Estimation of Tourist Travel Patterns with Recursive Logit Models based on Wi-Fi data with Kyoto City Case Study

GAO YUHAN

Abstract

Estimating traffic flows is the foundation of solving related problems in transportation. However, there are limited methods to do so economically and without invading privacy. In particular this is the case for investigating the behavior of specific population groups other than residents or goods. This research establishes a methodology that allows understanding and predicting tourist flows by relatively low-cost Wi-Fi sensing technology. These sensors are designed to detect and record all electronic devices with Wi-Fi function enabled within an average radius of around 40 meters. All detected devices are identified by an anonymized label encrypted from the device's MAC address. In addition to the hashed MAC address, Wi-Fi packet sensors also record timestamp, packet sensor ID, Received Signal Strength Indication (RSSI), and the device's vender ID. Multiple detections and the above attributes allow the identification of routes of any individual who carries a detectable device to be observed.

Given the importance of tourism in Kyoto as well as the relatively sparse literature on estimating tourism travel patterns within city, Wi-Fi data are used to understand these. Based on two data surveys conducted in Kyoto City, Recursive logit (RL) models are employed to describe their temporal and spatial aspects of travel.

Data of the first experiment was collected by twenty sensors distributed in the street blocks of Higashiyama Ward, Kyoto, Japan, which is one of the areas of Kyoto most frequented by tourists. Based on the data, we first discuss the extraction of tourist trip chains and the construction of a simplified time-space network with "stay" and "move" links. The network and link attributes are based on the location of the Wi-Fi sensors as well as map data. RL models are then employed to formulate tourists' route choice behavior. The existence of different kinds of "points of interests" (POIs) is shown to explain route choice as well as whether a person stays on a link for an extended period of time or passes through it in a short time.

For the second case study, data was collected by 39 sensors at key tourist points within Kyoto City. Because of the low resolution caused by sensor limitation, we first compare the quality of the obtained

Wi-Fi packet data with commercially available GPS footprints. Then a data fusion approach is proposed to enrich the Wi-Fi sample by GPS observations since either of them, by themselves, were found to be insufficient to interpret travel patterns at individual level. Besides the RL with Wi-Fi data modelling methodology and the specific case study findings, a third contribution of this dissertation is hence the data fusion of these two data sources. The fused data are referred to as "enriched Wi-Fi traces".

Based on these enriched Wi-Fi traces alternative RL models are established. In this case, time-dependent values measuring attractiveness of an area are defined. Further "Group to Outside-links", which are defined as cases when a person appears to visit locations different from those captured by the sensors, are defined. Their attractiveness is measured with a set of constants capturing these "errors". The results illustrate that temporal and spatial features of tourists' travel patterns over the city fairly well. The concluding discussion contains a general assessment as to the benefits of Wi-Fi data and, more specifically, recommendations as to when the usage of time-expanded networks versus simpler networks with "stay" and "move" links is beneficial in the modelling of the data since both structures have their benefits and shortcomings.

Keywords: Wi-Fi Packet Sensing, Recursive Logit, City Tourism, Route Choice, Activity Duration, Data Fusion

Acknowledgements

I feel very fortunate to have spent five years as a student in my current lab. I am truly grateful to everyone I have met in Kyoto University and for all the help you have provided. Without meeting you I would not be who I am.

I have my deepest gratitude for Dr. Schmöcker. You are my supervisor, my mentor and my friend. You encourage me to not to give up during years of my master's degree. Without you I would never have felt the joy of research and handled the challenges during my PhD. I am willing to share all the little things in my life with you, and you never tired of encouraging me, comforting me and supporting me. Wish our friendship will last forever.

My sincere gratitude to Prof. Yamada, for his support, encouragement, and trust on me. Thanks to Prof. Yamada and Prof. Fujii, and Dr. Schmöcker who are the committee member of my PhD dissertation and give me valuable comments.

Thanks to the current faculty of Intelligent Transportation Systems Laboratory, Prof. Yamada and Dr. Schmöcker, Dr. Nakao and lab secretary, Ms. Nishikawa, for the many help I received in my studies and life. Thanks to previous lab faculty, Prof. Uno in Kyoto University, Dr. Nakamura currently in Nagoya University. Also thanks to the previous lab secretaries, Ms. Shii, Ms. Nishimura and Ms. Nagata.

Thanks to the past and present students in our lab. My sincere appreciation to Mr. Kawakami, Mr. Nishigaki, Mr. Niwa and Mr. Sonobe. Meeting you guys is a great fortune in my life. When I first arrived in Japan, you gave me, a strange foreigner, unconditional help to settle into the lab. Thanks to Dr. Nakamura who gave me a precious opportunity to intern at IBS. Thanks to Dr. Zhang, Dr. Sun, and all the friends, seniors, juniors I met in Japan. I would never have chosen to live in Japan without you and your help.

Finally, I would like to thank my mother, Han Lu, my father, Gao Xiang, my grandparents, Han Changshi and Shi Zhaojing. I pride myself on being a very optimistic person, but I can't contain my sadness when I think that my grandfather, who has done everything for me, won't be able to see me graduate. This dissertation is dedicated to my family.

Preface

Parts of this dissertation have been presented in conferences or else submitted for review are as follows.

Conference presentations

I. **Gao, Y.** and Schmöcker, J.-D. (2019). Tourist Route Choices and Short-Term Flow Predictions in Tourist Areas Based on Wi-fi Packet Data. Presented at International Choice Modelling Conference (ICMC2019), Kobe, Japan, Aug 19-21. **(Chapter 4)**

II. **Gao, Y.** and Schmöcker, J.-D. (2019). Tourist route choices and short-term flow predictions in tourist areas based on Wi-Fi packet data. 60th Japan Infrastructure Planning Conference (Autumn Meeting), Toyama, Japan. Nov. 30 – Dec. 2. **(Chapter 4)**

III. Shen, K., Schmöcker, J.-D., **Gao, Y.** and Qureshi, A. (2019). Estimation of city tourism tours with survey data from Kyoto. 9th Int. Symposium on Travel Demand Management, Edinburgh, U.K., June 19-21. Also presented in similar form at the 60th Japan Infrastructure Planning Conference (Autumn Meeting), Toyama, Japan. Nov. 30 – Dec. 2. **(Chapter 3)**

IV. Wachtel, G., Schmöcker, J.-D., **Gao, Y.**, Nahum, O.E. and Hadas, Y. (2020). Planning for City Tourist Evacuation Routes: Collecting and Providing Information. 99th Annual Meeting of the Transportation Research Board. Washington D.C., U.S. **(Chapter 3)**

V. Hadas, Y., Ben-Moshe, B., Wachtel, G., Nahum, O., Schmöcker, J.-D., Sabashi, K. and **Gao, Y.** (2020). Assessing the navigation error characteristics of residents and tourists during evacuation – a combined simulation and virtual reality approach. 8th International Conference on Transport Network Reliability (Stockholm, 24-26 June). Postponed **(Chapter 3)**

Under review/ To be submitted

I. **Gao, Y.** and Schmöcker, J.-D. (2020). Estimation of Walking Patterns in a Touristic Area with Wi-Fi Packet Sensors. *Transportation research part C: Emerging Technologies* (under review). **(Chapter 4)**

II. **Gao, Y.** and Schmöcker, J.-D. (2020) Route choices and reliability of short-term flow predictions based on Bluetooth and Wi-fi sensors. 12th International Scientific Conference on Transport and Mobility (under review) **(Chapter 4)**

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background and research motivation

1.1.1 Investigating special groups in transportation

Understanding traffic flows is the foundation of solving related problems such as estimating spatial/temporal attraction, evaluation of transportation services and evacuation planning. Traffic flows have been investigated and analyzed based on large-scale questionnaire surveys such as the Person Trip (PT) surveys in Japan, a government-led survey that focuses on the movement of city-residents. However, facing the fast-growing urban traffic, the limitations of this traditional survey approach are becoming increasingly apparent.

One of the most essential limitations is the expensive cost. Budget constraints create limitations on understanding flow details. In Japan for example, PT surveys are conducted only once a decade and can only target the trips of respondents during one day. In order to capture the most general travel behavior by limited cost, the target day must be as "common" as possible. With this principle, flow details, and seasonal variations are completely sacrificed. A timely example is that the PT survey planned to be conducted in 2020 of the Kyoto-Osaka-Kobe metropolitans is postponed due to the Covid-19 crisis "significantly impacting the travel behavior" (MLIT, 2020). Facing the trade-off between obtaining data for long-term steady forecast and a current global incident, this postponement is representative for exposing shortcomings of conventional surveys.

Another drawback is the low accuracy and low respond rates of questionnaire surveys caused by the long answer steps. Although with the improvement of the information technologies there have been many efforts on easing answering burden such as applying online survey instead of posting, the response rate of the latest PT survey in the metropolitan areas of Japan is still less than 25% (MLIT of Japan, 2019). A low response rate impacts the quality of a survey, as the non-response ratio may be unequal among different population groups, such as age groups or income levels leading to biases. Such a low response rate further then indirectly affects again the cost of a questionnaire survey, since obtaining enough samples with a low response rate requires more labor to deliver and collect questionnaires more extensively and to stimulate respondents.

Furthermore, such infrequent surveys can clearly not capture the long- and short-term dynamics. In terms of dynamic traffic situation changes, it is difficult to quantify in a questionnaire either how a single episode of congestion affects traffic behavior, or how continued congestion gradually changes people's travel patterns.

Regardless of the above-mentioned shortcomings of questionnaires, questionnaire surveys are still the primary for researchers to investigate traffic flows on a large scale. For decades, even in recent years, it has maintained excellent reliability in urban transportation analysis and forecasting. However, besides residents there are specific population groups that behave differently, such as tourists and event participants. Although these groups represent a small percentage of the total, they sometimes behave differently than residents, and their particular behavior are likely to impact specific routes of urban transportation differently. They are surely necessary to be investigated, while questionnaires capture their behavior only to a limited degree. They are difficult to survey, or to be captured in general surveys.

Targeting such minor groups in transportation, the shortcomings of routine surveys are further magnified: they are generally less motivated to answer a questionnaire. More critically, the travel patterns of these populations are more influenced by seasonal changes, sudden congestions, and other dynamic conditions more than those of residents, which calls for more attention to their flow detail. However, understanding of behavioral response due to crowding at point-of-interests is limited with surveys since it is almost impossible to quantitative crowding in an interview. For these reasons, surveys specifically aiming at flows other than residents and goods flows are rare, for example the last tourist-trip survey in Kyoto City was conducted 14 years ago (Urban Planning Bureau of Kyoto City, 2006).

1.1.2 Electronic footprints

However, investigating specific minor groups like tourists and predict their demands becomes easier these days. For example, data for hotel booking can be straightforward knowledges for tourist trends, postings on social media and correspond social networks provides vast information for travel trends with socio-demography. Compared to daily transportation users like commuters, people with leisure

activities are more willing to review what they have visited. The informational society brings benefits for modelling tourist flows in multiple ways. First, information about the study subject is easier to collect, and second, the attractiveness of point-of-interest specific to the subject becomes easier to evaluate. With these advantages we are able to both precisely understand the choices tourists made, and enrich the characteristics of the alternative destinations.

The electronic footprints become new providers for obtaining tourism flow data thanks to the rapid development and spread of smartphones and wireless networks. The data commonly used comes from 1) Location information from smartphone applications, especially location-based services (LBS) and social networks, such as navigation and map service and electronic yellow pages, and 2) call detail record (CDR) from cellular networks provided by telecom companies. All of these data are passively emerging from users' daily activities, thus have superiority with respect to the richness of the sample size. In Japan, it is common for a popular application to have millions of users, not to mention the over 100 million cell phone penetration on cellular network. Also, with permission, these electronic footprints can be observed in detail over time and are useful for understanding detailed trips and long-term travel patterns. Further, these are highly time-sensitive, resulting in a naturally advantage for capturing flow dynamics and details.

In line with the above features, electronic footprints seem to fully compensate for the shortcomings of questionnaires, while in practice they are still subject to a number of limitations, both from the still high cost, from not yet established methodologies, and from increasingly stringent privacy regulations. In terms of cost, especially for CDR and LBS data, the cost is still high. Also, a study in use of electronic footprints usually becomes more or less data-driven because different electronic footprints have different features, there is a wide variety in terms of methodology. Moreover, in response to the serious threat to user privacy, the regulation for such research is getting stronger. For example, in Japan, CDR data on the market must be aggregated, and it is almost impossible to extrapolate anyone's daily trip from the data.

This research pays especially attention to data from Wi-Fi packet sensing, for its desirable cost-performance and superior privacy protection features. As a new technology not yet widely used, the

anonymized sensing data have advantages on 1) easier to obtain unified samples as obtaining information is not relying on any smartphone application; 2) ability to encrypt data locally for privacy protection and 3) lower cost compared to other electronic datasets. However, most of the studies on the sensing data are currently either at the aggregated level or small-scale experiments (e.g. Versichele et al. 2012; Fukuda et al. 2018; Ota et al. 2018).

In conclusion, this thesis presents a methodology that utilizing address-matched packet sensing technology to investigate special population groups in transportation in this study. Tourists are targeted as they are a representative type of special transportation user. Especially for cities like Kyoto, where tourism is a major industry, the tourist flow is no less important than the residential flow.

1.2 Research objectives, challenges

This research aims to achieve one major objectives: To establish and test a methodology that allows understanding and predicting tourist flows by relatively low-cost sensing technologies and respects user privacy.

More specifically, the challenges are:

- **Limited observable area:** Digital traces based on packet sensing are essentially sequences of passed nodes, since a sensor can only detect nearby devices. Routes taken between two observations are unknown. Although other electronic footprints such as GPS data and CDR data also have to some extent observation bias, challenges in terms of unobservable information in Wi-Fi sensing are even greater: depends on the sensor deployment, blank between each two records may be broad. Unlike GPS studies that usually try to fix the observation error at micro-level, an approach to address this issue by adapting the network to the distribution of sensors is explored.

- **Incomplete observation:** As a kind of microwave in 2.4GHz band, Wi-Fi signal are to some degree sensitive to the propagation environment. Mochizuki et al (2018), Ota et al (2018) confirm in experiments that Wi-Fi packet sensors have different (and sometimes considerably different) penetration rate compare to each other. This arises as an issue that a sequence

observed may randomly lost observations, and the probability of losing observations could be associated to sensor-specific penetration rates. A specific error term is considered necessary to correct the detection bias.

- **Uncertain user groups and behavior:** Compared to conventional travel surveys, passive collected digital footprints met difficulties in corresponding themselves to users' social-attributes, while modeling travelers require us to target at different user groups to avoid developing an over-homogeneous model that failed to capture features by travel purpose and social group characteristic. More extreme cases exist that signals from devices that are not carried by humans such as a driving recorder with internet functions may exist in raw samples. Extra approaches are required to 1) find out human travelers from raw logs, 2) distinguish their social groups e.g. tourist/pedestrians, and 3) identify activities.

- **Model specification:** A model that is able to capture characteristic of tourists as well as aforementioned sampling bias caused by sensing emoluments or errors in distinguishing activities is required.

In this study a methodology to use these Wi-Fi records for understanding and predicting tourist flows at a disaggregate level is proposed. The recursive logit model originally proposed in Fosgerau et al. (2013) to formulate tourists' route choice behavior is employed. The general concept is to derive the choice probabilities of the next decision points given attractiveness of the different downstream link options and whether traversing the link brings one closer to the destination. Though also stochastic path-based choice models are feasible, since there are often numerous paths and since many tourists might not choose the whole path but make choices *en-route*, we suggest this "recursive thinking" lends itself to this research's problem.

1.3 Research contributions

Given the overall objective, this research proposes a framework to predict tourist behavior based mainly on the Wi-Fi packet data. A complete set of processing starting from raw data cleansing to

population group analysis, tour-choice estimation, and finally tourist flow forecasting is proposed. Contributions are more specifically illustrated following the diagram shown at Figure 1.1 by stages.

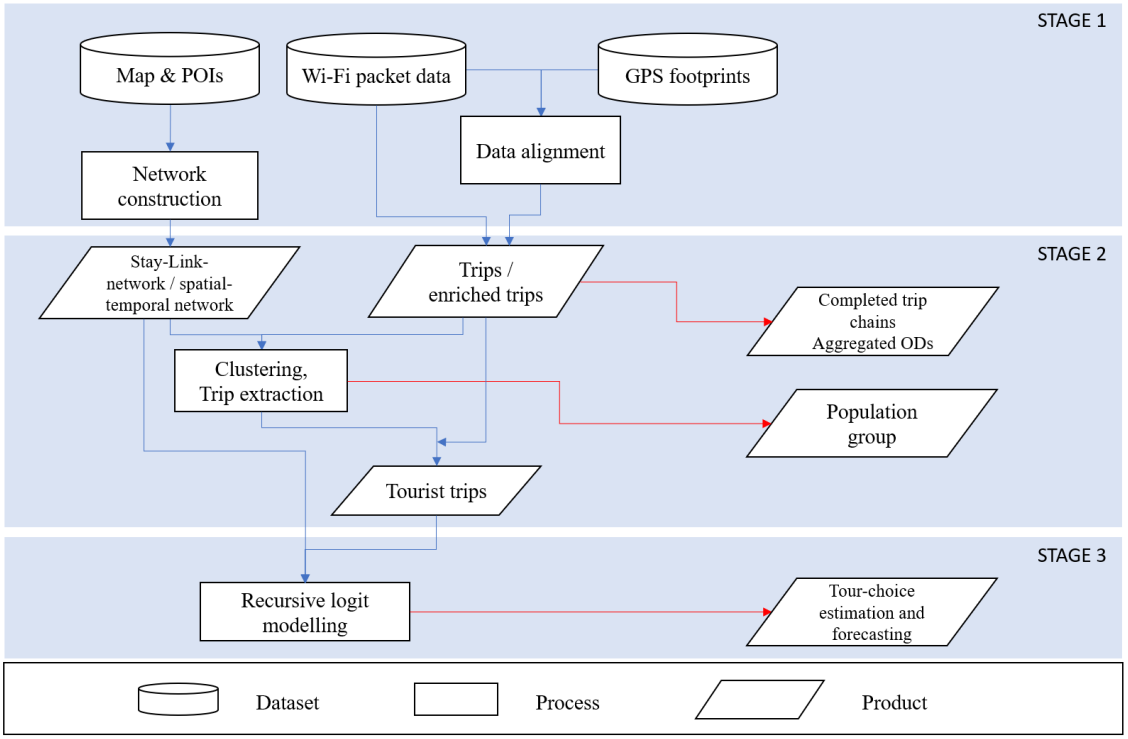


Figure 1.1 Research diagram

The first stage of this research is the data collection and pre-processing. Data required for the study includes the geographical information and the flow observation. Geographical information is collected from open sources data. Flow observation are obtained from electronic footprint surveys.

- By grouping up sensors with similar OD distributions and reducing the links on the network by features that affect tourists' behavior, the limited monitor resolution of packet sensing-based survey is addressed;
- Approach for organizing map data and flow observations into compatible routes and networks is proposed, and corresponded software is developed to convert raw observations into spatial or spatial-temporal trip-sequences.
- Measurement of POI attractiveness based on online reviews.

The second stage further looks at the characteristics of obtained routes. Trips and travel patterns are extracted from the sensing data, then obtain aggregated OD tables of surveys. For samples with

sufficient sensing density, the individual's travel path is recreated. For surveys that are difficult to interpret completed travel sequences due to sparse sensor coverage, we explore ways to enrich the observations.

- Clustering analysis is conducted to categorize different populations from trip chains, their characteristics are then identified.
- A time warping method is employed to fill the gaps in observations that are not informative enough to form trip chains, with aid of a small sample of complete GPS trips.

The third stage specifies the model. Different model specifications are presented in the two studies, one focuses on routes choices with possible brief stop en-route in a specific touristic area, the other focuses on individual general travel patterns at city-level. Specifically, works in this stage includes following contributions:

- A network structure with special links capturing the behavior of "stay at certain links for moments" without expending the network to temporal forms.
- Error terms addressing the observation missing caused by penetration rates.
- Warming up algorithm to find suitable initial guess to promote the estimation performance of the recursive logit.
- Comparison between non-temporal expanded network with stay links and temporal expanded network.

1.4 Outline of the dissertation

After the introduction of motivation and basic technology background of the packet sensor, chapter 2 reviews relevant literature of 1) characteristics of city-tourist behavior, methodology in studying pedestrian route choices; 2) in specific review regarding the recursive logit modelling (RL), for this is the major methodology employed in this dissertation; 3) implementation of electronic trace data in transportation research.

Chapter 3 introduces the tourism industry in Kyoto City, the subject of this study. Two Wi-Fi packet sensing experiments surveying street-block-scale and city-scale tourist trajectory are described. Additionally, for the city-scale data survey, GPS-based sample is introduced as the comparison object

since low resolution of citywide survey makes identifying travel pattern difficult. The data of the two experiments, as well as the auxiliary GPS sample, are used for further case studies.

Chapter 4 presents a methodology to utilize the relatively high-resolution Wi-Fi packet sensing data to model behavior of pedestrians in a touristic area. The obtained data sample is first categorized by two-step clustering for pedestrian identifications. Then network with "Stay-Links" to capture the behavior of brief stops during sightseeing is described. The recursive logit model is specified based on the characteristics of Wi-Fi packet data and the proposed network with Stay-Link. The estimates of the model capture the walking-tourists' behavior and well describe the difference in behavior between touristic-peak / off-peak hours.

Chapter 5 follows on from the conclusions for the city-scale data survey in Chapter 3. An approach for the enrichment of Wi-Fi data with auxiliary GPS sample is proposed. By similarity measurement, GPS trajectories are used to fill in the missing observations of Wi-Fi traces. The enrichment approach developed in this chapter lays the groundwork for tour-based modelling for urban tourist.

Chapter 6 proposes a methodology to model city-scale travel patterns in a day based on the enriched Wi-Fi packet traces introduced in Chapter 5. With specifically defined links to capture the error caused by monitoring area limitation, temporal-expanded network is employed, as in data preparation temporal variables describing travel cost and area attractiveness is obtained. The model estimation and limitation be found is discussed at last.

Chapter 7 summarizes main findings of this research. The comparison between non-temporal expanded network with stay links and temporal expanded network is specially discussed. Finally, future research directions in light of the contributions of this research are provided.

Chapter 2 Literature review

2.1 Characteristics of city-tourist behavior

Pedestrian routes, irrespective of whether the person is a tourist or not, have been studied by different approaches. Gipps and Markjo (1985) simulate pedestrians based on Cellular Automata. Inside the simulation, with random error, at each intermediate node an agent chooses which node to visit next obeying a set of rules, considering the perceived shortest path until s/he finally reaches the preset goal. With an experiment recording respondents' routes as well as a questionnaire to the same persons, Hill (1982) finds that pedestrians prefer the "simplest" route instead of the shortest one, as the directness and complexity affect their route choices. From a revealed preference survey asking respondents to choose their four most frequently chosen routes and their attributes Westerdijk (1990) finds that pleasantness also impacts the route choice significantly.

Then gradually researchers began to take behavior during walking into account in pedestrian route choice. Borgers and Timmermans (1986) introduce the concept of an "impulse stop" that randomly occurs during pedestrian walking to their preplanned destination. The probability of an impulse stop is affected by several environmental attributes of the route, for example, number and category of shops. Helbing et al (2002) simulate crowds by the "social force model", where pedestrians minimize their obstructions from other persons as well as infrastructure on their way to their destination. In this paper they demonstrate that this approach can lead to a clear understanding of observed pedestrian flow interaction in both normal and emergency situations. Variants of the model are implemented in for example VisWalk (PTV Group, 2011) and Legion (Legion Limited, 2012) and validation results have been in published in a number of publications (Martén et al., 2014; Chen et al, 2015). Social force models have been applied to both small area indoor modelling (e.g. Friis and Svensson 2013, Martén and Henningsson 2014, Daniel 2019), as well as to model pedestrians over longer time periods and spatially wider areas (Hänseler et al., 2014). A different approach to model pedestrians route choices is to utilize the Visibility Graph Analysis (VGA) that connects visible objects on a route with pedestrian movement in agent-based simulation (Turner et al., 2001). A fan-shaped vision field is widely applied for simulating how visible attractions affect individual route choice (e.g. Kitazawa et al., 2008; Asano et al., 2010; Park et al., 2013; Zhou et al., 2016). Wang et al (2014) developed an agent-based method taking visual attraction into account in pedestrian movement modeling.

The studies in this research are interested in understanding pedestrian behaviour over longer time periods, often several hours. Route choices are still estimated, but are less focused on detailed aspects such as interaction between pedestrians. For such a problem Ben-Akiva et al (1996) are a proponent of a "tour-based" perspective, arguing that any travel is a round trip from home with major and minor destinations making a connection between pedestrian's multiple movements in a day and that this should be the basis for route choice modelling (Ho and Mulley, 2013). Accordingly Bowman et al. (2001) extends this concept for modelling individual's travel schedule in one day by connecting the route choice modelling with activity modelling. Hoogendoorn and Bovy (2004) summarise that any action a pedestrian takes should be considered as providing utility, hence, both, which route they choose and how they behave during walking obeys the principle of utility maximization. In line with this, recent models jointly consider walking behavior and the goal of the trips.

Modelling pedestrian route choice as utility maximizing behaviour also allows the application of discrete choice models. Alivand et al (2015) apply a Path Size Logit (PSL) model for understanding relevant attributes and their relative importance of route choice between two touristic locations. More recently Lue (2019) develop a pedestrian route choice model for Toronto also using PSL but based on smartphone GPS data from 103 individuals. Further, though for cycling not walking, the study of Ghanayim et al (2019) apply mixed C-Logit, mixed, and mixed PSL for modelling route choice and obtain consistent results. One of their finding is a route "near sea" and "near park" is more attractive to the cyclers.

The above path-based discrete choice approaches assume that decision-makers plan their choices in advance, and that the consequent choices do not change during the journey. To relax this constraint, sequential choice models are established for forecasting individual trips from a less goal-oriented, more flexible perspective. Fu and Wilmot (2004) for example develop a sequential logit model with binary choices regarding whether to evacuate for a hurricane and estimate the model by considering all binary choices simultaneously instead of estimating multiple models along time intervals. Urata and Hato's (2013) work formulated the trade-off between delay penalties and engaging immediately in evacuation activities after a disaster occurred based on a logit model.

2.2 Recursive logit (RL) models

The aforementioned work of Fosgerau et al. (2013) then proposed the recursive logit approach as a link-based model which does not require generation of a choice set.

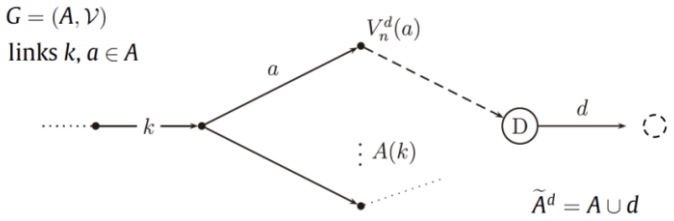


Figure 2.1: Concept of the recursive model. Adjusted from Fosgerau et al (2013).

The general concept of RL is shown in Figure 2.1 which follows the concept and notation of Fosgerau et al. (2013). At each link k individual n chooses the next link a among the set of outgoing links $A(k)$, maximizing the sum of instantaneous utility $u(a|k)$ and downstream utility $V_n^d(k)$. $u(a|k)$ and $V_n^d(a)$ is given as follows.

$$u(a|k) = v(a|k) + \mu\varepsilon_n(a) \tag{2.1}$$

$$V_n^d(k) = E\left[\max_{a \in A(k)} \left(u(a|k) + V_n^d(a) + \mu\varepsilon_n(a)\right)\right], \quad \forall k \in A \tag{2.2}$$

Equation (2.2) is known as the Bellman equation (Bellman, 1957). By the Markov property of traveler's sequential link choice, this dynamic choice model is proven to be equivalent to an MNL model with infinite alternatives (Fosgerau et al. 2013).

The decomposition method (DeC) developed by Mai et al. (2018) significantly improves the efficiency of the estimation process by solving the set of Bellman's equations corresponding to all destinations in one system. However, this method is only applicable for the "basic" form of RL. If the utility is specific to destinations, such as a "link size" attribute or a "discount factor", the estimation process has to be solved one OD pair at a time. Link size controls for the inclusion of a link in multiple paths similar to the PSL logit model (Fosgerau et al. 2013). The discount factor of Oyama and Hato (2017) instead controls for how myopic travelers are in that they discount downstream utility $V_n^d(k)$.

Following the recursive logit model, individuals' route choice is decomposed into a bunch of sequential link choices while with an extra term valued by Bellman equation that measures the utility downstream until a given destination. Mai et al. (2015) further develop the RL model into a nested recursive logit model to relax the independence of irrelevant alternatives (IIA) by giving scale parameters specific to each link. Oyama and Hato (2017) argue that decision makers have insufficient information for future decision-making states and introduce a discount factor to the expected downstream utility in a recursive logit approach for modelling people's myopic behavior in sequential route choices. Kaneko et al. (2018) introduce "Link Awareness" and show that it improves the stability of model estimations.

Closely related to this study are further the works of Zimmermann et al (2018), Hidaka et al (2019) and Vastverg et al (2020). Zimmerman et al use person trip survey data collected in Stockholm and model the location and/or activity change in 10min time intervals. Vastverg et al (2020) models the same data by a dynamic discrete choice model based on Bellman equations that allow some of the elements that effect the utility function to be stochastic. Finally, Hidaka et al (2019) propose a method combining RL modeling with a prior generated feasible time-space route set. The behavior of making a detour towards a POI while walking is modeled via numerical tests and generated data. All these three studies use a large time-expanded network definition where choosing a single "Stay-Link" means to remain a predefined amount of time in a specific state. In this dissertation instead, a simpler network is proposed to model if a person is staying at least a certain amount of time on a link due to attractions on the link. Chapter 4 returns to this discussion, when introducing the modelling approach.

2.3 Data sources for investigating pedestrian / tourist flows

Like many large-scale national wide surveys, the PT survey asks respondents to provide information on their movement in a weekday, as well as their household and personal demographics. Then the daily OD distribution is understood by social-demographic, purpose and modes. In Japan, the survey was first conducted in 1967 in the Hiroshima metropolitan area, and since then it has been conducted in metropolitan areas across the country to understand the current urban transportation, forecast future demand, and create master plans.

The choice of a specific modelling approach clearly also depends on the data available to the analyst. Above RL literature has mostly developed with the onset of detailed route tracking possibilities. Instead, most of the case studies describing tourist behavior are based on travel surveys. These bespoke surveys have limitation on revealing "actual" sequences of choice, as in general in trip diary type surveys short or secondary activities tend to be under reported (Aschauer et al. 2018, Thomas et al. 2018). In order to understand why tourists spend time in some touristic areas, however, such short, "minor" activities are important. This is one of the reasons why major "intelligent" data sources are attractive for studying tourist and pedestrian behavior in general.

One of the most powerful data are GPS traces. Passively collected GPS datasets can contain large amount of high-resolution digital traces without considerable bias. However, especially for collecting pedestrian data, such data are often hard to obtain, not least due to privacy issues. Either service providers with a certain market share have difficulties sharing data with researchers, or require persuading respondents to install a specific (and usually not commonly used) monitoring application. Consequently, although GPS data have a significant advantage in containing detailed choice sequences, the sample size is in most cases not much different from questionnaire surveys (e.g. Maruyama 2015, Lue 2019, Marra et al. 2019). Further, biases may occur depending on the target group of the add-on services of an application which the GPS tracking can be combined with. For example, tracking information collected by consensus from persons using a public transport journey planner will lead to the omission of all car users. If, instead tracking is based on dedicated recruitment without additional services, samples are likely to be disproportionately young (Lue 2019, Marra et al. 2019).

An alternative to GPS footprint data is the cellular data, usually referred to as "call detail record data" (CDR) cover millions of cell phone users. Its relatively poor spatial-temporal resolution but large sample size is suitable for large scale studies on activity-based modelling. Using anonymized CDR, Yin (2018) employed unsupervised and semi-supervised Input-Output Hidden Markov Models (IOHMM) for recognizing activity patterns. However, such data is so detailed that some of the governments impose regulations on it. Japan does not allow telecommunication companies to sell any disaggregated level record that can form a trace, even they are anonymized.

Instead, this study is based on Wi-Fi packet data. Collecting Wi-Fi as well as Bluetooth packet data requires deploying sensors as "packet sniffers" of probe requests. These requests for network connection are sent by Wi-Fi-enabled devices with frequency of 30 ~ 90 seconds and are used to search routers around, hence its transmission is independent as to whether the device is already connected to a network. The Probe Request Packet is "gossipy", including plenty of information that may lead to violation of privacy (Cunche et al 2012, Freudiger and Julien 2015). As a countermeasure both Google and Apple Inc. anonymized the MAC-address in probe request packets by randomizers since 2015, especially when a device is disconnected to any routers. However, within two years this randomization process has proven vulnerable and been suspended in some devices (Vanhoef et al 2016). Experiments in this research hence also detect a mix of randomized and non-randomized MAC-addresses. Every randomized address is unique, therefore without any processing, one smartphone would be observed as if there are multiple. It is, however, helpful that the randomized MAC-address can be easily distinguished by a specific flag inside the packet (at least during the time of data collection of this research).

One of the advantages of these location-based sensing methods is that the sensor has the ability of pre-processing data locally. With embedded CPU, before uploading observations to Internet, a sensor can hash and encrypt the observed MAC-addresses, or directly erase data that violates privacy. This property gives the data a natural advantage in respecting privacy also from the sensing side.

According to Mochizuki et al (2018), for smartphones in a pocket or bag, the detectable range of a sensor deployed in environment like urban streets is about 40 meters. The spatial resolution and type of analysis that can be conducted with Wi-Fi data clearly depends on the density of the sensor deployment. In most practical applications this is, however, not that high so that the studies are focusing on analyzing rather aggregate activity patterns. Among the, so far, relatively few journal publications using Bluetooth or Wi-Fi sensor data for travel pattern detection refer to following:

One of the first papers appears to be the one by Martchouk et al (2011) who analyse freeway travel time variability in Indianopolis, U.S. using Bluetooth data depending on, among others, time of day

and weather conditions. Nishida et al (2014) develop the Wi-Fi package sensor with anonymizing function, and show its potential of monitoring traffic volume by applying the sensor in a service area to measure the time vehicles stay at the service area. Ota et al (2018) propose an equation with empirical acquisition rate that reproduces the traffic volume from Wi-Fi probe request. Crawford et al (2018) use a large set of Bluetooth detections of car drivers in Wigan, U.K., to cluster these according to their temporal spatial travel patterns. They found that frequency as well as variability are good clustering variables. Further on clustering, but closer to the problem studied in this paper, Fukuda et al (2018) employ stochastic block models as a co-clustering method to classify the travel patterns of tourists by looking at the type of tourists and their destinations. With respect to pedestrian flow modelling, Versichele et al (2012) analyse flow patterns aggregately at Ghent Festival in Belgium using Bluetooth data.

As will be described further in Chapter 3, twenty sensors are deployed in a 1 km^2 area. The availability of Wi-Fi packet data in terms of chasing individuals' sequential choice behavior is explored. Given that tourist route choices in a touristic area include a large number of potential routes and the nature that tourists might indeed be strolling rather than aiming to fast complete a prior chosen route, a variant of the recursive logit model for this is employed. In Chapter 4, 39 sensors are deployed in key terminals and touristic attractions of Kyoto City. How to use this low-cost sensing technologies to capture tourist flow at city-scale is explored.

Chapter 3 Tourism in Kyoto City, data description

3.1 Introduction

3.1.1 Tourism in Kyoto City

As the target area of the entire research, this section gives basic knowledge of the tourism in Kyoto City. Kyoto was the capital of Japan from 794 until 1868, and remains one of the centers of Japanese culture until the present day. The long history of the city has resulted in a wealth of historical site. Buddhist temples and Shinto shrines are the main attractions throughout the city, of the attractions, 17 historic monuments are listed as the World Heritage (13 temples, three shrines and one Japanese castle). Some of the traditional folk festivals in Kyoto, such as the Aoi Festival and the Gion Festival, have been held for more than a thousand years and have become unique scenery in the city.

Being one of the most popular tourist destinations in Japan and around the world, over 55 million tourists visited Kyoto City every year since 2014, and the number has been expected to continuously grow (until the COVID-19 crisis), as shown in Figure 3.1. The Kyoto Tourism Map (Figure 3.3) published by Urban Planning Bureau of Kyoto City (2006) give an impression of how densely the attractions in the Kyoto City are. Since most of the tourist attractions in the city are historical sites, the map is still a very valuable reference. In a city of only 1.4 million inhabitants, problem does arise: tourism resources throughout the city have instead become one of the main causes in traffic congestions over the city. The extreme crowding not only reduces the standard of tourism services, but also impacts the travel of local residents. Satisfaction survey targeting tourists shows that “crowding and congestion” and “public transportation” are the top two categories in dissatisfaction about tourism in Kyoto City, and the disappointments deepens year by year (Figure 3.2, Kyoto City, 2016, 2017).

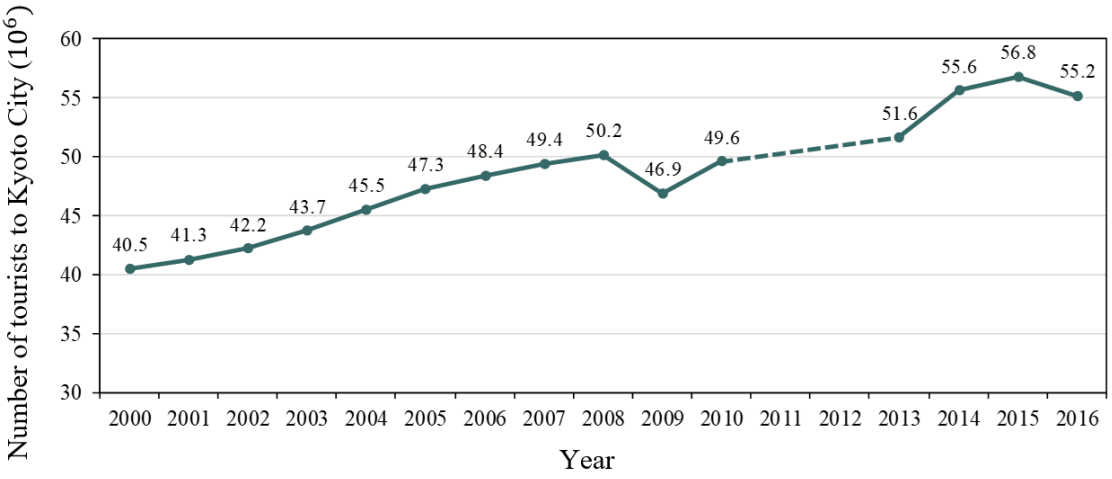


Figure 3.1 Annual number of tourists to Kyoto City (Source: Kyoto City, 2016, no data for dotted parts)

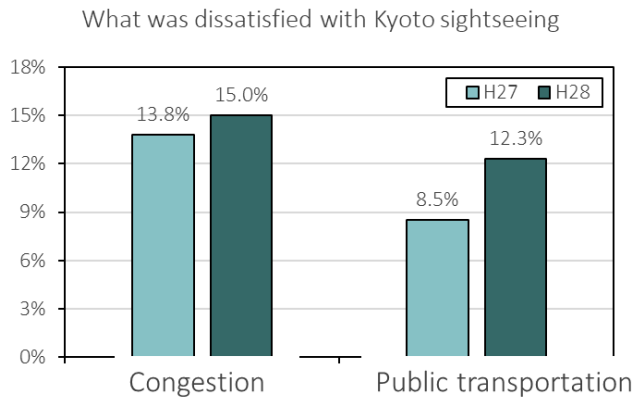


Figure 3.2 Dissatisfaction about tourism Kyoto City (Top two responses among 20 categories, excerpt from Kyoto City, 2016, 2017)

3.1.2 Investigating tourist flow in Kyoto City

The tourist-trip survey in 2006 is the last large-scale survey specifically revealed the tourists' person trips in Kyoto City. The survey asks respondents to provide information on their trip chains in a day, their modes and cost of each trip, given the 37 defined touristic areas in Figure 3.3 as reference. All of the respondents of the survey are Japanese. This was the last person trips survey of tourists, and

since then, the survey led by the city has shifted to tourist's destination choices and resulted satisfactions.

During the 14 years of survey blanks, Japan's tourism market has changed considerably. With the economic development of the emerging countries in Southeast Asia and policy deployment by the national government for attracting foreign tourists, the number of tourists visiting Japan has increased dramatically. It is almost impossible for a PT survey from 14 years ago to contribute to the analysis and prediction of tourist flows, while the city is in desperate need of information, such as how tourists move through tourist areas, how they plan their journey in a city, and, fundamentally, whether or not it can be understood to some extent by inexpensive methods are barely known for lack of large-scale investigation.

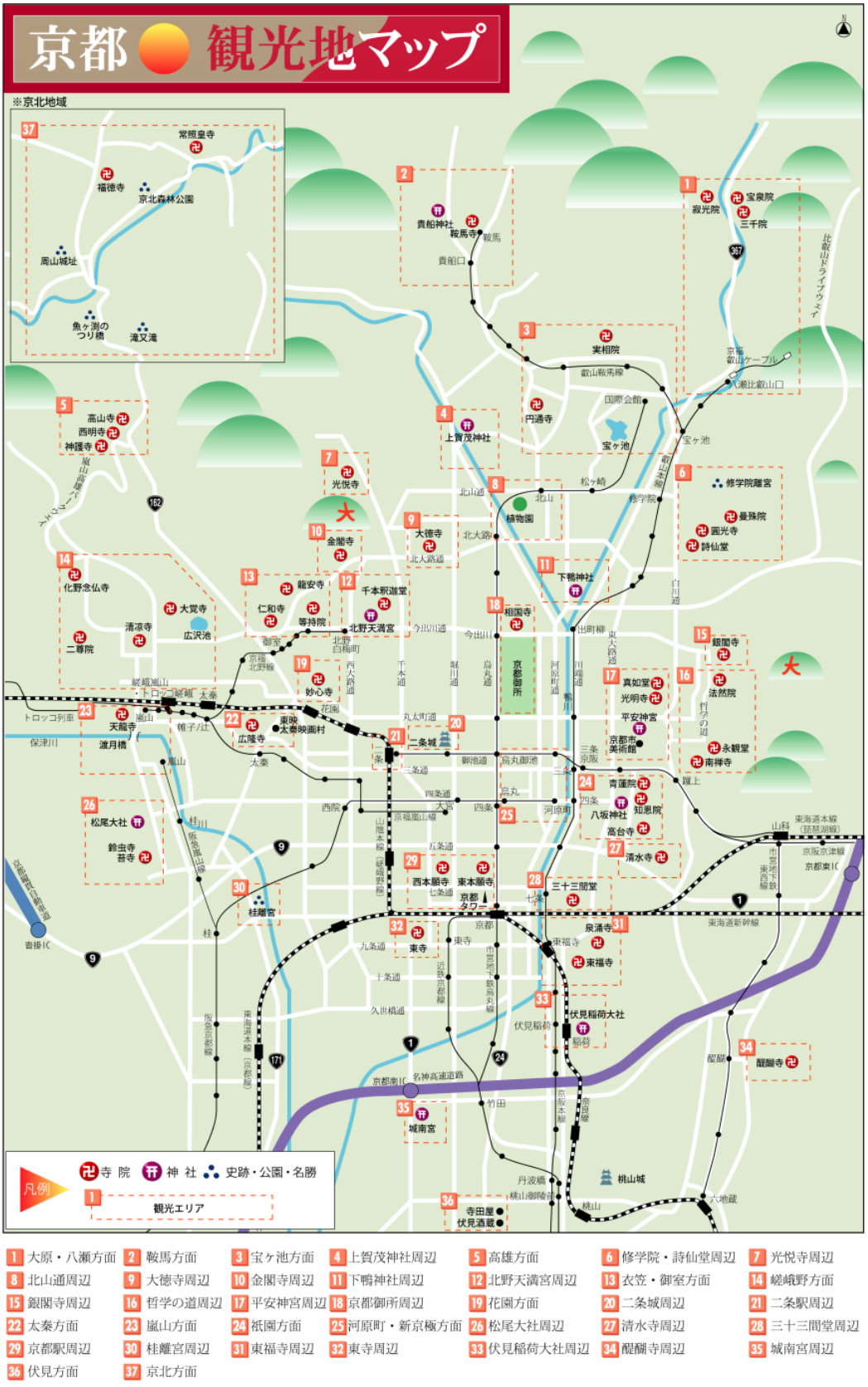


Figure 3.3: Kyoto Tourism Map (Kyoto City, 2006), English reference see Appendix 1.

3.1.3 Wi-Fi packet data

With the rapid development and spread of smartphones, Wi-Fi packet sensing is expected to provide new possibilities for obtaining tourism flow data. These sensors are designed to detect and record all electronic devices with Wi-Fi function enabled within an average radius of around 40 meters. All detected devices are identified by an anonymized label encrypted from the device's MAC address. In addition to the hashed MAC address, Wi-Fi packet sensors also record timestamp, packet sensor ID, Received Signal Strength Indication (RSSI), and the device's vender ID. Multiple detections and the above attributes allow the routes of any individual who carries a detectable device to be observed.

It is considered the several advantages of Wi-Fi packet data that help overcome traditional survey problems, including rich sample size, continuous monitoring, and real-time. Furthermore, compared to other electronic data sets such as GPS data, with Wi-Fi packet data it is easier to obtain unified samples, as obtaining information does not rely on any specific smartphone application.

Taking the benefits of Wi-Fi packet sensing, this chapter presents the results obtained from two experiments of different scales conducted in the city, a small area high monitoring density experiment and a wide area low monitoring density experiment. The two experiments test the practicality of utilizing the Wi-Fi packet sensing for flow investigation at different scenarios. The first survey monitors a one-kilometer a tourist area with a relatively simple demographic composition that most of the people inside are walking tourists to obtain relatively high-resolution location records of individuals inside.

The second survey expand the monitoring to the whole Kyoto City while not much increase the number of sensors. This city-scale survey obtains data for understanding the crowds at major public transportation points and famous attractions, and provides possibility for us to infer the travel patterns of tourists at city-level. Aiming at understanding such travel patterns, the current drawbacks of Wi-Fi packet sensing is firstly discussed. Through the comparison between Wi-Fi sensing and GPS-based location recognizing, we found that to a certain extent the two types of data are complementary, and further explosion towards the fusion of them is worth trying.

This reminder of this chapter is as follows. Section 3.1 reviews the survey targeting tourist in Kyoto City and explain why the Wi-Fi sensing survey is valuable for the city. Section 3.2 introduce the overview of the first Wi-Fi packet sensing survey in Higashiyama Ward, Kyoto. Section 3.3 introduce the overview of the second Wi-Fi packet sensing survey for the whole Kyoto city, with the comparison to a commercial GPS sample that will be involved for further study. Conclusions with the usefulness of data are drawn in section 3.4.

3.2 The Higashiyama experiments

3.2.1 Survey overview

Both experiments of this research use the same sensors as Mochizuki et al (2018) to collect users' footprints. The first data collection experiment was conducted in Higashiyama Ward (24th and 27th area of Figure 3.3), Kyoto City from November 2017 to March 2018. 20 sensors were installed in street blocks of this area, as shown in Figure 3.4 in purple. The area is about 700 meters long from west to east and 800 meters from north to south. Well-known sightseeing spots of Kyoto City are located inside this area. Sensor No.1 is set 50 meters south to the entrance of the often-visited Yasaka Shrine and sensors No.15 and 16 are near the entrances of Kiyomizu Temple, a world-heritage site.

Figure 3.4 also shows the actual network in the survey area. The area can be considered approximately as a grid area, and since the east of the Higashiyama area is a mountainous area with poor associability, most of the tourists enter this area by a north-south trunk road along sensors No.6, 10, 19 and 20. As presented in the figure, the sensors cover most of the intersections as well as the main "entrances" of the area. The uncovered are mainly private area with almost none point-of-interests (tourists-related POIs are projected in Figure 3.4). Although it is difficult to tell which route exactly an individual has taken, in terms of monitoring resolution, taking account of the uncomplicated road network structure of the area, the deployment gives enough knowledge for us to study route choices of travelers inside.

The ratio of travelers who turned on sensor as well as penetration rate of the sensor should be considered. Ota et al (2018) conduct a questionnaire survey within the area of Figure 3.4 investigating how many tourists carry a smartphone with the Wi-Fi function enabled. Among 360 Japanese tourists 86.6% and among 200 foreign tourists 97.1% answered that they have a smart phone. Further, among

those carrying a smartphone, 57.1% of the Japanese and 75.1% of the foreigners turned on the Wi-Fi function. Japanese often use the mobile network only whereas foreigners often purchase Wi-Fi packages or rely on free Wi-Fi services in the city. Thus, sample in this study seems to have a bias towards over-sampling foreign tourists.

During the data collection Ota et al. (2018) also evaluate the penetration rate of the sensors used in the study by comparing actual, counted flows with Wi-Fi records. They correlate these considering three factors: 1) surrounding environments; 2) height of sensor installation; 3) speed of tourist flow. With these factors they develop an empirical formula to estimate sensor specific penetration rates that vary between 42.2% and 76.7% (see Figure 3.4). The average error in their empirical formula for reproducing the flow of people compared to the actual flow of people is only about 8%. Therefore, the penetration rates for different sensors are estimate-able but should be considered in this study to reduce biases.

Sensor ID	Daily Flow (10^4)	Penetration Rate
1	1.7	76.7%
2	1.9	76.7%
3	1.0	47.8%
4	0.77	47.8%
5	0.96	42.2%
6	3.4	68.6%
7	5.9	76.7%
8	1.7	47.8%
9	1.6	42.2%
10	4.3	68.6%
11	1.6	60.6%
12	1.3	47.8%
13	1.8	60.6%
14	6.0	76.7%
15	4.6	53.4%
16	2.2	68.6%
17	2.1	53.4%
18	0.91	53.4%
19	3.2	60.6%
20	3.8	54.2%

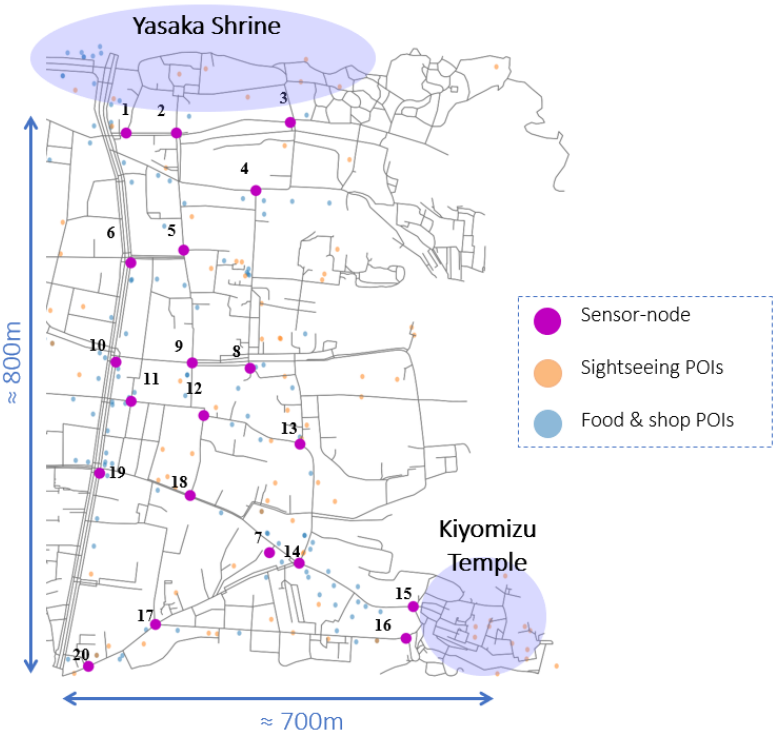


Figure 3.4 Survey area of Higashiyama experiment

3.2.2 Data description

The sensor is embedded with a MAC address hash-encryptor that does anonymous processing on its local CPU. Further, before uploading any records, information except for encrypted MAC address, timestamp, Received Signal Strength Indication (RSSI) and device's vender ID, is dropped.

By the hashed MAC address individuals are indexed. Their trajectories are inferred from timestamp of the record and by which sensor they are detected. During the experiment a total data amount of around one million probe requests are collected per day. Out of these about one hundred thousand are randomized MAC addresses and it is not possible to identify as to how many devices this corresponds to. To understand tourist behavior, clearly these randomized one-of records and other unique observed MAC addresses are not useful. On average around 90,000 non-randomized (but anonymized) MAC addresses are collected at least twice each day.

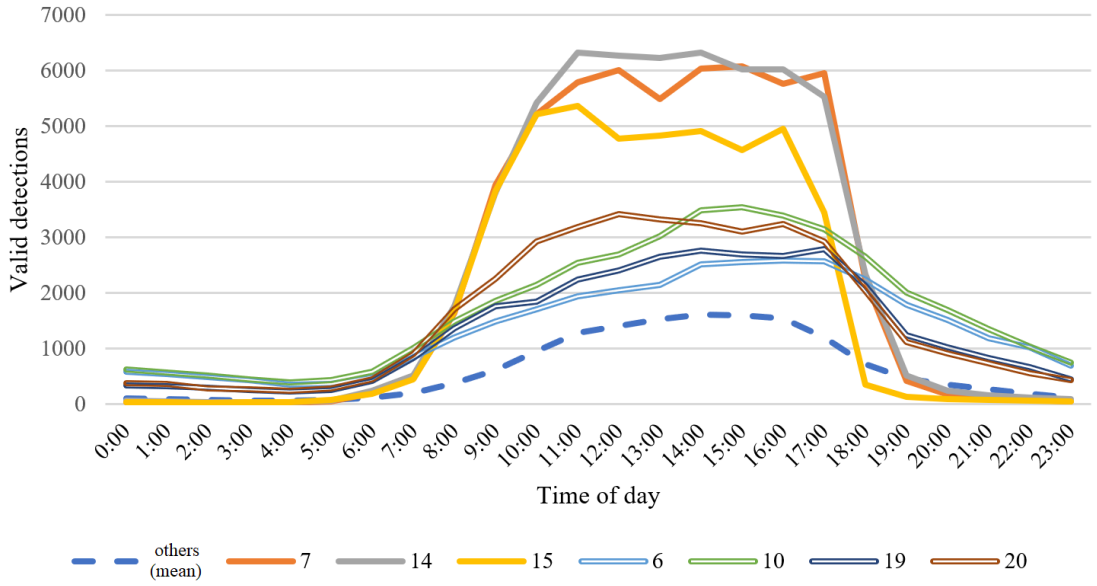


Figure 3.5: Detections by time of day for sensors shown in Figure 3.4

Figure 3.5 shows the daily counts of valid detections along hours by selected sensors and the remaining sensors ("others"). For each sensor the detection begins to increase fast around 8:00 and reaches a peak around midday. At daytime the counts from sensor No.7 (a bus parking area for Kiyomizu

Temple), No.14 (a path connecting the parking area and Kiyomizu Temple) and No.15 (the entrance of the temple) are far higher than the counts from other sensors. These counts drop quickly after 17:00 when the temple closes. Also, looking at the sensors along the main road in the survey area in Figure 3.5, (sensors No.6, 10, 19 and 20, plotted with double line), the counts are high compared to those at other sensors and, as expected, do not drop as sharply during evening hours. This aggregation implies that there are not only tourists among the population observed, but probably residents as well, and the need to be further studied.

In conclusion, in Higashiyama experiments the dense deployment of the sensors gives us enough knowledge to learn about the travel inside a touristic area in detailed resolution. A large sample of MAC-address matched individuals whose routes are feasible to be inferred are collected, while the data loss during a travel should be carefully considered. Since this is a typical tourist area with relatively little interaction with the daily activities of residents of Kyoto City, a large part of the observed population is presumed to be tourists. Preliminary statistics by location and time support this by finding out that in general people be captured behave like tourist, while also suggest that residents mixed in the sample should be to some extent distinguished.

3.3 The Kyoto experiments

3.3.1 Survey overview

The second data collection experiment conducted in 39 locations distributed in the whole Kyoto City, as shown in Figure 3.6. The deployment covers most of the transit station and major point of interest in Kyoto City. The details of deployment are summarized in Appendix 2. Locations that are under monitoring can be categorized to train stations and sightseeing spots. Unlike the first experiment in Higashiyama Ward, in this experiment sensors are installed at designated key locations in Kyoto City. Data resolution in this survey dropped while it is expected to capture people all over the city. Thus, the expected outcomes of the experiments shift from recognizing individual's detailed paths to inferring the travel patterns in the city.

As reference for sensor locations, this section again refer to the "Kyoto Tourism Map" in Figure 3.3. Out of 37 touristic areas designated in this survey, 16 of them are under the monitor of Wi-Fi packet

sensing. Due to permission reasons the Wi-Fi packet sensing did not take place at a few major tourist attractions, such as the Kinkakuji Temple and the Shimogamo Shrine, which are both world-heritage sites. Bias caused by lacking of monitoring in such areas is one of the key issues to be dealt with and will be addressed in following models.

Sensors at train stations are installed near the checking gate and transfer gate in order to capture the tourist as much as possible. Especially, for Kyoto Station, the major transit terminal placed in central Kyoto City, all of its nine ticket gates, seven belong to Japan Railway Group and two belong to Kyoto City Subway are covered. Apart from Kyoto Station, some other main transit stations in the city such as Kawaramachi Station operated by Hankyu Railway and Shijyo Station operated by Keihan Railway are also covered. Additionally, some train stations close to tourist attractions also under monitoring. Similar to the issue aforementioned for lacking of monitoring for several important tourist attractions, not all key terminals in the city are covered in the survey.

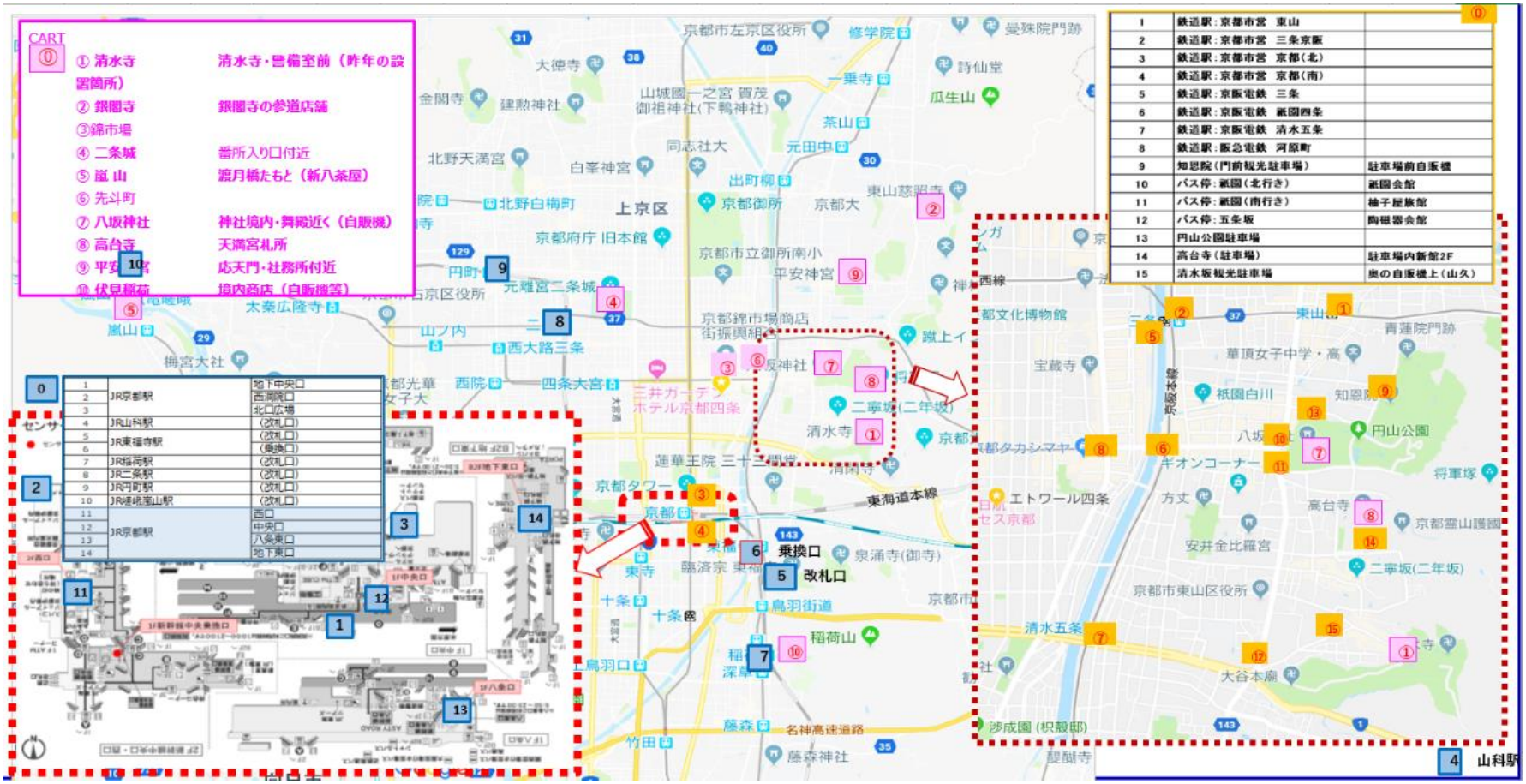


Figure 3.6: Locations of 39 sensors deployed in the Kyoto City. (Lower right are sensors in Kyoto Station)

Start from November 19, 2018, this data collection project has been running for over two years and is still ongoing (November 2020). The anonymous policy is as same as aforementioned Higashiyama Ward survey, that all collected MAC-address are hash-encrypted and the encryption key changes every 14 days. Figure 3.7 shows the daily number of observations along time from Saturday, December 1, 2018 to Thursday, January 31, 2020. At most more than seven million non-randomized MAC-address are captured in a day (the New Year's Day), around seven times as many as in the previous experiment.

Travel behaviors can be found reflected from the aggregated results. Eight peaks can be seen during this two-month period. Seven of the them are weekends, and the highest and long-lasted one is New Year's vacation. The number of observations at train stations increases or decreases slightly more sharply than the number of observations at tourist attractions. This is more obvious on a normal weekend. It is speculated that this is due to the fact that, compared to locals or Japanese tourists, travel plans of foreign tourists seems less rely on public transportation, and are less restricted by whether the day is a weekend or not. In other words, tourist trips observed on a weekday are more likely to be unrelated to public transportation.

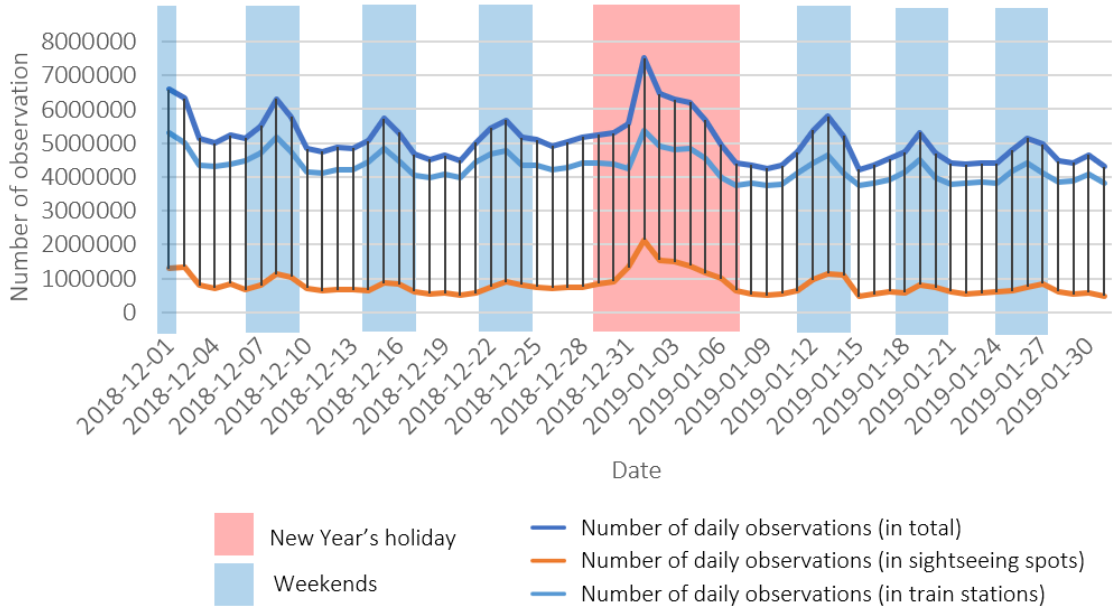


Figure 3.7: Number of daily observations from December 2018 to January 2019

Although the number of daily observations in this experiment is even much larger than the previous survey, a major drawback is considered in the obtained data: too few observation for each individual. For example, on 1st of December 2018, there are only 36% of the MAC-address (2,370,500 of 6,592,453) shows up more than once, and within those has been recorded multiple times, a large percentage of these people were recorded in the same location. Individuals with too few observations (or too few observed locations) are not informative enough for understanding travel patterns as the travel are hard to be reproduced.

Figure 3.8 aggregates the number of observed hours in a day of each individual. Subfigure on upper right is the aggregation without individuals only be capture once. Blue bars are the counts of records per individual, and orange bars are the counts of locations per individual. The number of individuals drops sharply, suggesting that most of the individuals only have been recorded less than six hours a day.

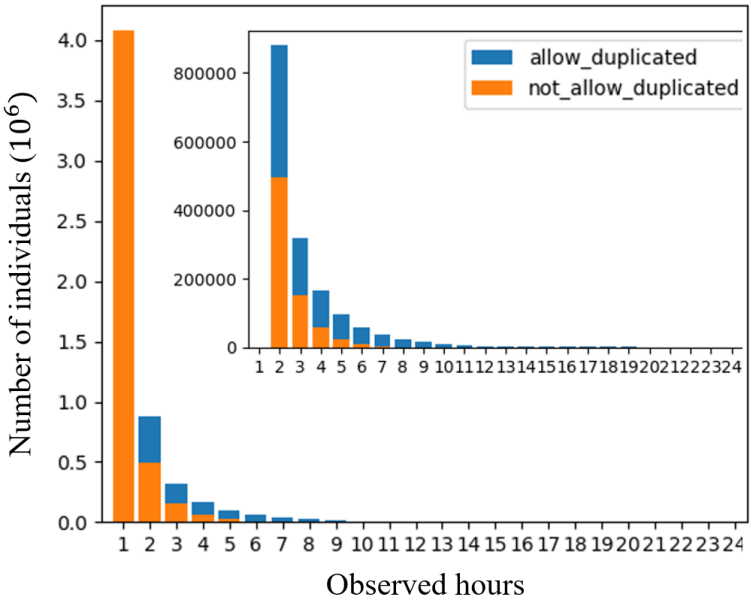


Figure 3.8: Number of individuals by observed hours in a day (Wi-Fi sample, first two weeks of Dec 2018)

To summarize this section, Kyoto data collection experiments collect (and are still collecting) huge size of sample for varies of key touristic locations in Kyoto City. The survey captures seasonal dynamics for each area, and can be a foundation for look at tourists' travel behaviors at disaggregated

level. However, it is considered necessary to have further discussions on where individuals have been to when they are outside of the monitoring.

3.3.2 GPS electronic footprints

The experiment carried in Kyoto City has fewer sensor per area unit compared to the survey conducted in Higashiyama ward, which causes excessive blanks that limited the understanding of tourist flow patterns in Kyoto City. A small size of commercially available GPS electronic footprint data is then involved to support the analysis and enrich the Wi-Fi observations. This section looks at the characteristics of the GPS data based on smartphone location information.

The GPS footprints use in this dissertation comes from a location-based service (LBS) named "Arukumachi Kyoto". The application, named after the Japanese words "Walk in Kyoto", provides Kyoto City-specific navigation service for users, such as navigation based on public transportation time table, as well as introduction/information of tourist attractions. Accordingly, the users of this application are presumed to be mainly tourists. The data includes the GPS location logs from September 2018 to February 2019 of those users who agree to share their locations with the service provider. The data is indexed by encrypted user-IDs. The encryption is similar to AMAC in Wi-Fi data while the hash function is different. In other words, with no means the researcher can know if someone detected by Wi-Fi packet sensors exists in the GPS database.

Figure 3.9 plots the number of observed hours in a day of each user in the whole 6-month GPS dataset. Compare to Figure 3.8, although the sample size is much smaller, the overall proportion of individuals with a long number of hours observed increased significantly.

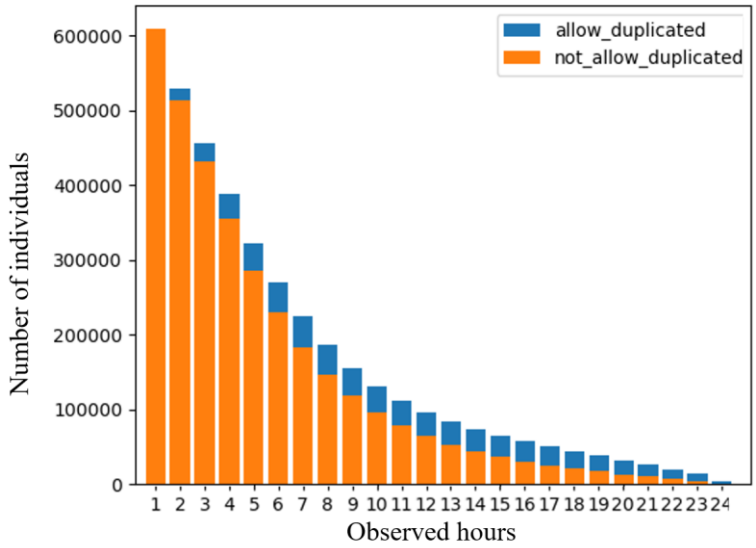


Figure 3.9: Number of individuals by observed hours in a day (GPS sample, from September 2018 to February 2019)

To further understand the GPS sample, the map is of Kyoto City is divided into a 200 x 150 meshes. Numbers of GPS records located in each mesh are plotted in Figure 3.10 by color gradient. For better description, the cumulated numbers are scaled by log10. The meshes with more than 10,000 records are mainly located at Kyoto Station and along Higashiyama area, and the blank meshes, which means less than 10 observation in half a year, are mountains surround downtown Kyoto. Compare to a real map of the Kyoto City (See Figure 3.3), this figure well describes not only famous tourist attractions, but also, surprisingly, trunk routes in the city. This raise a concern that if the GPS footprints can represent the actual origin and destination of trips.

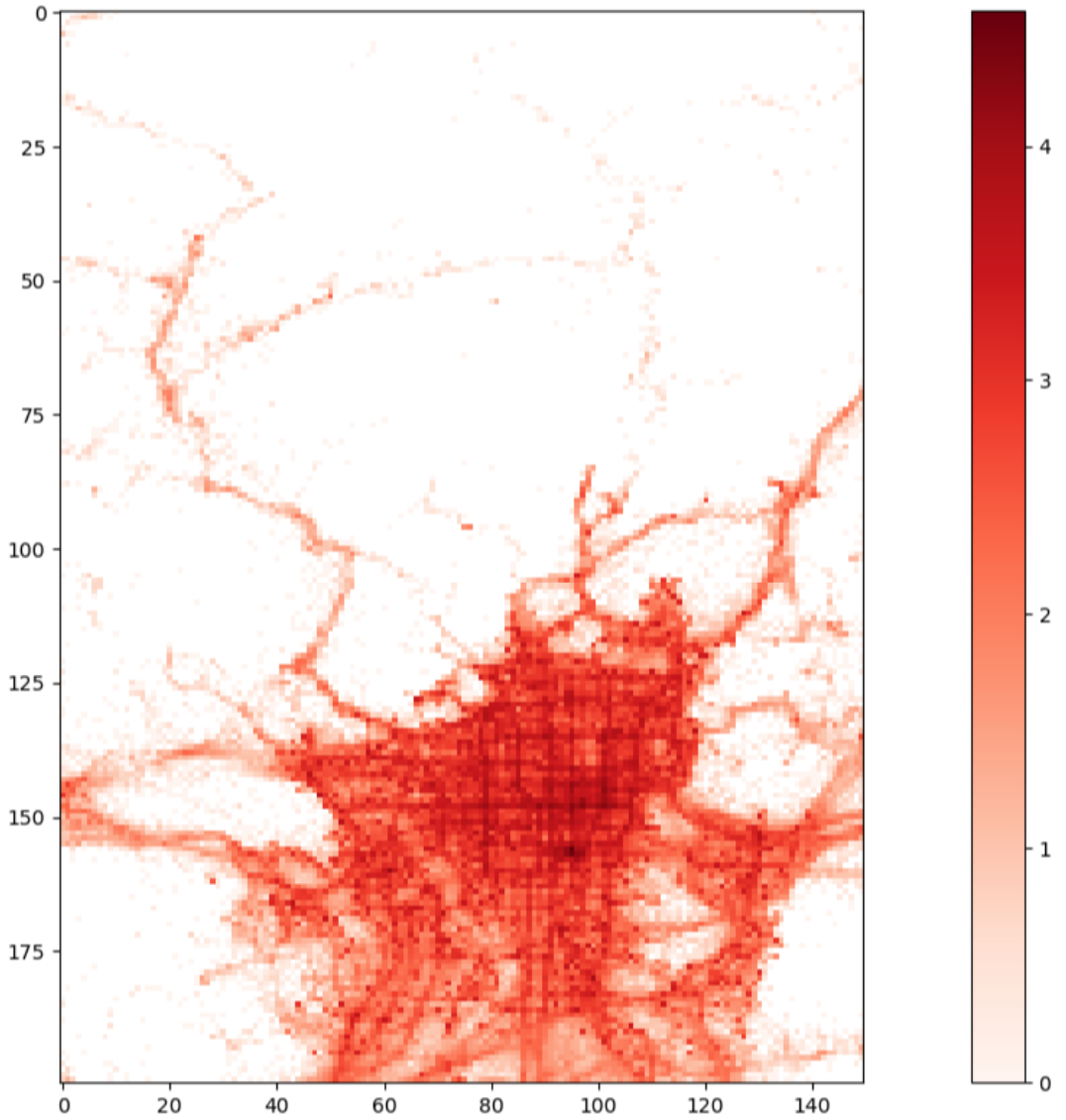


Figure 3.10: Heatmap with numbers of GPS observations, colorbar expresses the number of observations scaled by log10 in a mesh.

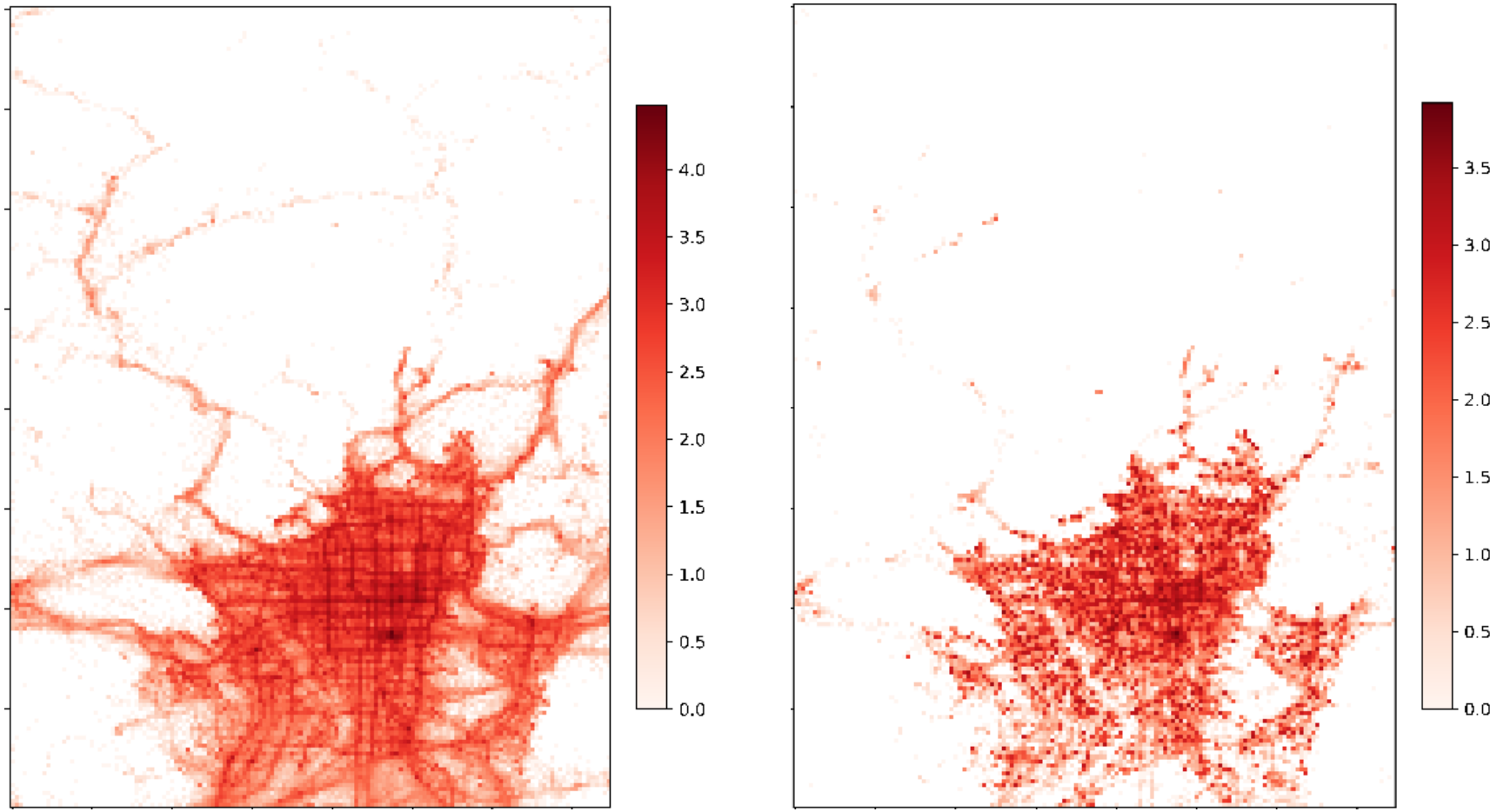
Like most of the location-based service in the world today, the "Arukumachi Kyoto" application provides service for both iOS and Android users. In the half-a-year dataset, 63% of the users use iOS and 37% use Android. The two operating systems have different privacy policies for reporting user locations (as of Feb. 2019), which lead to differences in collected electronic footprints. On iOS, the first time when users start the application, they will be asked if agree to upload the GPS records. They are able to choose among "No", "Yes, only when I using the application" and "Yes, always". Android users can only choose Yes or No towards whether to activated GPS reporting.

Even if users agree to report their location “always”, the reporting timing still varies by system. In Table 3.1 shows the number of observed hours of day from iOS users and Android users. Android users have noticeably more observable hours than iOS users. Note that more than 75% of the iOS users only be recorded for less than four hours in a day, which makes difficulty in inferring traveler’s activity chains.

Table 3.1: Statistics of observed hours in GPS “*Arukumachi*” dataset by operating systems

	ANDROID	**IOS**
Number of observations	468114	634708
Mean (std.) of observations from one device daily	9.40 (6.94)	4.53 (3.01)
Minimum of observations from one device daily	1	1
25%	3	2
50%	8	4
75%	15	6
Maximum of observations from one device daily	24	24

Moreover, tendency is found that Android system reports locations approximately once an hour, while iOS reports locations more frequently if the smartphone is travelling. This is an explanation for why in Figure 3.10 the trunk roads of the city are outlined. Figure 3.11 separately projects the number of observations onto the colored gradient by operating system. The outlines of roads are then no longer obvious in the projection of Android system while still clear for iOS.



(a) (b)

Figure 3.11: Heatmap in numbers of GPS observations of iOS(a) and Andorid (b), colorbar expresses the number of observations scaled by log10 in a mesh.

In summary, as data that is passively emerged, GPS data can certainly give a more detailed trace of travel than packet sensing, with a greater number of records for each trace, the data is considered valuable for filling the blanks in Wi-Fi observations. However, since the position reporting is almost random with not high frequency, whether the location recorded is the actual origin/destinations of an individual is difficult to determine. This drawback is further amplified by the presence of privacy protection mechanisms, as the iOS tend to report user location only during travelling and Android only report location once an hour. We close this section with noting that the purpose of the section is not only understanding a specific to dataset provided by a specific company. Instead, as a representative commercial data based on location service network, analysis for this data help us learn characteristics of commercially available location data under the current regulation.

3.3.3 Wi-Fi packet data vs GPS footprints

The last two sections learns the basic features of two electronic footprint datasets, one from the Kyoto City Wi-Fi packet sensing survey, the other from GPS footprints provided by a location-based service. This section compares the advantages and disadvantages of them and discuss the feasibility of integrating both types of data.

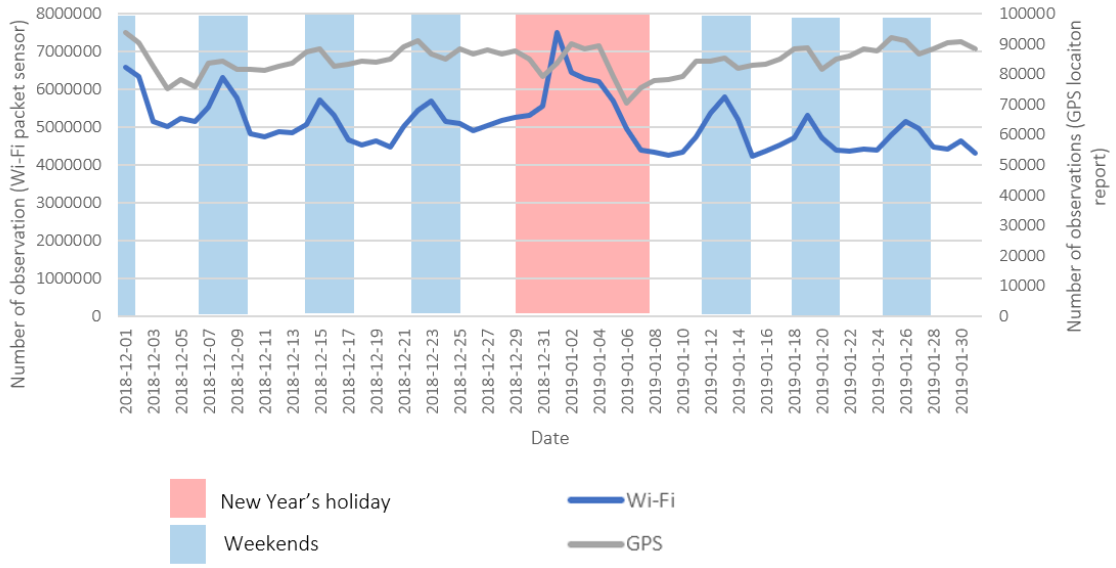


Figure 3.12: Number of daily observations in Wi-Fi and GPS dataset from December 2018 to January 2019

For basic comparison Figure 3.12 plots the daily obtained records from Wi-Fi and GPS surveys in the same period with Figure 3.7. Wi-Fi packet data has a sample size approximately 700 times larger than the GPS data in terms of size of records, while only 100 times larger in terms of individual. The GPS data can be found still have peaks when a weekend comes, while the swings are much less than Wi-Fi data. Similar to the conclusion in last section, these statistics suggests that the GPS samples are more informative than Wi-Fi samples for sufficient number of observations. This advantage comes from the passiveness of GPS data that gives us users' footprints over the city, unlike sensing method that limited by sensor coverage.

However, Wi-Fi sensing data has its indispensable advantages. The origin-destinations given by Wi-Fi sensing are more reliable. Only people who come within tens of meters of a sensor have a chance to be recorded, and the sensors are mostly deployed at "key" positions for its target places, such as at a checking gate of train stations or at the entrance of a tourist attraction (see Appendix 2) geological attributes associated with Wi-Fi sensors are expected to be more reliable. A Wi-Fi record gives relatively confident information of that people captured by sensors inside sightseeing spot must have visit the spot. For a GPS record, if a person is observed close to the station, there is no a simple way to know if he/she is visiting the spot or just passing by. And if the person is an iOS user, the chance of passing by further grows.

Also, backed by a huge sample size, there are still about 100,000 daily trips with observation times longer than six hours every week in Wi-Fi data, which be considered beyond the reach of sampling via location-based service. Such a large number of observations in a short period of time contributes to capture the seasonal dynamic of tourist flow over the city at low cost.

3.4 Summary

This chapter first reviews the status of tourism in the Kyoto City and the current methodology for the investigations of tourists flow in the city. With the limitation and drawbacks of existing survey, applying the Wi-Fi packet sensing for tourist flow investigations is believed to be an innovative and convincing attempt. Then procedure and outcome of surveys for obtaining MAC-address-matched Wi-Fi packet as electronic footprints are described. It is highlighted the superiority of the Wi-Fi packet

sensing for its ability to collect massive size of sample with considerably low-cost, while the survey error caused by characteristic of the Wi-Fi packet sensing should be address.

As the major data source of the research, the two surveys focus on different scenarios in investigating tourist flows. The Higashiyama experiments have dense deployment of the sensors that provides sufficient knowledge to understand how pedestrian travels in the one-kilometer square, laying the groundwork for modeling route choice in tourism areas. The Kyoto experiments conducted a city-scale survey by monitoring crowds at major public transportation points and famous attractions. The Kyoto experiments obtained even larger size of sample than the first survey. The huge sample size not only give us a clear vision to origin-destinations statistics among each area under monitoring but also a credible the description to seasonal dynamics.

This chapter also discusses drawbacks that found in the Wi-Fi package sensing survey as the technology is not yet fully mature, such as observation missing caused by the penetration rates that sensitive to surrounding environments or the low coverage of monitoring area. Especially in Kyoto City experiment there is a significant negative impact caused by lack of observation per individual on reproducing travel patterns.

To further understand the shortcomings of this technology a commercially available GPS data obtained by location-based service targeting tourists in Kyoto is introduced for comparison. The commercial GPS data is believed to be informative for understanding trips chains for its passiveness. However, the privacy policy and sample size limit the derivation of full vision of city-scale travel patterns based on GPS data.

In line with above discussion, although neither only using Wi-Fi nor GPS data separately is enough for interpreting completed tourist travel patterns at city level, the two type of data appear to be complementary to some extents: GPS footprints provide detailed trajectories over the city, while it might miss the information at where researchers might be most interested in. In contrast, the Wi-Fi data only provides observations from the key locations in the city, but might leave gaps in between.

In Chapter 5 the approach to combine the two datasets and retaining as much as possible of the strengths of each is explored.

Chapter 4 Small-scale, compact sensing experiment: investigating pedestrians in a touristic area of Kyoto City

4.1 Introduction

This chapter proposes and test a methodology for using these Wi-Fi records to understand and predict pedestrian flows, their time spent at different locations and less destination oriented decision-making patterns within a tourist area at a disaggregate level. Data input is collected from the first experiment in Higashiyama Ward, Kyoto mentioned in Chapter 3, which is one of the areas of Kyoto most frequented by tourists. Based on the data, features of individuals are firstly extracted, and clustering method is employed to distinguish the population in the sample. Then according to the clustering results, tourist trips is further extracted. On the other hand the construction of a simplified time-space network with "stay" and "move" links is conducted. The network and link attributes are based on the location of Wi-Fi sensors as well as map data.

This study employs the recursive logit (RL) model originally proposed in Fosgerau et al. (2013) to formulate tourists' route choice behavior. The general concept is to derive the choice probabilities of the next decision points based on the attractiveness of the different downstream link options and whether traversing the link brings one closer to the destination. In addition this study explores an approach to identify whether tourists walk a road without many delays or whether they stroll along the road, possibly stopping at souvenir shops, restaurants, and other attractions. The existence of different kinds of "points of interests" (POIs) is shown to explain route choice as well as whether a person stays on a link for an extended period of time or passes through it in a short time.

The finding in this chapter suggests that this behavior can also be modelled efficiently with the proposed approach, i.e. without relying on defined-in-advance path sets. Though there have been a number of contributions to RL modelling in recent years, while this study is one of the first to apply the RL approach to Wi-Fi-sensor data and to explicitly model "walk" or "stay" without explicit definition of a large time-expanded network.

The reminder of this chapter is organized as follows. Section 4.2 conduct the clustering for observed individuals. Section 4.3 describes the network simplification and construction in order to adopt sensing

data to survey area network. Section 4.4 describes the data preparation for the modelling, and Section 4.5 the model specification for both common network and network with stay- and move-links. Section 4.6 then discusses the resulting choice models for pedestrians' route choices in the target area. Conclusions are drawn in Section 4.7.

4.2 Clustering of observations

Dolniciar and Huybers (2010) discuss the heterogeneity of tourists in their destination perception based on a survey with Australian data leading to questions as to whether a city should target specific tourist groups. With the example of Heidelberg Freytag (2010) further discuss the economic benefits of tourists conducting more activities in a city. In line with this, the focus of this section is on the activities of tourists while walking within an attraction area as well as to preliminary distinguish the population group in the Wi-Fi sample set.

In this section the two-step clustering in SPSS is employed because the two-step approach is able to handle categorical variables and continuous variables in one model by using a likelihood distance to measure the distance.

Aiming at understanding the basic properties of population groups inside the sample, only limited variables is introduced into the cluster model. Also, to allow some generalizability, only a small amount of knowledge from the map is introduced as features for the clustering. The proposed approach is free from site specific characteristics (such as e.g. introducing a specific variable as to whether a tourist visited Kiyomizu temple, the most famous attraction in the case study area) so that it can be applied on most of the Wi-Fi trace samples collected within touristic areas.

Other variables introduced are described in Table 4.1. In particular, assuming some tourists are more exploring the area rather sticking to a shortest path, variable φ_n in Equation (4.1) is proposed as the measurement of how much detour an individual has taken during his/her journey.

$$\varphi_n = \frac{\sum_{i=0}^{I_n-1}(l(i,i+1))}{l(0,I)} \qquad (4.1)$$

where $I_n = \{i_n\}_{i=0}^{i=I}$ denotes the observed tour of an individual in a day as a sequence of the observed sensors i = 0 to I. l denotes the shortest distance index by two observations. Hence the denominator expresses the sum of actually traversed distance and the numerator the shortest distance between the first and last point where a tourist has been recorded. If a tourist was observed to make several tours in the area, as input feature the largest φ in a day is used as this is consider to be better represent the general behaviour of this person. In addition to φ a dummy variable is further introduced to give higher weight to those who have taken the shortest routes when heading for their destinations. For the computation of travelling distance we load the network attributed with route distance of the survey area by OSMnx (Boeing, 2017) and obtain the shortest path between sensor-nodes distances by Yen's algorithm (Yen, 1971).

From the perspective of time, the maximum duration of staying in the area in a day is measured by hour. Since the first/last observation does not fully represent the specific time an individual enters or leaves the survey area, this variable is rounded to nearest integers. Finally, a dummy variable representing if the individual is observed in only a single day or not is set as an indicator that help us distinguishing day-trip tourists.

Table 4.1: Variables of two-step clustering

VARIABLES	DESCRIPTION
LARGEST Φ IN A DAY	The largest value of daily φ obtained as described in Equation (4.1).
SHORTEST PATH USER DUMMY	Equal to one if the observed largest φ in a day of the individual is one, i.e. the shortest route is taken.
LONGEST OBSERVED DURATION IN A DAY	The longest observed duration for the individual in a day.
ONE DAY DUMMY	Equal to one if the individual has been observed only in a single day.

Data collected during the first week of the Wi-Fi survey is used as input. Pre-processing is required before the data enters classification. Those individuals who are observed by only one or two sensors are removed. These observations have very unique features: their observed durations are either

extremely short (less than a second) or extremely long (last for days). Both is presumed as noise in the data collection. It is inferred that the extremely short durations are caused by the same packet request captured by two different sensors, and the extremely long durations come from some fixed machines with Wi-Fi function such as routers. In addition, those individuals with an observed duration of less than 15minutes are removed as they are considered as inadequate observations. These might be pedestrians passing only the edge of the survey area or records from vehicles passing by the area without major stops. About 15% of individuals with 58% of records remain in the sample after passing these pre-possessing filters.

Results of the clustering is shown in Table 4.2 in descending order of percentage share. After testing five is selected as the appropriate number of clusters considering interpretability. The clusters are labelled as “fast tourist”, “shortest path user”, “slow tourist”, “resident/passing-by commuter” and “staff”, respectively. The “fast tourist” group takes over half of the samples. They are only observed in a single day and on average stay in the area for around two hours and take a certain amount of detour. Compared with the fast tourists, the “slow tourists” on average take more than three times of detours measured by φ, and their duration of stay in the survey area is also approximately to be three times longer than that of the fast tourist cluster.

The significant characteristic of the “shortest path users” is that they only use the direct path as the name indicates. These groups are inferred as possibly tourists who directly head for the major attraction of this area. (On closer observation, the study acknowledges, that in this cluster are further some, though only few, likely erroneous observations that pass the above described filtering. These are persons staying a long time in the area that were, however, only observed at very few sensors and hence might appear to be shortest path users.)

The remaining two groups of population do not behave like tourists. The forth group stay multi days in the area, while most of them take shortest paths when travelling, which can be inferred to be commuters who show up in commute hours or citizens living in the survey area. The last group is found characterized by taking a large number of detours as well as staying inside the area for more than half a day. They are considered as staff, such as security and police in the area.

To validate the clustering results, the pivot table of distribution of observations in percentage indexed by sensor and cluster is shown(Table 3). For the first three groups, they are found centralized on the route of sensors 7-14-15. This is the major route for tourism that starts at the major parking lot in the area and reaches the Kiyomizu Temple by passing Sanneizaka. These groups also visit sensor 13 relatively more often, the Nineizaka. Meanwhile the last two groups show up more on the route of sensors 1-6-10-19-20, the "Higashi-omichi", a longitudinal trunk line of Kyoto City. Remind that in the clustering the approach does not introduce "map knowledge" except than a dimensionless variable of φ, so that it is concluded that the clustering performs well showing how the groups distribute with different geographical characteristics.

In conclusion, through the clustering analysis finds that the pedestrians who enter the survey area with some kind of purpose (not just passers-by) may behave different, while most of them (more than 80%) are believed to be tourists. Among the observations labelled to be tourists, some prefer to take detours, others do not. Also, those who prefer to stay for a long time tend to take certain routes, while others, who are considered to leave the area as soon as finishing the visit major spots, are more observed on other routes. It is believed that this knowledge can also help, among others, with understanding the likely shopping behaviour of the different tourist groups. In section 4.6.4 this understanding is aimed to deepen with the following route choice model.

With part of the individuals that is removed in pre-processing still need to be consider as pedestrians, The sample is inferred to include 1) walking tourists 2) passing-by residents 3) staffs and 4) other "noises" that should be excluded for further modelling such as routers or vehicles. Since the clustering method is not entirely deterministic, for further behavioral modelling, a rule-based filter is built based on findings in this section to reduce noise and include every kind of tourists as many as possible.

Table 4.2: Results of two-step clustering

Cluster	**Fast tourist**	**Shortest path user**	**Slow tourist**	**Resident / passing-by commuter**	**Staff**
Share (size)	52.8% (77929)	18.0% (26564)	13.6% (20069)	10.6% (15632)	4.9% (7295)
Variables	**Variable center (categorical variable share)**				
Is a one-day-trip	1 (100%)	1 (100%)	1 (100%)	0 (100%)	0 (99.6%)
Observed duration in a day	1.5	2.25	4.46	4.97	13.54
Largest φ	2.66	1	9.48	3.08	10.76
φ is one	0 (100%)	1 (100%)	0 (100%)	0 (100%)	0 (100%)
Silhouette measure	0.50				

Table 4.3: Distribution of observations by sensor and cluster

Sensor ID	Fast tourist	Shortest path user	Slow tourist	Resident / passing-by commuter	Staff
1	1.8%	3.9%	2.4%	5.0%	6.4%
2	2.8%	5.5%	2.9%	4.6%	4.1%
3	1.8%	4.8%	1.5%	2.2%	2.3%
4	1.6%	2.2%	1.5%	2.1%	1.9%
5	1.2%	1.6%	1.6%	1.8%	4.8%
6	3.5%	5.6%	5.1%	8.5%	11.5%
7	16.1%	13.0%	13.5%	8.1%	4.4%
8	3.3%	3.6%	3.8%	4.2%	3.6%
9	2.1%	2.4%	3.3%	4.2%	3.5%
10	4.7%	6.6%	6.2%	9.5%	13.6%
11	1.9%	1.9%	2.5%	3.2%	4.2%
12	2.5%	2.7%	3.5%	2.5%	1.4%
13	4.5%	5.0%	5.2%	3.7%	2.0%
14	14.8%	9.4%	14.3%	8.6%	6.9%
15	13.6%	11.5%	12.0%	6.2%	3.5%
16	5.1%	3.4%	3.9%	3.4%	2.0%
17	4.5%	3.2%	3.6%	3.5%	4.8%
18	2.1%	1.2%	2.0%	1.6%	2.7%
19	4.8%	5.8%	5.4%	8.7%	9.8%
20	7.5%	6.9%	6.1%	8.5%	6.6%

4.3 Network construction

As the data collection experiment has a series of fixed locations where is able to observe persons but does not have sensors at all junctions, the network needs to be transformed into a reduced one that reflects the observed decisions. As illustrated in Figure 4.1, nodes in this reduced network are the sensor locations, and the link attributes are concluded by set of links associated to pairs of nodes. Furthermore, dummy nodes reflecting that a person has reached a destination are added. The study highlights this approach for its difference to studies based on GPS data where usually more complete paths are observed and the real network can be used.

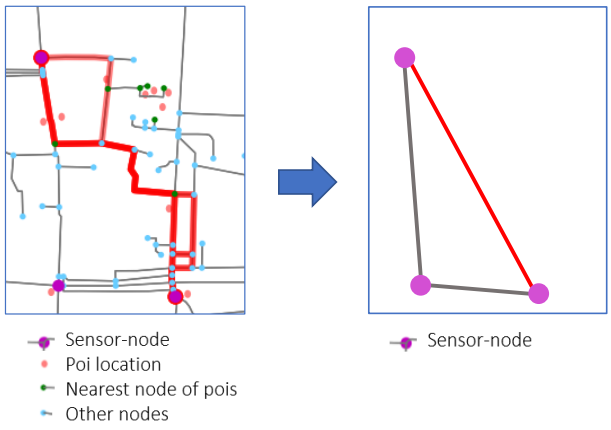


Figure 4.1: Illustrate example of transforming actual network to reduced network

The reduced network is constructed based on the data retrieved by OSMnx (Boeing, 2017) from the OpenStreetMap (OSM) database provided by OpenStreetMap contributors (2015). Points of interests (POI) information provided by Google Map (2020) is integrated to the network, as the POI records in OSM are found incomplete for the target area. Individuals' route choices are then extracted from Wi-Fi packet data referring to the reconstructed network.

4.3.1 Network extraction

The original OSM network for the area of interest contains more than 900 nodes. The network is simplified into a directed graph by only picking up 20 nodes associated to the location of the sensors. Then 118 directed physical links are generated describing the connectivity inside the target area

between “neighboring” sensor nodes, that is sensors that can be visited without passing another sensor. All the connected nodes are bidirectional connected.

For each link its “attractiveness” is defined with a set of attributes. These are length, return penalty, a “Kiyomizu Temple dummy”, a trunk road dummy, and number of different types of POIs. The variables are obtained by analysis from the actual, full network. Table 4.4 summarizes the description of the attributes.

Table 4.4: Link attributes

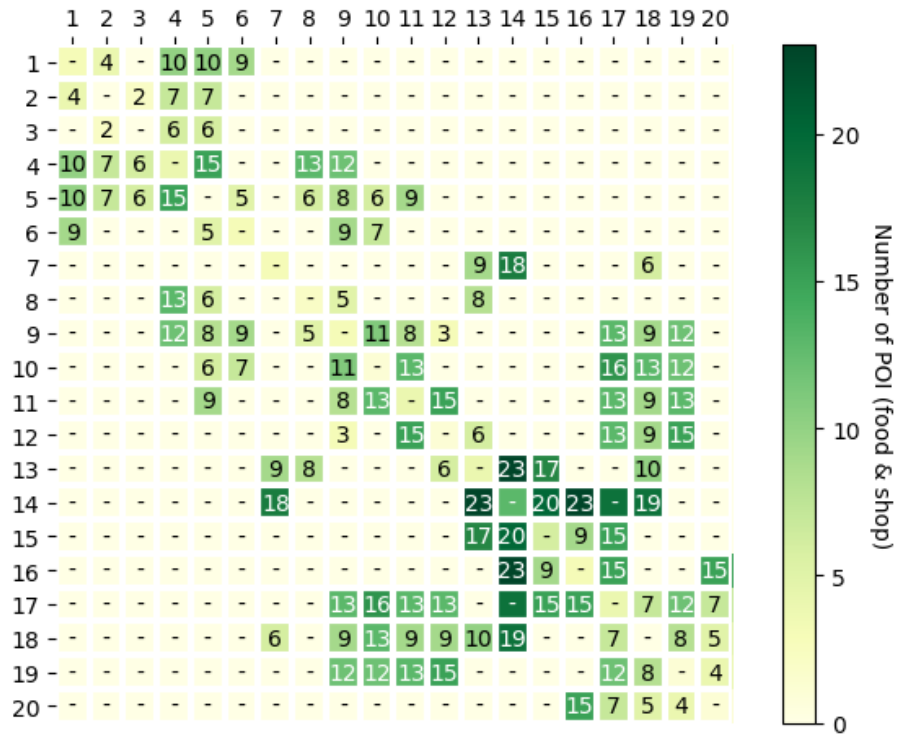
ATTRIBUTE	DESCRIPTION
LENGTH	Distance of shortest path between two sensor-nodes
RETURN PENALTY	Equal to one if the next visited sensor is the same as the one visited immediately before the current one.
NUMBER OF RESTAURANT AND SHOP POIS	Number of POIs, such as restaurants, convenience stores and markets
NUMBER OF SIGHTSEEING POIS	Number of POIs, such as parks, museums, shrines and temples
KIYOMIZU DUMMY	Equal to one if an entrance to Kiyomizu Temple is on the link
TRUNK ROAD DUMMY * LENGTH	Interaction of link length and if the link is associated with the trunk road
STAY CONSTANT (ONLY IN STAY NETWORK)	Equal to one if a link is a Stay-Link

Firstly, even though tourists might be “strolling”, it is expected that the further the distance, the less likely tourists take this route. To obtain the distance between sensors a shortest path algorithm is utilized. Arguably one could also use a set of paths constructed on the full network and weigh these according to the flows. However, since the experiment does not have other pedestrian flow data that show the routes between the sensors, weights cannot be calibrated.

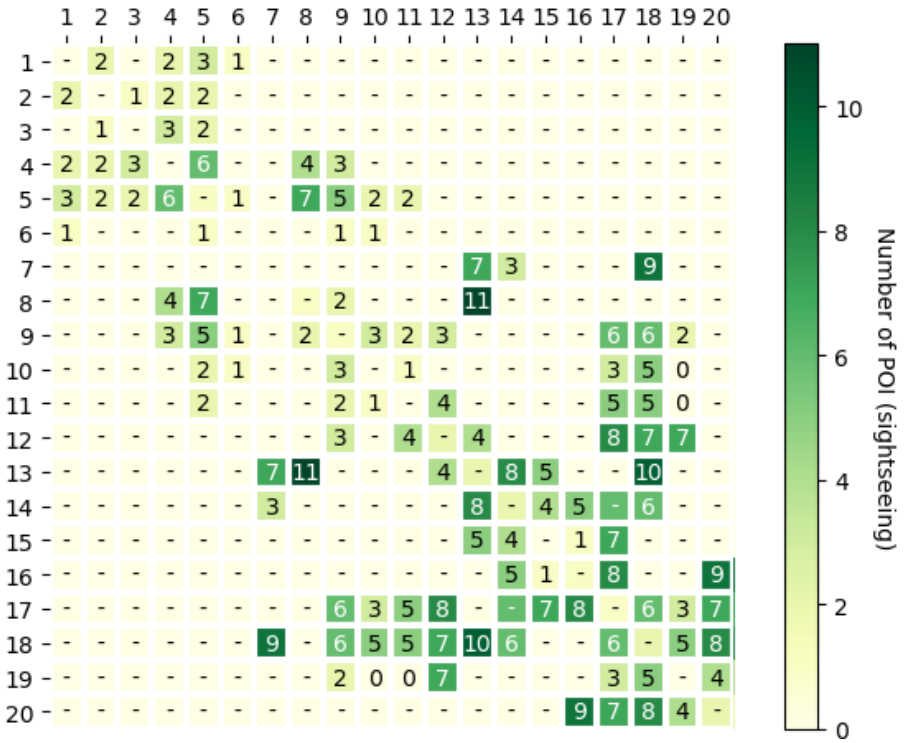
A “return penalty” is further included as one is expected that to be not likely to return on the link one has just traversed. This variable, that is only useful in this study’s context of sequential decisions, has also been used in other recursive logit model formulations (see Zimmermann et al 2017). Furthermore, due to the resolution of Wi-Fi traces, it is difficult to create detailed attributes such as left turn / right turn to reflect how "simple” a route is, which is an attribute that has been found useful in other studies to explain pedestrian route choice as discussed in the literature review.

Besides distance and route directness, the POIs indicated in Figure 3.4 are considered as important decision criteria. In general it is expected that the more POIs, the more a person is attracted to take this link even if his main destination is a different POI. To transform these POIs into link attributes for the reduced network, following framework is utilized. Firstly, the map nodes within a vicinity of 10m are grouped into a single node. Between all neighboring sensor nodes a set of *k*-shortest paths with k=8 following Yen’s algorithm (Yen, 1971) on the full network is created. Then, for each POI following actions are performed: The POI is associated with its nearest node in the full graph. If this node is the end-point of a cul-de-sac, then it is associated with the nearest non-endpoint node. Finally, the POI is associated with a link between two sensors (i.e. a link in the reduced network) if the POI node is on any of the afore identified *k*-shortest paths between these two sensor nodes.

The very detailed POI information provided by Google Maps allowed us to find and distinguish over 20 kinds of touristic POIs located in the target area. To avoid an excessive number of variables as well as correlation issues, the POIs are divided into two main categories, Restaurants & Shops and the Sightseeing Places. The distribution of these POI variables by node pairs is shown in Figure 4.2.



(a)



(b)

Figure 4.2: POI counts distribution over node pairs. a) Restaurant & shopping POIs; b) Sightseeing POIs

Since Kiyomizu temple has such a dominant role in the network and is the main target of most tourists, this POI is treated differently. A dummy variable equal to one is included for the links with Sensors 15 and 16 as head or tail node. Finally, “Higashi-omichi”, one of the trunk roads in Kyoto City, as well as the only trunk road in the survey area is represented by another dummy variable. This trunk road dummy is further interacted with length to give larger weight to longer trunk road sections. Other available information such as average slope on links were further tested but found to be insignificant.

4.3.2 The “Stay-Link”

This study aims to obtain also whether a person is performing an activity on a link, such as eating out, shopping and/or visiting a sightseeing spot. In order to do so, between each connected node a “Stay” connection is introduced. The “Stay” connection is said to be taken by the tourist if s/he spends a significant more time to reach the next detection point than would be required if one would keep moving. By doing so, alternatives of “stay on current link” can be created. Each stay link includes a constant to make it distinct from afore defined links which this dissertation hereafter refers to as “Move-Links”.

4.4 Tourist extraction

In this and the following sub-section the extraction of appropriate observations for the modelling is discussed. Firstly, the attempt to extract walking tourists followed by assembling link sequences that include “stay” and “move” from the records is described.

A sensor will record not only electronic devices carried by pedestrians, but also pick-up probe requests from other moving or installed devices, such as printers or routers. The fixed devices that repeatedly emit probe requests can be easily distinguished. To further distinguish residents from tourists, all the MAC addresses that show up more than 4 days in the data is removed.

Occasionally it is observed that the same MAC address almost simultaneously at more than one sensor. This occurs if the devices are in between the sensors and the signal strength is strong enough to be recognized by both sensors. To control for this, a 60 second constraint is introduced, that is, among

two detections at different sensors within one minute, only the detection with highest signal strength is kept. The threshold of 60sec appears appropriate as it is too short for walking tourists to travel between two sensors. It also means that fast moving vehicles on the trunk road are excluded. For example cars being detected at sensors 1, 6, 10 (see Figure 3.5) within one minute, will hence be only stored once, and since MAC addresses recorded only once are excluded, these cars will not enter the dataset. To note is that car traffic is not prohibited in all parts of the target area, and this approach cannot distinguish slow moving cars from pedestrians. However, pedestrians clearly dominate and the traffic volume remains very low during the whole day inside the area.

4.4.1 Definition of pedestrian routes

For reminding, in contrast to GPS traces, Wi-Fi sensing only observe tours within a predefined area, i.e. do not know where tourists come from and go afterwards. This study therefore sets the destination as the dummy node associated with the last detected sensor.

It is further necessary to prevent two observations that are too long apart from each other from being considered as one route, as otherwise some records with large time blanks such as commuter or workers coming and returning to their home or work place would be recognized as one route. However, introducing too short time limits between two detections would break routes apart, due to, for example, longer sightseeing. As this is contrary to the objectives of modelling “stay” and “move” together, a relatively long timeout limit of six hours is set for subsequent detections so that not many routes are broken apart.

The data cleaning and route extraction process can be summarized as follows:

1) Sort Wi-Fi packet records by time ascendingly;
2) Remove consecutive location-duplicated observations except for the first occurrence;
3) If there is a bunch of observations recorded within 60 seconds, keep only the one with highest signal strength.

4) For each observation, if the time interval until its subsequent observation is more than 6 hours, set the current location as the end of the current route, and start a new route from the location of subsequent observation.
5) Check if the locations of an observation and its subsequent one are physically connected. If not, set the current location as the end of the current route, and start a new route with starting point as the subsequent observation.
6) Repeat Steps 2 to 6 until all rules are valid.
7) Finally, all tours that contain less than four different observations and that are not starting within the time period of interest are removed.

Regarding the minimum route length in the final step, four is chosen as this guarantees that users enter the tourist area and not just stay on the main road. Lower thresholds are also tested and to note is that they lead to similar parameter estimates and even higher model fits. Nevertheless omitting too short routes is considered appropriate as doing otherwise would mean including persons not walking through the area but just passing along the edges.

The distribution of duration and visited sensors of the extracted routes after dropping individuals with fewer than four observations are shown in Figure 4.3. 27% of the routes passed four sensors, and the number of samples decreases gradually with the number of sensors passed. The distribution is monotone falling for number of sensors but not in terms of detection duration. 24% of the routes end within half hour, while the second largest group in terms of duration is 60-90 mins. This is explainable as engaging in sightseeing activities in the area likely leads to stays longer than an hour. Those devices detected less than 30min are likely those that just pass through the area without major stops.

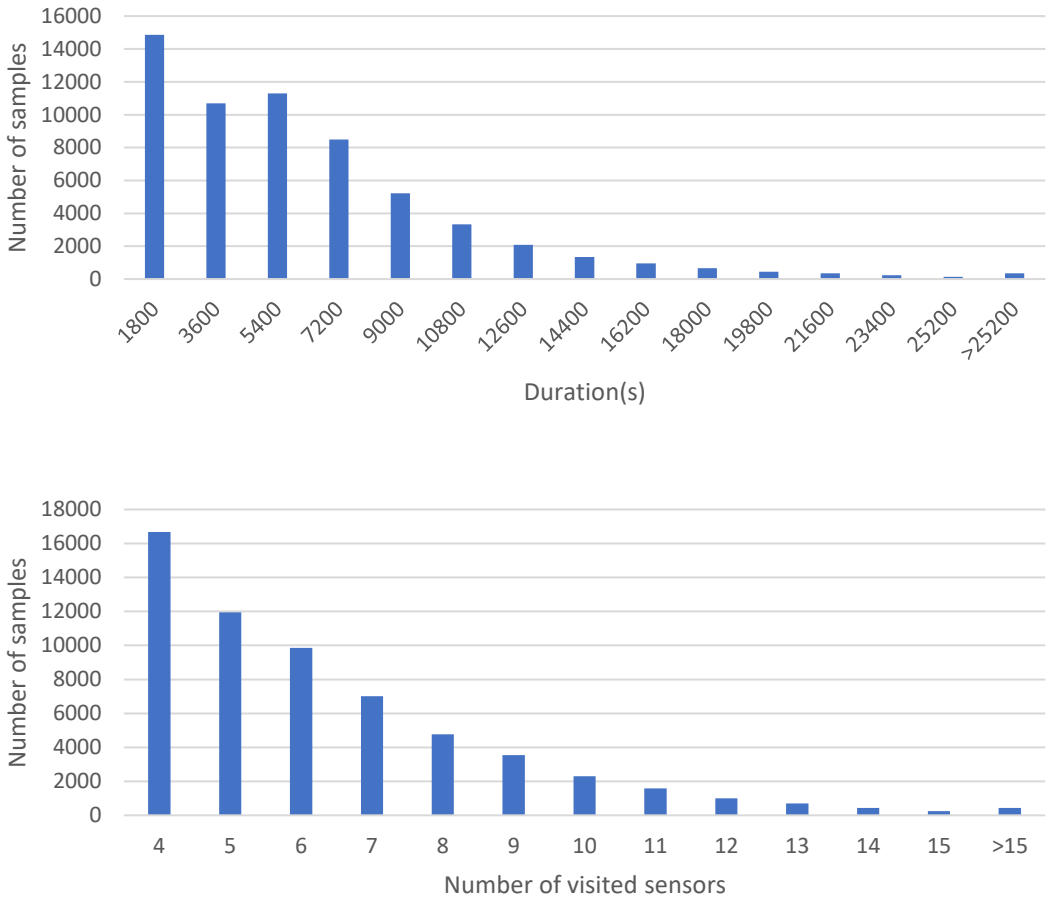


Figure 4.3: Distribution of extracted tours' length measured by duration (upper) and visited sensors (bottom). Tours with less than four links are omitted as these are presumed to be incomplete observations or tours without touristic purposes.

4.5 Recursive logit model with POI counts

4.5.1 The RL model with Stay-Links

The route choice is formulated by a sequential choice model where tourists make decisions on which link to visit next when they pass any not-absorbing link on the reconstructed network.

As discussed in the literature review, the behavior of "staying" within an RL framework has recently been modelled in some contributions by a time-space network representation. This means the physical network needs to be expanded by a factor corresponding to modelling periods / modelling interval times. The complexity of the network directly impacts the computational time and the estimation

performance. Addressing this problem, Zimmerman et al. (2018) restrict alternative links by additional time constraints for certain activities such as work, and by setting utilities of impossible alternatives during working hours to negative infinity. To avoid this problem, instead introduce "Stay-nodes" and associated "Stay-Links" are introduced to the network. Furthermore, the estimates of Stay-Link and Move-link specific variables provide then a clear comparison of how a variable influences the choice between "move" and "stay" for an easily changeable time period.

The reconstructed network and the extracted routes are adjusted for applying a RL model. Twenty dummy destination-nodes and corresponding destination-links are further generated as absorbing stages with link attributes all set to zero. To implement the "Stay-Link" concept, and in order to avoid multiple links with the same attributes between the same nodes which possibly disturb the estimation, "Stay-Nodes" S_{ij} are introduced in between all sensor-nodes i and j which are also connected by a "Move-Link". The resulting two Stay-Links between the sensor nodes and the Stay-Node are assigned each half of the attributes of its corresponding Move-Link. Further, for any link $\{i,j\}$ connecting sensors i and j, its return penalty exists on both of its reverse link, i.e. the links connecting j and i as well as j and S_{ij}. Figure 4.4 illustrates the concept of Stay-Nodes and Stay-Links in the reconstructed network. With this, there is a total of 354 links (118 Move-Links, 236 Stay-Links). Each link has on average 10.06 subsequent links with a maximum of 21. Different thresholds of what defines "Stay" are tested, as will be discussed in the case study.

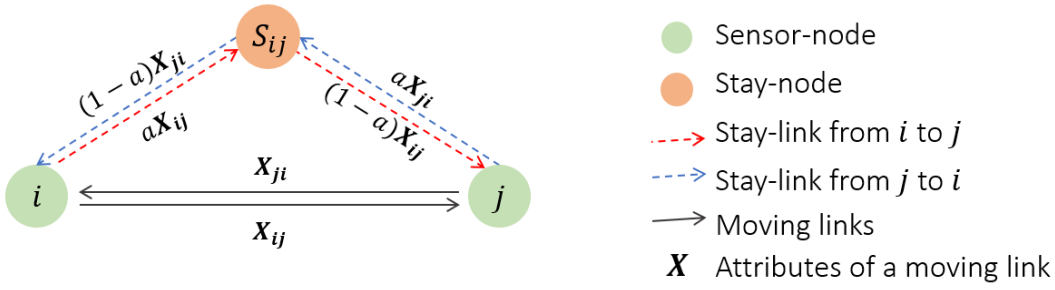


Figure 4.4: Concept of Stay-Nodes and Stay-Links

4.5.2 Data input

As shown in Figure 3.5 clearly the main period of tourist activities occurs during the day, however, also in the evening hours restaurants are frequented by some tourists. In order to compare the difference between touristic peak and off-peak hours, the sample from 8:00 to 18:00 and from 18:00 to 0:00 are estimated.

Data collected during the first week of the Wi-Fi survey is used to extract routes. A sample without stay choices included are first created, and then based on this sample, three samples with stay choices are generated based on stay-thresholds of 15, 30 and 60 minutes. Basic statistics of these samples are presented in Table 4.5. As expected the routes during the evening times tend to be shorter and include less stay activities. (Note that the summation of the number of routes of the two time periods is larger than the total routes due to routes starting before 6pm but extending after 6 being included in both partial sets.)

In addition, the routes identified by clustering analysis in section 4.2 are also input into the model. Models over three samples are estimated, the complete sample, the fast tourist group and the slow tourist group. For the sake of brevity, in the dissertation, only from 8:00 to 0:00 with a 15-minute stay threshold are presented. The statistics of samples by different tourist groups are shown in Table 4.6. Note that after this filtering, only 1% (128/26564) of the routes in the shortest path user group are left in the completed sample, confirming the suspicion that there are a large number of uncompleted observations in this group.

Table 4.5: Basic statistics of samples

	Non-Stay-Links	15 min	30 min	60 min
	8:00 – 0:00			
Number of routes	60,173			
Mean choice per route (std.dev)	5.34 (2.52)			
Mean stay links per route (std.dev)	-	1.35 (1.20)	0.82 (0.86)	0.37 (0.60)
	8:00 – 18:00			
Number of routes	57,464			
Mean choice per route (std.dev)	5.35 (2.50)			
Mean stay links per route (std.dev)	-	1.36 (1.20)	0.82 (0.85)	0.36 (0.59)
	18:00 – 0:00			
Number of routes	4,669			
Mean choice per route (std.dev)	4.38 (1.80)			
Mean stay links per route (std.dev)	-	0.5 (0.67)	0.33 (0.53)	0.18 (0.4)

Table 4.6: Basic statistics of samples by tourist groups, 8:00 – 0:00, 15-minute as stay thresholds

	Complete sample	Fast tourists	Slow tourists
Number of routes	60,173	36,662	13,282
Mean choice per route (std.dev)	5.34 (2.52)	5.28 (2.91)	6.10 (3.91)
Mean 15-minute stay links per route (std.dev)	1.35 (1.20)	1.36 (1.09)	1.63 (1.36)

4.5.3 Model specification

The bias caused by different penetration rates among sensors is addressed by introducing an error term valued by the natural logarithm of sensor-specific penetration rates into the Bellman equation, as formulated in Equation (4.2), where a^+ denotes the head node of link a and z_{a^+} is the penetration rate of a sensor deployed at a^+.

$$V_n^d(k) = E\left[\max_{a \in A(k)} \left(u(a|k) + V_n^d(a) + \theta \ln z_{a^+}\right)\right], \quad \forall k \in A \tag{4.2}$$

The logarithmic transformation of $\boldsymbol{z}$ has the advantage that the penetration rates along routes becomes additive, negative and their magnitude relationship is maintained. Therefore the bias caused by different penetration rates among sensors without violating the assumptions of the recursive logit model is captured. It has to be acknowledged that this approach only addresses part of the problem, since an assumption has to be made that the observed destination is the true destination, i.e. penetration of all destinations is one. One might be able to address this problem by building an outer loop around the RL probability estimations until convergence between estimated destination specific route choice probabilities, observed routes and penetration rates is obtained.

In the network without Stay-Links, given current link k, the instantaneous utility of choosing link a as the subsequent link is:

$$v(a|k) = \beta_L L_a + \beta_{UT} UT_{a|k} + \beta_{MR} MR_a + \beta_{KT} KT_a + \boldsymbol{\beta_{NP}} \boldsymbol{NP_a}^{\boldsymbol{T}} \tag{4.3}$$

where L_a is length of link a (100m), $UT_{a|k}$ is U-turn penalty associated with current link k and subsequent link a, and NP_{ap} is the number of POIs of type p, in this case $p = \{Food\&Shop, Sightseeing\}$ associated with link a. $\boldsymbol{NP_a}$ is scaled by 0.01 for better performance of the estimator. Correlation between variables is tested but any was found to be significant.

In networks with Stay-Links, the utility function is modified by interacting $\boldsymbol{NP_a}$ with the link type utilizing an indicator ${\delta_a}^{stay}$ that takes one for Stay- and zero for Move-links. In this way the model estimation aims to capture the different attractiveness of the different POIs for spending a significant amount of time on links.

$$\begin{aligned} v(a|k) = {} & \beta_L L_a + \beta_{UT} UT_{a|k} + \beta_{mainroad} MR_a + \beta_{kiyomizu} KT_a \\ & + {\delta_a}^{stay} \boldsymbol{\beta_{NP}}^{\boldsymbol{stay}} \boldsymbol{NP_a}^{\boldsymbol{T}} + (1 - {\delta_a}^{stay}) \boldsymbol{\beta_{NP}}^{\boldsymbol{move}} \boldsymbol{NP_a}^{\boldsymbol{T}} + \beta_{stay} {\delta_a}^{stay} \end{aligned} \tag{4.4}$$

Note that both the route extraction method and utility specification are consistent for networks with and without Stay-Links. That is to say, the network without Stay-Links is equivalent to a network with an infinite threshold for "stay".

As discussed in Fosgerau et al (2013), the probability of observing individual *n* choosing an outgoing link from current link *k* given destination *d* is

$$p_n^d(a|k) = \frac{e^{\frac{1}{\mu}v_n(a|k;\,\beta,\theta)+V_n^d(a;\,\beta,\theta)}}{\sum_{a\prime\in A(k)} e^{\frac{1}{\mu}v_n(a\prime|k;\,\beta,\theta)+V_n^d(a\prime;\,\beta,\theta)}} \tag{4.5}$$

where the value function for the downstream utility V_n^d of link *k* is recursively given by

$$V_n^d(k;\beta,\theta) = \mu \ln\left(\sum_{a\in A(k)} e^{\frac{1}{\mu}v_n(a|k;\,\beta,\theta)+V_n^d(a;\,\beta,\theta)}\right) \tag{4.6}$$

except for the destination where the utility is fixed to $V_n^d(d) = 0$. Let I_n be the number of links traversed by person *n*. Then the probability of observing individual *n* choosing a path $\sigma = \{k_i\}_{i=0}^{I_n}$ is as in (4.6) where μ is a scale parameter that is assumed to be equal to one unless nesting of options is assumed.

$$P_n^d(\sigma;\,\beta,\theta) = e^{-V_n^d(k_0;\,\beta,\theta)} \prod_{i=1}^{I_n} e^{\frac{1}{\mu}v_n(k_{i+1}|k_i;\,\beta,\theta)} \tag{4.7}$$

Thus, the log likelihood function can be obtained from the sequence of link choice probabilities as in Equation (4.8) where (4.7) is utilised in the second equality.

$$\begin{aligned} LL(\beta,\theta) &= \sum_{n=1}^{N}\left(\sum_{i=1}^{I_n-1}\left(ln\, p_n(k_{i+1}|k_i;\beta,\theta)\right)\right) \\ &= \sum_{n=1}^{N}\left(\sum_{i=1}^{I_n-1}\left(v_n(k_{i+1}|k_i;\beta,\theta) - V_n^d(k_0;\,\beta,\theta)\right)\right) \end{aligned} \tag{4.8}$$

4.6 Estimation results

4.6.1 Estimation initialization

Since the value function in this study is not specific to destination, for efficiency of the estimation further employs the decomposition (DeC) method. First introduced by Mai et al (2015), this method let it possible to solve all value functions associated to all destinations in the network within a linear system, which considerably promote the computing efficiency for maximum likelihood estimation of recursive logit model.

The DeC method is reproduced on Python integered with limited-memory BFGS algorithm as an external non-linear optimizer (Liu et al, 1989). Unlike many other logit models that have been developed, the estimation of recursive logit has much to be improved. To obatain stable estimation performance, a "warming up" algorithm is developed to reduce gradient descent failure and improve the speed to converge by finding a reasonable initial guess for maximum likelihood estimation (MLE).

The full network is reloaded and MLE is carried out after an initial guess is obtained by Algorithm 1. The logic of the algorithm is to find approximated solutions of the MLE by reducing the size of the network as well as the number of variables. Based on this "warming up" algorithm. Although this method does not fully guarantee the success of estimation, over 150 model estimations in this study including the cross-validation as will be mentioned later is estimated by this method, and combine with proper variable selection and scaling, the success chance of estimation significantly increased.

Algorithm 1: "Warming up" procedure for Log-likelihood maximization of recursive logit model

1. Remove unchosen links in network;
2. Input the instantaneous utility fromulation $v(a|k;\ \boldsymbol{\beta})$;
3. Set initial guess for each coefficient $\boldsymbol{\beta_0}$ = -1;
4. **Repeat**
5. Derive log likelihood function $LL(\boldsymbol{\beta})$ by $v(a|k;\ \boldsymbol{\beta})$;
6. Execute estimation by Limited-memory BFGS optimizer;
7. **if** estimation failed **then**
8. **if** more than one linear term in $v(a|k;\ \boldsymbol{\beta})$ **then**
9. Remove one variable and its associated parameter from $v(a|k;\ \boldsymbol{\beta})$;
10. **else**
11. **return** "estimation failure";
12. **else if** succeed **then**
13. **if** $v(a|k;\ \boldsymbol{\beta})$ is NOT identical to $v'(a|k;\ \boldsymbol{\beta})$ **then**
14. set $\boldsymbol{\beta_0}$ as estimates $\widehat{\boldsymbol{\beta_0}}$ except for what have been removed;
15. **else**

    //warming up succeed
16. **return** estimates $\widehat{\boldsymbol{\beta_0}}$;
17. **End repeat**

4.6.2 Result overview

Estimation results are shown in Table 4.7. All estimated parameters significantly effect the link utility. The average likelihood per choice in the stay network drops only slightly from the non-stay network's results despite the chocie set being doubled. The raise of the stay threshold leads to an increase in the final likelihood, as a long stay is easier to predict than a short one.

All parameters have reasonable signs. Except for the POI related parameters, estimates in non-stay networks are close to those estimated for networks with stay options, showing a high performance in terms of consistency. Parameter of length estimates are stable across all 12 models, in the range of -1.06 ~ -1.08. From the results the of non-stay network the overall effects of POI variables can be

inferred. Generally during a day both two POI types have a positive effect on the link utility and the link being chosen. Considering only quantity (count), one sightseeing spot's attractiveness is approximately seven times larger than that of a store or restaurant. Further, both variables are more significnat during day-time hours as one would expect.

The inclusion of Stay-Links provides us additional information. Before discussing these in more detail to note is that the POI parameter estimates obtained for the network without Stay-Links fall in between those obtained for the respective parameter estimates of Move-Links and Stay-Links with a tendency of being closer to the Move-Link estimate. This in general confirms the definition of Move- and Stay-Links. That the estimates are close to the Move-Link observation will be partly due to the fact that stay is a much rarer event. Therefore also the Stay-Link constant has a negative sign and its negative effects becomes stronger with an increasing time threshold for defining the duration as "Stay". The estimate reflects, all else being equal, the resistence of choosing to stay than moving, and this resistence grows when the threhold raises.

As expected, all POI related parameters are found to have a much stronger positive effect to the utility of Stay-Links compared to Move-Links. The existence of sightseeing POIs affect the likelihood of a link to be taken positively, whether just in passing or staying. Links with food and restaurant options attract pedestrians to stay on these links rather than to just pass through. Restaurants and shops are observed that add very few utility to move links as one would expect. In fact slightly negative values are observed for shorter thresholds. In other words, those who want to move fast will avoid links crowded with restaurants and shops.

Note that a stay threshold of 15 minutes means that for most links one has to keep walking without major interruption to reach the next sensor within the threshold and hence be assigned to the Move-Link. Therefore the 15minute threshold appears to be most useful if one wants to distinguish those with some interest in the attraction in the area from those who focus on walking towards their major goal. Looking at the effect of different Stay-thresholds, the effects of Sightseeing POIs for choosing to Stay reduces from 17.2 to 5.67 units if the threshold raises from 15 minutes to 30 minutes, suggesting that sightseeing spots also have more positive effects on attracting short term stays. This

might at first appear contradictory but the extracted sightseeing POIs are not only major attractions with entrance fees, but, for example, also memorial stones or “art studio stores” etc, i.e. POIs that one looks at a short time and then moves on. Meanwhile the effect of restaurants and shops on the utility of Stay-Links remains large and significant and only slightly reduces with longer thresholds, showing stable atractiveness to people who consider whether to stay. Overall, these findings can be concluded that support the assumption that the route choice behavior in toursitc areas is less destination-oriented.

To note is finally that the penetration rate parameter is also significnatly positive. This is in line with the expectations of this study. This term captures a “fake” positive impact that a less attractive link appears to be seemingly often chosen, which is, however, only due to a relatively higher penetration rate. To add is that models estimated without penetration error variable lead to essentially similar conclusions regarding the other variables but an overall lower model fit.

4.6.3 Comparison by time periods of day

Estimates for different time periods in Table 4.8 and Table 4.9 illustrate that Kiyomizu Temple shows greater positive effect to choices during peak hours (8-18) than off-peak hours (18-0) in line with the expectation. Parameter of the main road changes in the opposite way, attracting more pedestrans during off-peak hours. Also, restaurants and shops are much more attractive in peak hour as its parameter estimates are much larger than those in off-peak hours regardless of different stay thresholds. These observations are inferred that due to attractions inside the area being closed earlier than restaurants and shops along the main street. Souveniour shops and casual restaurants inside the area generally close around 6pm since the Kiyomizu temple closes its gate at 5pm meaning that by 6pm many tourists will have left the area (see also Figure 3.4 and Figure 3.5).

During off-peak hours, it is found that with the increase of the stay threshold, the importance of restaurants and shops for staying decreases and turns from positive to negative. This effect is, however, not observed for the effect of sightseeing POIs. This is explainable as parks and some of the temple areas remain attractive for walking in evening hours even if they can not be entered. (Running an additional model for late hours was considered, i.e. for those starting after 10pm, when truly all shops

are closed, to confirm this further. However, the sample size was not sufficient for this to model to converge as too many links remained unchosen by any sample.)

4.6.4 Comparison by tourist characteristics

Estimation results are shown in Table 4.10. All estimated parameters significantly affect link utility. Compared with the complete sample, the average likelihood per choice of the fast tourist group raises while that of slow tourist group drops, suggesting that the model captures more of the characteristic of fast tourists.

Most of the parameters have reasonable signs. Refer to the complete sample, link length has more negative impact on fast tourists and less negative effect on slow tourists, similar results are also seen for return penalty. These estimates well express the characteristics of the two tourist groups, suggesting that the fast tourists are more focus on using shorter paths than others while slow tourists are the opposite and care less about return to the former area.

In line with expectation, the existence of Kiyomizu Temple encourages more fast tourists to walk links that lead to the temple, while this positive effect is not as influentail on slow tourists, since the slow tourists by any means walk through the whole area. The trunk road are not that attractive to both tourist groups, as this road are considered more used by residents such as commuters.

POI related parameters are found to have a much stronger positive effect on the utility of Stay-links than on the utility of Move-links. The existence of sightseeing POIs increases the likelihood of a person walking a link, whether just in passing or staying. For choosing to stay, these POIs are most attractive to slow tourists, followed by fast tourists, and less so for the whole sample, in line with expectation. For choosing to only walk through a path in 15 minutes, among the three samples sightseeing POIs are most effective for fast tourists and least effective for slow tourists. This is interpreted as the slow tourists are more tend to stay on every link, while others may only take a brief look at the minor sightseeing places.

The estimates fo stay constants measures the willingness of choosing to stay at current link given other variables controlled. The slow tourists are found to be less resistant to stay while fast tourist is the other way around.

The error term is found by penetration rate estimates significantly less for slow tourist group than others. This is because the observed slow toruists are more possible to be complete observations hence less likely to require for the error term to correct the bias.

Table 4.7: Estimation results 8:00 – 0:00

	Stay-Link network						**Non-Stay-Link network**		
	15min		**30min**		**60min**				
Attributes	**Est.**	**t-test**	**Est.**	**t-test**	**Est.**	**t-test**	**Attributes**	**Est.**	**t-test**
Length	-1.07	-317.08	-1.07	-316.39	-1.06	-314.9	Length	-1.06	-313.91
Kiyomizu Temple	1.48	242.48	1.47	242.02	1.47	241.22	Kiyomizu Temple	1.48	241.87
Main road	0.4	103.24	0.38	100.28	0.36	96.45	Main road	0.36	97.96
Return penalty	-1.16	-173.1	-1.18	-176.35	-1.18	-176.45	Return penalty	-1.18	-176.43
Food & shops (Stay-Link)	12.97	138.53	11.97	99.9	10.14	57.54	Food & shops	0.75	31.69
Sightseeing (Stay-Link)	17.2	53.27	5.67	13.56	5.19	8.52	Sightseeing	12.84	121
Food & shops (Move-link)	-0.97	-35.87	-0.2	-7.96	0.41	16.78			
Sightseeing (Move-link)	12	107.3	13.03	120.21	12.8	119.66			
Stay constant	-2.66	-239.84	-2.89	-215.22	-3.59	-183.95			
Err. by Penetration rate	1.4	144.85	1.44	152.72	1.42	154.19	Err. by Penetration rate	1.39	153.46
Max LL	-466009		-434846		-388676		MaxLogLL	-322028	
Number of routes	60173		60173		60173		Number of routes	60173	
Number of choices	321571		321571		321571		Number of choices	321571	
Max LL per route	-7.744		-7.227		-6.459		MaxLogLL per route	-5.352	
Max LL per choice	-1.449		-1.352		-1.209		MaxLogLL per choice	-1.001	

Table 4.8: Estimation results 8:00 – 18:00

	Stay-Link network						**Non-Stay-Link network**		
	15min		**30min**		**60min**				
Attributes	**Est.**	**t-test**	**Est.**	**t-test**	**Est.**	**t-test**	**Attributes**	**Est.**	**t-test**
Length	-1.08	-311.28	-1.08	-310.54	-1.08	-309.04	Length	-1.08	-307.98
Kiyomizu Temple	1.5	240.47	1.5	239.87	1.5	239.1	Kiyomizu Temple	1.51	239.43
Main road	0.39	96.57	0.37	93.49	0.35	89.73	Main road	0.35	90.07
Return penalty	-1.19	-172.81	-1.21	-176.02	-1.21	-176.08	Return penalty	-1.21	-176.08
Food & shops (Stay-Link)	13.4	140.29	12.72	103.06	11.58	62.71	Food & shops	0.82	33.76
Sightseeing (Stay-Link)	17.83	53.84	5.76	13.27	5.83	9.02	Sightseeing	13.15	119.84
Food & shops (Move-link)	-0.94	-33.99	-0.16	-6.29	0.46	18.36			
Sightseeing (Move-link)	12.38	107.22	13.39	119.73	13.15	118.98			
Stay constant	-2.71	-235.92	-2.98	-211.83	-3.78	-180.58			
Err. by Penetration rate	1.4	139.95	1.43	147.43	1.42	148.92	Err. by Penetration rate	1.4	149.29
MaxLogLL	-442943		-412301		-366828		MaxLogLL	-305268	
Number of routes	57464		57464		57464		Number of routes	57464	
Number of choices	307576		307576		307576		Number of choices	307576	
MaxLogLL per route	-7.708		-7.175		-6.384		MaxLogLL per route	-5.312	
MaxLogLL per choice	-1.44		-1.34		-1.193		MaxLogLL per choice	-0.992	

Table 4.9: Estimation results 18:00 – 0:00

	Stay-Link network						**Non-Stay-Link network**		
	15min		**30min**		**60min**				
Attributes	**Est.**	**t-test**	**Est.**	**t-test**	**Est.**	**t-test**	**Attributes**	**Est.**	**t-test**
Length	-1.06	-74.1	-1.06	-74.14	-1.07	-74.43	Length	-1.08	-74.62
Kiyomizu Temple	1	21.54	1.01	21.85	1.02	22.08	Kiyomizu Temple	1.04	22.41
Main road	0.47	37.92	0.47	38.77	0.48	39.13	Main road	0.48	39.83
Return penalty	-0.73	-26.87	-0.74	-26.96	-0.74	-26.96	Return penalty	-0.73	-26.8
Food & shops (Stay-Link)	-1.48	-1.89	-4.88	-4.82	-7.81	-5.36	Food & shops	0.07	0.56
Sightseeing (Stay-Link)	12.53	7.17	11.93	5.63	6.52	2.22	Sightseeing	10.77	25.05
Food & shops (Move-link)	0.09	0.69	0.17	1.39	0.15	1.23			
Sightseeing (Move-link)	9.92	22.13	10.35	23.41	10.61	24.34			
Stay constant	-2.28	-39.35	-2.49	-35.07	-2.86	-29.65			
Err. by Penetration rate	1.49	38.93	1.44	38.24	1.35	36.51	Err. by Penetration rate	1.25	34.16
MaxLogLL	-24459.1		-23065.5		-21561.8		MaxLogLL	-18807	
Number of routes	4669		4669		4669		Number of routes	4669	
Number of choices	20448		20448		20448		Number of choices	20448	
MaxLogLL per route	-5.239		-4.94		-4.618		MaxLogLL per route	-4.028	
MaxLogLL per choice	-1.196		-1.128		-1.054		MaxLogLL per choice	-0.92	

Table 4.10: Estimation results 8:00 – 0:00, by tourist groups, 15-minute as stay thresholds

	Complete sample		Fast tourists		Slow tourists	
Attributes	**Est.**	**t-test**	**Est.**	**t-test**	**Est.**	**t-test**
Length	-1.07	-317.08	-1.12	-248.79	-1.03	-159.92
Kiyomizu Temple	1.48	242.48	1.62	199.84	1.23	108.39
Main road	0.4	103.24	0.38	70.05	0.37	49.57
Return penalty	-1.16	-173.1	-1.31	-146.18	-1	-82.15
Food & shops (Stay-link)	12.97	138.53	13.63	113.22	12.28	69.04
Food & shops (Move-link)	-0.97	-35.87	-0.93	-26.44	-0.73	-14.05
Sightseeing (Stay-link)	17.2	53.27	17.36	41.16	18.27	29.91
Sightseeing (Move-link)	12	107.3	13.48	89.64	10.02	47.35
Stay constant	-2.66	-239.84	-2.71	-184.7	-2.55	-120.63
Err. by Penetration rate	1.4	144.85	1.44	109.06	1.27	71.38
MaxLogLL	-466009		-272889		-128106	
Number of routes	60173		36662		13282	
Number of choices	321571		193744		81007	
MaxLogLL per route	-7.744		-7.443		-9.645	
MaxLogLL per choice	-1.449		-1.409		-1.581	

4.6.5 Cross-validation

The performance of prediction is evaluated by cross validation at different time periods. By each random drawing, 80% of the observations are used for estimation and 20% of the observations are used as holdout samples to evaluate the log-likelihood loss. Loglikelihood loss per route (Mai et al. 2015) as well as per choice (Oyama and Hato 2017) are used for measurement. Figure 4.5 shows the cross-validation result for the whole sample. The LL loss becomes stable when the cross-validation repetitions increase, showing that the modelling results are independent from sampling.

$$\overline{err_i^r} = -\frac{1}{|HS_i|}\sum_{n\in HS_i}\sum_{j=1}^{J_n-1} ln\, p\left(a_{j+1}|a_j;\widehat{\beta}_\iota\right) \tag{4.9}$$

$$\overline{err_i^c} = -\frac{1}{\sum_{n\in HS_i}|n|}\sum_{n\in HS_i}\sum_{j=1}^{J_n-1} ln\, p\left(a_{j+1}|a_j;\widehat{\beta}_\iota\right) \tag{4.10}$$

where,

$\overline{err_i^r}$: loglikelihood loss per route

$\overline{err_i^c}$: loglikelihood loss per choice

HS_i: Set of the holdout sample

$|HS_i|$: Number of choices in holdout sample

n: Route in holdout sample

J_n: Size of path n

a_j: j th Link in path n

$\widehat{\beta}_\iota$: Estimated coefficients

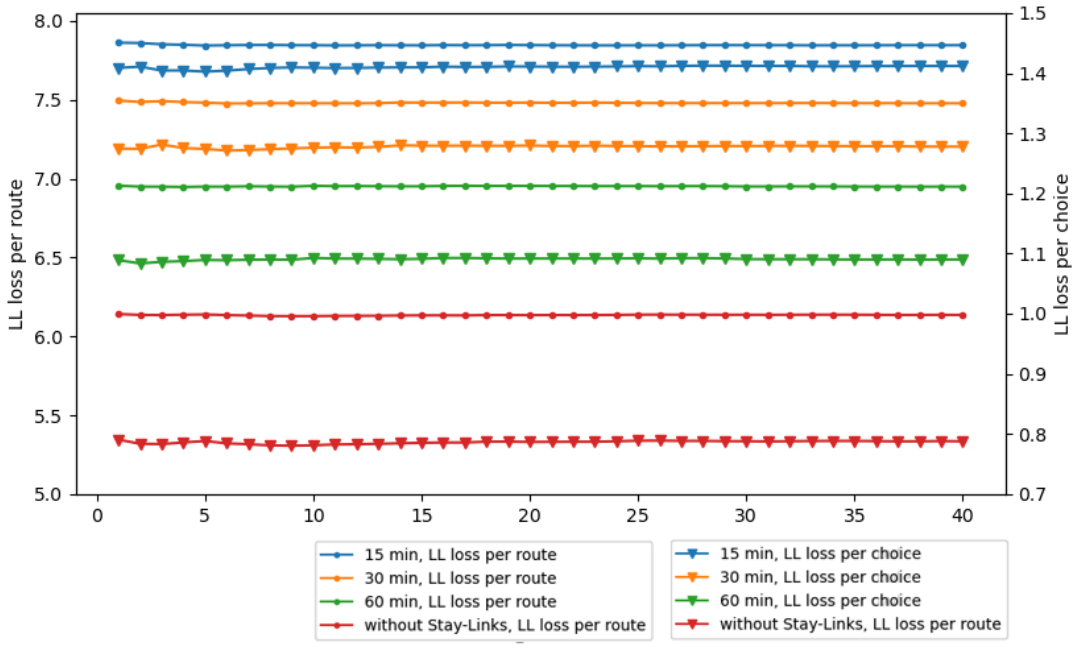


Figure 4.5: Average of test error per choice and per route over holdout samples

4.7 Summary

This study utilizes Wi-Fi packet sensors to model behavior of pedestrians in a touristic area. The research framework can be summarized to Figure 4.6. Pedestrians' routes are extracted from Wi-Fi packet data to identify trips, including when, where and how long a person stays. A network capable of modelling digital traces captured by location specific Wi-Fi packet sensors is abstracted from real geodata. For forecasting, data required contains the model, the map knowledge and ODs. The methodology is able to be applied in larger scale areas where GPS tracking becomes costly or indoor / semi-indoor areas where GPS tracking accuracy is too low.

The recursive logit (RL) model that avoids path enumeration is employed. Though for this problem with 20 sensors enumeration would still be feasible, installation of a larger number of sensors (mainly Bluetooth) in areas becomes more common. To note is that, for example, that in a Belgian shopping mall 56 Bluetooth sensors are deployed to extract spatial-temporal profiles of shoppers (Oosterlinck et al., 2016). Further, the Mall of Asia, the largest shopping mall in Manila, uses 700 Wi-Fi access points and over 600 Bluetooth sensors to guide shoppers (Quain, 2018) and this information could also be used to understand consumer behavior.

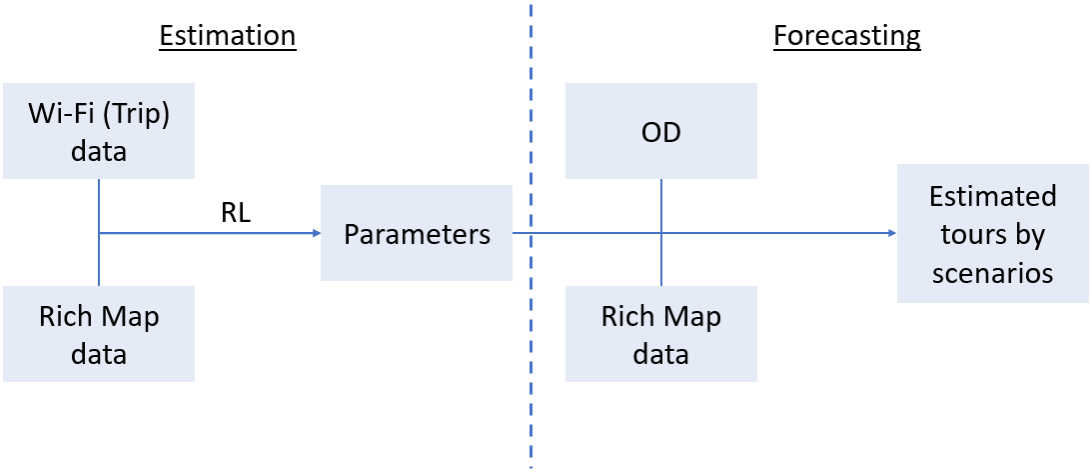


Figure 4.6: Reasearch framework of modelling route choices based on packet sensing data

To capture the decision of making a stop on a link this study proposes "Stay-Links" that create separate connections between nodes. In the modelling the number of two types of POIs is introduced as a variable to reflect link attractions that make a person's travel less destination-oriented. The estimates are consistent and have reasonable signs and magnitude. By assigning these variables to both Move- and Stay-Links the study provides knowledge how different POI types attract tourist flows. The estimates further describe the difference in behavior between touristic-peak / off-peak hours. The estimates can be used for planning as they provide understanding how the addition or removal of "average" or "typical" shops, restaurants or other attractions can influence flows in an area. Clearly "typical shop" or "typical restaurant" is a vague definition but this approach is suggested still meaningful. For example, in touristic areas one can often find a multitude of souvenir shops selling similar items and persons without prior knowledge of the area might not target specific souvenir shops.

Models over three different stay thresholds are presented, which are 15, 30 and 60 minutes, and one without Stay-Links, and conclude that the 15 minute-threshold is good to distinguish whether a person is just walking or visiting some of the souvenir and casual restaurants. If the main purpose is to distinguish persons entering more formal restaurants or spending a longer time at sights, using longer thresholds is recommended. More generally, to choose the appropriate threshold a pre-survey is

suggested that provides information on the average duration to complete the activity of main interest to the planner.

This study closes the discussion on appropriate Stay-threshold with two remarks. Firstly, it should be acknowledged that the stay behavior cannot be perfectly recognized. In particular, the shorter the stay threshold, the more likely “slow” moving will be classified as stay. Too long thresholds are instead not useful to distinguish specific activities. Secondly, also to note is that if the distinction of staying versus moving pedestrians is not important, the non-stay network is recommended. Not only because the model fit is better, but also because the network is simpler and it becomes easier to include additional attributes, such as distinguishing more POI types.

Chapter 5 Inferring City-scale trips based on Wi-Fi sensing with aid of small sample of GPS footprints

5.1 Introduction

As a foundation for further modeling this chapter aims at set up an approach for the inference of daily travel trajectories. So far, Chapter 3 has found that although Wi-Fi and GPS data both have difficulty reproducing individual travels independently, an integration of them may provide further knowledges. Accordingly this chapter proposes a methodology to use the Wi-Fi records to reproduce tourist flows, time spent at different locations over the city at a disaggregate level at city-scale with the assist of GPS footprints.

Firstly both GPS and Wi-Fi raw datasets are converted into a same temporal sequence format presenting the spatial-temporal activities for each individual. Then an similarity evaluation method is proposed to search the similar sequences from GPS sample as source sequence for each Wi-Fi sequence. Following the objective, time-warping is employed. Moreover, fuzzified distances for the measurement is introduced for avoiding the overfitting while keeping the variety of obtaining sequence as well. The resulting source sequence is then used to enrich the defective Wi-Fi trips.

The reminder of the paper is organized as follows. Section 5.2 describe the data alignment. Section 5.3 propose an approach to searching similar trip patterns for the aligned data, presents the achievement of the enrichment and validation. Conclusions are drawn in Section 5.4.

5.2 Data alignment

Aiming at the alignment of these two datasets, the raw logs are turned into a set of location sequences by discrete timespans. Doing so helps understanding the daily travel behavior along time as well as lays the groundwork for further comparison and eventually data fusions.

Denote the number of timespans as T, for observations belong to the same devices in one day, their records become a sequence with T elements (elements can be empty). If more than one observation is recorded in the raw data, the last observation replaces the others. Hence, more or less information from

the raw log is lost depending on the size of *T*, while on the other hand, the conversion can be seen as a hash procedure that provide privacy protection.

Both raw logs of Wi-Fi probe and GPS footprint are converted into space-time forms. Process of the conversion is illustrated in Figure 5.1. As the GPS function in Android has the most stable once-an-hour report cycle among others, this study chooses *T*=24 i.e. each day be divided by one hour to show statistics of the two datasets in space-time format. In this example, the first row is discarded as another observation shows up later in the same time span. Meanwhile, the eleventh element of id_1 is empty since none observation is recorded during 11:00 to 11:59. In addition, during the alignment, individuals who be recorded less than three times have been removed because of not informative for further analysis.

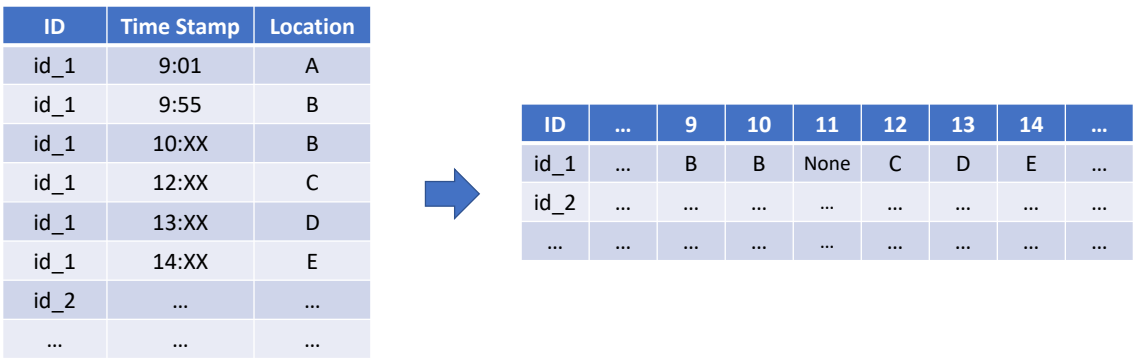


Figure 5.1: Processing raw observation panel log to space-time sequences

Figure 5.2 summaries the daily number of observations after alignment when *T=24*. Horizontal axis is the number of observations in percentage for an individual device in a day. In both samples, the greater the number of records, the smaller the proportion of the sample. For GPS sample, this curve gradually reduces, with more than 50% of the sequence having more than five observations, suggesting good conditions for mapping out daily movements. In contrast, only about 20% of the individual Wi-Fi sample have more than five observed hours, which seems not very sufficient to interpenetrate daily travel patterns for individuals.

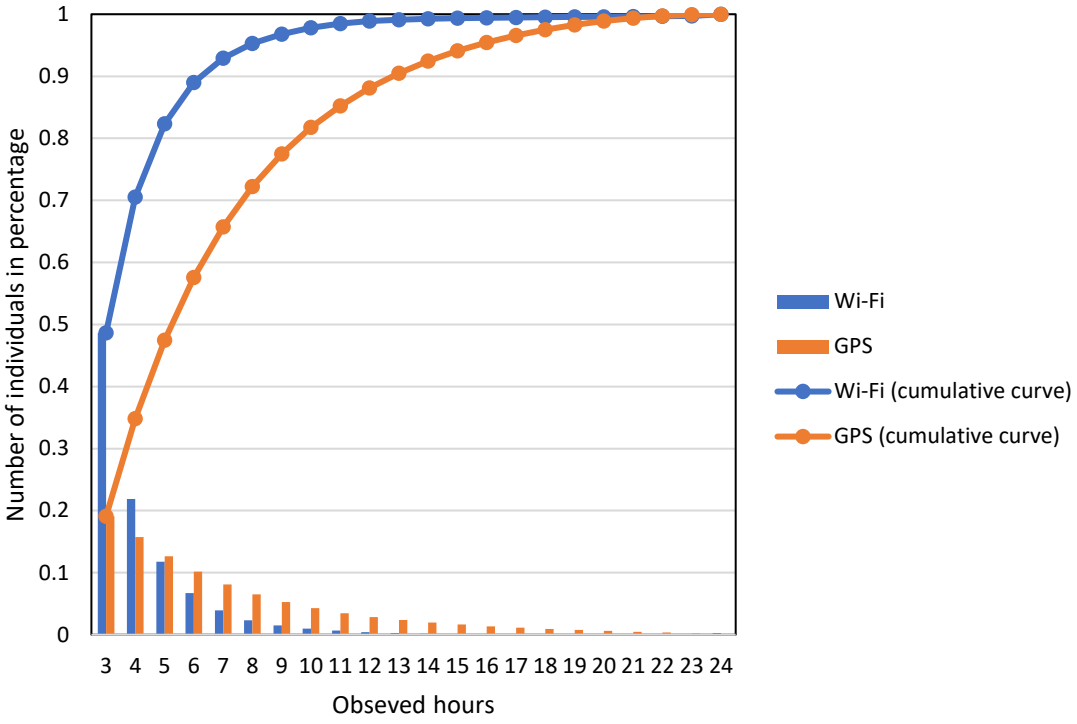


Figure 5.2: Number of observed timespans daily

5.3 Enriching Wi-Fi traces

As explained in Chapter 3, neither only using Wi-Fi nor GPS data separately is enough for interpreting completed tourist travel patterns. An approach is considered to combine the two datasets and retaining as much as possible of the strengths of each. Focusing on a large number of Wi-Fi samples, a method to reasonably enrich the defective Wi-Fi trips with the help of the small number of GPS samples is developed.

5.3.1 General process

Based on the temporal aligned sequence interpreted in last section, a time-warping method is used. Aiming at searching for similar and relatively more detailed sequences to a target trip, firstly the distance between two traces is measured. The illustration of the comparison between a target sequence with missing location and a completed sequence is presented in Figure 5.3. The general logic of the processing is to measure the distances between a target sequence and a source sequence by each

timespan in "comparison windows", which is defined as the timespans corresponding to the observation between the first and the last non-empty observations of the target sequence.

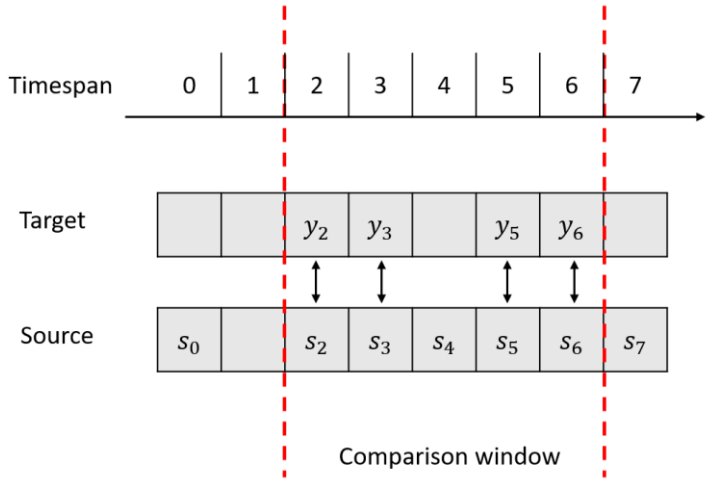


Figure 5.3: Comparison between target sequence and source sequence. The source sequence is required to be completed in the comparison window.

Denote target (to be filled-in) sequence as $Y_t = \{y_t(lat, lon)\}_{t=0}^{T}$, and source sequence $S_t = \{s_t(lat, lon)\}_{t=0}^{T}$, where (lat, lon) is the latitude and longitude of observation. The distance between each corresponded observation is given by:

$$d(y_t|s_t) = \begin{cases} \infty, & s_t = \emptyset \\ 0, & y_t = \emptyset \ \& \ s_t \neq \emptyset \\ hav(y_t(lat, lon), s_t(lat, lon)), & y_t = \emptyset \ \& \ s_t \neq \emptyset \end{cases} \quad (5.1)$$

where $hav(y_t(lat, lon), s_t(lat, lon))$ is the distance calculated by the Haversine formula (Robusto, 1957) which converts latitudes and longitudes into radians approximating the Earth as a sphere with a radius of 6373 km. This operation fixes the distance from a target sequence to a source sequence that is defective in the comparison window to infinity, and ignore the distance from each empty observation in target sequence. The total distance of the two sequence then can be calculated by:

$$D_{Y,S} = \sum_{t=0}^{T} d(y_t|s_t) \quad (5.2)$$

In the application this study sets each Wi-Fi daily sequence as target, and GPS daily sequence as source. However, the coordinates of a Wi-Fi sensor do not directly represent the coordinates of the

associated area in which it is located, therefore, a person who exactly passed the sensors should not have be considered a better match compare to a person who is a certain distance away from the sensor but still in the associated area. This issue is illustrated in Figure 5.4, where the red observation happens to be extremely close to sensors and the blue one does not, while they should have same or at least similar distance to the three sensors associated with same area.

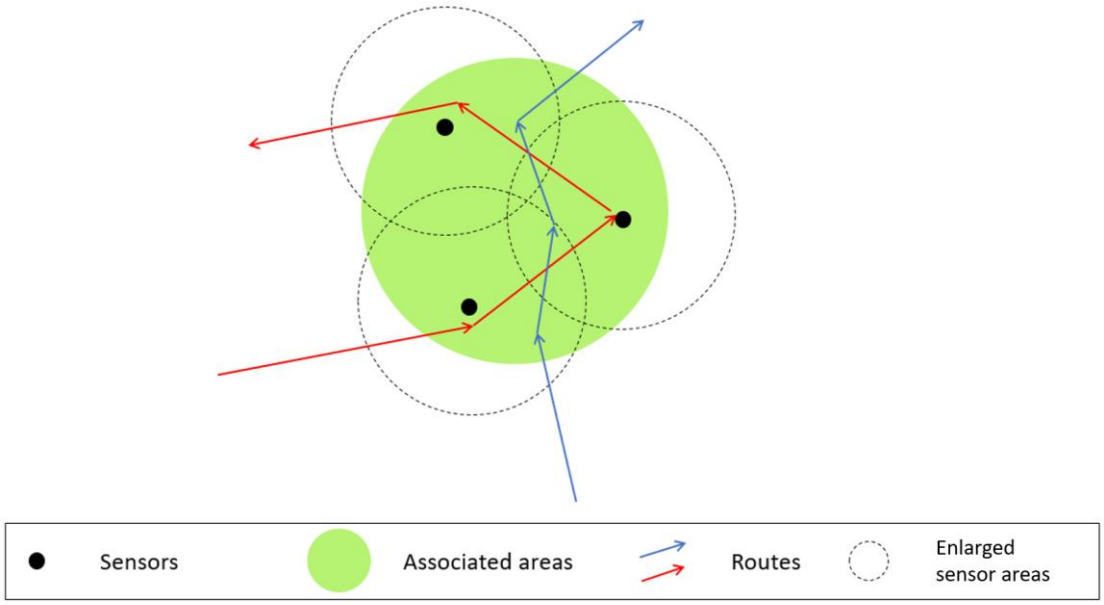


Figure 5.4: Illustration of fuzzy distance. Two route routes in the direction of the arrow have same fuzzy distance to the sensors in the enlarged sensor locations area.

Addressing the problem, the distance d is fuzzified from a GPS record to a Wi-Fi sensor less than a reference distance d_0 to be even. In other words, as shown in Figure 5.4 , all records located in a circle with sensor as the center and radius as d_0 are presumed to have a same fuzzy distance. Denote target (to be filled-in) sequence as $Y_t = \{y_t(lat, lon)\}_{t=0}^{T}$, and source sequence $S_t = \{s_t(lat, lon)\}_{t=0}^{T}$, where (lat, lon) is the latitude and longitude of observation. The fuzzy distance is then by approximated odd ratios given by floor division as in Equation (5.3) and (5.4).

$$w(y_t|s_t) = (1 + \pi d(y_t|s_t)^2 // \pi d_0^2)^{-1} \quad (5.3)$$

$$W_{Y,S} = \prod_{t \in T} w(y_t|s_t) \quad (5.4)$$

where $w(y_t|s_t)$ denotes the weight given by approximated odd ratios of one record at *t* being within a unit of distance d_0. Following Equation (5.3), any distance $d(y_t|s_t)$ is transform to an odd ratio to a reference circle with threshold radius of d_0 . An example is presented in Figure 5.5. Any observation located inside the circle of radius d_0 is weighted by one according to Equation (5.3), such as the red dot or the blue dot. In this way, observations on the green dot which are inside the circle of radius $\sqrt{2}d_0$ (which is twice the area of the circle of radius d_0) to the $\sqrt{2}d_0$ while outside the circle of radius d_0 is weighted by 1/2. And then similarly the orange dot is weighted by 1/3.

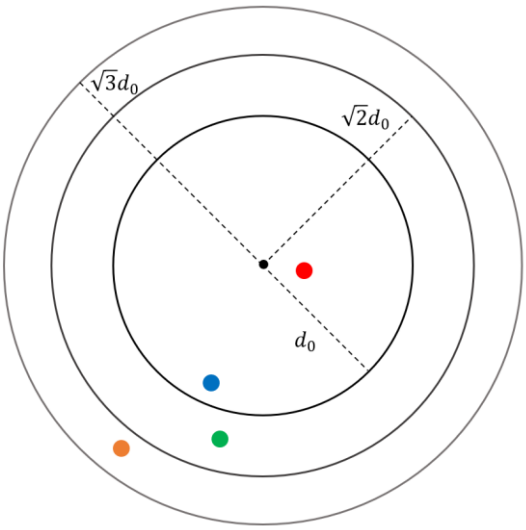


Figure 5.5: An example of fuzzy distance given by approximated odd ratios

Since the Wi-Fi sensor captures much more individuals than GPS application, to keep the variety of reproduced sequences, this study defines the source sequence corresponded to the first 10% percent of high $W_{Y,S}$ of each target sequence as the candidate of supplement source. Then for each target sequence, a supplement sequence is drawn in respect of normalized weights given by $W_{Y,S}$. After the best matched source sequence is picked, the target sequence is used to complete the source sequence as follows:

$$a_t(lat, lon) = \begin{cases} s_t(lat, lon), & s_t \neq \emptyset \\ y_t(lat, lon), & otherwise \end{cases} \tag{5.5}$$

$$A_t = \{a_t(lat, lon)\}_{t=0}^{T} \tag{5.6}$$

To note is that this approach (operations following Equation (5.1), (5.2),(5.3),(5.4),(5.5),(5.6)) is easily transformed into vectorized operations, meaning that the operations are able to applied to the whole set of source sequences instead of one individual sequence by another. This provides large advantages in searching matched sequence in a big dataset, as for each target only one operation is needed to obtain its distance to all the alternative sources.

5.3.2 Fused data

The GPS data is tripled through shifting forward/backward by one hour for obtaining a better sample size for the source sequences. Figure 5.6 shows the number of observed hours in percentage for the integrated data refer to the Wi-Fi and GPS dataset. Number of observed hours of filled-in dataset can be found considerably increased to a reasonable distribution compare to raw from of Wi-Fi sample.

To give an example for effect of the filled-in method Figure 5.7 plots the distribution of destinations from the eight sensors inside Kyoto Station along time. It can be found that before the fill-in processing, population of arriving at JR Arashiyama Station (the last row of the colored matrix) is unreasonably higher than any others. After the enrichment operations, the number of travelers arriving at other popular areas raised.

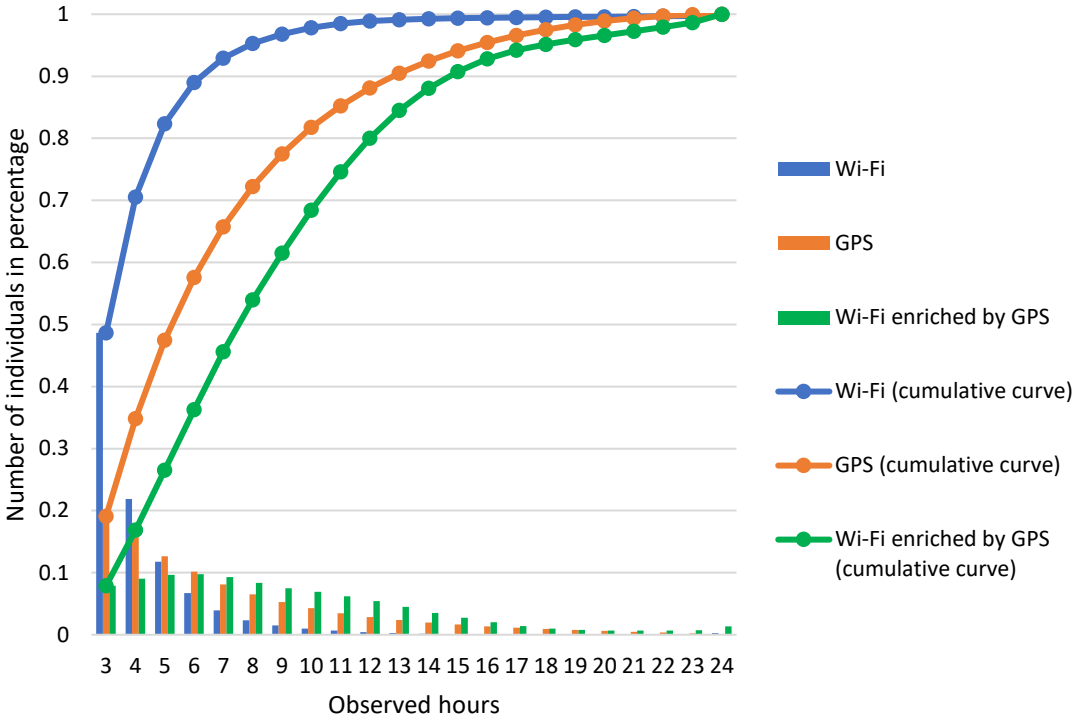
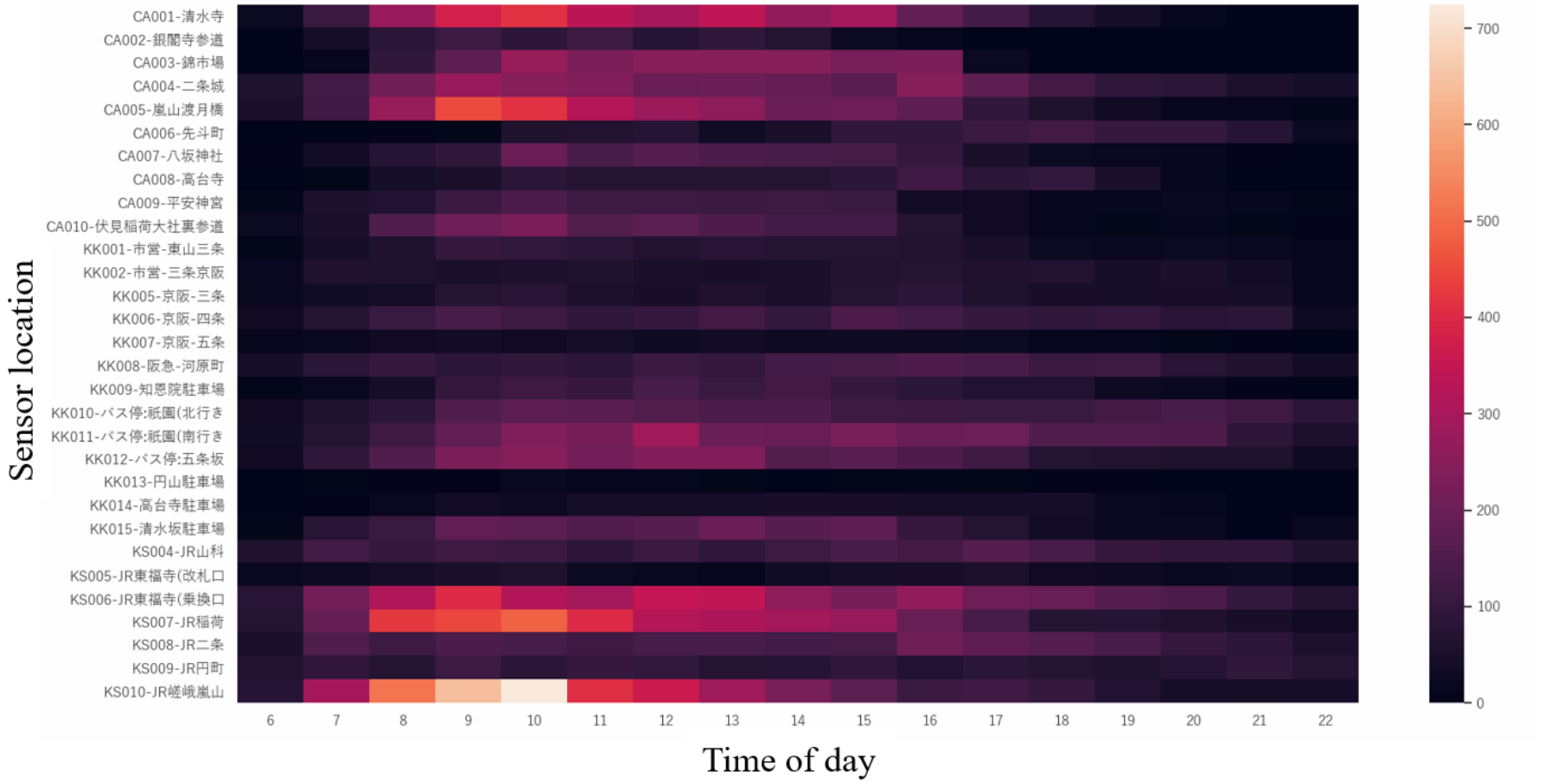
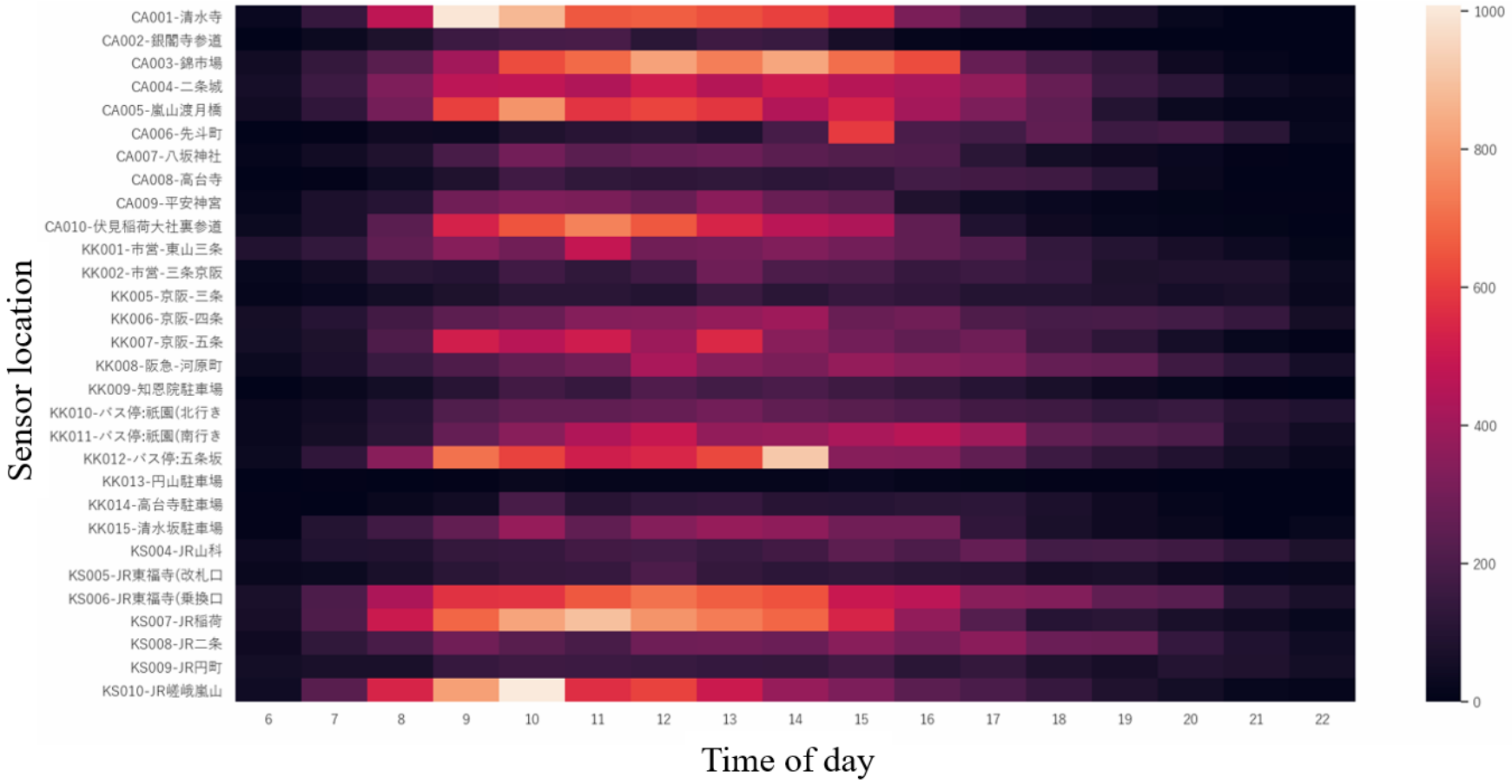


Figure 5.6: Number of observed hours in percentage for the integrated data refer to the Wi-Fi and GPS dataset



(a) before data enrichment



(b) after data enrichment

Figure 5.7: Number of persons arriving from Kyoto Station per hour along time of day, before and after the Wi-Fi traces enrichment

5.3.3 Cross-Verification

To verify the data supplement, the cross-validation method is again employed. 40 rounds of verification are conducted to avoid bias caused by sampling hold-out data from a set of GPS sequences that are completely recorded every hour from 08:00 to 19:00. In each round of verification, 20% of these sequences are picked up randomly as hold-out samples. Then for each hold-out sample a random number of the observations is removed while the first, last and one more record is kept in order to provide a minimum level of information. The removed observations are predicted by the rest of the data, including the remaining 80% of the "perfect" data and all remaining data. The error is measured by:

$$err = \frac{D_{Y,S}}{n_Y} \tag{5.7}$$

where $D_{Y,S}$ is the distance between true trace and the predicted trace calculated by Equation (5.1) and (5.2), and n_Y is the number of predications made for each uncompleted trace. Another measurement projects the true locations and predicted locations to tourism area defined in aforementioned Kyoto Tourist Survey and verify the approach by success rate of area identification. Denote n'_Y as the number of correct predictions, the success rate is

$$r = \frac{n'_Y}{n_Y}. \tag{5.8}$$

The cumulative average of err and r are shown in Figure 5.8. Indeed, the cumulative average of the two measurements being stable when rounds increases. The lower the value the better is the predication performance for both measurements. Generally, a better error measure is found associated with a higher identification rate. A prediction is approximately 0.85km far from the correct location and about 84.2% of the touristic area are successfully identified. No significant bias can be found by the condition of sampling hold-out data. Figure 5.9 shows one of the results after sorting the err values in ascending order. Around 75% of the prediction errors are under 1km yet some of the predictions have very large error (about 10km).

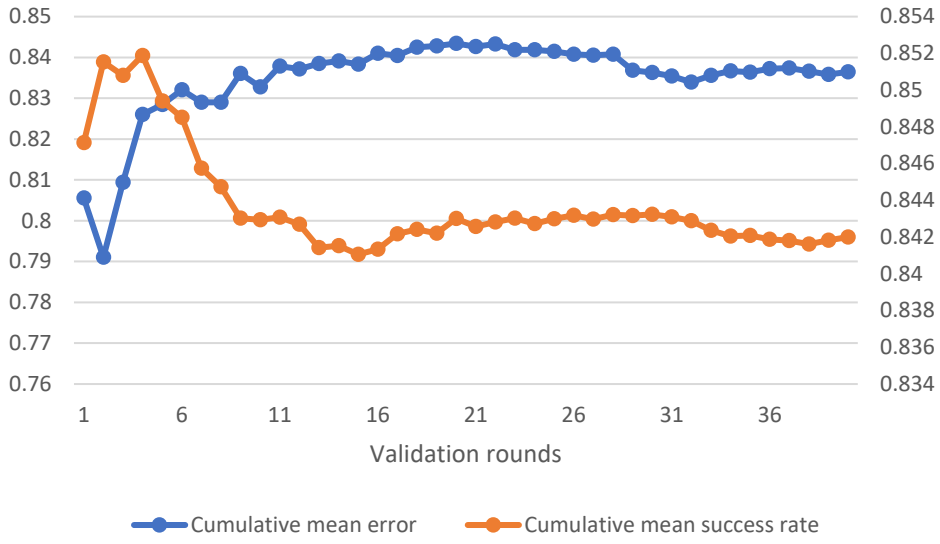


Figure 5.8: Number of observed timespans daily

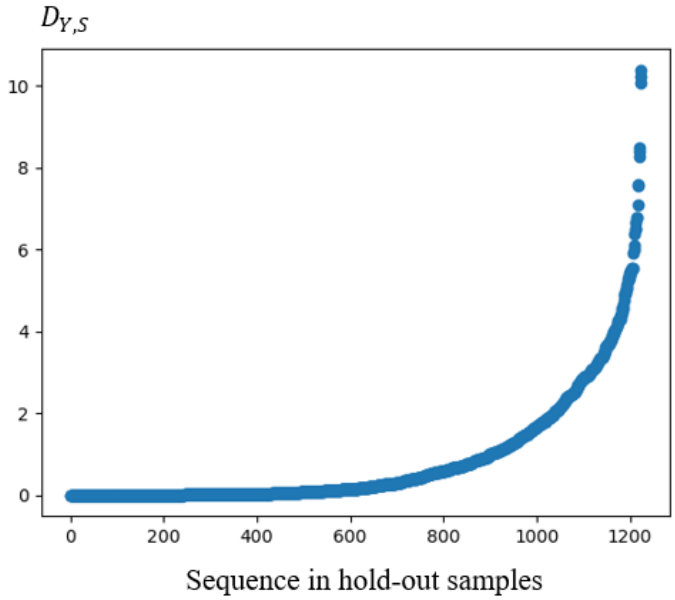


Figure 5.9: error of individuals in one validation (by ascending order)

5.4 Summary

This chapter proposes a methodology to use the Wi-Fi records to reproduce tourist flows, time spent at different locations over the city at a disaggregate level with the assist of GPS footprints. Firstly, the data alignment between Wi-Fi and GPS logs is done by converting the raw observations into sequences

that share the same format in space-time. The conversion also provided benefits for learning how many observations missing in the sensing refer to a location-based method. Based on the temporally aligned sequence interpreted vectorized operations are implemented for efficient similarity evaluations between trajectories inferred by Wi-Fi and GPS observations. Further addressing the problem of "overfitting" when using real physical distances due to the low resolution of Wi-Fi sensing data, the distances are converted to fuzzified odd ratios. Despite the fuzzification some of the trips still dominate the pattern for fill-in missing information. To avoid this, the source sequence set is enlarged by shifting each source sequence one hour forward/backward.

The improvement is visualized and the enrichment method is verified by cross-validation. The enriched Wi-Fi traces have a greater number of observed hours as well as more reasonable OD distributions along time compare the two raw data. The cross-validation is conducted to use 80% of the completed GPS sample to reproduce the rest 20% hold-out sequence shows acceptable results.

As the outcome of this chapter, the enriched Wi-Fi packet traces keep the advantages of rich sample size and the ability to capture seasonal dynamics to traditional survey. Furthermore, compared to common Wi-Fi packet datasets, it contains much more detailed information for the trip chains of each individual. This data enrichment provides an economical, large, and detailed dataset which is expected to become the base for tour-based modelling for urban tourist, and one of the examples is presented in Chapter 6.

Chapter 6 Modelling tourist flow in Kyoto City

6.1 Introduction

Chapter 4 models pedestrians' route choices in street blocks with relatively high-resolution Wi-Fi packet data samples. Chapter 6 here expands the research objectives to explore the possibility of modeling the tourist flow over the whole Kyoto City with a limited number of sensors.

Accordingly, this chapter models city-scale travel patterns. Instead of the decisions on tourists' detail routes, the study focuses on modelling their travel patterns in a day. The enriched Wi-Fi packet traces introduced in Chapter 5 have a large sample size and ability of capturing seasonal dynamics with reasonable cost. Hence, a methodology is explored for applying the model on these filled-in trips derived by observations of Wi-Fi sensing and GPS positioning.

The framework of network extraction, Wi-Fi record processing and recursive logit (RL) modeling (Figure 4.6) remains to be employed, as the RL models that maximize both the instantaneous utility and downstream utility for decision makers are considered valuable in formulating tourists' travel patterns along the time of day. To represent the attractiveness of each area, measurement is developed based on the open data reviewing POIs. The attractiveness of each POI is further associated to their opening hours and accordingly evaluation of attractiveness that is specific to both geographic of area and time of day is obtained. In order to take full benefit from these temporal variables, different from the Higashiyama Ward study, this study reconstructs the city network into a space-time network and assign the time dependent attributes to each links as their alternative-specific variables. Further dummy-links are introduced to the network and expect them to capture the error caused by missing from the area under monitoring.

The reminder of the chapter is organized as follows. Section 6.2 illustrate the process of network construction including sensor grouping, link attribute preparation, and the assignment of constructing time-expanded networks. Section 6.3 describes the model specification with the input networks and samples. 0 discusses the estimation results. Conclusions are drawn in Section 6.5.

6.2 Network construction

6.2.1 Grouping

As shown in Figure 3.6 the sensors are unevenly deployed. Some of the sensors are very close to each other, such as the sensors inside a train terminal. The objective of distributing sensors in this way is to capture people who appear in a station as complete as possible, while in studying persons' trip chains in city-scale, small specific movements becomes less necessary. Consider also the not high penetration rate as well as the possibility that a single signal may be acquired by multiple sensors at the same time, grouping some sensors is reasonable and necessary for further learning of the travel patterns.

The challenge then comes to how to logically group the sensors. Although the real distribution is unknown, A sensor grouping is sought which minimizes the differences in the distribution of the previous/next area visited derived by the grouped GPS and Wi-Fi data . For this purpose, the Kullback-Leibler divergence (Kullback and Leibler, 1951) is employed as a response to the similarity between two distributions. Between two discrete probability distributions *P* and *Q* in the same probability space Ω, their difference can be measured by:

$$D_{KL}(P||Q) = \sum_{\omega \in \Omega} P(\omega) \ln\left(\frac{P(\omega)}{Q(\omega)}\right) \tag{6.1}$$

where *Q* is the reference and *P* is the observation. If the distribution of *P* is close to *Q*, the measurement will be close to zero. This divergence measure is applied to the problem of sensor grouping by defining *P* and *Q* as the destination-distributions derived by GPS and Wi-Fi data.

All GPS signals within 500 meters of a sensor are associated with their nearest sensor, further, hourly OD transition matrices between each sensor-group can be derived for both Wi-Fi and GPS data. Given a sensor-group combination as $\boldsymbol{G}$, the derived OD matrices can be denoted as $\boldsymbol{Q_G^{WiFi}}$and $\boldsymbol{Q_G^{GPS}}$. Hence each row in $\boldsymbol{Q_G^{WiFi}}$ and $\boldsymbol{Q_G^{GPS}}$ represents the destination-distributions that sum up to one, which can be applied to Equation (6.1) for quantitating similarity, as shown in Equation (6.2).

$$D_{KL}(Q_g^{GPS}||Q_g^{WiFi}) = \sum_{i \in \boldsymbol{G}} Q_g^{GPS}(i) \ln\left(\frac{Q_g^{GPS}(i)}{Q_g^{WiFi}(i)}\right), \qquad g, i \in \boldsymbol{G} \tag{6.2}$$

where $Q_g^{Wifi}(i)$ and $Q_g^{GPS}(i)$ is the ratio from group g to group i refer to the destination-distributions.

Accordingly, denote $\boldsymbol{s}$ as the set of all the sensors and $\boldsymbol{X}$ as the vector of sensor combinations projected by $\boldsymbol{s}$ and $\boldsymbol{G}$, for example, $\boldsymbol{X}$ = {1,2,3,1,3,3} refers to separating six sensors to three groups with the first and fourth sensor in group one, the second sensor in group two, the third, fifth and sixth sensor in group three. Then with predefined number of groups, a combinational optimization problem expanded from Equation (6.2) is formulated as:

$$\underset{(\boldsymbol{s},\boldsymbol{G})\to\boldsymbol{X}}{Min} \sum_{g\in\boldsymbol{G}} \left(\frac{N_g^{GPS}}{N^{GPS}} + \frac{N_g^{WiFi}}{N^{WiFi}}\right) D_{KL}(Q_g^{GPS} || Q_g^{WiFi}) \tag{6.3}$$

Subject to:

$$\neg(g_i(lon^s) < s(lon) < g_i(lon^n)) \wedge (g_i(lat^w) < s(lat) < g_i(lat^e))), \qquad \forall s \notin g_i \tag{6.4}$$

$$hav(s_i(lat, lon), s_j(lat, lon)) < \sigma, \qquad \forall s_i \in g_i, s_j \in g_i \tag{6.5}$$

$$|g_i| \geq 1, \ \ \forall g_i \in \boldsymbol{G} \tag{6.6}$$

Equation (6.3) looks for a group combination that minimizes the sum of K-L divergence given by destination-distributions of each origin for GPS data and for Wi-Fi data separately, which is weighted by the number of observations at the origin. Denote group i as g_i, the boundaries of g_i by north latitude, east longitude, south latitude, west longitude as $g_i(lon^n, lat^e, lon^s, lat^w)$, constraints (6.4) ensure that the areas covered by any two groups do not overlap each other. Constraints (6.5) guarantee that any two sensors in the same group must not be more than σ apart, where $hav(s_i(lat, lon), s_j(lat, lon))$ is the distance calculated by longitudes and latitudes through the Haversine formula (Robusto, 1957). In this case σ is defined as 2 kilometers considering the size of the Kyoto City. Constraints (6.6) ensure each group contains at least one sensor.

Genetic algorithm (GA) is employed for the minimization. The $\boldsymbol{X}$ is used directly as the chromosome, hence the order of genes corresponds to the series of sensors, and genes with same value indicate a

same group. A transpose algorithm that randomly swaps a random length of two slices in vector $\boldsymbol{X}$ is used as mutation operation. Besides, a specific crossover operation illustrated in Figure 6.1 is developed for this problem. Following transpose algorithm and the adjusted crossover operation, at least one sensor is kept in a sensor-group.

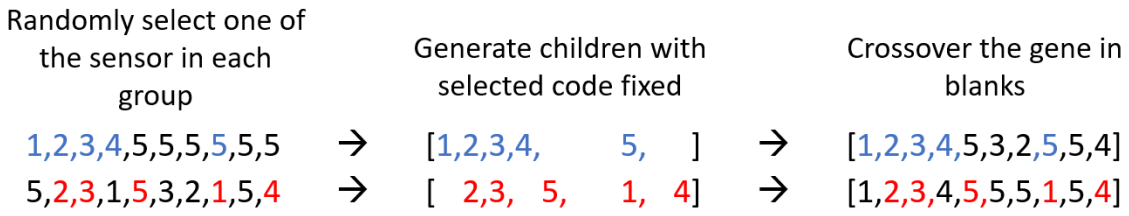


Figure 6.1: Illustration of the adjusted crossover operation

With a population size = 50, mutation probability = 90%, and number of sensor groups = 16, the resulting combination obtained by the GA is presented in Figure 6.2 using the Kyoto tourism map already shown in Figure 3.3. A relatively high mutation probability is used, since the mutation algorithm plays the main role for generating combinations. The number of groups is determined after alternative numbers have been tested, and 16 appears to be a good trade-off for interpretability and sufficient but not too many groups for the model. It should be noted that the problem space is highly non-convex and non-continuous so that with even longer search alternative solutions that are as good might be found. Nevertheless, the current grouping appears to be reasonable and sufficient for the purposes of this study. Clearly in further work one could refine the constraints and the search as discussed in Chapter 7.

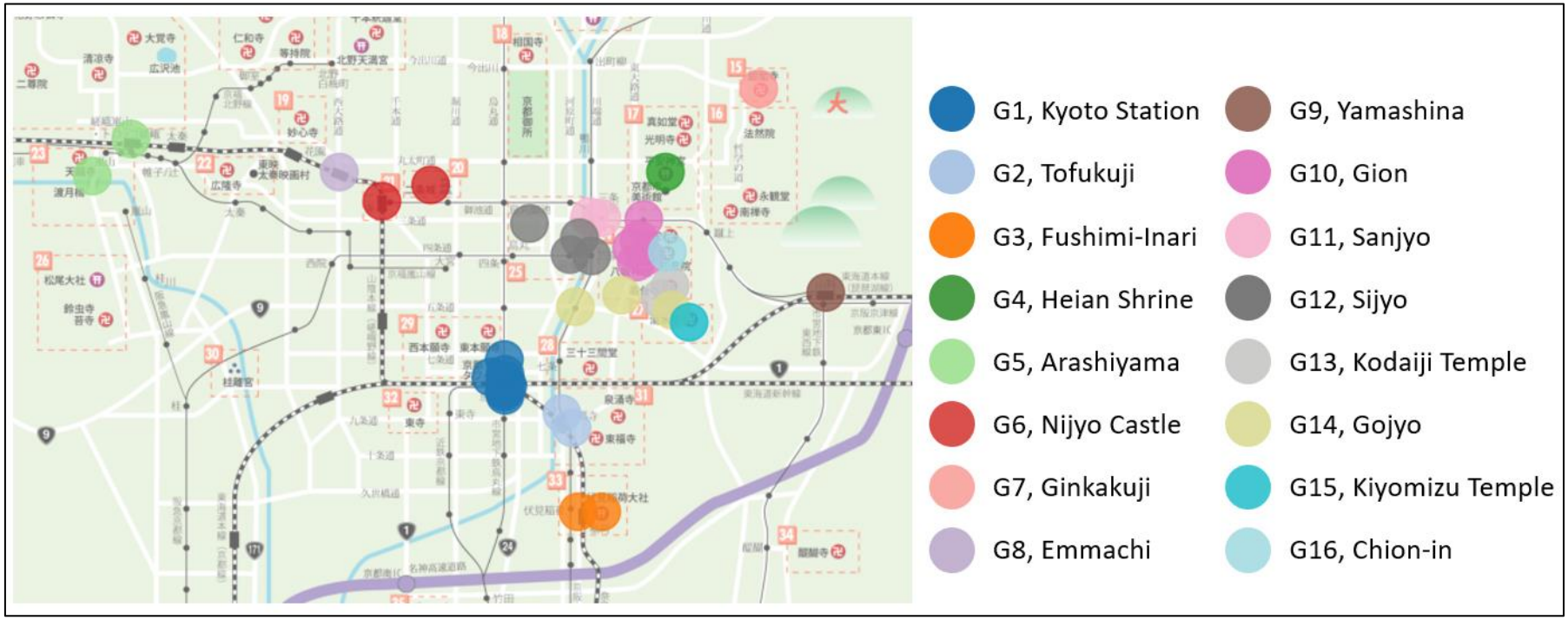


Figure 6.2: The 16 sensor-groups used in modelling refer to the tourism map of Figure 3.3

6.2.2 Collecting geographical attributes

Following the similar framework as the Higashiyama study in Chapter 4, adopting the Wi-Fi sensing data requires simplifying the network of Kyoto City. As the foundation of the network reconstruction, a variety of information is collected via Google APIs and web-crawlers. Based on the obtained data, a methodology to evaluate the attractiveness is explored.

Monetary and time cost is collected to travel between each sensor-pairs through the Google Direction API. Because of the access restrictions of the API the study does not collect very detailed time and fare tables. The fare and travel time table of public transport is inquired at 7:30, 12:30, 17:30 and 22:30 at a weekday, as these timings represent morning peak, daytime off-peak, evening peak and nighttime off-peak, respectively. Data collection then approximate the cost from 6:00 to 10:59 as the cost of 7:30, the cost from 11:00 to 15:59 as the cost of 12:30, the cost from 16:00 to 20:59 as the cost of 17:30, and the cost from 21:00 to 23:59 as the cost of 22:30.

POIs are presumed to be a strongly related to attractiveness of area. Especially in the era of social networking service, temporal information of POIs is believed to be highly effective for tourists in terms of determining daily trips. Via Google Place API, coordinates of POI location , possible types, average rating scores and cumulative number of reviewed users are obtained. Additionally, aiming at well-describing the geographical attractiveness along time, a web-crawler is developed to hourly gather the information of whether the POI is open. For better understanding, an example of information obtained from the Google Place is given in Figure 6.3.

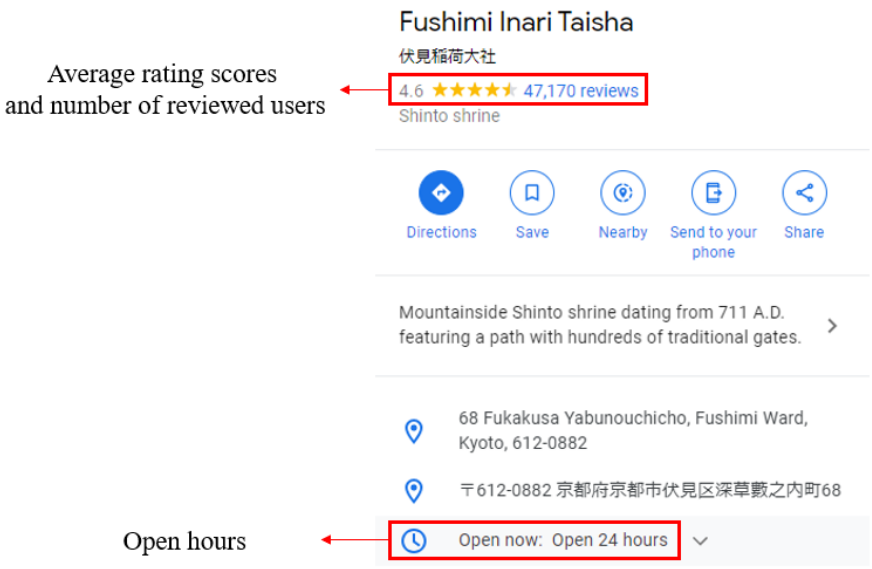


Figure 6.3: An example of Google Place informations. The average reating scores, number of review users and open hours are collected.

Based on data from Google, there are more than 60,000 POIs inside Kyoto City with possible functions be categorized. Each POI is usually grouped into from one to more than three categories by function, and there are hundreds of these categories. In total there are hundreds of POI types which are not hierarchically organized, causing the way of categorization is sometimes tedious and inaccurate, for example a store with ATM inside selling food may be labelled as both store, food and financial. To further select the POIs that may affect the choice of a tourist, this study focuses on the POIs that include features regarding open time, for the reason that a POI with open time is inferred to be more likely to belong to service industries. Figure 6.4 plots the active number of POIs collected by categories along time. The night amusements here refer to bars, pubs, night clubs and amusement such as bowling alley, and movie theaters that tend to be more attractive at nighttime. A drop can be seen for active number of restaurants in the afternoon, for most of the restaurants in Japan have afternoon breaks. Shops and sightseeing generally open at similar timing, sightseeing spots usually close around 17:00, and shops usually close later at night. Night amusements can be found gradually open after 18:00. Overall, the active POIs are found can be a considerable description to temporal attractiveness for each area.

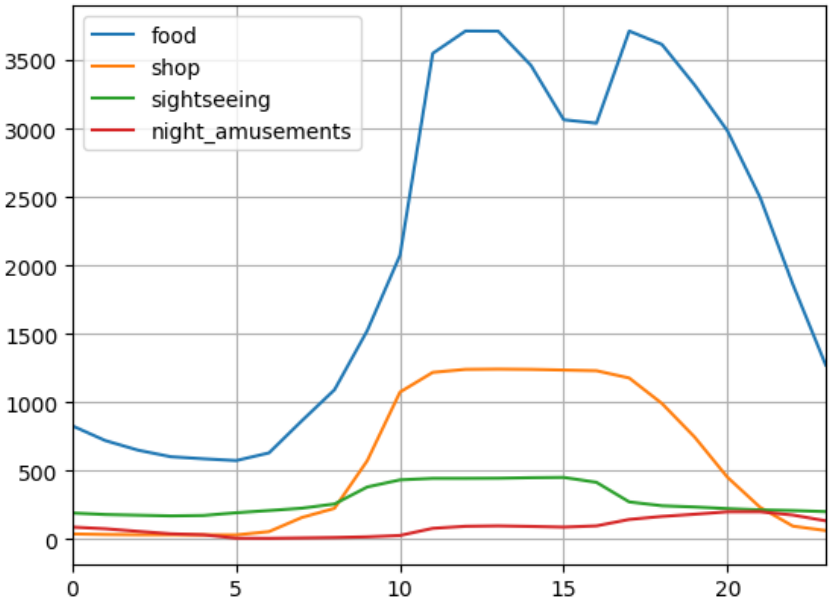


Figure 6.4: Active point-of-interst in Kyoto City along time

6.2.3 Network extraction

As there are clearly far less sensors than road links, the road and public transport network is transformed into a coarse network with the sensor groups as nodes. The links between these nodes describe the travel costs between the groups and the attractiveness of the head node (the "next area attraction") similar to the approach described in Chapter 4. However, different from the last study, the network needs to be reconstructed temporal-spatially, because time-varies variables prepared in the last section have to be assigned to time-specific links. an adjusted space-time network structure is proposed for the survey that attempts to model the flow over a large area with a fairly limited number of sensors. The illustration of a space-time route on the network is shown in Figure 6.5 where space is reduced to a single dimension. On the network individuals are only allowed to travel forward along time. Regular nodes on the network refer to the sensor-groups, and the sensor groups are assumed to be spatially completely connected. A special treatment is further introduced to overcome the limitation due of Wi-Fi data not being ubiquitous. The role of different type of links are detailed as follow.

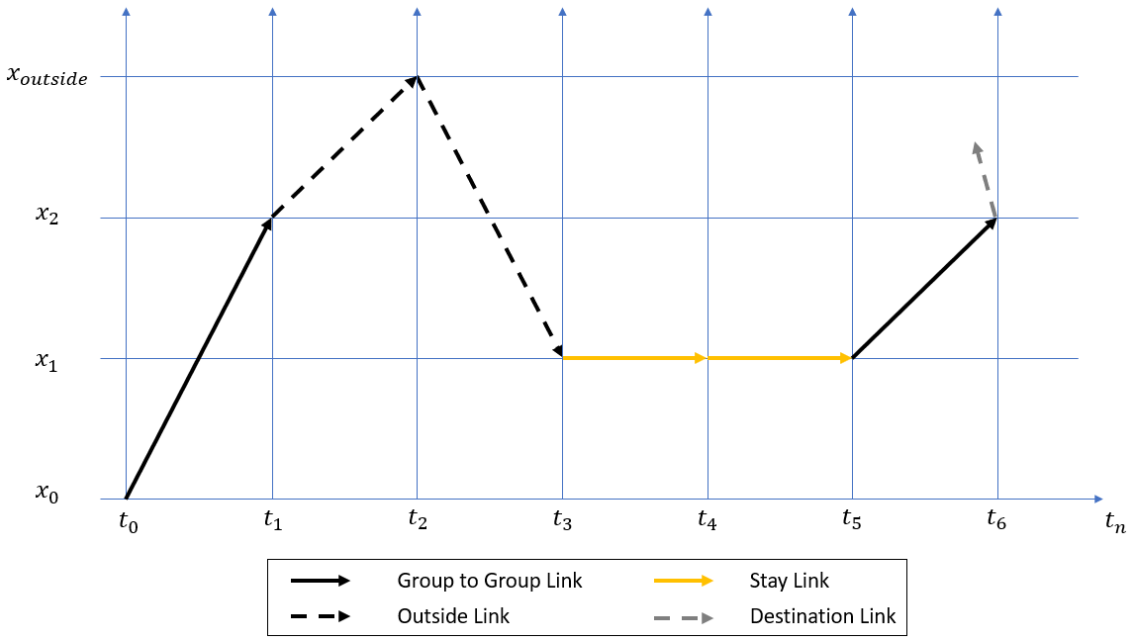


Figure 6.5: Illustration of trips on an adjusted space-time network

- **Group to Group (GtG) Links:** The links that represent travel from one area covered by a sensor-group to another. The links are forward in time on the space-time network as illustrated by solid lines in black in Figure 6.5. These links are characterized by geographic knowledges of correspond sensor-group e.g. time-specific travel duration, fare and POI-related variables. How to obtain and associate the geographic attributes to each sensor-group meanwhile respecting the addictiveness of the recursive logit will be introduced in next section.

- **Stay Links:** Unlike the network in the study of Chapter 4, a space-time network does not need specific Stay-Links paralleled with move-links. The behavior of choosing to remain at the current location for the next time period on space-time network is equivalent to choosing a link where head node and tail node area spatially same, as the yellow links that are shown in Figure 6.5. To capture the specific utilities given by choosing to stay at current area, instead area-specific constants of "stay" which equal to one when head node and tail node of a link is spatially same is introduced. This series of dummies are expected to capture the utility of staying at certain links.

- **Outside Nodes and Links:** The data experiment has limited observable areas in the city. To overcome this shortage the concept of "Outside nodes" is proposed as illustrated in the top row of Figure 6.5. Corresponding "Outside Links" are illustrated by dot lines since these links do not refer to a trip between two physical areas. The Outside node is used to denote all the locations given by GPS enriching process that are a certain distance away from far from a Wi-Fi sensor.

  Individuals are presumed all go to a dummy area namely the "outside" when they leave areas covered by Wi-Fi sensing. Making such presumptions in the model is equivalent to presuming that all of the areas that are away from the sensors are homogeneous and equidistance from all sensors. Thus, in order to be able to capture at least to some extent the area-specific influence for departure for "outside", area-specific binary constants are set for each outside links. On an outside link dummies of "stay outside", "move to outside" and "back from outside" are defined. Afore binary constants are set to capture the factor that cause a person to choose to disappear from/return to the survey area on the space-time network. The "move to outside" dummies equal to one when the head node of the link is outside and the tail node of the links belong to one of the under monitoring. This vector is expected to capture the incentive/disincentive that approaching the "outside" from one of the areas on the network except from outside itself. Similar to afore mentioned stay and continuously stay dummies, the "stay outside dummies" equals one when the head node and the tail node of link are both outside node, while the "continuously stay outside dummies" equals one when the outside link is connected with another outside link. They are expected to capture the utility of staying at the outside links.

- **Destination links:** Representing the absorbing stage of the daily trip, destination links are set that have no attribute connected with each physical link of the network, as illustrated in Figure 6.5 by dot lines in gray. Outside links therefore do not connected with a destination link, correspondingly in route extraction, individuals' routes are presumed to end up at the last non-outside links.

6.2.4 Link attributes

With map data being prepared and the network structure in place, the attributes are assigned to each link on the space-time network by building a mapping between sensors and sensor-groups. Denote s as a sensor, two groups of sensors $G_I = \{s \mid s \in G_I\}$ and $G_J = \{s \mid s \in G_J\}$, then the travel cost $c_{I,J}$ (both monetary and time cost) between groups G_I and G_J is calculated by

$$c_{I,J} = \frac{\sum_{s_j \in G_J} \sum_{s_i \in G_I} c_{i,j}}{|G_I| \times |G_J|} \tag{6.7}$$

where $c_{i,j}$ is the cost between sensor s_i and s_j. By this way the unweighted mean of cost are assigned to each link. Further the cost of travelling inside a group is approximated as zero. Thus, there is no time or monetary cost for a stay link. Whether this means that overall the attractiveness of stay-links is over- or under-estimated is unclear, as it depends on whether remaining in the area is more associated with the pleasure gained from the attractions or the costs incurred by travelling in between some of these attractions inside the area.

In order to obtain variables that can clearly indicate the attractiveness of the area given by POI, their basic characteristics should be firstly discussed. One may still consider the number of POIs as a measurement as in Chapter 4, but this study suggests that the aggregated number of POIs is not as good a measure in the city scale network because of the larger size of each zone. For example, a tourist is not significantly more attracted to an area if there are more than a certain number of souvenir shops. Instead it might be the quality or fame of some of these attractions that are a better measure.

To address this concern, this study considers two available variables related to attractiveness: the aggregated rating based on user review from 1.0 to 5.0, and historical accumulative number of rated users. In Kyoto City until October 2020, the most rated place is the Fushimi Inari that have over 46,000 reviews, followed by Kinkaku-ji (more than 32,000 reviews), and Kiyomizu Temple (more than 30,000 reviews). Presuming the attractiveness of a place is indicated by both the aggregated rating and number of reviews, this study purposes the product of normalized rating and Equation (6.8) as an indicator of this attribute:

$$A_x = log_{10}(NoR_x') \times R_x' \quad (6.8)$$

where A_x denotes the attractiveness of POI x, NoR_x' and R_x' is the number of rated user and average rating of POI x normalized into the range [0,1] by min-max scaling. Logarithm transformation is employed to close the difference of NoR_x, otherwise the gaps between famous attractions and other places will be too large to compare.

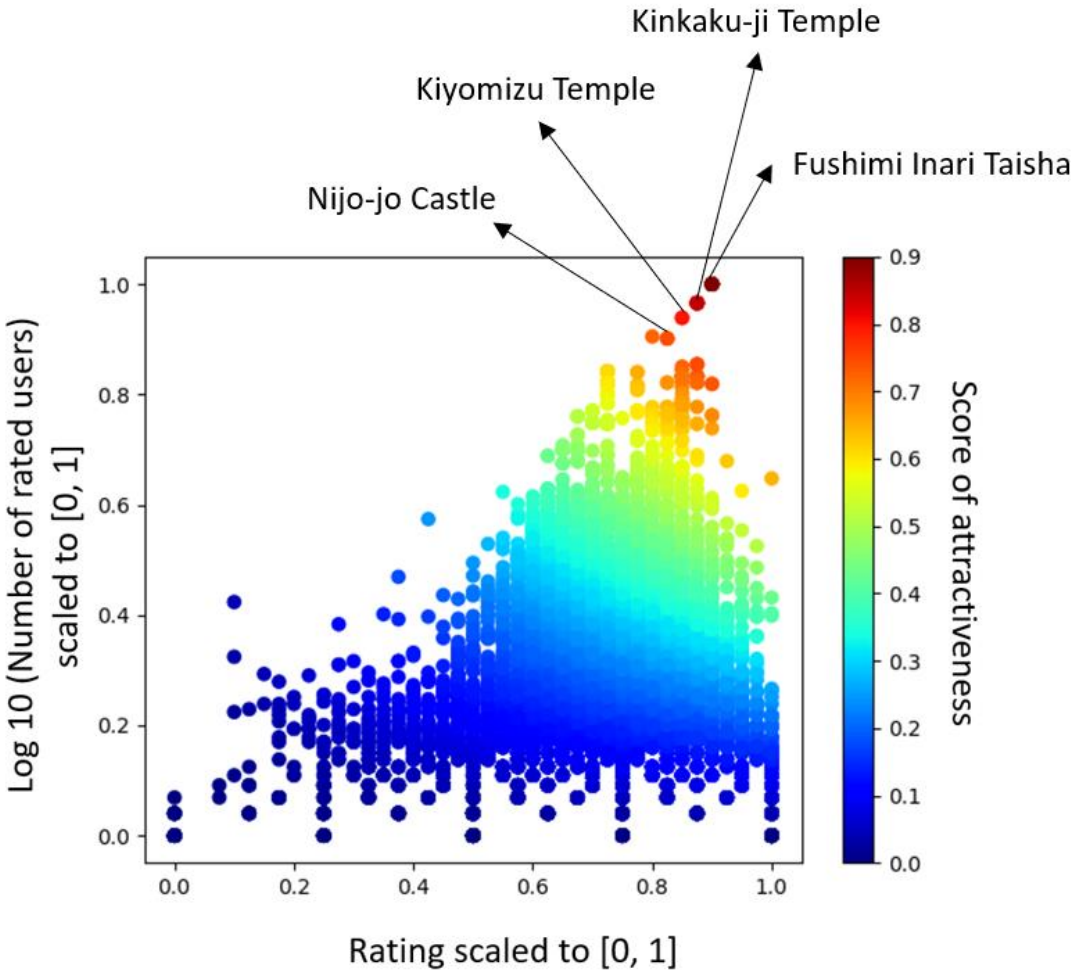


Figure 6.6: Ditribution of the review scores derived by Equation (6.8). Top scored locations are marked on the figure.

Figure 6.6 is the scatter plot of all POIs obtained in 6.2.2. Horizontal axis represents the R_x' in Equation (6.8), vertical axis represents $log_{10}(NoR_x')$, and each dot is colored by attractiveness-measurement A_x. It can be told from the coloring that the measurement for attractiveness is generally reasonable: a place is expected to have a high score of attractiveness if it is both highly rated and also

reviewed by enough people. Both the number of reviews and the rating contribute to the score of attractiveness, and the closer the point to the upper right, the higher the score. In this way attractiveness of different POIs is evaluated. Figure 6.7 shows the obtained aggregated attractiveness scores. For most of the area, two peaks are found for the curves of attractiveness from restaurants and shops, which is the lunch and dinner time. Sightseeing POIs are found more attractive during daytime.

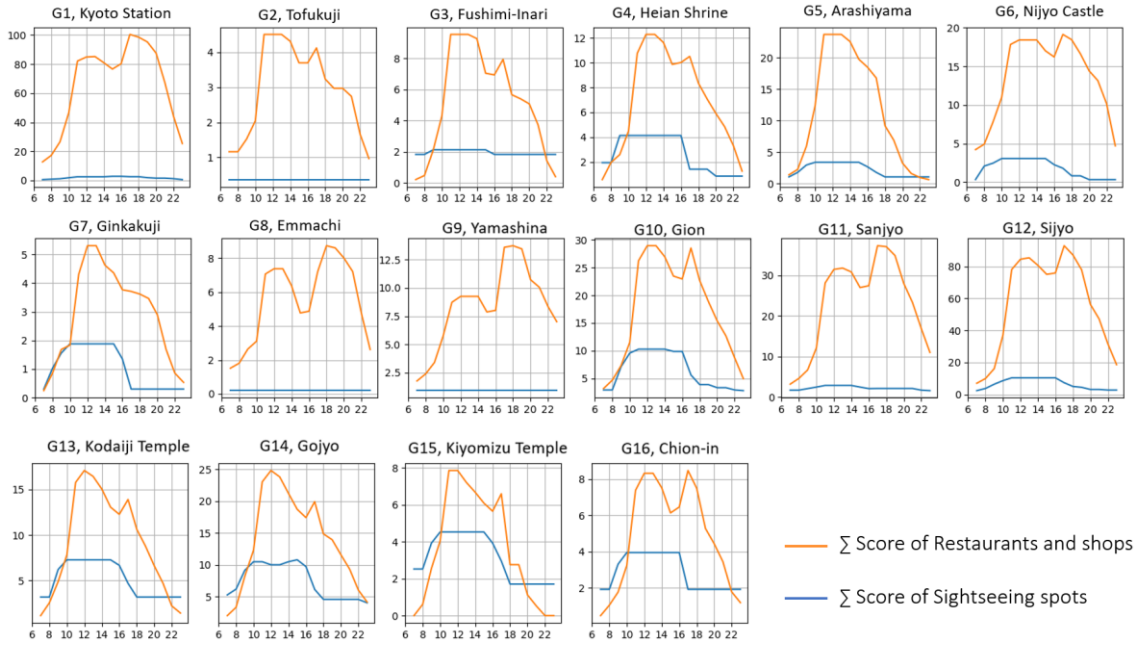


Figure 6.7: Attractiveness score aggreated by sensor-gruops.

Link attributes on the stay links need to be carefully handle in order to follow the additive feature of recursive logit model. In the previous study on no temporal expanded network, for those attributes that is not naturally additive, they are transformed by taking the product of the attribute itself and corresponded link length. Comparatively the situation changes on the space-time network. All behavior on the temporal-special network has an implicit time component.

To obtain POI related variables each sensor-group is associated with the POIs within 500m of one of the sensors in the group with this group. Denote the set of these sensor-corresponding POIs as X_s, accordingly, POIs associated to a sensor-group is $X_G = \bigcup_{s \in G} X_s$. Then the POI- attractiveness of a sensor-group G is determined by the aggregation of the attractiveness-indicators of POIs in X_G.

$$A_{X_G}^t = \sum_{x \in X_G} A_x^t \quad (6.9)$$

To note is that this attribute is also applied to stay links as well as to attributes that are not time-specific. If an attribute does not change by time, following Equation (6.9), it will become zero on a stay link thus give zero utility to those who choose to stay.

All link attributes are summarized in Table 6.1 with a brief description.

Table 6.1: Link attributes of Kyoto City space-time network

ATTRIBUTE	DESCRIPTION
Fare	Monetary cost obtained from Google Maps between two sensor-groups, given by Equation (6.7).
Time	Time cost obtained from Google Maps between two sensor-groups, given by Equation (6.7).
Attractiveness of active sightseeing POIs	Sum of score of attractiveness of sightseeing-type POIs, such as tourist attractions, given by Equation (6.9).
Attractiveness of active restaurant and shop POIs	Sum of score of attractiveness of restaurant- and shop- POIs, such as tourist attractions, given by Equation (6.9).
Stay constant	Equal to one if a link is a Stay-Link and not an outside link.
Outside constant	Equal to one if the head-node of a link connected with an outside link.
"From outside" constant	Equal to one if the tail-node of a link connected with an outside link.
Stay outside constant	Equal to one if a link is a Stay-Link and an outside link.

In addition, an alternative approach exists when attributes need to be assigned regardless of the implicit time component of the time-space network, which means that an individual will only obtains the utility associated to the attribute once at the moment of making a choice. In this case the network additivity should be respect. A space-time network requires for maintaining not only the spatial but also the temporal additivity of each property in the network as some attributes are time-specific. Under this assumption, when an individual enters a new area, he/she obtains an incentive measured by the utility, and after a period of time if he/she chooses to stay at current area, the incentive will not be obtained again. Accordingly, the index of POIs to a link refers to a trip from G_I at t to G_J at t_1 is:

$$X_{I,J}^{t_0,t_1} = X_{G_J}^{t_1} - X_{G_I}^{t_0} \quad (6.10)$$

Indexed by Equation (6.10), attractiveness attributes on each link is:

$$A_{I,J}^{t_0,t_1} = \sum_{x \in X_{I,J}^{t_0,t_1}} A_x \quad (6.11)$$

Whereas this presumes that all attributes are associated with the implicit time component of temporal links. These attributes satisfy the additivity and are the direct inputs to the recursive logit model. The usage of Equation (6.7) and (6.8) will be discussed in Chapter 7.

6.3 Model specification

6.3.1 Input network

As input network the 39 sensors are divided into 16 groups following the discussions in section 6.2.1, and divide one day into 24 time periods but ignore the midnight of 0:00 to 5:59. Figure 6.2 shows the grouping of sensors, and the naming based on area features for better narrative. Accordingly, in the network there are 16 physical stages, one absorbing stage, one outside stage spatially 18 timespans in chronological terms. The network is then generated following the descriptions in Section 6.1 With this, there are totally 5940 links, 273 destinations (the first timespan does not have absorbing stage). The physical links and the outside links are completely connected spatially.

6.3.2 Input samples

Data collected in first week of December 2019 is input to the model. The sample is separated by weekday and weekends. The basic statistics of the three samples are shown in Table 6.2. Note that this study does not divide long-time missing into several short trips, instead use GPS-fill in approach to infer where an individual has been to during missing. Thus, compare to samples of the study in Higashiyama, the number of routes decreases, and number of choices made in a route increase.

Table 6.2: Basic statistics of samples

	The first week of December 2018 (weekdays)
Number of routes	60,450
Mean choice per route (std.dev)	7.04 (3.13)
	The first week of December 2018 (weekend)
Number of routes	30,990
Mean choice per route (std.dev)	6.97 (3.11)

6.3.3 Model specification

Since this experiment does not conduct a survey regarding the penetration rate, there is not an error term for it. Hence, we use the standard Bellman equation for RL, as formulated in Equation (6.12).

$$V_n^d(k) = E\left[\max_{a \in A(k)} \left(u(a|k) + V_n^d(a) + \mu\varepsilon_n(a)\right)\right], \quad \forall k \in A \tag{6.12}$$

In the spacital-temporal network, given current link k, the instantaneous utility of choosing link a at timespan t as the subsequent link is:

$$\begin{aligned} v(a|k) = \beta_{MC} MC_{a|t} + \beta_{TC} TC_{a|t} \; &+ {\delta_a}^{stay} \boldsymbol{\beta_{AP}}^{\boldsymbol{stay}} \boldsymbol{AP_{a|t}}^{\boldsymbol{T}} \\ &+ (1 - {\delta_a}^{stay}) \boldsymbol{\beta_{AP}}^{\boldsymbol{move}} \boldsymbol{AP_{a|t}}^{\boldsymbol{T}} + \boldsymbol{\beta_{FromOutside}} \boldsymbol{FO_a}^{\boldsymbol{T}} \\ &+ (1 - {\delta_a}^{stay}) \beta_{OS} {\delta_a}^{outside} + \beta_{StayOutside} {\delta_a}^{stay} {\delta_a}^{outside} \end{aligned} \tag{6.13}$$

where $MC_{a|t}$ is monetary cost of link a (1000 JPY) at time *t*, $TC_{a|t}$ is time cost of link a (hour) at time *t*, and $\boldsymbol{AP_{a|t}}$ is the summation of attractiveness measured by Equation (6.9) of POIs of type *p*, in this case, still same as the Higashiyama Ward study, $p = \{Food\&Shop, Sightseeing\}$ associated with link a. $\boldsymbol{AP_{a|t}}$ is scaled by 0.01 for better performance of the estimator. Regarding variables related to the outside-links, ${\delta_a}^{outside}$ is the afore described constant of Outside links, which can also be interpreted as a penalty of entering outside-links, and $\boldsymbol{FO_a}^{\boldsymbol{T}}$ is the penalty for entering a specific physical area from outside.

Different from to the previous model without time-delayed expansion, considering there are enough specific (both temporal and spatially) variables for each link, a stay constant does not directly enter

the model. Instead on stay-link and move-link the effect of $\boldsymbol{AP_{a|t}}$ is separated by interacting $\boldsymbol{AP_{a|t}}$ with stay-indicator ${\delta_a}^{stay}$.

As discussed in Fosgerau et al (2013), the probability of observing individual *n* choosing an outgoing link from current link *k* given destination *d* is

$$p_n^d(a|k) = \frac{e^{\frac{1}{\mu}v_n(a|k;\,\beta,\theta)+V_n^d(a;\,\beta,\theta)}}{\sum_{a\prime\in A(k)} e^{\frac{1}{\mu}v_n(a\prime|k;\,\beta,\theta)+V_n^d(a\prime;\,\beta,\theta)}} \tag{6.14}$$

where the value function for the downstream utility V_n^d of link *k* is recursively given by

$$V_n^d(k;\beta,\theta) = \mu \ln\left(\sum_{a\in A(k)} e^{\frac{1}{\mu}v_n(a|k;\,\beta,\theta)+V_n^d(a;\,\beta,\theta)}\right) \tag{6.15}$$

except for the destination where the utility is fixed to $V_n^d(d) = 0$. Let I_n be the number of links traversed by person *n*. Then the probability of observing individual *n* choosing a path $\sigma = \{k_i\}_{i=0}^{I_n}$ is as in (6.8) where μ is a scale parameter that is assumed to be equal to one unless nesting of options is assumed.

$$P_n^d(\sigma;\,\beta,\theta) = e^{-V_n^d(k_0;\,\beta,\theta)} \prod_{i=1}^{I_n} e^{\frac{1}{\mu}v_n(k_{i+1}|k_i;\,\beta,\theta)} \tag{6.16}$$

Thus, the log likelihood function can be obtained from the sequence of link choice probabilities as in Equation (4.8) where (4.7) is utilized in the second equality.

$$\begin{aligned} LL(\beta,\theta) &= \sum_{n=1}^{N}\left(\sum_{i=1}^{I_n-1}\left(ln\, p_n(k_{i+1}|k_i;\beta,\theta)\right)\right) \\ &= \sum_{n=1}^{N}\left(\sum_{i=1}^{I_n-1}\left(v_n(k_{i+1}|k_i;\beta,\theta) - V_n^d(k_0;\,\beta,\theta)\right)\right) \end{aligned} \tag{6.17}$$

6.4 Estimation results

Same as the study in Chapter 4, the DeC method is still employed for the recursive logit model estimation. To note is that the DeC method becomes even more necessary for a temporal-expanded network, as a general estimation methodology can solve the value equation for only one destination at a time, and the number of destinations has increased by a multiplier equal to the number of time periods.

Estimation results are shown in Table 6.3. Most of the estimates significantly effect the link utility, and all parameters have reasonable signs. Parameter of pubilic transport fare estimates are slightly different between weekdays and weekend, which are -1.23 and -1.37, respectively. Parameter of cost of time by pubilic transport remains same between the two sample. The value of time for tourists can be derived by these two parameters as approximately -4.1/-1.3 ≈ 3150 JPY/hour, which is consider higher than the actual. This might be caused by 1) tourists are less concerned about fare; 2) the data collection process generally undervalues the time cost. Not every tourist take public transport, and for those who do not use public transport the cost of fare may be underestimated; 3) insuffcient correction for the effect from correlation of time and monetary cost variables.

Regarding the coefficicents of temproral area attractiveness variables, both POI types have a significant positive effect on the link utility and the link being chosen, while sightseeing POIs have much more impact by per unit of attractiveness-score. Both types of POIs are basically more attractive during weekends, while the difference is not very obvious since the dominant group in the sample are tourists for which both types of attractions are important. The interation with stay-indicators give us additional informaiton for how effecitve the temporal POI attractiveness is for people to one's decision to stay at current area or not. Compared to travelling between areas, the attractiveness from POIs of restaurants and shops are found to provide considerable more positive utitlities to decision of stay, as similar to the findings in the Higashiyama Ward study when the network is not expanded by time but instead Stay-Links are introduced.

Regarding the outside-link related variables, they are expected to capture the penalty of entering and leaving the area of “outside” to some extent since there is no other data as measurement to describe the cost of “missing from monitoring”. The estimates are found indeed entering an outside link gives

negative utilities to the decision-maker refer to travel to any physical areas, which also illustrate from the side that the sensing covers most of the major area for tourists in Kyoto.

The returning from outside penalty are values refer to G1, the Kyoto Station, which is fixed to zero. Areas with similar geographic properties to the reference group e.g. G10 (Sijyo area) show no significance difference with the reference. This penalty responds to the resistance to return to areas that are under monitoring. They are expected to be less negative or more positive if an area is easy to access from other areas. The groups with significantly negative estimates are G13 and G16, two small attractions located right in the Higashiyama Ward, with good access from both train stations to their south and north. On the other hand, an area that requires relative long duration to reach will reduce the resistance of returning from Outside-links, for a high chance of missing from sensing since travelling likely requires more than one hour. G5 (Arashiyama area) is a good example, as its coefficient is in fact positive. It has to be acknowledged that there are multiple factors effecting this coefficient and some of them are offsetting each other, while there is no promising way to separate these effects so far.

The estimation of the model is time consuming. Trade-off has to be made between the number of variables and estimation performance/efficiency. For example, the outside-constant can be indeed area-specific, giving more information on the resistance of entering outside areas for each area. Moreover, each coefficients of the dummy constants can be even time-dependent. After test, we choose to leave these parameters out to save the estimation time since most of them are not significantly impacting the utility of links.

In addition, if an area-specific constant is associated to the utility of links, the model fit becomes better while the variables of POIs become much less critical. This suggests that there must be still latent effects this research can not specify.

Table 6.3: Estimation results without area specific constants

	Weekdays		Weekend	
Attributes	**Est.**	**t-test**	**Est.**	**t-test**
Monetary cost (1000 JPY)	-1.37	-209.85	-1.23	-147.66
Time cost (hour)	-4.10	-236.91	-4.11	-172.33
Attractiveness measurements				
Sightseeing POIs	2.80	40.52	3.04	31.5
Sightseeing POIs (staying)	2.27	21.08	3.86	25.31
Food and shop POIs	0.09	9.44	0.19	14.46
Food and shop POIs (staying)	1.03	85.37	0.72	41.36
Outside-link related constants				
Arrive at Outside	-2.25	-119.33	-2.4	-79.84
Stay Outside	0.42	31.53	0.28	13.12
Outside to G1 (Kyoto Station)	Ref	---	Ref.	---
Outside to G2 (Tofukuji)	-1.66	-33.91	-2.38	-21.37
Outside to G3 (Fushimi-Inari)	-1.07	-25.29	-0.71	-11.91
Outside to G4 (Heian Shrine)	-0.85	-19.81	-0.45	-7.51
Outside to G5 (Arashiyama)	0.62	20.92	0.93	22.88
Outside to G6 (Nijyo Castle)	-0.09	-3.05	0.35	8.1
Outside to G7 (Ginkakuji)	-0.53	-9.82	0.02	0.36
Outside to G8 (Emmachi)	-0.96	-18.72	-0.18	-3.06
Outside to G9 (Higashiyama)	-2.5	-29.18	-2.56	-17.93
Outside to G10 (Gion)	-1.12	-29.02	-0.5	-9.83
Outside to G11 (Sanjyo)	-2.15	-33.54	-1.5	-19.04
Outside to G12 (Sijyo)	-0.20	-7.55	0.01	0.24
Outside to G13 (Kodaij Temple)	-3.4	-25.75	-2.63	-17.51
Outside to G14 (Gojyo)	-1.47	-30.85	-1.46	-18.48
Outside to G15 (Kiyomizu Temple)	-1.63	-24.69	-1.27	-13.38
Outside to G16 (Chion-in)	-3.4	-27.02	-3.25	-16.67
MaxLogLL	-702,382		-360,040	
Number of routes	60,450		30,990	
Number of choices	425,469		215,905	
MaxLogLL per route	-11.619		-11.618	
MaxLogLL per choice	-1.651		-1.668	

Table 6.4: Estimation results with area specific constants

	Weekdays		**Weekend**	
Attributes	**Est.**	**t-test**	**Est.**	**t-test**
Monetary cost (1000 JPY)	-1.18	-129.96	-1.08	-86.52
Time cost (hour)	-3.57	-145.84	-3.64	-106.39
Attractiveness measurements				
Sightseeing POIs	1.33	7.79	2.19	9.27
Sightseeing POIs (staying)	-1.04	-5.79	0.11	0.45
Food and shop POIs	-0.10	-3.79	-0.13	-3.44
Food and shop POIs (staying)	0.60	23.5	0.49	13.16
Area specific constants				
Arrive at G1 (Kyoto Station)	Ref.	---	Ref.	---
Arrive at G2 (Tofukuji)	-1.72	-80.23	-1.79	-56.26
Arrive at G3 (Fushimi-Inari)	-0.57	-27.63	-0.71	-23.23
Arrive at G4 (Heian Shrine)	-1.22	-52.19	-0.94	-29.45
Arrive at G5 (Arashiyama)	0.43	24.7	0.43	17.01
Arrive at G6 (Nijyo Castle)	-0.43	-23.1	-0.48	-17.6
Arrive at G7 (Ginkakuji)	-0.28	-11.24	-0.32	-8.92
Arrive at G8 (Emmachi)	-1.73	-59.83	-1.88	-44.6
Arrive at G9 (Higashiyama)	-2.34	-82.83	-2.23	-57.57
Arrive at G10 (Gion)	-0.15	-6.64	-0.21	-6.79
Arrive at G11 (Sanjyo)	-1.6	-88.21	-1.48	-58.74
Arrive at G12 (Sijyo)	-0.35	-27.49	-0.36	-20.73
Arrive at G13 (Kodaij Temple)	-0.94	-40.48	-0.99	-29.38
Arrive at G14 (Gojyo)	-0.17	-7.32	-0.31	-9.01
Outside to G15 (Kiyomizu Temple)	0.05	2.35	-0.04	-1.27
Outside to G16 (Chion-in)	-1.72	-69.77	-1.79	-50.19

Table 6.4: Estimation results with area specific constants (continued)

	Weekdays		**Weekend**	
Outside-link related constants	**Est.**	**t-test**	**Est.**	**t-test**
Arrive at Outside	-3.38	-78.22	-3.4	-53.41
Stay Outside	3.16	99.21	3.15	68.4
Outside to G1 (Kyoto Station)	Ref.	---	Ref.	---
Outside to G2 (Tofukuji)	-0.16	-2.74	-0.07	-0.86
Outside to G3 (Fushimi-Inari)	-0.26	-5.71	-0.15	-2.25
Outside to G4 (Heian Shrine)	0.35	6.72	0.27	3.84
Outside to G5 (Arashiyama)	0.41	12.62	0.47	10.26
Outside to G6 (Nijyo Castle)	0.58	16.38	0.69	13.84
Outside to G7 (Ginkakuji)	-0.25	-4.04	-0.02	-0.23
Outside to G8 (Emmachi)	1.05	20.72	1.12	15.25
Outside to G9 (Higashiyama)	-0.02	-0.24	0.11	1.04
Outside to G10 (Gion)	-0.47	-11.24	-0.43	-7.17
Outside to G11 (Sanjyo)	0.11	2.08	0.21	2.92
Outside to G12 (Sijyo)	0.17	5.57	0.24	5.53
Outside to G13 (Kodaij Temple)	-1.23	-15.11	-0.93	-8.79
Outside to G14 (Gojyo)	-0.86	-17.58	-0.85	-11.53
Outside to G15 (Kiyomizu Temple)	-1.39	-22.04	-1.2	-13.36
Outside to G16 (Chion-in)	-0.66	-7.87	-0.61	-4.87
MaxLogLL	-627,869		-324,339	
Number of routes	60,450		30,990	
Number of choices	425,469		215,905	
MaxLogLL per route	-10.387		-10.466	
MaxLogLL per choice	-1.476		-1.502	

6.5 Summary

This chapter proposes a methodology to model city-scale travel patterns in a day based on the enriched Wi-Fi packet traces introduced in Chapter 5. The recursive logit model is still employed, as its feature of maximizing both the instantaneous utility and downstream utility for decision makers are considered fit for the purpose of formulating tourists' travel patterns along the time of day. Also, the recursive logit allows us to handle a large number of sensors as it does not require the generation of choice sets.

This study still follows the data preparation framework presented in Figure 4.6. Additionally, before network construction several antecedent issues in the data survey are firstly addressed, for example, the unevenness of sensor deployment by combinational operation that makes the OD distributions derived by the grouped GPS and Wi-Fi data be most similar. This problem is specified based on K-L divergence as a measurement for the similarity of destination-distributions. Genetic algorithm (GA) and specifically adjusted crossover and mutation operation of the GA to search a reasonable way to group the sensors.

During the network construction, to quantify the attractiveness of links, Google Place API and webpage crawler are used to obtain information including travel cost in terms of time and money, reviews and open hours for POIs. Specifically, the information of POIs in Kyoto City is used to represent the attractiveness of each link in the network. A comprehensive measurement of attractiveness is developed by combining number of rated users and ratings of each POI. This attractiveness is further associated to the opening hours of each POI. Finally, through aggregations by area, attractiveness that specific to area geographic and time of day is evaluated.

In order to take full benefit of the temporal variables, different from the Higashiyama Ward study, this research expands the reduced physical network into a space-time network instead of introducing external Stay-Links. This way time dependent attributes can be assigned to each time-specific links as alternative-specific variables. Further except for the regular links between each physical area, this study introduces dummy-links namely the Group to Outside links to the network. Such Outside links

are associated to several constants that specific to their former or subsequent link for capturing the errors caused by of missing from the area under monitoring.

Samples collected over weekdays and weekends are modelled respectively with and without area-specific variables. The estimates in of model without the area-specific variables suggest strong positve impact from POI attractivness to route choice, and the influence is stronger at the weekend. However, the impact are weakened if the area specific variables are introduced. To draw a concludsion, there are still latent variables that are not specified affect the tourists' behavior.

Chapter 7 Conclusions

7.1 Summary of research

The main objective of this research is to establish a methodology that allows understanding and predicting tourist flows by relatively low-cost Wi-Fi sensing technology, for its desirable cost-performance and superior privacy protection features. Targeting Kyoto City, two experiments are conducted, one at small-scale for investigation of detailed route choices; the other at large scale over the city for obtaining general travel patterns.

Chapter 2 reviews the research of characteristics of city-tourists and clarify the research directions. The "impulse stop" of tourists during travelling has been considered one of the focuses of research since the early years. With the development of statistical models and computing, researchers have applied a variety of methods to model and simulate city-tourists. Types of data source for studying travelers in the city especially tourists are also discussed. Taking into account the characteristics of the Wi-Fi packet sensing data, the main problem of this dissertation is characterized as understanding tourists' behavior over time periods of several hours to a day. Thus, aim at estimating route choices with jointly consideration of touristic behavior and the goal of the trips, the recursive logit (RL) model is employed.

Chapter 3 presents a general introduction to the tourism industry in Kyoto City, the subject of the studies in this dissertation. The tourist resources all over the city bring huge economic benefits to the city and at the same time become the reason of citywide congestion, while the tourist flow that brings congestions cannot be effectively investigated. This leads to two Wi-Fi packet sensing experiments in Kyoto City. One of these monitors small-scale specific route choices, and the other on city-scale travel patterns. The overview of resulting datasets is described. The first experiment successfully captures the trips inside the touristic area. However, in the second experiment, the limited observable area leads to limited understanding for detailed travel patterns inside the city. Accordingly, a commercially available GPS-based sample is introduced for the comparison with the Wi-Fi sample. Conclusions are drawn that although neither only using GPS nor Wi-Fi data separately is enough, a fusion of the two data may help interpret completed tourist travel patterns.

Chapter 4 proposes the methodology to utilize Wi-Fi packet sensors to model behavior of pedestrians in a touristic area. As preparation, a reduced network fit for sensing survey is reconstructed from the real network. Tourist tours with information of travelling and staying are identified based on data cleansing and clustering analysis. Then the formulation of the RL model is specified by introducing 1) Stay- and Move-links to capture the decision of making a stop on a link; 2) variables of number of accessible POIs to reflect link attractions given by POIs that make a tourist's travel less destination-oriented; 3) specific error terms to correct the bias caused by penetration rate difference among sensors. The estimates of the model have reasonable signs, and all further describe the difference in behavior between different periods of day and tourist clusters. The proposed Stay-Links perform well in both reducing computational consumptions and describing impulse stops.

Chapter 5 proposes a data fusion approach for the enrichment of Wi-Fi data based on a small sample of GPS observations. After data alignment, vectorized operation for similarity measurement is conducted between trajectories inferred by Wi-Fi and GPS observations. Then one of the most similar GPS trajectories is used to fill in the missing observations for its target Wi-Fi trajectory. The enrichment approach developed in this chapter lays the groundwork for tour-based modelling for urban tourist.

Chapter 6 proposes a methodology to model city-scale travel patterns in a day based on the enriched Wi-Fi packet traces introduced in Chapter 5. Modeling framework presented in Chapter 4 is still followed. In this case, the temporal-expanded network instead of Stay-Link network is employed, as the study in this chapter obtains values measuring attractiveness that specific to area geographic and time of day, and the temporal-expanded network has compatibility towards time-dependent variables. Further the Outside-links with a set of constants capturing the error caused by leaving the sensing area are introduced into temporal-spatial network. The model estimation shows strong positive effect from POI attractiveness, while also suggest there are unspecified latent factors contributing city-tourists' decision making.

7.2 Contribution to existing knowledge

This dissertation is one of the first to apply the RL approach to data obtained from Wi-Fi packet sensing. A complete set of processing starting from raw data cleansing to population group analysis, tour-choice estimation, and finally tourist flow forecasting is provided.

In terms of data derivation, the two major studies in this dissertation present different approaches to achieve the extraction of tourist trajectories. In the case study for Higashiyama Ward, a two-step clustering method helps to establish a role-based filter to remove the noises from collected samples and accordingly pick up the pedestrian tour choices. In the case study for Kyoto City, the data of individual travel patterns are derived with the assistant of a small sample of commercially available GPS data. For this case a methodology for efficiently matching similar travel patterns based on the characteristic of Wi-Fi packet data and GPS data is developed.

Both of the proposed data derivation approaches utilized the advantage of anonymized Wi-Fi packet sensing technology, that allow capturing tourists' tour choices with large sample size and seasonal dynamics. Especially, the data enrichment approach developed for the city-scale survey provides an economical, large, and detailed dataset which is expected to become the basis for tour-based modelling for urban tourist. In addition to the above main points, this dissertation also contributes to data derivation of quantitative POI-attractiveness based on open data sources, and the traveler clusters of Wi-Fi packet sensing sample.

Regarding recursive logit model specification, the methodology proposed in the case study of the Higashiyama Ward experiment is the first to apply the RL approach to Wi-Fi-sensor data and to explicitly model "walk" or "stay" without explicit definition of a large time-expanded network. In contrast, for the Kyoto City case, the temporal-expanded network is introduced for better interpretation of the time-varying variables. A particular contribution of this dissertation is the definition of both "stay-links" and modified time-expanded networks; the comparison of these two network structures will be discussed in Section 7.3.

Regarding model estimation, the contribution of this dissertation starts with a warming up algorithm that is proposed for finding appropriate initial values. Looking at the results, models applied in both experiments of the dissertation give reasonable interpretations of tourists' tour choices. Restaurant and shops related POIs and sightseeing POIs are shown to significantly attracts tourists, and one "unit" of sightseeing POIs is found several times attractive among several measurements. Moreover, these attractiveness from POIs, especially from restaurants and shops are found to provide considerable more positive utitlities to decision of stay, as similar to the findings in the Higashiyama Ward study when the network is not expanded by time but instead Stay-Links are introduced.

Specifically, in response to the observation loss of Wi-Fi packet sensing, this dissertation contributes several treatments during the RL model specification. For high-resolution monitoring survey, logarithm of sensor-specific penetration rates is proposed as the error terms which respect RL definition. For low sensor-coverage survey, dummy links with sets of physical area specific constants are introduced.

7.3 Non-temporal expanded network with stay links vs temporal expanded network

As the two network definitions are a central contribution of the thesis, this section compares the spatial-temporal network proposed in Chapter 6 with the non-temporal expanded network with Stay-Links (hereinafter referred to as the Stay-Link network) proposed in Chapter 4 and discuss the advantages and disadvantages of the two networks for recursive logit modelling. We compare these two networks mainly with respect to two aspects: efficiency and interpretability.

The Stay-Link network dominates in terms of computational efficiency. Without temporal-expansion, the complexity of a Stay-Link network is in general much lower than a temporal-spatial network given the same number of physical nodes. More importantly, we note the computational duration for recursive logit model estimation is directly impacted by the number of destinations, especially when the DeC method is not employed, as other method needs to solve systems specific to each destination one by one. Temporal-expansion of a network has to multiply the number of destinations by the number of time periods. Researchers have to make extra effort such as limiting the alternative links and destinations to address this problem (Zimmerman et al 2018). In contrast, the introduction of Stay-

Links to network does not increase the number of destinations while to some extend keeps the ability of modelling route choice with temporal information with the help of stay threshold. From this aspect, it is clear that the stay-link network contributes significantly to the improvement of computational efficiency.

In terms of interpretability, both approaches have their merits. The spatial-temporal network is better in the case of owning many explanatory variables: it provides the most straightforward environment for assigning time-dependent attributes, such as the temporal POI attractiveness derived in Section 6.2.4. On the other hand, even when the attributes are not time-dependent, researchers can still separate the parameter by time periods to capture different time-specific effects from attributes. These advantages are partly the explanation for the good compatibility between recursive logit modelling and spatial-temporal networks, as survey-based research employ spatial-temporal networks for RL, especially the activity-based model with hard is found demand for the interpretability along change of time (e.g. Zimmermann et al, 2018 and Vastverg et al, 2020).

While the finding of this research suggests that the stay-link network has its own strengths in the comparison of the two opposing behaviors: the estimates of Stay-link and Move-link specific variables then provide a clear comparison of how a variable influences the choice between “move” and “stay” for an easily changeable time period. Combined with the computational efficiency, this feature provides benefits to produce handy and practical applications for operators to understand travelers’ behavior with easily changeable time period.

In line with above discussion, the recommendation of network requires comprehensively discussion on the purpose of the study. To a large extends, it depends on 1) whether the model has time-varying data as variables 2) whether the focus is more on efficiency or interpretation. If the temporal change of variables is not important for the study, or the model aims at a high-efficiency solution, the Stay-Link network is recommended. Considering the timeliness of Wi-Fi sensor-data, this network has potential to provide a real-time prediction for city-tourist flow. If the study aims at fully “offline” estimation with a large number of explanatory variables being prepared, the temporal expanded network is more appropriate.

7.4 Future research directions

Chapter 4 concludes with pointing out shortcomings and future work directions. For one, to acknowledge is that the data extraction methodology cannot fully exclude all non-tourist observations. Overall, the results suggest that the methodology may possibly extract too many observations. For example, slow-moving vehicles are not caught by the filter. This may result in an overestimate of the utility of the trunk road, which is termed "main road dummy." Because the target area is not very large and the population composition is relatively simple (mainly walking tourists), a prior clustering of the samples was not employed. For future applications it might be useful to apply unsupervised learning methods (or even semi-supervised methods if ground-truth data are available) for sequential data. Hidden Markov models or some of the clustering approaches mentioned in the literature review that have been used with Bluetooth and Wi-Fi data may help to better distinguish walking tourists and recognize their behavior patterns.

Further, corrections regarding the correlation between subsequent links might be considered. A link size attribute that has been proposed to improve RL model results was not included. One reason is that the computational fast DeC approach proposed in Mai et al, (2018) utilized for the study does not allow the inclusion of such destination-specific attributes. To further suggest is that, since the network is close to a grid network, these attributes may not be as important as in the networks where they have been shown to be useful. Nevertheless, in future work, methodologies that are able to relax the "independence of irrelevant alternatives" (IIA) assumption, such as nested recursive logit and mixed recursive logit models, should be considered. Finally, to address the bias introduced by sensor specific penetration rates (or incomplete GPS records), it is believed that the introduced error term can to some degree solve the problem but acknowledged that alternative methods within the RL context should be explored that correct for the fact that the true destination is not known.

In Chapter 5, one of the future works is to improve the alignment of sequence data. The improvement can be destabilizing the time period separations, as well as introducing the ideology of dynamic time warping which eliminates the effect of travelling speed. Besides, the assignment of Wi-Fi traces enrichment is lack of validation for absence of ground truth. An alternative way for validating the

methodology is to compare the result (at least the aggregated ODs) with a third data source, such as the CDR data provided by telecom companies.

In Chapter 6, firstly, the grouping problem can be expanded. As concluded at the end of section 6.2, the problem space is highly non-convex and non-continuous so that with even longer search alternative solutions that are as good might be found. The problem can be further specified by, for example, whether a GPS record is associated to a sensor-group or not can be determined by knowledge on map, instead of by distance to the closet sensor. Also, heuristic algorithms for combinational optimization other than GA may be worth trying.

Secondly, a more flexible approach for associating POIs sensor-groups should be introduced. Each sensor-group can have different radiuses to associate POIs. Also, assuming sensors as the center of each group is not necessary if more data is available, e.g. “Photo-location data” from location-based social network service.

Thirdly, as discussed in section 6.2.4, alternatively there is a network with POIs’ attractiveness variables of each link that is derived by taking the difference from the corresponded variables of previous link. On this network an individual will only obtains the incentive associated to an attribute once at the moment of making a choice. Such network is more applicable when the duration of behavior is more detailed, for example, precise to the 10-minute, then the instantaneous incentive given by different links as well as the utility of staying/moving per time period is worth to be estimated for better interpretation.

Last but not least, the Wi-Fi packet sensing survey is considered also capable for obtaining data for activity modelling if the survey resolution is more detailed, for example the tourist for shopping can be distinguished directly if sensors are installed inside a shopping mall. Considering the budget limit, we suggest a survey start with a scale similar to the Higashiyama survey but with a greater number of sensors. By enough information for inferring activities, pattern recognition methodology for sequences such as the HMM can be employed to derive individuals’ activity chains, then the RL can be applied on activity modelling.

References

Alivand, M., Hochmair, H., & Srinivasan, S. (2015). Analyzing how travelers choose scenic routes using route choice models. Computers, Environment and Urban Systems, 50, 41-52.

Asano, M, Iryo, T., & Kuwahara, M. (2009). A pedestrian model considering anticipatory behaviour for capacity evaluation. In: Lam W., Wong S., Lo H. (eds) Transportation and Traffic Theory 2009: Golden Jubilee. Springer, Boston, MA.

Aschauer, F., Hössinger, R., Axhausen, K. W., Schmid, B., & Gerike, R. (2018). Implications of survey methods on travel and non-travel activities: A comparison of the Austrian national travel survey and an innovative mobility-activity-expenditure diary (MAED). European Journal of Transport and Infrastructure Research, 18(1), 4-35. Bellman, Richard. "A Markovian decision process." Journal of mathematics and mechanics (1957): 679-684.

Ben-Akiva, M., Bowman, J. L., & Gopinath, D. (1996). Travel demand model system for the information era. Transportation, 23(3), 241-266.

Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Computers, Environment and Urban Systems, 65, 126-139.

Borgers, A., & Timmermans, H. (1986). A model of pedestrian route choice and demand for retail facilities within inner‐city shopping areas. Geographical analysis, 18(2), 115-128.

Bowman, J. L., & Ben-Akiva, M. E. (2001). Activity-based disaggregate travel demand model system with activity schedules. Transportation Research Part A: Policy and Practice, 35(1), 1-28.

Chen, Y., Cai, Y., Li, P., & Zhang, G. (2015). Study on evacuation evaluation in subway fire based on pedestrian simulation technology. Mathematical Problems in Engineering, 2015. Crawford, F., Watling, D.P., & Connors, R.D. (2018). Identifying roader user classes based on repeated trip behaviour using Bluetooth data. Transportation Research Part A, 113, 55-74.

Cunche, M., Kaafar, M. A., & Boreli, R. (2012, June). I know who you will meet this evening! linking wireless devices using wi-fi probe requests. In 2012 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM) (pp. 1-9). IEEE.

Fosgerau, M., Frejinger, E. & Karlstrom, A. (2013). A link based network route choice model with unrestricted choice set. Transportation Research Part B: Methodological, 56: 70-80.

Freudiger, J. (2015). How talkative is your mobile device? An experimental study of Wi-Fi probe requests. In Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks (pp. 1-6).

Friis, C., & Svensson, L. (2013). Pedestrian Microsimulation. A comparative study between the software programs Vissim and Viswalk (Master's thesis).

Frytag, T. (2010). Visitor Activities and Inner-City Tourist Mobility: The case of Heidelberg. Chapter 11 in Analysing International City Tourism by Mazanaic and Wöber (Eds). Springer.

Fu, H., & Wilmot, C. G. (2004). Sequential logit dynamic travel demand model for hurricane evacuation. Transportation Research Record, 1882(1), 19-26.

Fukuda, D., Ihoroi, N., Nakanishi, W., Arimura, M., Asada, T., Uchida, K. E., & Suga, Y. (2018). Wi-Fi based Continuous Monitoring of Tourists' Travel Behavior: Results of Two Large-Scale Field Experiments in Japan (No. 305). EasyChair.

Ghanayim, M., & Bekhor, S. (2018). Modelling bicycle route choice using data from a GPS-assisted household survey. European Journal of Transport and Infrastructure Research, 18(2).

Gipps, P. G., & Marksjö, B. (1985). A micro-simulation model for pedestrian flows. Mathematics and computers in simulation, 27(2-3), 95-105.

Google Map (2020). Google Map API. Available from https: https://cloud.google.com/maps-platform/ [Accessed Feb 2020].

Hänseler, F. S., Bierlaire, M., Farooq, B., & Mühlematter, T. (2014). A macroscopic loading model for time-varying pedestrian flows in public walking areas. Transportation Research Part B: Methodological, 69, 60-80.

Helbing, D., Farkas, I. J., Molnar, P., & Vicsek, T. (2002). Simulation of pedestrian crowds in normal and evacuation situations. Pedestrian and evacuation dynamics, 21(2), 21-58.

Hidaka, K., Hayakawa, K., Nishi, T., Usui, T., & Yamamoto, T. (2019). Generating pedestrian walking behavior considering detour and pause in the path under space-time constraints. *Transportation Research Part C: Emerging Technologies, 108*, 115-129.

Hill, M. R. (1982). Spatial structure and decision-making aspects of pedestrian route selection through an urban environment.

Ho, C., & Mulley, C. (2013). Tour-based mode choice of joint household travel patterns on weekend and weekday. Transportation, 40(4), 789-811.

Hoogendoorn, S. P., & Bovy, P. H. (2004). Pedestrian route-choice and activity scheduling theory and models. Transportation Research Part B: Methodological, 38(2), 169-190.

Kaneko, N., Oka, H., Chikaraishi, M., Becker, H., & Fukuda, D. (2018). Route choice analysis in the Tokyo Metropolitan Area using a link-based recursive logit model featuring link awareness. Transportation Research Procedia, 34, 251-258.

Kitazawa, K., & Fujiyama, T. (2010). Pedestrian vision and collision avoidance behavior: Investigation of the information process space of pedestrians using an eye tracker. In Pedestrian and evacuation dynamics 2008 (pp. 95-108). Springer, Berlin, Heidelberg.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. The annals of mathematical statistics, 22(1), 79-86.

Kyoto City. (2017) Kyoto City Official Website. “平成 28 年 京都観光総合調査”. (In Japanese). Available from <https://www.city.kyoto.lg.jp/sankan/page/0000222031.html>. [Accessed July 2020].

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. Mathematical programming, 45(1-3), 503-528.

Legion SpaceWorks. (2012). Functional Description. London: Legion Limited. (Software)

Lue, G., & Miller, E. J. (2019). Estimating a Toronto pedestrian route choice model using smartphone GPS data. Travel behaviour and society, 14, 34-42.

Mai, T., Bastin, F., & Frejinger, E. (2018). A decomposition method for estimating recursive logit based route choice models. EURO Journal on Transportation and Logistics, 7(3), 253-275.

Mai, T., Fosgerau, M., & Frejinger, E. (2015). A nested recursive logit model for route choice analysis. Transportation Research Part B: Methodological, 75, 100-112.

Martchouk, M., Mannering, F., Bullock, D., 2011. Analysis of freeway travel time variability using Bluetooth detection. Journal of Transportation Engineering 137(10), 697-704.

Marra, A. D., Becker, H., Axhausen, K. W., & Corman, F. (2019). Developing a passive GPS tracking system to study long-term travel behavior. Transportation research part C: emerging technologies, 104, 348-368.

Martén, J. B., & Henningsson, J. (2014). Verification and Validation of Viswalk for Building Evacuation Modelling. Journal, Vol, 5, 135-144.

Maruyama, T., Sato, Y., Nohara, K., & Imura, S. (2015). Increasing smartphone-based travel survey participants. Transportation Research Procedia, 11, 280-288.

Ministry of Land, Infrastructure, Transport and Tourism. (2020). 近畿圏パーソントリップ調査（実態調査）の延期について, https://www.kkr.mlit.go.jp/plan/pt/index.html

Monte Malveira, D. (2019). Analysis of Walking and Route-Choice Behavior of Pedestrians inside Public Transfer Stations: A Study on how pedestrians behave in the approaching vicinity of level-change facilities, and how it affects their walking and route-choice behavior. Independent thesis, KTH Stockholm. Available from <http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-264755>. [Accessed July 2020].

Nishida, J., Adachi, T., & Makimura, K. (2014). Traffic Flow Analysis by the Use of Wi-Fi Packets Receiver. In 1st IRF Asia Regional Congress & Exhibition.

Oosterlinck, D., Benoit, D. F., Baecke, P., & Van de Weghe, N. (2017). Bluetooth tracking of humans in an indoor environment: An application to shopping mall visits. Applied Geography, 78, 55-65.

OpenStreetMap contributors. (2015). Data file from June 2018. Retrieved from https://planet.openstreetmap.org. [Accessed November 2019].

Ota K., Omura M., Tsujido F., Asao K. & Nishida J. (2019) Wi-Fi 歩行者流動センサによる計測値からの 実数推定手法. 第 57 回土木計画学研究発表会・講演集 02-07.

Oyama Y. and Hato E. (2017). A discounted recursive logit model for dynamic gridlock network analysis. Transportation Research Part C: Emerging Technologies, 85: 509-527.

Park, J. H., Rojas, F. A., & Yang, H. S. (2013). A collision avoidance behavior model for crowd simulation based on psychological findings. Computer Animation and Virtual Worlds, 24(3-4), 173-183.

PTV Group. (2011). PTV Viswalk. Karlsruhe: PTV Group. (Software)

Quain, J.R, (2018). Cities looking to get smart take a lesson from an iconic shopping mall. Digital trends. Online magazine. Available from: <https://www.digitaltrends.com/ features/bluetooth-beacons-and-rfid-bands-the-mall-of-america-is-a-really-smart-city/>. [Accessed July 2020].

Robusto, C. C. (1957). The cosine-haversine formula. The American Mathematical Monthly, 64(1), 38-40.

Thomas, T., Geurs, K. T., Koolwaaij, J., & Bijlsma, M. (2018). Automatic trip detection with the Dutch mobile mobility panel: towards reliable multiple-week trip registration for large samples. Journal of Urban Technology, 25(2), 143-161.

Turner, A. (2001). Depthmap: a program to perform visibility graph analysis. Proceedings of the 3rd International Symposium on Space Syntax. Atlanta, GA, Georgia Institute of Technology, Vol. 31.

Urata, J., & Hato, E. (2013). Modeling social interactions between households for evacuation behaviors in the devasted areas. In European Transport Conference 2013Association for European Transport (AET).

Urban Planning Bureau of Kyoto City (2006). Trend survey of tourists in Kyoto city. Questionnaire survey. Internal report of Kyoto City. (In Japanese).

Vanhoef, M., Matte, C., Cunche, M., Cardoso, L. S., & Piessens, F. (2016). Why MAC address randomization is not enough: An analysis of Wi-Fi network discovery mechanisms. In Proceedings of the 11th Asia Conference on Computer and Communications Security (ASIA CCS'16), 413-424.

Versichele, M., Neutens, T., Delafontaine, M., & Van de Weghe, N. (2012). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. Applied Geography, 32(2), 208-220.

Västberg, O. B., Karlström, A., Jonsson, D., & Sundberg, M. (2020). A dynamic discrete choice activity-based travel demand model. *Transportation Science*, *54*(1), 21-41.

Versichele, M., Neutens, T., Delafontaine, M., & Van de Weghe, N. (2012). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent Festivities. *Applied Geography*, *32*(2), 208-220.

Wang, W. L., Lo, S. M., Liu, S. B., & Kuang, H. (2014). Microscopic modeling of pedestrian movement behavior: Interacting with visual attractors in the environment. Transportation Research Part C: Emerging Technologies, 44, 21-33.

Westerdijk, P. K. (1990). Pedestrian and pedal cyclist route choice criteria. Yen, J. Y. (1971). Finding the k shortest loopless paths in a network. Management Science, 17(11), 712-716.

Zimmermann, M., Mai, T., & Frejinger, E. (2017). Bike route choice modeling using GPS data without choice sets of paths. Transportation Research Part C: Emerging Technologies, 75, 183-196.

Zimmermann, M., Västberg, O. B., Frejinger, E., & Karlström, A. (2018). Capturing correlation with a mixed recursive logit model for activity-travel scheduling. *Transportation Research Part C: Emerging Technologies*, *93*, 273-291.

Appendix

Appendix 1: Defined tourist area in English

No.	Area (In Japanese)	Area (In English)
1	大原・八瀬方面	Ohara&Yase
2	鞍馬方面	Kurama area
3	宝ヶ池方面	Takaragaike
4	上賀茂神社周辺	Kamigamo Shrine
5	高雄方面	Takao
6	修学院・詩仙堂周辺	Shugakuin&Shisendo
7	光悦寺周辺	Koetsu-ji Temple
8	北山通周辺	Kitayama Dori
9	大徳寺周辺	Daitokuji Temple
10	金閣寺周辺	Kinkaku-ji Temple
11	下鴨神社周辺	Shimogamo Shrine
12	北野天満宮周辺	Kitano Tenmangu
13	衣笠・御室方面	Kinugasa&Omuro
14	嵯峨野方面	Sagano Area
15	銀閣寺周辺	Ginkaku-ji Temple
16	哲学の道周辺	The Path of Philosophy
17	平安神宮周辺	Heian Jingu Shrine
18	御所周辺	Kyoto Imperial Palace
19	花園方面	Hanazono area
20	二条城周辺	Nijo Castle Area
21	二条駅周辺	Nijo Station Vicinity
22	太秦方面	Uzumasa area
23	嵐山方面	Arashiyama Area
24	祇園方面	Gion Area
25	河原町・新京極方面	Kawaramachi
26	松尾大社周辺	Matsuo Taisha area
27	清水寺周辺	Kiyomizu-dera Temple
28	三十三間堂周辺	Sanjusangendo Area
29	京都駅周辺	Kyoto Station Vicinity
30	桂離宮周辺	Katsura Imperial Villa
31	東福寺周辺	Tofukuji Temple Area
32	東寺周辺	Toji Temple Area
33	伏見稲荷大社周辺	Fushimi Inari Shrine
34	醍醐寺周辺	Daigoji Temple Area
35	城南宮周辺	Jonan-gu Shrine Area
36	伏見周辺	Fushimi Area
37	京北方面	Keihoku Direction

Appendix 2: Detail profile of sensor deployment.

Marker on the map	Location (In Japanese)	Location	Installation details
Train stations			
1 in yellow	東山三条	Higashiyama Sanjyo Station, Kyoto City Subway	At gate
2 in yellow	三条（市営地下鉄）	Sanjyo Station, Kyoto City Subway	At gate
3 in yellow	京都（市営地下鉄）	Kyoto Station, Kyoto City Subway	At central gate
4 in yellow	京都（市営地下鉄）	Kyoto Station, Kyoto City Subway	At south gate
5 in yellow	三条（京阪電鉄）	Sanjyo Station, Keihan Railway	At gate
6 in yellow	祇園四条（京阪電鉄）	Shijyo Station, Keihan Railway	At gate
7 in yellow	清水五条（京阪電鉄）	Gojyo Station, Keihan Railway	At gate
8 in yellow	阪急河原町	Kawaramachi Station, Hankyu Railway	At gate
1 in blue	JR 京都-地下中央口	Kyoto Station, Japan Railway	At underground central gate
2 in blue	JR 京都-西洞院口	Kyoto Station, Japan Railway	At “Nishinotoin” gate
3 in blue	JR 京都-北口広場	Kyoto Station, Japan Railway	In square of north gate
11 in blue	JR 京都-西口	Kyoto Station, Japan Railway	At ear west gate
12 in blue	JR 京都-中央口	Kyoto Station, Japan Railway	At Central gate
13 in blue	JR 京都-八条東口	Kyoto Station, Japan Railway	At “Hachijo” gate
14 in blue	JR 京都-地下東口	Kyoto Station, Japan Railway	At Underground east gate

4 in blue	JR 山科	Yamashina Station, Japan Railway	At gate
5 in blue	JR 東福寺	Tofukuji Station, Japan Railway	At gate
6 in blue	JR 東福寺	Tofukuji Station, Japan Railway	At transfer gate
7 in blue	JR 稲荷	Inari Station, Japan Railway	At gate
8 in blue	JR 二条	Nijo Station, Japan Railway	At gate
9 in blue	JR 円町	Marmachi Station, Japan Railway	At gate
10 in blue	JR 嵯峨嵐山	Saga Arashiyama Station, Japan Railway	At gate
Tourism spots			
1 in pink	清水寺	Kiyomizu Temple	Near square of the main entrance
2 in pink	銀閣寺	Ginkakuji Temple	In front of a shop along the approach to Ginkakuji Temple
3 in pink	錦市場	Nishiki market	On the eaves of the first floor of a store
4 in pink	二条城	Nijo Castle	Near entrance
5 in pink	嵐山渡月橋	Arashiyama Togetsukyo Bridge	In front of a shop along the footpath to Togetsukyo Bridge
6 in pink	先斗町	Ponto-cho	In front of a restaurant located at midway of ponto-cho's cobbled alley
7 in pink	八坂神社	Yasaka Shrine	On the top of a vending machine near Mai-dono hall
8 in pink	高台寺	Kodaiji Temple	Near the ticket office
9 in pink	平安神宮	Heian Shrine	Near the entrance

10 in pink	伏見稲荷大社	Fushimi Inari shrine	On the top of a vending machine along the approach to Ginkakuji Temple
9 in yellow	知恩院	Chionin Temple	On the top of a vending machine in parking lot
10 in yellow	祇園	Gion	Inside the northbound bus stop
11 in yellow	祇園	Gion	Inside the southbound bus stop
12 in yellow	五条坂	Gojo-zaka	Inside the bus stop
13 in yellow	円山公園	Maruyama Park	Inside parking lot
14 in yellow	高台寺	Kodaiji Temple	Inside parking lot
15 in yellow	清水坂	Kiyomizu-zaka	On the top of a vending machine inside parking lot