patterns [1, 47] and other secondary information such as household survey data and POI (point of interest) in the area [68, 2]. However, these datasets also tend to be unavailable or inaccurate. In order to estimate the trip patterns and monitor urban changes without such limitations, a new method is needed.

Changes in urban space such as the development of new commercial facilities often have different impacts on human mobility depending on the trip purpose (e.g. commuting, leisure). This makes it crucial to understand the break-down of the number of human flow to understand urban changes. By the same token, decomposing human flow in a city would enable us to detect and understand such urban changes.

This chapter investigates a method for estimating the numbers of visitors for different trip purposes using mobility data collected through passively monitoring the passenger flow that reflects the actual human movement (e.g. smart card data of public transportation). In this study, we use the smart card data of public transportation in Japan. Specifically, we propose a method called EAT-CD (Extraction of Activity Types and Change Detection) that estimates the volume of passengers by each activity and detects changes in the number of visitors for each activity; e.g. increase in shopping trips triggered by the development of a new commercial facility.

Our research questions are summarized as below.

- **RQ1**: Can a method be developed to estimate the number of visitors to each place by their trip purposes using mobility data (e.g smart cards records from public transportation) without individual trip records?

- **RQ2**: Can a method be developed to detect changes in activity trends from the estimates of the number of visitors by each activity?

The remainder of this chapter comprises the following. Section 5.2 reviews previous studies, followed by the introduction of the proposed method (Section 5.3) for estimating the volume of trips by activities, and detecting changes in the trend of urban activities. Section 5.4 explains the dataset used and the analysis carried out in the empirical study. Section 5.5 discusses the results of the analysis, and Section 5.6 is the conclusion.

5.2 Related studies

There have been many studies on estimating trip purposes, land uses, and activities, and understanding urban changes, on the basis of human mobility data. The following sections give an overview of studies on inference of purpose, land use, and activity which focus on understanding urban changes, and a gap in the literature and how we might address it.

5.2.1 Inference on trip purposes, land uses, and activities

Trip purpose, land use, and activity are known to affect one another, which means that studies on their inference also have a considerable overlap, employing similar types of data and methodologies.

Inference based on individual travel patterns and additional secondary data

Individual travel patterns have been used in several studies for inferring the respective trip purpose [1, 47, 53, 122, 130, 138]. For instance, Alexander et al. [1] infer trip purpose from call detail records (CDRs) which are collected through the use of mobile phones that contain time-stamped geo-coordinates. They estimate the location of each mobile phone user and classify them into their home, work, and other places depending on the frequency of observation, day of the week, and time of the day. Han et al. [47] propose a method to derive the sequence of activities for each trip chain using a continuous hidden Markov model (CHMM) against smart card data and land use characteristics.

Others use POI data [20, 121, 128, 135], as they tend to offer information on specific trip purpose, land use and activity. For instance, Wang et al. [121] infer subway station functions by applying the Doc2vec model [65] to smart card data and POI data. Zhong et al. [135] propose a method for inferring building functions in Singapore using smart card data and POI data. Their study applies a Bayesian model on the passengers' mobility patterns to estimate their trip purposes. It then infers building functions by linking daily activities to the buildings surrounding the stops based on spatial statistics.

Supplementary information such as land use data and household survey data is also used for inferring land use, activity, and trip purpose [68, 2, 72]. For example, Long et al. [72] combine smart card data with household travel surveys, as well as a parcel-level land use map to identify job-housing locations and commuting trip routes in Beijing.

The above studies use detailed individual trip records which are usually not open to the public. In addition they use additional secondary data that tend to be inaccurate or not frequently updated. In this study, we develop a method that uses hourly number of human inflow in each area.

Inference based on patterns of temporal distribution of population or trips

Another strand of literature has focused on inferring the land uses and activities within each area using patterns of temporal distribution of population and trips. The advantage of using such methods is that they can identify daily land use of each area without any additional information such as POIs. For instance, Nishi et al. [87] extract area-by-area and daily land use patterns using location data obtained from mobile phones. Their method creates a 24-dimesional vector for each area and each day, which retains information about hourly numbers of human inflow. Each vector is then normalized by dividing all 24 elements by the total number of the elements. Their study uses the infinite Gaussian mixture model (GMM) which incorporates the Dirichlet process (DP) to cluster the vectors, which are labeled collectively with the respective land use type. Similarly, Frias-Martinez et al. [39] use the number of posts of Twitter users for clustering the same land use types within each area. Each area is vectorized by 144 elements, namely the numbers of posts every 20 minutes on weekdays and weekends. Land use of each area is estimated by applying spectral clustering to the vectors. Chen et al. [26] propose to delineate areas with urban functions based on social media data aggregated to the building-block level. The underlying assumption is that social media activities in buildings of similar functions will likely share similar spatiotemporal patterns.

The above studies identify land uses / activities without individual trip records or any additional information. However, they label single land use or activity to each area, and their methods cannot capture mixed land use. Therefore, our study develops a method for capturing multiple activities in each area.

5.2.2 Analyses of urban changes

Use of mobility data for interpreting changes in urban space and its usage is becoming an increasingly prominent research topic. Studies on the topic often apply analytical methods based on machine learning or spatial networks to measure changes in the landscape and usage of urban environment.

These studies are often confined by the limited range of attributes available in the mobility data. In order to estimate detailed patterns and changes in urban space, we need to find a way to decompose the human flows or population into different categories of activities and travel purposes such as commuting and leisure. Several studies have focused on this aspect.

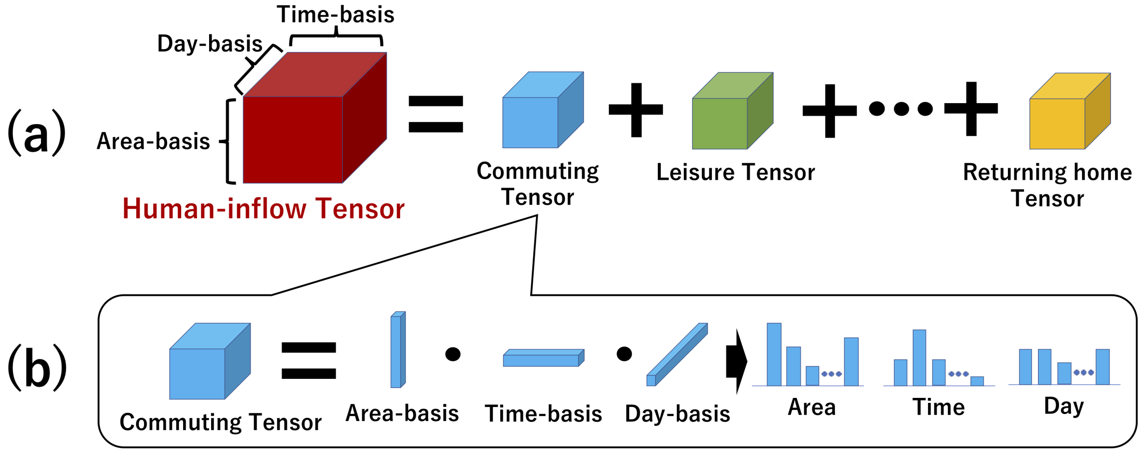


Fig. 5.1 Application of non-negative tensor factorization to human mobility data proposed in a previous study [37].

For instance, Fan et al. [37] propose a tensor factorization approach to modeling city dynamics. The study utilizes non-negative tensor factorization (NTF) [28] to decompose a human flow tensor obtained from GPS log data into basic life pattern tensors such as commuting, working, and entertaining. It applies the method for modeling the fluctuation in human flow before and after the Great East Japan Earthquake. Another study (Wang et al. [120]) models time-evolving traffic networks into a 3-order origin-destination-time tensor, which detects the spatial clusters, temporal patterns and the associations among such networks.

Spatial network analysis is another important method for detecting the dynamics of urban structure. Zhong et al.[134, 137] propose quantitative measures to evaluate the centrality of locations. Their study applies this method to mobility data collected in Singapore and conclude that the city-state has been rapidly transforming to adopt a polycentric urban form.

In our study, we develop a method to detect changes in each type of activity (e.g. commuting and leisure) within each area of a city. For example, the number of visitors for shopping would increase in a place where a shopping center opens. Our study aims to detect such changes by estimating the break-down between such activities within each area.

5.2.3 Gap in the literature

Although various studies have attempted to estimate trip purposes and land uses as well as changes in human activities and urban landscape in general, to our knowledge, no study has proposed a method for measuring and detecting changes in the trend of activities within each area.

Our research was originally motivated by Nishi et al. [87] and Fan et al. [37]. In Nishi et al. [87], the daily land use of each area is derived from the number of population recorded at each hour. However, most areas in a city may have mixed land use and activities, and their method would not be suitable for estimating the break-down between such activities within each area. On the other hand, Fan et al. [37] decompose human population into several activities by using non-negative tensor factorization (Fig. 5.1). Their tensor consists of area-basis, time-basis and day-basis, which respectively denote population by area, hour and day. This tensor is then converted and decomposed into a set of tensors. Each of the new tensors represents an activity type such as commuting and leisure (Fig. 5.1-(a)). It allows us to estimate the temporal and spatial distribution of each activity (Fig. 5.1-(b)), and have an overview of the trend of each activity across the entire study area. However, as their method estimates the temporal and spatial patterns of each type of activity at the aggregate level only, it is impossible to see the trend of each activity for each area separately. For instance, the case study discussed in their study features the impact of the Great East Japan Earthquake on human activity, but the method cannot show area-specific trends.

To understand the trend of each activity by area in a city, we need a new method that satisfies the following criteria: (1) it incorporates the temporal patterns of the distribution of the population within each area; (2) it decomposes the distribution of population into some activity types; and (3) it shows the chronological changes in activities for each area.

This chapter aims to propose a method that estimates the activity trends within each area of a city by extending the method proposed by Fan et al. [37]. The method will also be designed to detect changes in each activity for each area. The method will be tested with an empirical case study that utilizes smart card data of public transportation in western Japan, but the proposed method is applicable to any other mobility data including GPS log data.

5.3 Methods

In this section, we propose a method called EAT-CD (Extraction of Activity Types and Change Detection) for detecting and estimating changes in the trends of activities within each area of a city. Fig. 5.2 shows the procedure involved in carrying out EAT-CD. First, EAT-CD decomposes the temporal distribution of the numbers of visitors' arrivals at each place for each day into a set of activity types such as commuting and leisure (Fig. 5.2-(1)). The underlying assumption is that such distribution is represented by the linear sum of these basic distributions. This allows us to construct the trend of each activity type in each place (Fig. 5.2-(2)). EAT-CD automatically detects changes in the trend of each activity type in each place. EAT-CD may detect changes caused by events such as the construction of a new commercial facility and the beginnings and the ends of school terms (Fig. 5.2-(3)).

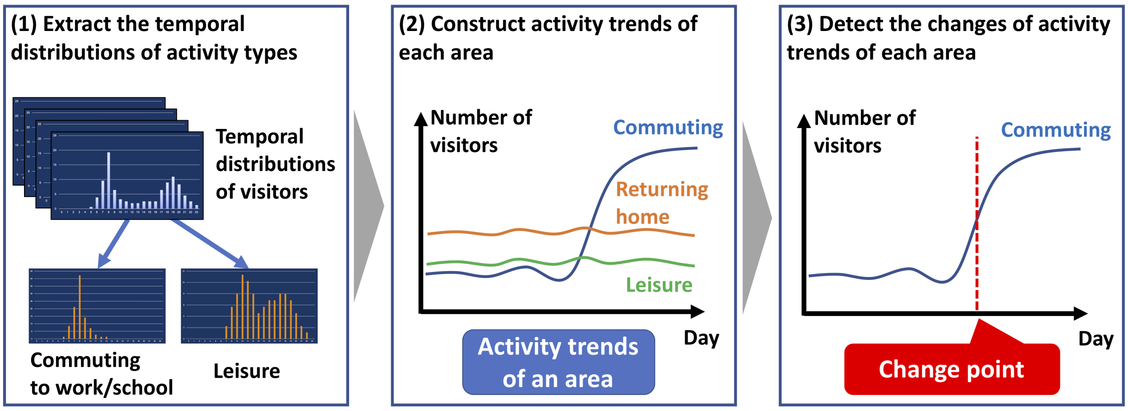

(1) Extract the temporal distributions of activity types
Temporal distributions of visitors
Commuting to work/school
Leisure
(2) Construct activity trends of each area
Number of visitors
Commuting
Returning home
Leisure
Day
Activity trends of an area
(3) Detect the changes of activity trends of each area
Number of visitors
Commuting
Day
Change point

Fig. 5.2 A flow-chart of the processes involved in carrying out our proposed method.

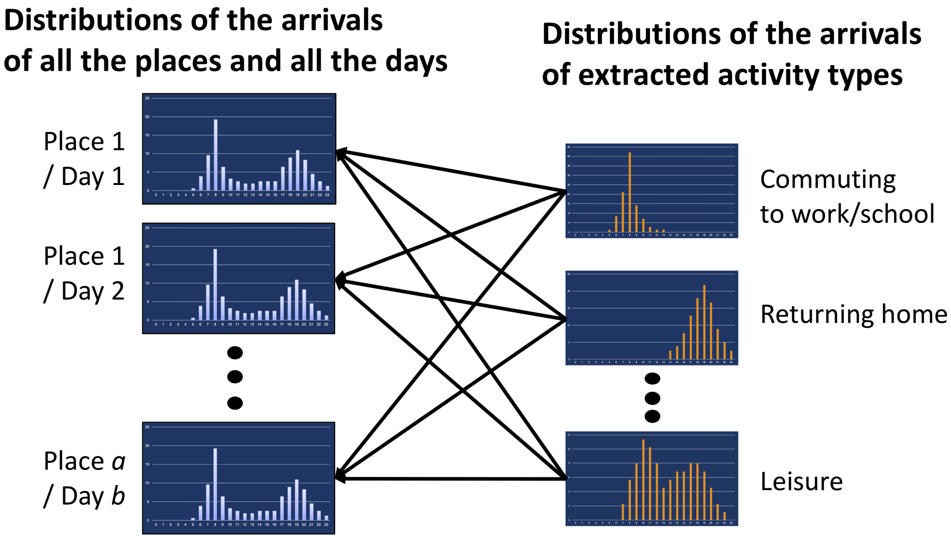

Distributions of the arrivals of all the places and all the days
Distributions of the arrivals of extracted activity types
Place 1 / Day 1
Place 1 / Day 2
Place a / Day b
Commuting to work/school
Returning home
Leisure

Fig. 5.3 Decomposition of the number of visitors' arrivals into different activity types.

5.3.1 Decomposing the temporal distribution of visitors' arrivals into activity types

We assume that the temporal distribution of the number of visitors' arrivals at each place for each day consists of superposition of multiple basic distributions, as shown in Fig. 5.3. For example, if the area containing a railway station is characterized as a mixture of a business area and a residential area, the number of passengers' arrivals on a weekday may have an acute concentration around 8 AM to 9 AM, and a more gradual increase around 6 PM. The distribution can be regarded as the superposition of two basic distributions, namely the passengers commuting/schooling to this area in the morning and the passengers returning to their home in this area in the evening. The temporal distribution of passengers' arrivals

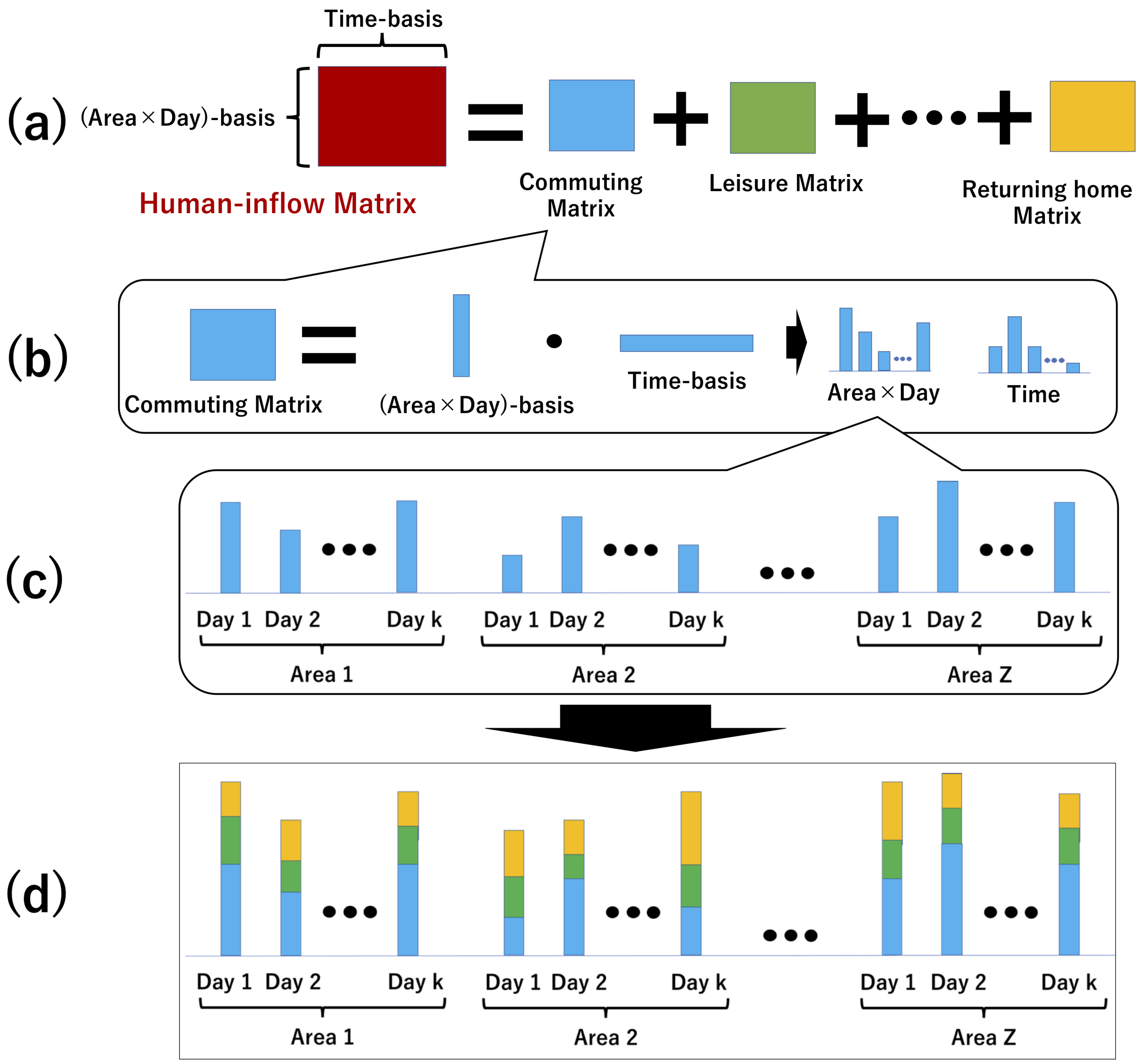


Fig. 5.4 Proposed application of non-negative matrix factorization to human mobility data.

at every place on every day can be expressed as the linear sum of a finite number of distributions that collectively represent the activity types such as commuting/schooling and leisure.

Fig. 5.1 illustrates the framework of the method proposed by Fan et al. [37]. By extending their method, we propose the method detailed in Fig. 5.4. First, we create a matrix which consists of (area×day)-basis and time-basis, and each element of the matrix denotes the number of visitors to an area during one time-slice in each day. Each row denotes an aggregate of all the visitors to one area in a day across all time-slices, and each column denotes visitors to all areas during one time-slice in a day. This matrix is then decomposed into a set of matrices. Each of the new matrices represents a unique activity type such as commuting or leisure (Fig. 5.4-(a)). Each matrix is represented as a matrix product of a single-column vector and a single-row vector (Fig. 5.4-(b)). The single-column vector provides information about the number of people who visit each area each day for a particular activity (Fig. 5.4-(c)). In this manner, we obtain the number of visitors to each area for each activity. Finally, we obtain the break-down of the number of visitors arrival (Fig. 5.4-(d)). Our method uses non-negative matrix factorization (NMF) [67] for decomposing the number of visitors.

The objective of non-negative matrix factorization

The purpose of non-negative matrix factorization is to represent a matrix as the product of two non-negative matrices. The variables we use here are listed in Table 5.1. We create matrix $\mathbf{V}$ that has information about the number of visitors for each place, each day, and each time-slice. We discretize the time by dividing one day into n time-slices. If the length of one time-slice is 1 hour, then $n = 24$. Let a denote the number of all places in the data, b denote the number of days in the data, and define m as $m = ab$. Let $m \times n$ matrix $\mathbf{V}$ denote the numbers of visitors' arrivals at all a number of places for b number of days across n instances of time-slices (i.e. each row vector of $\mathbf{V}$ indicates a place and a day, and each column vector indicates a time-slice). By using non-negative matrix factorization, Matrix $\mathbf{V}$ is represented as the product of the two non-negative matrices, $\mathbf{W}$ and $\mathbf{H}$: $\mathbf{V} = \mathbf{WH}$, as shown in Fig. 5.5. Each row vector of matrix $\mathbf{H}$ provides a break-down of the distribution of the numbers of visitors' arrivals for a trip activity, and is expressed in a vector form whose sum of all its elements is 1. The number of columns of matrix $\mathbf{W}$ and the number of rows of matrix $\mathbf{H}$ are equal to the number of activity types, and this number is determined arbitrarily. In this research, the number is increased by one unit at a time. When two similar basic distributions are detected, the process of increasing the number is stopped. The number is decided as the maximum value of matrix $\mathbf{H}$ under the condition that no similar data on trip distributions are available.

The algorithm of non-negative matrix factorization

We use an algorithm of NMF originally proposed by Lee and Seung [67] to decompose the temporal distribution of the numbers of visitors' arrivals at each place each day. The following explanation of the algorithm of NMF is based on Sawada et al. [100]. The objective is to minimize the difference between matrix $\mathbf{V}$ and matrix $\mathbf{WH}$. The difference is defined by Equation 5.1.

$$D(\mathbf{V}, \mathbf{WH}) = \sum_{i=1}^{I} \sum_{j=1}^{J} d(x_{ij}, \mathbf{t}_i^{\mathrm{T}} \mathbf{v}_j) \tag{5.1}$$

Table 5.1 Definition of variables.

Variable	Definition
a	The number of places in the data (e.g. the number of stations in the smart card data of public transportation).
b	The number of days in the data.
m	$a \times b$.
n	The number of time-slices in one day. For example, if the length of one time-slice is 1 hour, $n = 24$. If the length of one time-slice is 30 minutes, $n = 48$.
$\mathbf{V}$	A matrix that has information about the number of visitors for each day, each place, and time-slice. Each row indicates a specific day and place. Each column indicates a specific time-slice.
$\mathbf{W}$	A matrix that has information about the inferred number of visitors for each place and each day by activity type. Each row indicates a specific day and station. Each column indicates a specific activity type.
$\mathbf{H}$	A matrix that has information about time-series distribution of visitors for each activity type. Each row indicates a specific activity type. Each column indicates a specific time-slice.

where x_{ij} denotes an element of $\mathbf{V}$, $\mathbf{t}_i$ denotes the i-th row vector of $\mathbf{W}$, and $\mathbf{v}_j$ denotes the j-th column vector of $\mathbf{H}$.

There are three popular definitions of d, namely Euclidean distance [67], Kullback-Leibler distance [67], and Itakura-Saito distance [38]. Euclidean distance is used in our analysis for the sake of simplicity.

$$d_{\mathrm{Eu}}(x_{ij}, \mathbf{t}_i^{\mathrm{T}}\mathbf{v}_j) = (x_{ij} - \mathbf{t}_i^{\mathrm{T}}\mathbf{v}_j)^2 \tag{5.2}$$

Using $\{\hat{x}_{ij} \mid \hat{x}_{ij} = \mathbf{t}_i^{\mathrm{T}}\mathbf{v}_j\}$, matrix $\mathbf{W}$ and matrix $\mathbf{H}$ are obtained by repeating the following calculations until they converge.

$$t_{ik} \leftarrow t_{ik}\frac{\sum_j x_{ij}v_{kj}}{\sum_j \hat{x}_{ij}v_{kj}}, \quad v_{kj} \leftarrow v_{kj}\frac{\sum_i x_{ij}t_{ik}}{\sum_i \hat{x}_{ij}t_{ik}} \tag{5.3}$$

The procedure to introduce Equation 5.3 is as follows:

Equation 5.4 is obtained by omitting constants from Equation 5.1.

$$F_{\mathrm{Eu}}(\mathbf{T}, \mathbf{V}) = \sum_{i,j} \left[(t_{ik}v_{kj})^2 - 2x_{ij}\mathbf{t}_i^{\mathrm{T}}\mathbf{v}_j\right] \tag{5.4}$$

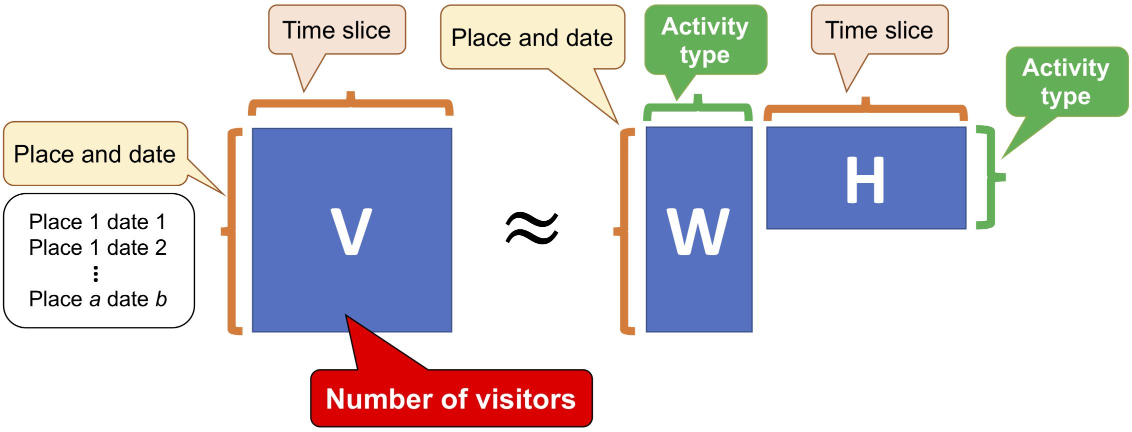


Fig. 5.5 A diagram showing the objective of NMF in this chapter. Matrix **V** indicates the numbers of visitors' arrivals across all places among all time-slices. Matrix **V** is represented as the product of the two non-negative matrices, **W** and **H**.

An auxiliary function, $F^{+}_{\text{Eu}}(T,V,R)$ is defined by adding auxiliary variables r_{ijk} that satisfy Equation 5.6.

$$F^{+}_{\text{Eu}}(\mathbf{T},\mathbf{V},\mathbf{R}) = \sum_{i,j}\left[\sum_{k}\frac{(t_{ik}v_{kj})^2}{r_{ijk}} - 2x_{ij}\mathbf{t}_i^{\mathrm{T}}\mathbf{v}_j\right] \tag{5.5}$$

$$r_{ijk} > 0, \quad \sum_{k=1}^{K} r_{ijk} = 1 \tag{5.6}$$

Function F^{+}_{Eu} satisfies Equation 5.7 and 5.8.

$$F_{\text{Eu}}(\mathbf{T},\mathbf{V}) \leq F^{+}_{\text{Eu}}(\mathbf{T},\mathbf{V},\mathbf{R}) \tag{5.7}$$

$$F_{\text{Eu}}(\mathbf{T},\mathbf{V}) = \min_{\mathbf{R}} F^{+}_{\text{Eu}}(\mathbf{T},\mathbf{V},\mathbf{R}) \tag{5.8}$$

The minimum value of F_{Eu} can be obtained by minimizing F^{+}_{Eu}. Equations 5.9, 5.10, 5.11, and 5.12 are derived by using Lagrange multipliers method.

$$L(\mathbf{T},\mathbf{V},\mathbf{R},\mathbf{\Lambda}) = F^{+}_{\text{Eu}} + \sum_{i,j}\lambda_{ij}\left(\sum_{k} r_{ijk} - 1\right) \tag{5.9}$$

$$\frac{\partial L}{\partial r_{ijk}} = -\frac{(t_{ik}v_{kj})^2}{r_{ijk}^2} + \lambda_{ij} = 0 \tag{5.10}$$

$$\frac{\partial F_{\mathrm{Eu}}^{+}}{\partial t_{ik}} = 2t_{ik}\sum_{j}\frac{v_{kj}^{2}}{r_{ijk}} + 2\sum_{j}x_{ij}v_{kj} = 0 \tag{5.11}$$

$$\frac{\partial F_{\mathrm{Eu}}^{+}}{\partial v_{kj}} = 2v_{kj}\sum_{i}\frac{t_{ik}^{2}}{r_{ijk}} + 2\sum_{i}x_{ij}t_{ik} = 0 \tag{5.12}$$

Equation 5.13 is derived from Equations 5.6 and 5.10.

$$r_{ijk} = \frac{t_{ik}v_{kj}}{\mathbf{t}_i^{\mathrm{T}}\mathbf{v}_j} = \frac{t_{ik}v_{kj}}{\hat{x}_{ij}} \tag{5.13}$$

Equation 5.14 is derived from Equations 5.11 and 5.12.

$$t_{ik} = \frac{\sum_j x_{ij}v_{kj}}{\sum_j \frac{v_{kj}^2}{r_{ijk}}}, \quad v_{kj} = \frac{\sum_i x_{ij}t_{ik}}{\sum_i \frac{t_{ik}^2}{r_{ijk}}} \tag{5.14}$$

Finally, Equation 5.3 is derived from Equation 5.13 and 5.14.

5.3.2 Constructing trends of activities by place

By applying NMF, matrix **W** and matrix **H** are obtained. Matrix **H** has information about the time-series distributions of each activity type (the distributions correspond to the graphs in the right side of Fig. 5.3). The temporal resolution of the distributions is the given length of a time-slice. The time-series distribution of passengers for each station and each day is expressed as the linear sum of the time-series distributions of activity types. Matrix **W** has the information about the correlation to the basic distribution for each station and each day. Using the information, the trends of activity for each place are expressed. The temporal resolution of an activity trend is one day.

Let vector w_i denote the i-th column vector of **W**. The vector contains information on the volumes of visitors' arrivals for one travel activity across all places and during the entire duration of the study period. By taking elements of place s from vector w_i and sorting those elements in chronological order, series $x_{i,s} = \{a_j\}_{j=1}^{n} = \{a_1,..,a_j,..,a_n\}$ is obtained where n is the number of dates included in the data, and j denotes a day. This series represents the trend of the visitors' arrivals for trip activity i at place s.

5.3.3 Change point detection

In this section, we describe how EAT-CD finds the dates when the trend of the visitors' arrivals for each trip activity at each place has changed. Detecting a change point can present

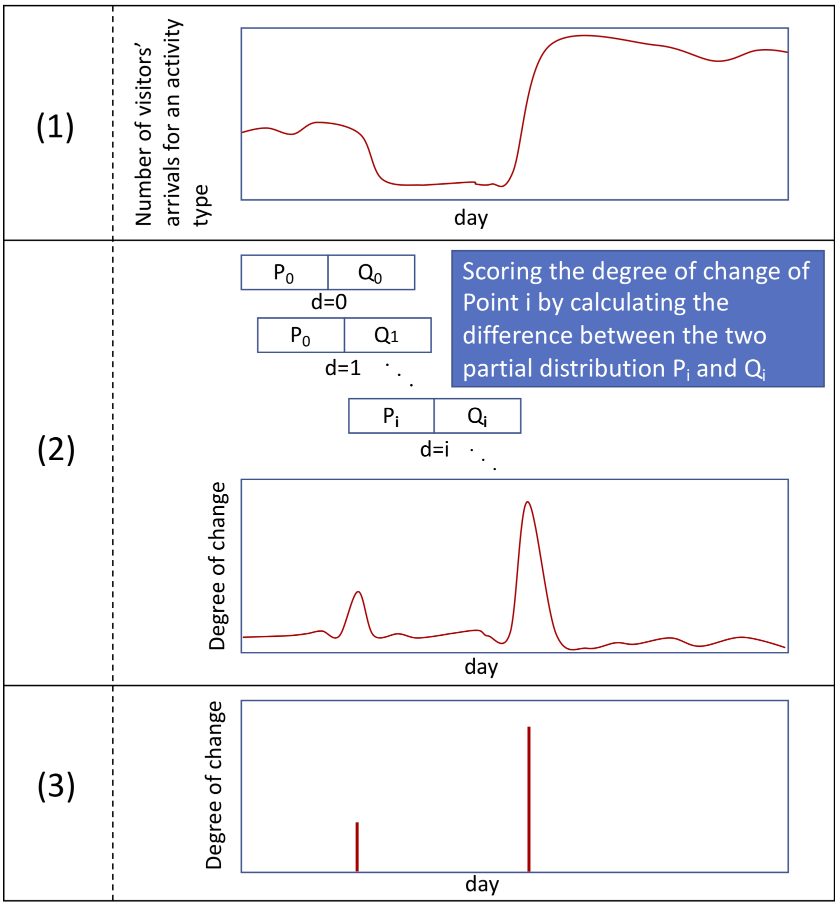

(1)
Number of visitors' arrivals for an activity type
day
(2)
P0
Q0
d=0
P0
Q1
d=1
Pi
Qi
d=i
Scoring the degree of change of Point i by calculating the difference between the two partial distribution Pi and Qi
Degree of change
day
(3)
Degree of change
day

Fig. 5.6 Procedure to detect changes of the trend of activity in an area: (1) The proposed method detects changes of the inferred number of visitors' arrivals for each activity type. (2) The method scores the degree of change of point i by calculating the Jensen-Shannon divergence between the two partial distributions P and Q. (3) The method ignores the scores of points that are located near the maximum points.

a challenge, as sudden changes may occur in a time-series distribution. Many efficient methods have been proposed (e.g. [108, 59, 76]). The objective of the studies is to find change points, and to develop an effective method to record the degree of change for each point. The degree of change can be derived by comparing partial distributions before and after the point.

The objective is to detect the change point from series $x_{i,s} = \{a_j\}_{j=1}^{n} = \{a_1, .., a_j, .., a_n\}$. Assuming that series $x_{i,s}$ draws a time-series distribution as shown in Fig. 5.6-(1), Day d_c is the date when a sudden change has occurred. The change point can be identified by investigating when the greatest degree of change is recorded for $j = c$.

The degree of change of point i is calculated by measuring the difference between series $\{a_j\}_{j=i-r}^{i-1} = \{a_{i-r},..,a_{i-1}\}$ and series $\{a_j\}_{j=i}^{i-1+r} = \{a_i,..,a_{i-1+r}\}$ using Jensen-Shannon Divergence [69] (Fig. 5.6-(2)).

Jensen-Shannon divergence is a method used for measuring the similarity between two probability distributions. It is a modification of Kullback-Leibrer Divergence [64]. The Kullback-Leibler divergence from distribution Q to distribution P is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \tag{5.15}$$

where p and q denote the probability density functions of P and Q, respectively. Assuming $P \sim N(\mu_1, \sigma_1^2)$ and $Q \sim N(\mu_2, \sigma_2^2)$,

$$D_{\mathrm{KL}}(P \| Q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{5.16}$$

Since this is asymmetric with respect to P and Q, Jensen-Shannon divergence is defined as follows to ensure symmetry.

$$D_{\mathrm{JS}}(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M) \tag{5.17}$$

where $M = \frac{P+Q}{2}$. Since M follows $M \sim N(\frac{(\mu_1+\mu_2)}{2}, \frac{\sigma_1^2+\sigma_2^2}{2})$,

$$D_{\mathrm{JS}}(P \| Q) = \frac{1}{2} \log \frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2} + \frac{(\mu_1 - \mu_2)^2}{4(\sigma_1^2 + \sigma_2^2)} \tag{5.18}$$

Assuming that series $\{a_j\}_{j=i-r}^{i-1}$ and series $\{a_j\}_{j=i}^{i-1+r}$ follow the normal distribution, their variances and means are derived. The difference between series $\{a_j\}_{j=i-r}^{i-1}$ and series $\{a_j\}_{j=i}^{i-1+r}$ is calculated using Jensen-Shannon divergence. The score of change of point i is defined as follows:

$$S_i = D_{\mathrm{JS}}\left(\{a_j\}_{j=i-r}^{i-1} \middle\| \{a_j\}_{j=i}^{i-1+r} \right) \tag{5.19}$$

The score increases gradually before d_c and decreases gradually after d_c. Therefore, the exact change point can be identified by finding the peak value; i.e. if there is a point with a greater score than S_i in $[i-r, i+r]$, we ignore S_i (Fig. 5.6-(3)). The score of change can be thus summarized as

$$S_i^* = \begin{cases} S_i & \left(\text{if} \quad S_i = \max\limits_{i-r \le j \le i+r} S_j\right) \\ 0 & (\text{otherwise}) \end{cases} \tag{5.20}$$

Using this method, the change points for each place and each activity are extracted.

5.3.4 Assigning labels to activity patterns and reasons to detected changes

Activity patterns were not defined mechanistically in previous studies (Nishi et al. [87]; Fan et al. [37]). They identified each activity pattern by interpreting the spatio-temporal distributions of the activity pattern. This study adopts a similar approach in that EAT-CD was not designed to extract and label each activity pattern. Rather, it is through the interpretation of the spatio-temporal patterns within the area that will inform us with the initial labelling of the activity types. In addition, while our method detects changes in the travel patterns, it is not intended as a means to explain reasons for such changes. In Section 5.5, we will interpret the activity patterns extracted with EAT-CD and infer reasons for the detected changes to better understand the results.

5.4 Dataset

In principle, EAT-CD can be applied to any type of mobility data that have an hourly visitor count for each location/area, and it does not require detailed information on each trip. To explore the validity of EAT-CD, this chapter applies it to a set of smart card data from public transport in the Kansai Area, Japan.

The time is discretized by dividing one day into n time-slices. In this experiment, we set the length of each time-slice to 1 hour. From this data, we extract the hourly numbers of passengers' arrivals for each station, and we set the numbers to the values of $\mathbf{V}$. We use the information about alighting time and alighting station to extract hourly numbers of passengers' arrivals. We do not use information about trip ID, passenger ID, boarding time, or boarding station in this study. Since every train service ends before 2 a.m., we created daily temporal distribution of passengers' arrivals using data from 3 a.m. to 3 a.m. on the following morning. The period of the data is from April 2013 to March 2015.

5.5 Results and discussions

The daily temporal distributions of passengers' arrivals at all stations are decomposed, allowing us to detect changes in the patterns of activities at each station. This section shows the results of the application, namely the estimated break-down among different trip activities, followed by the activity trends for each station. The pattern of each trip activity is labeled through the observation of the time-series distribution of each trip activity and the overall trends at the main terminal stations. For example, a sudden surge in the passenger volume around 8 a.m. can be classified as commute to work/school. Also, if there is a rise in the number of trips to the residential areas in the evening, they are regarded as the returning leg of commuters.

5.5.1 Basic distributions extracted by NMF

Using EAT-CD, the smart card data of public transport were classified into nine types of trip activities. The number of activity types may differ when this method is applied to mobility data in other areas or countries. Fig. 5.7 shows the temporal distributions of these trip activities. EAT-CD enables us to extract basic distributions and construct activity trends of each station without relying on local land use data. However, identifying each activity type requires either a priori knowledge of people's life style in the region, or a heuristic process to interpret information about the region. In this study, we labeled each distribution as an individual activity by empirically interpreting the shape of the distribution. In this experiment, each distribution is labeled as one activity manually by interpreting the shape of the distribution.

In Fig. 5.7, trip Type 1 represents commute to work/school, showing a sudden increase in the volume during the rush hour at 8 am. Types 2 and 3 are labeled as returning home, which are dominant among stations in residential areas. Type 2 is particularly prevalent in those stations from Monday to Thursday, while Type 3 prevails on Friday. It is considered that workers usually go out for social activities on Friday because they do not need to wake up early the following morning. Types 4, 5, 6, and 7 are dominant among stations that serve as access nodes for concert halls and stadiums. They tend to exhibit a steep rise at a particular time of the day and are less periodic in their chronological patterns. These are labeled as events. The difference in the time of a steep rise reflects the beginning time of events. Types 8 and 9 are labeled as leisure activities. They show a milder change in their volume over the course of the day than Types 4, 5, 6, and 7 do and also tend to resonate with the change in the season. These include the viewing of cherry blossoms and autumn leaves, which are popular activities in Japan during April and September. For such events, visitors do not need to arrive at a specific time. Thus, forming a milder curve in their volume.

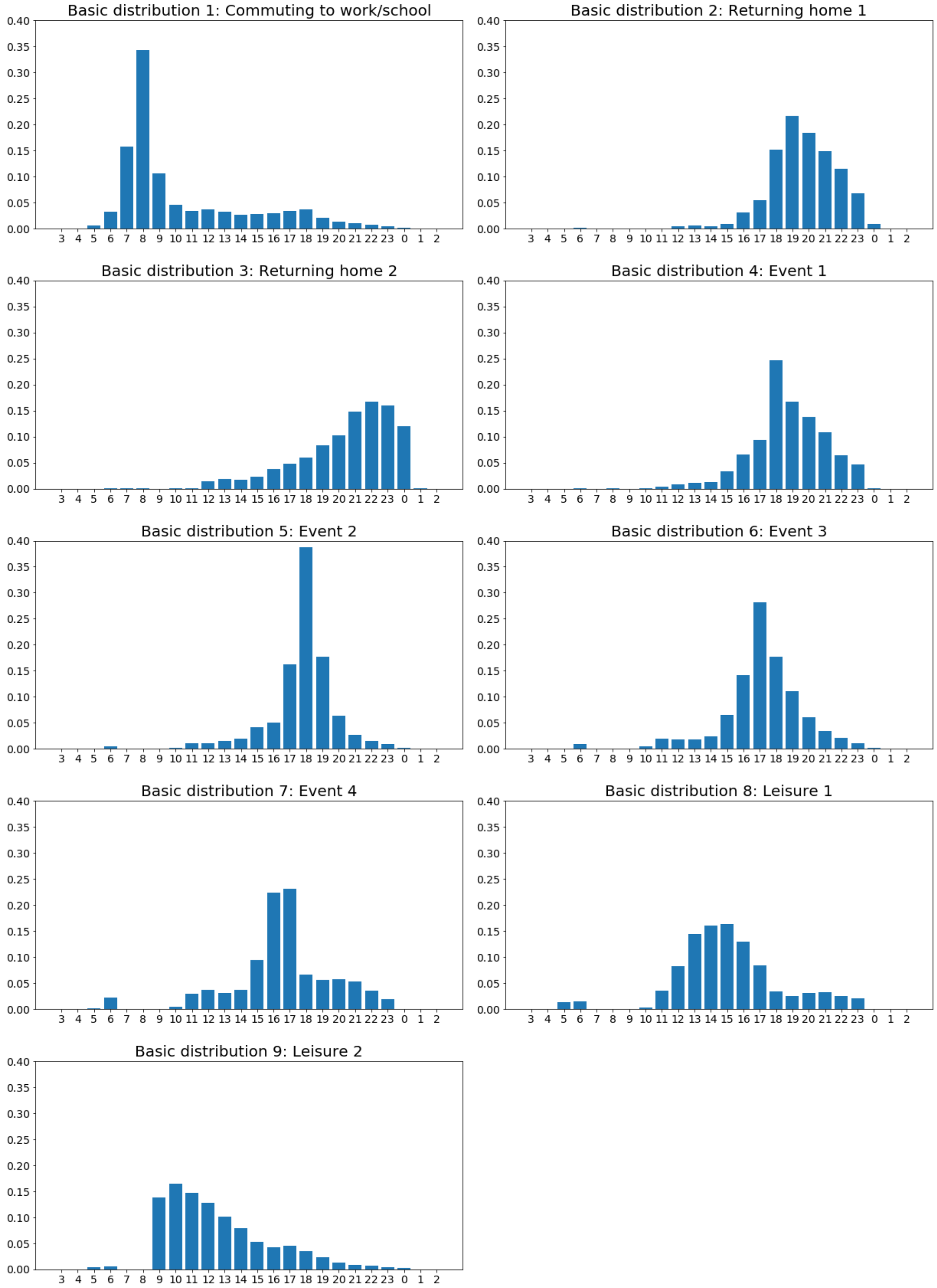


Fig. 5.7 The temporal distributions of the extracted activity types.

5.5.2 Trends of the activity types in the areas of stations

The trends of the trip activities near a station are extracted from **W**. Each column vector of **W** contains information about the volume of the arriving passengers for each travel activity across all places for the entire duration of the study period. To extract the distribution of a single activity type for each station, one column vector of **W** is selected, then elements of the place are selected. We obtain the activity trend by sorting the selected elements in chronological order.

Fig. 5.8 shows an illustrative example of the trends of the trip activities near a station that is located in a business area. In this area, Type 1 (commute to work/school) prevails as the dominant activity. The period between 11 August and 18 August falls on a Japanese Buddhist holiday week and shows a decline in the number of trip activities. There is also a sudden increase for Type 5 trips (Event 2) on 25 July 2014. This is due to a famous annual firework festival held that day along the river near that station, attracting a large number of spectators.

Fig. 5.9 shows another example of trip activities at a station in a residential area. In this area, Types 2 and 3 (returning home) are the dominant activities. On Fridays, the number of trips decreases for Type 2 and increases for Type 3. This is because people often go out for social activities and return home late on Fridays.

Fig. 5.10 shows the trends of trip activities at a station in an area known for the scenic beauty of autumn leaves. Viewing of autumn leaves is a popular activity in Japan. In this area, Type 9 trip (leisure 2) is the most dominant activity, especially on Saturday and Sunday when the number of Type 9 trips shows a rapid increase.

Fig. 5.11 shows the trends of trip activities at a station near a baseball stadium. In this area, Type 6 and 7 trips (events 3 and 4) are the most dominant activities. The dates when the trip activities suddenly increase coincide with the days of baseball matches.

Fig. 5.12 shows the trends of trip activities at a station near a university. In this area, Type 1 trips (commute to work/school) mark the most dominant activity. Type 9 trips (Leisure 2) shows a one-off surge on 23 and 24 November when the university campus hosts a festival.

Fig. 5.13 shows the trends of trip activities at a station in an area of mixed land use comprising business and residential functions. In this area, Types 1 (commute to work/school), 2 (returning home 1), and 3 (returning home 2) are the most dominant activities, and the decomposition process is successful in capturing multiple activities in the area.

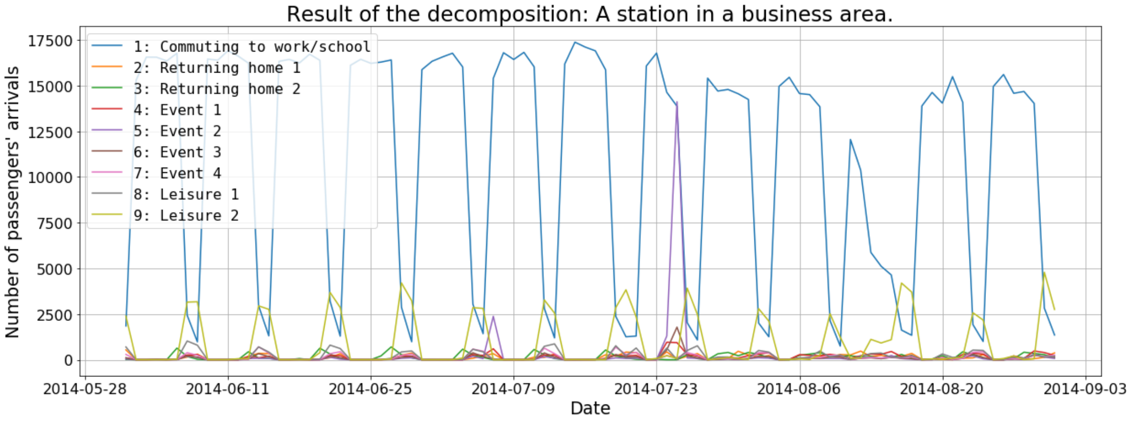


Fig. 5.8 The activity trend of a business area.

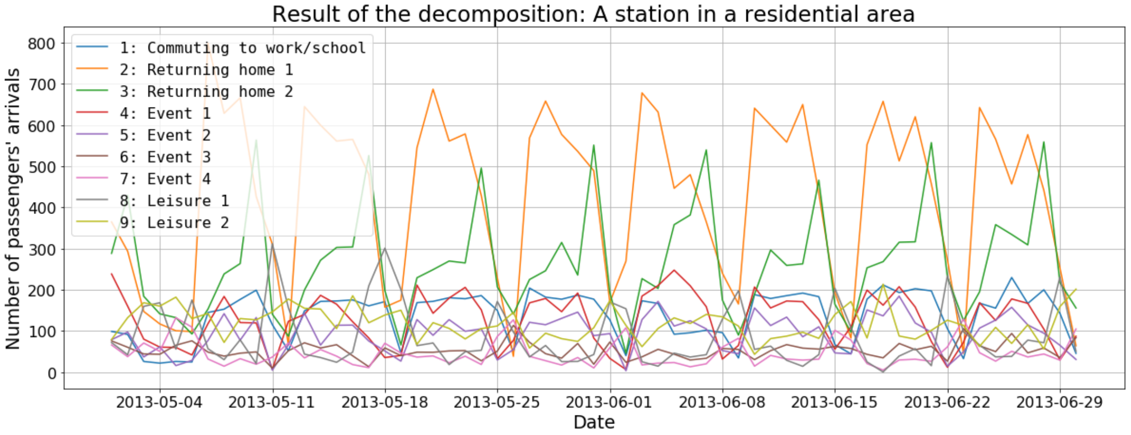


Fig. 5.9 The activity trend of a residential area.

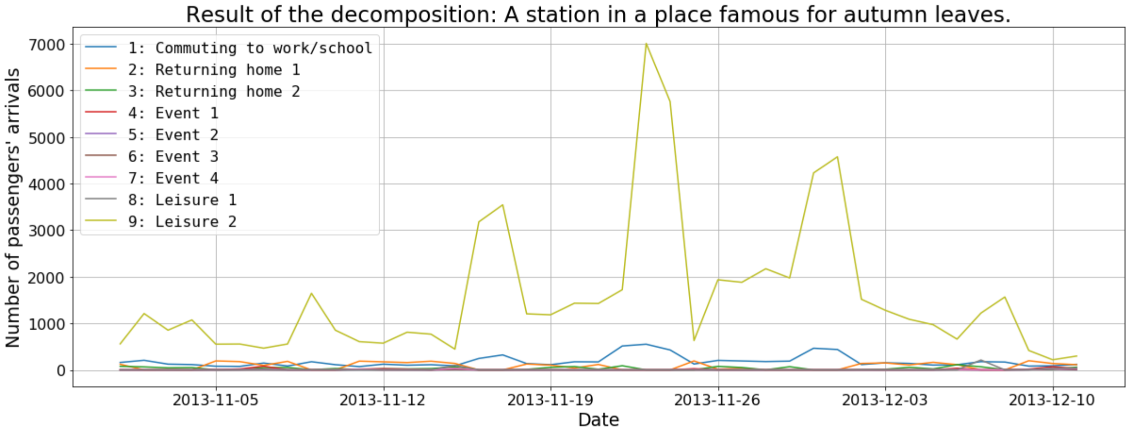


Fig. 5.10 The activity trend of an area famous for autumn leaves.

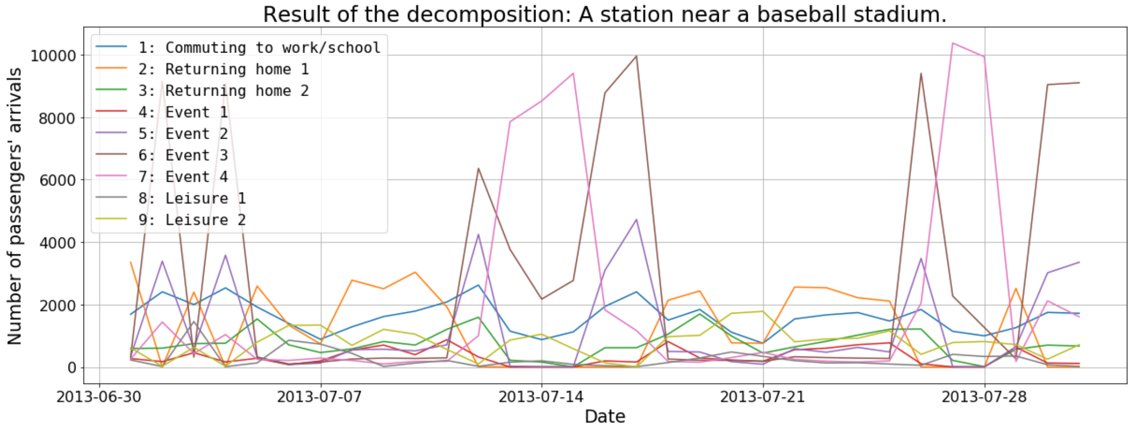


Fig. 5.11 The activity trend of an area with a baseball stadium.

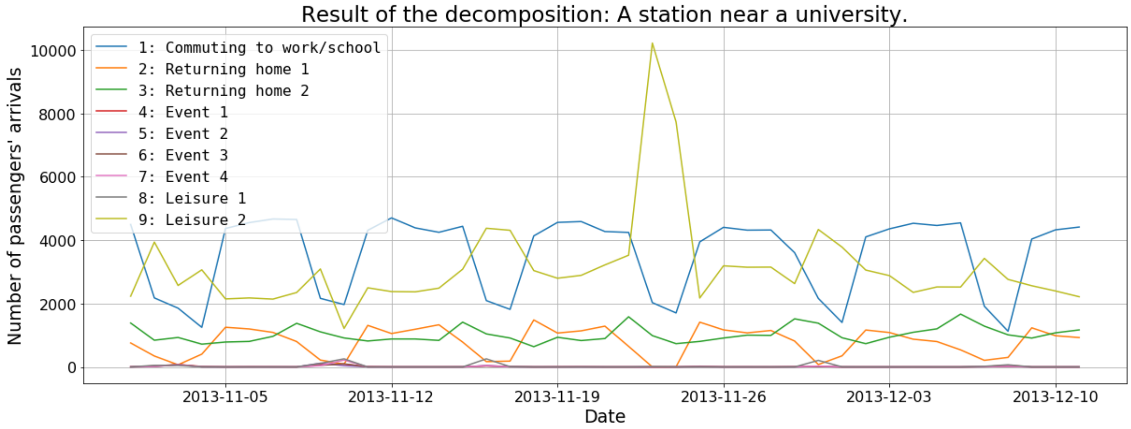


Fig. 5.12 The activity trend of an area where a university is located.

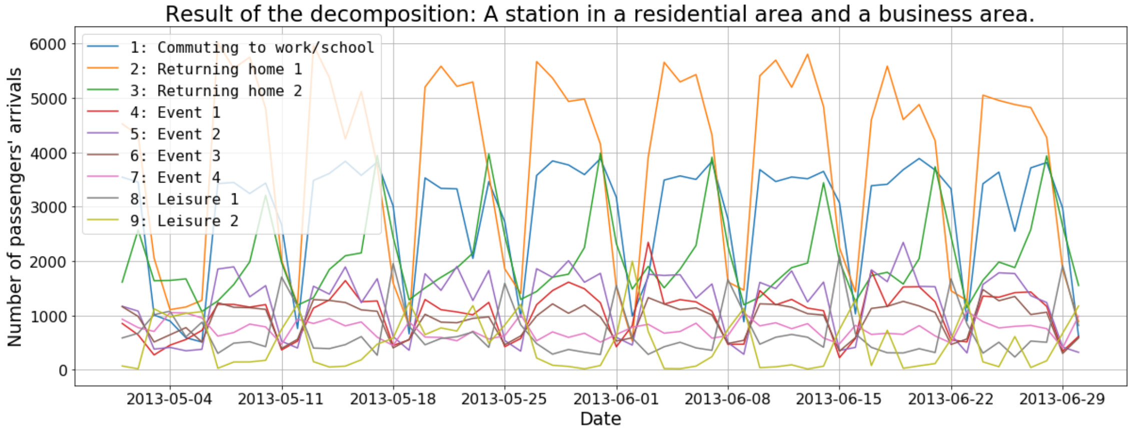


Fig. 5.13 The activity trend of an area of mixed land use for business and residence.

Table 5.2 Redefinition of trip purpose for evaluation.

Redefined purposes	Travel survey data	EAT-CD
Commuting to work/school	(4) Commuting to work (5) Commuting to school	Activity type 1 (Commuting to work/school)
Private purposes	(3) Private purposes	Activity type 4 (Event 1) Activity type 5 (Event 2) Activity type 6 (Event 3) Activity type 7 (Event 4) Activity type 8 (Leisure 1) Activity type 9 (Leisure 2)
Returning home	(1) Returning home	Activity type 2 (Returning home 1) Activity type 3 (Returning home 2)

Table 5.3 Comparison between the result of our method and the travel survey data.

Purpose	Day of the week	Correlation coefficient	p-value
Commuting to work/school	Weekday	0.878	5.32×10^{-163}
Commuting to work/school	Holiday	0.774	6.54×10^{-96}
Private purposes	Weekday	0.857	6.93×10^{-148}
Private purposes	Holiday	0.927	1.78×10^{-218}
Returning home	Weekday	0.812	3.83×10^{-121}
Returning home	Holiday	0.856	5.86×10^{-148}

5.5.3 Evaluation of the results of the decomposition

Performance of EAT-CD is evaluated by comparing the results from the NMF with travel survey data collected by the Ministry of Land, Infrastructure, Transport and Tourism in Japan. The survey is carried out every 10 years. Respondents are given questionnaires, and they record travel histories on the questionnaires. Respondents record the destination, origin, departure time, arrival time, trip purpose, and the means of transportation of each trip. The detailed data is not open to the public, but aggregated data collected in the Kansai Area is available online [109]. We use the aggregated data that contains information on trip purpose, exit stations, classification of weekday, weekend and holiday, and the number of passengers. The items of trip purposes in the survey data are (1) returning home, (2) business, (3) private purposes, (4) commuting to work, and (5) commuting to school. We group information about trip purposes into three types: commuting to work/school, private purposes, and returning home. We aggregate travel survey data and the results of EAT-CD by redefining trip purposes listed in Table 5.2. The evaluation process was carried out separately for weekdays and weekends/holidays, and by different trip purposes. The numbers of arriving passengers at all stations recorded in the survey data is compared against the estimates obtained through this study. Pearson's correlation coefficient is calculated for measuring the accuracy of EAT-CD. We note that there is no activity type in the results of EAT-CD that is equivalent to the trip purpose named "(2) business" in the survey data. Therefore, there is imbalance between the results of EAT-CD and the survey data: We use all the results from the smart card data, while we exclude the survey data whose trip purpose is "(2) business". The trip purpose means work outside the regular workplace of each person. EAT-CD is designed for extracting activity types that have particular temporal distributions of visitors' arrivals. We consider that human movements for the trip purpose named "(2) business" in the survey data do not have such characteristics. They could be included in activity type 8 or 9 (Leisure 1 or 2), but further study is necessary to confirm it. We consider that it does not have strong impact on our evaluation because the number of movements for the purpose is limited, and such human flows concentrate on specific stations.

There is no travel survey data collected after 2010. Therefore, we compare the result of EAT-CD from smart card data collected in 2014 with travel survey data collected in 2010. Since those two types of data are from different periods, we consider that land use of some areas may have changed between the two periods. These differences reduce the correlation coefficients between the two types of data, and the correlation coefficients would increase if we could compare data collected in the same period. Therefore, we assume that it is sufficient if the correlation coefficient between the two types of data from different periods is high enough.

Table 5.3 shows the result of evaluation. As shown in this table, correlation coefficients are higher than 0.77, and p-values are lower than $6.54\times\ 10^{-96}$. Therefore, we conclude that the correlation coefficients are sufficiently high with low p-values to confirm that the outcome of EAT-CD sufficiently reflects the survey data.

5.5.4 Detecting changes over time

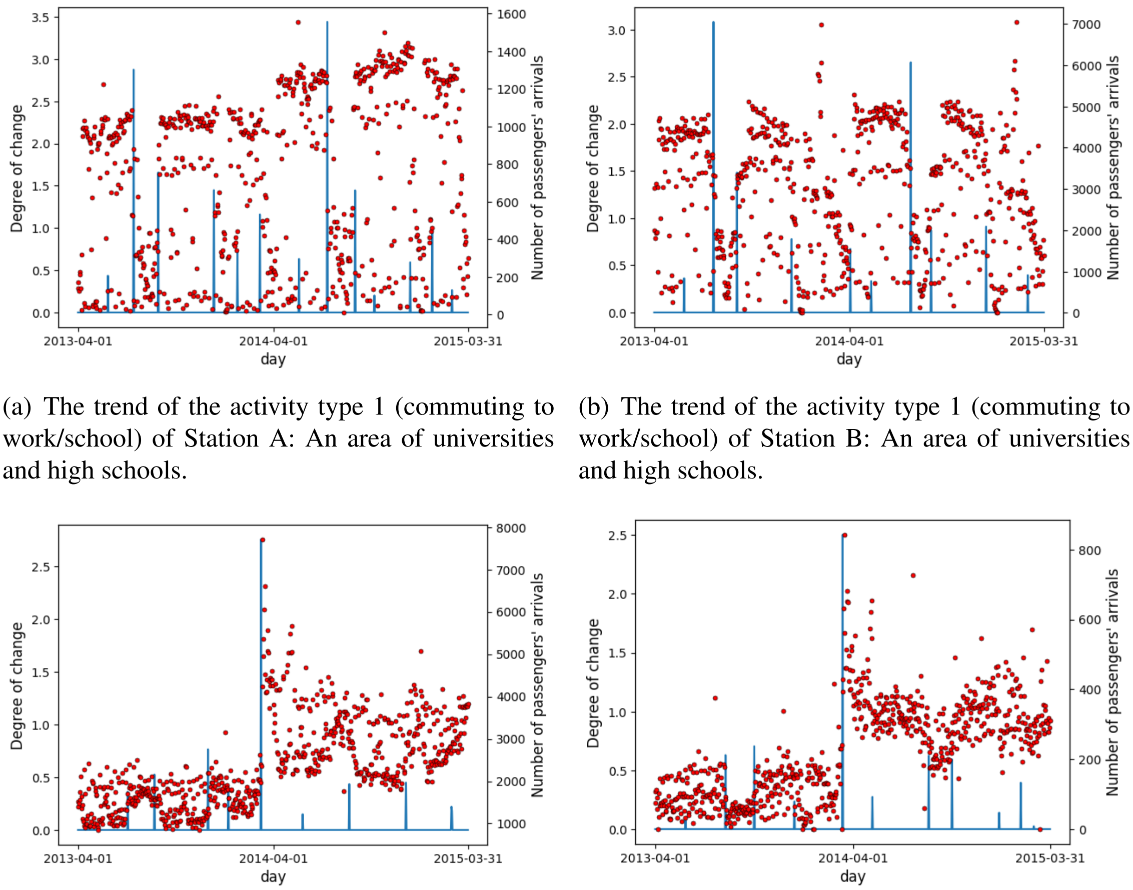
Table 5.4 shows the ranking of scores of change. Of the 15 highest scores, two of the nine activity types, namely Types 1 (commute to work/school) and 9 (leisure) are included. EAT-CD is designed for detecting changes in the activity trends, and it is not intended as a means to infer the reasons for such changes. We infer the reasons for the changes by the neighborhood characteristics for better understanding of the results. Thirteen changes in Activity Type 1 are caused by the beginnings or endings of academic semesters. Two changes in Activity Type 9 are caused by the development of a new shopping center and extension of an existing center, respectively.

Fig.s 5.14-(a)(b)(c) and (d) show the trends reflecting the 1st, 2nd, 7th and 12th changes respectively from Table 5.4 and their degrees of change. The red points illustrate the inferred numbers of passengers' arrivals for the Type 1 (Figs. 5.14-(a) and (b)) and Type 9 (Figs. 5.14-(c) and (d)). As shown in Figs. 5.14-(a) and (b), seasonal changes linked to academic semesters are detected, while non-seasonal changes are also detected (Figs. 5.14-(c) and (d)). Fig. 5.15 shows the graph of the trend of activity Type 9 (leisure) of Station F for a short period (1 month) near the detected change. The shopping center reopened on March 12, 2014. However, some shops started one day before the official reopening of the shopping center. EAT-CD captures the change on March 11, 2014.

Fig. 5.16 shows the trend of all trip activities in the area whose change is ranked 12th in Table 5.4. In this area, a large shopping center opened. The dashed red line indicates the opening date of the shopping center (March 16, 2014). The number of passengers commuting to work/school rapidly increases two days before the opening dates but sees a sharp decline after the opening dates. On the other hand, the number of passengers for leisure (Type 9) increases on the opening date and remains higher than it was before the opening date. It is considered that the increase of commuters is caused by the opening staff and temporary helpers for the launch of the shopping center, whereas the consistently high number of passengers for leisure purpose confirms that the shopping center continues to attract people after the opening date. However, without showing the break-down by trip purposes, the change in the total number of arriving passengers could mislead us to think that the shopping center attracted people on the opening date only.

Table 5.4 Ranking of scores of change. The top fifteen highest scores are dominated by Types 1 (commute to work/school) and 9 (leisure) activities only, which confirms the abrupt nature of the changes in the two types of trip activities.

	Score of change	**Activity type**	**Station ID**	**Characteristic of the area**	**Date**	**Reason for the change**
1	3.439	1	Station A	Universities and high schools	2014-07-13	End of semester
2	3.082	1	Station B	Universities and high schools	2013-07-23	End of semester
3	3.041	1	Station C	Universities and high schools	2014-07-18	End of semester
4	2.875	1	Station A	Universities and high schools	2013-07-16	End of semester
5	2.829	1	Station D	Universities and high schools	2014-07-22	End of semester
6	2.788	1	Station E	Universities and high schools	2013-07-14	End of semester
7	2.757	9	Station F	Shopping center	2014-03-11	Extension of a shopping center
8	2.737	1	Station E	Universities and high schools	2014-09-03	Beginning of semester
9	2.725	1	Station G	Universities and high schools	2014-09-03	Beginning of semester
10	2.717	1	Station H	Universities and high schools	2014-08-28	Beginning of semester
11	2.653	1	Station B	Universities and high schools	2014-07-27	End of semester
12	2.503	9	Station I	Shopping center	2014-03-15	Opening of a shopping center
13	2.367	1	Station C	Universities and high schools	2013-07-16	End of semester
14	2.342	1	Station D	Universities and high schools	2013-07-20	End of semester
15	2.197	1	Station E	Universities and high schools	2014-07-12	End of semester



(a) The trend of the activity type 1 (commuting to work/school) of Station A: An area of universities and high schools.

(b) The trend of the activity type 1 (commuting to work/school) of Station B: An area of universities and high schools.

(c) The trend of the activity type 9 (leisure) of Station F: An area of a newly extended shopping center.

(d) The trend of the activity type 9 (leisure) of Station I: An area of a newly opened shopping center.

Fig. 5.14 Detected changes in activity trends: The blue-colored graphs are the degrees of change. The red points are the inferred number of visitors' arrivals for commuting to work/school (a)(b) and leisure (c)(d). In areas of universities and high schools, changes are detected from the trends of the activity type 1 (commuting to work/school) according to the beginnings and endings of semesters. In areas of shopping centers, changes are detected from the trends of the activity type 9 (leisure) according to the opening and extension of the shopping centers.

5.5.5 The significance of the proposed method in the context of studies on inference of land use / activity

The proposed method (EAT-CD) succeeds in addressing the following points that the previous studies have not:

- The method extracts the temporal distributions of activity types.
- The method uses only hourly numbers of visitors' arrivals in each area.

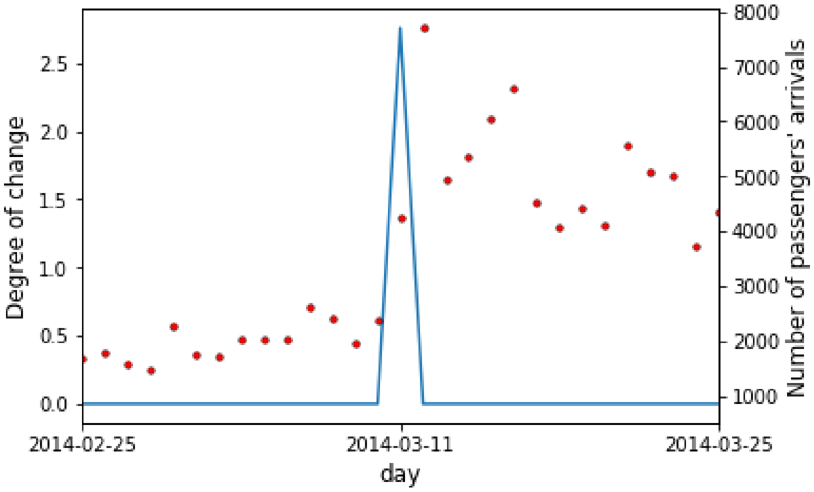


Fig. 5.15 The graph of the trend of the activity type 9 (leisure) of Station F in a short period (one month) near the detected change: An area of a newly extended shopping center.

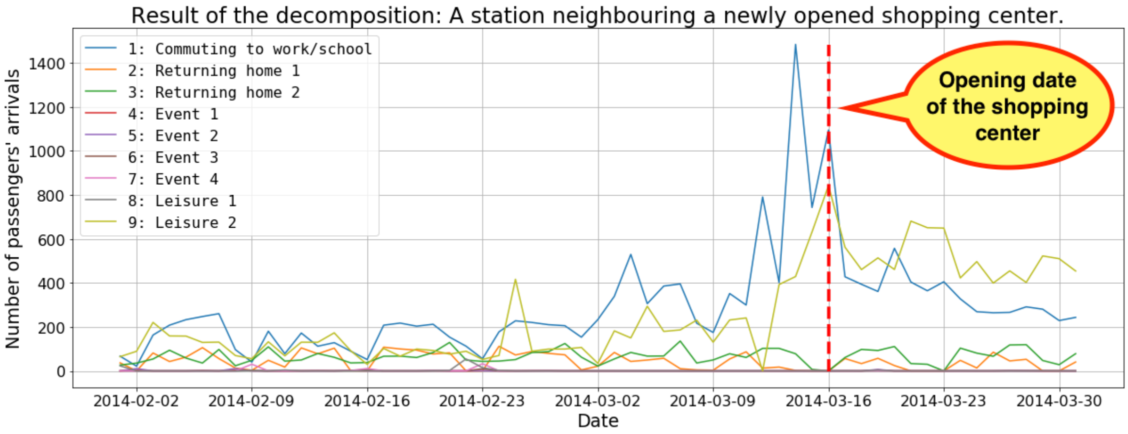


Fig. 5.16 The activity trends of the area of a newly opened shopping center.

- The method decomposes the number of visitors' arrivals in each area and each day into the activity types, and it constructs the trends of the activity types for each area.
- The method scores the degree of changes for each activity and each area, and it detects significant changes.
- It is possible to analyze the causes of changes by looking at the trends of the activity types.

With some adjustment, EAT-CD can be also applied to other types of locational data such as GPS data collected from mobile phones. Nishi et al. [87] and Fan et al. [37] analyze time-series distributions of people's visits using GPS data collected from mobile phones. To make time-series distributions of people's visits, they discretize the time and coordinates. They divide the study area of a city into grid cells and the duration of one day into time-slices. The number of human inflows to each tile based on human coordinates is used to create time-series distributions of visitors. By using the numbers, it is possible to perform EAT-CD for GPS data collected from mobile phones. Since EAT-CD can automatically detect urban changes by using mobility data such as smart card data of public transportation and GPS data collected from mobile phones, it will be possible to monitor urban changes globally. This is significantly advantageous for urban planners.

5.6 Conclusion

In this chapter, we proposed a method called EAT-CD to decompose the number of passengers arriving at each station into activity types, to construct activity trends for each area, and to detect changes in the land use patterns of urban areas. On **RQ1**, we confirmed that the numbers of visitors to each place can be estimated for different activity types by applying NMF to the hourly numbers of visitors recorded in the mobility data. On **RQ2**, we confirmed that a method can be developed for detecting changes in the activity trends using the estimated numbers of arriving passengers for each trip activity by using Jensen-Shannon divergence. EAT-CD only requires the total number of passengers by the hour. Validity of EAT-CD was examined by applying it to a set of smart card data from public transport in the Kansai Region, Japan. The results showed that the activity trends were successfully derived and significant changes in the patterns of travel or land use were well detected.

We note a limitation of our work. The number of activity types is determined manually in this study, and labeling of activity types is also conducted manually. Decomposing the number of visitors and detecting changes requires no prior knowledge of the region or the subjective setting of the trip types. However, labeling each activity type and finding the reasons for changes requires prior knowledge. It is necessary to modify the method to automatically execute the procedure.

EAT-CD can be applied to any kinds of data that have information about time-series distribution of visitors and would prove beneficial to planners who work on developing cities.

The ultimate goal of our research is to develop a method to capture urban changes that cannot be detected by other means. EAT-CD can detect changes that are known and established (e.g. the start and end of a semester, and opening of a shopping center), as well as those arising from previously unknown factors. EAT-CD has succeeded in capturing the effect caused by such big changes. We have shown that EAT-CD is capable of capturing changes in the number of visitors for leisure purposes to the station close to a newly opened shopping center and that this increase stays on where the number of passengers remain higher than it was before the opening (Fig. 5.16). The main advantage of EAT-CD is that it can capture changes that cannot be captured without decomposing the number of visitors into activity types. We expect that it is possible to develop a method to detect and understand the gradual changes of urban characteristics using EAT-CD.

Chapter 6

Measurement of Opportunity Cost of Travel Time for Predicting Future Residential Mobility

6.1 Introduction

Urban planners are making efforts to make cities livable for residents; They are also making efforts to predict future population change. Therefore, it is important to detect the shrinkage or growth of a population and to estimate the effects of urban development on future populations. However, it is difficult quickly following urban development to estimate the effects of urban development on future residential mobility.

Previous studies have shown that urban form influences both non-work travel behavior [25, 93, 131, 36] and residential mobility [92, 32]. The amenities of residential neighborhoods reduce the need for non-work travel to distant places and increase the number of in-migrants. For example, if shopping centers and parks are located in a neighborhood, residents do not have to go to shopping centers and parks located far away. Such convenience also contributes to an increase in residential movements to such a location.

We consider that the analysis of non-work travel contributes to the prediction of future residential mobility. Fig. 6.1 shows our hypothesis on the causal relationship between urban form, non-work human mobility, and residential mobility. Changes in urban form (e.g., the amenities of a neighborhood) could be a cause of changes in non-work human mobility (e.g., average travel time and diversity of destinations). Changes in non-work human mobility could be a cause of changes in residential mobility (e.g., the number of people who relocate to each place). Therefore, we consider it possible to create an index based on mobility

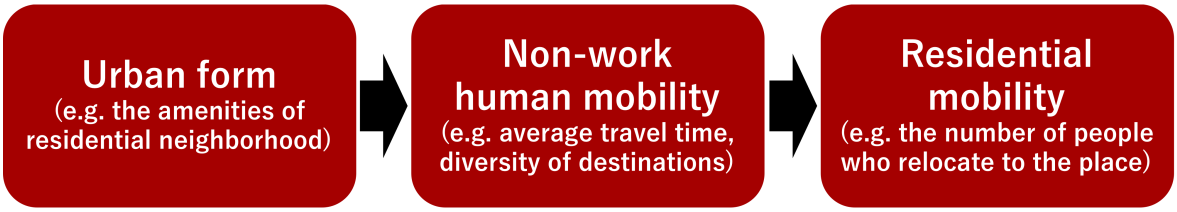

Urban form
(e.g. the amenities of residential neighborhood)
Non-work human mobility
(e.g. average travel time, diversity of destinations)
Residential mobility
(e.g. the number of people who relocate to the place)

Fig. 6.1 Hypothesis on the causal directions between urban form, non-work human mobility, and residential mobility.

data to estimate the influence of urban development on residential mobility. Since human mobility changes quickly according to urban changes, such as the opening of shopping centers, we consider that mobility data enable us to quickly predict the outcome of urban development. Therefore, the causal influence of changes in urban form on future residential movements can be quickly predicted, if it is possible to create such an index based on mobility data collected within a short period. Opportunities for analyzing massive real-time mobility data are increasing due to the automatic collection of mobility data such as smart card data of public transportation. Therefore, we consider that massive collected mobility data can contribute to the prediction of future residential mobility.

This study attempts to investigate a method for creating an index that not only correlates with future residential movements but also has causal influence on future residential movements. Creating an index based on human mobility data will make it possible to predict the influence of urban development on future residential movements.

We propose a method called *travel cost method for multiple places* (TCM4MP) by extending the conventional travel cost method (TCM). In previous studies, TCM was used for inferring the benefit of a recreational site [111, 77] and the opportunity cost of travel time [119, 19]. We consider that the opportunity cost of travel time on non-working days reflects the convenience and amenities of a neighborhood.

Opportunity cost is the loss of potential gain from other alternatives when one alternative is chosen. In this chapter, we focus on the opportunity cost of travel time on non-working days. When people choose to travel to a distant place from the location of their residences, they lose potential gain that could be obtained by engaging in activities in the residential neighborhood. Therefore, the opportunity cost of travel time is high if various types of amenities are accessible in the residential neighborhood. For example, the opening of a shopping center increases the opportunity cost of travel time, because people can have an enjoyable time at the shopping center. People prefer to spend time in the area where they live rather than to travel to distant places, if they can engage in enjoyable activities there.

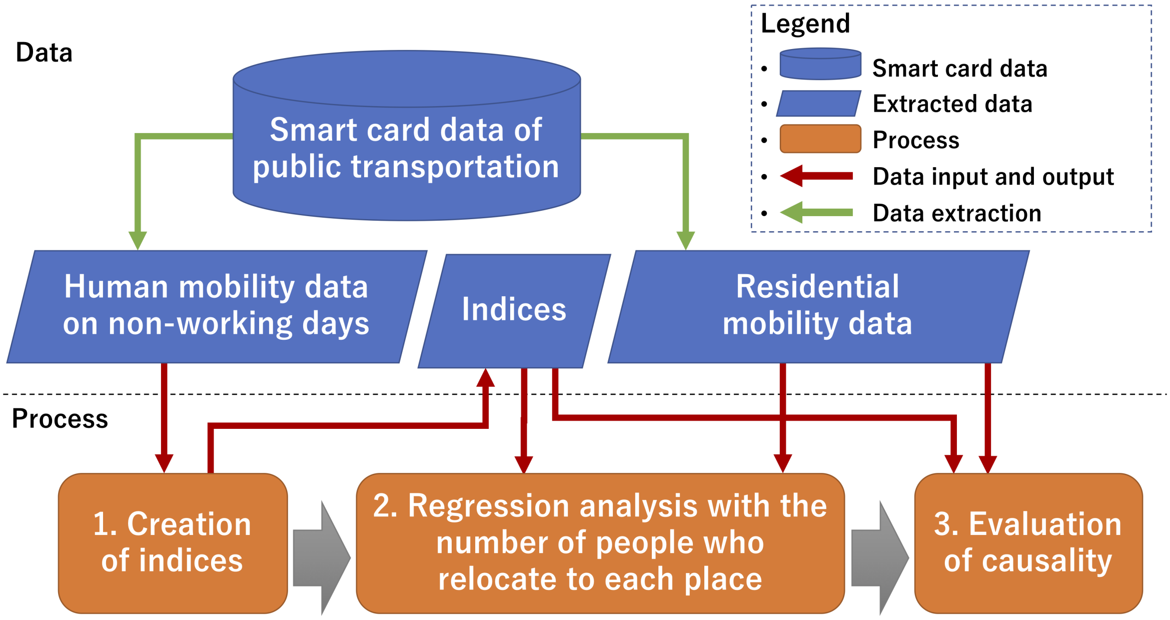

Data
Smart card data of public transportation
Legend
Smart card data
Extracted data
Process
Data input and output
Data extraction
Human mobility data on non-working days
Indices
Residential mobility data
Process
1. Creation of indices
2. Regression analysis with the number of people who relocate to each place
3. Evaluation of causality

Fig. 6.2 Flow-chart of the processes involved in this chapter.

TCM4MP is proposed to estimate the opportunity cost of travel time that varies according to the departure place. Conventional TCM does not assume that the opportunity cost of travel time varies according to the departure place. We consider such estimation as possible due to the use of massive mobility data. We assume that the opportunity cost of travel time on non-working days reflects the convenience and amenities of a neighborhood. Therefore, we consider that the opportunity cost of travel time has a causal influence on future residential movements.

The contributions of this chapter are summarized as follows:

- We propose a method to infer the opportunity cost of travel time on non-working days that varies according to the departure place by extending the conventional travel cost method.
- We examine the extent to which the opportunity cost of travel time contributes to the prediction of future residential movements.

We compare the contribution of opportunity cost to the prediction of the number of people who relocate to each place with other types of indices derived from smart card data of public transportation. It is insufficient to examine the contribution of the opportunity cost of travel time to the prediction of future residential movements only through the correlation to the number of people who relocate to each place. We infer a causal relation between the number of residential moves, the opportunity cost of travel time, and other indices that highly correlate with the number of residential moves. Thus, the research questions that we address in this chapter are as follows:

- **RQ1:** Does the opportunity cost of travel time calculated by an extended travel cost method contribute to the prediction of the number of people who relocate to each place compared to the current population and other indices?
- **RQ2:** Does the opportunity cost of travel time have a causal influence on the number of people who relocate to each place compared to the current population and other indices?

Fig. 6.2 is a flow chart of this study. We extract human mobility data on non-working days and residential mobility data from smart card data of public transportation. Indices for the prediction of residential movements are created. We create a regression model for each index using human mobility data on non-working days. Then, we perform the evaluation of causality between the number of people who relocate to each place and the created indices.

The remainder of this chapter is organized as follows: Section 6.2 reviews previous studies. Section 6.3 describes the proposed method. Section 6.4 explains the baselines and evaluation method. Section 6.5 explains the data that we use. Section 6.6 describes the data preprocessing method. Section 6.7 reports on the results. Section 6.8 discusses the implications and limitations of our study. Section 6.9 draws our conclusions.

6.2 Related studies

6.2.1 Travel cost method and opportunity cost of travel time

Travel is considered as a demand derived from the desire to engage in activities at destinations [57]. According to Becker's theory of the allocation of time [11], people allocate their limited time to activities to maximize utility. Since travel time is considered wasteful or unproductive [19], it is interpreted as a necessary evil to obtain benefits from activities at destinations. The travel cost method (TCM) is widely used for measuring the benefit of recreational sites [111, 77] and for measuring the opportunity cost of travel time [119, 19]. The method assumes that the ratio or the number of visits to a site and the cost of travel, including time and money, presents a demand curve. Regression analysis is performed for measuring the benefits of destinations and the opportunity cost of travel time. On the other hand, some studies [75, 74, 89] note that people spend time on productive or enjoyable activities while traveling using rapidly growing information and communications technologies (ICTs) such as smart phones, laptop computers, portable music players, and gaming devices. Therefore, the opportunity cost of travel time may differ based on such activities during traveling. In regard to methodologies developed in the behavioral sciences, discrete choice models are used to analyze individuals' destination choices of non-working trips, incorporating the opportunity cost of travel time into the utility models [13]. The methodologies are also useful for the analysis of choices of residential locations.

In our research, we measure the opportunity cost of travel time for multiple departure places using the smart card data of public transportation. Such measurement has never been performed in previous studies, and it requires an extension of the conventional TCM. In addition, we measure the opportunity cost of travel time for different demographic groups using gender and age data. The influence of the development of ICTs could be reflected in the difference in the opportunity cost of travel time among different demographic groups.

6.2.2 Interaction between urban form and travel behavior

In regard to the interaction between urban form and travel behavior, previous studies argue that urban form affects non-work travel behavior [25, 93, 131, 36]. According to these studies, the ratio of people who walk on non-working days is strongly related to land use diversity, intersection density, and the number of destinations within walking distance. On the other hand, while other studies [60, 49, 5, 15, 23] also acknowledge that urban form influences non-work travel behavior, they argue that attitudes toward travel behavior and preferences in residential location are more strongly associated with travel than are land use characteristics. These studies suggest that such preferences could affect residential choice. For example, people who prefer to drive may choose to live in suburban areas and travel by car, and people who prefer to walk may choose to live in areas with mixed land use and higher neighborhood accessibility. Mokhtarian and Cao [80] note the difficulty in finding causality between attitudinal factors, residential locations, and non-work travel behavior. They recommend the usage of longitudinal structural equations modeling for causal analysis. Krizek [62] performs a longitudinal analysis of the travel behavior of the same households to find causal relations between urban form and travel behavior. The findings of the study suggest that households change travel behavior when exposed to differing urban forms. In particular, locating to areas with higher neighborhood accessibility decreases vehicle miles traveled (VMT) and person miles traveled (PMT). In addition, Handy et al. [48] acknowledge that residential self-selection plays an important role in influencing individuals' travel decisions, but they also suggest that mixed land uses tend to discourage auto travel.

Regarding our research, public transportation was the dominant measure for moving in the area where data that we use were collected. We assume that the opportunity cost of travel time increases when amenities and convenience in the neighborhood are high. In addition, we assume that convenience and amenities in the neighborhood contribute to an increase in the number of people who relocate to the place. Our study does not assume that preferences for residential locations and travel behavior differ among people. Our study does not incorporate automobile travel, nor does it incorporate the difference in preferences for travel behavior and residential location. Therefore, it is necessary to further develop our method if applied to areas where people have varying preferences for travel behavior.

6.2.3 Influential factors on residential mobility

Influential factors on residential mobility and residential demand have long been studied [92, 32]. By applying hedonic modeling, the influence of crime [73], consumption amenities [63, 94, 91], neighborhood parks [56], social capital [58], walkability [41], and air quality [50] have been discussed. The relations between lifestyles, neighborhood characteristics, and the choice of residential location have been studied by modeling individual or household choices regarding residential location [46, 61, 14].

Jacobs [54] qualitatively argues that the ideal neighborhood is one that is walkable with many mixed uses and a diverse community. Walk Score [118], a publicly available website, has been introduced to quantitatively assess neighborhood walkability. Walk Score assigns neighborhood walkability scores based on distance to amenities and pedestrian friendliness (e.g., population density, road metrics) using data from Google, Education.com, Open Street Map, the U.S. Census, and Localeze. According to Gilderbloom et al. [41], Walk Score has positive impacts on housing values, crime rates, livability, and economic resilience, so it is an efficient tool to plan new urban forms to make a city livable and active.

We consider that the opportunity cost of travel time reflects consumption amenities, neighborhood parks, and walkability. However, crime rate and air quality are not reflected in mobility data. We consider that the benefit of our method is that it only requires mobility data for predicting residential mobility and that it does not require additional secondary data such as information about neighborhood amenities and road metrics.

6.2.4 Gap in the literature

In previous studies, data were collected through surveys. It is costly and time consuming to collect survey data. The availability of a large amount of mobility data collected from mobile phones and smart cards of public transportation has been rapidly increasing.

Recent studies create indices for estimating urban characteristics using automatically collected mobility data. For example, Smith et al. [104] create an index for estimating community well-being by calculating the diversity of destinations people visit based on the smart card data of public transportation. Zhong et al. [137] create indices of the centrality of mobility for each station using a network science approach from the smart card data of public transportation. Yabe et al. [129] measure the fragility of people flows to appropriately plan future investments in infrastructure based on locational data collected from mobile phones. The benefit of such indices is as follows: First, we can detect urban changes without surveys that cost time and money. Second, we can evaluate the result of investments in urban development and predict future outcomes. Third, we can obtain knowledge about the nature of urban systems using those indices.

To the best of our knowledge, no method has been proposed to create an index based on mobility data to predict future residential movements. We propose a method called TCM4MP (travel cost method for multiple places) to infer the opportunity cost of travel time that varies according to the given place as an index for predicting the number of people who relocate to each place.

Our study assumes that the opportunity cost of travel time on non-working days reflects the convenience and amenities of the neighborhood. The opportunity cost of travel time is the potential loss of benefit by traveling to distant places. Therefore, the opportunity cost of travel time is affected by the convenience and amenities of the neighborhood. However, we note that this assumption has a limitation. According to Walsh et al. [119], the opportunity cost of travel time on non-working days is also affected by other factors. For example, individuals with flexible work hours appear to have a lower opportunity cost of travel time than those whose work hours are fixed. Similarly, persons who are in school, retired, or unemployed have a lower opportunity cost of travel time. In addition, the opportunity cost of travel time is proportional to wage rate. If such factors differ widely according to places, the opportunity cost of travel time is affected by those factors. It is necessary to evaluate the relation between the opportunity cost of travel time and neighborhood amenities. We consider that such a comparison is possible using Walk Score.

Our study is motivated by previous studies in the domain of transportation engineering and urban economics; however, the aim of our study is practical use. The final objective of our study is to contribute to the detection and prediction of changes in the trends of residential mobility. We consider that the current population of each place also correlates with the number of people who relocate to each place. However, this changes slowly compared to changes in travel behavior; therefore, it does not contribute to the early detection and prediction of changes in the trends of residential mobility. We compare our proposed method with other indices derived from smart card data in terms of correlation and causality to the number of people who relocate to each place.

6.3 Method

6.3.1 Measurement of opportunity cost of travel time

Conventional studies using TCM do not assume that the opportunity cost of travel time varies according to the departure place. In addition, these studies usually calculate the benefit of one recreational site. Equation 6.1 is the equation of a demand curve defined in a conventional travel cost method.

$$v_i = C - Bt_i \quad (i \in D) \tag{6.1}$$

In this equation, v_i denotes the visitation rate to a recreational site per person who lives in place i, C denotes the benefit one can obtain by visiting the recreational site, t_i denotes the time it takes to travel from place i to the recreational site. B denotes the opportunity cost of travel time per unit time, and D denotes the set of departure places.

We propose a method called TCM4MP by extending Equation 6.1 to measure the values of the opportunity cost that differ according to the departure place. Equation 6.2 is derived by assuming that the opportunity cost of travel time differs according to the departure place and by incorporating the values of the benefits of multiple arrival places.

$$v_{i,j} = C_j - B_i t_{i,j} \quad (i \in D,\ j \in A) \tag{6.2}$$

In this equation, B_i is a parameter that represents the opportunity cost of people who live in departure place i. C_j is a parameter that represents the benefit that one can obtain by visiting arrival place j. $v_{i,j}$ denotes the logarithm of the average number of movements from departure place i to arrival place j per person, and A denotes the set of arrival places.

It is impossible to determine the values of B_i by simply applying linear regression analysis. We assume that the parameters B_i $(i \in D)$ and C_j $(j \in A)$ follow normal distributions as in Equations 6.3 and 6.4.

$$B_i \sim Normal(\mu_B, \sigma_B^2) \quad (i \in D) \tag{6.3}$$

$$C_j \sim Normal(\mu_C, \sigma_C^2) \quad (j \in A) \tag{6.4}$$

In these equations, μ_B, σ_B, μ_C, and σ_C are hyperparameters. All the parameters and hyperparameters are inferred by applying the hierarchical Bayesian inference method to Equations 6.2, 6.3, and 6.4 using mobility data on non-working days. In this study, we use statistical software, Stan [24], for Bayesian inference, which is based on No-U-Turn sampler (NUTS) [52], an extension of Hamiltonian Monte Carlo (HMC) [82]. Thus, the values of the opportunity cost of travel time B_i are obtained.

The opportunity cost B_i can be obtained separately according to people's demographic information. Our method divides mobility data according to demographic information about their gender and age, and we calculate B_i for each demographic group. We let k denote a demographic group, and $B_{i,k}$ denotes the opportunity cost of people living in place i in demographic group k.

6.3.2 Decision of prediction model

We assume that the opportunity cost of travel time reflects the convenience of the place. We also consider that the number of people who relocate to a place correlates with the convenience of the place. In addition, we consider that the convenience of a place varies depending on gender and age. Therefore, we use the opportunity cost of the travel time to a place for the prediction of the number of people who relocate to that place.

We create a power regression model φ_{O} for **opportunity cost (O)** to predict the amount of future residential mobility by Equation 6.5.

$$m_{i,k} = \varphi_{\mathrm{O}}(k,i) = \beta_k B_{i,k}^{\alpha_k} \tag{6.5}$$

Note that $m_{i,k}$ denotes the number of people in demographic group k who relocate to place i. The parameters α_k and β_k are unconditioned to place i. We consider that power regression is suitable because of the assumption that the number of residential movements is proportional to the product of the population of two areas (a gravity model [139]) and the assumption that the distribution of a population follows the power law (Zipf's law [40]). The parameters are calculated using the Levenberg-Marquardt algorithm [81].

6.4 Baselines and evaluation methods

6.4.1 Baselines

We evaluate how much the predictor using the opportunity cost (φ_{O}) contributes to the prediction of residential mobility by comparing it with other indices. We compare it with predictors using the following three indices:

- **Population (P)**: This index is simply the number of current residents in place i. We let $P_{i,k}$ denote the number of residents of demographic group k living in place i.

- **Average travel time (A)**: This index is a simplification of the opportunity cost of travel time. We let $T_{i,k}$ denote the multiplicative inverse of travel time per person.
- **Entropy (E)**: According to Smith et al. [104], the diversity of places people visit reflects the well-being of the community. The index is as follows:

$$diversity(u) = \frac{-\sum_{j \in S_u} w_{u,j} \log(w_{u,j})}{|S_u|} \tag{6.6}$$

$$H_i = \frac{1}{|M_i|} \sum_{u \in M_i} diversity(u) \tag{6.7}$$

In Equation 6.6, S_u is the set of places user u visited, and $w_{u,j}$ is the proportion of all u's visits to place j. The numerator in this equation is the Shannon entropy. In Equation 6.7, M_i is the set of users who live in place i. We let $H_{i,k}$ denote the entropy of places visited by users in demographic group k living in place i.

As with Equation 6.5, prediction models φ_{P}, φ_{A}, and φ_{E} are determined by power regression.

6.4.2 Evaluation of correlation and causal relation between the number of relocations and the predictors

First, we evaluate the correlation between a predictor and the number of people who relocate to each place. If we find predictors that highly correlate with the number of people who relocate to each place, we perform causal inference between the predictors and the number of residential moves.

We apply linear non-Gaussian acyclic model (LiNGAM) [102] for the causal inference. By applying LiNGAM, causal relations between variables are obtained as described in Equation 6.8.

$$x_i = \sum b_{i,j} x_j + e_i \tag{6.8}$$

In this equation, x_i is an observed variable. The variable e_i is an exogenous variable (random variable) having a non-Gaussian distribution, and $b_{i,j}$ is the strength of the causal connection from x_j to x_i. The objective of causal inference is to determine $b_{i,j}$.

The key difference between LiNGAM and earlier works on causal inference is that LiNGAM assumes that exogenous variables are non-Gaussian. Under this assumption, it is possible to estimate a causal ordering of variables using passive observational data alone without any prior information on a causal ordering of the variables.

In our study, x_i denotes either the number of residential moves or the value of a predictor for predicting the number of residential moves. By applying LiNGAM, we obtain the causal networks between the number of residential moves and predictors that highly correlate with the number of residential moves. We use causal networks obtained using LiNGAM for the evaluation of the causal influence of each predictor on the number of residential moves.

6.5 Data

We use the smart card data of public transportation collected from March to April 2016 and from March to April 2017. There are two types of smart cards in this area. One type of smart card requires applicants to submit their personal information when applying to obtain the card. We use data collected from that type of smart card. The data include information about the user's ID, user's age, user's gender, the postal code of the location of user's residence, boarding station, boarding time, alighting station, and alighting time. The data include 1,024 stations held by 14 railway companies and 3 agencies of city governments.

The demographic composition of the card holders is given in Table 7.3. The number of male card holders and that of female card holders exceeded one million in both 2016 and 2017.

We compare the demographic composition of the smart card data and that of the data published by the government of Osaka City. We categorize the smart card data by gender and ages in groups of 10 years. Osaka City has 24 wards. We therefore further divide the smart card data according to the ward in which each card holder's residence is located. We compare the number of residents by ward for each demographic group. Table 6.2 shows the correlation coefficients between the number of each demographic group of the smart card data and that of the governmental data. We assume that smart card data reflect real demographic compositions.

In this chapter, we use smart card data collected from March to May 2016 for obtaining predictors for the amount of future residential mobility. We recognize residential moves between 2016 and 2017 by the changes in stations at which card holders most frequently board for the first time in a day.

We do not use data on individuals younger than 20 or older than 79 because we assume that most of these people do not relocate to other areas by their own will. In the following, we divide card holders into demographic groups by gender and in age groups of 10 years.

Table 6.1 Demographic composition of card holders

	2016		**2017**	
Age	**Male**	**Female**	**Male**	**Female**
0-9	11,331	12,274	12,252	13,308
10-19	127,968	141,913	133,691	147,673
20-29	146,365	215,936	148,487	217,400
30-39	198,082	251,226	193,508	248,822
40-49	275,129	310,716	279,966	322,486
50-59	219,646	225,320	230,026	242,502
60-69	143,823	151,027	157,696	169,407
70-79	133,427	182,992	130,555	180,812
80-89	40,429	53,125	43,489	58,318
90-99	1,557	1,840	1,786	2,083
Total	1,297,757	1,546,369	1,331,456	1,602,811

6.6 Data pre-processing

This section explains the data pre-processing procedure for the smart card data that we used. Some thresholds used in this section are arbitrarily set, and we do not investigate a method to automatically find suitable such thresholds in this study. Further research is needed to investigate a method to automatically find suitable thresholds.

6.6.1 Decision of travel time between two stations

We determine the time it takes to travel from departure station d_1 to arrival station a_1 by $t(d_1, a_1)$ defined in Equation 6.9.

$$t(d_1, a_1) = \mathrm{Med}(H(d_1, a_1)) \tag{6.9}$$

In this equation, $H(d_1, a_1)$ denotes a set of all the values of durations for traveling from departure station d_1 to arrival station a_1. $\mathrm{Med}(H(d_1, a_1))$ denotes the median of $H(d_1, a_1)$.

6.6.2 Grouping departure stations, arrival stations, and smart card holders

We group departure stations by the physical proximity between stations and similarity of boardings by smart card holders. Departure station d_1 and departure station d_2 are grouped

Table 6.2 Correlation coefficients of the comparison of 24 wards in Osaka City between the population of smart card holders and the information about the demographic composition published by the government of Osaka City.

Age	2016	2017
0-9	0.787	0.823
10-19	0.937	0.948
20-29	0.959	0.969
30-39	0.936	0.946
40-49	0.925	0.943
50-59	0.927	0.942
60-69	0.898	0.927
70-79	0.951	0.950
80-89	0.948	0.941
90-99	0.883	0.904

together when Equation 6.10 or Equation 6.11 is satisfied.

$$l(d_1, d_2) < 300 \tag{6.10}$$

$$\Big(|V(d_1) \cap V(d_2)| > 0.1 \cdot |V(d_1)|\Big) \wedge \Big(|V(d_1) \cap V(d_2)| > 0.1 \cdot |V(d_2)|\Big) \tag{6.11}$$

In Equation 6.10, $l(d_1, d_2)$ denotes the distance between departure station d_1 and departure station d_2 in meters. In Equation 6.11, $V(d_1)$ denotes a set of smart card holders who had boarded at departure station d_1 for the first time on a day during the period from March to May 2016 (we assume that a station where a user boards for the first time in a day is located near the user's residential location). Equation 6.10 means that the distance between departure station d_1 and departure station d_2 is less than 300 meters. Equation 6.11 means that the number of smart card users who have boarded at both departure station d_1 and departure station d_2 is greater than 10% of the number of smart card holders who had boarded at departure station d_1 and that of smart card users who had boarded at departure station d_2.

We group arrival stations by the physical proximity and walkability between stations. Arrival station a_1 and arrival station a_2 are grouped together when Equation 6.12 or Equation 6.13 is satisfied.

$$l(a_1, a_2) < 300 \tag{6.12}$$

$$\min W(a_1, a_2) < 30 \tag{6.13}$$

In Equation 6.13, $W(a_1,a_2)$ denotes the set of all the values of the duration between a user alighting at station a_1 and the time they subsequently board at station a_2. Equation 6.13 means that the minimum period between when users alight at station a_1 and when the users subsequently board at station a_2 is less than 30 minutes. We assume that this means that people can walk from station a_1 to station a_2 within 30 minutes. This is based on the fare system employed by many Japanese railway companies whereby additional fare is not required for a transit of less than 30 minutes.

Finally, 599 groups of departure stations and 541 groups of arrival stations are obtained. In the following, the group of departure stations is simply referred to as the departure place, and the group of arrival stations is simply referred to as the arrival place.

We determine the time it takes to travel from departure place i to arrival place j by Equation 6.14.

$$T(i,j) = \min_{d \in F_{\mathrm{D}}(i),\ a \in F_{\mathrm{A}}(j)} t(d,a) \tag{6.14}$$

In this equation, $F_{\mathrm{D}}(i)$ denotes the set of all stations in departure place i, and $F_{\mathrm{A}}(j)$ denotes the set of all stations in arrival place j.

We also group smart card holders. Smart card holders are grouped by the stations at which they most frequently board for the first time in a day.

$$\operatorname*{argmax}_{k}\ n(u,k) = i \ \Leftrightarrow\ u \in U(i) \tag{6.15}$$

In this equation, $n(u,k)$ denotes the number of days user u boards at place k. This equation means that user u is grouped into set $U(i)$ when the number of days user u boards at departure place i is the highest among all the departure places.

In addition, we exclude the data of user $u \in U(i)$ if place i is far from the location of his/her residence estimated by his/her postal code. In this study, we set the threshold at 5,000 meters.

6.6.3 Removing records of arrival stations each user frequently visits

We apply the travel cost method to the data of non-work travel. Therefore, we apply the method to data collected on holidays and weekends. However, some people commute to work on holidays or weekends. To remove such records, we count the number of days each user has alighted at each station. We ignore the records of the stations at which a user frequently alighted. The records of arrival place j of user u that satisfy Equation 6.16 are ignored.

$$\frac{g(u,j)}{s} \geq \frac{2}{7} \tag{6.16}$$

Table 6.3 Valid card holders

Age	Male	Female
20-29	115,926	176,803
30-39	153,922	206,359
40-49	217,804	266,191
50-59	171,413	192,379
60-69	117,933	130,967
70-79	114,981	160,849

In this equation, $g(u, j)$ denotes the number of days (including weekdays) user u has alighted at arrival station j. s denotes the number of all the days in the period of the data.

Table 6.3 shows the number of valid users obtained after data preprocessing.

6.7 Results

6.7.1 Results of regression of the predictors to the number of residential moves.

Table 6.4 shows the results of the regression analyses between the number of residential movements $m_{i,k}$ and predictors φ_{O}, φ_{P}, φ_{A}, and φ_{E}. The predictor using opportunity cost (O) is superior to other baselines for the data of male individuals of all ages and female individuals less than 60 years old. On the other hand, the predictor using population (P) scores the highest for the data of female individuals who are 60 or older.

The coefficients of determination of the predictors using opportunity cost (O) and population (P) are higher than 0.6, but the other predictors are lower than 0.6.

6.7.2 Causal influences on the number of residential moves

We measure causal influences from the two predictors using opportunity cost (O) and population (P) on the number of residential moves using linear non-Gaussian acyclic model (LiNGAM) [102]. LiNGAM assumes that exogenous variables are non-Gaussian. We test whether the predictors and the number of residential moves satisfy this assumption. It is impossible to test the distributions of exogenous variables. Therefore, we test the distributions of observed variables instead.

The Shapiro-Wilk test [101] is used to test normality. The Shapiro-Wilk test tests the null hypothesis that a sample came from a normal distribution. The null hypothesis of this test is that the population of the data is normally distributed. If the p-value is less than the chosen alpha level, then the null hypothesis is rejected, and there is evidence that the data tested are not from a normally distributed population. Table 6.5 shows the results of the Shapiro-Wilk test. Every p-value listed in Table 6.5 is small enough to conclude that the predictors O and P and the number of residential moves to each place (M) do not follow a normal distribution.

In addition to non-Gaussianity, LiNGAM assumes that relations between variables are linear. The linearity between the number of residential moves ($m_{i,k}$) and the predictor O (φ_{O}) and the linearity between the number of residential moves ($m_{i,k}$) and the predictor P (φ_{P}) have already been confirmed, as shown in Table 6.4. In regard to the relation between the two predictors (φ_{O} and φ_{P}), the correlation coefficient is shown in Table 6.6. It is high enough to assume that the relation between the two predictors is linear. Therefore, we consider that the values satisfy the assumption of LiNGAM.

Fig. 6.3 shows the results of the application of LiNGAM. The directions of the obtained causal networks are the same for every demographic group. The number of people who relocate to each place (M) is affected by both the predictor using the opportunity cost of travel time (O) and the predictor using the current population (P). The strength of causality from O to M is greater from P to M except for the demographic group of female individuals between 60 and 69 years of age. In addition, the strength of causality from P to O is greater from P to M except for the demographic group of female individuals between 60 and 69 years of age.

We measure the reproducibility of the results. First, we examine the reproducibility of the directions of causal networks. We repeat the experiment by making 1000 causal networks for each demographic group by randomly taking 400 departure places out of 599 departure places and making causal networks using the data of 400 selected departure places. The results are shown in Table 6.7. The first type of network has the same causal directions as the network that we obtain using all the departure places. The number of individuals in the first type of network is 1000 for the demographic groups of male individuals less than 70 years of age and female individuals less than 60 years of age. On the other hand, the number of individuals in the first type of network for the groups of male individuals in their 70s is 772. Therefore, there might be missing causal factors for the demographic group; thus, it is necessary to seek other causal factors to incorporate.

Next, we examine the reproducibility of the order of the strength between the causalities to the number of residential moves from the opportunity cost of travel time and current population. We conduct the same random sampling until we obtain 1000 networks whose directions of causality are the same as the network that we obtain using all the departure places. Fig. 6.4 shows the ranges of the strengths of causality. Comparing the medians of the strengths of causality to the number of residential moves, the orders are consistent with the results obtained using all the departure places, as shown in Fig. 6.3. The ranges of the causalities (O→M and P→M) overlap only in regard to the demographic group of female individuals 60 years or older. Therefore, we conclude it is certain that the causal strength of O→M is greater than P→M, and the causal strength of P→O is greater than that of P→M, except for the demographic group of female individuals 60 years or older. We cannot draw a conclusion for the demographic group of female individuals 60 years or older.

Table 6.4 Coefficients of determination (R^2) between the number of people who relocate to each place ($m_{i,k}$) and predicted numbers (φ_O, φ_P, φ_A, φ_E).

	Male						Female					
Predictor	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**
O (φ_O)	**0.677**	**0.682**	**0.703**	**0.687**	**0.695**	**0.677**	**0.641**	**0.686**	**0.686**	**0.670**	0.647	0.622
P (φ_P)	0.637	0.626	0.639	0.627	0.633	0.662	0.632	0.655	0.683	0.663	**0.681**	**0.693**
A (φ_A)	0.476	0.462	0.501	0.485	0.420	0.397	0.517	0.459	0.506	0.460	0.382	0.391
E (φ_E)	0.345	0.337	0.398	0.411	0.397	0.230	0.325	0.344	0.389	0.373	0.345	0.227

Table 6.5 The result of the Shapiro-Wilk test.

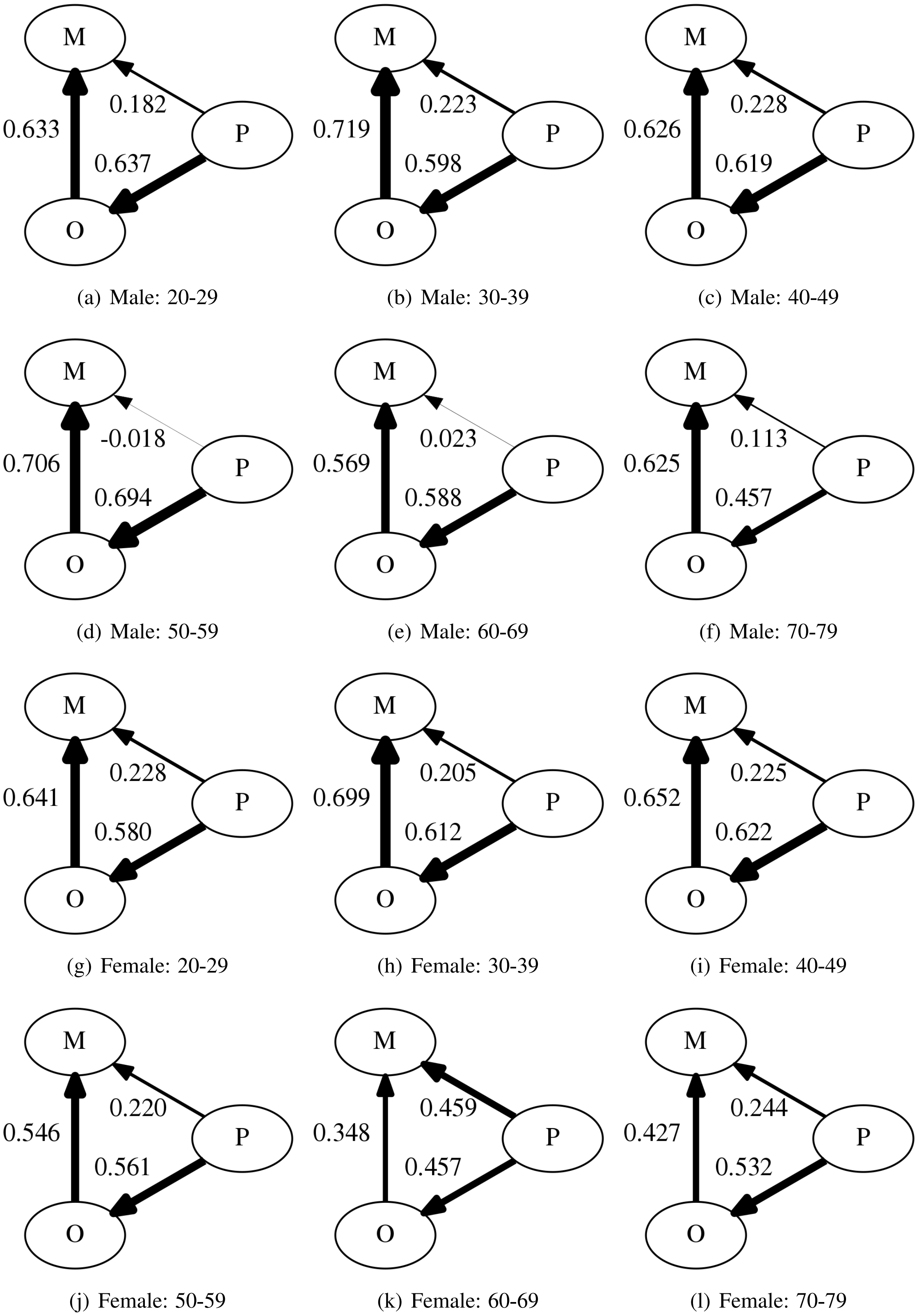
	Male						Female					
Valuable	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**
O (φ_O)	1.18×10^{-25}	9.34×10^{-28}	5.78×10^{-25}	9.99×10^{-23}	1.59×10^{-24}	9.59×10^{-32}	1.04×10^{-25}	1.45×10^{-27}	9.53×10^{-24}	5.21×10^{-24}	9.02×10^{-25}	2.44×10^{-30}
P (φ_P)	7.73×10^{-27}	5.09×10^{-28}	1.13×10^{-26}	6.38×10^{-26}	9.00×10^{-27}	1.46×10^{-31}	5.26×10^{-28}	1.93×10^{-28}	9.89×10^{-28}	5.59×10^{-28}	3.06×10^{-29}	4.13×10^{-32}
M ($m_{i,k}$)	2.02×10^{-35}	2.11×10^{-36}	5.82×10^{-34}	1.75×10^{-33}	9.16×10^{-34}	3.88×10^{-37}	1.57×10^{-36}	4.94×10^{-36}	4.03×10^{-34}	1.18×10^{-34}	4.11×10^{-35}	2.78×10^{-37}

Table 6.6 Correlation coefficient (R) between the predictor using opportunity cost (φ_O) and the predictor using current population (φ_P).

Male						Female					
20-29	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**
0.828	0.794	0.783	0.798	0.805	0.850	0.823	0.835	0.817	0.807	0.814	0.847

Table 6.7 The reproducibility of causal networks obtained using LiNGAM: We repeat the experiment by making 1000 causal networks for each demographic group by randomly taking 400 departure places out of 599 departure places and making causal networks using the data of 400 departure places.

	Male						**Female**					
Direction of causality	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**
1: O to M, P to O, P to M	1000	1000	1000	1000	1000	772	1000	1000	1000	1000	923	975
2: O to M, O to P, P to M	0	0	0	0	0	228	0	0	0	0	77	19
3: O to P, M to O, M to P	0	0	0	0	0	0	0	0	0	0	0	6



(a) Male: 20-29 (b) Male: 30-39 (c) Male: 40-49

(d) Male: 50-59 (e) Male: 60-69 (f) Male: 70-79

(g) Female: 20-29 (h) Female: 30-39 (i) Female: 40-49

(j) Female: 50-59 (k) Female: 60-69 (l) Female: 70-79

Fig. 6.3 Causal networks between the predictor using the opportunity cost of travel time (O), the predictor using population (P), and the number of residential moves to each place (M).

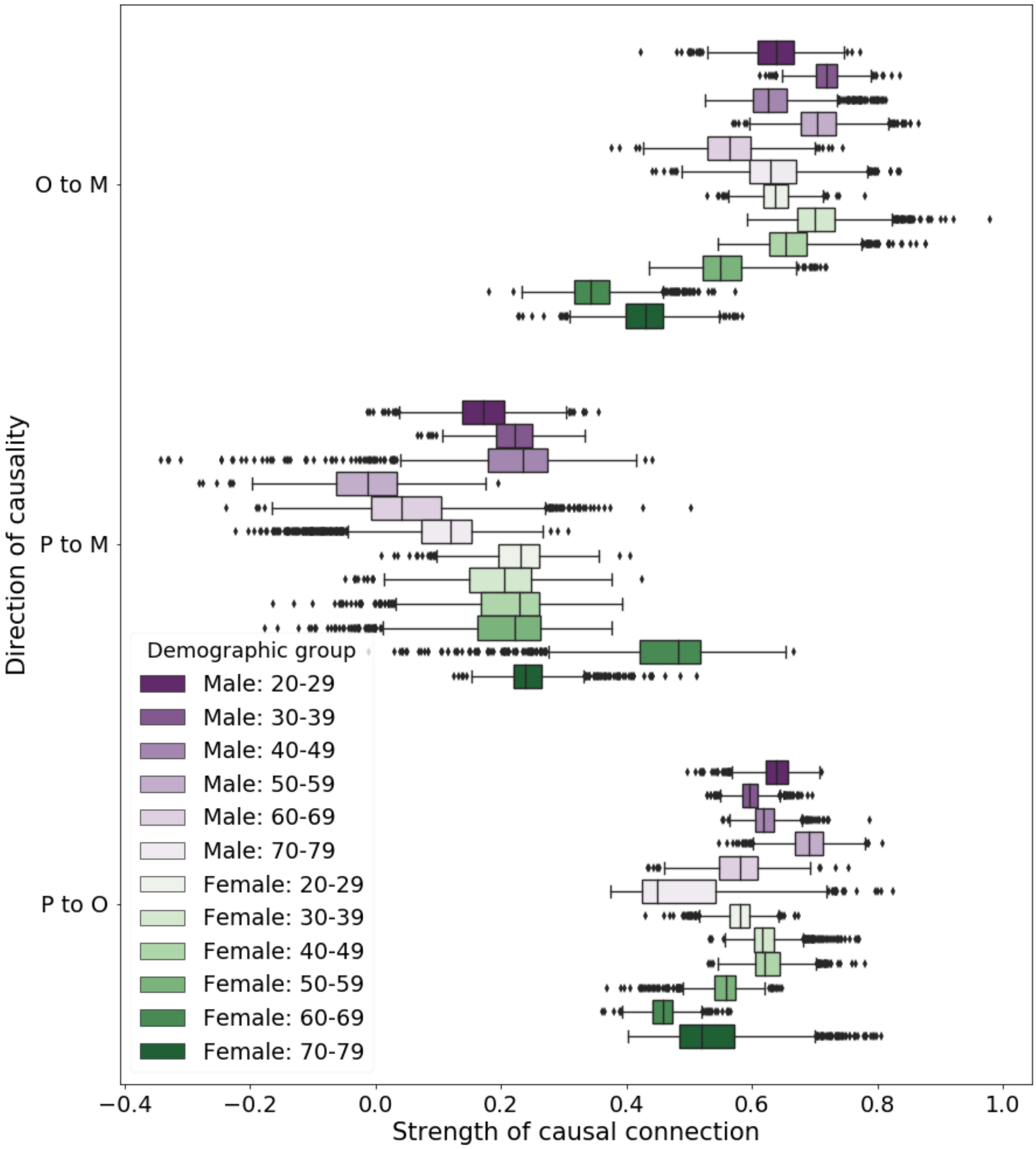


Fig. 6.4 The reproducibility of the strength of causalities.

6.8 Discussion

We have observed that the opportunity cost of travel time has strong causality to the number of people who relocate to each place. In addition, the index is influenced by the current population in each place. We assume that this is because a more heavily populated area will be more convenient because commercial developers and residential developers focus on such areas. This leads to the conclusion that most of the influence from the current population on residential mobility is mediated indirectly through the opportunity cost of travel time. The opportunity cost of travel time is calculated from non-work travel behavior in a short period of time (3 months), and it directly reflects the effects of urban development such as the construction of parks and the development of commercial facilities. Therefore, the opportunity cost of travel time is an efficient and effective index for predicting future residential mobility.

We can list many other factors that are not investigated in this study. Fig. 6.5 includes some factors related to residential mobility that were not investigated in this study. The factors in green are not investigated in this study.

Human mobility on non-working days is influenced by urban form such as amenities of the neighborhood and land use diversity. Urban form is influenced by investment from both industries and governments. Industries tend to invest in areas with high population densities. Therefore, investment from industries is influenced by the current population. Governments can change the trend of residential movements by investing in urban development.

There is also causality from the current population to the number of people who relocate to each place that is not mediated through non-work human mobility. Residential movement is influenced by land prices and the social capital of the local community. These factors are influenced by the current population. We assume that these causalities could be why the causality from O to M is not always greater than that from P to M in regard to the demographic group of female individuals 60 years or older.

There are also factors influencing residential mobility not being mediated by human mobility on non-working days and the current population. For example, closeness to business areas and air quality could have strong causality to residential mobility.

We note that further investigation needs to test the settings of thresholds and their influence on the results. In this study, we arbitrarily set thresholds for (1) grouping stations and smart card holders, (2) extracting non-working trips, and (3) selecting predictors for a causal analysis. In regard to grouping stations and card holders, the results may change according to the thresholds. In regard to extracting non-working trips, many business areas also have many commercial facilities in the neighborhood. Therefore, it is difficult to distinguish non-working trips using smart card data of public transportation. The results of our

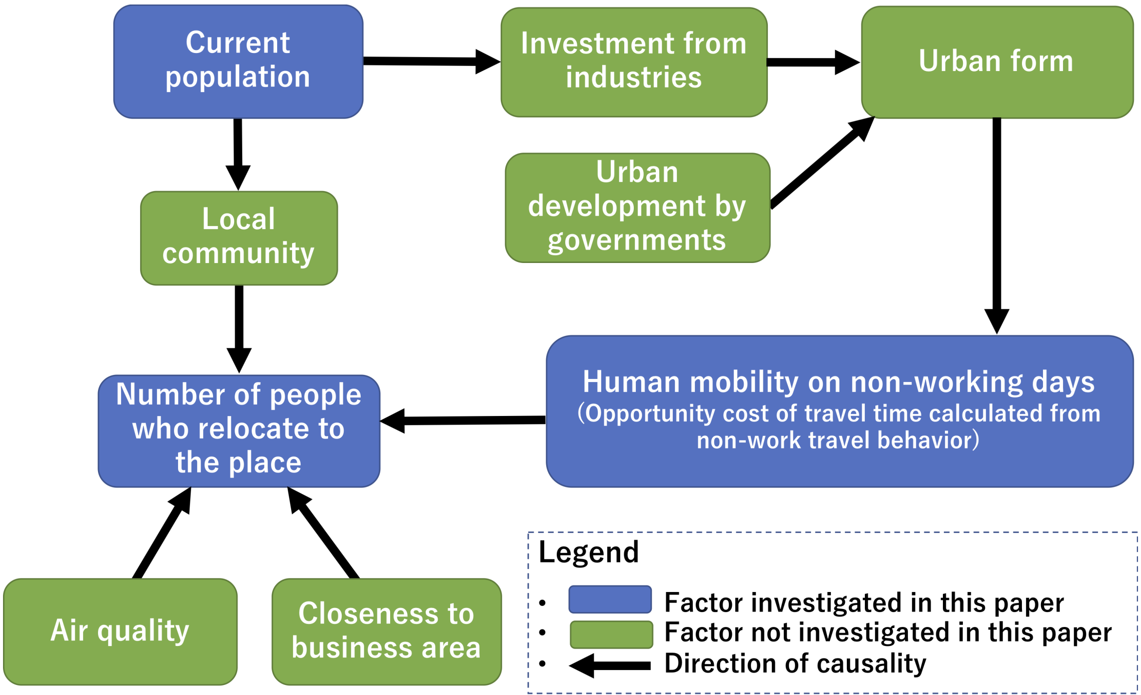

Current population
Investment from industries
Urban form
Urban development by governments
Local community
Number of people who relocate to the place
Human mobility on non-working days
(Opportunity cost of travel time calculated from non-work travel behavior)
Air quality
Closeness to business area
Legend
Factor investigated in this paper
Factor not investigated in this paper
Direction of causality

Fig. 6.5 Detailed hypothesis on the causalities.

study may be biased by such factors. In regard to selecting predictors for a causal analysis, we select predictors whose correlation coefficients with the number of people who relocate to each place is high. It is difficult to determine whether the relation between two variables is linear. In our analysis, the correlation coefficient between every combination of the three variables (P, O, and M) is very high, but it is manually determined. When our method is applied to other data, the selection of variables for causal analysis must be carefully done.

6.9 Conclusion

In this work, we have investigated a method for creating an index based on mobility data to estimate the effects of urban development on residential movements. The contributions of this work are summarized as follows:

- We have extended the conventional travel cost method to estimate the opportunity cost of travel time as a function of the departure place.
- We have confirmed that both the current population and the opportunity cost of travel time contribute to the prediction of the number of relocations.

- We have confirmed that most of the causal influence from the current population to residential mobility is mediated indirectly through the opportunity cost of travel time. Therefore, the opportunity cost of travel time is more effective at estimating changes in residential mobility caused by urban development.

Our method is aimed at predicting future residential mobility and detecting future changes in residential mobility. The method can also be used to evaluate urban development. In planning of a new urban form, discrete choice models for destination choices [13] and housing choices [66] are suitable means. In addition, Walk Score [118] can also contribute to planning a new urban form.

We note some limitations of our work. First, the opportunity cost of travel time is influenced not only by urban form but also by other factors such as people's disposable time and activities using ICTs while traveling (e.g., mobile phones and potable gaming devices). Second, residential mobility is determined by other factors such as closeness to business areas and air quality. Third, the causal networks are not theoretically derived; therefore, it is necessary to analyze the causalities from a theoretical perspective.

Further research is needed to investigate which factors of urban form influence the opportunity cost of travel time. In addition, further research is needed to extend the method to include other types of transportation such as automobiles.

Recent smart city initiatives are aimed at using a mixture of various types of data to solve a mixture of various urban problems, such as transportation, energy use, and economy [83]. Analyzing the relationships between human mobility and residential mobility marks an important step towards understanding urban dynamics and, thereby, helps planning and managing public transport and economic development. Despite the above limitations, our proposed method is beneficial to urban planners for estimating the effects of urban development and detecting the shrinkage and growth of populations.

Chapter 7

Extraction of the community structure of a city

7.1 Introduction

A city is a complex system where areas of a city interact with each other through the medium of people, things, and information. Cities evolve by the feedback loop of such interactions. Research into urban systems has been fostered by the availability of the massive data of spatial interactions in urban spaces, such as telecommunication data and human mobility data [133]. Such data are advantageous because they enable us to see both temporally and spatially detailed aspects of cities, and they are useful for urban planning and assessing the outcome of urban development. For example, network communication data are used for investigating the economic development by area in the United Kingdom [34], and smart card data of public transportation are used for analyzing the polycentric structure of London [98].

It has been observed that cities have community structures [98], as with many other types of complex systems. Network theory provides effective measures to discover the community structures from relational data of complex systems. Newman's community detection method [86, 84, 30] partitions an entire undirected network into tightly connected subnetworks (communities). The Infomap algorithm proposed by Rosvall and Bergstrom [97] handles directed networks using the probability flow of random walks. One of the advantages of the above community extraction methods is that they do not require prior information about the number of communities in advance. The methods determine the most suitable number of communities.

Community extraction methods have played important roles in understanding the structures of cities using human mobility data [3, 70, 31, 96, 134] or telecommunication data [22, 95, 105, 16]. Each node represents a station or a discretized space (e.g grid cell, polygon) of a city, and each edge represents human flows or telecommunications between two areas. The community structure of a city is obtained by partitioning such networks. Such information is useful for the analysis of urban dynamics, urban planning, and assessing the outcome of urban development. Zhong et al. [134] apply the Infomap algorithm to smart card data collected in Singapore and conclude that the city-state has been rapidly transforming to adopt a polycentric urban form. Amini et al. [3] apply Newman's community extraction method to mobility data collected from mobile phones in the Ivory Coast and Portugal. One of their findings is that the communities identified for Portugal exhibited high similarity with the official administrative boundaries, and those for Ivory Coast are strongly affected by tribal borders.

Despite the progress of the studies in extracting community structures of a city, the functional relations between areas of a city have not been considered in previous studies on extracting community structures of cities. In regard to human mobility, a trip is manifested through the need to access places and participate in activities such as working and shopping [57]. Therefore, human mobility reflects the functional relations between areas of a city. Such an assumption has been used to understand the social function of each area by classifying origin-destination relationship using human mobility data [71]. Previous studies on extracting community structures of a city have not distinguished between locations of people's residences and destinations where they participate in activities. They create networks based on the number of trips between areas of a city. We consider that it would be possible to obtain more detailed information about community structures of a city by distinguishing between locations of people's residences and destinations where they participate in activities.

Furthermore, we consider that bipartite network clustering methods are suitable to analyze functional relations between areas of a city. A bipartite network is constituted of nodes that are divided into two non-overlapping sets, and every edge connects one node in a set to another node in the other set. Guimerà et al. [44] suggest that a directed unipartite network can be represented as a bipartite network, where each node i is represented by two nodes A_i and B_i. A directed link from i to j is represented as an edge connecting A_i to B_j in the bipartite network. We consider that a place can be represented by two nodes, one as an origin and the other as a destination, and bipartite network clustering methods enable us to handle origins and destinations separately. Unipartite network clustering methods such as Newman's method and Infomap cannot cluster areas as shown in Fig. 7.1, because the weights of edges between areas of the same function are likely small. Bipartite network clustering methods are originally aimed at clustering two different types of entities such as researchers and their

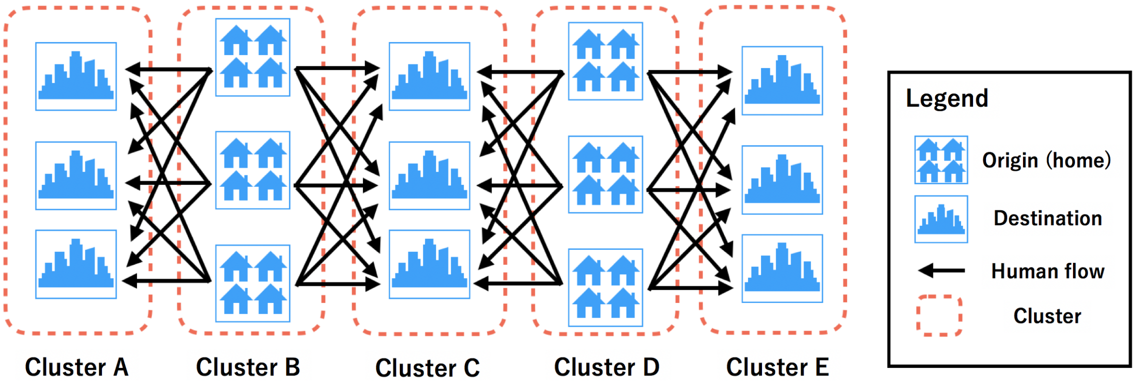


Fig. 7.1 Human mobility and functional clusters of areas of a city.

(co-authored) research papers. Therefore, they are able to cluster areas of a city as shown in Fig. 7.1. There are two types of bipartite clustering methods: The first one assumes that one cluster is strongly connected to only one cluster of the other side. The second method assumes that a cluster is connected to many clusters of the other side. We consider that the latter one is suitable to analyze community structures of a city. For example, cluster B in Fig. 7.1 is strongly connected to clusters A and C.

Gaps in previous studies are summarized by the following two points:

1. Previous studies have not created networks by distinguishing between locations of people's residences and destinations where they participate in activities.
2. Bipartite network clustering methods have not been used in previous studies.

In this chapter, we compare two types of origin-destination matrices generated from the smart card data of public transportation in Japan for constructing networks:

- OD Type 1: Origins and destinations respectively denote boarding stations and alighting stations of all trips. Each element of this type of matrix is the number of human flows between two stations.
- OD Type 2: Origins denote stations in passengers' residential neighborhoods, and destinations denote stations in areas where passengers participate in activities (e.g. work, shopping). Each element of this type of matrix is the number of passengers.

and we also compare four types of clustering methods:

1. Unipartite undirected network clustering method (Newman's method)
2. Unipartite directed network clustering method (Infomap)

3. Bipartite network clustering method considering single-facet characteristics of each cluster

4. Bipartite network clustering method considering multi-facet characteristics of each cluster

Another contribution of this chapter is our proposed evaluation metrics to evaluate the results of community detection from human mobility data. It is impossible to know or determine ground truth about the community structure of a city. Some research attempts to evaluate the results of community extraction from human mobility data by comparing them with administrative boundaries [31], but the communities extracted from a network generated from human mobility data are not always in line with administrative boundaries [3]. It is necessary to make evaluation metrics for investigating which community detection method is the most suitable to extract the community structure of a city. In this chapter, we propose three metrics to evaluate the results of community detection methods for human mobility data. The first metric is based on the geographical cohesiveness of each extracted cluster. Previous studies [95, 134, 105, 22] note that the extracted clusters show high cohesiveness. The second metric is based on how a clustering method captures the similarity of the community structures from two different weekdays. The third metric is based on the amount of information the result of a community detection method shows.

The remainder of this chapter is organized as follows: Section 7.2 explains the two types of OD matrices compared in this chapter. Section 7.3 reviews unipartite and bipartite network clustering methods and their applications to human mobility data. At the end of the section, we introduce the methods compared in this chapter. Section 7.4 describes our proposed evaluation methods for extracted communities from human mobility data. Section 7.5 explains the smart card data of public transportation that we use. Section 7.6 reports on the results. Section 7.7 discusses the implications and limitations of our study. Section 7.8 draws our conclusions.

7.2 Origin-destination matrix creation

We compare two types of origin-destination (OD) matrix for constructing networks, called OD Type 1 and OD Type 2.

OD Type 1 has information about the number of human flows between stations. Element $m_{i,j}$ of OD Type 1 matrix $\mathbf{M}$ represents the number of all trips from station i to station j within a day.

In regard to OD Type 2, an origin is defined as the station at which a smart card holder boarded first within a day. Destinations are all the stations at which a smart card holder alighted within the day excluding the origin station. OD Type 2 is aimed at representing the relation between the locations of passengers' residences and places where they participated in activities such as working and shopping. Element $m_{i,j}$ of OD Type 2 matrix $\mathbf{M}$ denotes the number of passengers who boarded at station i first within the day and alighted at station j within the day.

The difference between OD Type 1 and OD Type 2 is as follows: Suppose that four trips of a person are recorded within a day: The trips are $\{(s_1 \rightarrow s_2), (s_2 \rightarrow s_3), (s_3 \rightarrow s_2), (s_2 \rightarrow s_1)\}$ where s_1, s_2, and s_3 denote stations. In regard to OD Type 1, $m_{1,2}$, $m_{2,3}$, $m_{3,2}$, and $m_{2,1}$ increase by 1 respectively. In regard to OD Type 2, $m_{1,2}$ and $m_{1,3}$ increase by 1 respectively. In the latter case, the origin is s_1, because it is the station at which the person boarded first within the day. We count every trip for OD Type 1. On the other hand, the number of trips to the same destination by one person is counted as 1 for OD Type 2.

OD Type 1 has been used in previous studies to extract community structures from human mobility data [3, 70, 31, 96, 134], while OD Type 2 has not been used in previous studies.

7.3 Network clustering methods

7.3.1 Unipartite network clustering methods

Community extraction methods for unipartite networks have been used for analyzing community structures of a city. A unipartite network $G = (V, E)$ is a pair of a node set (V) and an edge set ($E \subset V \times V$). A clustering $\mathbf{C} = \{C_1, ..., C_k\}$ for a network $G = (V, E)$ is a partitioning of vertexes into the disjoint subsets of clusters.

Newman's modularity-based method [85] is proposed to extract communities from undirected unipartite networks, and the Infomap algorithm [97] is proposed to extract communities from directed unipartite networks.

Newman's modularity-based clustering method

Newman's modularity-based method [86, 84, 30] is proposed to extract communities from undirected unipartite networks. It extracts communities maximizing the modularity defined in Equation 7.1:

$$Q_{\mathrm{N}} = \sum_{C_i \in \mathbf{C}} \left(\frac{||C_i \rightarrow C_i||}{|E|} - \frac{||C_i \rightarrow V||^2}{|E|^2} \right) \tag{7.1}$$

where $||C_i \to C_i||$ denotes the number of intra-cluster connections, and $|E|$ denotes the total number of edges in the graph. We note that we effectively count each edge twice ($(u,v) \in E$ and $(v,u) \in E$) to calculate $|E|$. This is because each element of an OD matrix denotes the value of a directed edge.

Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. The Louvain algorithm [17] is one of the most widely used algorithms to maximize the modularity. It is more scalable and gives better modularity values.

This method is used in the studies of community extraction from the networks of human mobility collected from telephones [3], worldwide air transportation [43], telecommunication patterns of mobile phones [22, 16, 95, 105], and worldwide geolocated Twitter data [51].

Infomap algorithm

The Infomap algorithm [97] is proposed to extract communities from directed unipartite networks. The Infomap algorithm uses information-theoretic techniques and random walks. The algorithm obtains network partition $\mathbf{C}$ minimizing $L(\mathbf{C})$ defined in Equation 7.2:

$$L(\mathbf{C}) = q_{\curvearrowright} H(Q) + \sum_{C_i \in \mathbf{C}} p^i_{\circlearrowright} H(P^i) \tag{7.2}$$

The first term of this equation indicates the entropy of the movement between modules, and the second term indicates the entropy of movements within modules. Here, $q_{\curvearrowright}$ and $H(Q)$ are defined by Equations 7.3 and 7.4:

$$q_{\curvearrowright} = \sum_{C_i \in \mathbf{C}} q_{i\curvearrowright} \tag{7.3}$$

$$H(Q) = -\sum_{C_i \in \mathbf{C}} \frac{q_{i\curvearrowright}}{q_{\curvearrowright}} \log\left(\frac{q_{i\curvearrowright}}{q_{\curvearrowright}}\right) \tag{7.4}$$

where $q_{i\curvearrowright}$ denotes the probability that the random walk exits module C_i. The weight $p^i_{\circlearrowright} = \sum_{\alpha \in C_i} p_\alpha + q_{i\curvearrowright}$ is comprised of two terms, the fraction of within-module movements in module C_i and the probability of exiting module C_i (p_α is the probability that the random walk visits node α). $H(P_i)$ is the entropy of the within-module movements in module C_i and exiting movements from module C_i defined by Equation 7.5.

$$H(P^i) = -\sum_{\alpha \in C_i} \frac{p_\alpha}{p^i_{\circlearrowright}} \log\left(\frac{q_\alpha}{p^i_{\circlearrowright}}\right) - \frac{q_{i\curvearrowright}}{p^i_{\circlearrowright}} \log\left(\frac{q_{i\curvearrowright}}{p^i_{\circlearrowright}}\right) \tag{7.5}$$

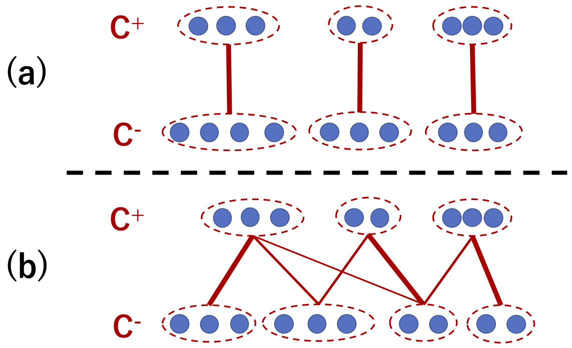


Fig. 7.2 Assumptions of bipartite network clustering methods: (a) One-to-one relationships. (b) Multi-facet relationships.

The network partition **C** minimizing $L(\mathbf{C})$ is obtained by using a deterministic greedy search algorithm with a simulated annealing approach.

This method is used in the studies of community extraction from the networks of taxi cabs' travel patterns [70], human mobility collected from telephones [3], GPS dataset collected from vehicles [31, 96], the smart card data of public transportation [134], and individual mobility pattern [6].

7.3.2 Bipartite network clustering methods

A bipartite network models the relationship between two different types of entities. Vertexes V is divided into two disjoint sets, V^+ and V^-. Every edge in a bipartite network connects a vertex in V^+ to a vertex in V^-. In our study, a station is represented by two nodes where the two nodes represent the station as an origin and a destination respectively.

There are two types of bipartite network clustering methods. The first assumes that each cluster of one side is strongly connected to one cluster of the other side, so the number of clusters is the same between the two sides (Fig. 7.2-(a)). The other relaxes the assumption, and it assumes that each cluster has multi-facet relations to clusters of the other side (Fig. 7.2-(b)).

To the best of our knowledge, no bipartite network clustering method has been applied to human mobility data to extract the community structure of a city.

Dormann-Strauss's modularity-based clustering method

Barber [7] modifies Newman's modularity-based clustering method to be able to cluster a binary bipartite network. Then Dormann and Strauss [33] extend it to deal with a weighted bipartite network. The method gives each node a label, which finally determines each node's cluster. The number of labels (clusters) is the same for V^+ and V^- (therefore, $|\mathbf{C}^+| = |\mathbf{C}^-|$). The modularity is defined by Equation 7.6:

$$Q_B = \sum_{C_k \in \mathbf{C}^+, C_l \in \mathbf{C}^-} \delta\left(S(C_k), S(C_l)\right) \cdot \left(\frac{||C_k \to C_l||}{|E|/2} - \frac{||C_k \to V|| \cdot ||C_l \to V||}{(|E|/2)^2}\right) \tag{7.6}$$

where $S(C)$ denotes the label of cluster C, δ is Kronecker's delta (i.e., $\delta(i,j) = 1$ if $i = j$, and $\delta(i,j) = 0$ if $i \neq j$), and $||C_k \to C_l||$ denotes the density of edges connecting from a cluster C_k to another cluster C_l, and $||C_i \to V||$ denotes the accumulated degrees for cluster C_i. Beckett [12] compares algorithms that are aimed at maximizing the modularity and find that an algorithm named DIRTLPAwb+ proposed by Beckett [12] marks the highest modularity.

Suzuki-Wakita's modularity-based clustering method

The above bipartite modularity model assumes that one cluster is strongly connected to only one of the clusters of the other side. Suzuki and Wakita [107] extend the method to incorporate multi-facet aspects of each community. It assumes that one community has connections to more than one cluster of the other side. The model does not assume that the number of communities is not always the same for V^+ and V^- (therefore, $|\mathbf{C}^+| \neq |\mathbf{C}^-|$). The modularity is defined by Equation 7.7:

$$Q_S = \frac{1}{2} \sum_{C_k, C_l \in \mathbf{C}} \frac{||C_k \to C_l||}{||C_k \to V||} \cdot \left(\frac{||C_k \to C_l||}{|E|/2} - \frac{||C_k \to V|| \cdot ||C_l \to V||}{(|E|/2)^2}\right) \tag{7.7}$$

7.3.3 Methods compared in this study

In this study, we compare the following methods for clustering human mobility networks:

- **Newman**[17]: A clustering method for undirected unipartite networks using Louvain algorithm to maximize Newman's modularity.
- **Infomap**[97]: A clustering method for directed unipartite networks using the probability flow of random walks.
- **DIRTLPAwb+**[12]: A clustering method for bipartite networks based on Dormann-Strauss's modularity capturing single-facet characteristics of communities.

- **Multi-Facet**[107]: A clustering method for bipartite networks based on Suzuki-Wakita's modularity for bipartite networks capturing multi-facet characteristics of communities.

7.4 Evaluation metrics for community extraction

There is no ground truth about community structures of cities, so it is impossible to evaluate the results of community extraction by comparing the results with labeled data. We propose evaluation methods for community extraction from human mobility data based on the following three points:

- Geographical cohesiveness of clusters.
- Similarity between the sets of clusters obtained from two weekdays.
- Amount of information in the result of community extraction.

7.4.1 Geographical cohesiveness of clusters

The results of previous studies [95, 134, 105, 22, 51] show that clustered areas exhibit geographical cohesiveness and boundaries. However, they do not quantitatively define the geographical cohesiveness of clusters.

We propose a method to quantitatively evaluate the geographical cohesiveness of clusters. Suppose that a clustering method divides stations into three clusters A, B, and C as shown in Fig. 7.3. Let stations in cluster A be further divided into clusters A_1 and A_2 using distance m, such that the distance between the two clusters is longer than distance m. We assume that cluster A is not geographically cohesive if distance m is too long.

We define an evaluation metric called **GC-Measure** to calculate the cohesiveness of clusters based on Algorithm 1. The set of clusters $\mathbf{C}$ is further divided into the set of clusters $\mathbf{G}$ by using distance m. The algorithm incrementally increases m per unit distance u. The incremental procedure ends when $\mathbf{C} = \mathbf{G}$, that is, when each cluster in set $\mathbf{C}$ is not further divided into smaller clusters by using distance m. GC-Measure is defined as the shortest distance m that does not further divide $\mathbf{C}$, which is obtained by Algorithm 1. In our definition, the higher the GC-Measure is, the less cohesive the set of clusters are. Therefore, a lower GC-Measure is preferable. The process of the lines 7-14 of Algorithm 1 is based on DBSCAN (density-based spatial clustering of applications with noise) proposed by Ester et al. [35]. By the lines 7-14 of Algorithm 1, all the pairs of nodes within distance m are clustered together.

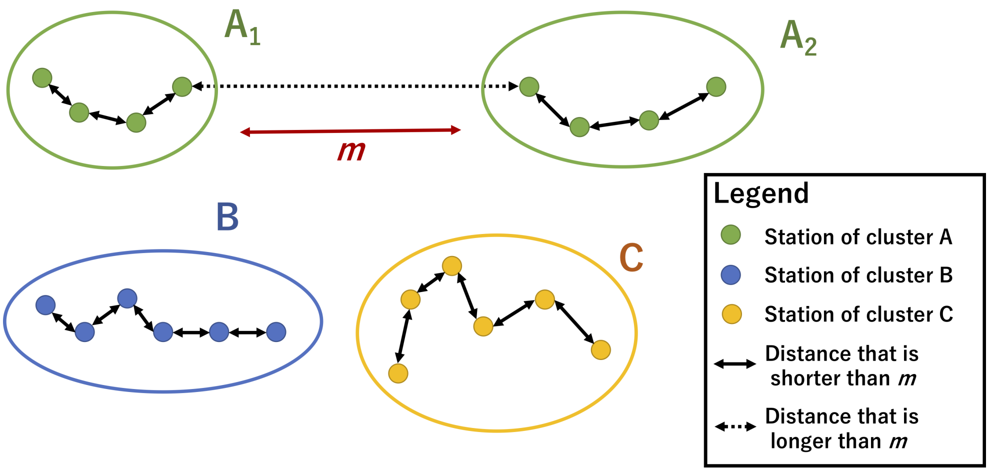


Fig. 7.3 Geographical cohesiveness of clusters: If distance m is very long, we conclude that the cohesiveness of cluster A is low.

7.4.2 Similarity between the sets of clusters obtained from two weekdays

Human mobility has high regularity such as trips between home and work/school [42]. Zhong et al. [136] analyze the similarity of human mobility between different days of the week. The results show the high regularity of human mobility patterns from Monday to Thursday. We consider that the regularity is due to the mobility pattern of commuters.

We measure the similarity of the two sets of clusters extracted from human mobility data on two different Mondays. To measure the similarity of the two sets of clusters, we use normalized mutual information (NMI). We define an evaluation metric named **R-Measure** that measures the regularity of human mobility defined by Equation 7.8 to measure the degree a clustering method captures the regularity of human mobility:

$$\text{R-Measure} = \text{NMI}\Big(\mathbf{C}(d_1), \mathbf{C}(d_2)\Big) \tag{7.8}$$

where d_1 and d_2 denote different weekdays (Monday), and $\mathbf{C}(d_1)$ and $\mathbf{C}(d_2)$ denote a set of clusters extracted from smart card data collected from day d_1 and day d_2 respectively. There are various definitions of NMI. Vinh et al. [115] perform a comparative analysis of the different definitions of NMI. A distance measure using the definition of Equation 7.9 satisfies the metric property according to their study, so we use the definition for calculating R-Measure.

$$\text{NMI}(\mathbf{C}_1, \mathbf{C}_2) = \frac{I(\mathbf{C}_1, \mathbf{C}_2)}{H(\mathbf{C}_1, \mathbf{C}_2)} \tag{7.9}$$

Algorithm 1 Algorithm to calculate GC-Measure

```
Input:   $\mathbf{C}$: the set of clusters
         $u$: unit distance
Output:  $m$: the shortest distance that does not further divide all the clusters in set $\mathbf{C}$
$m \leftarrow 0$
$\mathbf{G} = \{\}$
while $\mathbf{C} \neq \mathbf{G}$ do
    $m \leftarrow m + u$
    $\mathbf{G} = \{\}$
    for $C \in \mathbf{C}$ do
        $\mathbf{T} \leftarrow \{\{v\}|v \in C\}$ // Prepare singleton clusters
        while There is a change in $\mathbf{T}$ do
            for $T_1, T_2 \in \mathbf{T}$ do
                if $\exists v_1 \in T_1,\ \exists v_2 \in T_2,\quad distance(v_1, v_2) < m$ then
                    Concatenate $T_1$ and $T_2$
                end if
            end for
        end while
        Add every element $T \in \mathbf{T}$ to $\mathbf{G}$
    end for
end while
```

where $I(\mathbf{C_1}, \mathbf{C_2})$ denotes the mutual information of the sets of clusters $\mathbf{C_1} = \{C_1^1, \cdots, C_k^1\}$ and $\mathbf{C_2} = \{C_1^2, \cdots, C_k^2\}$, and $H(\mathbf{C_1}, \mathbf{C_2})$ denotes the joint entropy of the sets of the clusters. $H(\mathbf{C_1})$, $H(\mathbf{C_1}, \mathbf{C_2})$, and $I(\mathbf{C_1}, \mathbf{C_2})$ are defined by Equations 7.10, 7.11, and 7.12, using the contingency table illustrated by Table 7.1, where $n_{i,j}$ denotes the number of stations that are common to clusters $C_i^1 \in \mathbf{C_1}$ and $C_j^2 \in \mathbf{C_2}$, and a_i and b_j are respectively the number of stations in set C_i^1 and the number of stations in in set C_j^2.

$$H(\mathbf{C_1}) = -\sum_{i}^{k} \frac{a_i}{N} \log\left(\frac{a_i}{N}\right) \tag{7.10}$$

$$H(\mathbf{C_1}, \mathbf{C_2}) = -\sum_{i}^{k} \sum_{j}^{l} \frac{n_{i,j}}{N} \log\left(\frac{n_{i,j}}{N}\right) \tag{7.11}$$

$$I(\mathbf{C_1}, \mathbf{C_2}) = \sum_{i}^{k} \sum_{j}^{l} \frac{n_{i,j}}{N} \log\left(\frac{n_{i,j}/N}{a_i b_j / N^2}\right) \tag{7.12}$$

Table 7.1 Contingency table, $n_{i,j} = |C_i^1 \cap C_j^2|$.

	C_1^2	C_2^2	$\cdots$	C_l^2	Sums
C_1^1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,l}$	a_1
C_2^1	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,l}$	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_k^1	$n_{k,1}$	$n_{k,2}$	$\cdots$	$n_{k,l}$	a_k
Sums	b_1	b_2	$\cdots$	b_l	$N = \sum_{i,j} n_{i,j}$

7.4.3 Amount of information

We also evaluate network clustering methods by measuring the amount of the information of extracted communities. Human mobility exhibits similarity on the same days of the week, and also exhibits variabilities between different days of the week. We measure the amount of information in regard to the similarity and the variability of urban human mobility.

We create three metrics based on information entropy named **AI-Measures I**, **II** and **III**. They measure the amount of the information of the areas colored green in Figs. 7.4-(a,b,c) respectively. Fig. 7.4-(a) shows the intersection of the two sets of clusters extracted from different weekdays (Monday). Fig. 7.4-(b) shows the intersection of the two sets of clusters extracted from different weekends (Saturday), where e_1 and e_2 denote different weekends (Saturday). Fig. 7.4-(c) shows the difference between the green areas in Fig. 7.4-(a) and in Fig. 7.4-(b). The green area in Fig. 7.4-(c) represents the variability of urban human mobility between weekdays and weekends.

AI-Measures I, II and III are defined by Equations 7.13, 7.14, and 7.15.

$$\text{AI-Measure I} = I\big(\mathbf{C}(d_1), \mathbf{C}(d_2)\big) \tag{7.13}$$

$$\text{AI-Measure II} = I\big(\mathbf{C}(e_1), \mathbf{C}(e_2)\big) \tag{7.14}$$

$$\begin{aligned}\text{AI-Measure III} &= VI\big(\mathbf{C}(d_1) \cap \mathbf{C}(d_2), \mathbf{C}(e_1) \cap \mathbf{C}(e_2)\big) \\ &= I\big(\mathbf{C}(d_1), \mathbf{C}(d_2)\big) + I\big(\mathbf{C}(e_1), \mathbf{C}(e_2)\big) - 2 \cdot I\big(\mathbf{C}(d_1), \mathbf{C}(d_2), \mathbf{C}(e_1), \mathbf{C}(e_2)\big)\end{aligned} \tag{7.15}$$

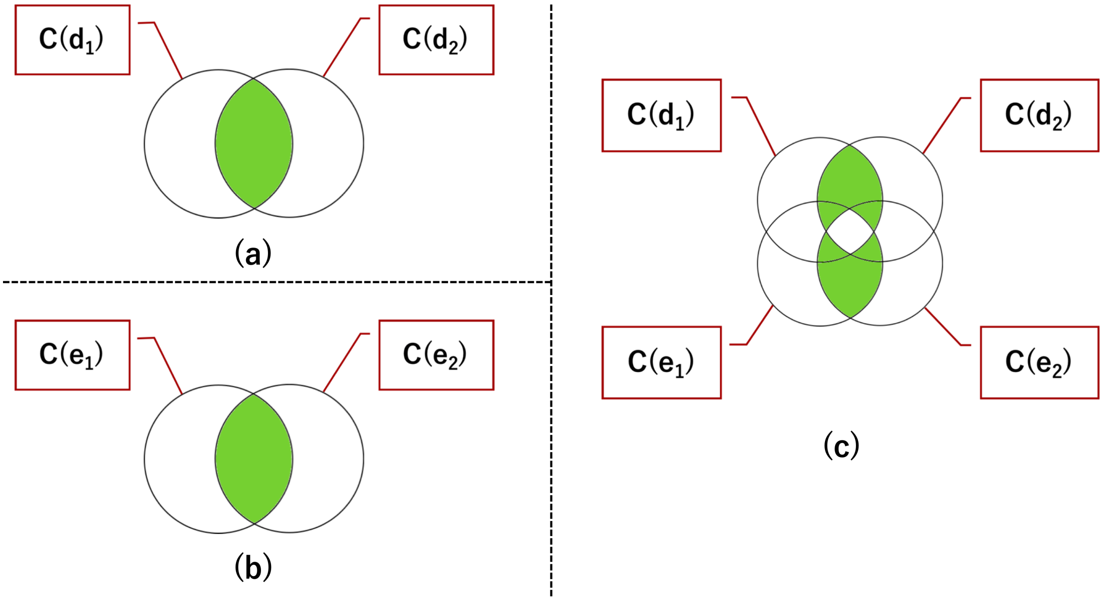


Fig. 7.4 Areas for which we calculate the amount of information.

where VI denotes variation of information. The definition of the mutual information of the four sets is the same as the definition for the two sets described in Equation 7.12.

We consider the functional relations between areas of a city vary between weekdays and weekends. AI measure Ⅲ is designed to evaluate how a clustering method can capture the variation of urban human mobility.

7.4.4 Summary of the measures

Table 7.2 shows the summary of the measures.

Table 7.2 Summary of the measures

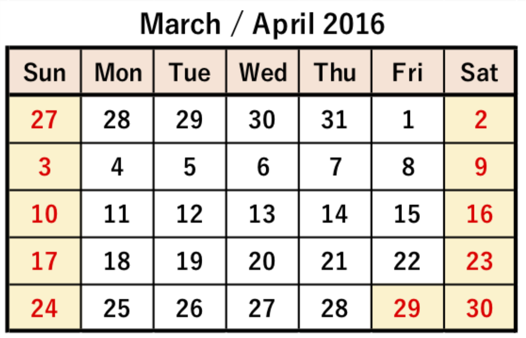
Measure	Explanation
GC-Measure	When the measure is low, the geographical cohesiveness of clusters is concluded to be high.
R-Measure	When the measure is high, the result of a clustering method is concluded to be able to capture the regularity of human mobility on working days.
AI-Measure I	When the measure is high, the result of a clustering method is concluded to exhibit rich information about the regularity of human mobility on working days.
AI-Measure Ⅱ	When the measure is high, the result of a clustering method is concluded to exhibit rich information about the regularity of human mobility on non-working days.
AI-Measure Ⅲ	When the measure is high, the result of a clustering method is concluded to exhibit rich information about the variabilities of human mobility between working days and non-working days.

7.5 Data

We perform three sets of experiments. The dates of the experiments are shown in Table 7.3. Fig. 7.5 shows working days and non-working days. April 29 is a national holiday. We consider that there is not much effect of the national holiday on the human mobility on the dates of the study.

Table 7.3 The dates of the study.

	d_1 **(Monday)**	d_2 **(Monday)**	e_1 **(Saturday)**	e_2 **(Saturday)**
Experiment 1	04 April 2016	11 April 2016	02 April 2016	09 April 2016
Experiment 2	11 April 2016	18 April 2016	09 April 2016	16 April 2016
Experiment 3	18 April 2016	25 April 2016	16 April 2016	23 April 2016



March / April 2016

Sun	Mon	Tue	Wed	Thu	Fri	Sat
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

Fig. 7.5 Working days and non-working days in the study period.

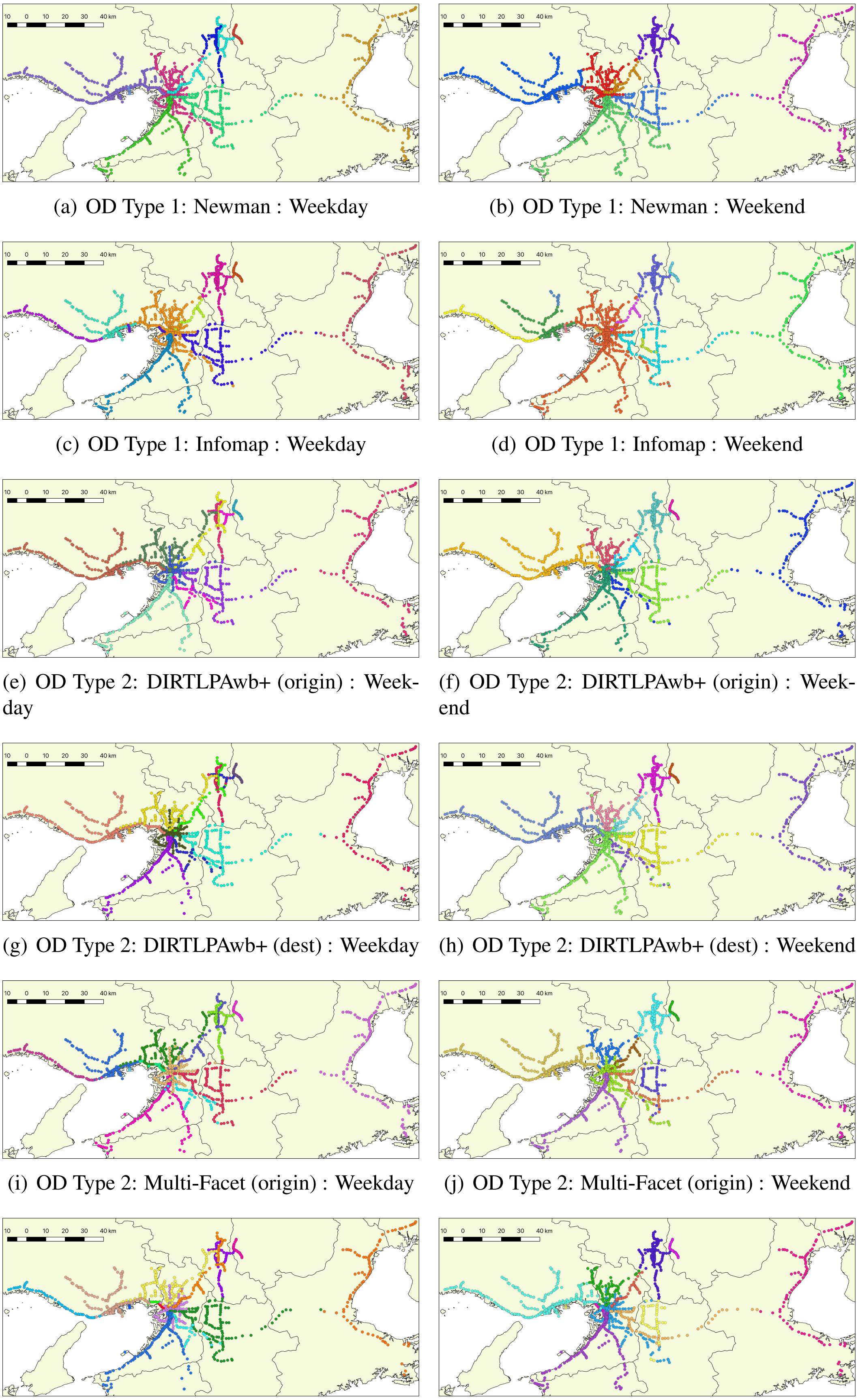
7.6 Results

7.6.1 Overview of the results

Table 7.4 shows the numbers of extracted clusters, and Fig. 7.6 shows the examples of the spatial distributions of extracted clusters using data on 11 April 2016 (Monday) and 9 April 2016 (Saturday). The difference of colors in Fig. 7.6 indicates the difference of clusters, and the colors are randomly selected. In regard to bipartite network clustering methods (DIRTLPAwb+ and Multi-Facet), the results are shown separately for origins and destinations. Similarities and differences can be seen among the spatial distributions, but it is difficult to compare or evaluate the results only from the figures. The following sections compare results by our proposed metrics.

Table 7.4 Number of extracted clusters

			Date (weekday)				Date (weekday)			
OD matrix	**Network**	**Method**	**04 April**	**11 April**	**18 April**	**25 April**	**02 April**	**09 April**	**16 April**	**23 April**
OD Type 1	Unipartite	Newman	8	9	8	9	8	7	9	7
	Unipartite	Infomap	12	11	12	12	14	12	13	12
	Bipartite(origins)	Multi-Facet	12	12	13	13	9	12	12	11
	Bipartite(destinations)	Multi-Facet	12	12	13	13	9	12	11	12
	Bipartite(origins)	DIRTLPAwb+	8	9	8	9	7	7	7	6
	Bipartite(destinations)	DIRTLPAwb+	8	9	8	9	7	7	7	6
OD Type 2	Unipartite	Newman	8	8	6	6	6	6	6	6
	Unipartite	Infomap	11	12	12	10	12	12	11	12
	Bipartite(origins)	Multi-Facet	15	14	17	16	10	12	11	11
	Bipartite(destinations)	Multi-Facet	16	15	17	16	11	13	14	14
	Bipartite(origins)	DIRTLPAwb+	10	10	10	14	5	9	6	8
	Bipartite(destinations)	DIRTLPAwb+	10	10	10	14	5	9	6	8



(a) OD Type 1: Newman : Weekday

(b) OD Type 1: Newman : Weekend

(c) OD Type 1: Infomap : Weekday

(d) OD Type 1: Infomap : Weekend

(e) OD Type 2: DIRTLPAwb+ (origin) : Weekday

(f) OD Type 2: DIRTLPAwb+ (origin) : Weekend

(g) OD Type 2: DIRTLPAwb+ (dest) : Weekday

(h) OD Type 2: DIRTLPAwb+ (dest) : Weekend

(i) OD Type 2: Multi-Facet (origin) : Weekday

(j) OD Type 2: Multi-Facet (origin) : Weekend

(k) OD Type 2: Multi-Facet (dest) : Weekday

(l) OD Type 2: Multi-Facet (dest) : Weekend

Fig. 7.6 Spatial distributions of extracted clusters

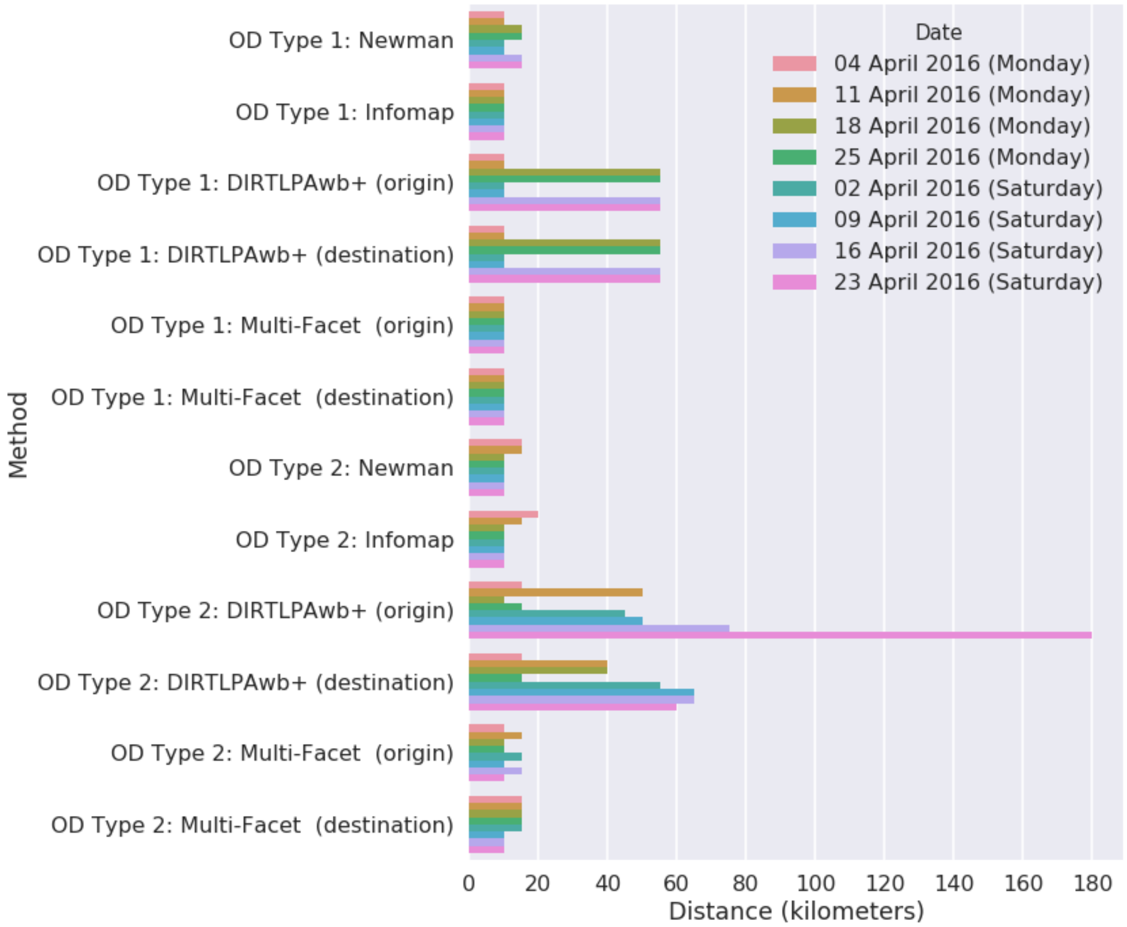


Fig. 7.7 Results of GC-Measure.

7.6.2 Results of GC-Measure

GC-Measure measures the geographical cohesiveness of the set of clusters. If the measure is low, the cohesiveness is high. Therefore, a lower GC-Measure is preferable. We set 5 kilometers to unit distance u, an input of Algorithm 1, to calculate GC-Measure. Fig. 7.7 shows the results of GC-Measure. In regard to Newman, Infomap, and Multi-Facet, the measure does not exceed 20 kilometers. However, many of the results of DIRTLPAwb+ are much higher than 20 kilometers. Fig. 7.8 shows one of the clusters whose geographical cohesiveness is low. It is the cluster of origin stations obtained from the data on 9 April 2016 (the same as Fig. 7.6-(f)). We, therefore, consider the extracted clusters of DIRTLPAwb+ are not geographically cohesive.

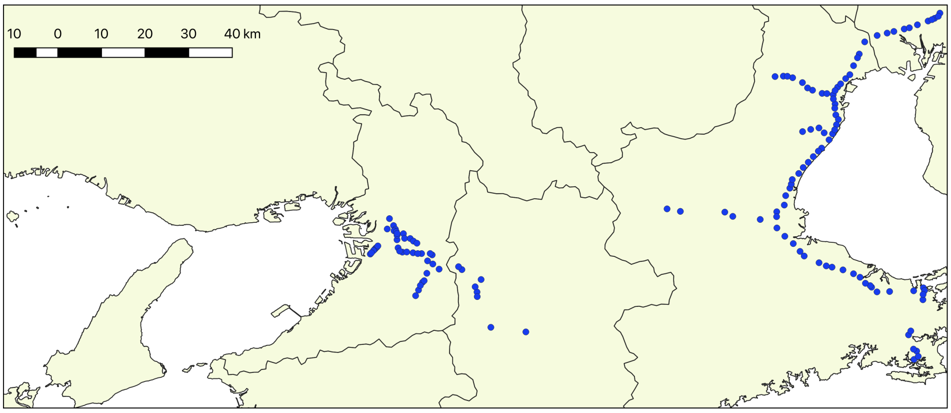


Fig. 7.8 A cluster created by DIRTLPAwb+ whose geographical cohesiveness is low.

7.6.3 Results of R-Measure

R-Measure measures the similarity of two networks constructed from human mobility data on two working days. It is intended to evaluate how a clustering method can capture the regularity of urban human mobility.

Fig. 7.9 shows the results of R-Measure. All the results of Multi-Facet using OD Type 2 score higher than those of the other methods. Therefore, it is concluded that Multi-Facet using OD Type 2 captures the similarity of human mobility on working days the best among all the four methods.

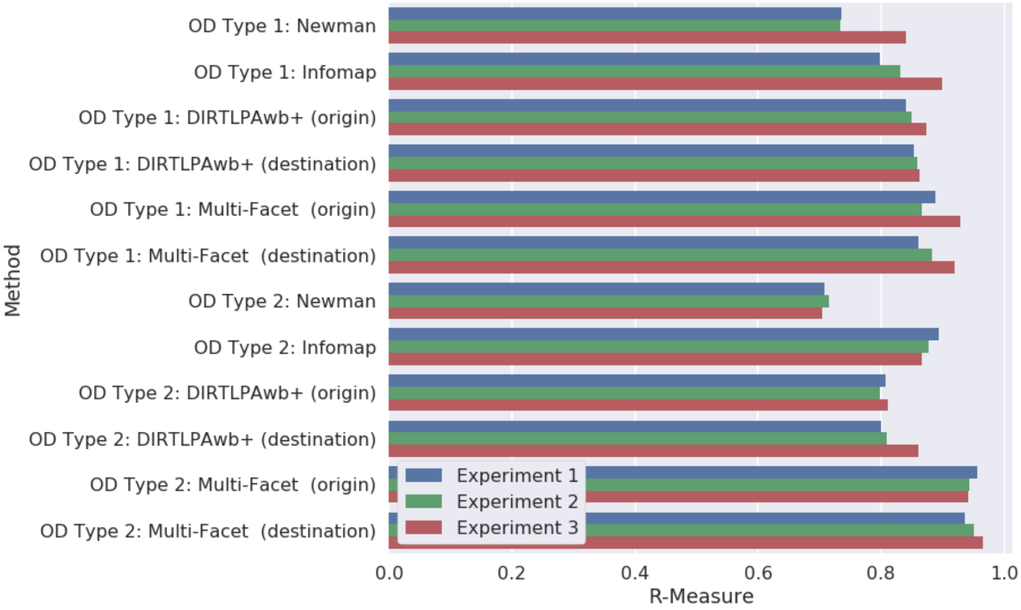


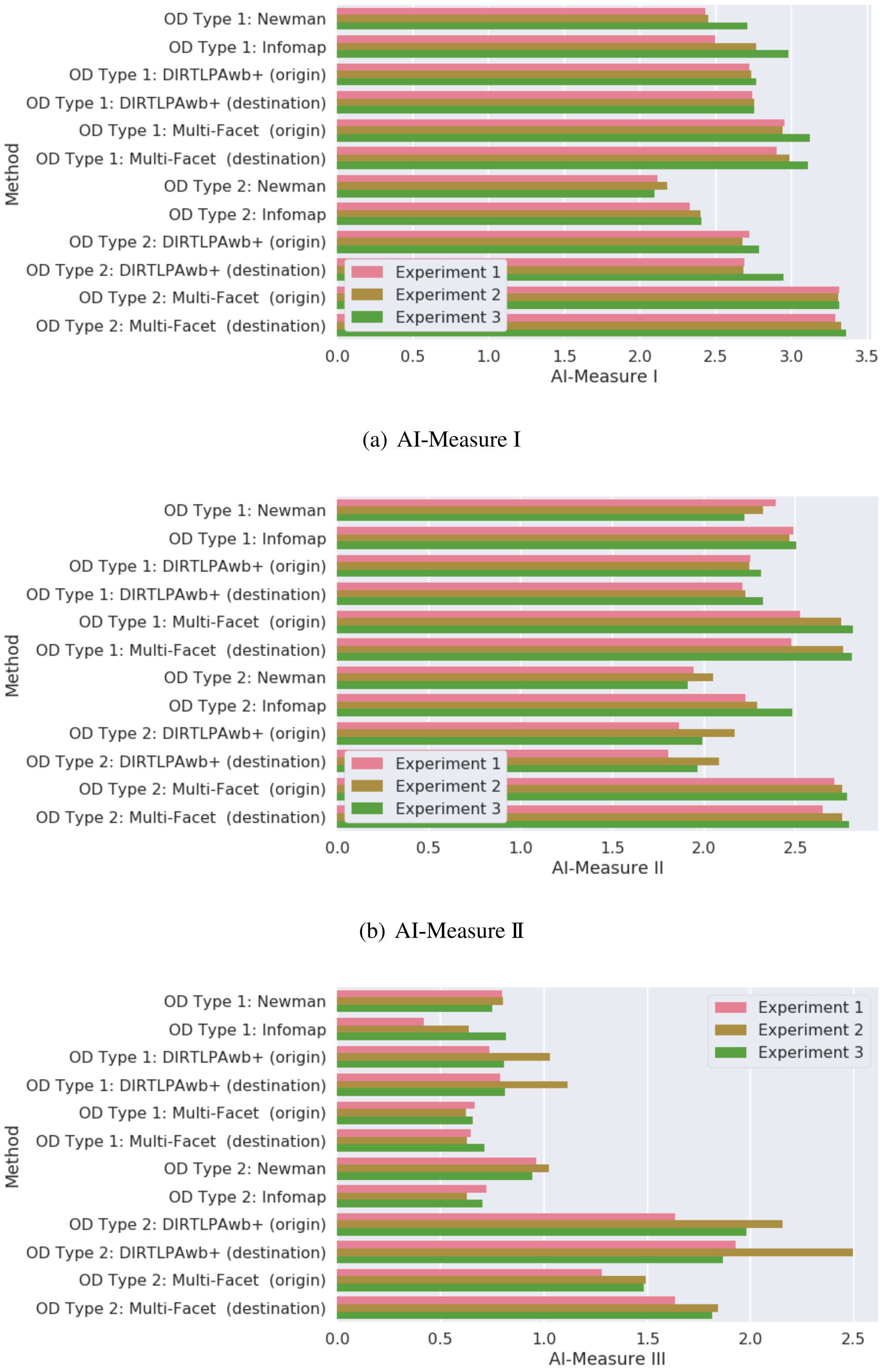
Fig. 7.9 Results of R-Measure.

7.6.4 Results of AI-Measures

AI-Measures measure the amount of information about the results of clustering. AI-Measure I and AI-Measure Ⅱ respectively measure the amount of information about the regularity of human mobility on working days and on non-working days. AI-Measure Ⅲ measures the amount of information about the variability of human mobility between working days and non-working days.

Figs. 7.10(a) and 7.10(b) show the results of AI-Measure I and Ⅱ respectively. In regard to AI-Measure I, the results of Multi-Facet using OD Type 2 exceed those of the other methods. In regard to AI-Measure Ⅱ, the results of Multi-Facet using OD Type 1 or 2 exceed those of the other methods. Therefore, it is concluded that Multi-Facet exhibits more information about the similarity of human mobility on both working days and non-working days than the other methods.

Fig. 7.10(c) shows the result of AI-Measure Ⅲ. The results of DIRTLPAwb+ using OD Type 2 exceed those of the other methods, and the results of Multi-Facet are the second highest.



(a) AI-Measure I

(b) AI-Measure II

(c) AI-Measure III

Fig. 7.10 The results of AI-Measures.

7.7 Discussion

We consider that Multi-Facet using OD Type 2 is the most suitable network clustering method for the following reasons:

- The clusters obtained by using DIRTLPAwb+ are not geographically cohesive according to the results of GS-Measure. Therefore, we consider DIRTLPAwb+ is not a suitable method.

- Multi-Facet using OD Type 2 scores the highest in regard to R-Measure and AI-Measures I and Ⅱ.

- Multi-Facet using OD Type 2 scores the second highest after DIRTLPAwb+ in regard to AI-Measure Ⅲ.

In regard to the two bipartite network clustering methods, it is necessary to take multi-facet characteristics of spatial interaction into consideration when analyzing urban human mobility. People move to various places located near to the location of their residence. Destinations also attract people from various places located near by. The bipartite network clustering method that considers single-facet characteristics of human mobility networks (DIRTLPAwb+) assumes that one-to-one relationships between the cluster of origins and that of destinations, so it is not suitable to analyze functional relations between areas of a city.

In regard to the difference between the two types of origin-destination matrices, AI-Measure Ⅲ of OD Type 2 is superior to that of OD Type 1 when using Multi-Facet. We consider this is because the functional relations of areas vary between weekdays and weekdays. In addition, the results of R-Measure and AI-Measures I obtained by applying Multi-Facet to OD Type 2 is superior to OD Type 1. We consider that OD Type 2 is more suitable for bipartite network clustering methods for the following reasons: OD Type 1 does not distinguish between trips from home to destination and trips back to home, and people usually make a round trip between home and destinations where they participate in activities. Therefore, it is likely that an OD Type 1 matrix is close to the transpose of the matrix, so it is unnecessary to apply bipartite network clustering methods to OD Type 1. Bipartite network clustering methods work effectively when there is a big difference between an OD matrix and the transposed matrix. Therefore, Multi-Facet works better using OD Type 2.

In regard to the undirected unipartite network clustering method (Newman), we consider that the size of each cluster is unnecessarily large. We consider that edges between areas that have similar functions tend to be small, because people usually move between residential areas and other types of areas. On the other hand, undirected network clustering methods assume that nodes in a cluster are densely connected to each other. Such an assumption is not suitable for networks created from urban human mobility, and it makes the size of each cluster unnecessarily large.

The Infomap algorithm is not suitable to analyze urban human mobility. It assumes that individuals move between many places in a network. Such an assumption is suitable when analyzing people's movement within a specific space such as amusement park.

The results in this study show that the bipartite network clustering method is able to show accurate and rich information about urban structures, taking into consideration the multi-facet characteristics of each cluster and functional relations between areas of a city.

7.8 Conclusion

Many studies have used network clustering methods for analyzing human mobility, but little has been explored about which method is suitable for extracting urban structures from origin-destination data. In this chapter, we have proposed evaluation metrics for evaluating network clustering methods based on the cohesiveness of extracted clusters, the regularity of human mobility on weekdays, and the amount of information about extracted clusters. The bipartite network clustering method marks good scores for these evaluation metrics, considering multi-facet characteristics of clusters and the functional relation between areas of a city. The previous studies on the extraction of urban structures from human mobility networks have used unipartite network clustering methods. The bipartite network clustering method shows accurate and rich information about urban structures. Therefore, it is recommended to use the method to analyze urban structures based on spatial interaction data.

It is necessary to choose the most suitable method when analyzing network data. We consider that the Infomap algorithm is the most suitable method for analyzing human mobility in a specific place such as an amusement park. Our future work will include the comparison of different types of human mobility. The evaluation metrics proposed in this study are useful for evaluating extracted urban structures. However, the metrics are not suitable for other types of human mobility data, such as human mobility in amusement parks. It is necessary to propose other metrics that are suitable for other types of human mobility data.

Chapter 8

Application of the proposed methods to urban design

The proposed methods have been developed for capturing and solving real problems in urban space, and they are designed to be applicable to any types of cities. In this chapter, we discuss the possible application of the proposed methods to real problems, by taking examples from Japan.

8.1 Current problems regarding urban development in Japan

Many cities in Japan are currently faced with serious and rapid changes. The population in the centers of cities has been rapidly increasing, and it causes many problems such as a shortage of elementary schools. In addition, commuter towns located far from city centers are faced with population aging and the shrinkage of the working population. Such changes cause an increase in vacant houses and the decrease of convenience.

One of the biggest reasons for these changes is the construction of high-rise buildings in city centers. Many high-rise apartments have been built in the centers of cities because of the deregulation of the floor area ratio according to the Building Standard Law of 1997. The deregulation is due to the advancement of building technology. In addition, many high-rise buildings that include many shops and offices have been built as a result of the Special Law for Urban Renaissance of 2002, which is aimed at stimulating the economy by creating new attractiveness in urban areas.

Economic changes also caused these changes. In the 1980s and early 1990s, Japan experienced a booming economy. Many private investments were poured into real estate. Old buildings were demolished, and vacant lots were left waiting for new redevelopments [55].

People cannot live in city centers because of the rise in land price. After the decades of economical regression, land price has decreased, and real estate developers have started constructing building in the centers of cities.

It is necessary to agglomerate the functions and population of each area of a city, and to connect areas of a city by transportation to overcome these difficulties. The Ministry of Land, Infrastructure, Transport and Tourism (MLIT) launched a priority measure called the Compact Plus Network to develop efficient cities [79].

8.2 Characteristics of urban development in Japan

The characteristics of land use policy and urban development in Japan can be described as below, according to the Japan International Cooperation Agency (JICA) [55].

In Japan, the utilization of land is considered to be left to land owner's will as long as there is no external nuisance. In addition, each lot is small and formless compared to Europe. These are the main reasons why it is difficult to control urban development in Japan, and they are also the main reasons for the rise in land prices in the 1980s and early 1990s. One of the objectives of urban planning is to regulate the construction of each building. Land Use Zoning and the District Plan are mainly used measures. Land Use Zoning is the most fundamental regulatory measure for controlling use, density, height and shape of buildings. Governments examine the applications of construction based on the regulatory rule according to the zone of the lot. The examination of applications is systematic and less time-consuming. The District Plan is a detailed and comprehensive land use planning system for wide areas of a city, for promoting the quality of urban environment. It is decided by the municipality and often drafted with initiative by land owners and residents. It includes the vision statement and the detailed district improvement plan.

8.3 Master planning and urban design

In urban planning, a master plan is created for designing the future layout of a city to guide future growth and development. A master plan provides analysis and proposals for land use, housing, and transportation, based on public input (citizen participation) and surveys. Detailed plans including the District Plan are made by breaking down a master plan.

In Japan, Basic Policy Concerning Municipal City Planning was stipulated by amendment of the City Planning Act in 1992 [55], and it emphasizes the significance of residents' participation in the process of making a master plan. Cities in Japan have created master

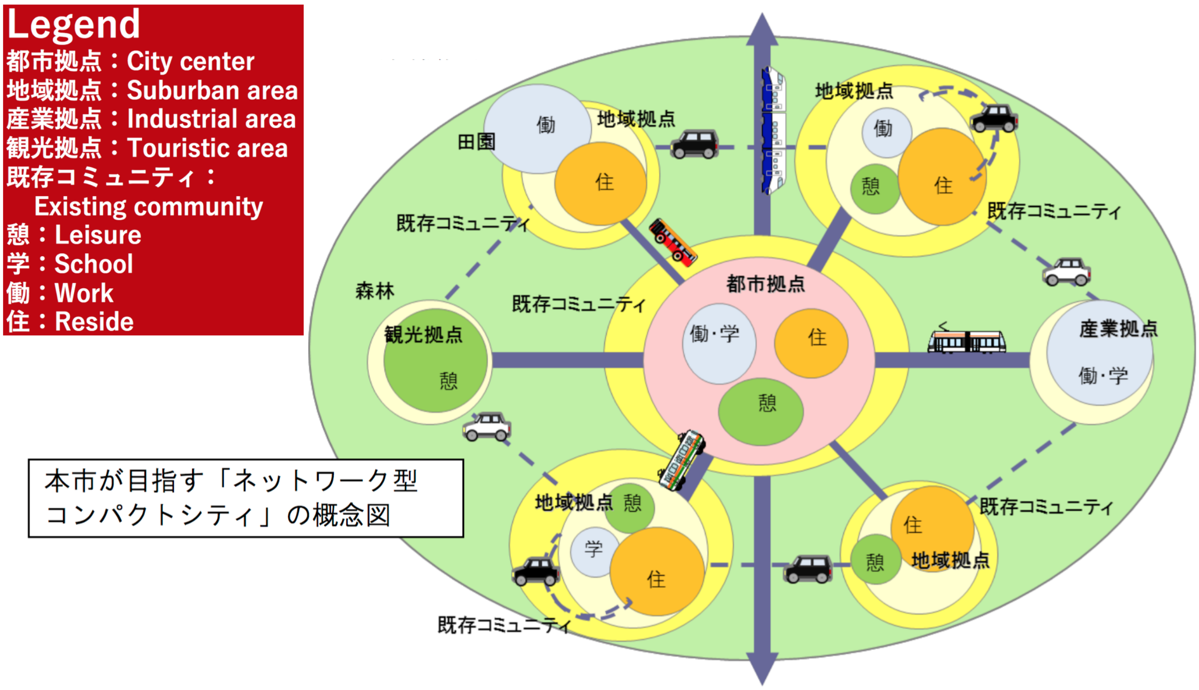


Fig. 8.1 The master plan of Utsunomiya City: Compact city forming a network [114].

plans, and they are continuously modified to cope with dynamic urban changes. For example, Toshima City has formed the Urban Planning Master Plan Formulation Exploratory Committee that consists of academics, residents and government officials, aiming at modifying the existing master plan. Documents including the minutes of the committee are open to the public and can be downloaded from the website of Toshima City [110].

Fig. 8.1 shows one of the most important measures proposed in the master plan of Utsunomiya City, aiming at developing a compact city that forms a network [114]. Green circles, orange circles, and gray circles respectively indicate functions of leisure, residence, and work/school. The plan is intended to agglomerate the functions and populations of each area of the city, and connect them by transportation.

8.4 Application of the proposed methods to urban planning

We consider that the methods proposed in this dissertation are useful for the design of a compact city considering the network of areas of a city. The proposed methods can be used for the analysis of the functional roles of each area, the evaluation of the outcome of urban development, and the analysis of fuctional clusters and their relations.

8.4.1 Understanding urban activities and detecting urban changes

The method proposed in Chapter 5 can be used for capturing the funtions of each area. It is necessary to understand the functions of each area in order to decide which function should be agglomerated for each area. In addition, the agglomeration of functions increases the attractiveness of an area and the number of visitors the specific activities. Therefore, the method can also be used for evaluating the outcome of agglomeration. The total number of passengers' arrivals was discussed by the Urban Planning Master Plan Formulation Exploratory Committee of Toshima City [110]. It does not provide information about the trip purposes, and it is impossible to evaluate the attractiveness and effectiveness of each area of the city. It is possible to discuss the attractiveness and effectiveness of each area of a city, using the method proposed in Chapter 5.

8.4.2 Predicting future residential mobility

The results of the method proposed in Chapter 5 provide information about which areas of a city are residential areas. It is necessary to enrich the amenities in residential neighborhoods and to agglomerate population to develop a compact city. Access to many types of amenities in walkable distance is the most important in developing residential areas. Such development increases the population of residents. The method proposed in Chapter 6 can be used for evaluating the effectiveness of development for increasing the quality of amenities and convenience in residential areas. It is important for reconsidering whether an existing plan is effecitve for developing residential areas.

8.4.3 Extraction of the community structure of a city

We have confirmed that a bipartite clustering method is effective to extract the community structure of a city, considering the functional relations between areas, in Chapter 7. It is useful for analyzing functional clusters of areas and their relations. It is effective to connect areas that play mutually complementary functions, by developing new public transportation. It is possible to make effective plans of public transportation that connect areas with different functions by using the results of the methods proposed in Chapters 5 and 6. Moreover, the network clustering method examined in Chapter 7 provides information about how much the new connection between two areas satisfies people's need for participating in activities.

Chapter 9

Conclusion

We have proposed methods to analyze urban systems based on human flows between areas of a city, and we have tested the methods using the smart card data of public transportation.

First, we proposed a method to estimate urban activities of each area of a city by decomposing the temporal distributions of visitors' arrivals. Then, we proposed a method to detect changes of urban activities in each area. The results have confirmed that the method is able to capture the changes of activities caused by events such as the openings of commercial facilities. In addition, the results show not only sudden changes of urban activities but also gradual changes. The method requires only human mobility data to infer the current state of each area.

Second, we have proposed a method to predict future residential mobility. The opportunity cost of travel time is estimated based on human mobility data on non-working days. We assume that the opportunity cost of travel time reflects the amenities and convenience of a residential neighborhood. Therefore, we expected that it would increase the number of movings-in to the place. The result shows that the estimated opportunity cost of travel time correlates to the number of people who relocate to the place. Then, we analyzed the causality between the opportunity cost of travel time, current population, and future residential mobility. The results show that the causal effect from current population to future residential mobility is mediated through the opportunity cost of travel time. We consider that our method is useful for quickly predicting the changes of future residential mobility caused by urban development because it only requires human mobility data from a short period.

Third, we have investigated which network clustering method is the most effective to discover the community structure of a city. We have proposed three evaluation methods to evaluate the results of community detection: (1) The geographical cohesiveness of extracted clusters; (2) How a method captures the regularity of human mobility on working days;

(3) The amount of information of extracted communities. We have shown that a bipartite network clustering method considering multi-facet characteristics of each cluster scores the best for these metrics. The clustering method has never been used for the analysis of urban structures based on human mobility data. Therefore, it is expected that the method will help better understand the urban strucures based on human mobility data.

Cities are faced with both rapid and gradual changes. Urban planners need to know the current and future state of areas and boundaries of subsytems. The significance of our research is that it enables us to find problems of cities and to predict the future outcome of urban development. Urban planning requires continuous monitoring of cities and modifying plans. We hope that the proposed methods will contribute to the improvement of urban planning.

Finally, the sharing of human mobility data among different stakeholders is usually difficult, because they are often highly confidential. The proposed methods would be more effective when they are applied to larger amounts of data. We have designed the methods to contribute to the public good. We hope that these methods provide opportunities to various stakeholders to collaborate with each other to solve social problems by integrating various types of urban data.

References

[1] Alexander, L., Jiang, S., Murga, M., and González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58:240 – 250. Big Data in Transportation and Traffic Engineering.

[2] Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., and Hickman, M. (2018). Public transport trip purpose inference using smart card fare data. *Transportation Research Part C: Emerging Technologies*, 87:123 – 137.

[3] Amini, A., Kung, K., Kang, C., Sobolevsky, S., and Ratti, C. (2014). The impact of social segregation on human mobility in developing and industrialized regions. *EPJ Data Science*, 3(1):6.

[4] Axhausen, K. W., Zimmermann, A., Schönfelder, S., Rindsfüser, G., and Haupt, T. (2002). Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2):95–124.

[5] Bagley, M. N. and Mokhtarian, P. L. (2002). The impact of residential neighborhood type on travel behavior: A structural equations modeling approach. *The Annals of Regional Science*, 36(2):279–297.

[6] Bagrow, J. P. and Lin, Y.-R. (2012). Mesoscopic structure and social aspects of human mobility. *PLOS ONE*, 7(5):1–6.

[7] Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Phys. Rev. E*, 76:066102.

[8] Batty, M. (2007). *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals*. The MIT Press.

[9] Batty, M. (2013). *The new science of cities*. Mit Press.

[10] Batty, M., Xie, Y., and Sun, Z. (1999). Modeling urban dynamics through gis-based cellular automata. *Computers, Environment and Urban Systems*, 23(3):205 – 233.

[11] Becker, G. S. (1965). A theory of the allocation of time. *The economic journal*, pages 493–517.

[12] Beckett, S. J. (2016). Improved community detection in weighted bipartite networks. *Open Science*, 3(1).

[13] Ben-Akiva, M. and Bierlaire, M. (1999). *Discrete Choice Methods and their Applications to Short Term Travel Decisions*, pages 5–33. Springer US, Boston, MA.

[14] Ben-Akiva, M. and Bowman, J. L. (1998). Integration of an activity-based model system and a residential location model. *Urban Studies*, 35(7):1131–1153.

[15] Bhat, C. R. and Guo, J. Y. (2007). A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B: Methodological*, 41(5):506 – 526.

[16] Blondel, V., Krings, G., and Thomas, I. (2010). Regions and borders of mobile telephony in belgium and in the brussels metropolitan zone. *Brussels Studies.*

[17] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10).

[18] Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. *Information Technology for Development*, 18(2):107–125.

[19] Bockstael, N. E., Strand, I. E., and Hanemann, W. M. (1987). Time and the recreational demand model. *American Journal of Agricultural Economics*, 69(2):293–302.

[20] Bohte, W. and Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: A large-scale application in the netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3):285 – 297.

[21] Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439.

[22] Calabrese, F., Dahlem, D., Gerber, A., Paul, D., Chen, X., Rowland, J., Rath, C., and Ratti, C. (2011). The connected states of america: Quantifying social radii of influence. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 223–230.

[23] Cao, X. J., Mokhtarian, P. L., and Handy, S. L. (2009). The relationship between the built environment and nonwork travel: A case study of northern california. *Transportation Research Part A: Policy and Practice*, 43(5):548 – 559.

[24] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

[25] Cervero, R. and Kockelman, K. (1997). Travel demand and the 3ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, 2(3):199 – 219.

[26] Chen, Y., Liu, X., Li, X., Liu, X., Yao, Y., Hu, G., Xu, X., and Pei, F. (2017). Delineating urban functional areas with building-level social media data: A dynamic time warping (dtw) distance based k-medoids method. *Landscape and Urban Planning*, 160:48 – 60.

[27] Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1082–1090, New York, NY, USA. ACM.

[28] Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S.-i. (2009). *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, New Jersey.

[29] Clarke, K. C. and Gaydos, L. J. (1998). Loose-coupling a cellular automaton model and gis: long-term urban growth prediction for san francisco and washington/baltimore. *International Journal of Geographical Information Science*, 12(7):699–714.

[30] Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70:066111.

[31] Coscia, M., Rinzivillo, S., Giannotti, F., and Pedreschi, D. (2012). Optimal spatial resolution for the analysis of human mobility. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 248–252.

[32] Dieleman, F. M. (2001). Modelling residential mobility; a review of recent trends in research. *Journal of Housing and the Built Environment*, 16(3):249–265.

[33] Dormann, C. F. and Strauss, R. (2014). A method for detecting modules in quantitative bipartite networks. *Methods in Ecology and Evolution*, 5(1):90–98.

[34] Eagle, N., Macy, M., and Claxton, R. (2010). Network diversity and economic development. *Science*, 328(5981):1029–1031.

[35] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231.

[36] Ewing, R. and Cervero, R. (2010). Travel and the built environment. *Journal of the American Planning Association*, 76(3):265–294.

[37] Fan, Z., Song, X., and Shibasaki, R. (2014). Cityspectrum: A non-negative tensor factorization approach. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pages 213–223, New York, NY, USA. ACM.

[38] Févotte, C., Bertin, N., and Durrieu, J.-L. (2009). Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830. PMID: 18785855.

[39] Frias-Martinez, V. and Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using twitter activity. *Engineering Applications of Artificial Intelligence*, 35:237 – 245.

[40] Gabaix, X. (1999). Zipf's law for cities: An explanation. *The Quarterly Journal of Economics*, 114(3):739–767.

[41] Gilderbloom, J. I., Riggs, W. W., and Meares, W. L. (2015). Does walkability matter? an examination of walkability's impact on housing values, foreclosures and crime. *Cities*, 42:13–24.

[42] González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–82.

[43] Guimerà, R., Mossa, S., Turtschi, A., and Amaral, L. A. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799.

[44] Guimerà, R., Sales-Pardo, M., and Amaral, L. A. N. (2007). Module identification in bipartite and directed networks. *Phys. Rev. E*, 76:036102.

[45] Haase, D., Lautenbach, S., and Seppelt, R. (2010a). Modeling and simulating residential mobility in a shrinking city using an agent-based approach. *Environmental Modelling & Software*, 25(10):1225–1240.

[46] Haase, D., Lautenbach, S., and Seppelt, R. (2010b). Modeling and simulating residential mobility in a shrinking city using an agent-based approach. *Environmental Modelling & Software*, 25(10):1225–1240.

[47] Han, G. and Sohn, K. (2016). Activity imputation for trip-chains elicited from smart-card data using a continuous hidden markov model. *Transportation Research Part B: Methodological*, 83:121 – 135.

[48] Handy, S., Cao, X., and Mokhtarian, P. (2005). Correlation or causality between the built environment and travel behavior? evidence from northern california. *Transportation Research Part D: Transport and Environment*, 10(6):427 – 444.

[49] Handy, S. L. and Clifton, K. J. (2001). Local shopping as a strategy for reducing automobile travel. *Transportation*, 28(4):317–346.

[50] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81 – 102.

[51] Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*.

[52] Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

[53] Hu, N., Legara, E. F., Lee, K. K., Hung, G. G., and Monterola, C. (2016). Impacts of land use and amenities on public transport use, urban planning and design. *Land Use Policy*, 57:356 – 367.

[54] Jacobs, J. (1961). *The death and life of great American cities*. Random House.

[55] Japan International Cooperation Agency (2007). Urban land use planning system in japan.
https://jica-net-library.jica.go.jp/library/jn325/UrbanLandUsePlanningSystem_all.pdf (accessed on 16 September 2018).

[56] Jim, C. and Chen, W. Y. (2010). External effects of neighbourhood parks and landscape elements on high-rise residential value. *Land Use Policy*, 27(2):662 – 670. Forest transitions Wind power planning, landscapes and publics.

[57] Jones, P. M. (1977). *New approaches to understanding travel behavior: the human activity approach*. Oxford University, Oxford.

[58] Kan, K. (2007). Residential mobility and social capital. *Journal of Urban Economics*, 61(3):436 – 457.

[59] Kawahara, Y. and Sugiyama, M. (2012). Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127.

[60] Kitamura, R., Mokhtarian, P. L., and Laidet, L. (1997). A micro-analysis of land use and travel in five neighborhoods in the san francisco bay area. *Transportation*, 24(2):125–158.

[61] Krizek, K. (2006). Lifestyles, residential location decisions, and pedestrian and transit activity. *Transportation Research Record: Journal of the Transportation Research Board*, (1981):171–178.

[62] Krizek, K. J. (2003). Residential relocation and changes in urban travel: Does neighborhood-scale urban form matter? *Journal of the American Planning Association*, 69(3):265–281.

[63] Kuang, C. (2017). Does quality matter in local consumption amenities? an empirical investigation with yelp. *Journal of Urban Economics*, 100:1 – 18.

[64] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

[65] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

[66] Lee, B. H. Y. and Waddell, P. (2010). Residential mobility and location choice: a nested logit model with sampling of alternatives. *Transportation*, 37(4):587–601.

[67] Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press.

[68] Lee, S. G. and Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport*, 6(1):1–20.

[69] Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

[70] Liu, X., Gong, L., Gong, Y., and Liu, Y. (2015). Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography*, 43:78 – 90.

[71] Liu, X., Kang, C., Gong, L., and Liu, Y. (2016). Incorporating spatial interaction patterns in classifying and understanding urban land use. *International Journal of Geographical Information Science*, 30(2):334–350.

[72] Long, Y. and Thill, J.-C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in beijing. *Computers, Environment and Urban Systems*, 53:19 – 35. Special Issue on Volunteered Geographic Information.

[73] Lynch, A. K. and Rasmussen, D. W. (2001). Measuring the impact of crime on house prices. *Applied Economics*, 33(15):1981–1989.

[74] Lyons, G., Jain, J., and Holley, D. (2007). The use of travel time by rail passengers in great britain. *Transportation Research Part A: Policy and Practice*, 41(1):107 – 120.

[75] Lyons, G. and Urry, J. (2005). Travel time use in the information age. *Transportation Research Part A: Policy and Practice*, 39(2):257 – 276. Positive Utility of Travel.

[76] Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

[77] McConnell, K. E. (1985). Chapter 15 the economics of outdoor recreation. In *Handbook of Natural Resource and Energy Economics*, volume 2 of *Handbook of Natural Resource and Energy Economics*, pages 677 – 722. Elsevier.

[78] Mellander, C., Lobo, J., Stolarick, K., and Matheson, Z. (2015). Night-time light data: A good proxy measure for economic activity? *PLOS ONE*, 10(10):1–18.

[79] Ministry of Land, Infrastructure, T. and Tourism (2018). Compact plus network. http://www.mlit.go.jp/toshi/toshi_ccpn_000016.html (accessed on 16 September 2018).

[80] Mokhtarian, P. L. and Cao, X. (2008). Examining the impacts of residential self-selection on travel behavior: A focus on methodologies. *Transportation Research Part B: Methodological*, 42(3):204 – 228. A Tribute to the Career of Frank Koppelman.

[81] Moré, J. J. (1978). The levenberg-marquardt algorithm: Implementation and theory. In Watson, G. A., editor, *Numerical Analysis*, pages 105–116, Berlin, Heidelberg. Springer Berlin Heidelberg.

[82] Neal, R. M. (2010). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162.

[83] Neirotti, P., Marco, A. D., Cagliano, A. C., Mangano, G., and Scorrano, F. (2014). Current trends in smart city initiatives: Some stylised facts. *Cities*, 38:25 – 36.

[84] Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133.

[85] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

[86] Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113.

[87] Nishi, K., Tsubouchi, K., and Shimosaka, M. (2014). Extracting land-use patterns using location data from smartphones. In *Proceedings of the First International Conference on IoT in Urban Space*, URB-IOT '14, pages 38 – 43.

[88] Noulas, A., Scellato, S., Lathia, N., and Mascolo, C. (2012). Mining user mobility features for next place prediction in location-based services. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 1038–1043, Washington, DC, USA. IEEE Computer Society.

[89] Ohmori, N. and Harata, N. (2008). How different are activities while commuting by train? a case in tokyo. *Tijdschrift voor economische en sociale geografie*, 99(5):547–561.

[90] Parker, D. C., Manson, S. M., Janssen, M. A., Hoffmann, M. J., and Deadman, P. (2003). Multi-agent systems for the simulation of land-use and land-cover change: A review. *Annals of the Association of American Geographers*, 93(2):314–337.

[91] Pope, D. G. and Pope, J. C. (2015). When walmart comes to town: Always low housing prices? always? *Journal of Urban Economics*, 87:1 – 13.

[92] Quigley, J. M. and Weinberg, D. H. (1977). Intra- urban residential mobility: A review and synthesis. *International Regional Science Review*, 2(1):41–66.

[93] Rajamani, J., Bhat, C., Handy, S., Knaap, G., and Song, Y. (2003). Assessing impact of urban form measures on nonwork trip mode choice after controlling for demographic and level-of-service effects. *Transportation Research Record: Journal of the Transportation Research Board*, (1831):158–165.

[94] Rappaport, J. (2008). Consumption amenities and city population density. *Regional Science and Urban Economics*, 38(6):533 – 552.

[95] Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R., and Strogatz, S. H. (2010). Redrawing the map of great britain from a network of human interactions. *PLOS ONE*, 5(12):1–6.

[96] Rinzivillo, S., Mainardi, S., Pezzoni, F., Coscia, M., Pedreschi, D., and Giannotti, F. (2012). Discovering the geographical borders of human mobility. *KI - Künstliche Intelligenz*, 26(3):253–260.

[97] Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123.

[98] Roth, C., Kang, S. M., Batty, M., and Barthélemy, M. (2011). Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLOS ONE*, 6(1):1–8.

[99] Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes Twitter users. In *Proceedings of the 19th international conference on World wide web - WWW '10*, page 851, New York, New York, USA. ACM Press.

[100] Sawada, H., Kameoka, H., Araki, S., and Ueda, N. (2013). Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):971–982.

[101] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

[102] Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030.

[103] Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J. A., Pitchford, J. W., James, A., Ahmed, M. Z., Brierley, A. S., Hindell, M. A., Morritt, D., Musyl, M. K., Righton, D., Shepard, E. L. C., Wearmouth, V. J., Wilson, R. P., Witt, M. J., and Metcalfe, J. D. (2008). Scaling laws of marine predator search behaviour. *Nature*, 451.

[104] Smith, C., Quercia, D., and Capra, L. (2013). Finger on the pulse: Identifying deprivation using transit flow analysis. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 683–692, New York, NY, USA. ACM.

[105] Sobolevsky, S., Szell, M., Campari, R., Couronné, T., Smoreda, Z., and Ratti, C. (2013). Delineating geographical regions with networks of human interactions in an extensive set of countries. *PLOS ONE*, 8(12):1–10.

[106] Stouffer, S. A. (1940). Intervening opportunities: a theory relating mobility and distance. *American sociological review*, 5(6):845–867.

[107] Suzuki, K. and Wakita, K. (2009). Extracting multi-facet community structure from bipartite networks. In *2009 International Conference on Computational Science and Engineering*, volume 4, pages 312–319.

[108] Takeuchi, J. and Yamanishi, K. (2006). A unifying framework for detecting outliers and change points from time series. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):482–492.

[109] The Ministry of Land, Infrastructure, Transport and Tourism in Japan (2010). Person trip data in the western area in japan (written in japanese). https://www.kkr.mlit.go.jp/plan/pt/index.html (accessed on 12 September 2018).

[110] Toshima City (2018). Toshima city in japan forms urban planning master plan formulation exploratory committee. http://www.city.toshima.lg.jp/295/kuse/shingi/kaigichiran/033814/033813.html (accessed on 16 September 2018).

[111] Trice, A. H. and Wood, S. E. (1958). Measurement of recreation benefits. *Land economics*, 34(3):195–207.

[112] United Nations Global Pulse (2012). *Big Data for Development: Challenges and Opportunities*.

[113] United Nations Global Pulse (2013). *Mobile Phone Network Data for Development*.

[114] Utsunomiya City (2014). Vision for developing a smart city forming a network. http://www.city.utsunomiya.tochigi.jp/_res/projects/default_project/_page_/001/006/078/140320sankoushiryou.pdf (accessed on 16 September 2018).

[115] Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854.

[116] Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., and Stanley, H. E. (1996). Lévy flight search patterns of wandering albatrosses. *Nature*, 381.

[117] Von Bertalanffy, L. (1968). *General system theory*. Braziller, New York, NY, USA.

[118] WalkScore (2007). Walk score. https://www.walkscore.com (accessed on 16 September 2018).

[119] Walsh, R. G., Sanders, L. D., and Mckean, J. R. (1990). The consumptive value of travel time on recreation trips. *Journal of Travel Research*, 29(1):17–24.

[120] Wang, J., Gao, F., Cui, P., Li, C., and Xiong, Z. (2014). Discovering urban spatio-temporal structure from time-evolving traffic networks. In Chen, L., Jia, Y., Sellis, T., and Liu, G., editors, *Web Technologies and Applications*, pages 93–104, Cham. Springer International Publishing.

[121] Wang, J., Kong, X., Rahim, A., Xia, F., Tolba, A., and Al-Makhadmeh, Z. (2017a). Is2fun: Identification of subway station functions using massive urban data. *IEEE Access*, 5:27103–27113.

[122] Wang, P., Fu, Y., Liu, G., Hu, W., and Aggarwal, C. (2017b). Human mobility synchronization and trip purpose detection with mixture of hawkes processes. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 495–503, New York, NY, USA. ACM.

[123] Weiner, N. (1948). *Cybernetics*. John Wiley, New York, NY, USA.

[124] White, R. and Engelen, G. (1993). Cellular automata and fractal urban form: A cellular modelling approach to the evolution of urban land-use patterns. *Environment and Planning A: Economy and Space*, 25(8):1175–1199.

[125] Wilson, A. (1967). A statistical theory of spatial distribution models. *Transportation Research*, 1(3):253 – 269.

[126] Wolfram, S. (1983). Statistical mechanics of cellular automata. *Rev. Mod. Phys.*, 55:601–644.

[127] World Bank (2014). *Big Data in Action for Development*.

[128] Xiao, G., Juan, Z., and Zhang, C. (2016). Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. *Transportation Research Part C: Emerging Technologies*, 71:447 – 463.

[129] Yabe, T., Tsubouchi, K., and Sekimoto, Y. (2017). Cityflowfragility: Measuring the fragility of people flow in cities to disasters using gps data collected from smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):117:1–117:17.

[130] Yao, D., Yu, C., Jin, H., and Ding, Q. (2015). Human mobility synthesis using matrix and tensor factorizations. *Information Fusion*, 23:25 – 32.

[131] Zhang, M. (2005). Exploring the relationship between urban form and nonwork travel through time use analysis. *Landscape and Urban Planning*, 73(2):244 – 261. Research on the Built and Virtual Environments.

[132] Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014a). Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol.*, 5(3):38:1–38:55.

[133] Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014b). Urban computing: Concepts, methodologies, and applications. *ACM Trans. Intell. Syst. Technol.*, 5(3):38:1–38:55.

[134] Zhong, C., Arisona, S. M., Huang, X., Batty, M., and Schmitt, G. (2014a). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, 28(11):2178–2199.

[135] Zhong, C., Huang, X., Arisona, S. M., Schmitt, G., and Batty, M. (2014b). Inferring building functions from a probabilistic model using public transportation data. *Computers, Environment and Urban Systems*, 48:124 – 137.

[136] Zhong, C., Manley, E., Arisona, S. M., Batty, M., and Schmitt, G. (2015). Measuring variability of mobility patterns from multiday smart-card data. *Journal of Computational Science*, 9:125 – 130.

[137] Zhong, C., Schläpfer, M., Arisona, S. M., Batty, M., Ratti, C., and Schmitt, G. (2017). Revealing centrality in the spatial structure of cities from human activity patterns. *Urban Studies*, 54(2):437–455.

[138] Zhou, Y., Fang, Z., Zhan, Q., Huang, Y., and Fu, X. (2017). Inferring social functions available in the metro station area from passengers' staying activities in smart card data. *ISPRS International Journal of Geo-Information*, 6(12):394.

[139] Zipf, G. K. (1946). The p1p2/d hypothesis: On the intercity movement of persons. *American Sociological Review*.

Acknowledgements

This dissertation is the compilation of several results from the analysis of the smart card data of public transportation, most of which were obtained in the second year and the third year of the doctoral program. This work would have been impossible without the support from a number of people.

I would like to thank Professor Kiyoshi Izumi for his generous support and guidance in writing this dissertation. I am also grateful to Professor Junichiro Mori for his support for each study and analysis in this dissertation. He helped me when I was in difficult times, and I received a great deal of guidance to achieve the goals of my research. I would like to express my deep gratitude to Professor Ichiro Sakata for his continuous encouragement and many valuable suggestions for my research. I would also like to sincerely express my gratitude to Professor Ryuichi Shibasaki and Dr. Hideki Fujii for carefully reading this dissertation and providing me with constructive suggestions.

I would like to extend my thanks to my research collaborators. I would like to thank Dr. Narushige Shiode, who supported my research with patience and tolerance. He guided me not only while I was studying at King's College London but also after I came back to Japan. I would also like to thank Dr. Chen Zhong for deepening our study. I am greatful to Mr. Tetsuo Sakimoto for providing me the opportunity to analyze the smart card data and for providing me with valuable advice. I am also grateful to Dr. Masanao Ochi and Ms. Izumi Hayashi for having discussion with me to deepen my research ideas.

I received tremendous support from the Leading Graduates Schools Program "Global Leader Program for Social Design and Management (GSDM)" run by the Ministry of Education, Culture, Sports, Science and Technology, Japan. My graduate study would have been impossible without this support.

I owe many thanks to my father Shiro, my mother Margaret, and my sister Catherine for their generosity and encouragement. My mother Margaret helped me improve the English of this dissertation. Finally and most importantly, I would like to express my deepest gratitude to my beloved wife Yuka, who is my constant source of strength and courage. I would not have had determination to pursue PhD without her encouragement.