Spatiotemporal Learning and Geo-visualization Methods for Constructing Activity-travel Patterns from Transit Card Transaction Data

By

Yi Zhu

M.U.P & M.S.
University of Wisconsin - Milwaukee
(2007)

ARCHIVES
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
OCT 07 2014
LIBRARIES

Submitted to the Department of Urban Studies and Planning
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Urban and Regional Planning

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2014

© 2014 Yi Zhu. All Rights Reserved

The author hereby grants to MIT the permission to reproduce and to distribute publicly paper and electronic copies of the thesis document in whole or in part in any medium now known or hereafter created.

Signature redacted

Author____
Department of Urban Studies and Planning
September 8, 2014

Signature redacted

Certified by _______________________
Professor Joseph Ferreira Jr.
Professor of Urban Planning and Operations Research
Dissertation Supervisor

Signature redacted

Accepted by__________________________
Professor Lawrence J. Vale
Chair, PhD Committee
Department of Urban Studies and Planning

Spatiotemporal Learning and Geo-visualization Methods for Constructing
Activity-travel Patterns from Transit Card Transaction Data

By

Yi Zhu

Submitted to the Department of Urban Studies and Planning
On September 2014, in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy in Urban and Regional Planning

ABSTRACT:

The study of human activity-travel patterns for transportation demand forecast has evolved a long way in theories, methodologies and applications. However, the scarcity of data has become a major barrier for the advancement of research in the field. At the same time, the proliferation of urban sensing and location-based devices generate voluminous streams of spatio-temporal registered information. These urban sensing data contain massive information on urban dynamics and individuals' mobility. For example, the transit smart card transaction data reveal the places that transit passengers visit at different times of day. As tempting as it appears to be, the incorporation of these urban sensing data into activity-travel study remains a big challenge, which demands new analytics, theories and frameworks to bridge the gap between the information observed directly from the imperfect urban sensing data and the knowledge about how people use the city.

In this study, we propose a framework of analysis that focuses on the recurring processing and learning of voluminous transit smart card data flows in juxtaposition with additional auxiliary spatio-temporal data, which are used to improve our understanding of the context of the data. The framework consists of an ontology-based data integration process, a built environment measurement module, an activity-learning module and visualization examples that facilitate the exploration and investigation of activity-travel patterns. The ontology-based data integration approach helps to integrate and interpret spatio-temporal data from multiple sources in a systematic way. These spatio-temporally detailed data are used to formulate quantitative variables for the characterization of the context under which the travelers made their transit trips. In particular, a set of spatial metrics are computed to measure different dimensionalities of the urban built environment of trip destinations. In order to understand why people make trips to destinations, researchers and planners need to know the possible activities associated with observed transit trips. Therefore, an activity learning module is developed to infer the unknown activity types from millions of trips recorded in transit smart card transactions by learning the context dependent behaviors of travelers from a traditional household travel survey. The learned activities not only help the interpretation of the behavioral choices of transit riders, but also can be used to improve the characterization of urban built form by uncovering the likely activity landscapes of various places. The proposed framework and the methodology is demonstrated by focusing on the use of transit smart card transaction data, i.e., EZ-Link data, to study activity-travel patterns in Singapore.

Although different modules of the framework are loosely coupled at the moment, we have tried to pipeline as much of the process as possible to facilitate efficient data processing and analysis. This allows researchers and planners to keep track of the evolution of human activity-travel patterns over time, and examine the correlations between the changes in activities and the changes in the built environment. The knowledge gained from continuous urban sensing data will certainly help policy makers and planners understand the current states of urban dynamics and monitor changes as transportation infrastructure and travel behaviors evolve over time.

Thesis Supervisor: Joseph Ferreira, Jr.
Title: Professor of Urban Planning and Operations Research

Acknowledgement

I would like to take the opportunity to express my gratitude to all the people who have helped and supported me during my study at the Massachusetts Institute of Technology.

I'm particular grateful for my advisor Professor Joseph Ferreira for his guidance and assistance throughout the period of my PhD study. He guided me to the areas of inquiry and inspired me to focus on the processing and visualization of urban sensing data in combination with geospatial data. Meanwhile, I have been influenced strongly by his attitude and passion on research and life. I'm truly honored to learn from him and work with him. I'm also very grateful for the other two members of my dissertation committee, Professor P. Christopher Zegras and Professor Marta C. Gonzalez, who have helped significantly in the course of my research and study. I was benefited intensively by their creative ideas and insightful comments.

The thesis work would not have been possible without the support from the SimMobility project funded by the Singapore National Research Foundation (NRF) through the Singapore-MIT Alliance r Research and Technology (SMART) Center for Future Mobility (FM). I want to thank for valuable inputs from Prof. Moshe Ben-Akiva and Francisco C. Pereira. I also acknowledge the contributions of other collaborators at MIT and in the FM team including Prof. Mi Diao, Xiaosu Ma and Pedro J.S. Gandol. In addition, we appreciate the support of Singapore Urban Redevelopment Authority (URA) and Singapore Land Transport Authority (LTA) on the EZ-Link data, HITS dataset, the REALIS dataset and other helpful information.

I want to thank Sandy Wellford, Kirsten Greco, and Sue Delaney for their distinguished administrative assistance and CRON for their technical and computing support.

My gratitude also goes to numerous colleagues, friends, and mentors beyond the dissertation committee and the SMART program: Karan Polensky, Jinhua Zhao, Michael Flaxman, Pengfei Ye and Xiaoting Jia, Mi Diao and Zuhua Cai, Weifeng Li and Jie Xia, Zhiyong Wang and Chunguang Wang among others.

Lastly, I would like to express my sincerest appreciation to my beloved wife, Zi Ye, for her unconditional love and caring. I would like to express special gratitude to my parents for their constant support. I dedicate my dissertation work to them.

Table of Contents

Chapter 1. Introduction 14

1.1 Background 14

1.2 Statement of the Problem 14

1.3 Research Framework and Objectives 16

1.4 Summary of Chapters 18

Chapter 2. Exploratory Analysis of Transit Smart Card Transaction Data 20

2.1 Introduction 20

2.2 Related Work 20

2.2.1 Urban sensing data and activity-travel research 20

2.2.2 Spatiotemporal features of activity-travel pattern 21

2.3 Urban development and transit system of Singapore 23

2.3.1 Heterogeneous urban spaces 23

2.3.2 Urban transit services 24

2.4 Smart Card Transaction Data 27

2.5 Exploratory Analysis of the EZ-Link data 28

2.5.1 Transit trip frequency 28

2.5.2 Temporal travel patterns 30

2.5.3 Spatial Pattern 33

2.5.4 Transfer Pattern 34

2.6 Summary 35

Chapter 3. Integration of Spatial-temporal Data for Activity-travel Learning 37

3.1 Introduction 37

3.2 Background and related work 38

3.2.1 Emerging data sources 38

3.2.2 Heterogeneous datasets 39

3.2.3 Existing data integration approaches 39

3.2.4 Ontology-based data integration 40

3.3 Data collection 41

3.4 Proposed approach of data integration for the activity-travel analysis 42

3.4.1 General framework of the proposed data integration approach 42

3.4.2 Creation of ontology to support semantic interoperability 43

3.4.3 Schema-mapping between data sources and ontologies 47

3.4.4 Data cleaning and normalization 49

3.4.5 Data quality assessment 50

3.4.6 Integration operations 53

3.5 Integrating building information for built environment measurement 53

3.6 Summary 60

Chapter 4. Measuring and Characterizing the Urban Built Environment 62

4.1 Introduction 62

4.2 Related work 63

4.2.1 Urban Planning and Physical Activity 63

4.2.2 Landscape Ecology 64

4.2.3 Urban Design 64

4.2.4 Gaps in the literature 65

4.3 Data and Methods 66

4.3.1 Catchment Areas of Transit Stations 67

4.3.2 Importance of Business Open Hours 69

4.4 Built Environment Measurement 72

4.4.1 Framework of the Spatial Metrics Computational Procedure 72

4.4.2 Land Use Mix 74

4.4.3 Buildings and Street Layout 78

4.4.4 Spatial distribution of Establishments 84

4.4.5 Transit Network Topology 89

4.4.6 Correlations among indicators 94

4.5 Summary 94

Chapter 5. Learning Transit-oriented Activity Types from a Household Travel Survey ... 97

5.1 Introduction 97

5.2 Methodology 98
5.2.1 Modeling Framework 99
5.2.2 Multinomial Logit Regression model 101
5.2.3 Conditional Random Fields model 102
5.2.4 The Relation between CRFs and the MNL model 104
5.3 Model selection 105
5.3.1 Spatial and temporal confounding effects 105
5.3.2 Overfitting 106
5.3.3 Imbalanced classes 107
5.3.4 Model validation 108
5.4 Data and variables 109
5.4.1 Activity categorization 109
5.4.2 Temporal interval 110
5.4.3 Imputing the boarding and alighting stops 110
5.4.4 Weather 111
5.5 Modeling results and assessment 111
5.5.1 Learning Models 111
5.5.2 Effect of spatiotemporally detailed urban built environment measures 113
5.5.3 Effect of model regularization 117
5.5.4 Effect of resampling of the minority class 120
5.5.5 Effect of the transition of activity types 121
5.6 Summary of Findings 122
Chapter 6. Activity Inference and Visualization 124
6.1 Introduction 124
6.2 Home Location Extraction 124
6.3 Activity Inference 129
6.3.1 Inference using MNL models 129
6.3.2 Inferring Activity Types using CRFs models 129
6.4 Examples of Exploring and Visualizing Activity-travel Patterns 129
6.4.1 Single Transit Trips 130

6.4.2 Individual Activity-Travel pattern 133
6.4.3 Activity Landscape of places 137
6.5 Implications to activity-travel research and planning practice 138
6.6 Summary 140
Chapter 7. Discussion and Conclusions 141
7.1 Major Findings 142
7.2 Research limitations and Future Work 144
7.2.1 Built Environment Measurement 144
7.2.2 Learning Models.. 145
7.2.3 Assumptions and Speculations 146
7.2.4 Data exploration and Visualization 147
7.3 Conclusions 147
REFERENCES 149
Appendix I. Building Type Categorizations in the Data Integration.............................. 158
Appendix II Sectorial Categorization of Establishments 160
Appendix III. Model Estimation Results 162

List of Figures

Figure 1-1.The proposed framework for using urban sensing data for activity-travel analysis 17
Figure 2-1. Land use plan (2008) of Singapore 24
Figure 2-2. Transit stations of the public transit system in Singapore 26
Figure 2-3. Number of vehicle trips along time of day broken down by the types of services 26
Figure 2-4. Percent of the number of days EZ-link cards were used from April 11 to April 17, 2011. 29
Figure 2-5. Average number of transit trips per day by the types of card holders 30
Figure 2-6. The temporal distribution of the first and the last transit trips of the transit smart card holders with at least two transactions on April 13, 2011 31
Figure 2-7. Temporal distribution of first and last transit trips for transit smart card holders with at least two transactions from April 11 to April 15, 2011. 32
Figure 2-8. Thematic maps of the transit stations by alighting passengers at different time of day. 33
Figure 2-9. The spatial distribution of bus-to-bus or bus-subway-LRT transfers in Singapore observed in the EZ-Link data 35
Figure 2-10. Distribution of the transfer time in seconds extracted from the EZ-Link data 35
Figure 3-1. Framework of the Proposed Ontology-based data integration approach 43
Figure 3-2. An example of the urban building ontology 46
Figure 3-3. Screenshots of the Protégé 47
Figure 3-4. Building footprints by categories from SLA. 54
Figure 4-1. Approximate walking distance of transit stations. 68
Figure 4-2. Results of the Random Forest (RF) model for open hour classification 72
Figure 4-3. Framework of urban built environment metric computations 73
Figure 4-4. Spatial distribution of land use entropy index at the transit station catchment area level. 75
Figure 4-5. Spatial distribution of the Contagion Index 77
Figure 4-6. Transit station catchment areas with the noticeable presence of parks 77
Figure 4-7. Spatial distribution of the average weighted mean shape index (AWMSI) ... 78
Figure 4-8. Building compactness index of the catchment areas of transit stations........82
Figure 4-9. Building centrality index of the catchment areas of transit stations 82
Figure 4-10. The nearest neighbor index of the catchment areas of transit stations 83

Figure 4-11. Density of Local Road (> 0.05 m per square meter) 84
Figure 4-12. Spatial distribution of establishments by INDCD categories 87
Figure 4-13. Time varying entropy index and the nearest neighbor index of the establishments in the catchment areas of three selected stations. 88
Figure 4-14. Illustration of the urban transit network topology representation 89
Figure 4-15. Frequency of transit services by road segments at different times of day.. 90
Figure 4-16. Correlation matrix of the selected built environment measures 96
Figure 5-1. Illustration of activity learning and inference from the EZ-Link data............. 99
Figure 5-2. Analytical framework of the activity-learning and inference module 100
Figure 5-3. Illustration of activity learning models 104
Figure 5-4. Illustration of SMOTE approach 108
Figure 5-5. 5-fold cross validation accuracy rate by activity types 116
Figure 5-6. Regularized MNL with five-fold cross validation. 118
Figure 5-7. 5-fold cross validation accuracy by activity types (from MNL5 to CRF1) 119
Figure 5-8. Activity transition coefficients from the linear-chain CRFs model 122
Figure 6-1. Illustration of the procedure of extract home locations from the EZ-link data 127
Figure 6-2. Spatial pattern of single transit trips in the EZ-link data on April 13, 2011. 131
Figure 6-3. Pattern of single trips ending at the Boon Lay bus interchange. 132
Figure 6-4. Visualizing individual activity travel pattern. 135
Figure 6-5. The temporal pattern of transit trips 136
Figure 6-6. Visualizing the activity-travel landscape of the Chinatown MTR station..... 138

List of Tables

Table 2-1. Average daily ridership 25

Table 2-2. The EZ-Link smart card transaction data sample........ 27

Table 2-3. Number of transfers per trip in the EZ-Link data 34

Table 3-1. Types of data heterogeneities and examples 39

Table 3-2. Types of correspondences specified in the urban modelling ontology example 44

Table 3-3. A look-up table for schema mapping between local datasets and global ontology 49

Table 3-4. A Summary of Data Quality Assessment Measure 52

Table 3-5. A list of datasets available for building information integration 55

Table 3-6. Data quality measures for the type of buildings 56

Table 3-7. Number of building records that are classified in the process of integration........ 59

Table 4-1. Walking time to public transportation stations 68

Table 4-2. Ten most prevalent open hour patterns 70

Table 4-3. Spatial metrics for land use patches 76

Table 4-4. Spatial metrics for buildings and street layout measurement 80

Table 4-5. Measures of spatial distribution of businesses 86

Table 4-6. Transit network measures 93

Table 5-1. Activity classification based on HITS 2008 dataset 109

Table 5-2. List of learning models and assessment measures 112

Table 5-3. List of variables included in the model specification 112

Table 5-4. Estimation and validation measures of traditional MNL activity-learning models 113

Table 5-5. Effect of the selected explanatory variables on activity types from the MNL5........ 116

Table 5-6. Effect of the selected explanatory variables on activity types from the MNL6........ 120

Table 5-7. SMOTE augmented samples 120

Table A1. Results of MNL1 model 162

Table A2. Results of MNL2 model 163

Table A3. Results of MNL3 model 166

Table A4. Results of MNL4 model 169

Table A5. Results of MNL5 model 172

Table A6. Results of MNL6 model 175

Chapter 1. Introduction

1.1 Background

Cities of the 21st century are increasingly concerned with ever growing challenges on urban ecology sustainability and disaster prevention, aging population as well as the impacts of economic recession, in addition to modern urban issues such as pollution, sprawl, and congestion. To urban policy makers, the complexity in decision making process is escalating as a result of various trends including decentralization, gentrification, deindustrialization, globalization and energy conservation occurring simultaneously in the context of urban management. Meanwhile, behaviors and preferences of people and firms have been changing rapidly as a result of the explosion of information, improved mobility and more differentiated lifestyles, which also require urban management strategies to be more adaptive and responsive to the issues exposed. Success in urban management calls for innovative decision making paradigms that can provide timely and informative analyses and forecasts grounded on a sound and profound understanding of the dynamics of urban systems.

From the point view of transportation and land use planning and management, knowledge of peoples' activity-travel patterns is important, because activities and activity-derived travels are indicative of demands for transportation services as well as other urban services and opportunities for various activities. Moreover, the spatiotemporal distributions of people and activities are also crucial for a series of urban management tasks like emergency planning, disaster management, infrastructure services and resources allocation (Krygsman et al, 2007). For individuals, households and firms, activity and travel preferences are correlated with the choices of job, housing and vehicle ownership, firm location as well as lifestyles and business operations, which together constitute the social, economic and cultural silhouette of a city's evolution. Hence, to capture a city's pulse, it is essential to understand the activity-travel patterns of its citizens.

1.2 Statement of the Problem

In the last couple of decades, trips and activities are mainly studied within the framework of travel demand forecasting, which has evolved a long way from the trip-based, aggregated models like the four-step model to the activity-based, disaggregated models (McNally, 2000). However, the complexity of activities and travels, and the enormous factors that may influence people's choices and behaviors pose a big challenge for predicting activities and travels analytically. For example, despite the energetic research effort on activity-travel theory and modeling, questions like the generation and scheduling of daily activities are still not well understood (Bowman and Ben-Akiva, 2001). Moreover, our understanding of the impacts of the urban built environment on individual decisions about activities is also inconclusive. The predicament of the travel demand forecasting paradigm is in part related to data. The household travel survey or the time use survey used in the four step modeling or the activity-based modeling, are typically expensive and time consuming to collect. As a result, the majority of household travel surveys or time use surveys only collect the trip and activity information of respondents for one day. But it has been well recognized that individual and households' activity-travel behaviors present not only daily variation but also day-to-day and weekday-weekend variations. Because of the high cost, the surveys are typically conducted once every a few years, which provide limited empirical evidence on how people respond to external changes of built environment or urban policies since many changes could have occurred between two surveys.

On the other side, the surveillance devices and wireless sensor networks in cities as well as the personal communication devices and location aware devices have been supplying massive data every day. These urban sensing data are generally spatially and temporally tagged, big in terms of sample size and having a longitudinal coverage. These characteristics are virtually not found in the conventional survey data, which makes urban sensing data distinctive in revealing the aspects of urban dynamics and activity-travel patterns that are not well predicted by activity-based modeling. For example, the large volume of urban sensing data allows researchers to explore the patterns resultant from people's activities and movement in different areas of city and at different time points, which helps to expose the aberrant movement patterns under various circumstances that are worth further inspection. Longitudinal data streams also make it possible to track individuals' behavioral responses to the changes in transportation policy or land use. As pointed out by Jin and Batty (2013), one important value of the emergent data for urban study is that these data may "stretch our notion of the system and problems of cities that we might model".

Nevertheless, urban sensing data are also imperfect for activity-travel research. In the process of investigation, we found the following challenges need to be confronted before trying to make urban sensing data useful for urban studies and planning.

1) Most urban sensing data lack the information on activities and trips in which people participate, as well as the social - economic characteristics of the travelers because of the anonymity nature of urban sensing data. This prevents urban sensing data from being interpreted and understood properly.
2) The value and meaning of urban sensing data are reduced if it is taken out of context. To make the best sense possible from urban sensing data, it is crucial to couple urban sensing data with additional information describing the contexts.
3) It is impractical to manually process and analyze the voluminous urban sensing data that expand rapidly. Thus, there is a need for the analysis and data processing techniques applicable in the big data environment.

Obviously, it is inappropriate to analyze these data through tradition means. In this study, we propose an innovative framework of analysis to address these challenges. The framework and the methodology is demonstrated by focusing on the use of transit smart card transaction data, i.e., EZ-Link data, to study activity-travel patterns in Singapore by putting emphasis on the movement trajectories of individuals, the activity landscapes of urban space, and the types of interactions among travelers and urban built environment.

1.3 Research Framework and Objectives

Figure 1-1 shows the proposed analysis framework for the exploration and data mining of urban sensing data for activity-travel study. The framework consists of six consecutive parts: preliminary exploration, integration of spatially and temporally detailed auxiliary data, and formulation of spatial metrics, activity learning models, activity inference from urban sensing data, as well as the visualization and interpretation of results. The idea is to make the best use of observable information in urban sensing data, i.e., location, time and sequence, along with additional data from other sources to reconstruct the contexts of the activities and trips contained in urban sensing data. Spatiotemporal variables, in particular the urban built environment measures, are generated to characterize the contexts. By using statistical learning approaches like probabilistic graphical models (Koller and Friedman, 2009) and visualization tools, we expect to extract the correlations between activity types and the spatiotemporal explanatory variables from the traditional household travel survey. Then, the learning model is applied to infer the unknown activity types associated with urban sensing data. The inferred activity types help to interpret the travel or movement patterns revealed by urban sensing data by making connections to the types of places travelers visited.

Indeed, the proposed analysis approach requires urban sensing data to be coupled with the data with high spatial and temporal resolutions and details. This is not only because spatially detailed data provide better characterization of the urban built environment, but also due to the fact that the majority of urban activities only take place at a small portion of urban areas, consisting of local concentrations of built-up lands, buildings and streets. Assessing a place using indicators calculated at zonal scale is likely to out-off-focus the parts of urban form that directly interact with humans.

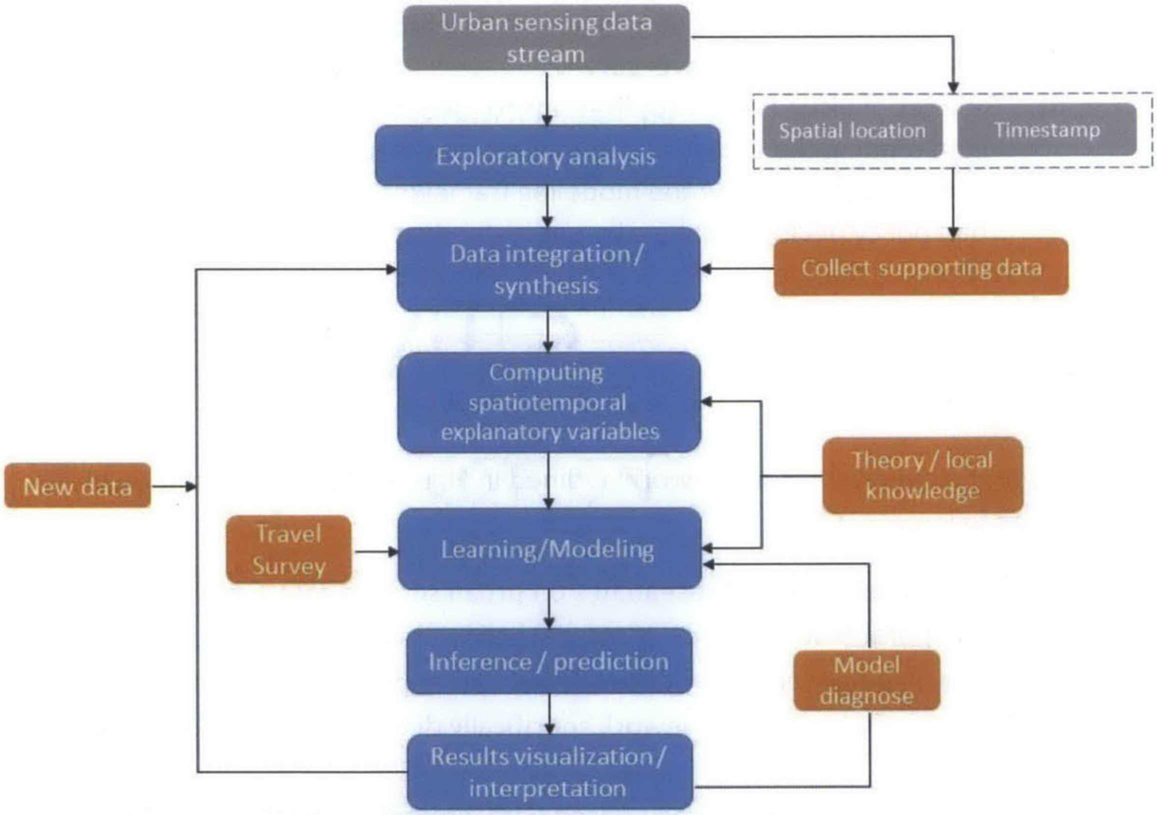


Figure 1-1.The proposed framework for using urban sensing data for activity-travel analysis

Detecting unknown activity types and other activity related components is intended to address the challenge 1). Integrating spatio-temporal data from multiple sources and develop spatial metrics to characterize urban built environment represents the effort to tackle the challenge 2). But the purpose of the study is more than those. The framework is designed to modularize and pipeline the fusing, processing, exploring and mining steps of urban sensing data and the auxiliary geospatial datasets, so that urban sensing data streams can be examined routinely without the need for heavy manual work. Meanwhile, we take into consideration the flexibility and scalability of the framework to make it more adaptable to the different needs of planning applications.

To be more specific, a number of research objectives are determined as following:

1) Propose a framework that enables repetitive exploring and learning of dynamic activity-travel behaviors from emerging data sources including both urban sensing data and other supporting datasets from a variety of sources.
2) Use spatio-temporal detailed data integrated from multiple sources, including official and crowdsourcing information to enrich the measurement of urban built environment.
3) Reconstruct unknown activity patterns by learning activity types from traditional household travel survey.
4) Use a set of visualization applications to assist the interpretation and understanding the activities and movement revealed by urban sensing data.
5) Provide insights on how the modeling framework can be used to inform urban policy decision makers by understanding activity-travel patterns in a responsive manner.

1.4 Summary of Chapters

According to the overarching framework outlined in Figure 1-1, the dissertation is organized as follows:

- Chapter 2 reviews related researches on urban sensing data and the spatiotemporal feature of activity-travel behaviors in the existing literature. Then, a preliminary analysis of the EZ-Link data is presented to provide a base for further analysis under a framework specifically developed for learning activities from big data.
- Chapter 3 proposes an ontology-based data integration mechanism (Gardner, 2005) to merge relevant urban spatial datasets from a variety of sources. The spatiotemporally detailed dataset obtained in this step will be used subsequently to support more precise characterization of the urban built environment and the improved analysis and learning efforts for the EZ-Link data.
- In Chapter 4, we review the literature on the urban built environment measurement in several fields. A comprehensive set of spatial metrics is then formulated to account for the four dimensions of the built environment: land use, street layout, business distribution and transit network.

- Chapter 5 presents the statistical learning methodologies to classify activity types encapsulated in a traditional household travel survey, the Household Interview Travel Survey (HITS) of Singapore in 2008, using the spatiotemporal detailed urban built environment measures and other exogenous variables as predictors. Statistical learning methods like logistical regressions and Conditional Random Fields (CRFs) are employed to examine the degree to which different activity types are dependent on the contexts described by the predictors.
- Using the estimated learning model, in Chapter 6, millions of trips recorded in the EZ-Link data are interpreted to identify likely activity types. Visualization applications are developed to explore resultant activity patterns, and scrutinize the way that the transit network and urban spaces are used by the predicted human activities.
- Chapter 7 concludes and discusses the significance and limitation of the study as well as some directions for future research.

Chapter 2. Exploratory Analysis of Transit Smart Card Transaction Data

2.1 Introduction

The emergence of urban sensing data has been stimulating growing interest in applying urban sensing data to activity-travel behavioral research. In this Chapter, we first review the related researches in this field. Then, an exploratory analysis of the transit smart card transactions data provided by the Land Transport Authority (LTA) of Singapore is presented to reveal the general temporal and spatial patterns of transit system ridership.

2.2 Related Work

2.2.1 Urban sensing data and activity-travel research

Most of the studies in this field focused on the GPS or cell phone data because of the prevalence of GPS devices and mobile phones. For the GPS data, considerable efforts have been made to detect travel related features like routes and modes from passive GPS logs. Papinski, Scott and Doherty (2009) developed a geographic information system (GIS) to collect planned route information of 31 survey participants and compared them with the personal GPS data gathered from real trips to understand the decision making process of real-time route choices. Tsui and Shalaby (2006) used the maximum and average speeds as well as the rate of acceleration derived from the GPS data to estimate travel mode and achieved moderate accuracy. The accuracy can be further improved if GPS trajectories are coupled with geospatial information of the study area such as public transportation network (Chung and Shalaby, 2005; Stopher et al., 2008). A few researchers have looked into GPS data to explore the possibility of extracting trip purpose. Wolf et al. (2006) used the GPS data from a Swedish study and match them with local point of interest (POI) and land use maps to infer destination types. Then, the researcher compare the result with the 2000 and the 2001 Swedish national travel survey to determine the most probable trip purpose conditioned by socio-demographic characteristics of respondents.

For cell phone data, many studies focus on the mobility patterns and anchor locations of users that can be extracted. Ahas et al., (2010) used cell phone data to study the temporal patterns of space consumption of the households living in the suburbanized areas of Estonia. From the perceptive of diffusion process, Gonzalez et al., (2008) focused on the movement trajectories of 100,000 mobile phone users tracked for a six-month period and found individual trajectories extracted from the cell phone data show a high degree of spatial regularity and temporal periodicity. Besides, they noticed that individual trajectories are characterized by the same two-dimensional probability distribution after normalizing the scale of each user's trajectory to unity. A limited number of more recent studies have started to search for ways of detecting activity patterns from mobile phone traces. For example, Phithakkitnukoon et al. (2010) developed an algorithm to identify the most probable activity associated with a specific location based on the spatial distribution of different types of points of interest. Jiang et al. (2013) presented extensive data processing steps to extract useful information from triangulated mobile phone traces and proposed to build probabilistic models to infer activity types conditional on land use type, time and daily mobility chain.

In comparison with GPS and cell phone data, transit smart card transactions not only reveal pattern of transit trips, but also reflect the performance of transit system. Thus, considerable literatures on the transit smart card transactions data have focused on using the information embedded in smart card transaction data to improve the quality of service of transit system. In a recent review of the application of transit smart card data, Pelletier et al. (2011) summarized that existing studies on the transit smart card transactions can be grouped into three categories. Strategic-level studies focus on the demands of different types of riders and the implication on long-term network planning (Agard et al, 2006; Chu and Chapleau, 2008). Tactical-level studies emphasize schedule adjustment, and longitudinal trip patterns. Operational-level studies employ transit smart card dataset to measure the performance of transit systems like schedule adherence and fare structure (Morency et al, 2007; Deakin and Kim, 2001). These previous studies have shown the great potential of using the transit smart card transactions data for in-depth activity-travel pattern analysis and mining, especially when more refined learning methods and detailed built environment and transportation data are available.

2.2.2 Spatiotemporal features of activity-travel pattern

Many researchers have been focusing on the variation of activity-travel behaviors within one day, in part due to the cross-sectional nature of household travel survey, which is the most common dataset used for activity-travel study. But it has been widely recognized that individual and households' activity-travel behaviors present not only daily variation but also day-to-day and weekday-weekend variations, which suggests the model built on a one-day survey might not capture the interdependencies of activities among multi-days. One of the evidences is from Huff and Hanson (1990). Based on the 1971 Uppsala Household Travel Survey which extended continuously for five weeks, they found the longitudinal travel records of most people exhibit not only repetition of regular trips like commuting but also variability of other trips from day to day. As they point out, "seven-day record of travel does not capture most of the separate behaviors exhibited by the individual over a five-week period, but it does capture, for most people, a good sampling of the person's different typical daily travel patterns (p1, Huff and Hanson, 1990) ". Their finding was confirmed by Schlich and Axhausen (2003), who investigated the regularity and variability in activity-travel behavior based on a six-week travel diary survey. Schlich and Axhausen (2003) also reported weekday behavior tends to be less variable than weekend behavior. Research also shows the heterogeneity in intra-personal variability of activity-travel behavior across population subgroups (Builung et al., 2008). These empirical evidences suggest the temporal regularity and variability of activity-travel behavior can be attributed to human habitual behavior, interdependencies of individual and household activities among multi-days, and the factor of day varying preference. Therefore, it is necessary to examine activity generation for multiple days in order to reveal the repetition and variability of activity-travel behavior.

Like the temporal variability of individuals' activity-travel behaviors, spatial variability of destination and path choices in activity generation is not ignorable. But less effort has been directed to questions concerning whether individuals make similar destination choices and similar path choices over time. According to the theory of habitual behavior, peoples' activity patterns should exhibit a high extent of spatial stability from day to day because of the uncertain cost and risk associated to explore new locations and paths. Susilo and Kitamura (2005) showed that workers and students demonstrate a quite stable spatial behavior on weekdays, in comparison with more variable action spaces of non-workers and all respondents in the weekend based on the 6-week survey mentioned above. Despite differences in urban form and planning polices, Buliung et al., (2009) found similar location-based repetitions in individuals' activity pattern using the data from Canada. This is what Huff and Hanson (1990) called "locational persistence". Understanding the stability of individual activity destinations over time provide insights on identifying the factors influencing peoples' location choices and activity-travel behavior analysis.

Activity-travel pattern intrinsically is a series of activities with sequential choices of locations, durations, transportation modes etc. These components are intermingled in the process of determining and planning the course of activities. The sequence of an activity-travel pattern not only signifies the priority of activities at different time of day, but also is indicative of lifestyle. Thus, when developing models of destinations, departing time and transportation modes, it is important to take into account the sequence of activities and their interdependencies with each other. Although many studies have dealt with particularly the activity generation, less effort was made to investigate the evolution of the activity demands over time and the activity-agenda formation considering explicit interactions with other components (Habib, 2007). More insights on the interdependencies between different activities within an activity-travel chain would help to increase the prediction capacity of activity learning models.

2.3 Urban development and transit system of Singapore

Before investigating the spatio-temporal patterns of transit ridership exposed by the transit smart card transaction data, it is necessary to briefly look at the characteristics of urban development mode and transit system in Singapore.

2.3.1 Heterogeneous urban spaces

As a "city state", Singapore is scarce in land and hence has a high level of urbanization. High-rise buildings are ubiquitous and expressway infrastructures frequently snake around buildings. Under the centralized planning process, Singapore presents a clear partition in the functionality of urban space. Heavy industry is concentrated in Jurong and light industry is distributed in several new towns in the North-western and Eastern parts of Singapore (See Figure 2-1). Public housing towns are developed according to the standard planning and development procedures. Each town has a local town center with clustered commercial establishments. Although under the pressure of upward extension and gentrification, Singapore has been effectively preserving many historic buildings and neighborhoods. For example, shophouses and the "five foot way" are still common in places like Chinatown and the neighborhoods around the Orchard Road. Consequently, urban spaces of Singapore can be distinguished by the age of buildings and infrastructures, the concentration of population by ethnicity, and the land use functionality. The various zoning boundaries like the Electronic Road Pricing (ERP) zone and school admission zones further exaggerate the level of space heterogeneity.

In cities like Singapore where mixed use and transit oriented developments prevail, characterizing urban form at the traffic analysis zone (TAZ) level has been inadequate to distinguish the impacts of places on human mobility and activity because many zones or towns present similar development patterns. On the one side, readily available business statistics aggregated at zonal level neither can tell the difference between shopping malls and commercial streets nor can differentiate restaurant clusters from local hawker centers. These differences are not likely to be revealed in the data of coarse spatial resolutions, but can have big impacts on location choices of people and businesses. On the other side, the city of Singapore has been growing rapidly in recent years. The changes in urban forms need to be captured for better understanding of changing activity-travel patterns revealed by urban sensing data stream. All these necessitate the search of spatially and temporally detailed data for the better characterization of urban built environment.

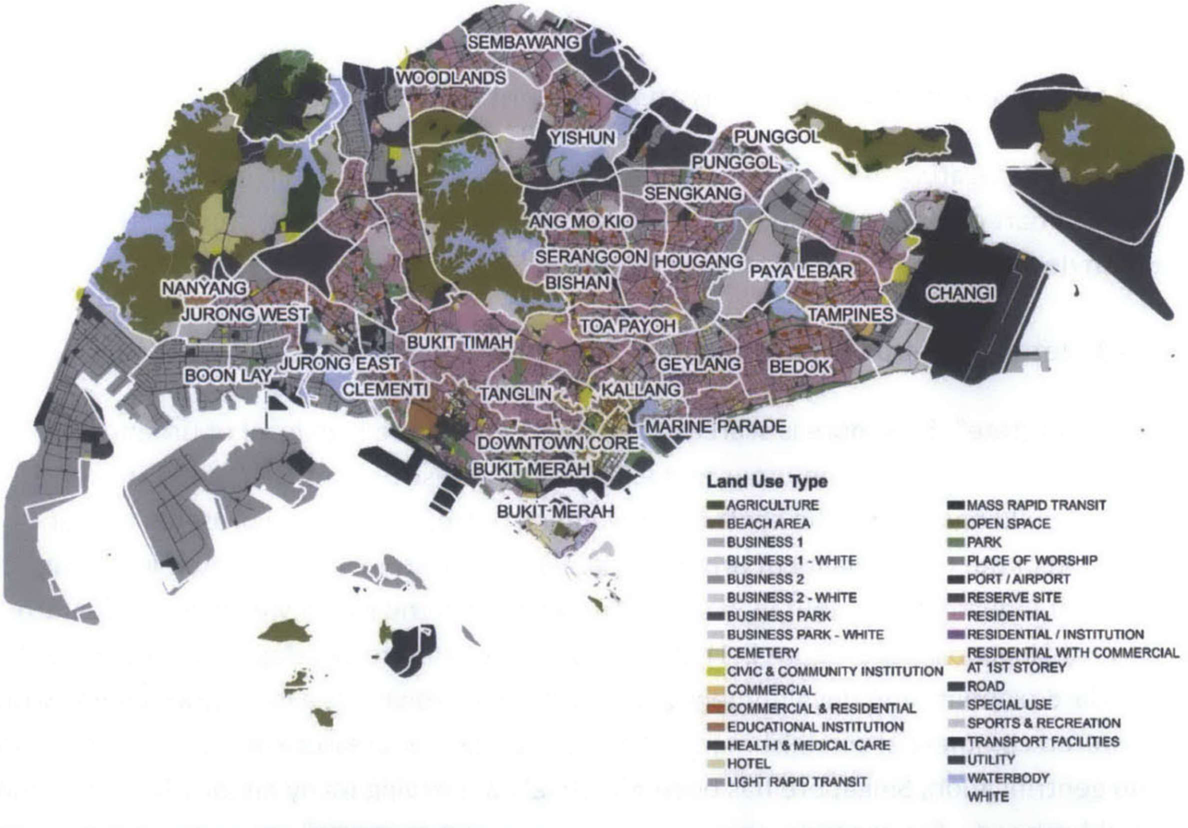


Figure 2-1. Land use plan (2008) of Singapore

2.3.2 Urban transit services

Singapore has one of the most extensive public transportation systems in the world. As of year 2011, the time when the EZ-Link data used in this study were collected, the public transit system of Singapore had been comprised of more than 4000 bus stops, 112 rapid transit stations serving four mass-rapid transit (MRT) lines and three light-rapid transit (LRT) lines. Figure 2-2 plots the spatial distribution of transit stations, which covered the majority of the built-up areas of the island. In the same year, the daily trips made on the public transit system reached around 5.8 million[1] (Table 2-1), making the transit system of Singapore one the busiest transit systems in the world. The mode split of public transportation also outweighs other modes. According to the HITS 2008 survey data, 54.37% the trips in the survey were made by bus or rail system, in comparison with 29.7% of car trips (including car passengers), 2.7% of taxi trips and 2.1% of motorcycle trips.

Table 2-1. Average daily ridership ('000 passenger-trips)

Year	MRT	LRT	Bus	Taxi
2011	2,295	111	3,385	933

Transit services running along transportation corridors and connecting major areas of the city generally have high frequency, which increases the attractiveness of public transit services. Figure 2-3 shows the number of vehicle trips along the time of day broken down by the types of services[2]. During most of the day (between 6am to 6pm), the system maintains a level of service having around 5,000 hourly vehicle trips. Services of trunk routes account for the largest proportion of vehicle trips, followed by the services from feeder routes. Besides the public transit system, there are a substantial number of shuttle services provided by institutions, private companies and shopping centers, which are important complement of the public transportation. Due to the issue of data availability, they are not counted in this study.

[1] Singapore Land Transport: Statistics In Brief 2011, available at http://www.lta.gov.sg/content/dam/ltaweb/corp/PublicationsResearch/files/FactsandFigures/Stats_in_Brief_2011.pdf

[2] Bus service types are collected from the SBS Transit and SMRT Buses websites.

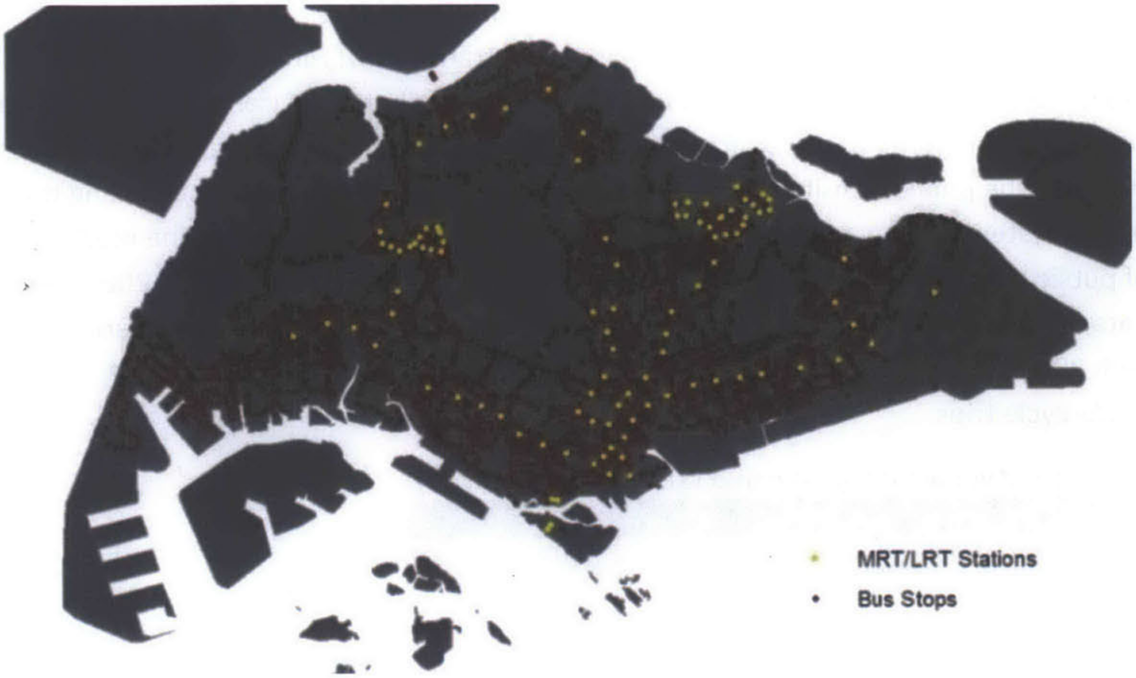


Figure 2-2. Transit stations of the public transit system in Singapore

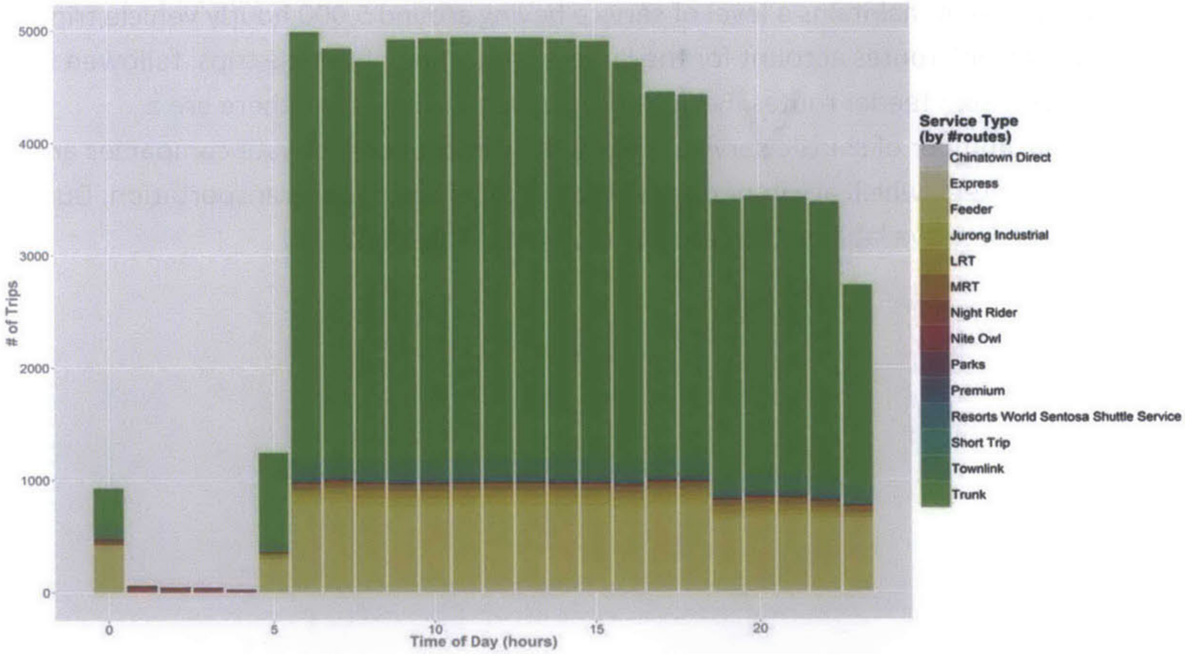


Figure 2-3. Number of vehicle trips along time of day broken down by the types of services.

2.4 Smart Card Transaction Data

The smart card transaction data used in the study is provided by the Land Transport Authority (LTA) of Singapore, covering trip logs of 15 days in April and May, 2011. This study mainly uses the EZ-Link data of the week between April 11, 2011 and April 17, 2011. The data originally came from the system for e-payments, which collects tap-in and tap-out transactions of the EZ-link cards for the bus and rapid transit system (RTS) systems. EZ-Link card is a contact less smart card with a tamper-proof IC chip and antenna built in. Almost all buses and RTS vehicles have been equipped with smart card readers. Because transit fare is charged based on the distance traveled, using smart card is usually more convenient and cheaper than paying by cash. This results in a very high penetration rate of EZ-link card in Singapore. Besides the EZ-Link card, NETS Flashpay card, which is far less common, can also be used to pay transit fare. Using the EZ-Link card, passengers can make up to 5 free transfers within a single trip, with a 45-minute allowance between each transfer. But in a trip, passengers can only take the same transit route once.

Table 2-2. The EZ-Link smart card transaction data sample

Field	Explanation	Example
TRIP_ID	ID of Trip	111111111111
CARD_ID	ID of EZ-link Card	1111111111111111
PassengerType	Type of EZ-link Card	Adult
TRAVEL_MODE	Transit Modes	RTS
BOARDING_STOP_STN	Boarding Station Code	STN Lavender
ALIGHTING_STOP_STN	Alighting Station Code	STN Eunos
RIDE_Start_Date	EZ-link card tap-in date	17/04/2011
RIDE_Start_Time	EZ-link card tap-in time	42:37.0
Ride_Distance	Distance of stage	4.8
Ride_Time	Time of stage	12.1
FarePaid	Fare paid by EZ-link card	0.91
Transfer_Number	ID of Stage	0

Source: EZ-link data, LTA

According to the jargon of transportation, a trip generally refers to a series of stages that made by travelers on different modes that connecting from origin to destination, where a particular activity will be conducted. By tapping in when boarding and tapping out when alighting, the EZ-Link card automatically identify trips taken by travelers and generate a non-duplicated trip id. Information related to trips like the boarding station code, alighting station code, ride start time and fare is also collected. In addition, each EZ-Link card has a unique card id, which enables analysts to uniquely identify a transit passenger and their likely social status based on the types of EZ-Link cards. Three types of EZ-Link cards are commonly seen: student/child concession card, senior concession card and regular adult card. Table 2-2 shows a sample of EZ-link data provided by LTA.

Despite of the high penetration rate of the EZ-Link data, two issues associated with the EZ-Link data could affect their usability for activity-travel research:

1) It is common that a passenger holds multiple EZ-Link cards.
2) A fraction of trip records in the EZ-Link data don't have alighting stations because the card holders failed to tap out when alighting.

Corresponding presumptions are made to circumvent the impacts of these data issues. For the multi-card problem, it is assumed that in a given day, each traveler only use one EZ-Link card. This presumption helps to justify the analysis of travel-activity patterns from the perspective of trip-chain as opposed to individual trip. For the second issue, it is assumed that the missing destination stations can be inferred from the most likely alighting stations of the trips on the same route with the same boarding stations but made on other observed days.

2.5 Exploratory Analysis of the EZ-Link data

2.5.1 Transit trip frequency

Because the EZ-Link data contain multiple-day trip information of the card holders, it is possible to detect the typical transit trip patterns of the card holders and their anchor locations like home and workplace. However, the accuracy of the detection may depend on the frequency of transit trips made by the card holders and recorded in the EZ-Link data. In other words, EZ-Link data can help to uncover the spatiotemporal distribution of daily activities of those travelers who use transit system frequently, but are less informative on infrequent transit users. Therefore, it is necessary to first examine the frequency of transit usage revealed by the EZ-Link data.

In the week of April 11, 2011, around 3.35 million EZ-link card ID numbers are uniquely identified, among which 79.54% are adult ticket type cards, 12.11% are child/student concession type cards, and the rest 8.35% are concession cards for senior citizens. Figure 2-4 plots the percentage distribution of the number of days that EZ-link cards were used in the week broken down by card types. This does not consider the possibility that transit passengers used different EZ-Link cards on different days. Around

20% of the EZ-Link cards were only used in one day, which implies a large number of Singaporeans use transit system infrequently. In contrast, less than 15% of cards were used on the daily basis, suggesting only a fraction of residents were completely transit captive for their daily activities. In terms of the weekday-weekend split, over 30% of cards were used in weekdays only and a little over 10% of cards were only used in weekends. Virtually, there is no significant difference in the frequency of EZ-link card usage among different groups of card holders. Senior concession card holders were more likely to take transits in only one day or two days while child/student concession card holders used transit system more frequently in the week.

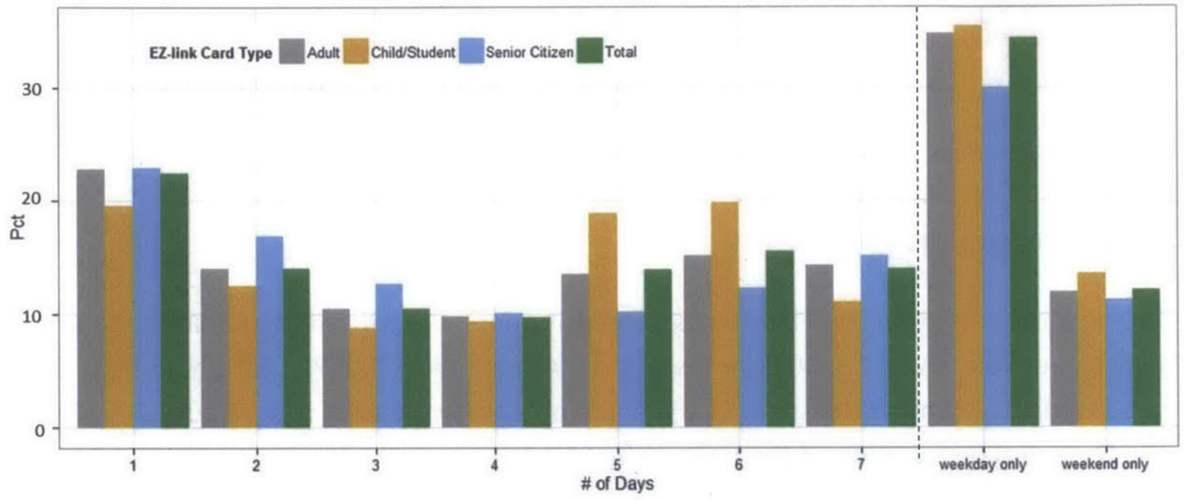


Figure 2-4. Percent of the number of days EZ-link cards were used from April 11 to April 17, 2011.

According to the EZ-Link data, a typical weekday (Monday to Thursday) in April, 2011 had 3.9 million passenger-trips on average, which is converted to 2.06 transit trips per card holder. On the Friday, the ridership increased to 4.3 million passenger-trips, and the transit trips per card holder rose to 2.14. That corresponding numbers are 3.5 million passenger-trips and 2.15 trips for Saturday, and 3.1 million passenger-trips and 2.1 trips per card holder for Sunday. Although the overall transit trips were fewer in the weekends, transit riders on average made more trips. Figure 2-5 plots the distribution of average number of trips of all card holders per day in the week from April 11, 2011 to April 17, 2011, broken down by the types of card holders.

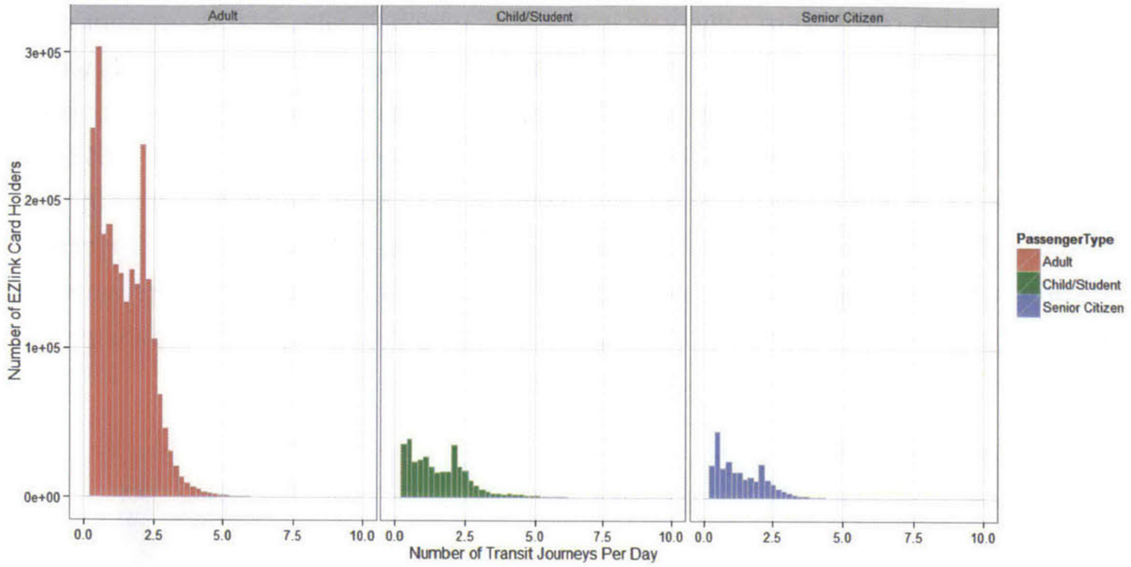


Figure 2-5. Average number of transit trips per day by the types of card holders

2.5.2 Temporal travel patterns

To show the temporal travel patterns of transit riders, the trips in the EZ-link data in a typical weekday (April 13, 2011) are selected. Among the 1,928,723 card holders who used transit system on that day, 61.3% made at least two transit trips. Figure 2-6 plots the temporal distribution of the pairs of first transit trip and last transit trip been made, divided by the types of card holders. The y axis represents the starting hour of the first transit trip and the x axis corresponds to the starting hour of the last transit trip. The thicker the red color, the greater the percentage of the corresponding trip pair characterized by the hours of the first trip and the last trip. As shown in Figure 2-6, most of adult riders started their first trip during the morning peak hours (6am-9am) and returned in hours between 5pm to 10pm, which are likely the commuting trip chains. The starting hours of the first trips of students were more concentrated, mostly at 6am and 7am. A subset of students, presumably having school scheduled at afternoon, started their daily transit trip at around 2 pm. In contrast, the distribution of the time of the last trips are more dispersed for students, ranging from 12pm to 10pm. Senior citizens were more likely to start their trips earlier and also end their trips earlier when compared to the temporal patterns of the other two groups. Besides, the time durations between the first trips and the last trips are mostly shorter.

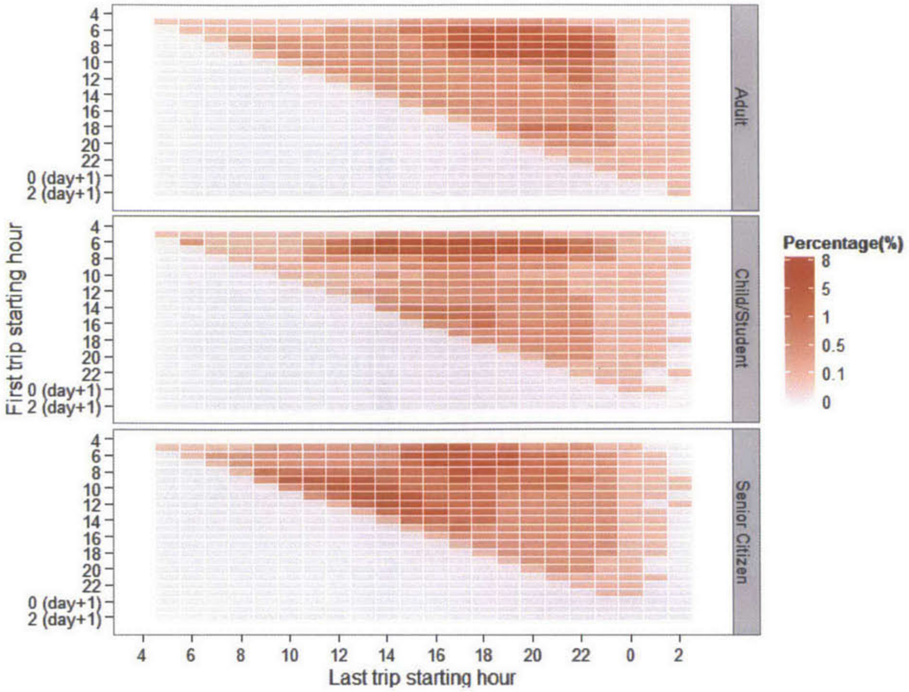


Figure 2-6. The temporal distribution of the first and the last transit trips of the transit smart card holders with at least two transactions on April 13, 2011 (Wednesday).

The regularity of the temporal-patterns of transit trips is shown in Figure 2-7, which plots the hours of the first transit trips and the last trips of transit passengers from April 11, 2011 (Monday) to April 15, 2011 (Friday). It can be seen that the temporal travel patterns are highly regular from Monday to Thursday, as a large proportion of daily transit trips are commuting trips. On Friday, there were more late trips, suggesting more active night lives on the last working day of a week.

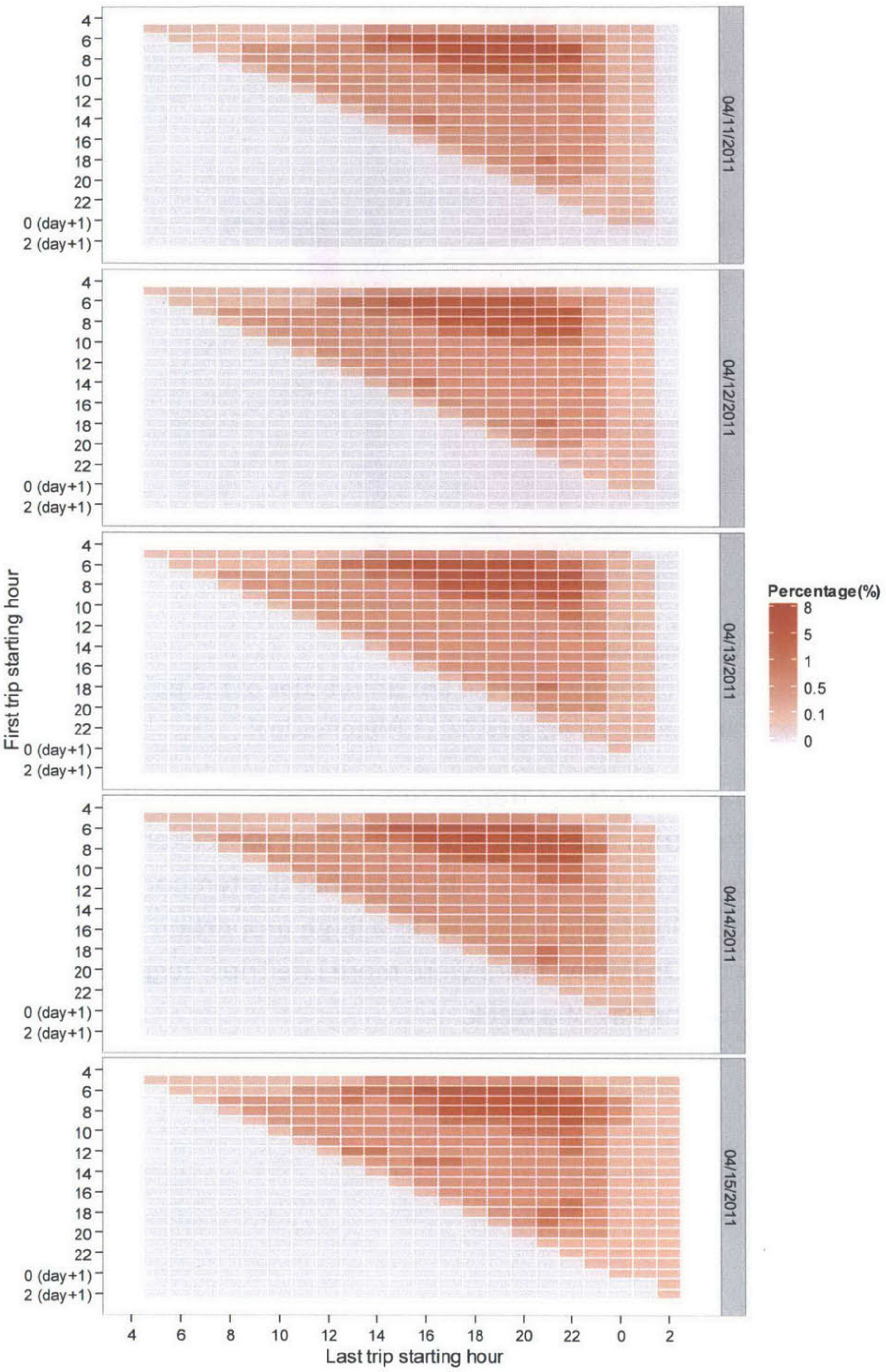


Figure 2-7. Temporal distribution of first and last transit trips for transit smart card holders with at least two transactions from April 11 to April 15, 2011.

2.5.3 Spatial Pattern

Figure 2-8 shows the spatial distribution of the hourly volumes of alighting passengers by stations at different time of day. Except for the time before 6am and after 12am, the transit system of Singapore has been quite busy throughout the day. Even in the late night, there are still transit riders take the transit system to major residential neighborhoods. Overall, the differences in the spatial distributions of hourly alighting passengers between 7am and 12am are insignificant. Generally speaking, more passengers are in travel during the morning peak hours from 7am to 10am and the evening peak hours from 7pm to 8pm. In the morning peak hours, a greater volume of passengers are observed to alight at the transit stations in the industrial zones in the west as well as the corridors along the Bukit Timah road. While in the evening peak hours, more passengers alight at the stations in the east part of Singapore, the Jurong area and Punggol (see Figure 2-1). But meanwhile, the city center and the Orchard Road are also attracting more visitors from buses and rails. The RTS stations, especially the MRT stations, generally served more passengers than bus stops. The spatiotemporal distributions of passengers clearly present the main transit corridors, the major workplace areas and residential neighborhoods in Singapore.

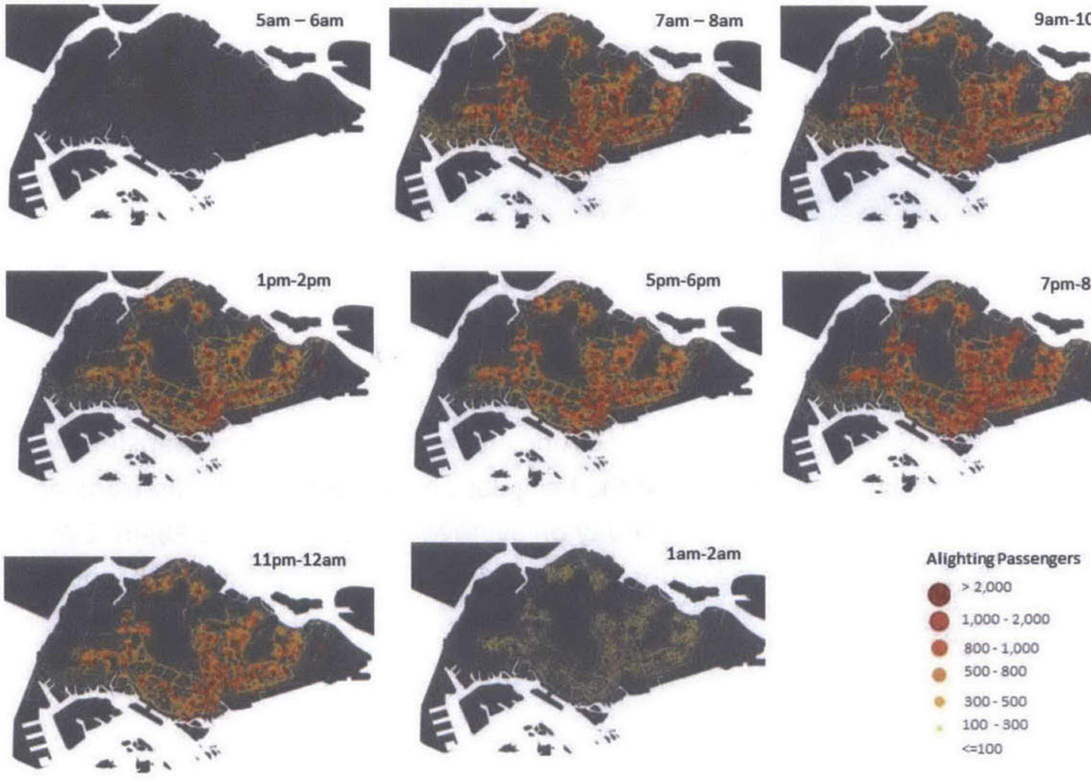


Figure 2-8. Thematic maps of the transit stations by alighting passengers at different time of day.

2.5.4 Transfer Pattern

Transfers are important parts of transit trips. Taking transfers into consideration when planning transit system is critical for balancing operational costs and passengers' convenience. From the perspective of passengers, easy transfers (short walking distance and good walking environment) among transit stations can influence the choices of routes (Guo and Wilson, 2011). From the perspective of location choice of businesses, having a position along major transfer corridors tend to benefit businesses because some pass-by passengers may be attracted by the intervening opportunities provided by businesses. This is especially the case in Singapore where many transfer routes are designed to wind around large shopping centers and underground pedestrian malls. Moreover, as will be discussed more in Chapter 4, transfers observed in the EZ-Link data can help to integrate bus network and rapid transit network, which otherwise can only be dealt with separately in the network analysis.

Table 2-3 shows the percentage of trips with different number of transfers observed in the EZ-Link data from all 15 days' of observations. Around 26.5% of the total transit trips involved transfers. But only less than 4% of the trips had two or more transfers. The statistics do not count for the internal transfers within RTS because no transactions are needed for transfers among subway lines and interchange stations.

Table 2-3. Number of transfers per trip in the EZ-Link data

Transfers	0	1	2	3	4	5
Percentage	73.46%	22.73%	3.44%	0.32%	0.04%	0.01%

The spatial distribution of the transfers in Singapore is displayed in the Figure 2-9. The map on the right zooms into the city center area. The degree of opaqueness and the width of the transfer link lines are proportional to the amount of transfers observed in the EZ-link data. To exclude the irregular transfer paths, only the transfer links that have at least two transfers per day on average are shown in the Figure 2-9. In the city center, activity transfers are observed between MRT stations and surrounding bus stops.

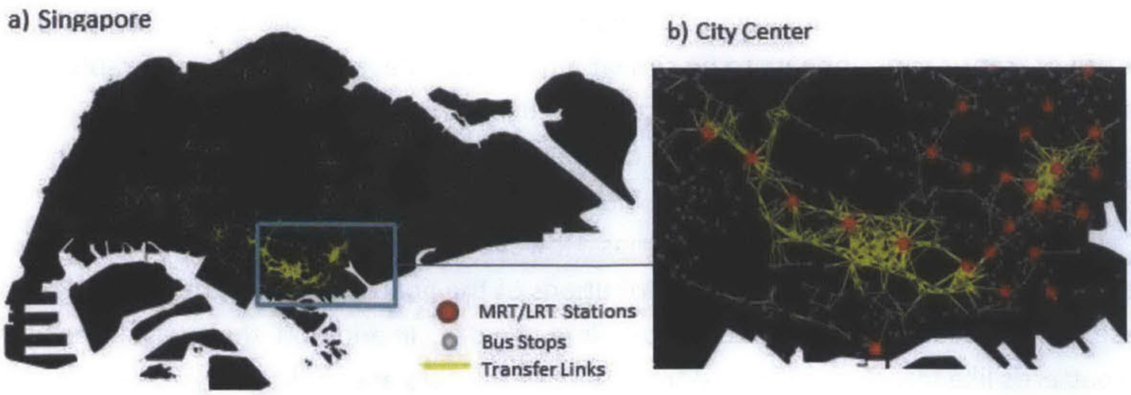


Figure 2-9. The spatial distribution of bus-to-bus or bus-subway-LRT transfers in Singapore observed in the EZ-Link data (15 days)

The e-payment system of the public transportation of Singapore allows a transfer up to 45 minutes. Transfer time refers to time between the moment that a rider tapped out at previous stage of trip and the moment that rider tapped in at next stage of trip. Transfer time typically include walking time and waiting time. Figure 2-10 shows the distribution of transfer time in seconds extracted from that EZ-Link data. The shape of the distribution approximates the normal distribution truncated at around 30 seconds. The distribution peaks at around 3 minutes.

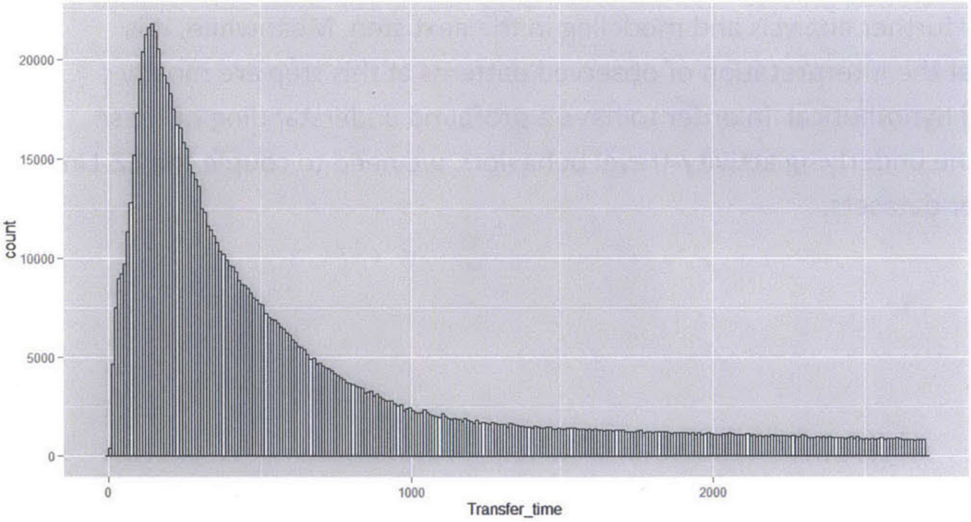


Figure 2-10. Distribution of the transfer time in seconds extracted from the EZ-Link data

2.6 Summary

This chapter provides a preliminary analysis of a week's transit smart card transaction data of Singapore. The data provide a potential to extract continuous profiles of transit use of different types of card holders in their daily lives. Although the overall temporal pattern of transit trips appears to be similar among weekdays, there is considerable spatial variability of the trips as evidenced by the low trip repetitiveness rate. This suggests that individuals have different choices of activities and destinations on each day within a week and necessitates the introduction of spatially detailed urban form measures to help infer the purpose of trips. However, when the trips are aggregated to the station level, the overall spatial distributions of the boarding and alighting passengers appear to be alike in the most time of a day. In addition, the special transit trip patterns like transfers observed from the EZ-Link data are unlikely to be learned from the traditional household travel survey data. This information is not only relevant to the activity-travel choices of individuals, but also reveals the spatial heterogeneity of places as a result of differentiated transit service provision and usage.

In a nut shell, the preliminary exploration is an important step for the analytics of urban sensing data. It helps the researchers to form a general knowledge of the data like the formats of the data, the types of information contained, as well as the strength and the weakness of the data. It also facilitates the identification of unanticipated patterns and findings that warrant special attentions in the following analysis steps. The general spatio-temporal patterns of transit trips exposed in this section provide a good foundation for further analysis and modeling in the next step. Meanwhile, it is recognized that the interpretation of observed patterns at this step are mostly superficial and hypothetical. In order to have a profound understanding of these patterns and the underlying activity-travel behaviors, we need to couple the EZ-Link data with other datasets.

Chapter 3. Integration of Spatial-temporal Data for Activity-travel Learning

3.1 Introduction

The value of the EZ-Link data is limited for human activity-travel research if only the information of the trips are presented but other important driving forces of travelers' choices such as the purposes of the trips and the durations of the stays are unknown. It is conceivable the activities associated with observed transit trips can be inferred based on the correspondences between the behaviors of travelers' and the characteristics of urban environment surrounding the origin and destination transit stations, this needs spatially and temporally detailed data to characterize the urban built environment, which are usually difficult to acquire. However, the pain of data collection and preparation can be somewhat alleviated thanks to the open data initiatives participated by many government agencies and the unprecedented temporal and spatial information embedded in the readily accessible data sources like web services and crowdsourcing. It is necessary investigate generic ways to assemble datasets from different emergent sources to support urban sensing data analytics that concern increasingly detailed and complicated urban policies and built environment settings.

This first calls for innovative approach to integrating datasets from multiple sources, which present heterogeneous formats, qualities and information. Most secondary datasets collected from online sources or provided by agencies are by-products of other tasks. Therefore, it is common that a dataset only provide partial information for the analysis and modeling. In addition, although available datasets are growing explosively in recent years, the format and quality of datasets become increasingly heterogonous as a result of more diverse data sources. It is rare that two datasets from different sources can be merged directly without pre-processing. This is especially for the case of emergent datasets, which usually have very different structures from the data from traditional sources. Therefore, it is a big challenge to resolve the compatibility issue between different datasets in the process of data fusion.

In this section, we propose an ontology-based data integration mechanism to evaluate, sift and integrate geospatial information from heterogeneous data sources. Ontology defines the vocabularies and concepts that are commonly understood in a domain. Building ontology is usually a collaborative effort requiring the involvement of domain experts to achieve consensus on concept definition. The focus of this study is not to build a comprehensive ontology for activity-travel research. Instead, we will present a prototype that allows semantic-level matching between local data and a simplified ontology to demonstrate the benefits and applicability of proposed data integration approach. The objective is to create an integrated dataset that absorb the reliable and detailed spatial information like building, business and land use from various sources, to facilitate the characterization of the urban built environment and the inference of activities from the EZ-Link data.

3.2 Background and related work

3.2.1 Emerging data sources

Ubiquitous sensors equipped in urban areas and location aware devices like smart phone and GPS have been used widely in travel behavior data collection as a way to supplement conventional household travel survey or activity-based time allocation survey (Wolf, 2006; Asakura and Hato et al., 2004; Forrest and Pearson, 2007). Lately, Hato (2010) proposed a behavioral context information measuring instrument, which integrates a variety of sensors like atmospheric pressure sensor, barometric pressure sensor, sound sensor and 3-dimension accelerator in addition to a basic GPS. The idea is not only to record the positions of device carriers, but also to collect detailed movement information like acceleration and characteristics of surrounding environments like noise level, which may help to infer personal travel behavior and activity.

The web has an enormous reserve of information. The open data initiatives have motivated many agencies and companies to share their data. Mashup or web-based services created by data providers result in more and more data available to public through internet. Meanwhile, the amount of user generated information accumulates at an unprecedented rate per day fuelled by proliferation of "social websites and applications" like Wikipedia, Facebook and Twitter. Users are also increasingly willing to share information tagged with location using their location-aware devices such as smartphones and tablets. Besides the social network data, volunteered geographic data like OpenstreetMap (OSM) has been widely used in many applications and are receiving growing attentions in academia (Haklay and Weber, 2008; Neis et al, 2011). The advancement of technology like recent efforts on semantic web and linked data is likely to empower data search, access, and use on the world-wide web in a more intelligent, interoperable and convenient way.

3.2.2 Heterogeneous datasets

However, only a small fraction of urban sensing and web open data is readily discoverable and accessible, much less usable. Whilst the emergent data appear to stimulate new efforts in urban modeling and human activity studies, their values can be extracted fully only when being coupled with other datasets. One of the biggest challenges in data integration is how to reconcile all types of heterogeneities stemming from different data sources. A diverse range of heterogeneities and conflicts can be found among datasets from different sources, as listed in Table 3-1.

Table 3-1. Types of data heterogeneities and examples

Type	Examples
Syntax Heterogeneity	• Different tuple (entity, identifier)
Structural Heterogeneity	• Different cardinality between entities
	• Different entity-identifier relationship
Sematic Heterogeneity	• Polysemy, Synonymy and homonym
	• Different abbreviation
	• Different categorization
Data Quality Heterogeneity	• Accuracy
	• Completeness
	• Source trustworthiness
Data Type Heterogeneity	• Different data format and type
	• Different spatial scale and representation
	• Different temporal scale and representation
	• Different measurement unit

While variations in data models and schema lead to syntax and structural differences among datasets, semantic difference is more common due to different naming or categorization conventions of data providers. Besides, heterogeneity in data quality, in data collection time, as well as in privacy and security requirement also hinder the integration of relevant information from different data sources.

3.2.3 Existing data integration approaches

To tackle with the issue of data interoperability and integration, in the field of computer science, several approaches are commonly used to deal with the data interoperability and integration. Software vendors and consortiums like W3C and OGC (Open Geospatial Consortium) set standards on data formats and representations.

Examples are XML for web service based data sharing and CityGML for the storage and exchange of virtual 3D city models. But these standards only addressed the syntax heterogeneity of datasets of the same type. Besides, data warehouse realizes data integration through a common data storage strategy. This approach focuses more on the physical location of datasets than on the fusion of information.

Multiple database based integration uses a global schema and mediators to reconcile semantic differences of local DBMS (Ziegler and Dittrich, 2004). Mediator-wrapper architecture is used to translate user defined queries to queries that can be carried out by wrappers built for local databases. Most applications of this type are generally based on relational data models and global schemas tightly coupled with tasks or applications (Ziegler and Dittrich, 2004). Schema matching is typically performed manually by database or domain experts. Besides, traditional data manipulation that is built on the basis of relational database schema becomes increasingly difficult to deal with semi-structured, unstructured data with increasing spatial and temporal complexity.

A deluge of urban sensing data and web-based information requires that an integration approach has capabilities to incorporate different types of data. But it is more demanding when data with various semantics, temporal and spatial representation, and quality need to be merged together. More complicated operations like comparing, cross-validating, correcting, aligning and aggregating become necessary in the data integration process.

Further, it is not necessary to require end users to know the nuts and bolts of every dataset. It is also unrealistic to expect all end users have the expert skills to tackle with these datasets. Providing a single entry point to the integrated data and information will greatly facilitate the analysis and modeling of human activities and travels. As the size and the variety of data accumulate over time, a capability for automatically handling changes is very necessary. Heavy manual intervention is not only time consuming, but also error prone. Successful data integration requires a system that does more than just exchange or merge data. The meaning, representation and structure of data need to be understood as well (Nathalie, 2009).

3.2.4 Ontology-based data integration

As alluded in the last section, data integration approach relying on mediator-wrapper architecture and middleware is usually task-specific and will become more complex and cumbersome when facing complicated urban modeling and analysis. Meanwhile, different users need to repeat the data processing and schema matching procedures to acquire the same datasets. It will be desirable if data integration approach can map local datasets from different sources to a purpose-generic but domain-specific global schema which provides a common representation and interpretation of the information needed for the research and application in the domain, and a single entry point for users. Ontology provides a promising solution for the idea.

As defined by Mars (1995), Ontology is “a structured, limitative collection of unambiguously defined concepts”. In general, Ontology explicitly defines the meanings, representations and the structures of the concepts in a domain, which forms the mutually understandable, sharable and reusable domain knowledge repositories. This definition implies the potential of ontology in reconciling the conflicts in the semantics and syntax of datasets arisen from unstandardized ways of data encoding. Many previous researches have shown ontologies can be a valid tool for addressing the interoperability of information from multiple heterogeneous sources (Partridge, 2002; Nathalie, 2009).

An integrated ontology from one or several domain-specific ontologies can benefit the community by providing guidance on generic vocabularies and correspondences to represent data and modeling related knowledge. However, although ontology-based approaches have reached a good level of maturity in other fields, its application in the domain of urban planning and modeling is limited (Benslimane et al., 2000).

3.3 Data collection

Activity-travel analysis requests comprehensive and very detailed information regarding the urban built environment, transportation network and service as well as people’s travel behaviors. Urban planning researchers are in particular concerned with the spatial and temporal changes of urban environment and the impacts on human behaviors. Geospatial data at a variety of geospatial levels like parcels, buildings and zones, transit stations, public parking lots and roads are all desired for the measurement of the urban built environment. To account for the temporal coherence of datasets, which is important for a rapid-changing urban environment, the temporal information of datasets is also required. Therefore, many efforts have been made to gather data from heterogeneous sources including government agencies, web services, crowdsourcing, and corporate data owners.

In addition to the official datasets provided by government agencies, a series of web-scrapping and web-service programs were developed and streamlined to systematically retrieve online information. For example, the data mall of Land Transport Authority allows the retrieval of real-time information of Carpark lot occupancy in the major shopping malls at the Marina Center, Orchard and Harbourfront areas of Singapore. Moreover, user contributed information from websites like FourSquare and OpenStreetMap are also collected to complement and enrich the characterization of contextual factors for analysis and modeling.

While most of the datasets are already secondary at the time they are gathered, the primary approaches of data collection are very diverse, from interview to urban sensing, from government registration record to volunteered contributed information. Therefore, these datasets show tremendous variations in the scope and granularity of details, the data types and formats and in particular the data qualities. Most of the datasets have limitations in one aspect of data quality or another such as reliability, completeness and timeliness, which restrict the extent of analysis and modeling can be carried out. But more notably, the data quality issue complicates the integration process. It is always a challenge to sift and assemble valuable information from the datasets with dubious quality.

3.4 Proposed approach of data integration for the activity-travel analysis

Activity-travel analysis deals with a large amount of objects and concepts in an urban system that can take on various representations and relationships. Different classification and naming conventions as well as different reference system adopted by different agencies adds to the complexity of the already intricate integration tasks. For example, buildings can in one case geometrically be represented as points or footprint polygons, but in other cases they are represented as nested polygons based on the aerial view from above. In the transit network analysis, node is usually considered as a synonym of vertex. Edges could refer to links, connections, and routes, which can be confusing for inexperienced data users. As the amount of data grows, it is not only desirable but also imperative to build a knowledge base to explicitly support interpretation and exchange of entity representations and correspondences in a field. The proposed ontology-based framework covers ontology building, ontology mapping, and custom tools to support integrating, sharing and exchanging of the information needed for the analysis of human activity-travel patterns. The proposed framework can also be transferred and applied to the data integration process in other fields.

3.4.1 General framework of the proposed data integration approach

Figure 3-1 illustrates the proposed framework of the ontology-based data integration approach used in this study. The framework centers on a domain ontology, which is used as an unambiguous and persistent knowledge reservoir consisting of the concepts and relationships used in the activity-travel analysis. In the domain ontology, concepts and relationships are elaborated and annotated by the definition and terminology that are generally accepted in the field. As outlined in Figure 3-1, the approach contains a series of sub-modules like data cleaning and standardization, schema mapping, data query and integration based on matched schemas, and the imputation of missing attribute. Each of these sub-modules is discussed in details in the following sections.

3.4.2 Creation of ontology to support semantic interoperability

Creation of a domain specific Ontology requires great efforts and collaborations on the design of knowledge representations as well as annotations on concepts, relationships and axioms. The focus of this study is not on developing a comprehensive generic ontology for urban planning and modeling, but on demonstrating the utility and applicability of the proposed approach that can generate more spatial-temporally precise and detailed information to support further analysis and modeling efforts with a small-scale task-oriented ontology example.

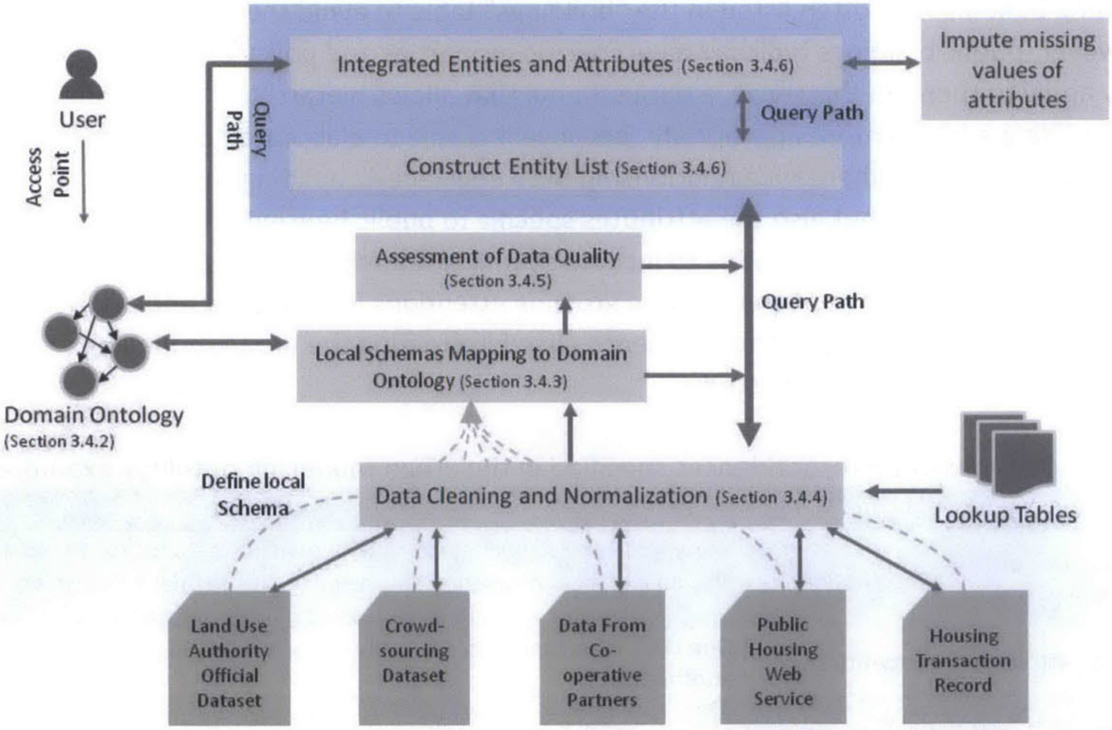

User
Access Point
Query Path
Integrated Entities and Attributes (Section 3.4.6)
Impute missing values of attributes
Query Path
Construct Entity List (Section 3.4.6)
Assessment of Data Quality (Section 3.4.5)
Query Path
Local Schemas Mapping to Domain Ontology (Section 3.4.3)
Domain Ontology (Section 3.4.2)
Define local Schema
Data Cleaning and Normalization (Section 3.4.4)
Lookup Tables
Land Use Authority Official Dataset
Crowd-sourcing Dataset
Data From Co-operative Partners
Public Housing Web Service
Housing Transaction Record

Figure 3-1. Framework of the Proposed Ontology-based data integration approach

The data model adopted in this ontology example is compliant with the Resource Description Model (RDF) triples <Subject, Predicate, Object>, a metadata model specified by the World Wide Web Consortium (W3C). The subject and object are classes representing real-world entities, and the predicate denotes to a relationship between the subject and the object. For example, to represent the notion "Car license has a period of validity of 10 years" in RDF, a subject is denoting "Car license", a predicate is denoting "has" and an object is used to represent "period of validity". "10 years" is the value for the object "period of validity". Based on this primitive RDF data model, RDF schema was specified to describe ontologies. It contains two basic elements: class and property. Class represents real world objects and property defines the correspondences and rules applied to classes. RDF is usually written in a XML-like language. Thus, it is machine-readable and can be easily shared and integrated with other applications and platforms, either online or offline.

Much of the power of ontology comes from its capability and flexibility in defining the correspondences between classes that capture complex relationships in the real world. In addition to entity-attribute association commonly seen in the traditional relational database, ontology often includes other types of correspondences such as subsumption, mereology, equivalence and constraint. These relationships enable a richer and more precise matching with increasingly sparse, complicated and heterogeneous data. For example, in the traditional relational database schema, public housing buildings are often listed in the "buildings" table to avoid the redundancy. However, public buildings typically have specific properties and policy constraints that only apply to them. In this sense, a data schema that allows hierarchical definitions of the building entity is more appropriate, because it is able to elaborate not only the superclass-subclass inheritance relationship between generic buildings and public residential buildings, but also the attributes specific to public housing buildings. Similarly, in ontology, it is possible to include spatial and temporal correspondences among entities that have been receiving growing attentions in modelling and analysis. Table 3-2 lists a number of correspondences used in creating the urban modelling ontology example used in the study.

Table 3-2. Types of correspondences specified in the urban modelling ontology example.

Type	Relationship	Description	Example
--subclass_of --superclass_of	subsumption	Every instance of subclass is also an instance of superclass.	HDB building is a *subclass of* residential building. Residential building is a *superclass of* condominium building.
--has_ attribute	entity-attribute	One class is a property of another class.	Building has height.
--has_key	Identifier	Every key uniquely identifies a class instance.	Building has key address.
--part_of --have	mereology	One class is a part of the other class.	Bridges are *part* of roads. Buildings *have* units.

--equivalent_to	equivalence	Two classes are equivalent.	HDB building is *equivalent to* public housing building.
--constrained by	constraint	One class is constrained by another class.	The classification of children is constrained by age (under 20).
--from_time --to_time	temporal correspondence	The temporal validity of a class.	Activities start *from time1*. Activities last *to time2.*
--located_in --contain	spatial relationship	The spatial containment relationship between two classes.	Buildings are located in zones. Parcels contain buildings.

One of the benefits of using ontology for data integration is that at the ontology creation stage, we can start with the relevant ontologies deigned by others because the modification and extension of ontology is relatively straightforward. A few ontologies have been developed for urban management and planning (Teller, Lee and Roussey, 2007). However, none of them are designed for the modeling and analytics of activity-travel research, especially for the analytics of emerging big data. Figure 3-2 presents an urban building ontology example targeted at the fusion of building information of Singapore. One of the major purposes of using ontology is to reconcile semantic conflicts. Many pairs of classes have an "equivalent as" relationship between them. For example, "year built" of buildings from one dataset is assumed to be equivalent to "lease commence year" from another dataset. "Transit" and "public transportation" is assumed to refer to the same concept. Also, as mentioned above, this ontology enables a more flexible representation of the hierarchical and intricate relationships among different types of buildings as well as the relations between buildings and other entities. For example, it is clearly exhibited in the diagram that "multiple-story car park (MSCP)" is a subclass of "building" but also a subclass of "parking". Ontology allows classes at the same level to be overlapped with each other unless the disjointness between classes is explicitly stated. This makes it easier to classify buildings like "shop-house", which has both residential and commercial uses.

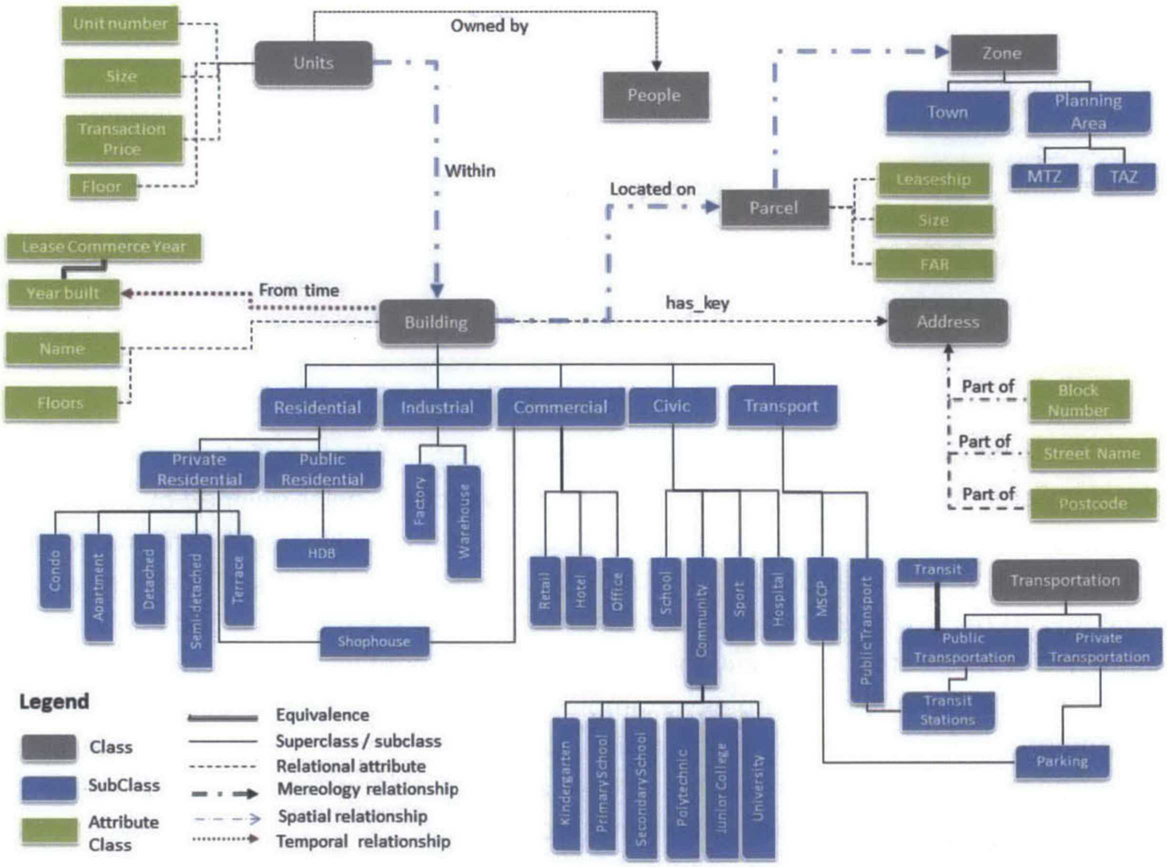


Figure 3-2. An example of the urban building ontology (as a part of the ontology for activity-travel analysis)

The example ontology used in this study is generated by using the Protégé, an interactive ontology editor developed by the Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine (Knublauch et. al, 2004, Tudorache et. al, 2008). Figure 3-3 shows the screen prints of four different components in the Protégé that can be used to assist ontology creation and reasoning. Figure 3-3a) displays the classes editing panel of the Protégé, which allows users to define the superclass, the subclass or the equivalent class of a given entity and to add annotations. The relationships between classes are specified in the "property matrix" panel as shown in Figure 3-3 b). Figure 3-3c) provides a graphical view of the relationships among different classes. The reasoning process of the Ontology is usually executed by the SPARQL query, which is also implemented in the Protégé, as shown in Figure 3-3 d). Readers can refer to Horridge et al. (2004) for a comprehensive introduction of the Protégé.

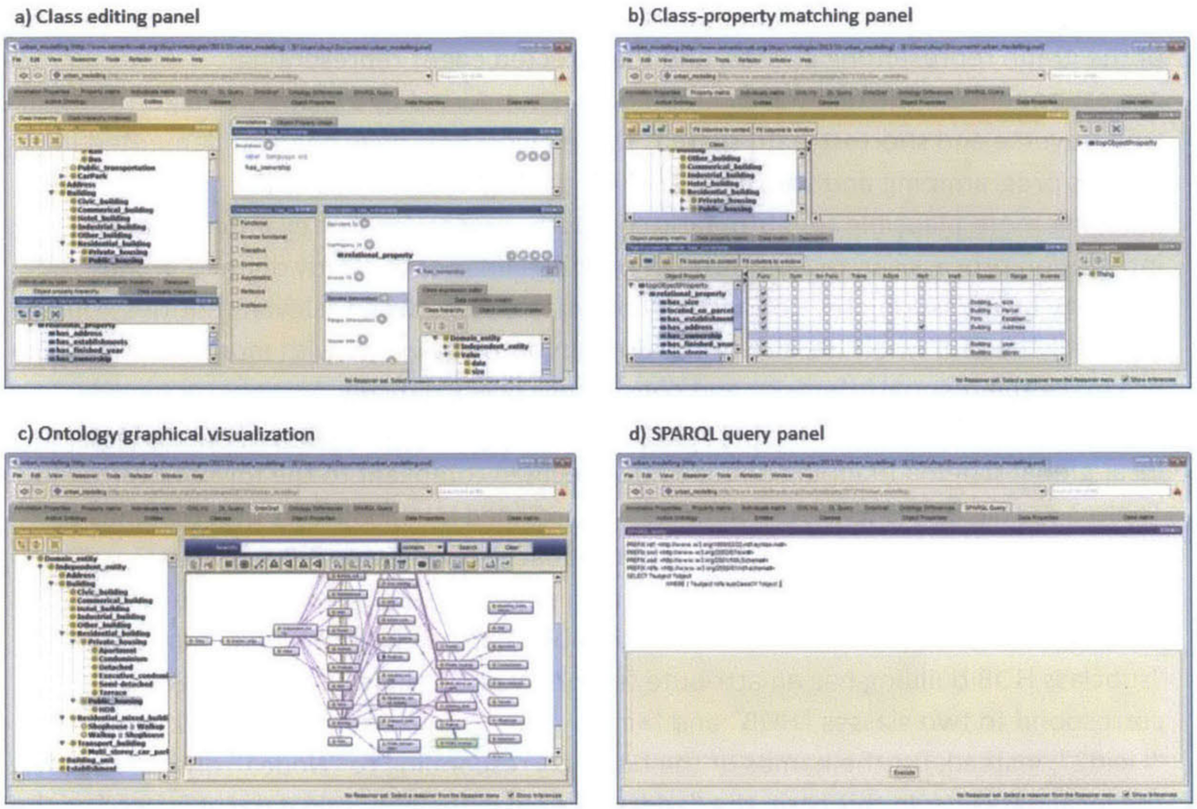

a) Class editing panel
b) Class-property matching panel
c) Ontology graphical visualization
d) SPARQL query panel

Figure 3-3. Screenshots of the Protégé

3.4.3 Schema-mapping between data sources and ontologies

One of the most critical challenges for data integration is to fit the information from individual data sources to the global data schema, i.e. ontology, designed for tasks. This challenge is addressed by schema mapping, which aims at finding correspondence between semantically related entities based on the mapping between local schemas and global ontologies. Many schema mapping approaches have been proposed (Kokla, 2006; Nathalie, 2009). These mapping approaches can be grouped into two types: lexical matching and structural matching. While lexical matching focuses on using linguistic and text techniques to detect the correspondence between ontology labels and local column names, structural matching relies on the structural relationship outlined in the ontology to sort out the relationships among information contained in various datasets. The proposed data integration approach adopts the structural matching between local datasets and the global ontology.

As shown in the ontology diagram for buildings in Figure 3-2, the ontology example has a tree-like structure incorporating the hierarchy of classes and the heterogeneity of relationships among classes. It is convenient and straightforward to convert to static ontology schema to a graph data structure metadata with the vertices of the graph representing classes, and the directed edges representing correspondences. The analysis of graph data is supported by a rich set of algorithms, like tree traverse and shortest-path search. These algorithms have been implemented in various programming and statistical analysis packages, which make the ontology mapping more plausible for automatic computer processing. For example, the data integration process can reason through relationships not explicitly defined in the ontology by measuring the connectivity or adjacency between relevant entities. This is especially helpful when the size of the domain ontology grows with more details, which will make the manual inference and consistency check difficult.

In order to facilitate the schema matching process, for each local dataset, a local-global schema mapping needs to be specified compliant to the format drafted in Table 3-3. The idea is to list the information provided by the dataset that matches the entity relationships specified in the ontology. Each row in Table 3-3 represents information from local dataset matches to an edge (correspondence) in the global ontology. For example, the first row of the table matches with the edge indicating "subclass HDB building has an attribute finished year". "Node1" and "Node2" correspond to two classes "HDB" and "Finish_Year" in the ontology. "Field1" and "Field2", instead, list the names of the fields corresponding to "Node1" and "Node2" in the local datasets. In this case, the "address" field in the hab_sale dataset is used as the primary key of HDB building and the "lease_commence" field is assumed to have the information on the finish year of buildings. The first row of Table 3-3 indicates that the information of "Finish_Year" attribute of "HDB" buildings can be found in "lease_commence" field of "hdb_sale" dataset with the HDB buildings identified by "address" field.

In the schema mapping look-up table, the temporal tags of the datasets are included in order to help users to retrieve the information that is temporally valid. The last column "DR_index" is a composite data quality index used to determine the sequence of retrieval when information from multiple datasets matches with the same edge (i.e. correspondence between classes). For example, the age of building information contained in different datasets may be in conflict with each other. Therefore, it is important to evaluate the quality of information and select the most accurate information into the integrated dataset first, followed by less reliable information. More discussion on this can be found in Section 3.4.5.

Table 3-3. A look-up table for schema mapping between local datasets and global ontology

Node1	Node2	Dataset	Field1	Field2	From_time_field	To_time_field	DR_index
HDB	Finish_Year	hdb_sale	address	lease_commence	lease_commence	NULL	0.9
HDB	Finish_Year	hdb_resale	address	lease_commence	lease_commence	NULL	0.9
Private_housing	Building_type	private_street_dir	address	type	NULL	NULL	0.7
Condominium	Building_type	condo_sax_dir	address	prop_type	year	NULL	0.7
Condominium	Building_type	condo_guru_dir	address	type	NULL	NULL	0.7
Private_housing	Building_type	realis_private	address	type	year	NULL	0.9

The schema mapping process is semi-automatic in this example. A fuzzy text matching method based on the generalized Levenshtein distance[3] is used to align table names and field names from source datasets to the semantically closest concepts predefined in the ontology. Meanwhile, equivalent concepts or entities are matched directly. A recommended mapping result will be provided for manual inspection and validation. The local-global schema mapping table only needs to be prepared once as long as the schemas of local datasets keep the same. New data from the same local sources can be automatically integrated based already established schema mapping. Therefore, this approach is especially useful for the datasets that update frequently such as urban sensing data.

3.4.4 Data cleaning and normalization

Before data from different sources are merged together, data cleaning and pre-processing routines need to be gone through to improve the coherence of values for the same entities and attributes. Tasks in this step include data type convention, attribute value normalization and hidden information extraction. Besides, data formats and constraints are sometimes included to improve the validity of dataset. For example, it is essential to make sure that every individual in the person data has an age value between 0 and 120. Some ontology editors like Protégé have incorporated a data property component, which is explicitly used to set constraints for the types and formats of classes.

[3] The generalized Levenshtein distance is measured by the minimal possibly weighted number of insertions, deletions and substitutions needed to transform one string into another string.

In this study, a series of look-up tables are prepared for the normalization of attribute values like street name and business name. The "street" of address information is important for buildings or businesses. But in some datasets, "street" names come with abbreviated forms (e.g. "Ave" for Avenue and "Rd" for Road). A look-up table with street names and corresponding abbreviations was used to ensure the consistent "street" format across all datasets. There are also some routines targeted at ad-hoc data issues. For example, the building data from Singapore Land Authority (SLA) has the 6 digit postcode information encapsulated in the 30-digit building identifier numbers. Therefore, a simple function is included to extract the postcode information. Other common data issues like casing, categorization inconsistency, and inconsistency in spatial and temporal representations are also dealt with in this step. In the future, it is more desirable to incorporate some of these look-up tables and processing routines into the ontology. In a nutshell, the purpose of this step is to standardize the data types and formats to make datasets from heterogeneous sources more coherent for the following integration steps.

3.4.5 Data quality assessment

Data quality is an old problem rooted in data collection and processing. As the sources of datasets become increasingly heterogeneous and electronic data become more pervasive and easy to spread, data quality issues have acquired renewed attentions in data and information related disciplines like statistics (Karr, 2006; Winkler, 2004). It is recognized that data with poor quality can create significant economic inefficiency, cost and performance issues.

The data issues presented often include inconsistent data types, misspellings and mistakes during data entry, missing information, as well as outdated records, information duplication, inaccurate geographical tags and time stamps, erroneous records when urban sensing devices went wrong. These issues are more manifested in datasets from many emerging sources. In the volunteered geospatial data, errors, incompleteness, and redundancy are common (Mooney et al, 2010).

Therefore, it is essential to inspect the dataset before integration, singling out valid information and filtering out unreliable information. If the quality of inputs is not controlled, the "dirty data" from one source can contaminate other data in the process of integration. Nevertheless, a broader range of data sources also provide new opportunities to detect and correct data errors through cross-validation. In order to determine which information can be preserved and which cannot in the data integration process, it is an important prerequisite to evaluate the quality of datasets through preliminary dataset analysis (Rahm & Do, 2000). As argued by Boin and Hunter (2008), data quality measures can help users make informed choices towards reducing possible uncertainty in the data.

The data quality literature has provided a comprehensive conceptualization and classification of the dimensions pertinent to data quality issues, although there are discrepancies in the definitions. As shown in Table 3-4, data quality matrices can cover a long list of dimensions like accuracy, reliability, consistency, completeness, timeliness, unambiguousness, source trustworthiness, and understandability (metadata) and value-added. However, the quality of data is not independent of the needs of data users. The importance of the dimensionality of data quality is to a large degree depends on the task at hand. Therefore, user's input on the allocation of weights to different measures is critical in order to generate a composite data quality index (equation 3-1) to help guide the data retrieval and integration process.

$$DR = \sum_{i=1}^{N} w_i \, s_i \qquad (3\text{-}1)$$

Where *DR* is the data rank index; *N* denotes to the number of measures assessed; s_i is the rank of the dataset across all available datasets when assessed by measure *i*. $w_i$is the weight allocated to measure *i*.

The reputation and reliability of data sources are established through the confidence on data providers and preliminary data quality assessment. Data quality measures provide some factual basis for the evaluation. Dataset with low data rank index usually means more reliable information and high selection priority. In contrast, data with high data quality index indicates fewer added values for the tasks, which will result in reduced use. However, judgment and subjectivity are introduced into the data quality assessment in the steps like measures selection and weights assignment.

Table 3-4. A Summary of Data Quality Assessment Measure

Dimensions	Descriptions	Reference	Measures
Accuracy	Extent to which data are correct and reliable	Wang & Strong (1996); Bratini et al. (2009)	--
- Syntactic Accuracy	Weather value matches with corresponding definition domain in the real world.	Bratini et al. (2009)	The distance between the value stored in the database and the correct ones
- Semantic Accuracy	Accuracy of the semantics of attributes	Bratini et al. (2009); Girres & Touya (2010)	Levenshtein distance from the reference values
- Spatial Accuracy	Accuracy of the positions, geometries and spatial relationships of geographical features.	Girres & Touya (2010), Yang et al. (2013)	Number of valid values / total number of values expected
- Temporal Accuracy	Accuracy of temporal attributes and temporal relationships of features	Bratini et al. (2009), Girres & Touya (2010), Yang et al. (2013)	number of values with valid and complete temporal information / total number of values in dataset
Completeness	Extent to which a given data include corresponding real world features.	Bratini et al. (2009), Jark et al. (1995), Girres & Touya (2010)	--
- Omission	Absence of existing features	Girres & Touya (2010)	Ratio of the size of none null values in the dataset to the total number of values in the real world.
- Commission	Presence of excess features not existing in the real world	Girres & Touya (2010)	Ratio of excess values in the dataset to the total number of values.
Consistency	Degree that data conform to semantic or logic rules	Bratini et al. (2009)	--
- Logical consistency	Degree that data comply with logic rules and integrity constraints	Bratini et al. (2009), Girres & Touya (2010)	Number of consistent values /number of total values
- Semantic consistency	Degree that data comply with semantic rules	Bratini et al. (2009)	Number of consistent values /number of total values
Source Trustworthiness	Degree of confidence on data sources	Bratini et al. (2009)	max(0,1)
Unambiguity	Concerns whether the values are unambiguous and corresponds to clearly defined real world values	Bratini et al. (2009)	Number of unambiguous values / total number of values expected
Uniqueness	Concerns whether the information can only be provided by the given dataset		Number of values / Total Number of values from all datasets

3.4.6 Integration operations

Because ontology has a richer representation of the correspondences among classes, it enables users to formulate more specific and tailored queries based on custom defined transitivity rules. In this study, three operations are designed and developed to facilitate the integration of multiple datasets.

1. **Construct entity list:** construct the most comprehensive list for the entity given the datasets at hands and the targeted time periods. When building the list, the algorithms are designed to traverse the graph-structure ontology to include all instances of targeted entities and instances of its subclass entities that can be found in the datasets. For example, when constructing the list of private residential buildings, the function automatically includes instances of condominiums, apartments and detached-houses etc, which are subclasses of the private residential buildings.
2. **Retrieve attribute values**: select and complement values of the targeted attribute for an entity. Based on the list of instances belonging to the entity, the algorithms are implemented to search and retrieve corresponding attribute values from all relevant entity classes (including superclasses and subclasses). For example, when querying the "built year" of private housing buildings, in addition to search the datasets for the private housing buildings, the function would also look for the "built year" information in the building (superclass) datasets and as the apartment (subclass) datasets. The sequence of information retrieval is dependent on the quality of dataset evaluated in Section 3.4.5.
3. **Impute missing values:** three types of imputation functions are defined in the study. First, for the attributes of some geospatial objects, missing values of attributes of the targeted instances are substituted by the existing attribute values of instances spatially similar. Second, regression models are calibrated to estimate the possible values of an unobserved attribute based on its correlations with other observed variables. For example, Zhu and Ferreira (2013) use a multinomial logit model to infer the missing household income information based on the characteristcis of the building that the household occupy. Third, a classification model is developed to estimate the possible open hours of establishments. This imputation approach is discussed in details in Section 4.3.2.

3.5 Integrating building information for built environment measurement

Buildings are among the most important geospatial objects in cities. They are the places that people live, work and engage in other activities. However, buildings are complex because the classifications are various. Detailed and complete information on buildings is usually rare but very valuable for the measurement of urban built environment as well as in land use and transportation modeling. For example, Figure 3-5 shows the thematic map of buildings by types using the building footprint shapefile dataset from SLA. The categories "Block" and "Standard" are ambiguous in meaning and are clearly not consistent with other categories like "HDB" and "Industrial", which imply the main usage of buildings. These two types of buildings amount to 55,846 records, which makes up around 34% of total building records. Also, the types of buildings crucial to the characterization of urban spaces like "Commercial" or "Civic" are not classified. Besides the types, the dataset only have address information of buildings. Important building properties like time of completion, storey and number of units are all absent. Therefore, although the building directory from SLA has a detailed spatial representation, the attributes of buildings are very limited, which makes the data insufficient to support the level of urban built environment measurement and urban modeling desired by researchers. In this sense, it is essential to search for building information from other sources to complement the SLA dataset.

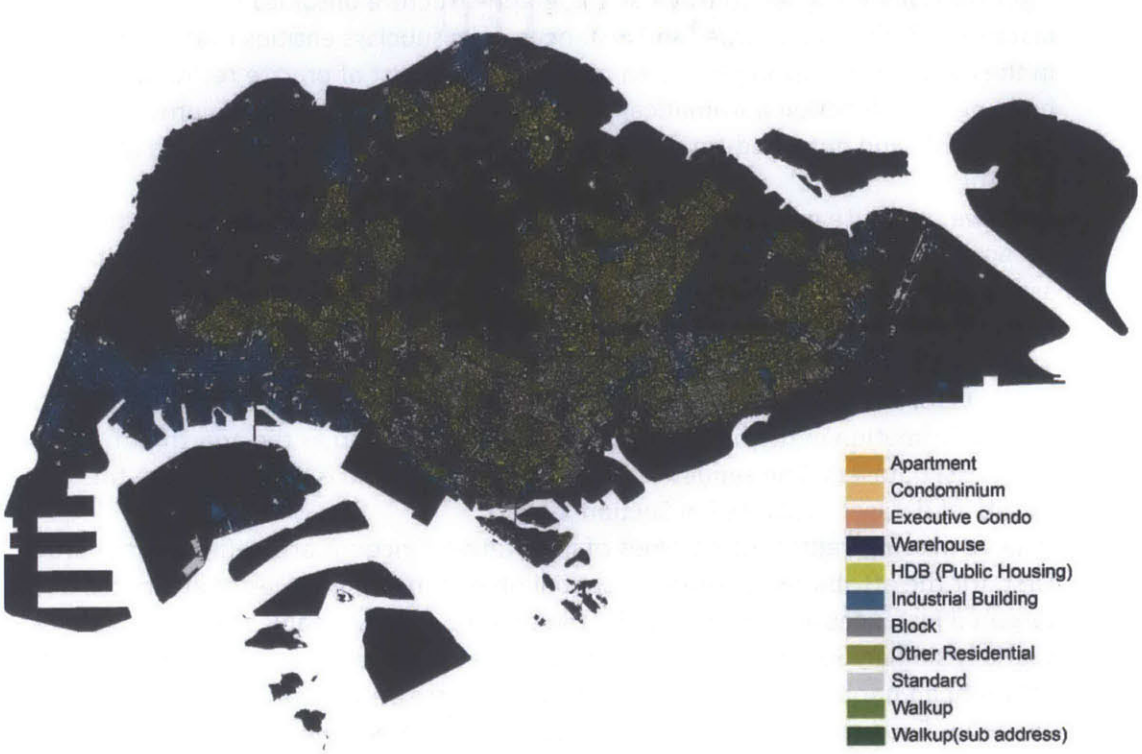


Figure 3-4. Building footprints by categories from SLA.

Unlike conventional ways of data gathering that is usually a one-time effort, the data collection and integration has increasingly become a continuous process, because the information from emergent sources like OpenStreetMap is updated frequently. In addition, over the time, one may discover new sources of building information become available for the analysis. Hence, the information of buildings needs to be dealt with in a flexible and evolving way, starting from a base ontology and a few data sources, which can be extended later by including a richer set of semantics and additional sources of information about buildings. In this section, the proposed approach is applied to integrate building information of Singapore for the activity-travel analysis. The objective is to generate a comprehensive list of buildings with essential attributes such as the number of floors, the number of units, year built and type of lease. Except for the ontology creation, the other steps of the integration practice are accomplished by processing the scripts and functions written in R.

Table 3-5 lists a number of datasets pertinent to buildings that have been gathered from different sources, including the building directory from the Singapore Land Authority (SLA), the housing unit transaction information, and the semi-structured building lists collected from various websites. These datasets differ in building types, numbers, attributes and data qualities. Although the official building data provided by SLA are very reliable and complete in terms of the number and the spatial locations (represented by address and postcode) of buildings, the attributes of buildings of great interest to analysts like the year built and the floor information are absent. Besides, the classification of building types in the dataset from SLA is ambiguous. Therefore, information like building types, capacities, finished year, property value and unit types and areas need to be retrieved from other datasets, which are considered less authoritative and complete but more error-prone.

Table 3-5. A list of datasets available for building information integration

Data type	Coverage	Source	Records	Attributes	Year
Building List	All types of building	SLA	162491	postcode, address, type	2008
Building List	All types of buildings	Street Directory	111591	postcode, address, type	2013
Building List	HDB, Condos and Apts	FourSqaure	14256	postcode, address, type, year	2013
Building List	HDB blocks	HDB web service	9276	postcode, address, type, year, unit type and amount	2013
Housing Unit Transaction Record	HDB and Private housing units	HDB web service/Realis	54375	address, type, lease_commence_date, floor area, transaction price, min/max level	2013
Building List	All types of buildings	EMPORIS	6868	Address, postcode, year, floor, height	2013

The first task is usually to identify the reference attributes of the entity class. In the case of this study, because addresses are available in most datasets and are semantically consistent, it becomes an ideal attribute for the cross-reference among datasets. Although in the real world, one building may have multiple addresses, it is assumed here that one address corresponds to a pseudo building instance, which can be a part of a large building complex.

As mentioned in the section 3.4.4, a series of data normalization routines were applied to the attributes like address, postcode, building types and year to ensure the values of the same attributes are coherent across different datasets. In addition, information in the housing unit transaction records was converted to information for the buildings. For example, the room numbers of units are parsed to estimate the number of floors and the total number of units in a building.

The quality of data is evaluated at the attribute level as opposed to the dataset level because there are significant variations in the reliability of information for different attributes from the same data source. When assessing an attribute, the values of the attribute from a dataset that is most trustworthy are picked out as reference for the attribute values from other datasets to compare against. The most important attribute of building is probably the functionality types. Types of buildings are tricky to deal with because of its hierarchical classification structure as shown in the building ontology diagram in Figure 3-2. For example, EMPORIS classifies buildings at a very detailed level (e.g. primary school, church, fire station and theme park etc). Building list from SLA has a mixed levels of classification with some buildings are tagged as "Residential" while some others tagged as "Walk-up".

Based on the analytic need, an accurate classification of buildings at medium aggregated level (HDB, Apartment, Factory, Warehouse, Retail and Office etc) is desired for the result of data integration. Therefore, the Housing and Development Board (HDB) building list collected from the web service of HDB official website is selected as reference because of its high reliability and completeness. After that, lists of buildings tagged as "HDB" in other datasets are compared with the reference HDB list. Table 2-5 shows the results of data quality measures used in the assessment. It appears that housing unit transaction dataset has the lowest semantic error (misclassification of building types), while EMPORIS has the highest error. This is because most of the HDB buildings in the EMPORIS dataset are simply tagged as "Residential", which indicates the quality of user contributed information is usually poorer than desired.

Table 3-6. Data quality measures for the type of buildings

Dataset	Semantic Error	Completeness	Commission	Omission	DR index
	$\frac{n(D1 \cap D2)}{N(D1 \cap D2)}$	*N1/N2*	$\frac{N(D1 \notin D2)}{N2}$	$\frac{N(D2 \notin D1)}{N2}$	
HDB web service dataset	--	--	--	--	base
Housing unit transaction	0.0005	0.9428	0.0414	0.0965	1.5500
SLA	0.0065	1.1523	0.2033	0.0493	2.8500
Streetdirectory	0.0134	1.1051	0.1298	0.0211	2.7500
FourSquare	0.0183	1.1341	0.1837	0.0525	3.8500
Emporis	0.9060	0.0524	0.0126	0.9590	4.0000

Note: D1-assessed dataset, D2-refernce dataset, N1-number of building records in D1, N2-number of building records in D2, $n(D1 \cap D2)$-Number of matched records with different "building type" values, $N(D1 \cap D2)$-number of matched building records, $N(D1 \notin D2)$- number of records in dataset D1 but not in dataset D2, $N(D2 \notin D1)$- number of records in dataset D2 but not in dataset D1.

When calculating the composite data quality, special attentions are given to value accuracy. Semantic error, which indicates the misclassification of building types, is in general less tolerable than other measures. Besides, commission of additional building records to the base HDB list may also indicate misclassifications. In contrast, completeness and omission are less important because the missing buildings have chances to be complemented by other datasets. Thus, the weights allocated to four measures: semantic error, completeness, commission and omission are 0.6, 0.05, 0.3 and 0.05. The resultant data rank index (DR_index) suggests that building type information should be retrieved sequentially from the HDB web service dataset, the housing unit transaction dataset, the streetdirectory dataset, the SLA dataset, the FourSquare dataset and at last the EMPORIS dataset.

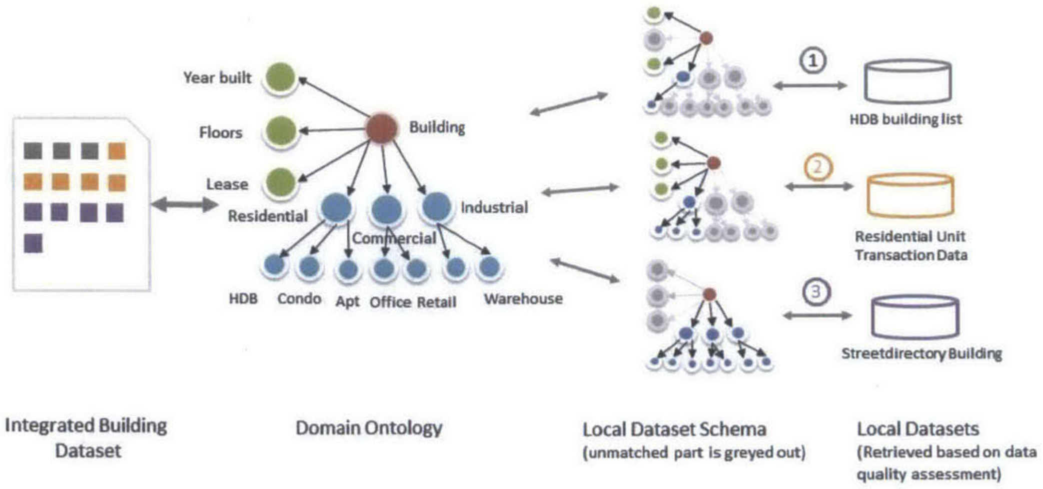


Figure 3-5. Illustration of building type integration process

Figure 3-5 illustrates the process of merging building information from multiple datasets based on the building list provided by SLA. In this case, categories of building types from local data sources are mapped to building classes and subclasses pre-defined in the example ontology. The building classes are semantically enriched by the building types that appear in local data sources (Appendix I). A generic R function is written and applied to carry out the fusion of building type information. Because the graphic-based schema mapping allows the data user to easily traverse the graph, the returned building types not only include the values from the inquired datasets but also their superclasses. For example, if an inquired building record is tagged as "primary school", then the returned building record will not only be tagged as "primary school", but also be tagged as "community" and "civic", which are the parent and grandparent classification types of "primary school" in the ontology.

Table 3-7 shows the contributions of different datasets in providing useful building type information for the integrated data. As we can see from the table, the streetdirectory appears to have the most significant contribution because of its very detailed classification of building types. In comparison, *Emporis* dataset provided very limited building type information. Figure 3-6 shows the thematic maps of integrated buildings by different levels of categorizations. Category 1 building types are very generic and only have 8 types, which leads to around 12.1% of buildings unclassified. In contrast, Category 3 building types are quite specific, which results in a large number of buildings with missing Category3 information. See Appendix I for the three levels of building type categorizations.

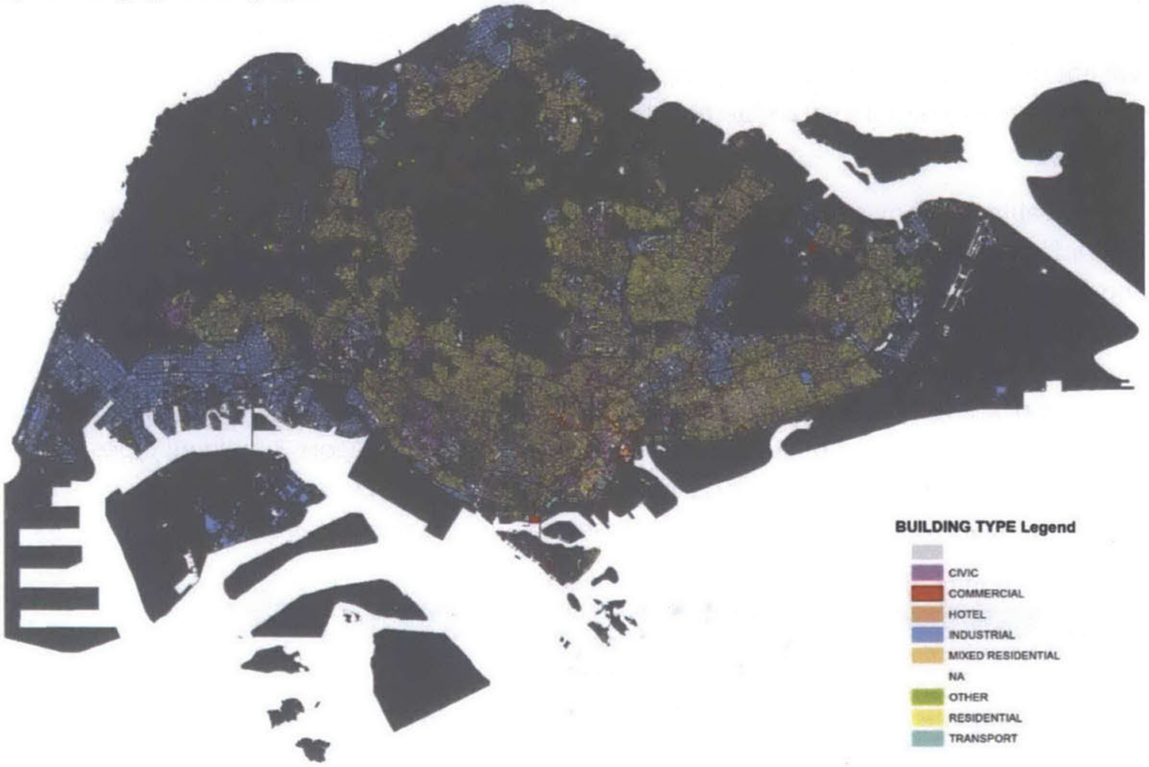
Table 3-7. Number of building records that are classified in the process of integration

Dataset	Category 1	Category 2	Category 3
HDB Web service data	9276	9276	9276
+ Housing Unit Transaction	33687	33687	18327
+ Streetdirectory	119768	84451	84451
+ SLA	144604	135329	95409
+ FourSquare	144900	135707	95929
+ Emporis	144908	135709	95929

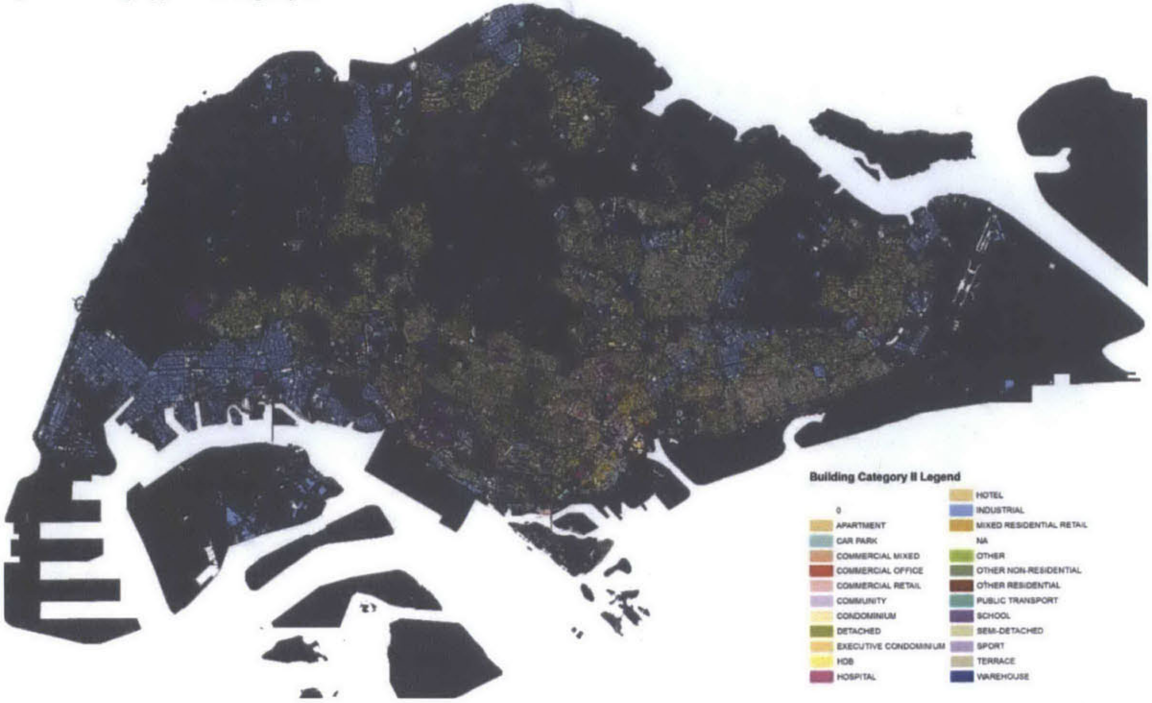
The opportunity for users to verify where data originates from and how it was combined and converted into its current form is critical for enabling users to distinguish between facts and assumptions and, in consequence, to establish trust in integrated data. Therefore, in the integrated dataset, an additional column called "building_type_source" is added in the integration process to address the lineage and traceability issues of integrated dataset.

Likewise, other attributes like building finished year are also merged following the same procedures. However, because 76.7% of buildings did not find finished year information from any of these datasets, a spatial similarity-based imputation is performed to infer the finished year of remained buildings. In this case, a set of heuristic rules were designed to determine the how the missing finished year values are filled by using the corresponding available information of nearby buildings. These rules dictate that the search will first be directed to the buildings of the same type in the same development project as the targeted building, followed by the buildings of the same type on the same street. If these two searches fail, the values for the nearest building of the same type will be used. This imputation approach helped to bring the percentage of buildings with missing completion year values down to 15.7%. Most of these buildings are industrial buildings in recently developed areas, as shown in Figure 3-6.

a) Building type - Category 1



b) Building type - Category 2



c) building type - Category 3

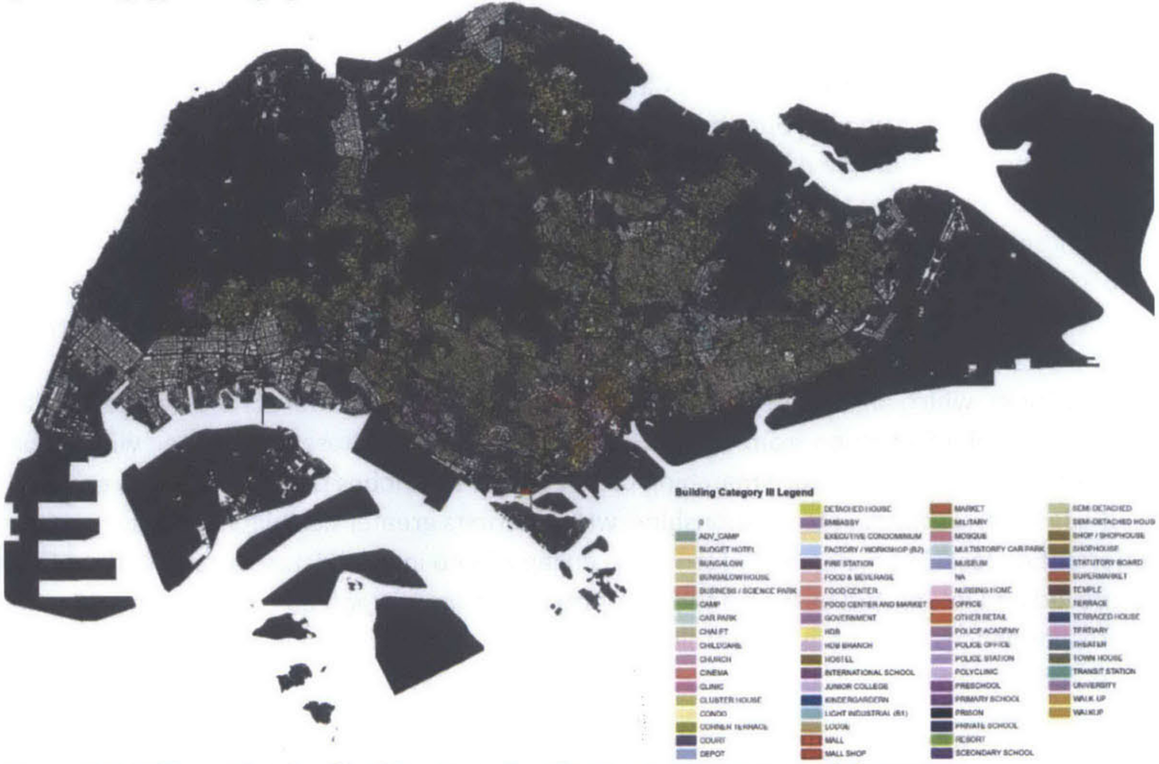


Figure 3-6. Different levels of building type classification in the integrated dataset.

Note: buildings with missing type information are filled in white.

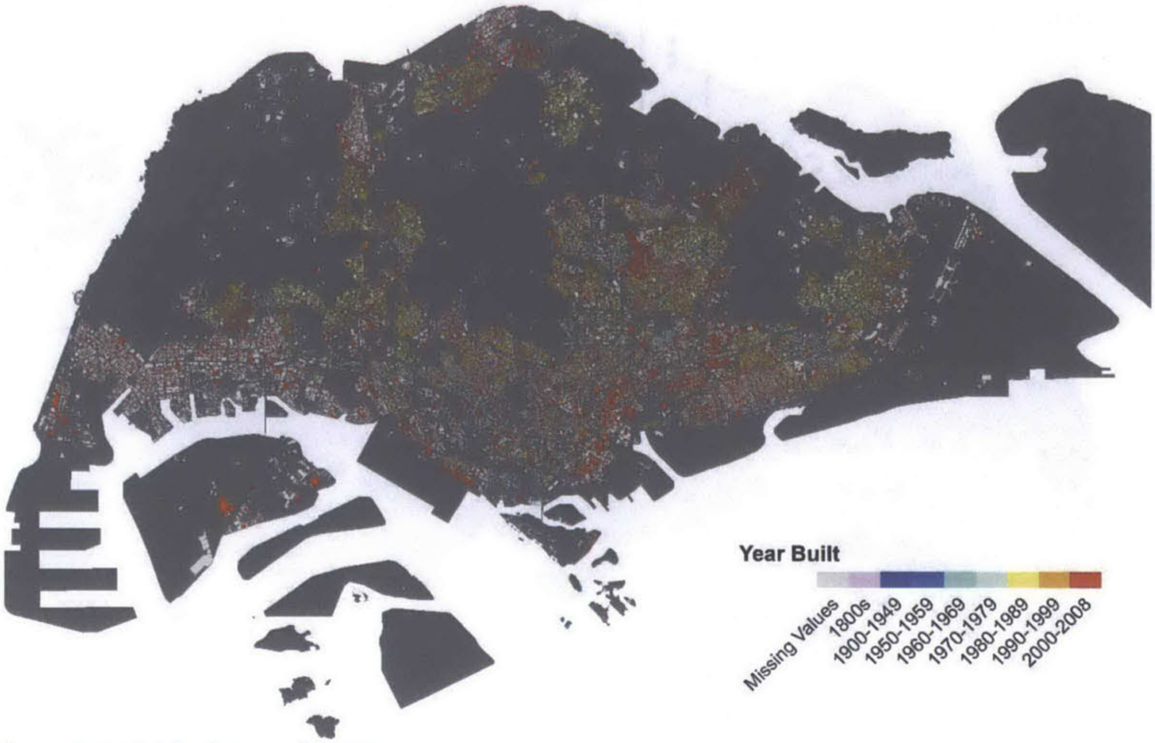


Figure 3-7. Finished Year of buildings

Note: buildings with missing year information are filled in white.

3.6 Summary

The scarcity of spatiotemporal data restricts the depth and scope of researches that can be conducted. Although data become available from more and more sources, merging information from multiple datasets is held up by the data interoperability and quality issues. This chapter presented an ontology-based schema mapping approach to reconcile the existing semantics and data representation conflicts between datasets. The approach explicitly builds the domain knowledge into a task-specific ontology to help integrate data for the urban built environment measurement and activity-travel research. The ontology is regarded as globe schema using the generic representation of entities and relationships among them, which are independent from the available data. To address the heterogeneity and scarcity of information from various data sources, a local dataset is matched with global schema via a graphic structure matching table consisted of nodes corresponding to entities and edges corresponding to relationships, which permits greater flexibility and less information loss. Then, needed information encapsulated in local datasets can be identified, retrieved and integrated based on the interpretation from the agreed concepts and correspondences laid out in the ontology. The quality of information in each data source is assessed in the process of data integration to avoid the contamination of problematic data.

We demonstrated the approach by applying it to gather information about buildings such as usage type, stories and completion year, which are all missing in the official building directory dataset. A significant improvement is observed in complementing the missing values of attributes for buildings, as evidenced by the examples of building type and completion year. In addition, one important reason of using the ontology-based approach is to make the data integration process more automatic and reliable, because the data integration is likely to be a repetitive process in the future planning and research practices as longitudinal data become increasingly available. Further, the ontology can also be used as a single entry point for users to query the integrated information without the need to have the knowledge of each individual dataset.

Further, as urban modeling and simulation becomes increasingly complicated and data demanding, the proposed ontology-based data integration approach may help to improve the efficiency and reliability of many aspects of urban modeling. For example, it is always a challenge to generate a good synthetic population for agent-based urban modeling. It requires reconstructing the current "state of the world" by estimating detailed spatial and aspatial attributes of agents like individuals, households and firms from limited – and not always consistent - data sources. In order to assign synthetic agents such as households to building units in a plausible fashion, good information on buildings are needed. In the case of Singapore, although the geometry of buildings and the transactions of some units are available, it is difficult to acquire information on the number of units in total in each building, nor the time of construction or last renovation from official sources (Zhu and Ferreira, 2014). To overcome the data limitation, we need to infer required information from existing data sources and gather imperfect information about the required data from additional datasets. Building age can be estimated if construction date from any unit in the building is available. Building height may be estimated from the highest floor level recorded in a sales transactions for the building. In this case, the information about buildings are dispersed in a number of imperfect datasets. It is not immediately clear to researchers what can be extracted from which dataset if one doesn't have a good knowledge of all datasets. Using ontologies to make these inferences more automated and reliable is especially important for the new round of urban simulation models that work at individual household and building scale.

In fact, ontology can play a more important role in the field of urban modeling and analysis by possibly providing a consistent knowledge base and hence stimulate more connections and collaborations among researches. As argued by Yin and Batty (2013), "the outstanding data and methodological challenges require modelers, who tended to work in isolation in the past, to develop an in-depth understanding across model types and styles regarding possibilities in concert with one another, so that models can be linked to each other and to the policy questions on which they are focused". In addition, Ontologies are a key component of the semantic web, which is defined by W3C as the "common framework allowing data to be shared and reused across application, enterprise and community boundaries". It is also likely to be a critical part of the web 3.0 initiative, which stresses on the shareability, interoperabity and connectivity of distributed datasets based on the semantic web and linked data technologies (Hendler, 2009). Thus, it is proactive and timely to consider the application of ontology-based data integration approach for both online and offline data fusion.

Chapter 4. Measuring and Characterizing the Urban Built Environment

4.1 Introduction

The urban built environment and its transportation facilities are important factors that influence the daily life functions, activity spaces, and travel options of urban residents. In this sense, activity-travel studies are required to quantify the obscure built environment components and to capture the reciprocal impacts between human activities and urban spaces. This situation is the case for Singapore, a rapidly changing city shaped by various trends from service and job decentralization to personal mobility improvement and new urban lifestyles that are increasingly influenced by information and technology as well as comprehensive urban policies. These trends have molded Singapore into a city with heterogeneous urban spaces but also have given rise to diverse and complex location and time preferences for different urban activities among residents. Although many previous studies have formulated built environment indicators with relatively coarse resolutions (Talen, 2002), the advancement of the research agenda in activity-based studies calls for improved urban built environment measures that can inspect and compare urban dynamics and human activity-travel patterns at more fine-grained resolutions, both spatially and temporally. This aim is becoming a possibility via the availability of more detailed, disaggregated, and comprehensive datasets from various sources, as described in Chapter 3.

The objective of this section is to formulate a group of generic metrics that are able to capture and measure the obscure components in the urban built environment that have differentiated impacts on the locational choices of different activities to enable the detection of the underlying activities in urban sensing data with the assistance of the selected indicators. However, the answer to the question of which built environment components correlate with activities is not as obvious as might be assumed. This situation prompts a search for good measures of the urban built environment. In the previous literature, some studies have considered the urban built environment solely in terms of the morphological patterns and fabrics of the urban form. Others include the socioeconomic aspects of the urban form, i.e., population density and average land price. In this study, we include not only the physical structure of the locations but also the spatial distribution of businesses and the topology of transit network in the urban built environment measurements. These dimensionalities of the built environment display different degrees of association with the behaviors and choices pertinent to human activities. The measurement process draws on spatial and temporal detailed datasets from the data integration processes described in Chapter 3. Therefore, special attention is focused on the spatial and temporal representation of measurements because these two dimensions are generally well recorded in urban sensing data.

4.2 Related work

Measures of urban built environment have for years been an important topic in the urban planning literature. It is generally acknowledged that the social, economic and environmental outcomes of urban form can only be captured by evaluating the relative differences in the type, pattern, and intensity of development in different urban areas, although considerable ambiguity remains in the definition and understanding of urban built environment (Clifton et al., 2008). In addition to the urban planning and urban design, fields like land ecology, public health and network analysis have growing interests in formulating built environment measures from the perspectives of their research scope, which help to enrich the defining dimensions and measurements of urban form. In this part, I will briefly summarize the previous research efforts in built environment measurements in these fields, and bring to light the indicators that correlate with human activities.

4.2.1 Urban Planning and Physical Activity

In the context of urban planning, built environment has been investigated extensively in two areas: quantifying urban development pattern (Song and Knaap, 2004; Galster el al., 2001) and investigating the effects on travel behaviors (Cervero and Kockelman, 1997; Crane, 2000). The measures developed in former tend to emphasize land use composition, street network topology and accessibility at neighborhood level or bigger scales. For example, Galser et al. (2001) identified eight dimensions of land use, namely density, continuity, concentration, clustering, mixed uses, proximity, centrality and nuclearity.

In contrast, researches emphasizing the effect of urban built environment on human activities usually focus on constructs like accessibility, mobility, and density as well as the social-economic aspects of urban form. For example, Cervero and Kockelman (1997) classified urban built environment measures into three big categories: density, diversity and design. Detailed geospatial information like points of interest (POIs) and streetscapes including sidewalks, parking and lighting were used to construct the built environment variables associated with travel demand. Krizek (2003) proposed an operationalizable neighborhood accessibility measure that is composed of three sections: density, mixed land use and streets/design. Krizek argued that the variables under three dimensions are simple, parsimonious and significant in influencing travel behaviors at neighborhood level. In an attempt to pinpoint the correlation between vehicle miles travelled (VMT) and the built environment of residency location, Diao (2010) included a group of accessibility variables measuring distances to nearest main activity centers like grocery store, school, mall and theater, other than the variables describing local road and transit systems.

Similarly, in the field public health, it is believed that environmental intervention may lead to changes in physical activity, which can help to reduce the risk of having obesity and promote healthier lifestyle. Studies in the field usually resort to the planning principles and measures similar to those developed in the urban planning community but have a conscious emphasis on the walkability, and accessibility to open space and sporting sites (Durand et al., 2010).

In general, these measures in the urban planning literature are designed not only to quantify urban form variability but also to embody the key features of development schemes. For this reason, the measures are readily to be used for evaluating and comparing the impacts of alternative planning scenarios.

4.2.2 Landscape Ecology

Landscape ecology is concerned with the expansion of urban areas and the potential risk it may pose on the natural environment and wild species (Durand, 2011). Therefore, the evolution of land covers is of particular interest to the researchers in this field. Data on land cover changes are usually come from aerial photography and satellite remote sensing images. In order to identify the types of land cover in these datasets, landscape measures focus on depicting the shapes, the compositions, and other morphological features of patches like continuity and compactness. The morphology of land patches also implies the natural and artificial geographical boundaries of areas like rivers, hills, railways and expressways.

Urban form measures of this field have been used widely to evaluate environmental topics like the loss of agricultural lands as a result of urban area expansion, and the correlation between landscape and animal distribution. However, they are rarely applied to assess the effect of urban landscape on human activity and movement. As pointed out by McGarigal and Marks (1995), that isolated patches usually have fewer species, and isolated land use in the urban area may also have lower attractiveness to human activities. Also, a large, single-use urban area and a fragmented mixed use urban area may have different attractions to urban residents. In this sense, bringing in some of the meaningful measures of landscape ecology to characterize the land use dimension of built environment will be helpful.

4.2.3 Urban Design

Urban design focuses on the design details of urban built environment and their effects on how peoples perceive and experience space. Physical features and design elements of a place are considered to be correlated with a variety of physical activities and social interactions, usually by shaping the walkability, accessibility and attractiveness of places (Clifton, et al, 2008). Good urban design creates a distinctive sense of place and attracts more people to visit. For example, street features like wide sidewalks, sufficient coverage of tree shade, and the enclosure from building facades, the presence of streetlight are usually considered to be associated with good walkability. In an effort to rethink and measure urban intensity, Sevtsuk et al. (2013) looked at specific street layout and building design variables including sheltered walkways, building setbacks and entrances, and ground-level floor heights in addition to three common urban form variables: gross floor area, spacing between buildings, and pedestrian network of a place. Evidently, quantitative analysis in this field deals with built environment measurement at very fine-grained resolution.

Within the field, a considerable number of researches focus on quantitative models of the configuration of urban space and the ways that human use space. This area is typically known as space syntax. From a topological stance, space syntax analysis is based on the geography theory, and is regarded as an extension of network analysis into architecture and urban planning (Hillier, 2007). One important aspect of space syntax analysis is axial analysis, which inspects the geometrical properties of road network on the distribution of movements of both vehicle and pedestrian. Typical measures like depth and number of turns from starting points are used to gauge the physical environment of different paths from origin to destination in an urban space like a building, a park or a neighborhood place.

However, as critiqued by Ratti (2004), axial analysis based on the topological representation of urban spaces is less compelling because it underestimates or even ignores other metric characters of spaces. Clearly, simplified geometric properties of spaces like lines and grids are only part of the complex urban system and their effect on human behavior is inconclusive without the support from additional empirical evidences. To overcome this limitation, in the study investigating the spatial location choices of retail and eating, Sevtsuk (2010) not only used the geometric features along the shortest paths toward buildings like distance, betweeness, number of interactions and number of turns to measure accessibility, but also the indicators of land use and businesses reachability at building level.

In the literature of urban design, the importance of individuals' perceptions and experiences of built environment and attitudes of places has been generally recognized. Handy et al. (2006) differentiated the objective measures of built environment with the perceived measures and found perceptions of safety, attractiveness of place, and distance played even more significant roles in walking and cycling travel than the corresponding objective measures, suggesting how individuals perceive the neighborhood space and scale, and how they filter spatial information are very critical when making spatial choice decisions.

4.2.4 Gaps in the literature

The urban built environment is a comprehensive concept. Each reviewed field has formulated measures used to evaluate the aspects of the built environment that are of interest to the respective field. Most of the measures are simplified and less sensitive to the components in the built environment that can lead to spatial variability. Although a growing number of voices advocate for more comprehensive and cross-disciplinary measures for empirical analysis (Ratti, 2004; Clifton, et al, 2008), the response has been inadequate (Sevtsuk, 2010), primarily due to a lack of awareness of the measures used in other fields and the limited data availability.

In reality, most of the data required to measure the built environment are only available at the aggregated level. One problem arises when studies compare the activity-travel behaviors of individuals across judgmentally pre-defined neighborhoods, which are usually surrogated by census tracts and traffic analysis zones (TAZs) in which urban form data are more readily available and easily matched to travel data. However, a slight change in the scale of the geographical analysis could give rise to divergent measurement results. This situation is known as the modifiable areal unit problem (MAUP), and the problem is more evident if different built environment attributes display different spatial extents of influence on travel choices. The question of how to address this problem has become an important challenge for researchers (Horner and Murray, 2002, Zhang and Kukadia, 2005).

Although urban design methods measure space at a notably high level of resolution, the sources of data required for computing these measures are difficult to acquire at the scale of a city. Data are often collected through field observations or interviews (Clifton et al., 2008), and a significant amount of manual efforts and resources must be invested. Hence, most studies that use urban design measures focus on a site or a neighborhood in the city.

Furthermore, most of these spatial metrics have traditionally treated urban space as stationary, and the temporal changes of various elements of the urban built environment are minimally accounted. As argued by Kevin Lynch (1960, p108), "we need fresh thought on the theory of forms which are perceived as a continuity over time...". Despite the stationary physical components of the urban built environment, the city is dynamic due to such moving elements as people, activities, and time-varying urban services and opportunities. In recent years, increased interest has arisen in urban dynamics and longitudinal evolution. However, the scope of research has been limited to investigation of the longitudinal evolution of the urban form using relatively coarse satellite land use images.

The current study draws on interdisciplinary research efforts, innovative data collection methods, and an integrated approach to examine different dimensionalities of the urban built environment that are deemed to affect activity-travel patterns. Using the geospatial data at the disaggregated level, each built environment variable was computed at the local area (the catchment area of transit stations) to describe the "micro-environments" that people experience in the locations in which they engage in daily activities.

4.3 Data and Methods

To capture the richness and diverse dimensions of the urban built environment and their impacts on human activities, the measurement process uses the datasets for buildings and establishments fused in the data integration process described in Chapter 3. In addition, the land use plan data for the year 2008 provided by the Urban Redevelopment Authority (URA) of Singapore is used to gauge the landscape mix of the catchment areas of the transit stations. This work also uses transit schedule information to assess the position of each transit station in the entire transit network.

The temporal dimension of the urban built environment is primarily reflected by the temporal variation of the available opportunities and services on the supply side. For businesses, the open hours of venues taken from FourSquare (FSQ) were applied to establishments via a matching procedure. For transit service, the operational frequencies and headways of transit routes in the schedule were used to mimic the temporal variability of transit services. Moreover, the synchronicity among various datasets is usually desirable for analysis and modeling efforts. However, it is unlikely that all supporting datasets for the same years as those of the travel survey data and EZ-Link data will be acquired. The household survey data are taken from the year 2008, and the available EZ-Link data are from the year 2011. However, the transit services schedule and the FSQ dataset were collected in 2013. In this case, we retain the FSQ venues for the buildings that existed in 2008. Similarly, for the RTS stations, only those that were in operation in 2008 are retained. Because it is difficult to sift out the stops that were added after 2008, all bus stops in the transit schedule data are retained in the analysis.

4.3.1 Catchment Areas of Transit Stations

Because the EZ-Link transit transaction dataset only records the boarding and alighting stations of transit trips, it is unlikely to determine the types of origins and destinations of the trips. Instead, it is possible to classify the types of places in the local areas surrounding the transit stations. Therefore, the main purpose of the built environment measurement in this section is to quantitatively characterize these local areas surrounding the stations or the catchment areas. The catchment areas of the transit stations are defined as the areas surrounding bus stops or RTS stations that are more accessible by walking from the station than by transferring to another transit route. The catchment area of a transit station also implies the range of potential activity spaces of a station that transit passengers may explore or are planning to visit when alighting at that station.

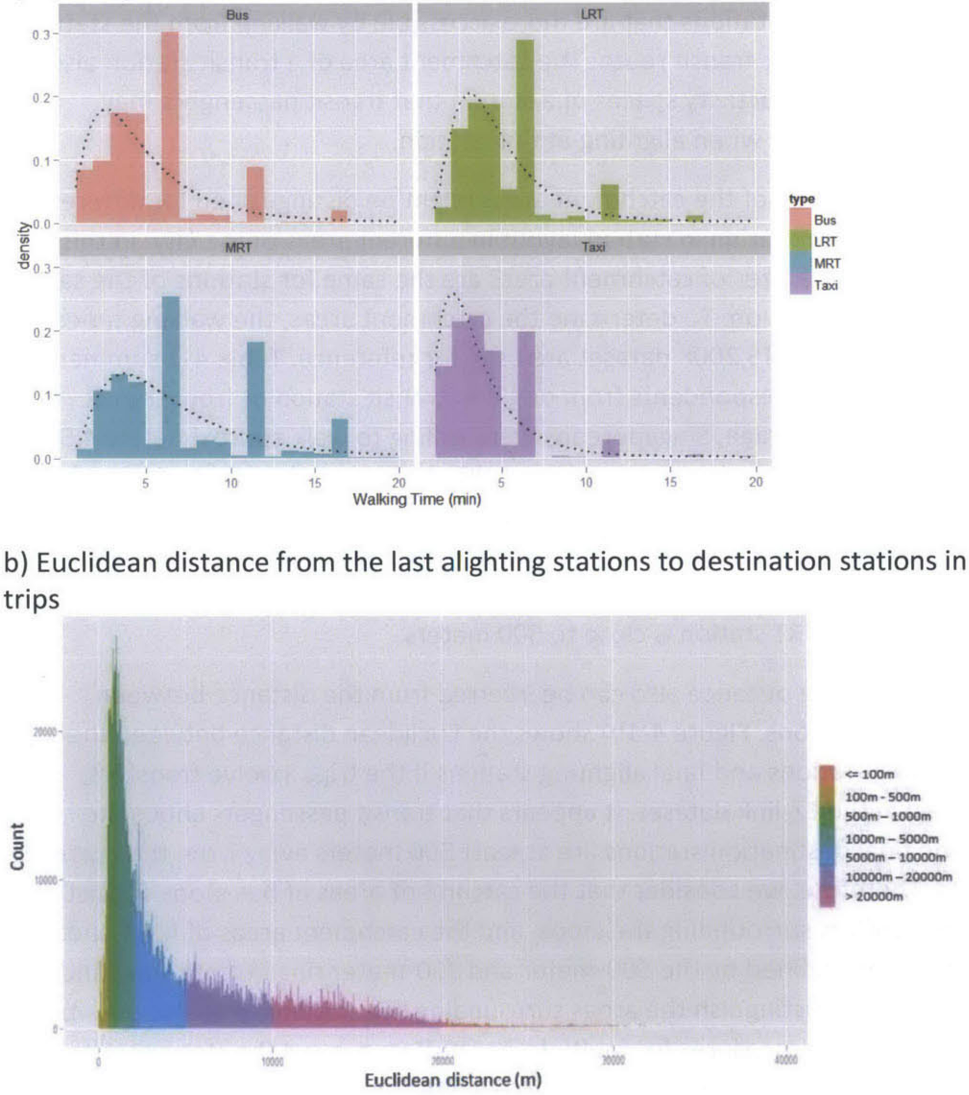
Theoretically, the sizes of the catchment areas might be distinguished for different transit stations due to the variation in station layout in different areas of the city. In this study, it is assumed that the sizes of catchment areas are the same for stations of the same type, i.e., bus stop or MRT station. To determine the catchment areas, the walking times of transit trips reported in the HITS 2008 dataset are used for reference. Table 4-1 summarizes the average walking time for respondents from origin to transit station or from transit station to destination. On average, Singaporeans were willing to walk approximately 4.5 minutes to or from bus stops, 6 minutes to or from MRT stations, and 4 minutes to or from LRT stations in the year 2008. Assuming that the average walking speed is 5 km per hour, the average walking distance to a bus stop observed in HITS trips is approximately 400 meters. The average walking distance to the LRT station is approximately 350 meters, and the mean distance to the MRT station is close to 500 meters.

The range of walking distance also can be inferred from the distance between boarding and alighting stations. Figure 4-1b) shows the Euclidean distance between the intermediate alighting stations and final alighting stations if the trips involve transfers, which are observed in the EZ-link dataset. It appears that transit passengers choose to transfer only if the final destination stations are at least 500 meters away from the current alighting stations. Therefore, we consider that the catchment areas of bus stops consist of the 400-meter ring buffers surrounding the stops, and the catchment areas of MRT and LRT stations are respectively defined by the 500-meter and 350-meter ring buffers surrounding these stations. To further distinguish the areas surrounding the stations, it is also possible to generate built environment measures for multiple ring buffers around the stations, although this topic is not covered in this study due to the time limitation.

Table 4-1. Walking time to public transportation stations

	To Bus Stop	To MRT Station	To LRT Station	Taxi	Transfer
Mean (min)	4.486	6.041	3.934	3.066	4.208
Standard Deviation (min)	3.196	4.303	2.680	2.313	2.878
Minimum (min)	1	1	1	1	1
1st Quarter (min)	2	3	2	2	2
Median (min)	4	5	3	2	3
3rd Quarter (min)	5	10	5	5	5
Maximum (min	35	35	15	30	15
No. of observations	22622	11640	729	1609	53

Source: Household Interview Travel Survey in 2008



a) Walking time distribution to different types of public transportation stations.

b) Euclidean distance from the last alighting stations to destination stations in the transfer trips

Figure 4-1. Approximate walking distance of transit stations.

Note. In a), the distributions of walking time are compared with the truncated normal distributions (dashed lines).

4.3.2 Importance of Business Open Hours

Presumably, the temporal availability of services and opportunities will cause the attractiveness of certain places to vary within a day, which directly influences the location and time choices for activities. However, this temporal variation on the supply side is seldom accounted for in the urban built environment measurement, primarily due to data scarcity. Although crowdsourcing websites, such as FSQ, provide detailed information on the open hours for most listed venues, because most of the information is sourced from users, the data quality is often dubious. For example, names and addresses of businesses are usually inconsistent in terms of semantics. The convenience store "seven-eleven" can be spelled in different ways, e.g., "7-11" or "Seven Eleven". Many buildings and businesses have nicknames known only to local residents. In addition, FSQ uses its own classification system for venues. Semantic matching is required. Therefore, the data integration approach discussed in Chapter 3 is applied to integrate the FSQ venue lists with business data from other sources.

Another issue in crowd-sourced data is incompleteness. In this study, the open hours of 15,600 venues are found in FSQ, which accounts for only approximately 1/10 of the amount of establishments collected from other sources, i.e., streetdirectory. Using concatenation of the name and postcode of establishments as the primary identifier, 4,430 out of 15,600 venues collected from FSQ have counterparts listed in Streetdirectory. To impute the open hours of the remaining businesses that are not found in the FSQ, relatively strong assumptions are adopted, and a series of sequential heuristics rules are applied:

1) Establishments in the government and manufacturing sectors have open hours between 8 am to 6 pm.
2) The same types of the businesses located in the same building have the same open hours.
3) Businesses in malls, markets, food centers, and theme parks comply with the operational hours of these venues.
4) An open-hour classifier is developed based on the FSQ data and is applied to impute the open-hour patterns of the remaining establishments.

The first three steps of the procedure successfully determined the open hours of approximately 54,000 establishments out of nearly 146,000 records in total. To predict the operational hours of the remaining establishments, it is practical to focus on a group of general patterns, i.e., open at noon and afternoon as opposed to specific hours and minutes. One method is to discretize the open hours. To limit the number of combinations in the classification model, the temporal resolution is reduced to a 4-hour period (1 am to 5 am) and ten two-hour periods between 5 am to 1 am on day+1. This categorization is consistent with the classification of within-day time periods used in the activity learning models (see Section 5.3.2). At the same time, a filter is applied to retain only the open hour patterns that have an ignorable presence in the FSQ dataset (greater than or equal to 30 observations), which reduces the number of open hour classes to 43. Table 4-2 lists the most common open hour patterns found in the FSQ dataset.

Table 4-2. Ten most prevalent open hour patterns

Open hours	Time period encoding	Percentage
11am – 11 pm	00001111110	10.86%
11am – 9 pm	00001111100	9.74%
11 am – 3 pm & 5 pm – 9 pm	00001101100	5.83%
11 am –3 pm & 5 pm – 11 pm	00001101110	5.81%
5 pm – 11 pm	00000001110	5.30%
5 pm – 1 am	00000001111	5.17%
9 am – 9 pm	00011111100	5.11%
7 am – 9 pm	00111111100	4.74%
9 am – 11 pm	00011111110	3.89%
1 pm – 11 pm	00000111110	3.36%

Note: The open hours of venues are on Wednesdays.

This work used a supervised learning technique based on the Random Forest (RF) decision tree to learn the classification of open hour patterns. The advantage of Random Forest over other classification methods (i.e., SVM and Naive Bayes) is that it is able to address the complex (non-linear) relationships between classes and explanatory variables if insufficient instances are available for training. In addition, this approach performs reasonably well even if a large proportion of data are missing. Furthermore, this approach contains a mechanism to balance errors from different classes with imbalanced data (see additional discussion on this topic in Chapter 5).

In general, RF creates multiple classification and regression trees (CART), and each is trained on a bootstrap sample of the training data. CART is a type of decision tree that involves greedy and recursive partitioning of the instance space into disjoint areas. The CART tree grows by generating binary branches that split at the terminal nodes, which correspond to the conjunction of variables. The Gini measure of impurity is one of the most common criteria used to determine whether to split at a node. For a given node v and variable inputs X, the Gini measure takes the form of (Khalilia et al., 2011):

$$Gini(v, X) = \sum_{i=1}^{K} \frac{n_{x_i}}{N} I(v_{x_i}) \qquad (4\text{-}1)$$

$$I(v_{x_i}) = 1 - \sum_{c=0}^{C} \left(\frac{\tilde{n}_{x_i,c}}{n_{x_i}}\right)^2 \qquad (4\text{-}2)$$

where n_{x_i} is a subset of samples with values x_i, and $\tilde{n}_{x_i,c}$ is a subset of samples that have values x_i and belonging to class c. N is the total number of observations in the training dataset, and K is the total number of descendent at node *v*. After computing the Gini measure for all possible terminal nodes and variables, the split that maximally decreases the Gini measure is selected. The process repeats until all terminal nodes contain very few cases or the samples at the nodes all belong to same class.

During the training process, instead of including all explanatory variables, only a randomly selected subset of variables will enter each decision tree to avoid overfitting of model. At the stage of prediction, each CART will predict a class output for each set of new inputs. The output that obtains the majority vote from the outputs of all trees becomes the predicted class of the Random Forest model (Breiman, 2001).

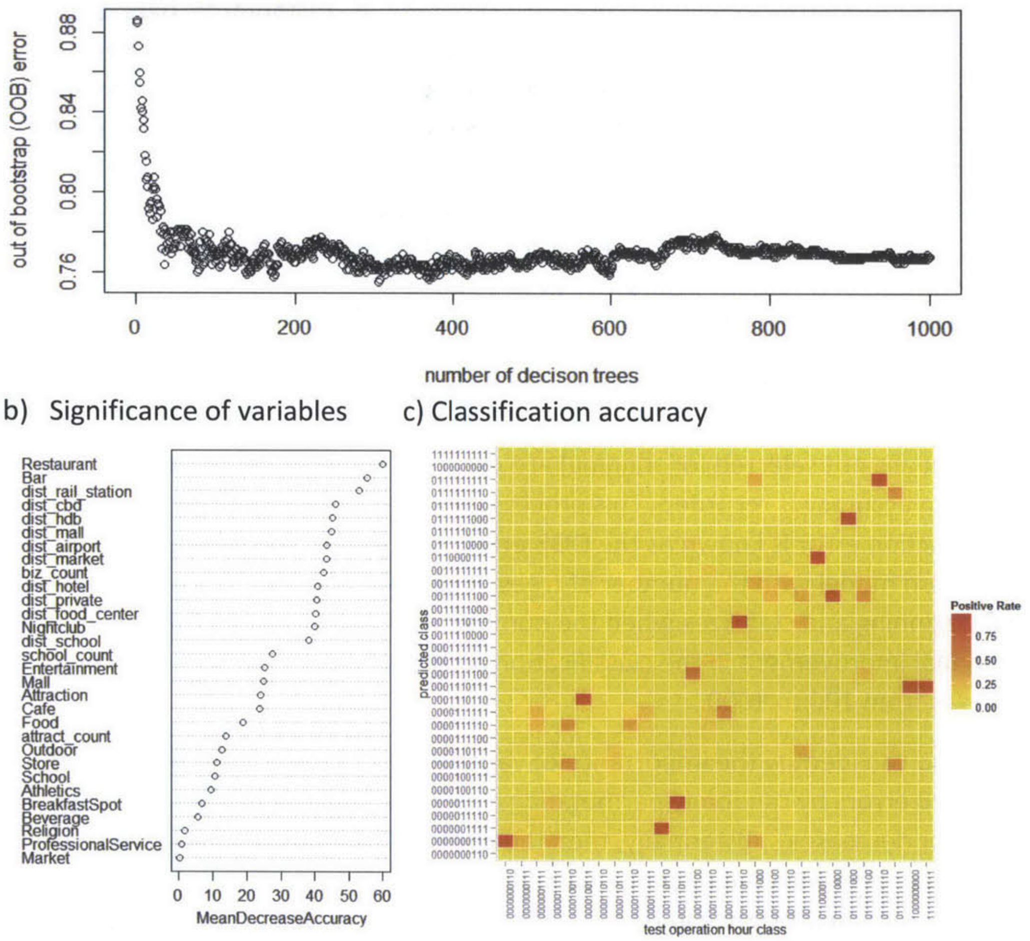
A RF approach implementation in the R package, randomForest (Liaw and Wiener, 2002), is used to classify the open hour patterns. This work incorporates 32 explanatory variables into the RF model, including 16 types of businesses (i.e., restaurants, bars, and cafes) and accessibility measures for businesses (i.e., distance to CBD, distance to the nearest mall, and distance to the nearest food center) as well as the number of different types of businesses within 50 meters of the given business location. The RF classifiers with different parameter configurations with the number of decision trees {500, 1000, 2000} and number of variables for each tree {4, 5, 6, 7} are randomly selected to decide the best variable for splitting at each node. The results of the classifier that uses 1000 trees and four variables was reported in this study because it showed good performance in terms of both prediction accuracy and time efficiency.

Because RF is a bootstrapping type of learning approach and only samples a subset of observations each time to train the decision trees, the out-of-bag (OOB) samples (observations that are not selected for training) are used as if they were novel test samples, and the prediction accuracy is recorded. This process is carried out for all OOB samples, and the average prediction error is known as the OOB error. Figure 4-2a) plots the OOB error of the RF classification model with the number of trees. It appears that the OOB error declines rapidly in the beginning and stabilizes when the number of trees increases to approximately 100. Only marginal improvement in the classification accuracy is observed when the number of trees increases from 100 to 1000.

The RF also can identify the importance of variables as measured by the average change in the OOB errors by randomly permutating a variable from the model. If the classification error decreases substantially, then the permuted variable is considered to have a strong association with the response variable. Figure 4-2b) lists the top 30 variables sorted by the estimated importance in terms of the average decreased accuracy. It appears that the dummy variable "Restaurant" is the most important variable in the RF model. In other words, restaurants have relatively more predictable open-hour patterns.

Figure 4-2c) plots the confusion matrix between the predicted responses and observed responses averaged over all OOB tests. The cells on the 45-degree diagonal line from the bottom left to the top right indicate the true positive rate, whereas the other cells show the false positive rate, which represents the proportion of observed classes that are incorrectly marked as other classes. Overall, the model has a true positive rate of 0.257, which means that approximately 25.7% of the predicted classes of observations match with the observed open hour classes. Considering the large number of classes that must be distinguished and the lack of evident and separable boundaries among classes, a model with this accuracy rate is acceptable. Taking a closer look at the plot of the confusion matrix in

Figure 4-2c), we observe that many misclassifications occur between similar open hour patterns. For example, a high proportion of businesses with the pattern 00000001100 (open between 5 pm and 9 pm) are classified as 00000001110 (open between 5 pm and 11 pm). Therefore, the classification accuracy can be further improved if the open hours are grouped into larger time intervals (e.g., every four hours) or additional predictors (i.e., explanatory variables) are included in the RF model.


a) Number of decision trees in the RF models and the corresponding OOB errors

b) Significance of variables

c) Classification accuracy

Figure 4-2. Results of the Random Forest (RF) model for open hour classification

4.4 Built Environment Measurement

4.4.1 Framework of the Spatial Metrics Computational Procedure

Urban built environment measurement generates indicators and variables that can be used to comparatively analyze urban forms such that better plans and policies can be enacted. For activity-travel research, these spatial metrics are important explanatory variables that can assist in revealing the effects of urban spaces on people's behaviors and choices. Traditionally, due to the limitation of data availability, only stationary spatial metrics are formulated. Conceivably, it is important to develop measures that can best describe the context in which each individual makes decisions or choices in daily life. In other words, for different activities, the built environment indicators should also differ because these activities occur under different spatio-temporal environments. However, much ambiguity exists in determining the spatial and temporal effects of built environment components. One approach to tackling this problem is to experiment with indicators with various spatial and temporal scales/scopes (e.g., various sizes of catchment areas). In addition, as new datasets become available and accumulate, built environment measurement will become a routine task for various analytical efforts. To cope with the need to generate different sets of spatial metrics, it is crucial to streamline the computing procedure for spatial metrics to reduce the amount of repetitive manual work.

Figure 4-3 shows the procedure for computing the urban built environment metrics in this study. At this point, a set of the geo-processing tasks, i.e., buffering of transit stations and the interactions between land use patches (or buildings) and the buffered areas, are accomplished in ArcGIS 10.1. The remainder of the indicator computation work is carried out by four routine R script programs, each generating a group of indicators that gauge one dimensionality of the built environment. For most indicators, the R scripts are also designed to automatically create summaries and plot distributions. The procedure is semi-automatic up to this point and requires inputs from users to specify the input locations and parameters, i.e., the time periods at which the urban built environment is measured.

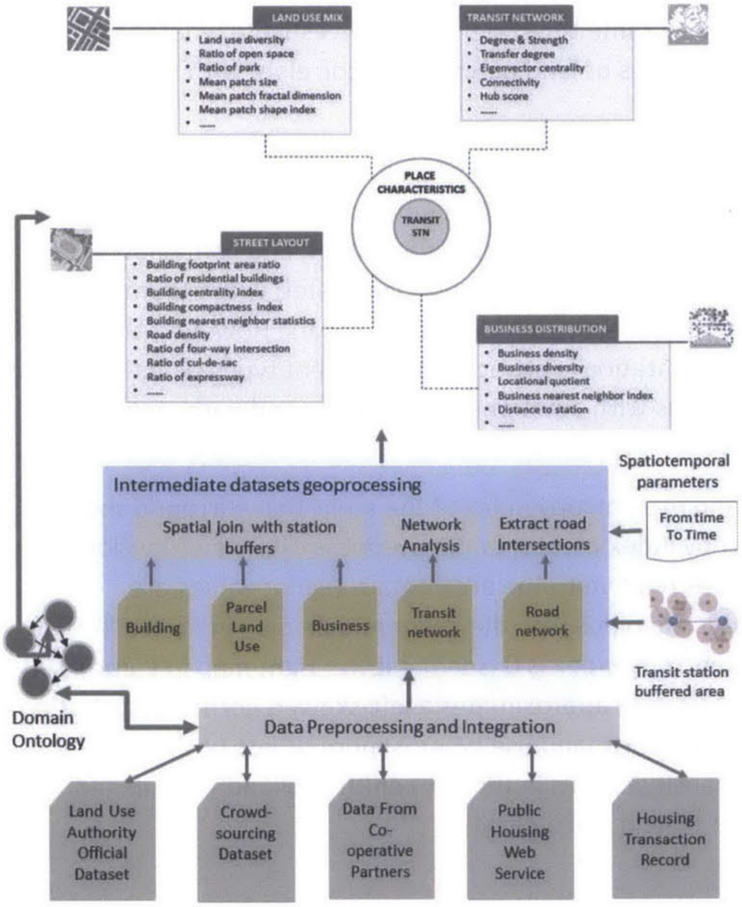


Figure 4-3. Framework of urban built environment metric computations

The urban built environment metrics formulated in this section are based on three aspects of consideration:

1) Preprocessed spatial data on land cover, buildings and establishments, and transit and road systems as a result of the data integration process described in Chapter 3.
2) A set of urban forms and built environment indicators identified by reviewing related work.
3) The indicators that may have an impact on people's daily activities and travels and can be derived from both the HITS dataset for model estimation and the EZ-Link dataset for activity inference.

The following sections describe four groups of indicators.

4.4.2 Land Use Mix

The measures developed in this section are primarily used to describe the land use composition in the catchment areas of transit stations. Land use composition refers to the mix of land patches for different uses in the area that may create synergy, which leads to improved attractiveness to the groups of people who might engage in particular activities. For example, a mix of commercial and residential land uses can promote short-distance shopping activities. Creation of large parking lots around a commercial strip will encourage more driving to shopping. Therefore, it is necessary to inspect whether the compositional relationships among different land uses result in increased attractiveness or increased repulsion in destination choices. At the same time, we want to investigate whether the extent of the mix and the shapes and sizes of land patches also correlate with the location choices of various activities.

Table 4-3 lists selected measures of land patches in the catchment areas of transit stations, including land use richness (RICHNESS), land use diversity, and area ratios of different types of land uses, i.e., open space/park area (OPR/PARKR) and contagion index (CONTAGI). The land use mix is commonly considered to be correlated with accumulative accessibility (Huang, 2006) and therefore leads to reduced use of cars. However, the type of mix and to what extent spatial fragmentation is attractive to different types of activities are ambiguous qualities. Presumably, areas with additional open spaces and parks will attract more sports and recreation activities.

Figure 4-4 shows the land use entropy index of the areas that surround the transit stations. The land use entropy index measures the evenness of the distribution of areas among different land use types relative to the areal totals. Higher values are representative of more diverse land use types (see the lower middle map in Figure 4-4). If only one land use type exists, the entropy index is 0 (see the lower right map in Figure 4-4), and the land use entropy index in this case approximates a left-skewed normal distribution, suggesting that the area distribution among land use types is more or less uneven. As expected, areas surrounding the transit stations in the city center, Holland Village, and Queenstown show a higher degree of land use diversity.

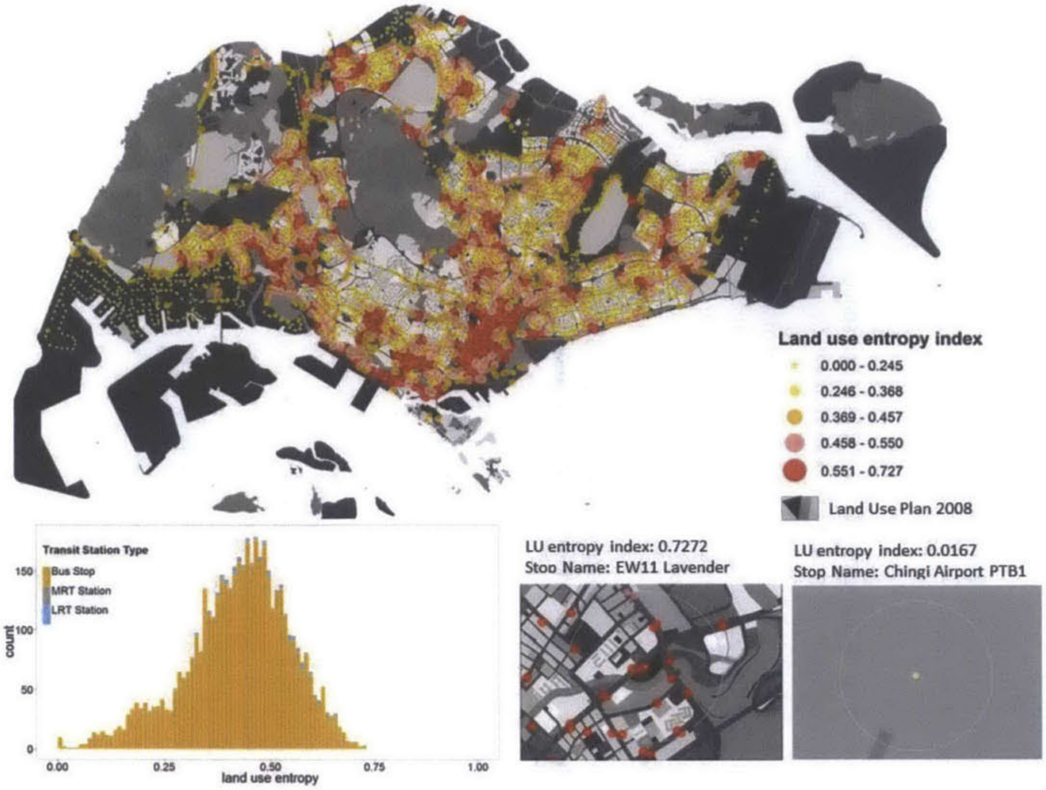


Figure 4-4. Spatial distribution of land use entropy index at the transit station catchment area level.

The contagion index measures the extent to which the land use types are separated in the catchment area. Higher values of contagion may result from landscapes with a few large and contiguous patches characterized by poor dispersion and interspersion of patch types. Figure 4-5 shows the distribution of the contagion index of the catchment areas and two examples corresponding to areas with high and low contagion indices. It is obvious that the land use entropy index is correlated with the contagion index because a large number of different land use types usually suggest that the land patches are more dispersed and separated. Figure 4-6 shows the catchment areas that have a notable presence of parks (the ratio of park area is greater than 33%). The station's proximity to parks may attract additional recreational and sporting activities.

Table 4-3. Spatial metrics for land use patches

Indicator	Formula	Range	Description	Citation
Part 1. LAND USE MIX (unit of analysis - land patches)				
Land use richness (LURICH)	n_l – Number of land use types considered	>1	Number of land use types considered in the catchment area.	Wu et al., 2002;
Mean patch size (MPS)				
Land use area ratio P_i	$P_i = A_i/A_{catchment}$ *- i is a land use type*	0-1	Area ratio of different land use types (open space, park, commercial, mixed use, waterbody etc).	Schwarz, 2010; Huang et al., 2007;
Land use diversity (LU_ENTROPY)	$LU\_ENTROPY = \sum_{i=1}^{n} P_i \ln(1/P_i)/\ln(n_l)$	>= 0	Measures the evenness of the distribution of areas among different land use types relative to zonal totals. Higher values are representative of more diverse land use types. 0 if only one land use type in the zone.	Krizek, 2003; Cervero & Kockelman 1997; Wu et al., 2002; Li and Yeh, 2004;
Area weighted mean shape index (AWMSI)	$AWMSI = \frac{\sum_{i=1}^{N} l_i/4\sqrt{A_i} \times p_i}{N_L}$ - l_i is the perimeter of patch *i*	>=1	Measures the regularity of land use types. It equals 1 for circular patch and increases as patch becomes more irregular.	Wu et al., 2002; Schwarz, 2010;
Area weighted mean patch fractal dimension (AWMFD)	$AWMFD = \sum_{i=1}^{N} 2 \times \ln(0.25 \times l_i)/\ln(A_i) \times p_i$	>=1	Measures the raggedness of patch boundaries. The fractal dimension approaches to 1 when shape has simple perimeters and approaches to 2 when shape becomes more jagged.	Huang et al, 2007; Wu et al., 2002; Longley and Mesev, 2000.
Contagion index (CONTAGI)	$CONTAG = 1 + \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left(P_i \frac{g_{ij}}{\sum_{j=1}^{n} g_{ij}}\right) \times \left(\ln(P_i) \frac{g_{ij}}{\sum_{j=1}^{n} g_{ij}}\right)\right]}{2\ln n}$ g_{ij} is the number of adjacencies between land use i and j.	0-1	Measures the extent to which the land use types are separated in the catchment area. Higher values of contagion may result from landscapes with a few large, contiguous patches characterized by poor dispersion and interspersion of patch types.	Wu et al., 2002; Herold et al., 2002;

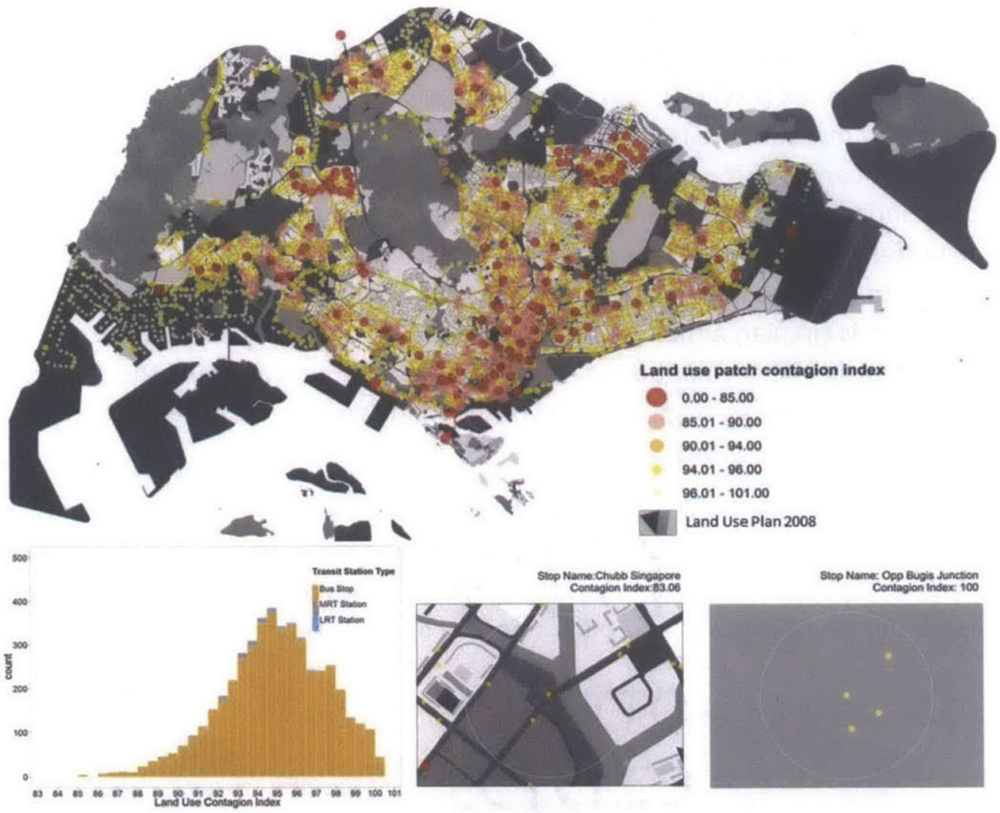


Figure 4-5. Spatial distribution of the Contagion Index

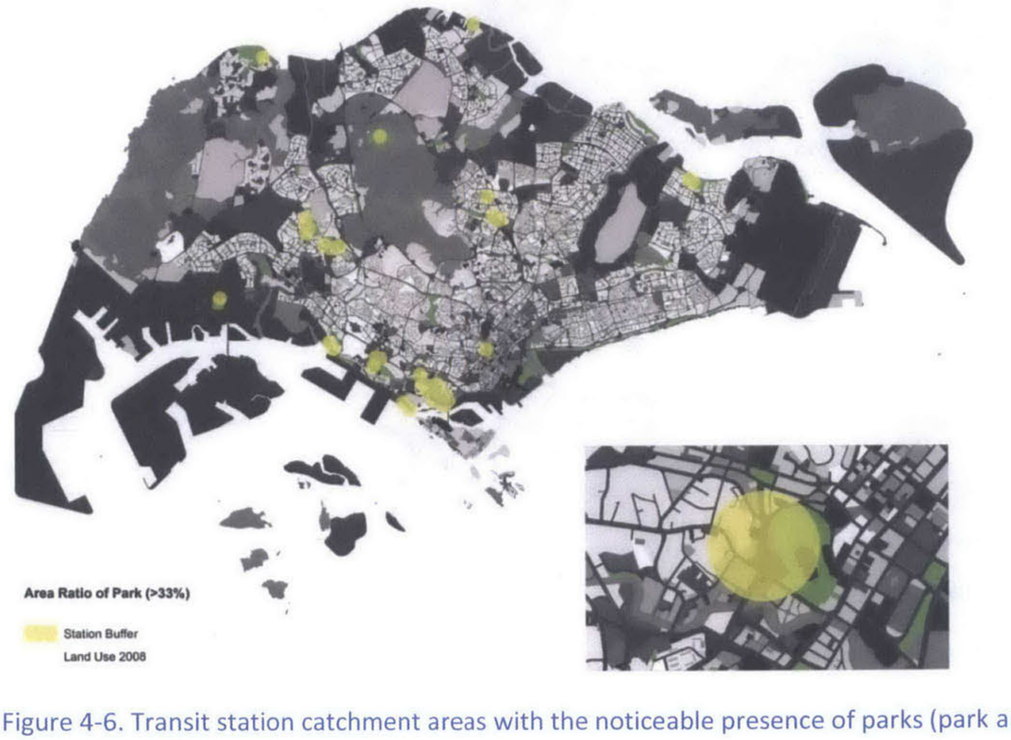


Figure 4-6. Transit station catchment areas with the noticeable presence of parks (park area ratio greater than 33%).

In addition to the land use mix of the patches, their shapes could also matter. Spatial indicators that include the average patch size (APS), landscape shape index (LSI), area weighted mean shape index (AWMSI), and area weighted mean shape fractal dimension (AWMFD) are used to measure the regularity and raggedness of patch shapes. It is worth noting that the unit of analysis for computing these indicators is the intersected pieces of land use patches and transit station buffer areas as opposed to the original land use patches. If the buffer area of a transit station lies within a large land patch, then its AWMSI value will be equal to 1, which means that the shape of the patch within the catchment area of that transit station is circular, although the shape of the real patch could be quite irregular. In other words, the circular boundaries of the buffer areas can make the shapes of lands look more regular and less ragged, which consequently biases shape measures such as AWMSI and AWMFD to lower values.

For example, Figure 4-7 plots the AWMSI values of the station catchment areas. As the values approach 1, the shapes of the land patches become more circular and regular on average. Otherwise, because the values are larger, the shapes of patches become more irregular. Overall, the AWMSI values approach a normal distribution.

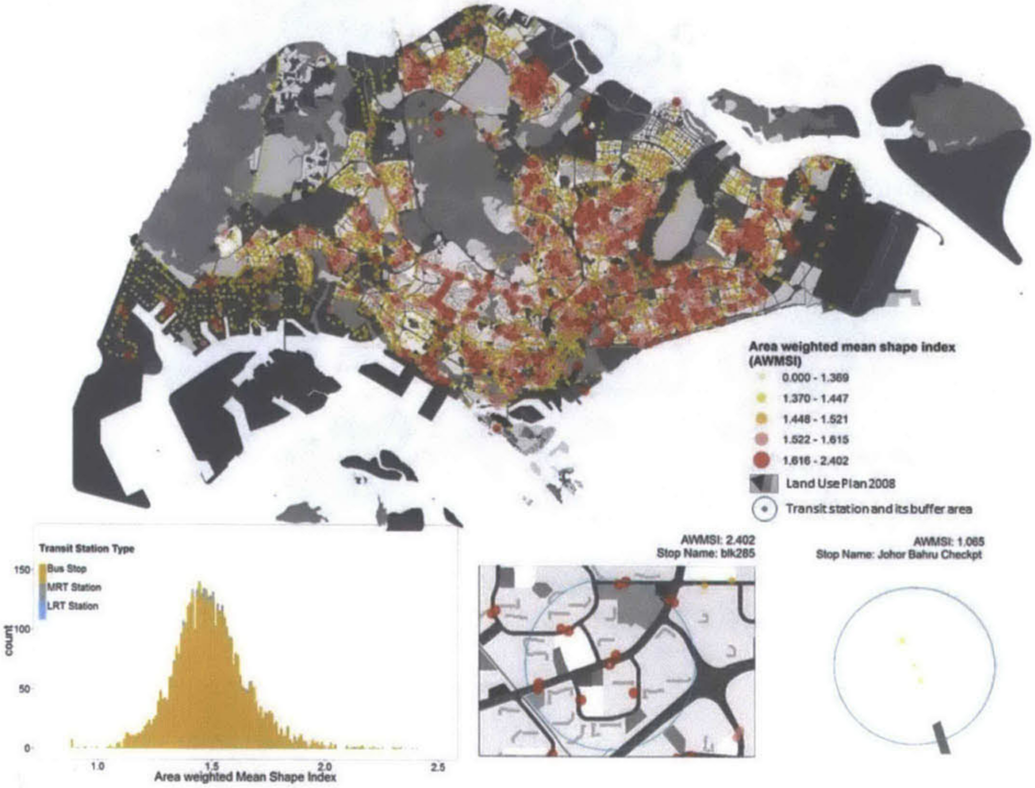


Figure 4-7. Spatial distribution of the average weighted mean shape index (AWMSI)

4.4.3 Buildings and Street Layout

Street layout and streetscapes are critically related to the mobility and walkability of places. Better connectivity of local road networks facilitates additional walking and bicycling activities. In contrast, cul-de-sacs are less desirable from the point of view of pedestrian.

Expressways usually block the continuity of landscapes and fragment the local neighborhood. Drawing on the data on building footprint geometries and road networks provided by SLA together with the integrated attributes from other data sources, spatial metrics on the spatial distribution of buildings and street layouts are intended to evaluate the local street landscape at a more detailed level. The density, compactness, and street circulation of the local built-up areas around transit stations are of particular interest. In recent years, the urban form of Singapore has been subjected to the combined efforts of redevelopment for higher densities and transit-oriented new development on vacant lands. Transit-oriented development requires easy pedestrian access at the areas near transit stations such that people can access the places where they live, work, shop, and conduct other activities via public transportation and walking. Empirical studies report that a compact urban form with well-connected road system encourages walking and bicycling trips (Benfield et al., 1999; Talen, 2002) and thus aids in improving the prosperity of the area by providing additional opportunity for social interaction for people and more intervening opportunities for local businesses.

Table 4-4 lists the indicators selected to measure the building layout and street connectivity in local areas of transit stations. The compactness of buildings is estimated based on the comparison between the perimeter of the footprint of each building and the corresponding perimeter of a circle that has the same area (Li and Yeh, 2004). The lower the compactness index (COMPACT), the greater the compactness of local urban form will appear from the point of view of existing buildings. On the contrary, the more regular the shape of buildings and the smaller the building numbers, the higher the compact index value. As shown in Figure 4-8, the compact urban forms in Singapore are mostly located at the central and eastern areas such as city center, little India, Geylang and Hougang.

The centrality measures the proximity of other buildings relative to the building with the largest footprint size in the catchment area and aids in characterizing how other buildings are spatially distributed around a node building such as a mall, MRT station, or local stadium. The more elongated the distribution pattern, the higher the centrality index (Huang et al., 2007). As shown in the Figure 4-9, the catchment area with a single large building (e.g. airport terminal) tends to have high centrality value while the catchment areas with many like-sized buildings have lower centrality values.

Table 4-4. Spatial metrics for buildings and street layout measurement

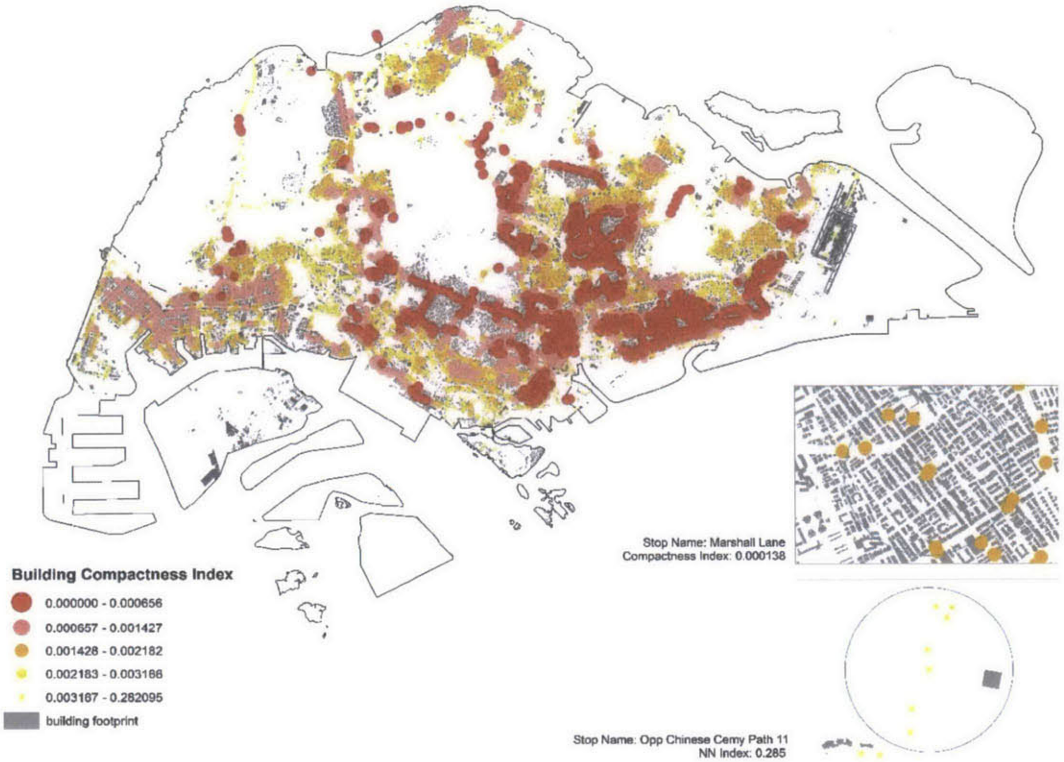
Part 2. Spatial Distribution of Buildings (unit of analysis – building footprints)				
Building density (BLDDen)	$BLDDen = N_{bld}/A_{developed}$ - N_{bld} is the number of buildings	>=0	The number of buildings per developed unit area.	
Building footprint size ratio (BLD_FP_RATIO)	$BLD\_FD\_RATIO = \sum_{i=1}^{N} FS_i \,/A_{developed}$ - FSi is the footprint size of building i	0-1	The ratio of total building footprint area to the size of the developed lands in the catchment area of transit stations	
Largest building index (LBI)	$LBI = \max(FS_i)\,/A_{developed}$	0-1	The ratio of the area of the largest building (path) to the size of the catchment area.	Wu et al., 2002;
Mean building footprint size (MBFS)	$MBFS = \sum_{i=1}^{N_b} FS_i \,/N_{bld}$	>=0	Average building footprint size.	
Normalized building footprint size standard deviation (BFSSD)	$BFSSD = \frac{\sqrt{\sum_{i=1}^{N_b}(FS_i - MBFS)^2}}{MBFS}$	>=0	Higher values means greater variation in building footprint size	

Compactness Index (COMPACT)	$COMPACT = \frac{\sum_j 2\sqrt{(FS_j / \pi)} / l_j}{{N_{bld}}^2}$	0-1	Measures the compactness of buildings in the catchment area. The more regular the shape of buildings and the smaller the building numbers, the higher the compact index value.	Li and Yeh, 2004;
				Herold et al., 2002;
				Schwarz, 2010;

Centrality Index (CENTRALITY)	$CENTRALITY = \frac{\sum_{i=1}^{N-1} D_i / (N-1)}{\sqrt{A / \pi}}$
	- D_i is the distance of building i to the largest building in the zone.
		0-1	Measures the closeness (average distance) of other buildings to the building with the biggest footprint size.	Li and Yeh., 2004;
				Herold et al., 2002;
				Schwarz, 2010;

Nearest Neighbor Statistics (NN_{bld})	$NN_{bld} = \frac{\sum_j d_j}{n} / 2\sqrt{n / A}$ - d_j is the distance of building j to the nearest building	>=0	Measures the spatial distribution of building footprint polygons. A clustering tendency has values approaching 0. 1 means completely random distribution and perfect uniformity corresponds a theoretical value of 2.15.	Mesev, 2005;
HDB units density (HDB_UNIT_DEN)	$\mathrm{HDB\_UNIT\_DEN} = N_{HDB\_unit}/ A_{catchment}$	>=0	A proxy for measuring public housing population density.	

Part 3. Walkability (unit of analysis: road segments and intersections)				
Intersection density (INT_DEN)	INT_DEN=Number of interactions / road length	>=0		
Ratio of four-way intersections (INTER4_RATIO)	Inter4_R $= N_{Inter_{4way}}/(N_{Inter_{4way}} + N_{Inter\_3way})$	0-1	Ratio of four-way intersection to the total number of four-way and three-way intersections.	Song & Knaap, 2004;
Ratio of Expressway (EXPRESS_RATIO)	$EXPRESS\_R = L_{CAT(A,B)}/L)$	0-1	Ratio of the length of Category A (Expressway) and Category B (Highway and Arterial) to the total length of road network in the area.	
Ratio of Pedestrian Mall (PED_MALL_RATIO)	$PED\_MALL\_R = L_{Ped\_mall}/L$	0-1	Ratio of the length of pedestrian mall to the total length of road network in area.	
Ratio of Cul-de-sac (CUL_DE_SAC_RATIO)	$CUL\_DE\_SAC\_R = L_{cul\_de\_sac}/L$	0-1	Ratio of the length of cul-de-sacs to the total length of road network in area.	Song & Knaap, 2004;



313

Figure 4-8. Building compactness index of the catchment areas of transit stations (the compactness index is lower in highly compact areas as shown by points in red)

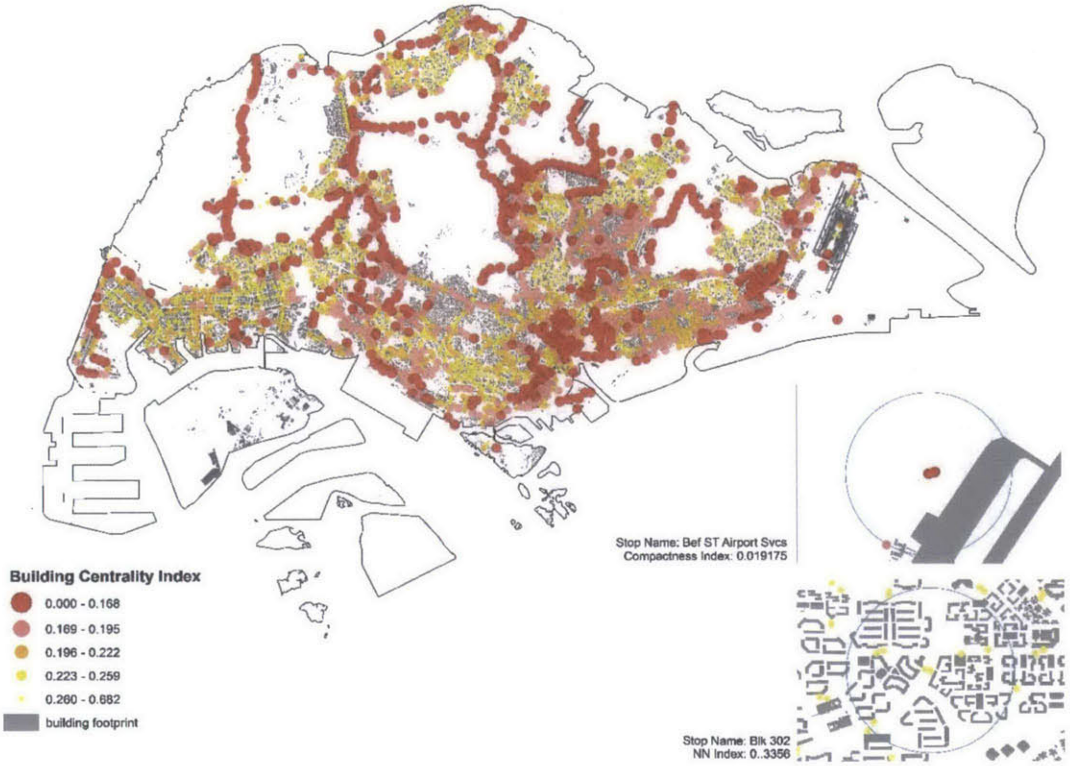


Figure 4-9. Building centrality index of the catchment areas of transit stations (the centrality index is lower in areas with more centrally distributed buildings as shown by points in red)

The nearest neighbor (NN_BLD) index measures the spatial distribution of buildings relative to the random spatial distribution. According to Mesev (2005), the nearest-neighbor indices compare the observed average distance between neighboring building centroids to the expected neighboring distance among building centroids under the random distribution condition. If the nearest-neighbor index approaches 0, the tendency of spatial clustering becomes stronger. If the value of the nearest-neighbor index approaches 1, the locations of buildings show a random distribution pattern. If the value approaches 2.15, then the spatial distribution is prone to be uniform. The three small maps in Figure 4-10 show situations in which buildings are clustered, randomly distributed, and approach to a uniform distribution.

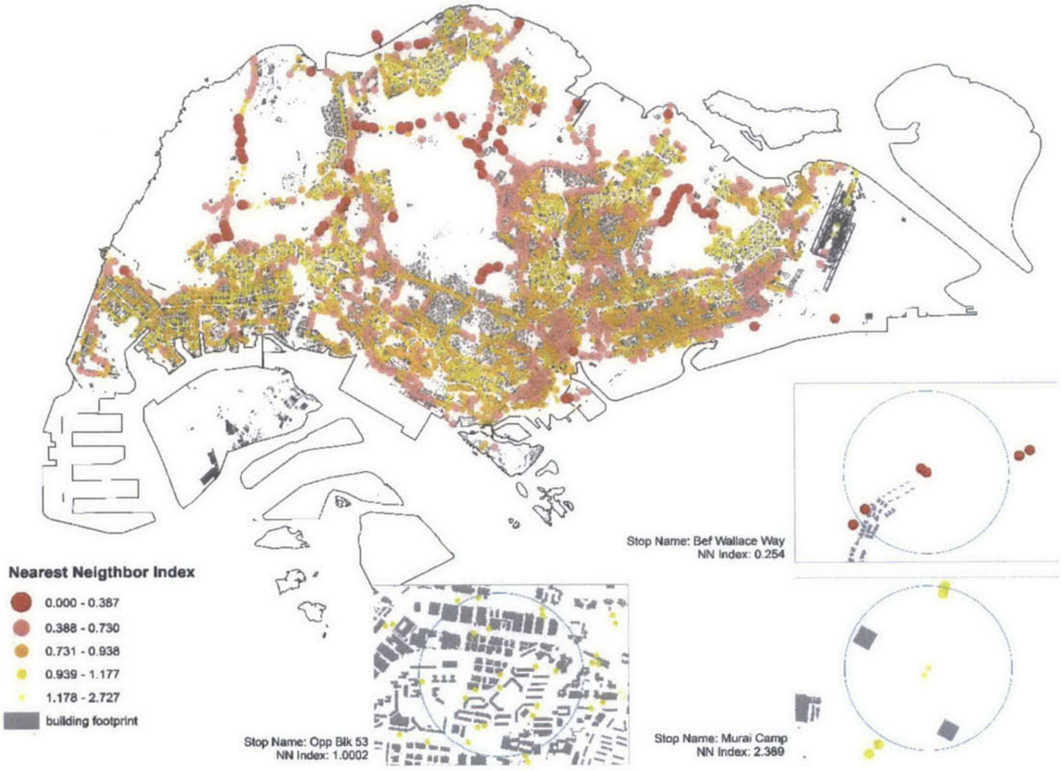


Figure 4-10. The nearest neighbor index of the catchment areas of transit stations

The density of intersections, especially four-way intersections, has been used widely as an important indicator for street connectivity and walkability. A greater number of intersections on a given length of street suggest shorter street blocks and better street circulation, which will potentially offer additional attractions for transit riders who are more likely to explore the destination area by foot. Similarly, a high density of local roads (class A and B roads) imply good walkability and connectivity for pedestrians (Figure 4-11). In contrast, areas characterized by street networks with a high proportion of expressways and cul-de-sacs tend to attract additional car users.

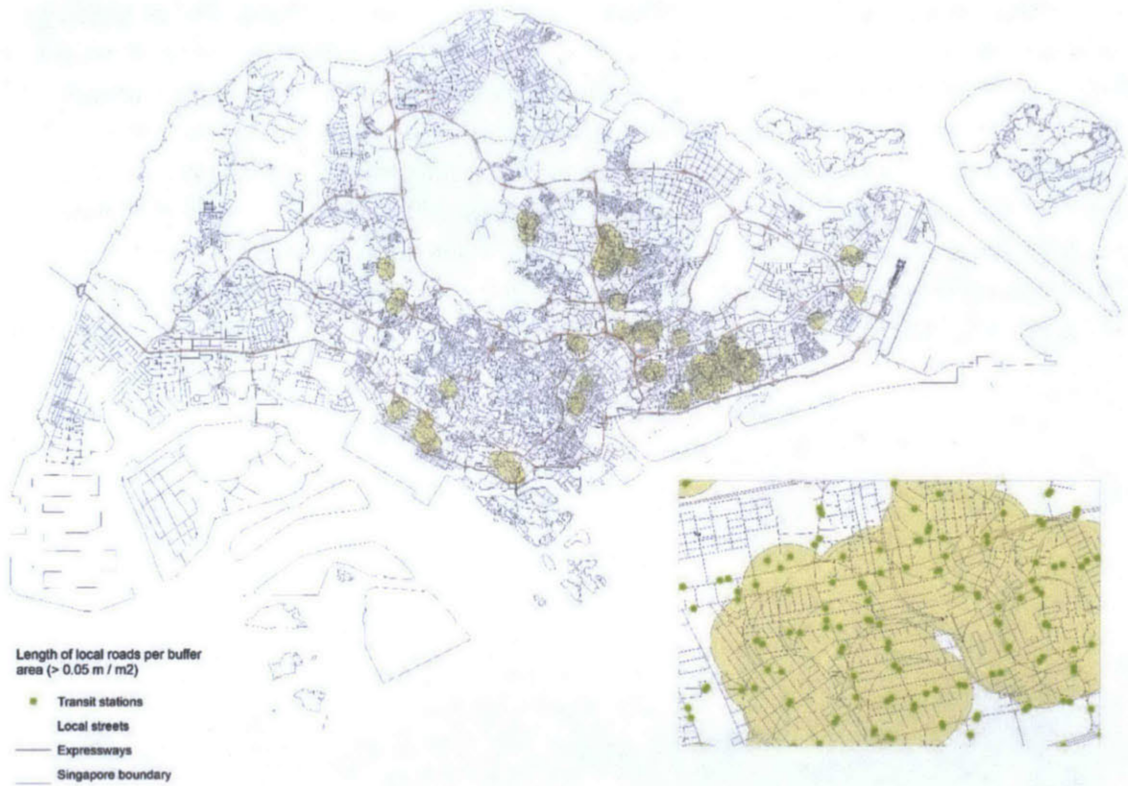


Figure 4-11. Density of Local Road (> 0.05 m per square meter)

4.4.4 Spatial distribution of Establishments

The magnitude of services or opportunities is usually measured using the number of jobs by economic sectors. Because it is difficult to obtain employment data at disaggregated levels as buildings or establishments, the information of establishments are used accordingly to gauge the spatial distribution of opportunities for different types of activities. Usually, when an individual makes destination choices for an activity, both the amounts and the mixture of the relevant types of establishments can matter. As argued by Fotheringham et al. (2001), spatial location choice involves a hierarchical process in which a large destination area is first selected and the alternatives within the area are subsequently evaluated. For example, for a dining trip, travelers may first pick a food center that contains clusters of restaurants and subsequently choose a particular restaurant. In this sense, all establishments within the catchment areas of the transit stations constitute the choice set of activity destinations. The size of the choice set is also relevant.

Hence, the indicators constructed in this section are intended to measure the levels of spatial agglomeration of establishments of individual sector as well as the degree of mixture of various sectors of establishments. The sectorial categorization of establishments is built based on the two-digit Singapore Standard Industrial Classification 2010 (SSIC) categories (Appendix II). In reality, the availability of commercial opportunities does not only vary spatially but also temporally. The temporal dimension is captured in the measurement using the open hours of establishments as constraints on availability.

To consider the agglomeration effect of businesses, the locational quotient is used to measure the sectorial specification of establishments. The measure is based on a sectorial share of the number of establishments within the catchment area relative to its share in all of Singapore. The mixture of businesses is measured once again using Shannon's entropy index. If all types of establishments are equally common in a catchment area, then the Shannon index takes the value *ln (number of establishment types)*. The more unequal the quantities of the types, the larger the weighted geometric mean values will be, and the smaller the corresponding Shannon entropy. If practically all establishments tend to belong to the same type, the Shannon entropy approaches zero.

In addition to the concentration and diversity of establishments, the spatial distribution of business points is also important to consider. Whether establishments are all clustered in a single building (i.e., a mall) or spread along the streets might imply different activity patterns. Additionally, the distances of the businesses from transit stations might suggest different development patterns of urban form. Similar to the spatial distribution of buildings, the nearest-neighbor indices are used to distinguish different spatial distribution patterns, from clustering to random distribution to diffusion.

Table 4-5 lists the measures of the distribution of establishments in the catchment areas of the transit stations considered in this study. In addition to the spatial variability, we examine the effect of the temporal variability of in-operation businesses on the activity-travel patterns. The open status of establishments is obtained by learning from the FSQ samples, as described in Section 4.3.2. Figure 4-12 displays the geographical distribution of establishments according to our categorization based on the Singapore Standard Industrial Classification (SSIC) 2010 (Appendix II). The map on the left plots all establishments, and the maps on the right show the open establishments during three different time periods of a weekday. In the early morning (5 am – 7 am), most available services belong to the category of food and beverage service activities (red dots on the maps). This observation is in line with the reality that restaurants and cafes that provide breakfast always open early. At noon (11 am – 1 pm), most establishments involved in retail and wholesale activities (blue dots on the maps) are open. In the late evening (11 pm -1 am), the establishments that are still open are mostly restaurants or businesses in the category of creative, art, and entertainment activities (orange dots).

Figure 4-13 zooms into the catchment areas of three transit stations located in different regions of Singapore: Former Tang Village, Changi Airport, and Parkway Hotel. In addition, the entropy index and the nearest neighbor index of the establishments are labeled on the map in the catchment areas of three selected stations at different times of day. The indices vary both spatially and temporally. For example, in the catchment area of the Parkway Hotel bus stop, the entropy index changes from 0 in early morning (5 am – 7 am) to 0.839 at noon (11 am – 1 pm) and 0.793 at night (9 pm – 11 pm), indicating that the degree of business mix of the local area begins with zero in the morning when only food

Table 4-5. Measures of spatial distribution of businesses

Indicator	Formula	Range	Description	Citation
Unit of Analysis: Establishments				
Business Density (BIZ_DEN)	$BIZ\_Den_E = N_E/A_{developed}$ - N_E is number of establishments in the developed lands in the catchment area	>=0	Average number of establishments per square kilometer	Cervero and Kockelman, 1997;
Normalized Mean distance to station (BIZ_AVG_DIST)	$BIZ\_AVG\_Dist_E = \sum_{i=1}^{N_E}(d_{Ei}/d_{Buff})/N_E$ - d_{Ei} is the distance of the establishment *i* to the corresponding transit station; d_{Buff} is the radius of the catchment area of the corresponding transit station;	0-1	Normalized average distance of (all, food, retail) businesses to the nearest transit station	
Normalized Distance standard deviation (BIZ_DIST_SD)	$BIZ\_DIST\_SD_E = \sqrt{\frac{1}{N_E}\sum_{i=1}^{N_E}(d_{Ei} - Dist_E)^2}/d_{buff}$ - $Dist_E$ is the mean distance to station	>0	Normalized standard deviation of the distance of (all, food, retail) businesses to the nearest transit station	
Establishment Diversity (BIZ_ENTROPY)	$EI_E = \sum_{i=1}^{n} p_{Ei}\log(1/p_{Ei})/\log(n_E)$ $p_{Ei} = N_{Ei}/N_E$	>= 0	Higher values are representative of more diverse business types.	Kumar et al, 2007;
Specialization index (LOC_QUO)	$LOC\_QUO_E = \frac{N_{Ei}/N_E}{NE_i/NE}$ - *NE_i* and *NE* are the total number of establishments of type *i* and the total number establishments in Singapore	>=0	Measure the concentration of a particular type of establishment in the area. Higher values indicate agglomeration economy?	de Bok & van Oort, 2011; Garcia-López, & Muñiz, 2013;
Competition Index (COMPI)	$COMP_E = \ln(\frac{1}{Herf_i}), Herf_i = N_{Ei} \times \left(\frac{1}{NE_i}\right)^2$	>=0	Measure the degree of competition each establishment in sector i faces in a catchment area.	Martin et al, 2011
Nearest Neighbor Index (NN_BIZ)	$NN_{bld} = \frac{\sum_j d_j}{n}/2\sqrt{n/A}$ - d_j is the distance of establishment j to the nearest establishment	>=0	Clustering when NN<1.0 and dispersion when NN>1.0.	Mesev, 2005

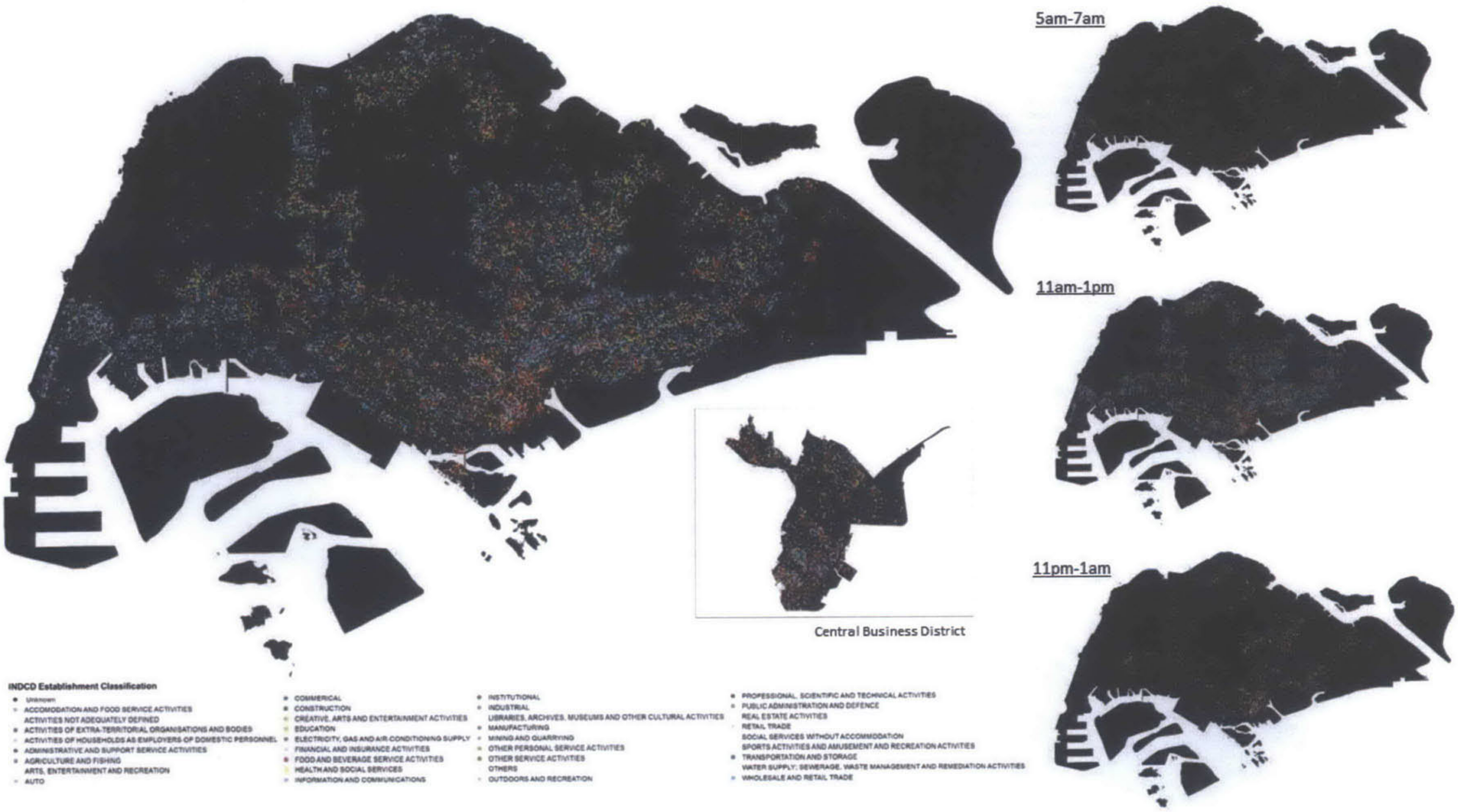

5am-7am
11am-1pm
11pm-1am
Central Business District
INDCD Establishment Classification
Unknown
ACCOMODATION AND FOOD SERVICE ACTIVITIES
ACTIVITIES NOT ADEQUATELY DEFINED
ACTIVITIES OF EXTRA-TERRITORIAL ORGANISATIONS AND BODIES
ACTIVITIES OF HOUSEHOLDS AS EMPLOYERS OF DOMESTIC PERSONNEL
ADMINISTRATIVE AND SUPPORT SERVICE ACTIVITIES
AGRICULTURE AND FISHING
ARTS, ENTERTAINMENT AND RECREATION
AUTO
COMMERICAL
CONSTRUCTION
CREATIVE, ARTS AND ENTERTAINMENT ACTIVITIES
EDUCATION
ELECTRICITY, GAS AND AIR-CONDITIONING SUPPLY
FINANCIAL AND INSURANCE ACTIVITIES
FOOD AND BEVERAGE SERVICE ACTIVITIES
HEALTH AND SOCIAL SERVICES
INFORMATION AND COMMUNICATIONS
INSTITUTIONAL
INDUSTRIAL
LIBRARIES, ARCHIVES, MUSEUMS AND OTHER CULTURAL ACTIVITIES
MANUFACTURING
MINING AND QUARRYING
OTHER PERSONAL SERVICE ACTIVITIES
OTHER SERVICE ACTIVITIES
OTHERS
OUTDOORS AND RECREATION
PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES
PUBLIC ADMINISTRATION AND DEFENCE
REAL ESTATE ACTIVITIES
RETAIL TRADE
SOCIAL SERVICES WITHOUT ACCOMMODATION
SPORTS ACTIVITIES AND AMUSEMENT AND RECREATION ACTIVITIES
TRANSPORTATION AND STORAGE
WATER SUPPLY; SEWERAGE, WASTE MANAGEMENT AND REMEDIATION ACTIVITIES
WHOLESALE AND RETAIL TRADE

Figure 4-12. Spatial distribution of establishments by INDCD categories

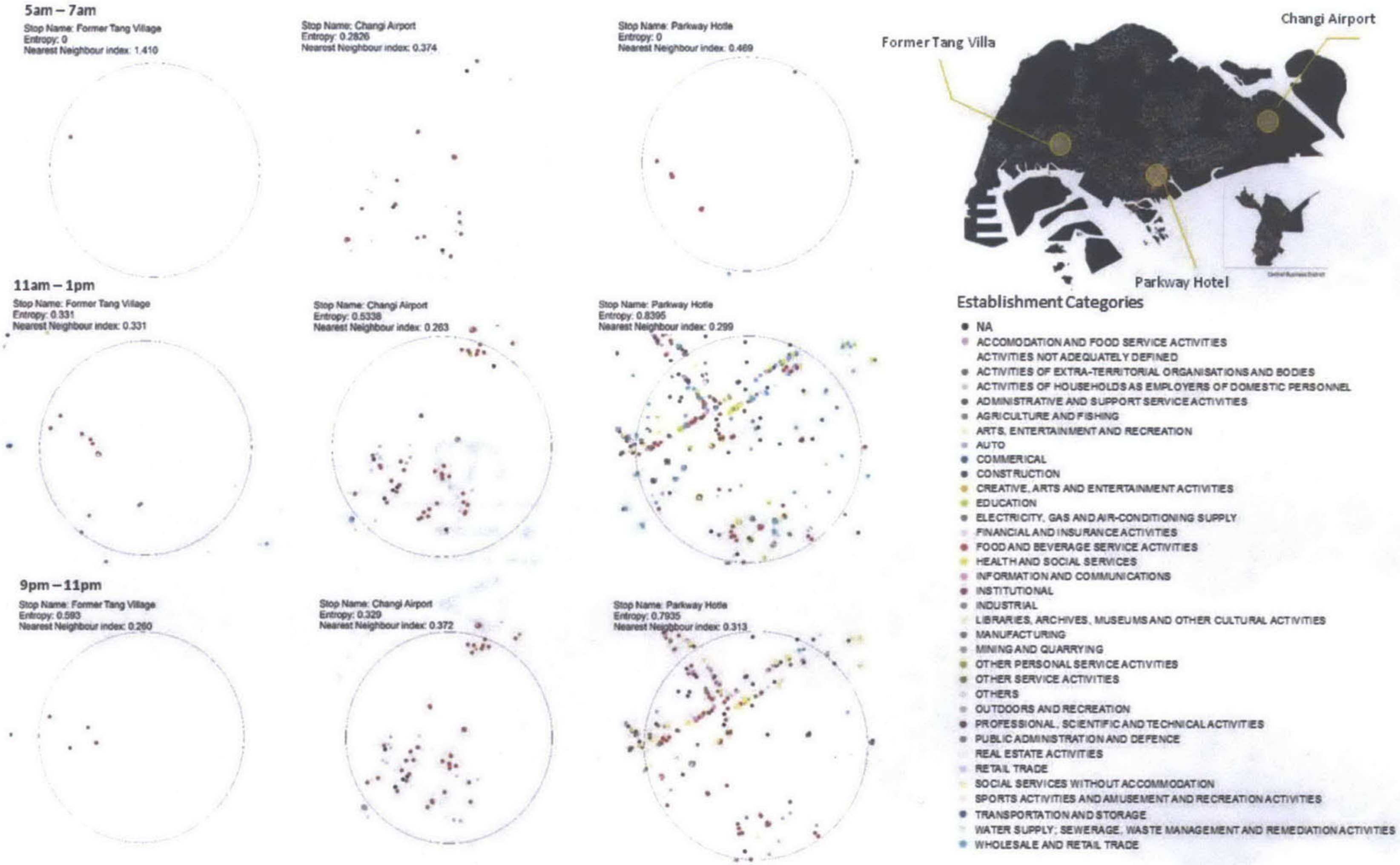

5am – 7am
Stop Name: Former Tang Village
Entropy: 0
Nearest Neighbour index: 1.410
Stop Name: Changi Airport
Entropy: 0.2826
Nearest Neighbour index: 0.374
Stop Name: Parkway Hotle
Entropy: 0
Nearest Neighbour index: 0.469
Former Tang Villa
Changi Airport
Parkway Hotel
11am – 1pm
Stop Name: Former Tang Village
Entropy: 0.331
Nearest Neighbour index: 0.331
Stop Name: Changi Airport
Entropy: 0.5338
Nearest Neighbour index: 0.263
Stop Name: Parkway Hotle
Entropy: 0.8395
Nearest Neighbour index: 0.299
9pm – 11pm
Stop Name: Former Tang Village
Entropy: 0.593
Nearest Neighbour index: 0.260
Stop Name: Changi Airport
Entropy: 0.329
Nearest Neighbour index: 0.372
Stop Name: Parkway Hotle
Entropy: 0.7935
Nearest Neighbour index: 0.313
Establishment Categories
● NA
● ACCOMODATION AND FOOD SERVICE ACTIVITIES
ACTIVITIES NOT ADEQUATELY DEFINED
● ACTIVITIES OF EXTRA-TERRITORIAL ORGANISATIONS AND BODIES
● ACTIVITIES OF HOUSEHOLDS AS EMPLOYERS OF DOMESTIC PERSONNEL
● ADMINISTRATIVE AND SUPPORT SERVICE ACTIVITIES
● AGRICULTURE AND FISHING
● ARTS, ENTERTAINMENT AND RECREATION
● AUTO
● COMMERICAL
● CONSTRUCTION
● CREATIVE, ARTS AND ENTERTAINMENT ACTIVITIES
● EDUCATION
● ELECTRICITY, GAS AND AIR-CONDITIONING SUPPLY
● FINANCIAL AND INSURANCE ACTIVITIES
● FOOD AND BEVERAGE SERVICE ACTIVITIES
● HEALTH AND SOCIAL SERVICES
● INFORMATION AND COMMUNICATIONS
● INSTITUTIONAL
● INDUSTRIAL
● LIBRARIES, ARCHIVES, MUSEUMS AND OTHER CULTURAL ACTIVITIES
● MANUFACTURING
● MINING AND QUARRYING
● OTHER PERSONAL SERVICE ACTIVITIES
● OTHER SERVICE ACTIVITIES
● OTHERS
● OUTDOORS AND RECREATION
● PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES
● PUBLIC ADMINISTRATION AND DEFENCE
● REAL ESTATE ACTIVITIES
● RETAIL TRADE
● SOCIAL SERVICES WITHOUT ACCOMMODATION
● SPORTS ACTIVITIES AND AMUSEMENT AND RECREATION ACTIVITIES
● TRANSPORTATION AND STORAGE
● WATER SUPPLY; SEWERAGE, WASTE MANAGEMENT AND REMEDIATION ACTIVITIES
● WHOLESALE AND RETAIL TRADE

Figure 4-13. Time varying entropy index and the nearest neighbor index of the establishments in the catchment areas of three selected stations.

services are available, reaches a maximum point at noon, and decreases at night because businesses such as retail are closed. In parallel, the nearest neighbor index changes from 0.469 to 0.299, and again to 0.313, which suggests the extent of spatial clustering of businesses increases from morning to noon and decreases from noon to night.

4.4.5 Transit Network Topology

For urban residents who rely on the transit system for their daily trips, knowledge of transit network not only helps them to decide on the destinations for accessible activities but also assists in scheduling activity chains during a day by taking into consideration the destinations that can be conveniently connected by transit services. Unlike air transportation networks or social networks in which the links between nodes represent direct connections, the urban transit network contains two levels of network, i.e., the underlying infrastructure network and the operation network served by buses and trains. Accordingly, the concept of the spaces ***L*** and ***P*** proposed in (Guimera, et al., 2005) was used in a series of analyses that focused on the public transit network (Sen, et al., 2003). Space ***L*** represents the physical network of transit system composed of rails and roads. The transit network is typically formulated as space ***P*** in which the edge between two nodes indicates a bus or a train route that serves the two nodes (Figure 4-14). In this sense, the degree ***K*** in the urban transit network corresponds to the number of stations directly reachable by a single bus or train route without transfer. In this study, we use the transit schedule information from 2013 provided by LTA to measure the transit network topology of Singapore since the corresponding data from 2008 were not available. Most urban residents plan for trips according to the transit schedule information. Therefore, the measures of the transit system derived from the schedules better reflect the transportation knowledge of the people who use the system.

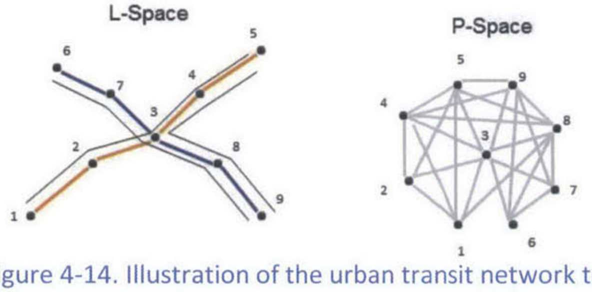


Figure 4-14. Illustration of the urban transit network topology representation

4.4.3.1 Temporal variability of transit service

On the supply side, the frequency and coverage of urban transit services usually vary by time of day. Typically, services are more frequent at peak hours but less frequent at night. The frequency of service is directly related to the waiting time for transit trips and may influence the transportation mode and destination choices of an activity. According to the transit schedule data in 2013, Figure 4-15 shows the maps of the frequency of transit services for road or railroad segments at different times of the day. It can be observed that many bus and RTS services are in operation early in the morning. At the morning peak hour, the frequency of services reaches the maximum for the day. In addition, it appears that certain long-distance services connect the northwestern region of Singapore to the downtown area only during the morning peak hour. The levels of service are quite stable during the day between 11 am to 7 pm, and frequency begins to decrease in general after 7 pm. Only selected night services are available (e.g., from downtown to the airport) that run between 1 am and 5 am. A lack of empirical evidence exists on how the temporal variability of transit service affects activity, but to examine the effect of transit network topology on activity, the temporal variation of service is accounted for in the transit network measurement in this section.

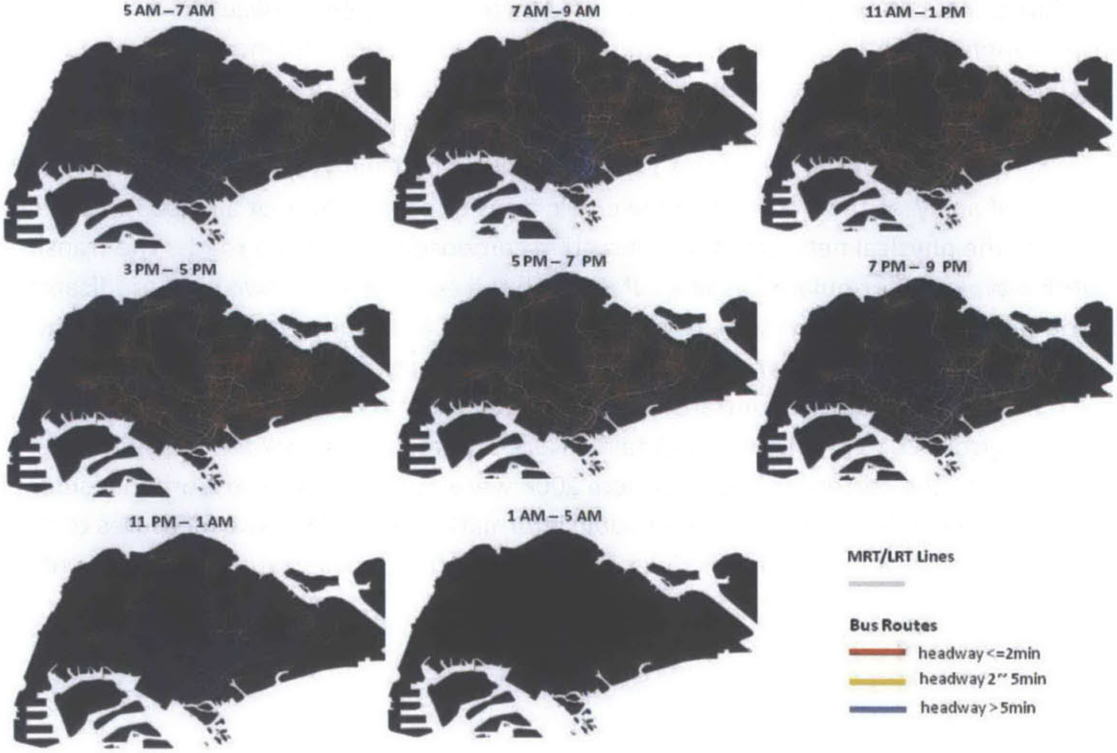


Figure 4-15. Frequency of transit services by road segments at different times of day.

4.4.3.2 Transit Transfer

Most early research on network analyses focused on the properties of network components, such as nodes and edges formed by a single type of connection, i.e., airlines or co-author relationships. This situation gives rise to two issues if the methods are applied to study urban transit systems. First, the urban transit network in large cities usually contains both bus networks and rail networks. The two different networks are treated separately in most of the network analysis literature. However, the transit system consists of these two networks that have an inseparable relationship in transportation planning and operation. From the view of travel demand, it is impractical to presume that transit passengers consider only the bus or only RTS when planning trips. Therefore, it is essential to combine the bus and rail networks when evaluating the transit network. Second, unlike airports, which are usually distantly located and relatively independent (airports in the same city are usually clustered together in the network analysis), urban transit stations have strong spatial dependencies. Nearby stations that serve different transit routes often do not have connected transit services due to proximity. However, these stations together shape the transit accessibility of the local area. In this sense, it is desirable to account for the spatial dependences among neighboring stations in the network analysis.

One of the methods for addressing these two issues consolidates the transit network by introducing transfers as virtual links among stations. Whether a transfer link exists between two stations depends on many factors, i.e., the existence of walkable paths. Fortunately, in this study, we are able to add virtual transfer links into the network analysis based on the observed transfers in the EZ-link dataset. To single out the infrequent transfers, only those station pairs with at least two observed daily transfers are considered for virtual links. Based on this prerequisite, approximately 10,238 transfer links are directly observed and are added to the transit network.

4.4.3.3 Connection Power

To avoid the situation in which all connections among transit stations are counted equivalently in the network analysis, weighting of the connections is introduced to distinguish the connectivity and accessibility of connections caused by different quality of services. Following Mishra et al. (2012), the weight of edges between transit stations are formulated as:

$$P_{ij,t} = \alpha(F_{ij,t}) \times \beta \frac{1}{T_{ij,t}} \qquad (4\text{-}3)$$

where *i*, *j* are two transit stations, $F_{ij,t}$ is the frequency of the transit services between station *i* and *j* during time period *t*, $T_{ij,t}$ is the average scheduled travel time of the transit services between station *i* and *j* during time period *t*, α is the scaling factor coefficient calculated as the reciprocal of the average frequency of all edges, and β is the scaling factor coefficient represented by the average travel time of all edges in the network.

4.4.3.4 Transit network measure

Table 4-6 lists the selected indicators of the transit network analysis. These indicators are generally derived at the station level. The degree of centrality of a node is commonly measured by the number of other nodes to which it is linked. The strength of a node is the node degree weighted by connection powers of its edges or the sum of connection powers on the edges of the node. However, the traditional method of computing degrees and strengths does not necessary reveal the centrality of a node because RTS stations are only directly connected to other RTS stations, which are generally far fewer in the number of bus stops. According to the EZ-Link data, although the no-transfer trips account for 73% of total transit trips, 22% of the trips involve one transfer. Therefore, it might be more reasonable to calculate the degrees and strengths of nodes based on the connectivity involving no more than one transfer.

The connectedness of a station is represented not only by how many other stations it is connected to but also by how connected those stations are as well. A regional bus station junction with connections to other regional bus station junctions and MRT stations may be weighted more heavily than a bus stop served by same amount of bus routes. Eigenvector centrality considers the variance of the influence of nodes with different consecutiveness (Bonacich, 2007). Technically speaking, the eigenvector centrality of a node is the value of the first eigenvector of its adjacency matrix and can be interpreted as arising from a reciprocal process in which the centrality of each node is proportional to the sum of the centralities of those nodes to which it is connected. In other words, nodes with high eigenvector centralities are those that are connected to many other nodes, which are again connected to many others, and so on. The idea behind eigenvector centrality is that the importance of the nodes with which a node is connected influences the importance of that node in the network. Unlike other directed networks, in the case of the directed transit network in this study, eigenvector centrality is concerned with the importance of outgoing nodes as opposed to that of incoming nodes because people are more inclined to forward thinking when planning trips or activities in a day.

An extension of eigenvector centrality is hub and authority centralities, which are formulated to describe two types of roles played by nodes in the network. Stations with a high hub centrality value generally have many services linked to those stations with a high authority centrality value and thus present a large out-degree. In contrast, stations with a high authority centrality value generally have many services linking from those stations with high hub centrality value and thus present a large in-degree. In other words, a station is good hub if it has transit services oriented toward major transit stations in the city that are connected to and from many other stations. A station is a good authority if it has transit services from those stations that connect to many other stations. A single station may have high hub centrality and authority centrality at the same time. According to Newman (2010), hub centrality is actually the eigenvector centrality of the matrix $A^T A$, where A is the adjacency matrix. Authority centrality, in contrast, is the eigenvector centrality of the matrix AA^T. Unlike the eigenvector centrality, which only generates non-zero values for the strongly connected node cliques in directed networks, hub and authority centrality can generate non-zero centrality values even for weakly connected node cliques (Newman, 2010).

The closeness centrality is defined as the average shortest scheduled travel time of all nodes in the network to the given node. Closeness centrality is a natural measure of the spatial centrality of spatial networks such as the urban transit network. The assortativity coefficient measures the level of homophyly of the graph based on certain vertex labels or values assigned to vertices (Newman, 2010). If the coefficient is high, this means that connected vertices tend to have the same labels or similar assigned values.

Measures of centrality and connectivity typically describe a station position in the global network. In the local transit network, the role of a station is related to how many other neighboring stations passengers transfer to it or how many other neighboring stations passenger transfer from it. Analogous to degree centrality, the number of stations that have transfer connectivity with the target station is the named transfer degree in this case.

Table 4-6. Transit network measures

Indicator	Mathematical Formula	Range	Description	Citation
Centrality and Connectivity				
Routes		>=0	Number of routes	
Degree centrality	$D_c(n)=\frac{\sum_{i}^{N}\sum_{j\neq i}^{N}a_{ij}}{N-1}$ where $a_{ij}=1$ if *i* and *j* are directly connected and 0 otherwise.	0-1	Normalized degree based total number of direct connection to other stations	Newman, 2010; Soh et al., 2010; Barthelemy, 2010;
Strength	$S_c(n)=\frac{\sum_{i}^{N}\sum_{j\neq i}^{N}a_{ij}p_{ij,t}}{N-1}$ where $p_{ij,t}$ is the weight of connection between *i* and *j* defined by eq(4-1)	0-1	Weighted degree based on the frequency of services between station pairs	Newman, 2010; Barthelemy, 2010;
Eigenvector centrality	$E(i)=\frac{1}{\lambda}\sum_{j\in G}a_{ij}E(j)$ where λ is the greatest eigenvalue of $AE=\lambda E$, A is the adjancey matrix $\{a_{ij}\}$	>= 0	Measures the influence of nodes in the network.	Newman, 2010; Barthelemy, 2010;
Closeness centrality	$CC(n)=\frac{\sum_{i}^{N}\sum_{j\neq i}^{N}t_{ij}}{N-1}$ Where t_{ij} is the shortest travel time between station *i* and station *j*	>=0	Average shortest travel time between a station and all other stations in the network.	Newman, 2010
Transfer degree	$TD_c(n)=\frac{\sum_{i}^{N}\sum_{j\neq i}^{N}tr_{ij}}{N-1}$ where $tr_{ij}=1$ if *i* and *j* are directly connected by transfer links.	>=0	Number of stations connected with observed transfers.	
Transfer strength	$TS_c(n)=\frac{\sum_{i}^{N}\sum_{j\neq i}^{N}tr_{ij}pr_{ij}}{N-1}$ where $tr_{ij}=1$ if *i* and *j* are directly connected by transfer links. pr_{ij} is the trasnfer weight between i and j.	>=0	Number of stations connected by transfers weighted by number of observed transfers.	

The topological properties of a transit station may influence the decisions of people who schedule multiple activities. It is also recognized that these measures are generated based on the entire transit network and might lie beyond the scope of consideration of travelers who are more likely to only focus on a subset of the transit network related to their activity spaces. In addition, indices such as hub and authority score may be correlated with certain land use, building, and establishment measures.

4.4.6 Correlations among indicators

Although the indicators are formulated to measure different dimensions of the urban built environment, it is inevitable that selected indicators may have high levels of correlations. For example, AWMFP and AWMSI could be highly correlated with the ratio of open space. The more fragmented and complex the local landscape mosaic, the larger the open space compared with the total area. Because these place indicators are used to predict activity types, to avoid high co-linearity and to increase the parsimony of learning models, it is important to select the appropriate independent variables.

Figure 4-16 shows the correlations among the built environment indicators. From the correlation matrix graph, it is clear that the levels of correlations among most indicators are low or moderate with a few exceptions. A strong positive correlation exists between the centrality and compactness indices of buildings. In addition, as expected, two measures that measure the shape of land use patches are highly correlated, i.e., AWMFD and AWMSI. In addition, the competition indices of businesses in different sectors (food, retail, social, and entertainment) are positively correlated with each other. Certain transit network measures, such as hub score and authority score, also have a strong correlation.

In the previous literature on the built environment measurement, factor analysis is an approach commonly employed to use smaller number of factors to represent the variability of many correlated variables (Diao, 2010; Krizek, 2003). In this study, because some built environment metrics vary spatially but also temporally, we need a spatiotemporal clustering approach that can distill the correlations of the measures in two dimensions. Few clustering approaches have achieved satisfied performance on this problem due to the escalated complexity of two-dimensionality. Moreover, it is unclear whether the factor loadings derived from the factor analysis can be directly applied to calculate new factors for new data, which could have a different correlation structure due to the changes in urban built environment. More theoretical or empirical evidences are needed. In this study, the highly correlated built environment indicators are further scrutinized and selected when they are fed into the activity learning models.

4.5 Summary

The objective of this section is to formulate generalized spatial indicators that can capture and measure the rich set of spatial and temporal components in the urban built environment that have differentiated impacts on the locational choices of different activities. The intent is to facilitate the translation of transit card traces into meaningful daily activity patterns by using the built form indicators to estimate the types of activities that transit riders engage in. The measurement draws on the rich spatial information generated from the data integration processes described in Chapter 3. The temporal information of businesses like the open hours is imputed using a supervised learning model developed based on the Random Forest (RF).

The spatial metrics of built environment focus on four dimensionalities: land use composition, spatial layout of buildings and streets, spatial distribution of establishments and transit network topology. For example, the land use contagion index measures the extent to which the land use types are separated in the catchment area. The neatest neighbor index of establishments measures the level of spatial aggregation of establishments in the catchment areas, from clustering to random distribution to uniform distribution. Each measure was computed individually for the catchment areas of transit stations to describe the "local environments" that people experience when they traveled and engaged in activities. Special attention is paid to the spatial and temporal dimensions of measurement. Measures of the availability of establishments and urban transit services are differentiated by time of day based on estimated open hours and transit schedules respectively. The spatial metrics described in this chapter are able to provide a more comprehensive and time-sensitive measurement for the urban built environment surrounding transit stations, which might have more correspondences with the choices and behaviors of travelers.

The spatial scale effect of measurement is always an important concern for the built environment measurement since the spatial aggregation of geographical entities such as buildings and businesses varies as spatial scale changes. In this study, the spatial scale of the catchment areas of transit stations is approximated by the average walking distance observed in the Household Interview Travel Survey (HITS) of Singapore in 2008, differentiated by the type of transit stations. However, some built environment measures can be sensitive to changing spatial scales. It is more desired to include values of measures corresponding to multiple spatial scales of the catchment areas when characterizing urban forms. Since the spatial scale effect of measurement is not the focus of this study, we will examine this in future work.

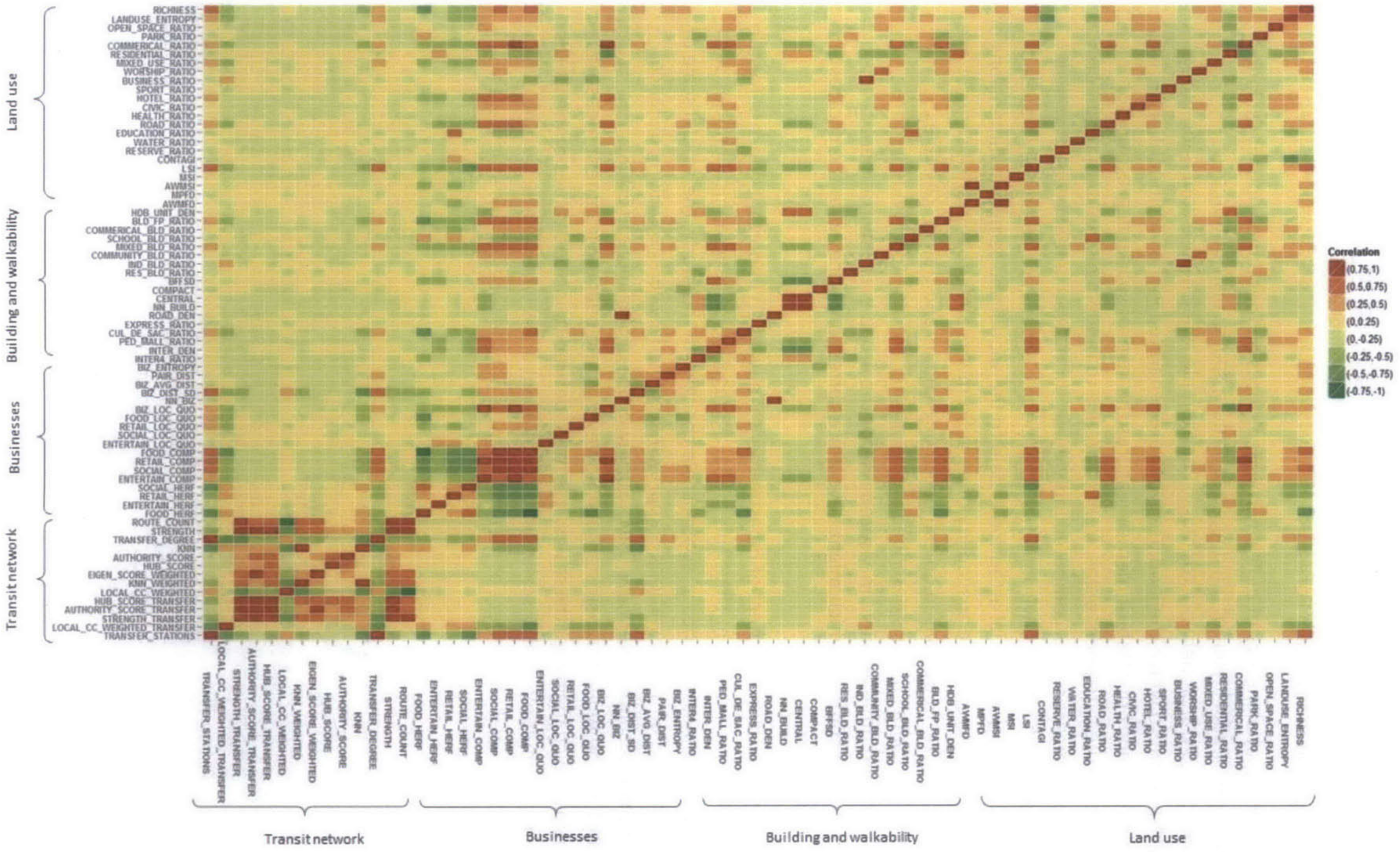


Figure 4-16. Correlation matrix of the selected built environment measures

Chapter 5. Learning Transit-oriented Activity Types from a Household Travel Survey

5.1 Introduction

Traditionally, household travel survey data or time-use survey data have typically been used for activity-based travel demand modeling. However, transit smart card transaction data have their advantages when used for the exploration of activity-travel patterns. The data are easy and cheap to obtain. Detailed transit trip information of users can be generated over a long period of time, as long as the electronic fare payment system is in operation. In addition, the data provide the possibility to capture the trajectories and movements of users through the spatial and temporal details of the transit trips recorded. Further, the data usually cover a much larger population size than do most surveys. In Singapore, the majority of frequent transit passengers use smart cards for convenience and, sometimes, economy. However, the drawbacks of the transit smart card transaction data are equally salient. The social-economic and demographic information of users, which is crucial to activity-travel behavioral research, are generally lacking in this type of data. We can only rely on the types of cards to differentiate passengers. Moreover, except for the in-vehicle travel time, the waiting time and walking time for trips are absent in the data. The spatial unit of the analysis is usually the transit station because only the boarding and alighting stations of the transit trips are recorded. However, one of the most outstanding problems that hinder the use of transit smart card transaction data in activity-travel research is that the purposes of the transit trips (i.e., the activities) are unknown. This gives rise to the question of whether it is possible to infer the unknown activity types from observed information in the data such as the time, the location and the sequence of trips.

Although it is of great interest to researchers, decision makers and service providers to extract the embedded activity patterns from the transit smart card transaction data, few studies have focused on the activity aspect of transit smart card transaction data. In this chapter, a group of statistical learning methods is investigated to learn the activity types from a household travel survey dataset, the HITS survey in year 2008. The models incorporate the detailed spatial and temporal explanatory variables including the urban built environment metrics discussed in Chapter 4 to determine potential activities.

5.2 Methodology

Because the information provided by the EZ-Link dataset is limited, it is a challenge to make the best use of the known information and, based on this information, to derive variables that can be used to forecast activity and travel from other data sources. Meanwhile, it is also crucial to find out proper ways to recognize unknown activities in future data. In this study, we investigate a set of machine learning models with different specifications and structures to estimate the associations between activities and assembled variables learned from existing data.

The decision-making process of activities and travels is very complex. Components such as activity types, destination choices, departing times, and activity durations, choice of transportation mode, interpersonal dependencies and social-economic characteristics of individuals are all simultaneously intertwined as a part of the decision-making process. In addition, an individual's choice of activity is considered to be subject to a number of constraints (Arentze and Timmermans, 2004):

- Household constraints: Household members share household obligations and resources, such as cars. Household incomes may constrain the number and types of activities.
- Situational constraints: The appropriate transportation mode must be in place to enable the actor to pursue a certain activity. Destinations lacking parking capacity and bicycle lanes will discourage car and bicycling trips.
- Institutional constraints: Working schedules and business hours of shops restrict the time allocations of certain types of activities.
- Time constraints: The time and duration of discretionary activities are conditioned by the time and duration of mandatory activities.
- Spatial constraints: Destinations are usually specific to a certain activity purpose. This means some activities cannot be conducted at certain locations. Moreover, the time constraints restrict the location choice of successive activities.

These constraints and dependencies provide a basis for activity type recognition, in light of a list of variables that are considered correlated with activities or travels but, in addition, can also be derived from the information in the EZ-Link dataset: specifically time, location, sequence and trip. Based on the time of trips, it is possible to determine the availability of opportunities and the weather conditions. Based on the location of origin and destination transit stations, the built environment of the surrounding areas can be characterized. Based on the sequence of trips and the locations, it is possible to learn home locations, workplaces and other frequently visited places. Based on the recorded trip information, the accessibility between the origin and the destination can be built into the learning model. In short, to maximally draw on the observed information of the EZ-Link data into activity-learning models, it is necessary to couple the EZ-Link data with auxiliary data from other sources to establish the spatially and temporally detailed contexts for the observed trips using quantitatively computable measures and variables. Then, the behavioral choices of travelers under reconstructed contexts can be estimated from training data, which is the household travel survey data in this case. Figure 5-1 is the conceptual interpretation of the learning and inference of activity types associated with the EZ-Link data.

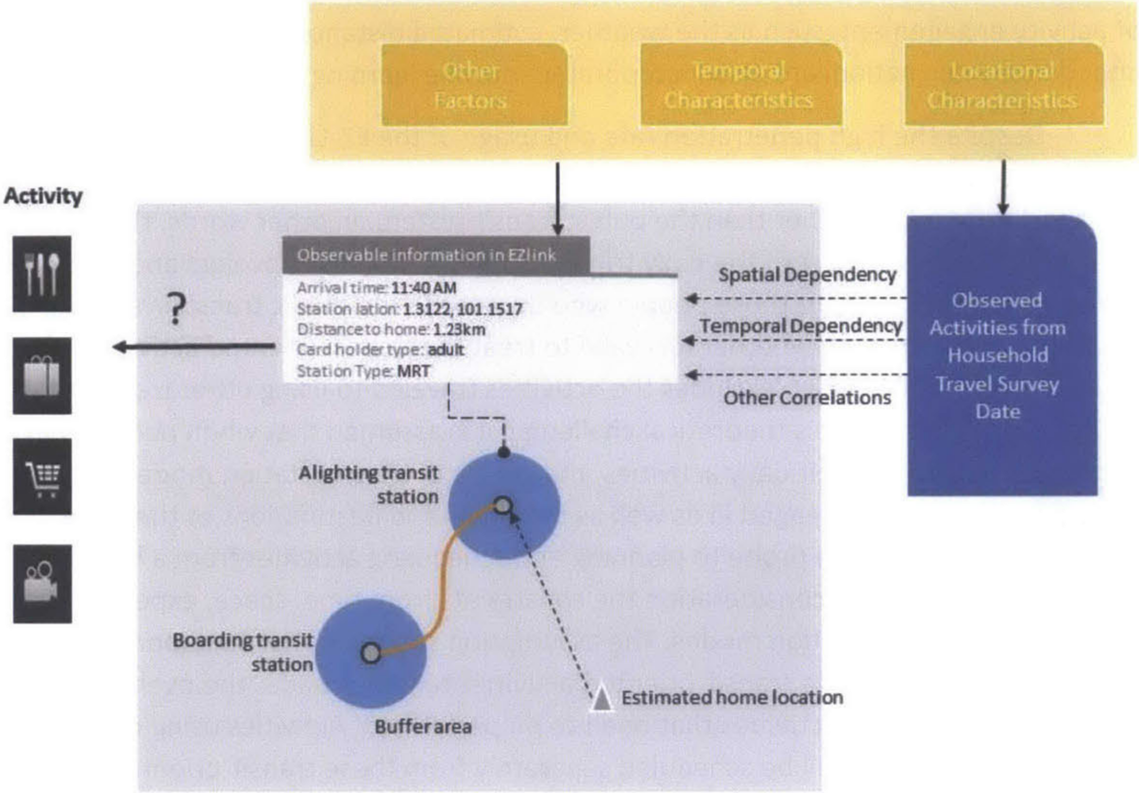


Figure 5-1. Illustration of activity learning and inference from the EZ-Link data

5.2.1 Modeling Framework

The proposed framework of learning and inferring activity-travel patterns from the transit smart card transaction data contains two pipelines, as shown in Figure 5-2. The first pipeline is to train the learning models by drawing on the transit trip and activity information in the HITS survey dataset from 2008. The second pipeline, which focuses on the activity inference of the EZ-Link data, is described in Chapter 6. To apply the learning models to the EZ-Link data, it needs to be ensured that the same set of explanatory variables or predictors can be generated from both datasets. However, unlike in the EZ-Link data, whereby both the boarding and alighting stops are known, only the boarding and alighting RTS stations can be found in the HITS 2008 dataset.

Information about the bus stops of transit trips is absent, although bus routes were reported. Therefore, the first task is to identify unknown bus stops in the HITS dataset from the observed destinations (recorded as postcodes) and bus routes. In addition, as the types of activities engaged are to a large degree correlated with the opportunities available at destinations as well as the surrounding built environments, pre-computed built environment measures for the catchment areas of transit stations are retrieved as a major part of the specification of the model. Other variables observable or extractable from the EZ-Link dataset and considered to be relevant for the decision-making process of activity engagement, such as the weather, estimated distance to home, and travel impedance information, are also incorporated into the learning models.

Despite the high penetration rate and usage of the EZ-Link smart card in Singapore, the trips recorded in the EZ-Link dataset do not include those using transportation modes other than the public transit system. In other words, the EZ-Link dataset only contains part of the daily trip information for most travelers and does not include information about those people who do not use the public transit system. This raises the question as to whether it is valid to treat the transit-oriented activities of an individual as a chain, which excludes the activities traveled to using other transportation modes. To circumvent this theoretical challenge, it is assumed that when people make decisions concerning their daily activities, the choice of transportation modes dictates the types of activities engaged in as well as the times and destinations of the activities. Theoretically, people are prone to planning and scheduling activities from a holistic perspective, taking into consideration the constraints from time, space, expense, family and available transportation modes. The assumption supposes that Singapore residents tend to plan and schedule transit-oriented activities together under the overall considerations of daily activities that need to be performed. Activities using other transportation modes will be scheduled separately from these transit-oriented activities. This makes transit trips a sub-chain of peoples' daily activity-travel sequences and makes the transitions between transit-oriented activities more interpretable.

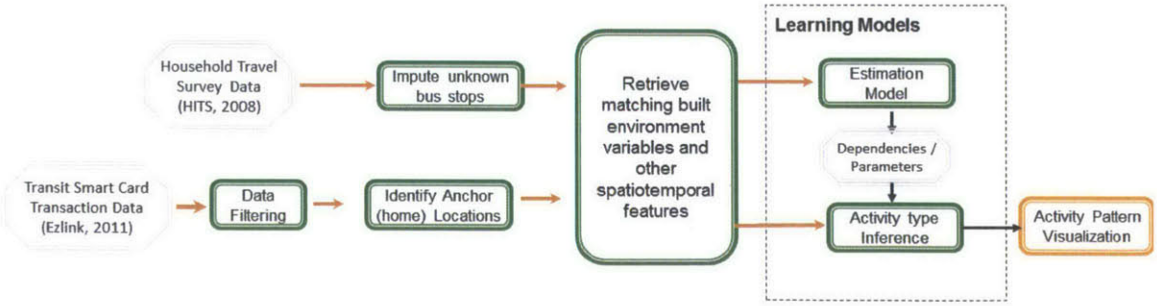

Learning Models
Household Travel Survey Data (HITS, 2008)
Impute unknown bus stops
Retrieve matching built environment variables and other spatiotemporal features
Estimation Model
Dependencies / Parameters
Transit Smart Card Transaction Data (Ezlink, 2011)
Data Filtering
Identify Anchor (home) Locations
Activity type Inference
Activity Pattern Visualization

Figure 5-2. Analytical framework of the activity-learning and inference module

In this study, we employ two types of learning models in light of two different activity generation and scheduling assumptions. The first type of learning model is the multinomial logit model (MNL), which focuses on individual transit trips without considering the dependencies among sequential trips and activities. The second type of model is a Conditional Random Fields (CRFs) model, which accounts for the sequence of activities in the form of Markov chains. The Conditional Random Fields model is a full-fledged sequence-learning model, which has been demonstrated to be effective in inferring sequences of unknown information from sequences of known observations. It has been used intensively in the field of pattern classification in the areas of, for example, gesture and speech recognition. Under the framework of CRFs, the activity-travel patterns of travelers are treated as sequences, and the transitions among different types of activities are explicitly represented.

5.2.2 Multinomial Logit Regression model

The multinomial logit model, also known as multinomial logistic regression, is a discrete choice model usually used to predict probabilities of different categorical outcomes of dependent variables contingent on given independent variables. It is also used extensively for classification and pattern recognition in the field of machine learning. Within the multinomial logit model framework, the probability that an individual *p* engages in activity type *y* near location *s* during the time period *t* conditional on mode choice being transit and given the individual, temporal, locational and trip-related attributes $\{X_p, X_t, X_{st}, X_{pst}\}$ can be formulated as

$$P(Y|X) = \frac{e^{(\beta_{0a}+\beta_{1a}X_p+\beta_{2a}X_t+\beta_{3a}X_{st}+\beta_{4a}X_{pst})}}{\sum_{i\in C} e^{(\beta 0_i+\beta 1_i X_p+\beta 2_i X_{t+}\beta 3_i X_{st}+\beta 4_i X_{pst})}} \qquad (5\text{-}1)$$

where β is an *n*-dimensional vector of coefficients corresponding to *n* independent variables and C is the collection of all activity types. Traditionally, β is estimated by maximizing the log-likelihood function of the specification

$$LL(\beta) = \ln \prod_{j=1}^{N} P_j(a|p,s,t)^{a_j}$$
$$= \sum_{j=1}^{N} s_j \ln \left(\frac{e^{(\beta_{0a}+\beta_{1a}X_p+\beta_{2a}X_t+\beta_{3a}X_{st}+\beta_{4a}X_{pst})}}{\sum_{i\in C} e^{(\beta 0_i+\beta 1_i X_p+\beta 2_i X_{t+}\beta 3_i X_{st}+\beta 4_i X_{pst})}} \right)_j \qquad (5\text{-}2)$$

where *N* is the sample size and a_j is the activity type of sample *j*. For *C* alternatives, only *C-1* alternative-specific constants and alternative-specific variables are estimable because for any values of the parameters $\{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4\}_a$, $\{\beta_0 + c, \beta_1 + c, \beta_2 + c, \beta_3 + c, \beta_4 + c\}_a$ gives the same probabilities (Ben-Akiva and Lerman, 1985), where *c* is a constant value. However, as we will see later, regularized MNL models are able to identify the coefficients of variables for all *C* alternatives by including a penalized regulation term that is sensitive to manipulation of coefficients as in $\{\beta_0 + c, \beta_1 + c, \beta_2 + c, \beta_3 + c, \beta_4 + c\}_a$ in the log-likelihood function. Using the estimated parameters β and equation (5-1), the MNL model can predict a set of probabilities representing the chances that an individual engages in different types of activities under the occasions described in the new data.

5.2.3 Conditional Random Fields model

A daily activity pattern is a series of activities based on sequential choices of locations, durations, transportation modes etc. These components are inter-mingled in the process of determining and planning a course of activities. The sequence of an activity-travel pattern not only signifies the priority of activities at different times of day but also implies a lifestyle. Regularly doing shopping on the way from a workplace to home and doing shopping after eating out at night indicate two different lifestyles. Hence, when investigating the correlations between activities and urban-built environments, it is important to consider the sequence of activities and their interdependencies. Because it is not straightforward to incorporate sequences of activities into traditional MNL models, a linear chain conditional random fields (CRFs) model is employed to formulate sequential dependent relations among activities.

A CRFs model is an undirected graphical model that has been increasingly used for temporal classification (Vail et al., 2007). The model represents the conditional probability of a particular sequence of states or classes Y given a sequence of observed independent variables X. Similar to the Hidden Markov Model (HMM), CRFs are doubly stochastic models whereby observations are modeled conditioned on a small number of discrete states, with Markovian transitions formulated among the states. However, unlike HMM, which models the joint distribution $P(Y, X)$, CRFs focus on the conditional probability $P(Y|X)$, which can be partitioned into computationally tractable sub-models. This avoids the complexity that dependencies among the input variables X need to be explicitly represented and hence afford the inclusion of a rich set of features and structures during model specification (Lafferty et al, 2001; Sutton and McCallum, 2006).

In the context of activity-travel studies, the daily activity chain of an individual can be considered as a stochastic process over a day. Changing from one activity to another is represented as a state transition within a Markov chain. Along with the activity chain, there are other synchronized sequences, such as destination chains, that can often be retrieved from household travel diaries as well as from the transit transaction data. In this study, activity types are considered as hidden states because they are not directly observable from urban sensing data. Locations are regarded as observable signals conditioned by the hidden states – activities (see Figure 5-3 c).

Two concerns arise when formulating a CRF model for learning activities from the HITS dataset and for inferring activities from the EZ-link dataset. First, the transition probabilities of activities and the correlations between urban spatial forms and choices of activities may vary across the time. Instead of specifying temporally varying dependencies, which will dramatically increase the number of parameters, the temporal variation in associations is expected to be captured by including within-day time periods as a group of dummy variables and using the time-varying built environment indicators to represent the dynamics of places. Second, when formulating the self-transition dependencies among activities, it is assumed that the choice of the current activity types is conditioned only by the last activity type. This assumption is consistent with the formulation of the first-order Markov chain. In reality, the current activity may depend on more than simply the last activity. This can be addressed by the skip-chain CRF model, which is not discussed in this study. Briefly, the main function of the CRF model is to learn the probabilities of activity types for each travelers given the time of day, the destination built environment, and the types of prior activities.

The linear-chain conditional random field specifies the conditional probability $P(Y|X)$ as follows:

$$p(y|x)=\frac{1}{Z(x)}\exp\left\{\sum_{m=1}^{M}\lambda_m f_m(y_k,y_{k-1},x_k)\right\}, \qquad (5\text{-}3)$$

where $f_m(y_{k,}y_{k-1,}x_k)$ is defined as a feature function representing the correlations between the observed variables of trip *k*, activity type of trip *k* and activity type of trip *k-1*. M is the number of feature functions in the model, including the relations between dependent variable and independent variables as well as the transitional dependencies among independent states. The feature function usually takes the form of an indication function. λ_m is the parameter for the feature function m. In the CRF model in this study, $f_m(y_{k,}y_{k-1,}x_k)$ can be decomposed into two types of relations $f_m(y_{k,}y_{k-1})$ and $f_k(y_{k,}x_k)$, where

$$f_m(y_{k,}y_{k-1})_{\{i,j,k\}}=\begin{cases}1 & if\ y_k=i\ and\ y_{k-1}=j\\ 0 & otherwise\end{cases}$$

and

$$f_m(y_{k,}x_k)_{\{i,j,k\}}=x_{m,k}$$

Each of the feature functions corresponds to a type of dependency in the model specification. $f_k(y_{t,}y_{t-1})$ is the transition from the last activity to the current activity, and $f_k(y_{t,}x_t)$ represents the dependency between the observed activity and the observed explanatory variables. The model is log-linear on the feature function. By changing the formulation of the feature function, it is possible to include more complex dependency relations in the learning model.

The conditional probability in Equation 5-3 requires calculating Z(x), which is a normalization function that guarantees that the sum of the probability distribution $p(y|x)$ is 1.

$$Z(x)=\sum_{y}\exp\left\{\sum_{m=1}^{M}\lambda_m f_m(y_k,y_{k-1},x_k)\right\} \qquad (5\text{-}4)$$

Note that Z(x) sums up all possible activity sequences, a number that is exponentially large. Fortunately, researchers have developed numeric approaches to compute Z(x) efficiently using the backward-forward algorithm (Lafferty et al, 2001; Sutton and McCallum, 2006). The coefficients are estimated by maximizing the log-likelihood function

$$LL(\lambda)=\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M}\lambda_m f_m(y_k^{(i)},y_{k-1}^{(i)},x_k^{(i)})-\sum_{i=1}^{N}\log Z(x^{(i)}) \qquad (5\text{-}5)$$

where *i* is the *i-th* individual in the dataset, *N* is the total number of individuals, and *K* is the total number of trips observed for individual *i*.

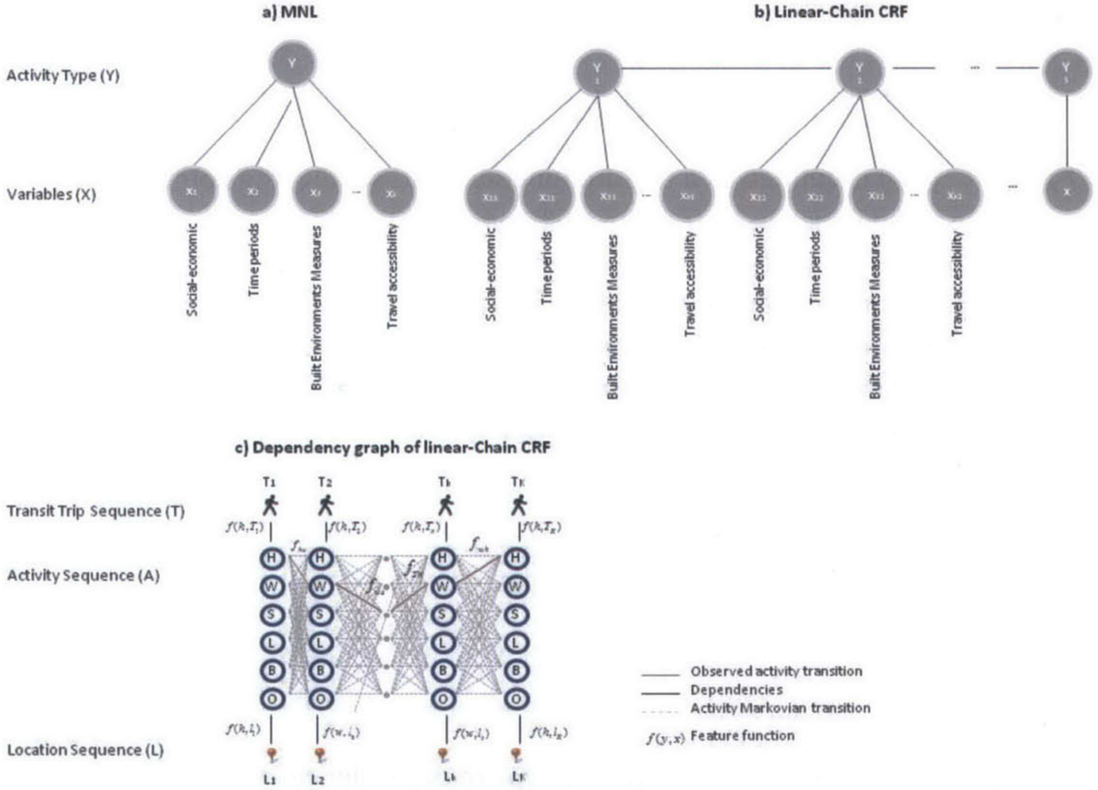


Figure 5-3. a) Generic structure of the MNL model. b) Generic structure of linear-chain CRFs. c) Illustration of a linear-chain CRF representation for a daily activity–travel (transit) chain.

Note: (in graph c: H – in-home activity, W - work, L – leisure, S - shopping, O - other, B - personal business)

5.2.4 The Relation between CRFs and the MNL model

Unlike Hidden Markov Models (HMMs) and Naive Bayes Models that assume that observations are independent of each other given the hidden states, both logistic regression models and CRFs make no assumptions about the correlations among explanatory variables. In other words, these models allow observed variables to be correlated. In this sense, the conditional random field (CRF) model is actually a generative form of the logistic regression model based on explicitly accounting for the dependencies between states. To see this, assume in the CRFs model that

$f_m(y_{k,}y_{k-1})_{\{i,j,k\}} = 0$

which means that the transitions among activities are not correlated. Then, eq. (5-3) takes the form of

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left\{\sum_{m=1}^{M} \lambda_m f_m(y_k, x_k)\right\} = \frac{1}{Z(x)} \exp\left\{\sum_{m=1}^{M} \lambda_m x_{m,k}\right\},$$

and $Z(x)$ becomes

$$Z(x) = \sum_{y} \exp\left\{\sum_{m=1}^{M} \lambda_m x_{m,k}\right\}$$

As a result, the probability of Y conditional on X becomes

$$p(y \mid x) = \frac{\exp\left\{\sum_{m=1}^{M} \lambda_m x_{m,k}\right\}}{\sum_{y} \exp\left\{\sum_{m=1}^{M} \lambda_m x_{m,k}\right\}},$$

This is exactly the form of how $P(Y|X)$ is formulated in the MNL model.

5.3 Model selection

In this section, four aspects of modelling are discussed for both the MNL model and the CRF model. Spatial and temporal confounding effects are a common issue faced by statistical models. Overfitting and imbalanced samples are common issues faced by classification and machine learning models. Model validation will be used to guide the comparison and selection of the model for activity inference from the EZ-Link dataset.

5.3.1 Spatial and temporal confounding effects

In statistical modeling, the confounding effect refers to the impact from unobserved variables that leads to biased estimates of the correlations between the dependent variables and the explanatory variables. Spatial and temporal confounding effects are likely to be present when the residuals of the statistical models are spatially or temporally correlated. Consider the effects of the time of day on the choices of activity types. Many variables explain the temporal variability of activities, including the working and school hours, open hours of businesses, weather conditions and the services provided by transit agencies. These variables correlate with the types of activities but also varying temporally. Similarly, spatial confounding variables are those variables correlated with the dependent variable but also vary spatially. If some of these variables are not included in the learning models, it will be difficult to distinguish the effect of observed spatial-temporal variables and the unobserved ones due to the correlations between the residuals and predictors. Therefore, it is necessary to include the spatially and temporally precise measures of the urban built environment as well as other spatio-temporal factors in the activity learning models to reduce the biases caused by the confounding effects.

5.3.2 Overfitting

Because this study applies models trained using the HITS dataset to predict the activities embedded in the EZ-Link dataset, there is a danger that the trained model may fit the noise of the training dataset as opposed to the true patterns by including irrelevant variables during the model specification. This is certainly the case when a group of urban built environment indicators and transit network measures are considered relevant to activity participation despite empirical evidence being limited. Therefore, it is most likely advantageous to select a simpler model that fits the training data worse over a more complex model that fits the training data better if the former model is restrained from matching noise in the training data.

To address the potential risk of overfitting, the regularized multinomial logit regression model (Bishop, 2006; Friedman et al, 2010) is employed for the learning task. Regularization typically refers to the trade-off between the goodness of fit and the model complexity. Unlike traditional MNL specification, the regularized multinomial logistic regression model includes extra regularization terms in the log-likelihood function to avoid model overfitting by penalizing an excessive number of variables and by detecting important predictors from redundant variables. According to Friedman et al., (2010), the regularized MNL includes the penalized log-likelihood (PLL) function in the form of

$$PLL(\beta) = \ln \prod_{j=1}^{N} P_j(a|p,s,t)^{a_j} - \lambda P_\alpha(\beta) \qquad (5\text{-}6)$$

$$P_\alpha(\beta) = \frac{(1-\alpha)\|\beta\|_2^2}{2} + \alpha\|\beta\|_1$$

where $P_\alpha(\beta)$ are the penalized terms used to prevent extreme values of variable coefficients; $\|\beta\|_1$ is usually known as an L1 regularization term taking the form of $\sum_{i=1}^{k}|\beta_i \,|$; $\|\beta\|_2^2$ is known as an L2 regularization term taking the form of $\sum_{i=1}^{k}\beta_i^2$; $\lambda(\geq 0)$ is a complexity parameter used to discount the complex model specification; $0 \leq \alpha \leq 1$ is an elastic-net mixing parameter used as a compromise between the L2 regulation ridge ($\alpha=0$) and the L1 regulation lasso ($\alpha=1$); and α is the hyper-parameter, which needs to be tuned in the process of model estimation. Because the regularized MNL model considers the size of the regression coefficients to be part of the error term, the coefficients are encouraged to be small.

Similarly, in this case of the CRF model, to avoid overfitting, regularization is also used, which is a penalty on coefficient vectors whose norm is large. A common choice of penalty is based on the Euclidean norm (L2 regulation term). Then, the likelihood function (5-5) becomes

$$LL(\lambda)=\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M}\lambda_m f_m(y_k{}^{(i)},y_{k-1}^{(i)},x_k^{(i)})-\sum_{i=1}^{N}\log Z(x^{(i)})\text{-}\sum_{m=1}^{M}\frac{\lambda_k{}^2}{2\sigma^2} \qquad (5\text{-}7)$$

where σ is a free parameter determining the strength of the penalty or the variance of the Gaussian prior distribution (Goodman, 2004).

5.3.3 Imbalanced classes

The problem of imbalanced data in the field of data mining refers to the fact that the learning process of statistical models will usually be dominated by classes with many instances while underestimating or even ignoring rare classes (Weiss and Provost, 2001). When the sample size is limited, the ability of models to learn regularities inherent in small classes is also restricted. As a result, observations belonging to the rare classes are more prone to being misclassified than are classes with higher proportions of observations. Some datasets are naturally imbalanced across classes, such as activities in this study (see Table 5-1).

To urban planners and researchers, it is more valuable to correctly detect small classes, such as recreation and social activities, in the activity-based learning model because these activities have greater implications on the demand of urban services than do working and in-home activities, which account for the majority of activities in the daily life of urban residents. Moreover, in comparison with working and in-home activities, the destination choices of other activities are more volatile and hence more difficult to understand.

There are two common approaches to addressing the imbalanced class problem. One is to achieve a balance through resampling from training data by either oversampling rare classes or by underdamping dominant classes. The other approach is to increase the cost of misclassifying rare classes, which is also known as cost-sensitive learning (Sun et al., 2009).

In this study, we used the Synthetic Minority Over-sampling Technique (SMOTE) to augment the training dataset to improve the classification accuracy of rare activity types learned by the MNL models. Rather than resampling existing data, SMOTE creates synthetic instances of the minority classes. By synthetically generating additional instances of the minority class, the learning models are able to broaden the regions of the high-dimensional feature vector spaces associated with the minority class.

The SMOTE samples are generated by linearly combining the target minority samples with the k-th nearest neighbors of the same class in the feature vector space (Blagus and Lusa, 2012). The new SMOTE samples of a minority class will have supporting features taking the values of

$$\tilde{X} = X + \alpha(X^R - X)$$

where X is the feature value of a target sample of the minority class, X^R is the feature vector of a sample randomly selected from the k-th nearest neighbors of the target sample, and α is a parameter between 0 and 1. According to Blagus and Lusa (2012), SMOTE includes the following theoretic features:

1) The expected value of the SMOTE-augmented minority class is not changed by newly created synthetic samples. Instead, the variability is decreased.
2) SMOTE does not introduce correlations among variables. However, it does introduce correlations between original samples and new samples.

SMOTE provides a plausible measure for the issue of imbalanced classes faced by MNL models. However, it cannot be applied to CRFs models, which learn based on the chains of activities as opposed to the individual activity.

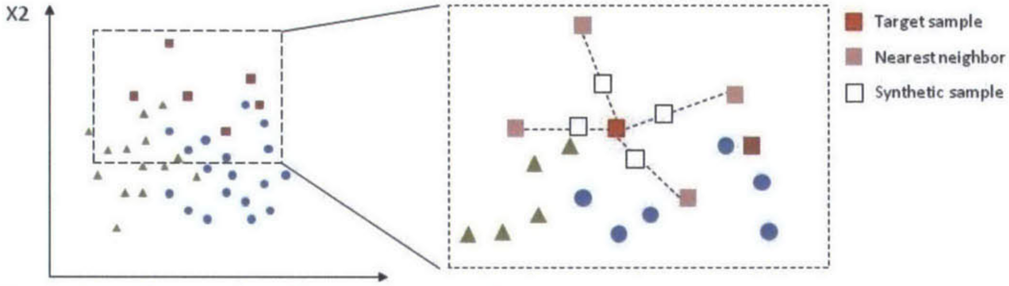


Figure 5-4. Illustration of SMOTE approach

5.3.4 Model validation

For traditional MNL models, the overall goodness of fit of the model is considered as a criterion during the model assessment. The overall goodness of fit is measured by the statistic commonly known as the adjusted likelihood ratio $\hat{\rho}^2$ or the adjusted McFadden R^2, which is computed as follows:

$$\hat{\rho}^2 = 1 - \frac{LL(\beta)/(n-p-1)}{LL(c)/(n-1)}$$

where $LL(\beta)$ is the maximum value of the log-likelihood function given the included explanatory variables, $LL(c)$ is the initial value of the log-likelihood function when only a constant is included, *n* is the total number of samples and *p* is the number of independent variables.

In the field of machine learning, the quality and predictivity of learning models are often validated and assessed using the results of a *k*-fold cross validation, which divides the dataset into *k* subsets. The model is also estimated and assessed *k* times. Each time, *k* -1 sets of the dataset are used for the estimation, and the remaining set of the dataset is used for validation. Next, the average error across all *k* trials is computed. The advantage of this method is that it is less dependent on how the data are divided. Every data sample is in a test set exactly once and is in the training set *k* -1 times. The disadvantage of this method is that the training algorithm has to be rerun *k* times, which means that it takes *k* times as many computations to perform an evaluation. Common choices of *k* are 5 and 10, depending on the size of the dataset and the computational cost. In this study, we choose the 5-fold cross validation, considering the estimation of some models is quite time consuming.

5.4 Data and variables

5.4.1 Activity categorization

The categorization of activities is mainly based on the trip purpose information reported by the respondents in the HITS 2008 dataset. As shown in Table 5-1, 17 trip purpose types are grouped into 10 activity categories. Multi-purpose trips are not included due to the absence of relevant information in the HITS survey.

Table 5-1. Activity classification based on HITS 2008 dataset

ID	Activity Category	Trip purpose type	Total	Transit subset[1]	Ratio (transit to total)
1	**Home**	Return home, return to another home	28591	13130	45.92%
2	**Working**	Go to work	14533	7324	50.40%
3	**Education**	Education	8372	3152	37.65%
4	**Leisure**	Entertainment, Recreation, Sports/exercise	666	311	46.70%
5	**Shopping**	Shopping	2037	1212	59.50%
6	**Eating**	Meal/eating break	1568	383	24.43%
7	**Social**	Social visit/gathering	1503	752	50.03%
8	**Personal Business**	Personal errand, Medical, Religious-related matters	1241	518	41.74%
9	**Business**	Work-related business	1318	318	24.13%
10	**Other**	To drop-off/pick-up someone, To accompany someone	5065	314	6.20%

Note. 1. A subset of trip samples in the HITS 2008 dataset using public transit means (bus, MRT or LRT).

The correlation found for using transit and conducting different types of activities is in part revealed by the ratios of trips using transit to total trips broken down by activity types (Table 5-1). For most activity types, the ratio for choosing transit is approximately 50%. However, for three types of activities, the chances of using transit are disproportionally low: *eating, business* and *dropping off* or *picking up* someone. Car drivers and car passengers together account for 37% of *business* trips, 55.9% of *eating* trips, and approximately 77.6% of *other* trips (mainly dropping-off/picking-up someone). Short-distance walking is another major mode of transportation for *business* (18.3%) and *eating* (10.2%). It is also noteworthy from Table 5-1 that the samples of different activity types are very imbalanced.

Because around half of the transit trips in the HITS 2008 are *returning home* trips, in the MNL model, *returning home* is not included as an alternative activity choice. Although *returning home* is comprised in the activity chains in the CRFs model to estimate the transitions among activities, it is explicitly tagged as known. For the *returning home* trips, the idea is to identify them by developing approaches to extract approximated home locations in the light of multiple day transit trip records in the EZ-Link data. This will be discussed in Chapter 6.

5.4.2 Temporal interval

Activity participation usually shows significant relevancy for within-day time periods. To capture this effect, we include dummy variables for the day of the week and the time of the day in the classification models. One day is divided into 11 time intervals: early morning (5-7 am), morning-peak hour (7-9 am), morning-work (9-11 am), noon (11 am-1 pm), afternoon-work (1-5 pm), afternoon-peak hour (5-7 pm), night (7-9 pm), and late night (10 pm-1 am). Most transit services stop before midnight, and only a few night buses provide services between 1 and 5 am. Consequently, there are a very limited number of trips made by transit and recorded in the EZ-Link dataset during this time period; therefore, it is not considered in the study. In addition, in this study, only weekdays are considered because the HITS survey data only covers weekdays. The weekend pattern is typically different from weekday pattern because most working and education activities do not occur on the weekends.

5.4.3 Imputing the boarding and alighting stops

In the HITS 2008 dataset, the information regarding the boarding and alighting bus stops of the trips is not disclosed if the trip was made using a bus. Instead, the route of the bus was reported by most survey respondents. This makes it possible to impute the bus stops associated with the public transit trips based on the route and the trip origin and destination. In this study, it is assumed that the stops on the reported bus route that have the smallest Euclidean distance to the origin and destination postcode locations are boarding and alighting stops, respectively. For records in the dataset where the bus route information is missing, all the bus stops within 400 meter buffers of the origin and destinations are considered as candidates, and subsequently, the transit network is skimmed to identify the fastest route connecting any one stop in the origin buffer with any one stop in the destination buffer. Accordingly, the origin and destination stops associated with the identified route are assumed to be the boarding and alighting stops.

5.4.4 Weather

Weather conditions vary across time and locations and could influence an individual's choice of activity. Although in Singapore, weather, such as thunderstorms, only influences local areas, only city-wide precipitation conditions and temperature information are available.

5.5 Modeling results and assessment

To learn the activity types from the transit trip and activity information encapsulated in the HITS survey, a group of models (as shown in Table 5-4) are specified to evaluate the following:

1) The effect of the time period;
2) The effect of built environment measures, especially concurrent built environment measures concerning the attractiveness of the catchment areas of the destination transit stations;
3) The effect of other concurrent factors, such as the weather and the travel time; and
4) The significance of transitions among different types of activities.

Meanwhile, on the modeling methodology side, we want to investigate whether the predictivity of the learning models will be improved by

1) Including regularized terms in the models to reduce the chance of model overfitting and
2) Including resampling strategies in the models to address the issue of imbalanced classes.

5.5.1 Learning Models

Table 5-2 lists the models estimated and assessed in this study. As described in section 5.2.3, models are compared on the basis of the likelihood ratio test as well as the classification accuracy resulting from a 5-fold cross validation for each model. For traditional MNL models, all exploratory variables, including the intercept term, enter into the model specifications as alternative-specific variables. Regularized MNL models are able to estimate coefficients of variables for all alternatives (Friedman et al., 2010). Table 5-3 lists the detailed variables considered in the learning models.

Table 5-2. List of learning models and assessment measures

Model name	Model description	Explanatory variables / features	Assessment
MNL1	Traditional Multinomial Logit Model (MNL)	Smart-card-holder types + zonal population and job densities	likelihood ratio test + classification accuracy by activity type
MNL2		Smart-card-holder types + stationary built environment measure of transit station catchment areas	likelihood ratio test + classification accuracy by activity type
MNL3		Smart-card-holder types + concurrent built environment measures of transit station catchment areas	likelihood ratio test + classification accuracy by activity type
MNL4		Smart-card-holder types + stationary built environment measure of transit station catchment areas + time period of day + transit trip + weather	likelihood ratio test + classification accuracy by activity type
MNL5		Smart-card-holder types + concurrent built environment measure of transit station catchment areas + time period of day + transit trip + weather	likelihood ratio test + classification accuracy by activity type
MNL6	Regularized MNL	Smart-card-holder types + concurrent built environment measure of transit station catchment areas + time period of day + transit trip + weather	Classification accuracy by activity type
MNL7	Regularized MNL + SMOTE Resampling	Smart-card-holder types + concurrent built environment measure of transit station catchment areas + time period of day + transit trip + weather	Classification accuracy by activity type
CRF1	Regularized CRF	Smart-card-holder types + concurrent built environment measure of transit station catchment areas + time period of day + transit trip + weather + transitions among activities	Classification accuracy by activity type

Table 5-3. List of variables included in the model specification

Variable type	Variable category		Variables
Dependent variable / response variable	Activity types		Activity
Independent variables / predictors	Smart card holder types		Senior, Student/Child, Adult (for reference)
	Time period of day		Early morning, morning peak, noon, afternoon peak, night, late night
	Built environment measures	Land use composition	Park area ratio, Open space area ratio, Mixed area ratio, Worship area ratio, Sport area ratio, Waterbody area ratio, Land use richness, Contagion index, AWMSI, Land use entropy, AWMFD
		Layout of buildings	HDB unit density, Nearest neighbor index, Centrality, Compactness, Ratio of residential buildings, Ratio of School buildings, Ratio of commercial buildings, Ratio of community buildings, normalized standard deviation of building footprint size

		Street layout	Road density, ratio of pedestrian mall, ratio of expressway, ratio of four-way intersections, number of intersections per km road length
		Business distribution	Nearest neighbor index, Business entropy index, normalized mean distance to station, standard deviation of distance to station, location quotient (food, retail, social, entertainment), number of business by types
		Transit Network	Transfer degree, strength centrality, hub score index, trip frequency, degree centrality,
	Transit Trip		In-vehicle time, transfers, MRT, dist (distance from alighting station to estimated home location)
	Weather		Rain (dummy variable), temperature

5.5.2 Effect of spatiotemporally detailed urban built environment measures

For models MNL1 through MNL5, traditional MNL models are used to classify activity types but using different sets of explanatory variables. In all five models, the working activity is used as a reference for the estimation. Table 5-4 presents the estimation and validation results of the MNL models.

In addition to the type of EZ-Link card holders, model MNL1 uses population statistics as well as jobs by different sectors at the Traffic Analysis Zones (TAZ) where transit stations are located as proxies of local attractiveness. The population and job statistics at aggregated zonal levels are common indicators of built environments and can be found in many activity-travel studies. As a baseline model in the study, MNL1 is used to show the extent to which the activity types can be recognized if additional spatially and temporally detailed built environment variables are not considered. The goodness of fit of MNL1 using the HITS 2008 dataset is 0.3276 in terms of the adjusted McFadden R^2. The average prediction accuracy of the 5-folder validation is 55.65%. However, most of explanatory power in MNL1 is actually contributed by the information on the types of card holders, which alone can achieve 0.275 of the adjusted McFadden R^2. The estimation result of MNL1 is reported in Table A1 of Appendix III.

Table 5-4. Estimation and validation measures of traditional MNL activity-learning models

	MNL1	MNL2	MNL3	MNL4	MNL5
Log-likelihood	-14,864	-14,143	-12,255	-11,285.94	-11,146.44
Adjusted McFadden R^2	0.3276	0.354	0.437	0.484	0.492
5-fold cross validation accuracy	55.65%	57.92%	60.53%	62.59%	62.96%

As opposed to the zonal population and job density information used in MNL1, model MNL2 uses stationary built environment measures of the catchment areas of transit stations, as described in Chapter 4. These measures are selected to capture characteristics such as land-use composition, spatial distributions of various types of buildings and establishments, street layouts and transit service connectivity at local neighborhoods, without considering the temporal variability. The estimation results are reported in Table A2 of Appendix III. According to Table 5-4, a moderate increase is observed in both the goodness of fit and the classification accuracy when comparing the result of MNL2 to that of MNL1.

Unlike the stationary built environment measures used in MNL2, MNL3 include the time-varying built environment measures, mainly the measures of establishments and transit services as a result of accounting for open hours and transit schedules in the measurement. This leads to a significant improvement in the goodness of fit, which rises from 0.354 (MNL2) to 0.437 (MNL3). However, as discussed later, this does not necessarily imply that the concurrent built environment measures are significantly better than the stationary measures in predicting activity types. Rather, it is more likely that the time of day, which is highly correlated with service availability at the supply side, and the time constraint of people at the demand side, play a role in the improvement of the explanatory power of MNL3.

The model specification of MNL4 is an extension of that of MNL2. In addition to the same set of stationary built environment measures of transit station catchment areas used in MNL2, explanatory variables in MNL4 also include variables concerning transit trips, such as in-vehicle travel time, and variables concerning weather, such as temperature. In particular, a set of dummy variables representing within-day time periods of trips enter the model specification of MNL4 to capture the different choices of activities at different times of day. As a result, the adjusted McFadden R2 increases to 0.484, and the accuracy of the 5-fold cross validation is raised to 62.6%.

In contrast to MNL4, MNL5 uses the built environment measures concurrent with the transit trips. The estimation and validation results indicate that the prediction accuracy and the goodness of fit of MNL5 are only marginally better than those of MNL4. When comparing these results with the significant improvement from MNL2 to MNL3, in which time period dummy variables are absent, it can be suggested that the availability of opportunities and transit services are highly correlated with within-day time periods. When the time period dummy variables are not included in the model specification, the temporal variations revealed through the concurrent built environment measures capture the temporal variations in the activity-type choices caused by other temporal factors, such as working hours, and thus help to significantly improve the explanatory power of the model. Factors such as transit service schedules are already well tuned with travel demands. However, it is dangerous to jump to the conclusion that the supply of opportunities and services has limited effects on peoples' choices of activities. In MNL4, a good portion of the temporal effects of opportunities have been represented by temporal dummy variables and other travel-related variables, which explains the evident improvement of MNL4 over MNL2. For a subset of travelers, especially those who have fewer time restrictions, their choices of activities and destinations can be contingent on the availability of relevant opportunities and services. This is the reason that MNL5 produces slightly more accurate predictions compared to MNL4. In addition, the estimates of built environment variables in MNL5 should be less biased than those in MNL4 if we accept the temporal availability of opportunities at destinations as a factor that people will consider when planning for travel and activities.

Additionally, it is recognized that the opening hours of businesses, which are used to measure the spatio-temporal distributions of opportunities, are the results of a simple classification model built on limited sample sizes. It can be expected that more precise information on the temporal availability of opportunities will make additional contributions to the learning models.

Figure 5-5 shows the classification accuracies broken down by activity types from the 5-folder cross-validation for MNL1 to MNL5. It is clear that all MNL models can achieve relative satisfied classification accuracy for *working* and *education* activities, in part due to explanatory power contributed by the types of card holders. But the prediction accuracies are less satisfactory for other activities. It also appears that the accuracies are closely related to the size of samples in the training data, as the activity types with small sample sizes like *leisure*, *eating* and *work-related business* generally have lowest classification accuracies.

The estimation result for the MNL5 model is presented in Table A-5 of Appendix III. The effects of selected explanatory variables on the likelihood of different activity types are summarized in Table 5-5. Types of card-holders and the time periods of days are significantly correlated with most of the activities. Destinations that are further from home may imply higher probability for *social*, *work-related* and *othe*r activities. It is also found that high temperature encourages most out-door activities except for *working* and *education*. This is because most *working* and *education* trips occur in the morning, when temperature is relatively low.

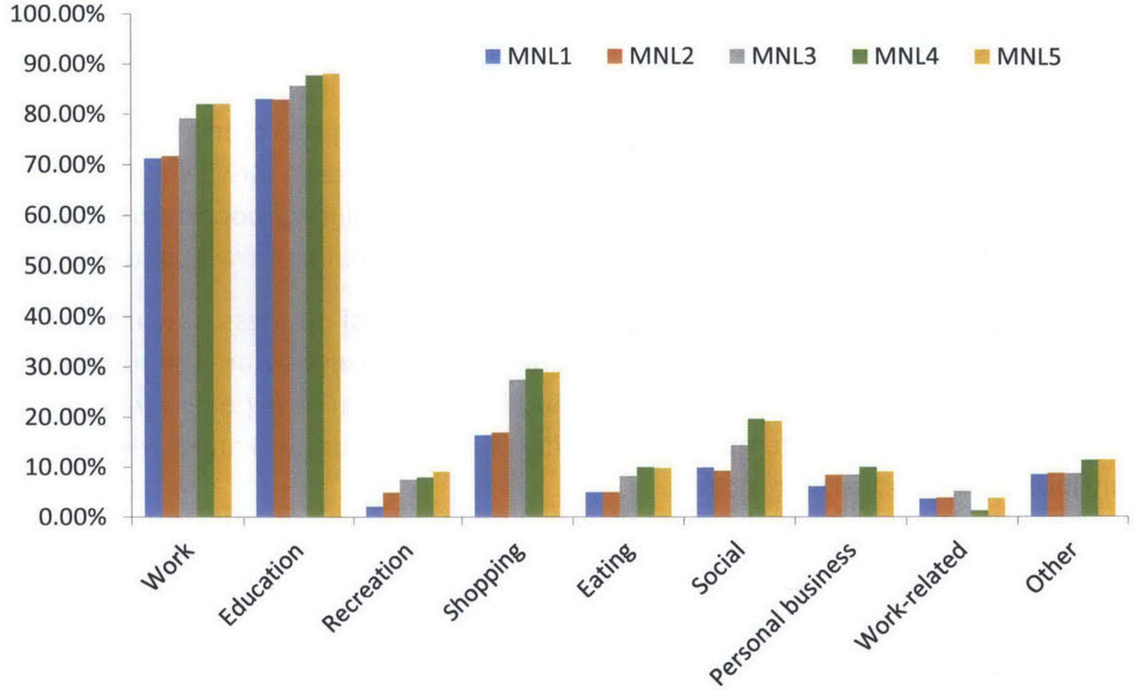


Figure 5-5. 5-fold cross validation accuracy rate by activity types (from MNL1 to MNL5)

Table 5-5. Effect of the selected explanatory variables on activity types from the MNL5

MNL5	Education	Leisure	Shopping	Eating	Social	Personal	Work-related	Other
Student/Child	+*	+*	+*	+*	+*	+*	+*	+*
Senior	-	+*	+*	+*	+*	+*	+*	+*
Morning (5am-11am)	+	-*	-*	-*	-*	-*	-*	-*
Noon (1am-1pm)	+*	-*	-*	+	-	-	-*	-
Afternoon peak (5pm – 7pm)	+*	+*	+	+*	+*	+	+	+*
Night (7pm-9pm)	+	+*	+*	+*	+*	+*	+	+*
Late Night (9pm-1am)	-	-	-*	+	+*	-	-	-
Friday (dummy)	-*	-*	+	+	-	+	-*	+
#Transfers	+	+	-*	-*	+*	-*	+	-
Distance to home (km)	-	-	+	-	+*	+	+*	-*
Temperature	-	+*	+*	+*	+*	+*	+	+
Contagion index	-	-*	-*	-*	-*	-	-	-*
Park area ratio	+	+*	+*	+	+*	+	-	+
LU entropy	+	-	+	+	-	+	-*	+
HDB unit density	+*	+	+*	+*	+*	+*	+	+*
Nearest neighbor index - Bld	+	-	+*	-	+	-	-	-
Centrality	-*	+	-*	-	-	-	+	+
Compactness	-	-	-	+	-*	-*	-	-*
Four-way intersection ratio	+*	+	+	+	+	-	+	+*
Intersection per km road	+*	-	-	+*	+	+	-	+
Entropy index –establishment	+	+*	-	+*	-	+	+	+*
Establishment Distance to station (stdev)	+	-	-	-*	-*	-	+	-*
Nearest neighbor index - establishment	+	+	-	-	+	-	-*	+*
Hub centrality (transfer network)	+*	-	+	-	+	-	-	-
Transit service frequency	+	-*	+*	+	-	-*	-	-*

Note: * significant at <0.1 level. + positive coefficient; - negative coefficient

Many built environment indicators are found to be significant in classifying one or more types of activities. A greater contagion index corresponds to less segmented land use. Thus, areas with a high contagion index are more likely to be study places, such as Science Parks and Industrial Zones. Catchment areas with high area ratios of parks are more likely to be associated with recreation, shopping and social activities. Land use mix does not appear to be a significant factor for most activity types. This could partly be because the area ratios of many types of land use are also included in the model. The HDB unit density is used as a surrogate of the population density of the areas. As expected, most activities, including education, are positively correlated with HDB unit density in comparison to working activities. The centrality measures the closeness (average distance) of other buildings to the building with the biggest footprint in the areas. Apparently, workplace buildings are likely to have bigger footprints and are likely to be more clustered. Thus, areas with high centrality attract working and work-related trips. Most transit network measures do not appear to be significantly associated with the types of activities. Travelers alight at stations with high hub centrality values, which have many out-connections, and tend to have a greater propensity to engage in education activities. In comparison with working, stations having frequent services are more likely to be associated with shopping trips but are less likely to correlate with recreation, personal business and other trips. The reason for this is most likely because transit services are more frequent in shopping centers, hawker centers and workplaces but less frequent in parks and in open spaces, where *leisure* activity are likely to occur.

5.5.3 Effect of model regularization

As mentioned in section 5.3.1, the major objective of introducing penalty terms to the likelihood function of the MNL models is to avoid overfitting. In this study, the regularized MNL model implemented in the *glmnet* package of R is used to learn activity types from the HITS survey data from 2008. The *glmnet* package can generate a series of runs that take different input values of the complexity parameter λ and determine the explanatory variables in the model. If the value of λ is sufficiently large, it will reduce the number of explanatory variables with extreme coefficients and those making smaller contributions to the model. On the other side, if λ is small and if too many explanatory variables are included in the model, the model may have a propensity to overfit the data. The decision for the model selection is usually based on the minimum deviation between two model specifications with different variables and λ:

$$Dev(y) = -2\big(\log\big(LL(y|x_1,\lambda_1)\big) - \log\big(LL(y|x_2,\lambda_2)\big)\big)$$

where $LL(y|x_1,\lambda_1)$ is the log-likelihood of the first model, which takes explanatory variables x_1 and the penalty variable λ_1, while $LL(y|x_2,\lambda_2)$ is the log-likelihood value of the second model. It is important to note that all variables, expect for dummy variables, are already standardized in MNL6.

Figure 5-6 shows the performance of the regularized MNL models with 5-fold cross validation using different λ values. The mean deviation is used as the measure of error. The mean cross–validated error curve and a one-standard-deviation band are plotted. The top axis of the plot indicates the number of explanatory variables included in the models. The left vertical line corresponds to the model with the smallest minimum error, and the right vertical line corresponds to the best model based on the "one-standard-error" rule, which acknowledges that the risk curves are estimated with error, therefore tending to the side of being conservative (Hastie et al, 2009; Friedman et al, 2010). In this case, the model following the "one-standard-error" rule is only marginally worse than the model of the minimum error but is more conservative. The value of λ corresponding to the "one-standard-error" line is 0.0028.

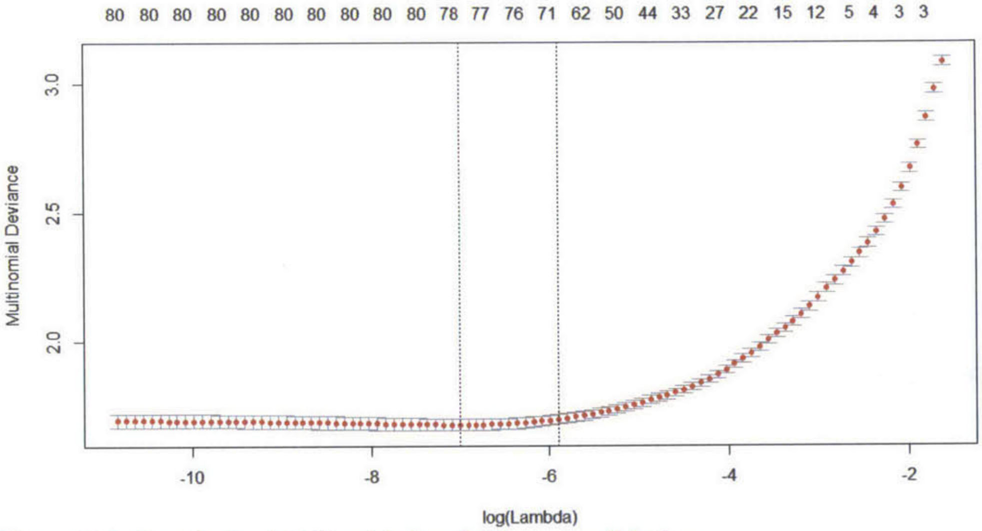


Figure 5-6. Regularized MNL with five-fold cross validation.

Based on the results of the 5-fold cross validation, the overall performance of MNL6 is slightly better that of MNL5, with an overall accuracy of 64.35%. The classification accuracy by activity type in MNL6 is close to that of MNL5, as shown in Figure 5-7. However, MNL6 is able to select from variables with high correlations and exclude variables having smaller contributions to the learning model. For example, the dummy variable of the time period "noon (11 am – 1 pm)" is excluded because its effect on the activity types is not significantly different than that of the base dummy variable "afternoon study (1 pm – 4 pm)". Similarly, a land use entropy index is also utilized because it correlates groups of area ratios with different types of land use. Among the two shape indices of land use patches, AWMSI and AWMFD, only AWMSI is used because of the high correlation among them. In this way, 10 variables are removed during the model specification to produce a more conservative learning model. The estimation results of MNL6 are reported in Table A6.

On the other hand, as mentioned earlier, because of the inclusion of the penalty term into the log-likelihood function of the regularized MNL model, the magnitude of coefficients will shrink because of the effort of minimizing penalties. Thus, coefficients cannot be scaled freely as those in the traditional MNL model can be. This extra level of constraint allows the identification of the coefficients for all alternative activity types, as shown in Table A6. In other words, no reference alternative is needed in the regularized MNL model. This helps to improve the performance of the model. Meanwhile, it can be observed from Table A6 that the magnitudes of coefficients of variables remain small, mostly ranging between -1 and 1.

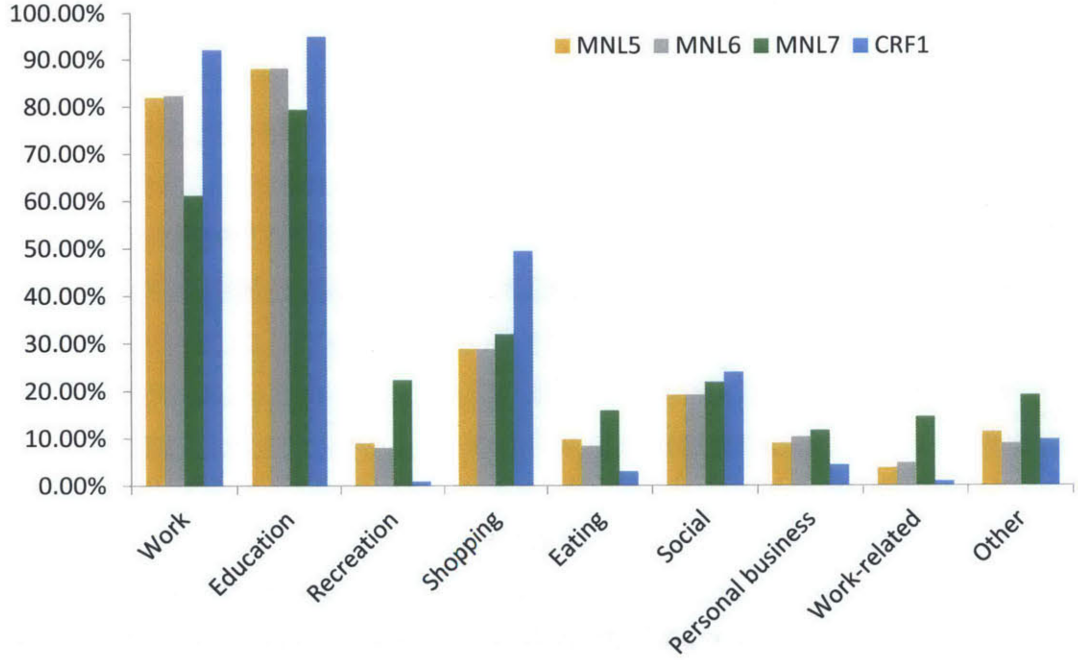
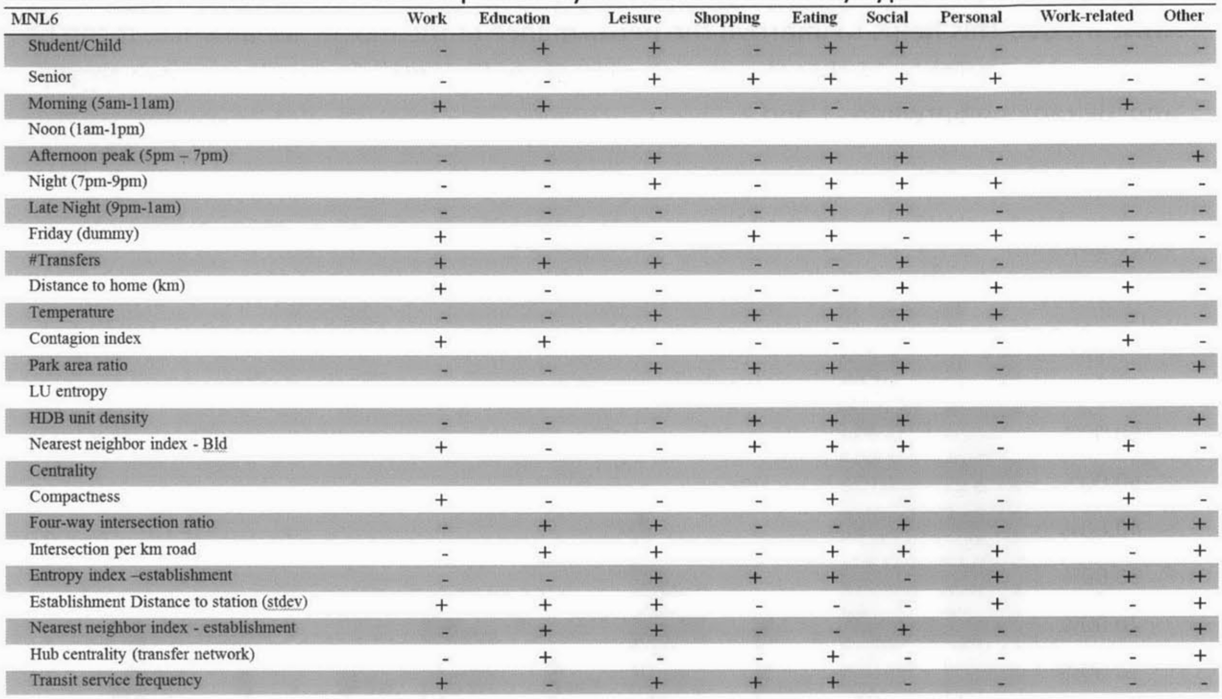


Figure 5-7. 5-fold cross validation accuracy by activity types (from MNL5 to CRF1)

Because of the difficulty in deriving the Hessian matrix in the process of estimation, the standard errors of the coefficients are not calculated, and hence, the significance of variables cannot be tested. Table 5-6 summarizes whether the effects of the selected variables from Table A6 are positive or negative effects. In comparison with the effects in Table 5-5 for MNL5, some changes between effects being positive and negative are observed in MNL6 after the regularization and standardization of the variables and the deletion of redundant variables. For example, according to Table 5-6, student and child concession card holders are less likely to be involved in activities such as shopping, personal businesses and work-related activities compared to adult card holders. However, the corresponding cells in Table 5-4 are all positive. The positive effect of student and child card holders on shopping activities is ambiguous in Table 5-4 because it is not clear whether the positive effect is derived based on the comparison between student/child groups and adult groups or based on the comparison between shopping activities and working activities, which is the alternative reference in MNL5. In this sense, the regularized MNL model can produce a learning model with better interpretability.

Table 5-6. Effect of the selected explanatory variables on activity types from the MNL6



MNL6	Work	Education	Leisure	Shopping	Eating	Social	Personal	Work-related	Other
Student/Child	-	+	+	-	+	+	-	-	-
Senior	-	-	+	+	+	+	+	-	-
Morning (5am-11am)	+	+	-	-	-	-	-	+	-
Noon (1am-1pm)									
Afternoon peak (5pm – 7pm)	-	-	+	-	+	+	-	-	+
Night (7pm-9pm)	-	-	+	-	+	+	+	-	-
Late Night (9pm-1am)	-	-	-	-	+	+	-	-	-
Friday (dummy)	+	-	-	+	+	-	+	-	-
#Transfers	+	+	+	-	-	+	-	+	-
Distance to home (km)	+	-	-	-	-	+	+	+	-
Temperature	-	-	+	+	+	+	+	-	-
Contagion index	+	+	-	-	-	-	-	+	-
Park area ratio	-	-	+	+	+	+	-	-	+
LU entropy									
HDB unit density	-	-	-	+	+	+	-	-	+
Nearest neighbor index - Bld	+	-	-	+	+	+	-	+	-
Centrality									
Compactness	+	-	-	-	+	-	-	+	-
Four-way intersection ratio	-	+	+	-	-	+	-	+	+
Intersection per km road	-	+	+	-	+	+	+	-	+
Entropy index –establishment	-	-	+	+	+	-	+	+	+
Establishment Distance to station (stdev)	+	+	+	-	-	-	+	-	+
Nearest neighbor index - establishment	-	+	+	-	-	+	-	-	+
Hub centrality (transfer network)	-	+	-	-	+	-	-	-	+
Transit service frequency	+	-	+	+	+	-	-	-	-

5.5.4 Effect of resampling of the minority class

To address the issue of imbalanced data in the MNL models, the SMOTE augmented samples are generated using the R package *DMwR* developed by Luis Torgo (2010). The *SMOTE()* function of the *DMwR* package enables the simultaneous addition of new examples for the minority class based on k nearest neighbors and interpolation and under-sampling the majority class examples. However, because *SMOTE()* only recognizes the class with the fewest observations in the dataset as the minority class, it is executed multiple times to achieve a more balanced sample number among all classes, as shown in Table 5-7. To realize this, different over-sampling rates were applied to different classes based on the original sample size of the training dataset. For example, personal business was over-sampled by a rate of approximately nine. In other words, for each original sample in the training dataset, nine additional synthetic instances of personal business activity were generated using the random interpolation between targeted samples and their five nearest neighbors.

Table 5-7. SMOTE augmented samples

	Work	Education	Leisure	Shopping	Eating	Social	Personal	Work-related	Other
Original training samples	5642	2474	257	937	299	603	401	243	386
SMOTE augmented samples	5690	2584	3027	4568	3371	3570	3600	2583	2983

Using the SMOTE-augmented samples, the average prediction accuracy of the regularized MNL model (MNL7) decreases to 58.03%, which is worse than most of the other models. However, as shown in Figure 5-6, the decrease in performance is mainly caused by the reduced prediction accuracy for study and education activities. For the minority activity types, such as leisure, eating, work-related and other activities, which are poorly handled by other models, significant improvements are observed in the results of MNL5, which uses the sample-balancing approach. For example, the predication accuracy for leisure increased from 8.08% in MNL6 to 22.39%, representing an almost three-fold improvement. Even for other activity types, such as shopping, social and personal business, there are still moderate improvements in the prediction accuracy. The experiment shows the effectiveness of the SMOTE approach in providing better classification results for the activity types with fewer observations.

5.5.5 Effect of the transition of activity types

By focusing on the sequence of transit trips as opposed to individual trips and by accounting for the transitions among different activity types, the CRF model is able to achieve the prediction accuracy of activity types to 73.8% on average in comparison with 64.35% for MNL6 and 53.05% for MNL7, based on a 5-fold cross validation. In comparison with the MNL models, the improvement in accuracy is mostly a result of the better performance in classifying activities such as work, education, shopping and social activities. For example, the prediction accuracy of shopping activities is 49.5%, and that of education activities is 92.11%. However, the improvement in the overall accuracy is at the expense of a worse performance for the minority classes such as leisure, eating and work-related activities. Obviously, the issue of imbalanced classes is also salient in the CRF model. However, because the unit of analysis in the CRF models is an activity chain or trip chain, common sample-balancing approaches, such as SMOTE, are not directly applicable. One potential strategy that may address this issue is to introduce distribution-sensitive prior information of classes into the model as the prior belief of the distributions of classes, which permits samples to have a balanced impact on the learning process (Song et al, 2013). An experiment using this approach will be included in future research work.

Figure 5-8 shows the coefficients of the activity transitions from the linear-chain CRF model. The higher the value, the greater the probability that the transition between activity pairs can be observed in the training data, keeping all other variables controlled. It appears that it is common that people engage in leisure activities (including sports) after school or in shopping activities after some social gatherings. In contrast, it is rarely observed that people study after social activities or go to work/school after finishing personal business. Again, note that the activities in the learning model are those participated by means of transit. At this point, the transitions among activities are considered stationary given the limited size of the training data. It is possible to make the transitions time-varying if additional training data are made available.

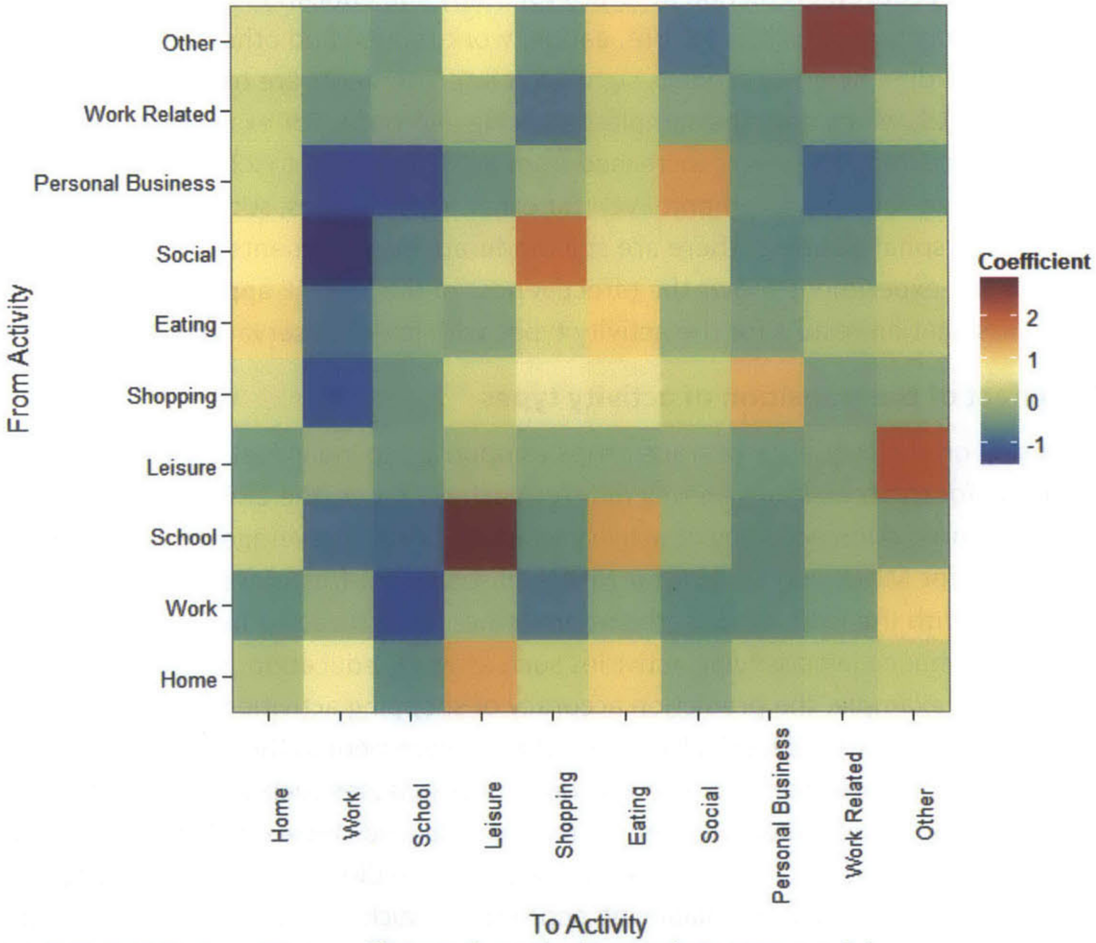


Figure 5-8. Activity transition coefficients from the linear-chain CRFs model

5.6 Summary of Findings

In this chapter, a group of activity learning models (i.e. MNL and CRFs models) are tested on different model specifications and structures. For MNL models, it appears that incorporating more comprehensive and spatially detailed built environment measures rather than zonal statistics into the model help to improve the goodness of fit moderately. Moreover, as shown from model MNL2 to model MNL3, when the built environment measures are changed from stationary ones to time-varying ones, there is a notable improvement in the classification accuracy as well as the goodness of the fit of the model. Nevertheless, when the time-of-day variables are present, the benefit of incorporating temporally varying built environment measures becomes insignificant, as shown by the comparison between model MNL4 and MNL5. The possible interpretation is that the impacts of many temporal covariates are mixed up together with the dummy variables of time periods. More empirical studies are needed to identify the variables mattered temporally and the appropriate temporal scales of the variables. Meanwhile, considering the imperfect information of business open hours is used to measure the spatiotemporal distribution of establishments, it can be anticipated that more improvements will be achieved with better temporal data.

To address the interdependencies among the activities, a linear-chain conditional random fields (CRFs) model is tested to estimate types of activities according to the conditional probabilities of observing a sequence of potential activities based on the sequences of observed trips and destination locations. The result of the CRFs model shows a notable improvement in the overall classification accuracy due to the inclusion of Markovian transitions among activities as additional variables in the model. However, because the CRFs model selects the most probable activity chain based on the sequences of observed variables as the predicted activities of the observed transit trip chain, which tend to bias toward activities with more samples, the classification accuracy rates of the minority activity types are generally worse in comparison with those from the MNL models.

The methodology discussed in this chapter is still cross-sectional in nature. Alternatively, the types of activities may also be learned based on the time and frequency of visits over a long period of time. The trip and location histories are especially valuable for the activity detection when considering certain activities like shopping or personal businesses are planned weekly rather than daily. Nevertheless, this requires the training data spanning over multiple days. In this study, the training data come from a one-day household travel survey, which make it difficult to incorporate the repetitiveness and variability of trips into the activity learning models. In Chapter 6, we describe an approach to extract home location from week long transit trip logs from the available EZ-Link data.

Chapter 6. Activity Inference and Visualization

6.1 Introduction

With the activity learning models estimated in Chapter 5, the results can be applied to the EZ-Link data to infer the hidden activity types and to explore the activity-travel patterns of millions of transit passengers and their interactions with the urban built environment. Because the in-home activity following the *"returning home"* trip is not included as an alternative activity type in the multinomial logit (MNL) models, the first task of the activity inference is to identify the approximate home locations of the transit passengers. The identified home locations enable to differentiate the returning home trips from other transit trips in the EZ-Link data. At the same time, we can compute the distance of alighting stations to the approximated home locations, which is one of the explanatory variables needed to estimate the probabilities of different activity types.

In the proposed analytical framework, activity inference is not the end of the process. Instead, it should be considered as a starting point for a more profound exploration and reasoning process to drill down to the human activity-travel trajectories at disaggregated level, to identify the irregular patterns that are not well predicted by the models and to extract more knowledge on the correlations between the complicated activity-travel patterns and the urban built environment. As we start to tackle with rapidly accumulated spatiotemporal information, and more complicated models, the conventional thematic mapping or geoprocessing procedures that relying on much manual work on post-processing effort become both too cumbersome and less sufficient to present the level of detail that we want to explore. There are imperative needs for new visualization approaches to account for these challenges.

In this chapter, we first discuss the algorithms used to extract the home locations of the EZ-Link card holders. Based on the estimated learning models and home locations, the activity types of the transit trips are inferred. Then, visualization examples are prototyped to illustrate the spatio-temporal dynamics of individuals' movements and activities, as well as their collective impacts on the catchment areas surrounding the transit stations. The built environment measures are included in the visualization to describe and differentiate urban places, and to help interpret the context-dependent choices of people.

6.2 Home Location Extraction

It has been noticed by researchers that the individual movement trajectories revealed by the multiple-day urban sensing data contain the locational information that could help to extract the important features of activity-travel patterns. Many sophisticated rules and algorithms have been developed to extract the mobility information from urban sensing data, although the exact implication of the recorded positions may not be clear. For example, Kim et al (2006) were able to identify the movement speed, pause times, destination transition probabilities, and waypoints between destinations by examining the WiFi traces of wireless users and explicitly define the physical properties of movement such as pause and flight. Rhee et al. (2008) has demonstrated that from GPS trajectories it is possible to extract information like moving direction, velocity and hotspot. Hariharan and Toyama (2004) also adopted a set of heuristic rules to parse locational information like stay, destination, trip and path from the GPS data. In their study, a stay is a single instance of place that an object spent some time. A destination is defined as a place that one or more objects had a stay. A trip is the movement between two successive stays and a path is composed by a set of trips connecting destinations. These characteristics of human mobility are commonly used to fit mobility models such as random walk models and Levy Walks.

In comparison with the other types of urban sensing data such as GPS or cell phone data, the information collected from the transit transaction smart card has the advantage of that all spatial information is sensible. It is either the boarding station or the alighting station of the trips. Considering the density of transit stations in Singapore and the transit trip records collected over multiple days, it is likely the spatial locations of homes can be extracted approximately if the spatial-temporal rules of the home-based or home returning trips can be detected. Because the EZ-Link data only record the positions of the transit passengers at the level of transit stations, the primary task of determining the home location of a passenger is to find out the transit stations surrounding the home location that the passenger can directly access from home by walking. Here, we define the concept "home station" as the transit stations used by a given transit rider that are more accessible from the home by walking than by connecting from other transit stations. Here, the accessibility is a composite measure that not only takes into account the distance but also considers the time, the cost, and the convenience of connections. Home station detection is not as easy as it appears to be. Note that a home is usually served by multiple home stations with which the transit passenger can travel to different destinations via different transit routes. Also, it is important to realize that the mostly visited transit station in the observed EZ-Link trips of an individual is not necessary his home station. For example, the MRT station close to the workplace may be visited more than the nearest bus stop to home if there are multiple "home stations" that divide the home-based trips and the returning home trips. In addition, the starting station and ending station of a home-based tour can be different because passengers may select the more frequent bus route if multiple transit routes are available between origin and destination.

To account for these complexities, a hierarchical nesting of heuristic rules are applied to extract home locations. The degrees of difficulty for detecting home locations are different for different passengers, depending on the volume of trips contained in the EZ-Link data. The more transit trips made by a passenger, the greater the likelihood that the approximate location of home can be detected. Therefore, as opposed to formulating generic rules that try to match with all card holders, it is more appropriate to have complex rules applied to those card holders with sufficient information but have simple rules applied for those who have limited information recorded. In brief, the approach is intended to extract the best possible result for each card holder using the differentiated rules based on the amount of information contained in the data.

The proposed approach to extract home location consists of two steps. In the first step, for each EZ-Link card holder k, a set of candidate "home stations" $\hat{S}_k$ are identified from the 15-day EZ-Link dataset by finding all the stops sufficing one the following conditions:

- Boarding stop of the first transit trip made between 5am and 10am on a weekday;
- Alighting stop of the last transit trip made between 6pm and 2am on a weekday;
- Boarding stop of the first trip on a weekend or a holiday;
- Alighting stop of the last trip on a weekend or a holiday.

These rules help to narrow down the set of potential transit stations surrounding home locations because most transit passengers normally start their daily trips from home and end daily trips at home. But for an infrequent transit user, the origin station of the first transit trip or the destination station of the last transit trip might be irreverent to his home location. Thus, time constraints were placed in the rules to reduce the possibility of falsely considering the stops of a transit trip neither starting nor ending at home as "home stations". According to the HITS 2008 dataset, 87.56% of the first transit trips of respondents are originated from home. But if the starting time of the first transit trip is constrained to the time period between 5am and 10am, the ratio increases to 97.07%. Likewise, if one looks into the last transit trips that end after 6pm, the chance that the trips end at homes is 93.97%. These statistics from the household travel survey provide evidence for the heuristic rules in detecting the "home stations". For the card holders that don't have candidate home stations identified in this step, the most frequently visited stops are used as the surrogates of "home stations". If multiple home station candidates are found, the mostly visited one is then considered as the "primary home station" PS_k. The remaining candidate stations that are within an appropriate distance d_s (e.g. 500m) from the primary home stop are considered as secondary home stations, which are also used, although not as frequent as the primary home station, for home-based transit trips. Figure 6-1 b) illustrates of the identification of candidate home stops from the transit trips recorded in the EZ-Link data.

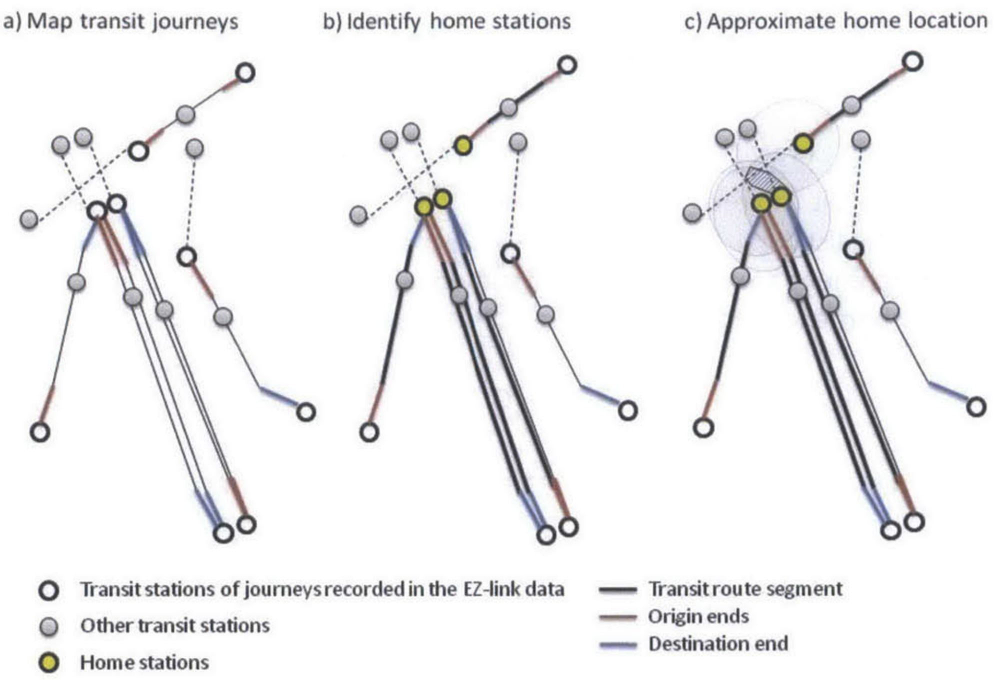

a) Map transit journeys
b) Identify home stations
c) Approximate home location
○ Transit stations of journeys recorded in the EZ-link data
● Other transit stations
● Home stations
Transit route segment
Origin ends
Destination end

Figure 6-1. Illustration of the procedure of extract home locations from the EZ-link data

In the first step of the approach, one or more "home stations" are identified for each EZ-Link card holder k and a "primary home station" PS_k is determined. In the second step, the spatial resolution of the potential home locations is reduced to building level by examining the potential residential buildings h_k surrounding the set of "home stations" $\hat{S}_k$. By combining the topology of transit network and the information contained in the transit trips, we can bind the home locations to a clear-cut geographical area (Figure 6-1 c). For each transit trip starting or ending at a home station, the previous stop and the next stop of the home stations are retrieved. Only those residential buildings that are closer to the identified home stations than to the previous or next stops of the recorded transit routes are considered as the eligible home locations. Then, a building is sampled from the eligible residential buildings in the area. This building is regarded as the approximate home location of the given card holder H_k. To protect the privacy of travelers, the locational information of home is not revealed and only the relative spatial relationships like distances between the destination stations and the estimated home locations are preserved.

A pseudo code of the home location extraction approach is given in Table 6-1. The inputs are the transit trips from the EZ-Link data recording the EZ-Link card id, trip starting time, boarding station and alighting station $TR\ \{c_i, t_i,\ s_i^{boarding}, s_i^{alighting}\}$, as well as the residential building directory B. $first\_trip(t_i, c_i)$ and $last\_trip(t_i, c_i)$ are the functions used to determine whether a trip sufficing one the first trip and last trip conditions mentioned above. The *dist()* function calculate the Euclidean distance between two spatial objects. $s_{\hat{i}}^{alighting+1}$ denotes to the next station of the alighting station in the transit route and $s_{\hat{i}}^{alighting-1}$ denotes to the previous stop.

Table 6-1. Pseudo code of extracting home locations from the EZ-Link data

```
Input: raw EZ-Link transit trip record $TR\ \{c_i, t_i,\ s_i^{boarding}, s_i^{alighting}\}$
   residential building list $B$
   radius of the catchment area around transit stations $d_s$
Output: estimated home location vector $H$

Initialize $k \leftarrow 1, i \leftarrow 1, \hat{S} \leftarrow \emptyset,\ PS \leftarrow \emptyset, H \leftarrow \emptyset, h \leftarrow \emptyset$
While    $k < N$
      $m \leftarrow count(\ c_i = card\ holder\ _k) + i - 1$
     $\hat{i} = i$
While    $i < m$
    If      $first\_trip(t_i, c_i) = TRUE$
         $\hat{S}_k \leftarrow \hat{S}_k \cup s_i^{boarding}$
    Else If  $last\_trip(t_i, c_i) = TRUE$
         $\hat{S}_k \leftarrow \hat{S}_k \cup s_i^{alighting}$
    End
         $i \leftarrow i + 1$
   End
      $PS_k \leftarrow \widehat{S}_k^{\max\_count}$
     $h_k \leftarrow B\ \ s.t.\ dist(B, PS_k) \le d_s$
   While $\hat{i} < m$
     If   $s_{\hat{i}}^{boarding} \in \widehat{S}_k$
        $h_k \leftarrow h_k\ s.t. \begin{cases} dist\left(h_k\ ,\ s_{\hat{i}}^{boarding}\right) \le dist\left(h_k\ ,\ s_{\hat{i}}^{boarding+1}\right) \\ dist\left(h_k\ ,\ s_{\hat{i}}^{boarding}\right) \le dist\left(h_k\ ,\ s_{\hat{i}}^{boarding-1}\right) \end{cases}$
     Else If $s_{\hat{i}}^{alighting} \in \widehat{S}_k$
        $h_k \leftarrow h_k\ s.t. \begin{cases} dist\left(h_k\ ,\ s_{\hat{i}}^{alighting}\right) \le dist\left(h_k\ ,\ s_{\hat{i}}^{alighting+1}\right) \\ dist\left(h_k\ ,\ s_{\hat{i}}^{alighting}\right) \le dist\left(h_k\ ,\ s_{\hat{i}}^{alighting-1}\right) \end{cases}$
     End
         $\hat{i} \leftarrow \hat{i} + 1$
    End
      $H_k \leftarrow sample(h_k, 1)$
End
```

6.3 Activity Inference

6.3.1 Inference using MNL models

Once the same set of explanatory variables are computed and assembled for the transit trips in the EZ-Link dataset, the estimated activity learning models can be applied to infer the hidden activity types associated with the transit trips. MNL learning models predict the probability of each activity type based on the pertinent trip information and the built environment measures of the areas surrounding the destination station according to the equation 5-1. Rather than choosing the activity type with the largest predicted probability, a Monte Carlo simulation method is used to decide the activity types. This avoids the situation that the minority activity types are absent in the predicted results because their probabilities are usually low. Meanwhile, it adds the stochasticity to the results, which to some extent reveals the randomness in the real world that is not captured in the learning models.

6.3.2 Inferring Activity Types using CRFs models

Considering now the parameters of a conditional random fields (CRFs) model have been estimated from training dataset, which means the estimates of the activity transition rates and the context dependency coefficients as well as the initial activity state distribution are known, the problem becomes how to infer the unknown activity types in the form of activity chain. Under the framework of CRFs, we need to find out a single best activity sequence that can maximize the joint probability of hidden activity types conditional on observed sequences of trips, locations and temporal factors. This problem can be effectively solved by the forward-backward algorithm, which is also referred to as Viterbi algorithm (Rabiner, 1989). Although the estimated CRFs model has a higher overall accuracy in predicting activities, it has been shown in the Chapter 5 that it performs worse in classifying the minority activity types like *leisure* and *personal business*. Therefore, in the visualization examples discussed below, we use the activities inferred from the regularized MNL model (MNL6).

6.4 Examples of Exploring and Visualizing Activity-travel Patterns

The abundant information contained in urban sensing data and the intertwining relationships of multi-facets of activity-travel patterns make it difficult to examine the data, the models, and the prediction results. The contextual information adds to the complexity of exploration and reasoning process. Therefore, there is an urgent need for effective visualizations to assist in understanding activity-travel patterns and interpreting the spatiotemporal associations among activities, trips and the urban built environment. Ideally, visualization examples should not only provide an overview of the data, but also allow exploration at a variety of geographical and temporal scales, because both the overall urban dynamics at the city level as well as the built environment at the neighborhood level are important for comprehending the behaviors and choices of individuals. Moreover, the visualizations must present individuals' activity trajectories together with the built environment information. In this study, three visualization examples are prototyped to demonstrate the strength of coupling EZ-Link data with built environment information to acquire insights on individual activity-travel trajectories, urban space usage, and the interactions between human behavioral choice and urban environment. By combining the analytical reasoning with visual exploration, it is expected to not only facilitate the detection of methodological issues, but also stimulate the knowledge discovery, which can be used to augment the analysis process.

6.4.1 Single Transit Trips

The large volume of information contained in the EZ-Link data provides opportunities for researchers to focus on the patterns that are relevant but less emphasized in the convention studies. One example is the pattern of transfers as a result of the choices of transit paths (see Chapter 2 for more description). Another example is the single transit trips made by travelers during any of the observed days.

In the process of exploration, we found around 24.9% of the transit riders only made one transit trip in a day. These single transit trips account for around 10.7% of the total transit trips. That means these travelers took transit system for a one-way trip and might use other transportation modes for the other legs of the trips if their trips were round trips or looped tours. Understanding these single transit trips can provide insights for the choices of multi-mode transportation for people's daily travel chain. The single transit trips are paired by origin and destination stations and are plotted on Figure 6-2. Each line in Figure 6-2 represents an OD pair flow with observed single transit trips. The width and the levels of opaqueness of the lines are proportional to the magnitude of the single transit trips. The lines are coded in three colors with the orange ends connecting to boarding stations and blue ends connecting to alighting stations.

The map is helpful in identifying places that produce or attract such trips. For example, the places like Changi airport and the Singapore-Malaysia border attract and generate significant amount of single transit trips because travelers just come from or prepare to travel abroad. These are all one-way trips. Jurong point mall appears to be another strong attraction for the single transit trips. The places like harbor front, downtown and the neighborhoods around the Jurong Point seem to be the main sources of the single transit trips. To understand why some places attract or generate such single transit trips, it is useful to explore the trips side by side with the urban form information of areas surrounding the origin and destination transit stations. Next, we dig into the example of Jurong point mall to demonstrate the exploration and reasoning process that can be facilitated by visualization examples.

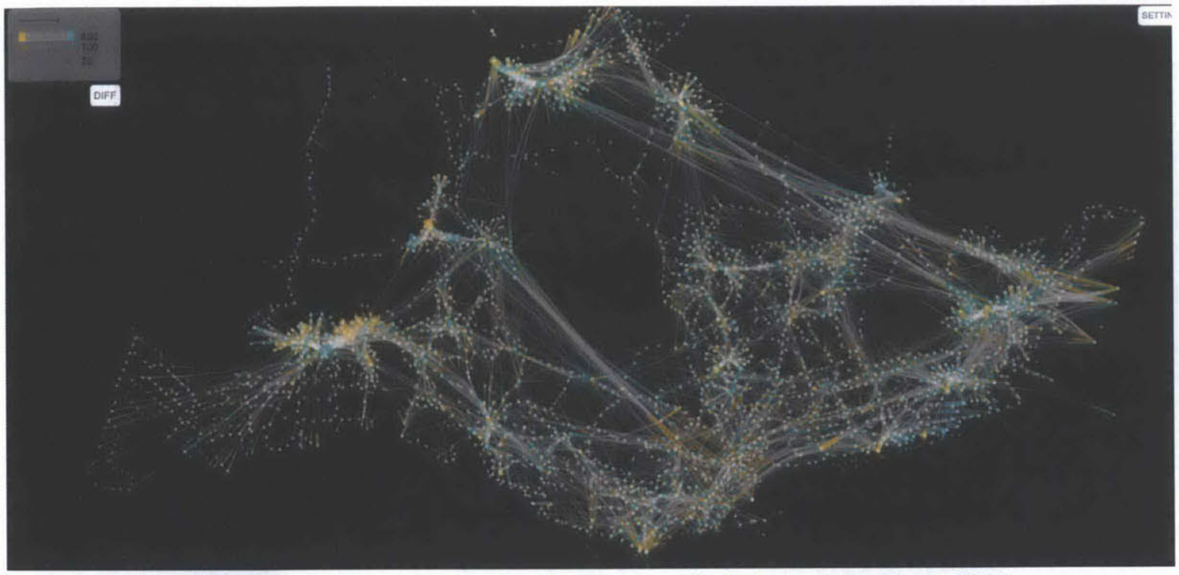


Figure 6-2. Spatial pattern of single transit trips in the EZ-link data on April 13, 2011.

When zooming into the Boon Lay bus interchange, the main bus stop next to the Jurong Point mall, it can be seen from Figure 6-3 a) that the single transit trips are mainly originated from the transit stations in the nearby neighborhoods and Jurong west as shown by the purple and red lines which indicate a relatively high volume of transit trips. In contrast, the number of single transit trips from the stations east to the Boon Lay bus interchange is generally low, mostly ranging from 1 to 9.

Figure 6-3 c) shows a parallel coordinate plot intended to depict the surrounding built environment of transit stations by a group of selected indicators. The colors of lines are corresponding to the volume break down of the single transit trips. The thick red line denote to the origin station that generate the maximum volume of single trips to the Boon Lay bus interchange. The thick blue line represents the Boon Lay bus interchange. As shown in Figure 6-3 c), it is relatively clear that the areas that generate more single transit trips to the Boon Lay bus interchange are those with less fragmented land use types (high contagion index), low presence of businesses indicated by relative low business entropy values and retail amounts, low node centrality in terms of transfer degree and strength, high industrial areal ratio and mid-high residential areal ratio. Coupling the map a) and the parallel coordinates c), a number of possible hypotheses might be contemplated to interpret the great number of single transit trips ending at the Boon Lay bus interchange. First, the shuttle services provided by the Jurong Point mall and other agencies enabled travelers to avoid the use of transit service for their returning trips. Second, there are four taxi stands around the Jurong Point mall, which make it much easier for travelers to get a taxi for the ride home when they may be burdened with shopping bags. Third, the relative low presence of retail and other businesses in the Western part of Singapore made the Jurong point not only a

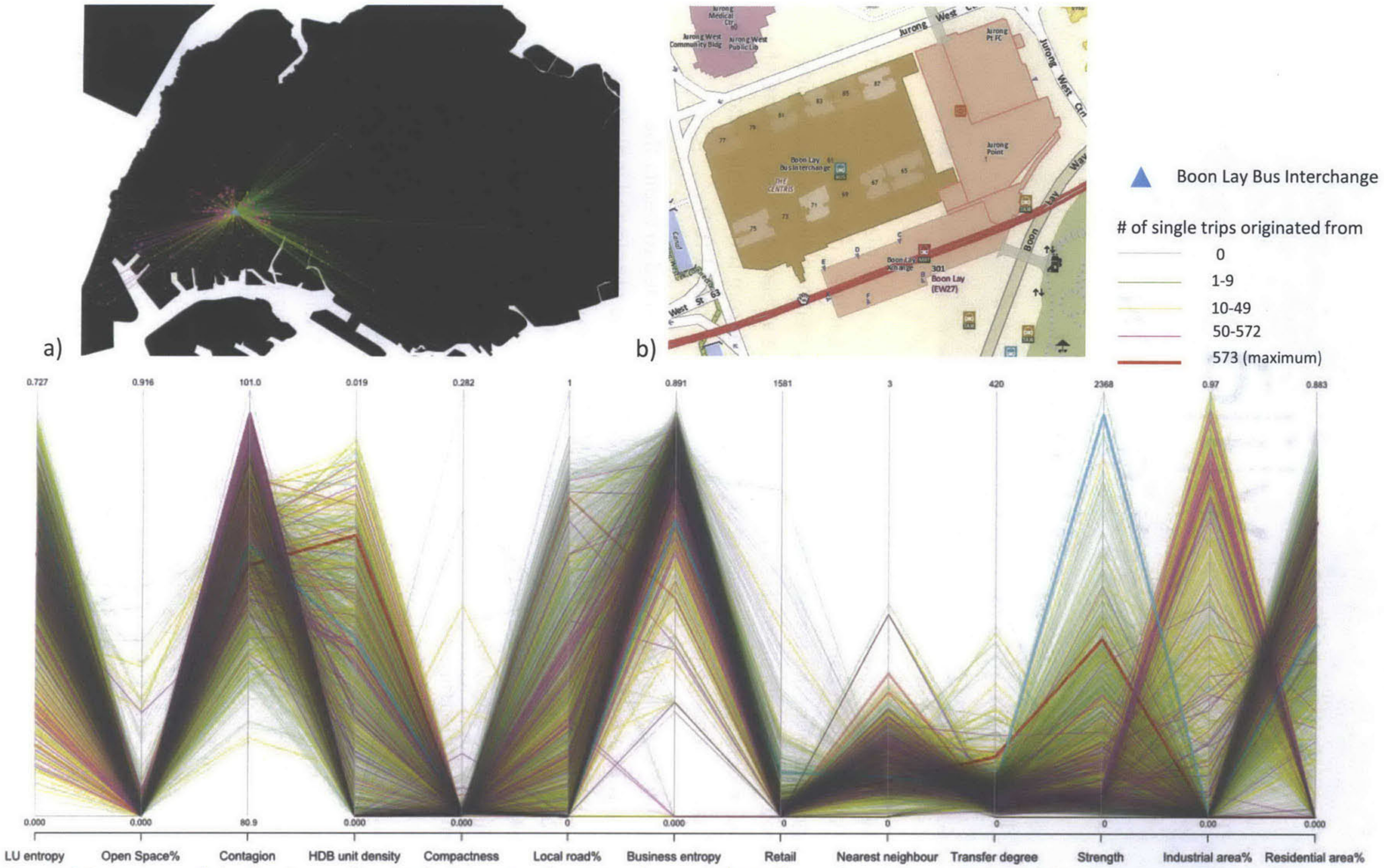


Figure 6-3. Pattern of single trips ending at the Boon Lay bus interchange. a) OD flow of single transit trips to the Boon Lay bus interchange. b) Screenshot of the Boon Lay bus interchange from onemap.com. c) Parallel coordinates of the selected BE measures of stations.

main attraction of various types of activities, but also a meeting point for people from different origins. Conceivably, a considerable number of returning trips might be accomplished by taking the cars driven by other travelers. These hypotheses are worth to be explored further as additional information become available. To sum up, the significance of the single transit trips call for attentions to underestimated modes in the transportation planning like shuttles, company vehicles, taxis etc.

6.4.2 Individual Activity-Travel pattern

The individual movement patterns in the EZ-Link data contain plentiful information to be explored. As an example, Figure 6-2 and 6-3 plot the transit trips that a subject made between April 11, 2014 and April 15, 2014 and the inferred activity types side by side. The subject is selected because he is among the passengers who made the most transit trips in the selected time period in the EZ-Link dataset. To protect privacy, the names of the stations are replaced by implicit station IDs and the coordinates of the locations are tweaked to make it difficult to identify the actual locations.

To represent the rich information of activity-travel pattern, a variety of graphical encodings are used in the visualization. The transit trip trajectories displayed in Figure 6-2 use different line styles to represent different days of the week. For example, the solid lines correspond to the trips made on Monday and dotted lines correspond to the trips made on Wednesday. In addition, for the trips between same origin and destination but were made on different days, offsets were added to the trip lines to show the repetitiveness of the trips. In Figure 6-2 b), one can see multiple parallel lines connecting among transit stations, which suggests the regular pattern of the trips. The types of transit stations are differentiated by the colors, with blue circles representing origins and yellow circles representing destinations. The sizes of the circles are proportional to the number of transit trips observed to originate from or designate at the given stations. The stack bars next to each yellow circle (i.e. trip destination) indicate the inferred probabilities of engaging different types of activities under the given circumstance. The number of bars is in line with the number of trips made between the same OD pair.

One can see that the predicted results of activity types are different for the same trip routine made on different day. This is because the contexts of the transit trips have changed due to the variation of temporal factors like weather, transit services and time of day. However, this gives rise to the question about whether the regular routine trips should always have same trip purposes. If the answer is yes, then the multiday regularity of the trips should be built into the activity learning models as an additional variable. Because the HITS data only cover the trips of the respondents on a single day, the question is left to be answered in the future.

As shown in Figure 6-4, the activity space of the selected subject concentrates in two areas connected by daily routine trips. One is at the northern part of Singapore, where the estimated home is located (Figure 6-4 b). The trips ending at the stations surrounding the estimated home locations are mostly returning home trips. The other is at the central part of Singapore (Figure 6-4 c), where the subject appeared to be very active in engaging social, eating and shopping activities. One may speculate that the workplace or school of the subject is in the second area because of the daily routine trips. However, the activity types inferred from the learning models suggest that the purposes of trips ending in the area are not very likely to be working or education. This motivates me to explore further by considering the temporal distribution of the trips.

Figure 6-5 shows the temporal pattern of the trips made by the selected subject. The X axis denotes the time period between April 11, 2014 and April 15, 2014 grouped by every two hours. The Y axis displays the station IDs of the boarding stations (blue dots) and the alighting stations (red dots) of the observed trips. The black lines linking pairs of boarding and alighting stations are the transit trips. It is clear that most of the transit trips were made at late afternoons (after 3pm) or in the evenings. No trips were made in the mornings.

Moreover, the subject was very active between 3pm to around 11pm, because he made around 7 transit trips on average on the observed days. Some trips are quite regular temporally. For example, the trips from station 37 to station 32, which can be observed at about the same time (3pm) on Monday, Wednesday and Thursday. The time and the levels of activeness of the observed transit trips both imply the purposes of routine trips from the estimated home location to the major activity destination are unlikely to be working or education. Instead, it is possible that the subject's parents or other important relatives were living in the second area, which may explain the regular transit trips made between the estimated home location and the major activity destination area.

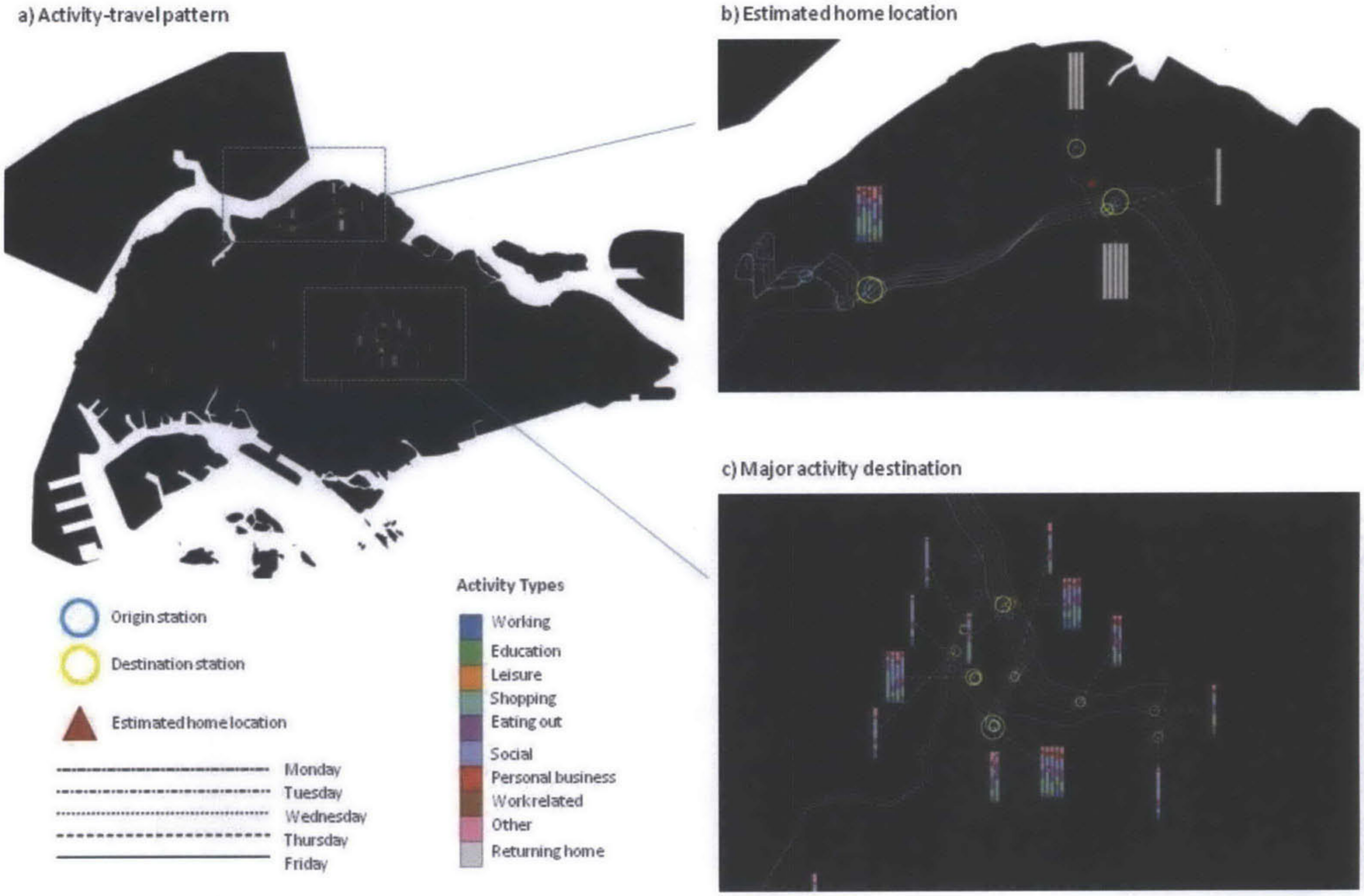


Figure 6-4. Visualizing individual activity travel pattern.

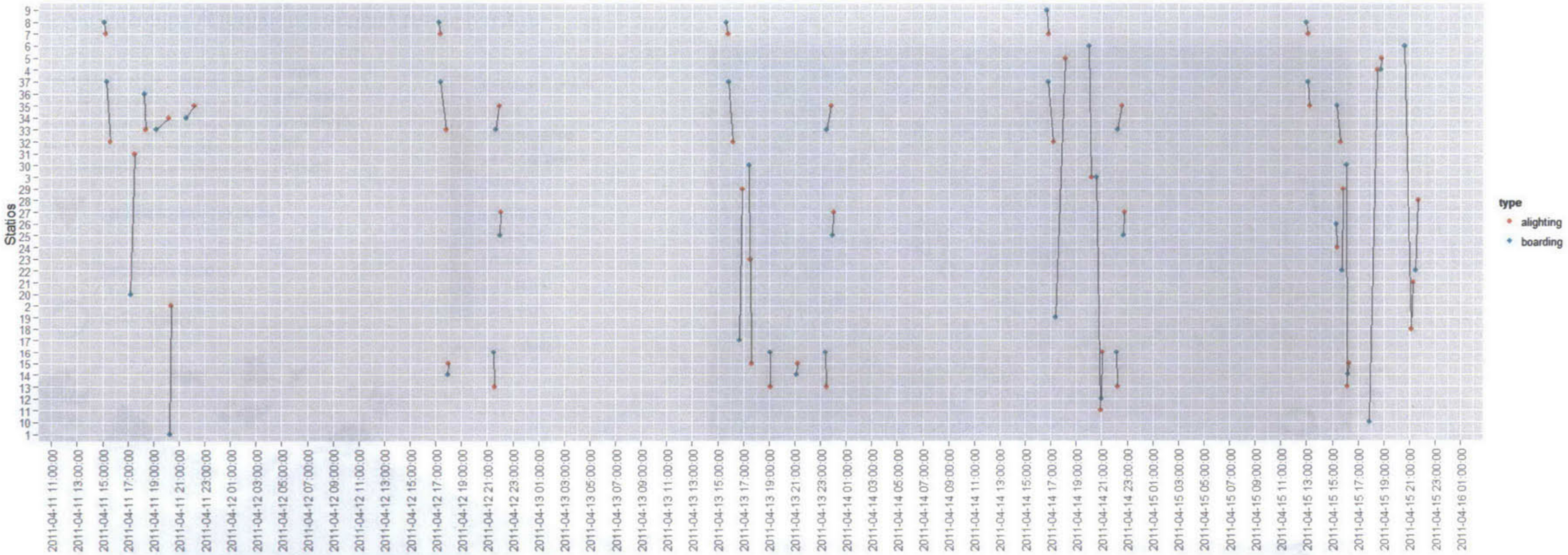


Figure 6-5. The temporal pattern of transit trips

6.4.3 Activity Landscape of places

Beyond revealing the activity-travel patterns of individual transit passengers, it is informative as well to use the inferred daily activity types to depict the activities profiles of places in the Singapore. Characterizing place is not only crucial for urban design but also important for location choice models in transportation and urban modeling. Traditionally, researchers rely on static land use data to differentiate locations. With the EZ-Link data and the activity learning models, it is possible to explore how urban spaces are utilized by transit passengers over time in a day. Figure 6-6 a) shows the magnitudes and the types of transit-oriented activities occurring in the areas surrounding the Chinatown MRT station throughout a weekday. A trellis display chart is presented to show the built environment measures the location (Figure 6-6 b).

In Figure 6-6 a), the estimated activity profile of the transit passengers alighting at the Chinatown MRT station is represented by a wind rose chart. Each bar column around the inner circle corresponds to the volume of passengers boarding or alighting at the station in every 10 minutes. The grey bars pointing inward to the circle center describe the volume of boarding passengers. The stack bar columns stretching outward represent the aggregated probabilities of different activity types of the alighting passengers. It is clear from the figure that most of the trips ending at the Chinatown MRT station during the morning peak hours were *working* trips. There were at most around 90 passengers per minute arriving in the area by the MRT trains. As time passes by, a higher proportion of transit passengers came to the area for the purposes of *shopping*, *eating* or *personal business*. At around 6pm, the volumes of both incoming trips and outgoing trips attained another peak. The main purposes of the incoming trips were *shopping*, *eating* and *social*.

The trellis display chart shown in Figure 6-6 b) displays the built environment measures of the place from four dimensionalities: land use, building layout, distribution of establishments and transit network connectivity. The values of the measures showed in the chart have already been standardized. For each selected built environment measure, the chart represents the values corresponding to the Chinatown MRT station in red dots, the minimum value in yellow dots, the maximum value in blue dots, as well as the values of other stations in grey dots. This enables us to know the relative position of the target place among the catchment areas surrounding all transit stations regarding particular aspects of the built environment. For example, the catchment area of the Chinatown MRT station generally has high land use diversity, but low presences of either park or open space. The standard deviation of the building footprint sizes is comparatively low, indicating relatively uniform building types in terms of the footprint size. Besides, the area has a high mix of the types of businesses and the MRT station has a high degree of transfer connectivity with the surrounding transit stations. Compared with the other locations, the businesses within the Chinatown area also tended to be more clustered as indicated by the nearest neighbor index. These measures help to interpret the attractiveness of the area to *shopping*, *eating* and *social* activities. Overall, these figures enable a better understanding of the locations within the Singapore and can provide better support for urban management and location choice modeling.

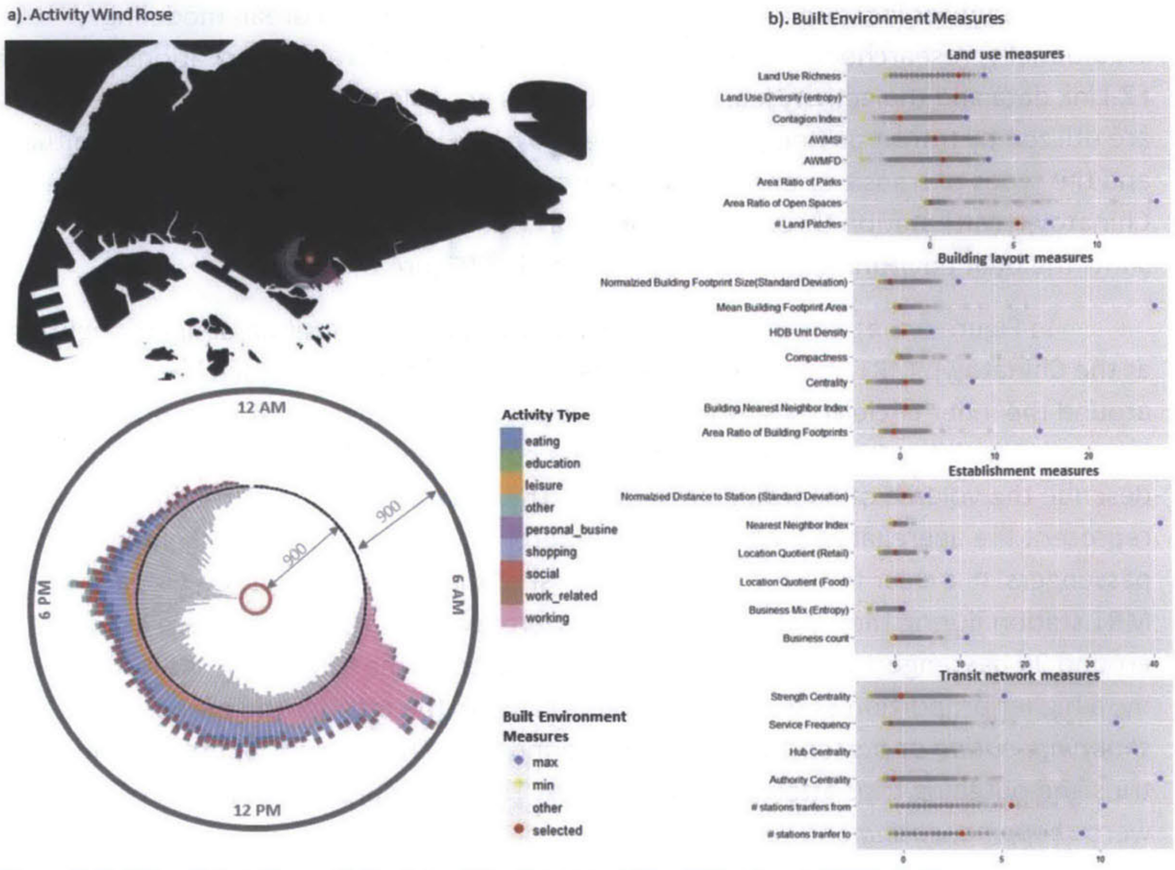


Figure 6-6. Visualizing the activity-travel landscape of the Chinatown MTR station.

6.5 Implications to activity-travel research and planning practice

This study investigated an analysis framework to incorporate the EZ-Link data into the activity-travel behavioral study with the assistance of improved measurement of context, in particular the urban built environment, by drawing on the spatiotemporally detailed data from multiple sources. Central to the proposed framework is the development of learning models to recognize the unknown activity types of the observed transit trips based on the estimated correlations derived from a household travel survey. It is crucial for policy makers and planners to know the activities associated with trips because activities explain the question why trips are made. In addition, recognizing various activities in a place provide rich information on interpreting how urban places are used. In this sense, the framework shows some potentials of using EZ-Link data to complement our understandings of activity-travel behaviors and urban dynamics.

Generally, EZ-Link data can extend the scope of the questions that can be investigated. The large sample size might help to expose the patterns that are not easy to be discovered from traditional surveys. We present two examples: transfers among transit stations and single transit trips. Both examples are important to consider from the perspective of mode and path choices in the travel demand forecasting but are not well addressed in the conventional activity-based modeling approach. On the one side, transfer is critical for the path choice of travelers. The place of intermediate stop may be considered in the process of activity scheduling for activities like grocery shopping or personal banking service. On the other side, understanding the transfer preference of travelers help to better organize the transit services and therefore reduce operational cost by removing or adjusting under-demanded route or link segments. Traditional mode choice models usually predict the same transportation mode for trips between the same origin and destination. Single transit trips indicate the importance of considering multiple modes and other contextual factors in the mode choice model since an unignorable portion of round trips might be made by different modes for each way. For example, some transit riders may switch to taxis when raining or accumulating shopping bundles. Similarly, additional insights on the travel patterns might be exposed by the EZ-Link data, such as the choices of rapid transit routes, trunk lines and other feeder lines. The difference in path choice can be analyzed to detect the problems with underused routes such as long detours and to help improve the efficiency of transit system by addressing the detected issues.

The second aspect of contribution results from the longitudinal coverage of EZ-Link data. At the individual level, we would be able to observe the regularity and variability of weekly transit trips, and detect the frequently visited activity destinations. This is important because most people plan and schedule their activities over multiple days. More significantly, the longitudinal data enable us to capture the behavioral changes of people in response to the land use change such as the conditions that a new subway line is in operation. It is also possible to track the evolution of activity spaces of people potentially shaped by their spatial knowledge. This can be useful to help determine the choice set of destinations in the destination choice model.

From the perspective of land use and transportation planning, the exploration of the transit trips recorded in the EZ-Link data and the correspondingly inferred activities help to reveal the usage of urban spaces for different purposes at different time of day and day of week. Meanwhile, for a particular place, we are able to find out where the visitors come from via transit service. This enables us to identify the spatial distribution of potential customers of a local mall or the spatial range of attractiveness of a community center. It is also possible to estimate the impact on an existing mall if a new mall is opened in the next town. Of course, this type of analysis needs to couple EZ-link data with spatio-temporally detailed urban built form information.

Further, EZ-link data can be used to test and validate activity-based models. This is an important research topic but beyond the scope of this study.

In summary, EZ-link data have the potential to enrich the understanding of the context dependencies of human activity-travel behaviors by assisting the detection of changes or irregular patterns. The possible causes of the changes might be distinguished, and the exploring and reasoning process can be facilitated, when coupling EZ-Link data with other spatio-temporally detailed data. This is helpful in expanding the knowledge base to support more realistic activity-based modeling and planning practices.

6.6 Summary

In this section, we demonstrated the possibility of detecting home locations based on the multiple-day trip records in the EZ-Link data and of inferring the hidden activity types using the estimated learning models. In addition to the home location, it is also likely to detect other anchor locations of individual travelers such as workplaces and schools based on the regularity of the trips, if the EZ-Link data are available for more days. The visualization examples illustrates the types of analytical reasoning that the proposed visualizations facilitate by presenting the spatiotemporal details of the transit trip trajectories alone with the predicted activity types of individuals. For a specific place, the visualization examples can be used for pursuing a wide range of inquiries (e.g. what type of people visit this place for what purposes at different time of day). This kind of place-based inquiry will contribute to our understanding of destination choice behaviors and location-activity interactions.

One important issue that must be confronted is privacy. As the EZ-Link logs accumulate over days, the spatially explicit information on the trajectories of individuals make it hard to anonymize subjects completely. Although the raw data access is strictly limited, it is important that the processed data are presented with appropriate attention to privacy by either aggregating the information or transforming the spatial and temporal tags of the data to make it impossible to identify any subject or the actual locations they visited (Gutmann et al., 2008). In this study, offsets are added to the transit stations and estimated home location to make it hard to recognize the actual locations from the visualization example.

Chapter 7. Discussion and Conclusions

The study of human activity-travel patterns for transportation demand forecast has evolved a long way in theories, methodologies and applications. However, the scarcity of data has become a major barrier for the advancement of research in the field. At the same time, the proliferation of urban sensing and location-based devices generate voluminous streams of spatio-temporal registered information. These urban sensing data such as the EZ-Link data contain massive information on urban dynamics and individuals' mobility. For example, EZ-Link data reveal the places that transit passengers visit at different times of day. As tempting as it appears to be, the incorporation of these urban sensing data into activity-travel study remains a big challenge, which demands new analytics, theories and frameworks to bridge the gap between the information observed directly from the imperfect urban sensing data and the knowledge about how people use the city. This study represents a step toward this objective.

We propose a framework of analysis that focuses on the recurring processing and learning of voluminous EZ-Link data flows in juxtaposition with additional auxiliary spatio-temporal data, which are used to improve our understanding of the context of data. The framework consists of an ontology-based data integration process, a built environment measurement module, an activity-learning module and visualization examples that facilitate the exploration and investigation of activity-travel patterns. The ontology-based data integration approach helps to integrate and interpret spatio-temporal data from multiple sources in a systematic way. These data are used to reconstruct the context under which the travelers made their transit trips. In particular, a set of spatial metrics are formulated to characterize urban built environment of the trip destinations. In order to understand why people make trips to destinations, we need to have a sense about the possible activities associated with trips. Therefore, an activity learning module is developed to infer the unknown activity types from millions of trips recorded in transit smart card transactions in Singapore by learning the context dependent behaviors of travelers from the traditional household travel survey. The learned activities not only help the interpretation of the behavioral choices of transit riders, but also can be used to improve the characterization of urban built form by uncovering the likely activity landscapes of various places.

Although different modules of the framework are loosely coupled at the moment, we have tried to pipeline as much of the process as possible to facilitate efficient data processing and analysis. This allows researchers and planners to keep track of the evolution of human activity-travel patterns over time, and examine the correlations between the changes in activities and the changes in the built environment. The knowledge gained from continuous urban sensing data will certainly help policy makers and planners understand the current state of urban dynamics and monitor change as transportation infrastructure and travel behavior evolve over time.

In addition to incorporating new data, the framework has feedback loops that facilitate the refinement of each module, such as better measurement of urban form, and the improved learning models. Therefore, the knowledge and ideas generated in the process of analysis are readily to be applied and tested.

7.1 Major Findings

Through the exploratory analysis, we found the overall transit travel patterns revealed by the multiple day EZ-Link data present regularity in both spatial and temporal dimensions. However, there is considerable variability at the level of trips as evidenced by the low trip repetitiveness rate within the observed week. This suggests that individuals have different choices of activities and destinations on different day of the week, which necessitates the investigation of multiday activity-trip pattern and the introduction of spatial detailed urban form measures to infer the purpose of trips.

In order to support finer analysis and modeling efforts of the EZ-Link data, spatial and temporal data relevant to the urban built environment are gathered from a variety of sources including government agencies, web services and crowdsourcing. However, these datasets from heterogeneous sources take on different structures, formats, naming conventions and data qualities, which make it difficult to merge into an integrated dataset. To overcome the barriers for data interoperability and integration, we demonstrated an ontology-based approach to integrate multiple datasets by matching the schema from each local dataset to a global schema, i.e., a task-specific ontology. Routine processing procedures are developed for the cleaning, evaluating and fusion of multiple datasets. These procedures enhance the spatio-temporal detail and attribute richness of a city's built form and socio-economic activity in ways that can be helpful for many aspects of urban planning and urban management. In our case, we focus on using the data fusion results to improve our interpretation of EZ-Link data.

The framework also facilitates the processing of the spatial-temporal detailed dataset and the production of a comprehensive set of spatial metrics to measure the transit station centric urban built environment from four dimensions: land use composition, building and street layout, spatial distribution of businesses, and transit network typology. In particular, the measures of transit network are included to gauge the centrality and connectivity of a station in the network. The transit network analysis is augmented by the transfers among stations observed in the EZ-Link data. Presumably, the destination choice may involve the consideration of accessibility of the destination, especially when multiple trips will be made. However, according to the results of the MNL activity learning models, the measures of transit network and transit service are mostly not significantly correlated with the activity types. This is probably due to the correlation between transit network measures and other station-centric built environment measures. Otherwise, it may imply the accessibility and centrality measures derived from the transit network analysis do not really capture the factors correlated with activity types from the perspective of individual travelers, since the activity designation may be only part of what motivated the destination choice and route choice.

Incorporating the temporal variation of establishments and transit services to the built environment measurement provides a better portrayal of the surroundings of each transit trip destination at the time the trip was made. By using the concurrent built environment measures, the goodness of fit as well as the prediction accuracy of the MNL learning model does increase, although only marginally. But considering the imperfect information of business open hours is used to measure the spatiotemporal distribution of establishments, it can be anticipated that more improvements will be achieved with better temporal data. In addition, the impacts of many temporal covariates are mixed up together in the learning models if only the dummy variables of time periods are included, to disentangle the mixed impacts and reduce the possible spaito-temporal confounding effect, we must use the built environment measures that are more precise both spatially and temporally.

To address the interdependencies among the activities, a linear-chain conditional random fields (CRFs) model is used to estimate types of activities according to the conditional probabilities of observing a sequence of potential activities based on the sequences of observed trips and destination locations. The result of the CRFs model shows a notable improvement in the overall classification accuracy due to the inclusion of Markovian transitions among activities as additional variables in the model. However, because the CRFs model selects the most probable activity chain based on the sequences of observed variables as the predicted activities of the observed transit trip chain, which tend to bias toward activities with more samples, the classification accuracy rates of the minority activity types are generally worse in comparison with those from the MNL models. This suggests we may need to consider higher order dependencies in the CRF chains or identify other factors that distinguish transit trip types in the EZ-Link data for which we could fit separate models.

In this study, we also discussed several methodological issues that are confronted by the activity learning models. Regularized MNL model is employed to avoid the model overfitting issue in the traditional MNL model by including extra penalty for the coefficients of explanatory variables. For the issue of imbalanced classes concerning the poor classification of the minority activity types in the learning models, we have experimented with a resampling approach, SMOTE, to attain relative balanced sample sizes of the training dataset for different activity types. The 5-fold cross validation shows that the classification accuracies of the minority activities are improved to different extents, at the cost of reduced prediction accuracies of working, education and shopping activities.

In a nutshell, the proposed analytical framework demonstrates that the use of urban sensing data such as EZ-Link transactions for urban planning purposes requires considerable data processing and juxtaposition with many other datasets about the city, including traditional household travel surveys and newly emergent online data sources. The rich context is needed to piece together the trips, the inferred activities and the urban built environments under which the activity-trip behaviors are observed in urban sensing data. What's more, the framework emphasizes the streamlining of the analytical modules, which can make the future analysis more adaptive for the new data, new learning models and new measures of the contexts of activities and trips.

7.2 Research limitations and Future Work

7.2.1 Built Environment Measurement

One of the objectives of this study is to formulate appropriate measures for the urban built environment, which is mainly used to infer the types of activities based on the correlations between the locational built environment and trip destinations. But more generally, measuring the built environment is a part of the effort to quantitatively describe the broader contexts under which the activities and trips are participated, which can include more spatio-temporal factors than what are included in this study. Although a set of spatial metrics are calculated to evaluate the built environment from different perspectives, some simplifications and assumptions in the process of the measurement can be further refined in the future.

Because employment data are not available at the disaggregated level, the information about establishments becomes important for characterizing the opportunities and attractiveness of urban places. While the temporal availability of the establishments is accounted for in the study, it is recognized that the classification model used to impute the open hours should be improved. To improve the model, we can experiment with additional predictors or search for extra training data from new sources.

In addition, the types of establishments were determined on the basis of two-digit Singapore Standard Industrial Classification (SSIC) categories, which is a much aggregated level of classification. For example, all businesses related to food and drinks are classified as "Food and Beverage Service Activities". Such a classification is not oriented to match with human activity types. Restaurants, pubs and coffee shops are of the same type and are considered to have the same effect on human activities. This certainly increases the ambiguity of the calculated spatial metrics. Therefore, a more detailed and activity-oriented classification of establishments is needed. More and more online information like yellow book and FourSquare provide concrete description of the businesses. Applying the ontology-based data integration approach and text mining approach may help to acquire the categorization that is more appropriate for the built environment measurement and activity learning.

Further, the role of a transit station in the transit network directly influence the connectivity and accessibility of the urban place it serves. The transit network analysis in this study addresses the issue of the integration of bus network and transit network by including transfers between stations as virtual links. The indicators focus on measuring the topological position of the stations in the network, such as centrality and connection strength. However, because the unit of analysis of transit network analysis is station as opposed to place, the centrality and accessibility of the individual station cannot be explicitly translated to the accessibility of places since one urban place may be served by multiple transit stations. From the point view of transit passengers, what needs to be measured is the accessibility of places, not only the accessibility from origins, but also the connectivity to next destinations if a chain of activities are scheduled. In this sense, although the transfers and frequencies of the services are accounted for in the current measurement, to measure the transit accessibility of a place (e.g. the catchment area of a transit station), it is necessary to identify other stations serving the same place and count the contributions from those stations on transit accessibility. Clearly, additional research effort is needed in order to transform the "network of stations" to the "network of places".

7.2.2 Learning Models

Good learning models are the ones that can extract relatively homogeneous relationships between activities and explanatory variables. This depends on two factors: the availability and quality of the data for both the dependent variables and the independent variables, and the structure and specification of model. The discussion in this section revolves around these two factors.

In this study, we adopt the trip purpose categorization used by the HITs survey as major activity types. However, this type of categorization is still too general and ambiguous. For example, among those activities categorized as social, the behaviors and choices of meeting with friends could be very different from meeting with relatives. Also, the variation between the grocery shopping and street shopping is not smaller than the variation between shopping and recreation. The multiple purpose trips are not accounted in the HITS data. Thus, it is necessary to further distinguish the types of activities to make it more specific and less ambiguous for the learning task.

Regarding the modelling structure, machine learning methods like logistic regression and conditional random fields (CRFs) are employed to learn the probability of various activity types contingent on the dynamically changing urban built environment quantified by a set of spatial metrics and other spatio-temporal variables. Both models belong to the generative graphic model. Although linear-chain CRFs model included the transitions among activities, it is limited at the first-order transition level, which only considers the constraints and associations between two consecutive activities in an activity chain. Given the complexity of choices involved in the activity planning, a number of improvements in the model structure and learning techniques should be investigated in the future:

1) Much of accuracy in predicting activities in the existing models comes from work and school trips. As more EZ-Link data become available, it is possible to separately estimate the locations of workplaces and schools of the card holders from the histories of trips, while leaving the model to focus only on predicting the location, characteristics, and timing of non-work locations.
2) Taking into consideration the interdependences among distant activities in an activity chain, which is potentially allowed by a skip-chain CRFs model.
3) Accounting for the latent status of the activities engaged (e.g. discretionary or mandatory) using the latent class choice model or Hidden-unit CRFs (van der Maa et al., 2011).
4) The dependencies or correlations in the current models are stationary over time, which may not match the reality. The dynamic interactions between human activities and urban environment conceivably can be better accounted for by using the time-varying dependencies in the model.
5) Class imbalance appears to be a major issue restricting the classification accuracy of the CRFs model. This necessitates more research on the appropriate class balance approach for sequential models. One example is the distribution-based balance approach (Song et al, 2013).

Meanwhile, more evidences from empirical studies are needed to distinguish important factors from less relevant factors to prioritize the modeling efforts.

7.2.3 Assumptions and Speculations

In addition to activity learning models, many assumptions are made at various stages of analytics. For example, a number of heuristic rules are employed to impute the missing values of variables such as the open hours of businesses and the home locations of the EZ-Link card holders. Not all the assumptions and speculations are verified by existing data or supported by empirical studies. The unjustifiable assumptions can potentially lead to biased estimates of activity learning models and misclassification and misinterpretation of activity-travel patterns. Therefore, a side process needs to be established to re-check the assumptions involved in the analytical procedure as more data become available.

7.2.4 Data exploration and Visualization

As the complexity of activity-travel patterns and the amount of datasets increase, new generation of visualization tools are needed to digest and visually represent the ever-updating information in a dynamic and interactive way to facilitate more timely exploration and analysis. The development of the visualization application will continue by adding consolidated exploratory features for locations. Meanwhile, I will also experiment with other visual metaphors and labeling strategies, and apply them to datasets of various sizes. Another imperative task is to investigate ways to link the front-end user interface to the data processing components (i.e. R) so that the data analysis methods can be integrated smoothly into visualization and exploration process.

Currently, we only have access to the EZ-Link data of 15 sample days, which is insufficient to carry out the types of study that may reveal the casual relationships between the built environment and human activity-travel behavior such as the before and after analysis of a new MTR line. But as we start to accumulate multiple types of urban sensing data including parking lots availability and taxi GPS traces, it is desired to investigate the value of these urban sensing data in the longitudinal studies of activity-travel behaviors, and to compare the analytical power of the longitudinal studies with the cross-sectional studies fed by the data from surveys.

In the future, novel visualization tools are designed to meet these requirements by focusing on features like web-based interactive interface, dynamic information updating and consolidated visual representation of information. These are expected to support the modeling and decision-making, and facilitate more effective participation in urban planning process. Therefore, I believe this research will not only benefit other researches on activity based modeling but also facilitate the pro-active decision making process of policy makers through continuously updated data and interactive visualization tools.

7.3 Conclusions

The growing data flows from emergent sources become available to urban planning and management today are generally too big and too unstructured for traditional means of analysis and modeling designed for data stocks. To convert massive information in these data to the knowledge meaningful for urban planning, new analytical framework is needed. The framework needs to have the capability of gathering, integrating, analyzing and interpreting data repetitively so that the aberrant patterns can be captured and the fast-break trends can be monitored. Then, the insights gained from these efforts can be linked to the operations of public services and urban planning to support urban management in a more responsive way. Much of the knowledge can also help to improve the existing research paradigm of urban studies. For example, the longitudinal information on individual mobility contained in new data sources can help us develop a richer description of activities that could improve activity modeling.

Towards this end, this study proposes an analytical framework for using the transit smart card transaction data for human activity-travel research. Through the serial processes of preliminary data exploration, data integration, urban built environment measurement, and activity learning and inference, we investigated the potential of using the transit smart card transaction data, i.e., EZ-Link data, for better understanding individuals' activity-travel pattern, as well as the activity profiles of urban spaces. As has been emphasized throughout the thesis, we consciously avoid fitting urban sensing data to the existing activity-based research paradigm because of the evident gaps between the information embedded in these emergent data and the information needed for the activity-based modeling. Instead, we focus on the recurring exploration of the activity-travel patterns and urban dynamics enabled by the streams of urban sensing data and other auxiliary datasets.

The framework is composed of relatively independent modules, most of which are generic and transferable. This enables the framework to be readily adapted to handle other types of urban sensing data and to examine other aspects of urban dynamics, since the processes like the data integration, urban form characterization, and data visualization are essential for the majority of studies falling into the category of urban analytics. Some of modules such as the ontology-based data integration and the urban built environment measurement are useful for many other applications beyond the academic research. For example, it is always a challenge for the public sector to fuse and manage the data collected by various agencies. The ontology-based data integration approach may have some implications in this regard. Besides, the spatial metrics that can provide better characterization of urban form are useful to planning agencies.

In a nutshell, the type of analysis of EZ-Link data enabled by the proposed framework opens the door for the possibilities of exploring many other aspects of human behaviors and urban dynamics on account of large size of samples and continuous data streams. The examples described in this study such as transfer patterns and single transit trips represents a small portion of what is possible (e.g. the spatial segregation of activities, the destination choices of the residents living in the same neighborhood). In the future, we expect a close look at the longitudinal evolution of individuals' activity-travel patterns along with the changes in the urban environment can enlighten a more broad scope of inquiries on the interactions between human activities and urban environment, and provide new insights to the human activity-travel research.

References

Ahas, R., Aasa, A., Silm, S., & Tiru, M. (2010). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. Transportation Research Part C: Emerging Technologies, 18(1), 45-54. Elsevier Ltd.

Bafna, S. (2003). SPACE SYNTAX A brief introduction to its logic and analytical techniques. Environment and Behavior, 35(1), 17-29.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR), 41(3), 16.

Batty, M. (2013). Big data, smart cities and city planning. Dialogues in Human Geography, 3(3), 274-279.

Ben-Akiva, M. E., & Lerman, S. R. (1985). Discrete choice analysis: theory and application to travel demand (Vol. 9). MIT press.

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

Beneventano, D., Dahlem, N., El Haoum, S., Hahn, A., Montanari, D., & Reinelt, M. (2008). Ontology-driven semantic mapping. In Enterprise Interoperability III (pp. 329-341). Springer London.

Benslimane, D., Leclercq, E., Savonnet, M., Terrasse, M. N., & Yétongnon, K. (2000). On the definition of generic multi-layered ontologies for urban applications. Computers, Environment and Urban Systems, 24(3), 191-214.

Benfield, F. K., Raimi, M. D., & Chen, D. D. (1999). Once there were greenfields : how urban sprawl is undermining Americas's environment, economy, and social fabric. New York : Natural Resources Defense Council

Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 1, p. 740). New York: springer.

Blagus, R., & Lusa, L. (2013). "SMOTE for high-dimensional classimbalanced data" BMC Bioinformatics, 14(1), p. 106, 2013.

Boarnet, M., & Crane, R. (2001). The influence of land use on travel behavior: specification and estimation strategies. Transportation Research Part A: Policy and Practice, 35(9), 823-845.

Bonacich, P. (2007). Some unique properties of eigenvector centrality. Social Networks, 29(4), 555-564.

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.

Boyd, D., & Crawford, K. (2011). Six provocations for big data. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. Social Sci. Res. Netw., New York. Available: http://ssrn. com/abstract, 1926431.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: density, diversity, and design. Transportation Research Part D: Transport and Environment, 2(3), 199-219.

Clifton, K., Ewing, R., Knaap, G. J., & Song, Y. (2008). Quantitative analysis of urban form: a multidisciplinary review. Journal of Urbanism, 1(1), 17-45.

Cottrill, C. D., Pereira, F. C., Zhao, F., Dias, I. F., Lim, H. B., Ben-Akiva, M. E., & Zegras, P. C. (2013). Future Mobility Survey. Transportation Research Record: Journal of the Transportation Research Board, 2354(1), 59-67.

Crane, R. (2000). "The Influence of Urban Form on Travel: An Interpretive Review." Journal of Planning Literature, Vol. 15, No. 1, pp. 3-23.

Chung, E.-H., and A. Shalaby. (2005). A Trip Bases Reconstruction Tool for GPS-Based Personal Travel Surveys, Transportation Planning and Technology, Vol. 28, No. 5, 2005, pp. 381–401.

de Bok, M., & van Oort, F. (2011). Agglomeration economies, accessibility and the spatial choice behavior of relocating firms. Journal of Transport and Land Use, 4(1), 5-24.

Diao, Mi (Sept. 2010). Sustainable Metropolitan Growth Strategies: Exploring the Role of the Built Environment. PhD dissertation

Durand, C. P., Andalib, M., Dunton, G. F., Wolch, J., & Pentz, M. A. (2011). A systematic review of built environment factors related to physical activity and obesity risk: implications for smart growth urban planning. Obesity Reviews, 12(5), e173-e182.

Ewing, R., & Cervero, R. (2001). Travel and the built environment: a synthesis. Transportation Research Record: Journal of the Transportation Research Board, 1780(1), 87-114.

Fotheringham, A. S., Nakaya, T., Yano, K., Openshaw, S., & Ishikawa, Y. (2001). Hierarchical destination choice and spatial interaction modelling: a simulation experiment. Environment and Planning A, 33(5), 901-920.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1), 1.

Gagnon, M. (2007). Ontology-based integration of data sources. In Information Fusion, 2007 10th International Conference on (pp. 1-8). IEEE.

Galster, G., Hanson, R., Ratcliffe, M. R., Wolman, H., Coleman, S., & Freihage, J. (2001). Wrestling sprawl to the ground: defining and measuring an elusive concept. Housing policy debate, 12(4), 681-717.

Garcia-López, M. À., & Muñiz, I. (2013). Urban spatial structure, agglomeration economies, and economic growth in Barcelona: An intra-metropolitan perspective*. Papers in Regional Science, 92(3), 515-534.

Gardner, S. P. (2005). Ontologies and semantic data integration. Drug discovery today, 10(14), 1001-1007.

Giese, Martin, Diego Calvanese, Peter Haase, Ian Horrocks, Yannis Ioannidis, Herald Kllapi, Manolis Koubarakis et al. "Scalable End-user Access to Big Data." Big Data Computing (2013): 205.

Girres, J. F., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. Transactions in GIS, 14(4), 435-459.

González, M. C., Hidalgo, C. a, & Barabási, A.-L. (2008). Understanding individual human mobility patterns. Nature, 453(7196), 779-82.

Goodman, J. 2004. Exponential priors for maximum entropy models. In Proceedings of the Annual Meeting of the Association for Computational Linguistics.

Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. Nature, 433(7028), 895-900.

Guo, Z., & Wilson, N. H. (2011). Assessing the cost of transfer inconvenience in public transport systems: A case study of the London Underground. Transportation Research Part A: Policy and Practice, 45(2), 91-104.

Gutmann, M. P., Witkowski, K., Colyer, C., O'Rourke, J. M., & McNally, J. (2008). Providing spatial data for secondary analysis: Issues and current practices relating to confidentiality. Population Research and Policy Review, 27, 639–665

Handy, S., Cao, X., & Mokhtarian, P. L. (2006). Self-selection in the relationship between the built environment and walking: Empirical evidence from Northern California. Journal of the American Planning Association, 72(1), 55-74.

Hariharan, Ramaswamy and Kentaro Toyama (2004) "Project Lachesis: Parsing and Modeling Location Histories". Geographic Information Science, Lecture Notes in Computer Science, 2004, Volume 3234/2004, 106-124,

Herold, M., Scepan, J., & Clarke, K. C. (2002). The use of remote sensing and landscape metrics to describe structures and changes in urban land uses. Environment and Planning A, 34(8), 1443-1458.

Hillier, B. (2007). Space is the machine: a configurational theory of architecture.

Horner, M. W., & Murray, A. T. (2002). Excess commuting and the modifiable areal unit problem. Urban Studies, 39(1), 131-139.

Huang, J., Lu, X. X., & Sellers, J. M. (2007). A global comparative analysis of urban form: Applying spatial metrics and remote sensing. Landscape and Urban Planning, 82(4), 184-197.

Hunter, G. J., Bregt, A. K., Heuvelink, G. B., De Bruin, S., & Virrantaus, K. (2009). Spatial data quality: problems and prospects. In Research trends in geographic information science (pp. 101-121). Springer Berlin Heidelberg.

Jang, W. (2010). Travel time and transfer analysis using transit smart card data. Transportation Research Record: Journal of the Transportation Research Board, 2144(1), 142-149.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E., & González, M. C. (2013, August). A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing (p. 2). ACM.

Jin, Y., & Batty, M. (2013). Applied Urban Modeling: New Types of Spatial Data Provide a Catalyst for New Models. Transactions in GIS, 17(5), 641-644.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making, 11(1), 51.

Kim, M., Kotz, D., & Kim, S. (2006). Extracting a Mobility Model from Real User Traces. Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications, 1-13. Ieee.

Klien, E., Lutz, M., & Kuhn, W. (2006). Ontology-based discovery of geographic information services—An application in disaster management. Computers, environment and urban systems, 30(1), 102-123.

Knublauch, H., Fergerson, R. W., Noy, N. F., & Musen, M. A. (2004). The Protégé OWL plugin: An open development environment for semantic web applications. In The Semantic Web–ISWC 2004 (pp. 229-243). Springer Berlin Heidelberg.

Kockelman, K. M. (1997). Travel behavior as function of accessibility, land use mixing, and land use balance: evidence from San Francisco Bay Area. Transportation Research Record: Journal of the Transportation Research Board, 1607(1), 116-125.

Koller, D., & Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.

Krizek, K. J. (2003). Operationalizing neighborhood accessibility for land use-travel behavior research and regional modeling. Journal of Planning Education and Research, 22(3), 270-287.

Krygsman, S., Jong, T.D., & Schmitz, P., (2007). Capturing daily urban rhythms: The use of location aware technologies. 10th International Conference on Computers in Urban Planning and Urban Management, 11-13 July, 2007.

Kumar, M., Bowen, W. M., & Kaufman, M. (2007). Urban spatial pattern as self-organizing system: An empirical evaluation of firm location decisions in Cleveland–Akron PMSA, Ohio. The Annals of Regional Science, 41(2), 297-314.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. Nature, 400(6740), 107-107.

Li, X., & Yeh, A. G. O. (2004). Analyzing spatial restructuring of land use patterns in a fast growing region using remote sensing and GIS. Landscape and Urban planning, 69(4), 335-354.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R news, 2(3), 18-22.

Longley, P. A., & Mesev, V. (2000). On the measurement and generalization of urban form. Environment and Planning A, 32(3), 473-488.

Lynch, K. (1960). The image of the city (Vol. 11). MIT press.

McGarigal, K., & Marks, B. J. (1995). Spatial pattern analysis program for quantifying landscape structure. Gen. Tech. Rep. PNW-GTR-351. US Department of Agriculture, Forest Service, Pacific Northwest Research Station.

McNally, MG. (2000). The Activity-based Approach. Handbook of transport modeling, Chapter four. 53-68. Elservier, Oxford, UK.

Métral, C., Falquet, G., & Vonlanthen, M. (2007). An ontology-based model for urban planning communication. In Ontologies for Urban Development (pp. 61-72). Springer Berlin Heidelberg.

Mesev, V. (2005). Identification and characterisation of urban building patterns using IKONOS imagery and point-based postal data. Computers, Environment and Urban Systems, 29(5), 541-557.

Mishra, S., Welch, T. F., & Jha, M. K. (2012). Performance indicators for public transit connectivity in multi-modal transportation networks. Transportation Research Part A: Policy and Practice, 46(7), 1066-1085.

Morency, C., Trepanier, M., & Agard, B. (2007). Measuring transit use variability with smart-card data. Transport Policy, 14(3), 193-203.

Newman, M. (2010). Networks: an introduction. Oxford University Press.

N. J. I. Mars, "What is an ontology?", in The impact of ontologies on reuse, interoperability, and distributed processing, (A. Goodall, ed.), pp. 9-19, Uxbridge, Middlesex, U.K.: Unicom, 1995.

Olszewski, P., & Wibowo, S. S. (2005). Using equivalent walking distance to assess pedestrian accessibility to transit stations in Singapore. Transportation Research Record: Journal of the Transportation Research Board, 1927(1), 38-45.

Palmer, J. R., Espenshade, T. J., Bartumeus, F., Chung, C. Y., Ozgencil, N. E., & Li, K. (2013). New approaches to human mobility: Using mobile phones for demographic research. Demography, 50(3), 1105-1128.

Papinski, D., Scott, D. M., & Doherty, S. T. (2009). Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. Transportation Research Part F: Traffic Psychology and Behaviour, 12(4), 347-358. Elsevier Ltd.

Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies, 19(4), 557-568.

Peter Mooney, Padraig Corcoran, and Adam C. Winstanley. 2010. Towards quality metrics for OpenStreetMap. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '10). ACM, New York, NY, USA, 514-517. DOI=10.1145/1869790.1869875

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., & Ratti, C. (2010). Activity-aware map: Identifying human daily activity pattern using mobile phone data. In Human Behavior Understanding (pp. 14-25). Springer Berlin Heidelberg.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

Ratti, C. (2004). Urban texture and space syntax: some inconsistencies. Environment and Planning B: Planning and Design, 31(4), 487-499.

Razniewski, S., & Nutt, W. (2013). Assessing the completeness of geographical data. In Big Data (pp. 228-237). Springer Berlin Heidelberg.

Rhee, I., Shin, M., Hong, S., Lee, K., & Chong, S. (2008). On the Levy-Walk Nature of Human Mobility. 2008 IEEE INFOCOM - The 27th Conference on Computer Communications, 924-932. Ieee. doi: 10.1109/INFOCOM.2008.145.

Saelens, B. E., Sallis, J. F., & Frank, L. D. (2003). Environmental correlates of walking and cycling: findings from the transportation, urban design, and planning literatures. Annals of behavioral medicine, 25(2), 80-91.

Schwarz, N. (2010). Urban form revisited—Selecting indicators for characterising European cities. Landscape and Urban Planning, 96(1), 29-47.

Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P. A., Mukherjee, G., & Manna, S. S. (2003). Small-world properties of the Indian railway network. Physical Review E, 67(3), 036106.

Sevtsuk, A. (2010). Path and place: a study of urban geometry and retail activity in Cambridge and Somerville, MA (Doctoral dissertation, Massachusetts Institute of Technology).

Sevtsuk, A., Ekmekci, O., Nixon, F., & Amindarnari, R. (2013). Capturing Urban Intensity. In Conference on Computer-Aided Architectural Design Research in Asia (CAADRIA 2013) (Vol. 551, p. 560).

Simon Razniewski and Werner Nutt. 2013. Assessing the completeness of geographical data. In Proceedings of the 29th British National conference on Big Data (BNCOD'13), Georg Gottlob, Giovanni Grasso, Dan Olteanu, and Christian Schallhart (Eds.). Springer-Verlag, Berlin, Heidelberg, 228-237. DOI=10.1007/978-3-642-39467-6_21

Song, Y., & Knaap, G. J. (2004). Measuring urban form: Is Portland winning the war on sprawl?. Journal of the American Planning Association, 70(2), 210-225.

Stopher, P., & Greaves, S. (2007). Household travel surveys: Where are we going? Transportation Research Part A: Policy and Practice, 41(5), 367-381.

Sun, A., Lim, E. P., & Liu, Y. (2009). On strategies for imbalanced text classification using SVM: A comparative study. Decision Support Systems, 48(1), 191-201.

Sutton, C., & McCallum, A. (2006). An introduction to conditional random fields for relational learning (Vol. 2). Introduction to statistical relational learning. MIT Press.

Talen, E. (2002). Pedestrian access as a measure of urban quality. Planning Practice and Research, 17(3), 257-278.

Tsui, S. Y. A., and A. S. Shalaby (2006). An Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems. In Transportation Research Record: Journal of the Transportation Research Board, No. 1972. Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 38–45.

Tudorache, T., Noy, N. F., Tu, S., & Musen, M. A. (2008). Supporting collaborative ontology development in Protégé. In The Semantic Web-ISWC 2008 (pp. 17-32). Springer Berlin Heidelberg.

Uitermark, H. T., van Oosterom, P. J., Mars, N. J., & Molenaar, M. (1999, January). Ontology-based geographic data set integration. In Spatio-temporal database management (pp. 60-78). Springer Berlin Heidelberg.

Vail, D. L., Lafferty, J. D., & Veloso, M. M. (2007). Feature selection in conditional random fields for activity recognition. In Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on (pp. 3379-3384). IEEE.

Weiss, G. M., & Provost, F. (2001). The effect of class distribution on classifier learning: an empirical study. Rutgers Univ.

Wibowo, SS & Olszewski, P, (2005). Modeling walking accessibility to public transport terminals: Case study of Singapore mass rapid transit, Journal of the Eastern Asia Society for Transportation Studies, volume 6, pp. 147-156, 2005.

Wolf, J. (2006). Application of New Technologies in Travel Surveys. In Travel Survey Methods—Quality and Future Directions (P. Stopher and C. Stecher, eds.), Elsevier, Oxford, U.K., 2006, pp. 531–544.

Wolf, J., S. Schönfelder, U. Samaga, M. Oliveira, and K. W. Axhausen. (2004). Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information. In Transportation Research Record: Journal of the Transportation Research Board, No. 1870, Transportation Research Board of the National Academies, Washington. D.C., 2004,pp. 46–54.

Wu, J., Shen, W., Sun, W., & Tueller, P. T. (2002). Empirical patterns of the effects of changing scale on landscape metrics. Landscape Ecology, 17(8), 761-782.

Yang, X., J. D. Blower, Lucy Bastin, Victoria Lush, Alaitz Zabala, Joan Masó, Dan Cornford, Paula Díaz, and Jo Lumsden (2013) "An integrated view of data quality in Earth observation." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371, no. 1983.

Zhang, M., & Kukadia, N. (2005). Metrics of urban form and the modifiable areal unit problem. Transportation Research Record: Journal of the Transportation Research Board, 1902(1), 71-79.

Zhu, Y., and Ferreira, J., (2014) " Synthetic Population Generation for Land Use and Transportation MicroSimulation with Multiple Limited Datasets" Transportation Research Record (In Press)

Ziegler, P., & Dittrich, K. R. (2004, August). Three decades of data integration-All problems solved?. In IFIP congress topical sessions (pp. 3-12)

Ziegler, P., & Dittrich, K. R. (2007). Data integration—problems, approaches, and perspectives. In Conceptual Modelling in Information Systems Engineering (pp. 39-58). Springer Berlin Heidelberg.

Appendix I. Building Type Categorizations in the Data Integration

CATEGORY I	CATEGORY II	CATEGORY III	CATEGORY I	CATEGORY II	CATEGORY III
CIVIC	COMMUNITY	CHURCH	COMMERCIAL	COMMERCIAL RETAIL	GAS STATION
		COURT			KIOSK
		EMBASSY			MERCANTILE
		FIRE STATION			RESTAURANT
		GOVERNMENT			SERVICE BRANCH(ES)
		HDB BRANCH			SHOPS
		MOSQUE			SHOPPING CENTER
		MUSEUM			THEATER/OPERA
		POLICE ACADEMY			VEHICLE SERVICE
		POLICE OFFICE			BUSINESS / SCIENCE PARK
		POLICE STATION			FOOD & BEVERAGE
		STATUTORY BOARD			MALL SHOP
		TEMPLE			OFFICE
		CITY HALL			OTHER RETAIL
		COMMUNITY CENTER	HOTEL	HOTEL	BUDGET HOTEL
		GOVERNMENTAL OFFICE			HOSTEL
		LIBRARY			REGULAR HOTEL
		OBSERVATORY	INDUSTRIAL	INDUSTRIAL	ASSEMBLY
		POLICE HEADQUARTERS			INDUSTRY (UNSPECIFIED)
		POWER DISTRIBUTION			MANUFACTURING
		POWER PLANT (NATURAL GAS)			DEPOT
		SENIOR LIVING			WAREHOUSE
		TEMPLE (BUDDHIST)		WAREHOUSE	FACTORY/WORKSHOP(B2)
		WASTE DISPOSAL			LIGHT INDUSTRIAL(B1)
	HOSPITAL	CLINIC	MIXED RESIDENTIAL	MIXED RESIDENTIAL RETAIL	SHOPHOUSE
		HOSPITAL			CONSERVATION HOUSE
		NURSING HOME			SHOP/SHOPHOUSE
		POLYCLINIC	OTHER	OTHER NON-RESIDENTIAL	CAMP
		CHILDCARE			MILITARY
	SCHOOL	INTERNATIONAL SCHOOL			PRISON
		JUNIOR COLLEGE			FOOT TRAFFIC
		KINDERGARDERN			GATEHOUSE
		PRESCHOOL			MILITARY APARTMENTS
		PRIMARY SCHOOL			MILITARY OFFICE
		PRIVATE SCHOOL			SWITCHING STATION
		SCEONDARY SCHOOL	RESIDENTIAL	APARTMENT	REGULAR APARTMENT

		SPECIAL SCHOOL
		TERTIARY
		UNIVERSITY
		TRAINING SCHOOL
		ADV_CAMP
	SPORT	BADMINTON
		FITNESS CENTER
		INDOOR SPORTS
		INLINE HOCKEY
		MULTI-USE SPORTS
		SQUASH
		SWIMMING
		TENNIS
		TRACK AND FIELD
		MEDICAL
COMMERCIAL	COMMERCIAL MIXED	COMMERCIAL OFFICE
	COMMERCIAL OFFICE	RESORT
		CONFERENCING SPACE
		CINEMA
		FOOD CENTER
	COMMERCIAL RETAIL	FOOD CENTER AND MARKET
		MALL
		MARKET
		SUPERMARKET
		THEATER
		THEME PARK
		AMUSEMENT
		CANTEEN
		CASINO
		CLUB HOUSE
		CONCERT LOCATION
		DRY LABORATORY
		EXHIBITION
		GALLERY

		RENTAL APARTMENTS
		SERVICED APARTMENTS
		CONDOMINIUM
	CONDOMINIUM	REGULAR CONDOMINIUM
		OFFICE CONDOMINIUM
		RESIDENTIAL CONDOMINIUM
		BUNGALOW
	DETACHED	PUBLIC HOUSING
		CHALET
		LODGE
		DORMITORY
	EXECUTIVE CONDO	RESIDENTIAL
	HDB	SEMI-DETACHED
	OTHER RESIDENTIAL	TERRACE
		BUNGALOW HOUSE
		CLUSTER HOUSE
		CORNER TERRACE
	SEMI-DETACHED	DETACHED HOUSE
		EXECUTIVE CONDOMINIUM
	TERRACE	GOOD CLASS BUNGALOW
		SEMI-DETACHED HOUSE
		TERRACED HOUSE
		TOWN HOUSE
		WALKUP
TRANSPORT	CAR PARK	CAR PARK
		MULTISTOREY CAR PARK
		PARKING
	PUBLIC TRANSPORT	AIRCRAFT TRAFFIC
		AIRPORT PASSENGER TERMINAL
		BUS STATION
		METRO STATION
		RAIL STATION
		TRANSIT STATION

Appendix II Sectorial Categorization of Establishments

SSIC 2010		Our Categorization
01	AGRICULTURE AND RELATED SERVICE ACTIVITIES	Agriculture
02	FORESTRY, LOGGING AND RELATED SERVICE ACTIVITIES	
03	FISHING, OPERATION OF FISH HATCHERIES AND FISH FARMS; SERVICE ACTIVITIES INCIDENTAL TO FISHING	
08	MINING AND QUARRYING	Mining
09	SERVICE ACTIVITIES INCIDENTAL TO OIL AND GAS EXTRACTION EXCLUDING SURVEYING	
10	MANUFACTURE OF FOOD PRODUCTS	Manufacturing
11	MANUFACTURE OF BEVERAGES	
12	MANUFACTURE OF TOBACCO PRODUCTS	
13	MANUFACTURE OF TEXTILES	
14	MANUFACTURE OF WEARING APPAREL; MANUFACTURE OF ARTICLES OF FUR; MANUFACTURE OF KNITTED AND CROCHETED APPAREL	
15	TANNING AND DRESSING OF LEATHER; DRESSING AND DYEING OF FUR; MANUFACTURE OF FOOTWEAR	
16	MANUFACTURE OF WOOD AND OF PRODUCTS OF WOOD AND CORK, EXCEPT FURNITURE; MANUFACTURE OF ARTICLES OF STRAW AND PLAITING MATERIALS	
17	MANUFACTURE OF PAPER AND PAPER PRODUCTS	
18	PRINTING AND REPRODUCTION OF RECORDED MEDIA	
19	MANUFACTURE OF COKE AND REFINED PETROLEUM PRODUCTS	
20	MANUFACTURE OF CHEMICALS AND CHEMICAL PRODUCTS	
21	MANUFACTURE OF PHARMACEUTICALS AND BIOLOGICAL PRODUCTS	
22	MANUFACTURE OF RUBBER AND PLASTIC PRODUCTS	
23	MANUFACTURE OF OTHER NON-METALLIC MINERAL PRODUCTS	
24	MANUFACTURE OF BASIC METALS	
25	MANUFACTURE OF FABRICATED METAL PRODUCTS EXCEPT MACHINERY AND EQUIPMENT	
27	MANUFACTURE OF ELECTRICAL EQUIPMENT	
28	MANUFACTURE OF MACHINERY AND EQUIPMENT	
29	MANUFACTURE OF MOTOR VEHICLES, TRAILERS AND SEMI-TRAILERS	
30	MANUFACTURE OF OTHER TRANSPORT EQUIPMENT	
31	MANUFACTURE OF FURNITURE	
32	OTHER MANUFACTURING	
35	ELECTRICITY, GAS AND AIR CONDITIONING SUPPLY	Public Service
36	WATER COLLECTION, TREATMENT AND SUPPLY	
37	SEWERAGE	
38	WASTE COLLECTION, TREATMENT AND DISPOSAL ACTIVITIES; MATERIALS RECOVERY	
41	CONSTRUCTION OF BUILDINGS	Construction
42	CIVIL ENGINEERING	
43	SPECIALISED CONSTRUCTION ACTIVITIES	
45	WHOLESALE AND RETAIL TRADE OF MOTOR VEHICLES AND MOTORCYCLES	Retail and Wholesale
46	WHOLESALE TRADE, EXCEPT OF MOTOR VEHICLES AND MOTORCYCLES	
47	RETAIL TRADE, EXCEPT OF MOTOR VEHICLES AND MOTORCYCLES	
49	LAND TRANSPORT AND TRANSPORT VIA PIPELINES	Transportation
50	WATER TRANSPORT	
51	AIR TRANSPORT	
52	WAREHOUSING AND SUPPORT ACTIVITIES FOR TRANSPORTATION	
53	POSTAL AND COURIER ACTIVITIES	
55	ACCOMODATION	Accommodation
56	FOOD AND BEVERAGE SERVICE ACTIVITIES	Food
58	PUBLISHING ACTIVITIES	

59	MOTION PICTURE, VIDEO AND TELEVISION PROGRAMME PRODUCTION, SOUND RECORDING AND MUSIC PUBLISHING ACTIVITIES	Information and Communication Service
60	RADIO AND TELEVISION BROADCASTING ACTIVITIES	
61	TELECOMMUNICATIONS	
62	COMPUTER PROGRAMMING, CONSULTANCY AND RELATED ACTIVITIES	
63	INFORMATION SERVICE ACTIVITIES	
64	FINANCIAL SERVICE ACTIVITIES, EXCEPT INSURANCE AND PENSION FUNDING	Financial Service
65	INSURANCE, REINSURANCE, PROVIDENT FUNDING AND PENSION FUNDING	
66	ACTIVITIES AUXILIARY TO FINANCIAL SERVICE AND INSURANCE ACTIVITIES	
68	REAL ESTATE ACTIVITIES	Real Estate Service
69	LEGAL AND ACCOUNTING ACTIVITIES	Professional Service
70	ACTIVITIES OF HEAD OFFICES; MANAGEMENT CONSULTANCY ACTIVITIES	
71	ARCHITECTURAL AND ENGINEERING ACTIVITIES; TECHNICAL TESTING AND ANALYSIS	
72	SCIENTIFIC RESEARCH AND DEVELOPMENT	
73	ADVERTISING AND MARKET RESEARCH	
74	OTHER PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES	
75	VETERINARY ACTIVITIES	Other Service
77	RENTAL AND LEASING ACTIVITIES	Administrative Service
78	EMPLOYMENT ACTIVITIES	
79	TRAVEL AGENCIES, TOUR OPERATORS AND RESERVATION SERVICE ACTIVITIES	
80	SECURITY AND INVESTIGATION ACTIVITIES	
81	CLEANING AND LANDSCAPE MAINTENANCE ACTIVITIES	
82	OFFICE ADMINISTRATIVE, OFFICE SUPPORT AND OTHER BUSINESS SUPPORT ACTIVITIES	
84	PUBLIC ADMINISTRATION AND DEFENCE	
85	EDUCATION	Education
86	HEALTH SERVICES	Social Service
87	RESIDENTIAL CARE SERVICES	
88	SOCIAL SERVICES WITHOUT ACCOMMODATION	
90	CREATIVE, ARTS AND ENTERTAINMENT ACTIVITIES	Entertainment and Recreation
91	LIBRARIES, ARCHIVES, MUSEUMS AND OTHER CULTURAL ACTIVITIES	
92	GAMBLING AND BETTING ACTIVITIES	
93	SPORTS ACTIVITIES AND AMUSEMENT AND RECREATION ACTIVITIES	
94	ACTIVITIES OF MEMBERSHIP ORGANISATIONS	Other Service
95	REPAIR OF COMPUTERS, PERSONAL AND HOUSEHOLD GOODS AND VEHICLES	
96	OTHER PERSONAL SERVICE ACTIVITIES	
97	ACTIVITIES OF HOUSEHOLDS AS EMPLOYERS OF DOMESTIC PERSONNEL	
99	ACTIVITIES OF EXTRATERRITORIAL ORGANISATIONS AND BODIES	
00	ACTIVITIES NOT ADEQUATELY DEFINED	

Appendix III. Model Estimation Results

Table A1. Results of MNL1 Model

MNL1	Education	Leisure	Shopping	Eating	Social	Personal	Work-related	Other
Intercept	-3.679***	-3.719***	-2.396***	-3.529***	-2.732***	-3.149***	-3.229***	-2.754***
	(0.133)	(0.125)	(0.069)	(0.112)	(0.085)	(0.097)	(0.111)	(0.096)
Card holder types								
Student/Child (dummy)	7.293***	3.246***	2.203***	2.535***	2.575***	2.237***	0.820**	1.570***
	(0.149)	(0.158)	(0.122)	(0.160)	(0.133)	(0.162)	(0.279)	(0.179)
Senior (dummy)	-0.437	1.978***	1.880***	1.656***	1.842***	2.207***	0.465*	0.965***
	(0.592)	(0.159)	(0.090)	(0.142)	(0.107)	(0.112)	(0.465)	(0.145)
Population and Job density								
Population density (person per m2)	50.847***	24.608*	45.463***	67.466***	92.554***	53.745***	50.075***	88.339***
	(6.343)	(10.271)	(5.389)	(7.676)	(5.579)	(7.083)	(8.730)	(6.375)
Job density (jobs per m2)	-125.03***	-33.796*	18.894**	7.646	-67.191***	-27.909*	-23.668’	-60.417**
	(20.051)	(15.330)	(5.765)	(10.065)	(15.279)	(12.359)	(13.638)	(18.409)
Retail job density (jobs per m2)	-126.76***	22.038**	42.887***	19.564**	-0.641	15.813**	12.660’	-2.088
	(13.715)	(6.720)	(3.136)	(6.024)	(7.280)	(6.091)	(7.314)	(9.057)
Manufacture job density (jobs per m2)	-196.38***	-140.83***	-357.64***	-317.5***	-151.5***	-82.56***	-54.446*	-193.72***
	(26.973)	(35.509)	(35.933)	(54.926)	(25.716)	(21.368)	(23.255)	(33.794)
Office job density (jobs per m2)	74.092*	40.943*	-73.775***	-34.919*	37.835’	9.002	14.304	-14.364
	(29.417)	(19.537)	(11.232)	(16.460)	(21.010)	(17.228)	(18.138)	(30.422)

Note. 1. Log-Likelihood: 14764; Adjusted McFadden R2: 0.3276; Significance codes: *** 0.001, ** 0.01, * 0.05, ’ 0.1;

2. Figures in the parenthesis are the standard error of coefficients.

Table A2. Results of MNL2 Model

MNL2	Education	Leisure	Shopping	Eating	Social	Personal	Work-related	Other
Intercept	-33.228*	-15.564	-37.052*	-76.02***	-26.746′	-13.854	-13.911	-35.145*
Card holder types								
Student/Child	7.477***	3.247***	2.270***	2.572***	2.596***	2.297***	0.892**	1.601***
Senior	-0.293	1.912***	1.923***	1.695***	1.880***	2.172***	0.549**	0.955***
Land use measures								
Contagion Index	-0.041′	-0.110***	-0.061***	-0.056*	-0.055**	-0.029	-0.037	-0.092***
LU entropy	1.995′	-1.995	-1.248	-0.475	-1.781′	-0.493	-1.755	0.487
Park area ratio	0.282	4.971***	4.033***	3.474*	3.672***	1.357	0.989	1.840′
Open space area ratio	-0.240	6.625	11.195	5.091	5.451	2.616	6.021	3.307
Commercial area ratio	-2.375′	3.621*	3.503***	-3.244*	-2.043′	-0.101	4.709**	1.122
Mixed use area ratio	-1.508	0.230	4.423***	2.170	0.449	1.793	0.232	-0.612
Worship area ratio	22.667**	24.763**	21.261**	13.284	14.917*	27.16***	22.47**	16.936*
Sport area ratio	-2.037	6.049***	1.895	-0.123	0.920	2.384′	-0.062	2.855′
Waterbody area ratio	-2.733	-0.024	-11.44***	-11.069***	-0.046	-5.672*	2.451	-5.590*
Reserve land area ratio	1.536′	-0.299	0.810	1.380	1.759*	-0.384	3.472***	1.315
AWMSI	-3.194′	-3.978	-5.399**	-8.223***	-2.878′	-0.349	-2.376	-3.428′
AWMFD	27.537′	23.879	41.673**	79.652***	28.601′	7.188	15.837	39.439*
LU Richness	-0.069′	-0.037	-0.071′	-0.045	0.020	0.001	-0.065	-0.075′
Building and Street layout								
Bld footprint area ratio	-1.520	1.967	-4.059***	1.659	0.159	-0.372	-1.911	-1.101
Normalized building footprint size	0.137	-0.131	0.237*	-0.038	-0.094	-0.030	-0.228	-0.100
Residential building ratio	2.139**	0.699	2.245***	2.515**	0.576	1.008	0.461	1.880**
School building ratio	5.655***	-1.388	1.733′	3.843**	0.205	3.052**	-3.161*	1.464
Commercial building ratio	-1.798	0.933	5.418***	1.101	-2.214	0.823	-6.659*	-0.204
Mixed building ratio	-2.286***	-0.060	1.488***	0.776	0.473	0.135	-0.459	-0.678
Community building ratio	-2.570	-7.000**	-3.209*	-6.556*	-0.672	1.676	-2.995	0.910
HDB density (units per m2)	0.162***	0.127	0.051***	0.190***	0.163***	0.100*	0.61	0.45***
Nearest Neighbor Index -building	5.292**	2.667	5.522***	3.342	4.113**	1.522	-1.177	0.461

Centrality Index	-26.578***	-13.641	-25.323***	-19.855’	-18.214**	-15.744’	7.432	-7.747
Compactness Index	0.159	-10.181	-73.556’	28.778’	-26.309	-6.832	-8.569	-102.990
Road density (1000 meters per m2)	-63.342*	-21.851	1.010’	0.142	-0.877’	-30.338	1.804*	-1.846**
Pedestrian mall length ratio	17.070**	3.212	-1.872	5.765	1.827	8.187’	-8.678	11.872*
Expressway length ratio	-0.102	-0.874’	-0.462	-0.092	0.526’	-0.291	-0.459	-0.018
Four-way Intersection ratio	0.326	0.228	0.230	0.083	0.626’	-0.131	0.053	1.279**
Intersection density (intersections/km)	0.251**	0.021	-0.046	0.216*	0.196**	0.076	0.019	0.139’
Establishments								
Entropy index	5.485*	3.703	7.216**	13.234*	5.613*	16.390***	3.297	13.429***
# Manufacture establishments	2.445	-6.661	-3.729	-1.209	-6.129*	-3.828	-0.829	-11.438*
# Financial establishments	0.001	0.001	0.006***	0.003	0.002	-0.001	0.000	-0.001
# Information establishments	0.024***	-0.007	0.011**	-0.008	0.010*	-0.001	0.003	0.002
# Professional service establishments	-1.595	1.213	-8.321***	-0.289	-4.087’	-1.586	-1.026	-0.594
# Cultural service establishments	-0. 321	-0.0959	-0.189	0.113	-0.434*	-0.904***	-0.091	-0.126
# Other services	3.066	6.917	1.725’	2.296	4.403	-0.772	6.369’	1.097
# Sport establishments	7.822	-36.992	-17.105	-22.168	-53.354’	-98.695**	-63.836	-109.94**
# Government establishments	13.796*	1.805	7.436*	-5.913	9.544*	-2.696	15.203**	7.591’
# Other establishments	-7.816’	1.742	-16.893***	2.397	-4.355	2.662	2.610	-4.872
# Outdoor POIs	0.055	-0.122’	0.045	-0.018	-0.032	-0.061	0.104’	0;055
Nearest Neighbor Index	1.933***	1.905*	-1.029’	-0.129	0.881’	-0.085	-1.880*	1.937**
Normalized sd distance to station	2.454’	3.324’	2.060’	-0.303	0.133	2.889’	0.963	-0.631
Normalized mean distance to station	1.488	0.028	3.537**	3.222’	0.738	4.434**	-1.960	0.833
Normalized average pair distance	-0.572	-1.574*	-0.742’	-0.954	-0.716’	-0.604	0.599	-0.546
Normalized average distance of retails	1.008	2.135’	-2.456**	-2.094	-0.847	-2.366*	0.625	-0.656
Normalized average distance of food	0.669	-1.154’	-1.781***	-1.818*	0.220	-0.499	0.880	0.521
Total establishment share	-2.058*	-0.322	0.896’	0.497	-0.362	0.644	-2.040*	0.429
Location quotient - food	0.231’	-0.221	-0.200’	0.042	-0.144	-0.183	0.037	-0.258’
Location quotient -social	0.088	-0.047	-0.032	-0.275	-0.019	0.145’	0.053	0.023
Location quotient -retail	0.870’	1.187;	1.650**	0.962	0.628	0.596	0.709	0.713
Location quotient -entertainment	0.031	0.093	0.046	0.073	0.139*	0.032	0.123’	-0.051

Diversity of food establishments	-0.571	1.157	0.986	1.635	1.057’	0.711	0.085	1.641*
Diversity of social establishments	0.130	-2.191	1.324’	-3.214	-3.854*	-0.051	-1.089	-0.667
Diversity of entertainment establishment	1.200	2.101	-2.258*	-0.522	2.141	-2.531’	1.164	-1.926
Diversity of retail establishment	-0.915’	-1.606’	-0.916’	-0.285	-0.338	-1.030	-0.526	-1.870*
Transit Network – station level indicators								
Transfer links - from other stations	-0.016	-0.044’	-0.001	0.000	-0.047**	-0.012	-0.015	-0.042*
Transfer links - to other stations	0.035’	0.062*	-0.014	0.014	0.058**	0.004	0.030	0.028
hub score (including one transfer)	2.738*	-0.464	0.590	0.021	1.158	-8.902*	3.468'	1.446
Strength (one transfer)	0.003	0.027	0.010	0.022	-0.006	0.080’	-0.016	0.018
Authority score (one transfer)	-0.396	-0.999	-1.378	-0.801	-2.523*	2.358	-3.438’	-2.472’

Note. Log-Likelihood: -14143; Adjusted McFadden R2: 0.354; Significance codes: *** 0.001, ** 0.01, * 0.05, ‘ 0.1;

Table A3.Results of MNL3 Model

MNL3	Education	Leisure	Shopping	Eating	Social	Personal	Work-related	Other
Intercept	-32.669’	18.836	-0.306	-43.370’	-7.131	-4.248	1.305	-24.628
Student/Child	7.439***	3.068***	2.052***	2.441***	2.467***	2.124***	0.80**	1.454***
Senior	-0.416	2.350***	2.273***	2.180***	2.327***	2.456***	0.741***	1.369***
Contagion index	-0.042’	-0.114***	-0.073***	-0.062*	-0.064**	-0.056*	-0.044’	-0.098***
LU entropy index	0.404	-3.346*	-0.522	-0.696	-2.316*	-1.072	-3.262*	-0.827
Park area ratio	1.93’	6.112***	3.890***	4.743***	4.709***	2.358*	1.479	3.298**
Open space area ratio	0.360	8.314**	10.32***	6.258’	3.769	-1.727	6.753**	2.126
Mixed use area ratio	-1.388	2.019	2.88*	2.759’	0.670	1.472	0.576	-0.144
AWMSI	-4.189*	-0.780	-2.758’	-6.031*	-1.433	-0.462	-1.172	-3.178’
AWMFD	34.432’	-9.218	8.356	52.02*	11.946	6.100	4.894	33.98’
Worship area ratio	28.337***	26.295**	26.652***	16.632’	14.794*	25.160***	25.282**	19.440**
Sport area ratio	-1.199	8.541***	1.741	0.367	2.487’	3.629*	1.190	4.753**
Waterbody area ratio	-2.377	2.800	-6.431**	-7.135*	2.359	-4.692’	4.464 *	-3.598
Reserve land area ratio	0.443	-0.736	0.609	1.891’	2.062*	0.691	3.048**	2.372*
Commercial area ratio	-3.024*	1.766	2.253*	-4.054**	-3.619**	-3.270*	3.421*	-1.468
Richness	0.000	-0.059	-0.089*	-0.043	0.022	0.027	-0.012	-0.017
Patches	0.000	0.002	0.001	-0.010**	-0.006**	-0.002	0.000	-0.008**
HDB unit density (unit/m2)	0.095***	0.031	0.212***	0.160***	0.115***	0.092***	0.082***	0.173***
Nearest Neighbor Index	3.550’	3.813	6.697***	3.719	5.319**	2.579	-1.057	2.347
Centrality Index	-18.055*	-17.130	-26.283***	-18.238’	-20.930**	-16.671’	5.309	-12.175
Compactness Index	-21.801	-9.393	-52.943	15.880	-33.269	-35.459	-11.096	-173.51’
Bld foot print area ratio	0.013	-1.525	-8.325***	-1.459	-4.116***	-3.984**	-2.879*	-5.295***
Residential building ratio	2.663***	-0.764	0.927’	1.455	-0.113	0.001	0.088	0.753
School building ratio	5.411***	-0.891	0.910	2.572’	0.898	3.022**	-3.026*	2.240*
Commercial building ratio	0.613	-0.163	4.171***	-2.554	-5.656**	0.848	-7.93**	-2.976
Mixed use building ratio	-1.371*	0.043	1.083*	0.356	-0.108	-0.125	-0.881	-0.881’
Community building ratio	-4.435*	-7.654**	-5.555***	-6.771	-2.219	0.035	-4.256’	-1.427
Normalized footprint size SD	0.022	-0.027	0.386***	0.133	0.086	0.179’	-0.175	0.048
Road density (1000m/m2)	-76.347**	-19.405	43.448’	11.946	50.862*	-11.136	45.154	27.299
Pedestrian mall length ratio	16.764**	2.040	-4.082	3.709	0.229	6.735’	-10.792’	12.624**
Expressway length ratio	-0.906*	-0.998’	-0.938*	-0.300	0.179	-0.562	-1.068’	-0.204

Four-way Intersection ratio	0.726′	0.609	0.332	0.420	1.033**	0.074	0.096	1.607***
Intersection per km road	158.220*	70.028	32.293	334.21***	243.64***	142.72′	-60.765	252.620**
Nearest neighbor index - establishment	0.476	0.976′	-0.179	-1.237′	0.669′	0.508	-1.462**	1.243**
Distance to station (standard deviation)	2.096′	1.208	0.491	-3.094′	-1.311	1.201	-0.331	-1.631
Entropy - establishment	-0.929	13.18***	13.519***	16.007***	12.954***	14.071***	4.087**	12.363***
Establishment share (1/1000)	-0.239	-0.397 *	-0.123***	-0.767***	-0.409**	-0.702***	-0.385	-1.275***
Location quotient - food	0.015	-0.276**	-0.269***	-0.281***	-0.276***	-0.276***	-0.121′	-0.261***
Location quotient -social	-0.043	-0.535***	-0.070*	-0.739***	-0.427***	0.006	-0.075	-0.031
Location quotient-retail	-0.159*	-0.350**	-0.287***	-0.230*	-0.162*	-0.221**	-0.111	-0.106
Manufacture establishments	-3.021	-7.190′	-4.412	-9.528*	-10.713***	-8.050**	-9.387**	-
Professional service establishments	-2.926	-4.354	-14.665***	-10.871**	-7.965**	-5.562′	-6.805′	-3.948
Cultural service establishments	-0.094	-0.386′	-0.159	-0.522*	-0.312′	-0.651***	-0.328	-0.087
Outdoor	0.087	-0.130′	0.131**	0.027	-0.121*	-0.039	-0.120′	0.072
Transfer links (from other stations)	-0.013	-0.045′	0.007	0.008	-0.042**	-0.032′	-0.021	-0.037′
Transfer links (to other stations)	0.014	0.026	-0.049**	-0.006	0.026	-0.003	0.022	-0.013
Hub score (one transfer network)	6.254***	-10.758***	-14.103***	-10.686***	-7.258***	-16.899***	-6.393***	-3.157
Strength (one transfer network)	-0.026	0.100	0.137	0.131	0.073	0.143	0.074	0.049
Authority score (one transfer network)	-1.706	2.207	3.512***	1.992	0.086	2.385*	-0.698	-0.558
Location quotient -entertainment	0.012	0.047	-0.243***	0.055	0.060′	-0.187**	-0.129′	-0.189***
Other services	-6.646′	4.587	11.463***	5.163′	2.571	10.520***	4.281	11.994***
Sport establishment	0.094′	0.223***	0.207***	0.216***	0.197***	0.166***	0.143**	0.230***
Government establishment	4.171	8.150′	15.470***	12.852**	12.550***	0.902	10.967*	11.952**
Other establishment	-2.194	0.180	2.028	3.546′	2.725′	7.209***	4.022′	9.721***
Average distance to station	0.801	2.311*	0.883	1.127	1.521*	2.652***	0.234	2.360**
Average pair distance	-0.191	-0.779	-0.543	-0.204	0.140	-0.692	0.270	-0.458
Average distance to station - retail	0.158	0.106	-0.996***	-0.675′	-0.660*	-1.023**	-0.383	-0.726*
Average distance to station - food	0.179	-1.611***	-0.379	-1.252*	-0.486′	-0.243	-0.320	-0.298
Average distance to station - entertainment	0.138	-0.992**	-0.868***	-0.976***	-1.219***	-0.835**	-0.992**	-1.622***
Transit trip frequency	-0.026	-0.184	-0.177′	-0.539**	-0.212	-0.353*	-0.297	-0.176
Financial establishments	-0.003	0.001	0.003′	0.003′	0.001	0.000	0.000	-0.002
Information establishments	0.022**	0.011′	0.026***	0.023***	0.021***	0.021***	0.013*	0.015*
Diversity - food sector	-0.035	-0.184	-0.701*	0.611***	0.337**	0.503***	0.198	0.244
Diversity - social sector	0.084	-13.20***	0.550*	-13.198***	-11.727***	0.428′	0.588**	0.408*

Diversity - entertainment sector	0.426	9.349**	-4.445***	7.663*	6.873***	-5.146***	-2.367***	-4.584***
Diversity - retail sector	-0.055	0.332	0.641***	-0.015	0.203	0.475*	0.328*	0.378*

Note. Log-Likelihood: -12255; Adjusted McFadden R2: 0.437; Significance codes: *** 0.001, ** 0.01, * 0.05, ' 0.1;

Table A4.Results of MNL4 Model

MNL4	Education	Leisure	Shopping	Eating	Social	Personal	Work-related	Other
Intercept	-28.072	-23.987	-11.329	-71.595**	-25.774’	-5.888	-20.656	-30.129’
Student/Child	7.401***	3.05***	2.044***	2.474***	2.493***	2.125***	0.703*	1.461***
Senior	-0.435	2.466***	2.339***	2.335***	2.511***	2.506***	0.737***	1.371***
Early morning	0.552	-3.403***	-5.266***	-3.148***	-5.218***	-3.506***	-1.869***	-3.356***
Morning peak	0.614*	-3.630***	-3.804***	-3.701***	-3.565***	-3.622***	-2.272***	-2.809***
Morning work	0.063	-2.224***	-2.046***	-1.989***	-2.065***	-1.647***	-1.753***	-2.451***
Noon	0.674**	-0.736***	-0.451**	0.263	-0.301’	-0.226	-0.698**	-0.344’
Afternoon peak	1.176***	0.847***	0.759***	1.466***	1.273***	0.069	0.296	1.573***
Night	0.826*	1.637***	0.868***	2.231***	1.518***	1.117***	0.288	0.807***
Late night	-17.540	-0.492	-1.239***	0.295	1.312***	-19.449	-0.772’	-0.269
In vehicle time (min)	-0.008’	-0.019***	-0.025***	-0.029***	-0.008*	-0.016***	-0.008’	-0.017***
MRT	0.077	0.395	-0.991***	0.006	-0.210	-0.470’	0.355	0.023
Transfer #	0.151	0.129	-0.615***	-0.452**	0.144’	-0.190’	0.035	-0.184
Rain	-0.491*	0.626*	-0.001	0.531*	0.187	0.457*	0.335’	-0.148
Temperature (>85)	-0.021	0.099***	0.044**	0.041’	0.051**	0.064***	0.030	0.009
Contagion index	-0.042’	-0.107**	-0.065**	-0.053’	-0.054*	-0.026	-0.026	-0.095***
LU entropy	0.856	-2.403’	-1.126	0.164	-1.72’	-0.525	-1.531	0.874
Park area ratio	0.859	4.437**	3.749***	2.709’	3.648**	1.048	-0.133	1.569
Open space area ratio	-1.253	8.197*	9.844***	4.285	4.670’	0.611	5.730*	1.955
Mixed use area ratio	-1.292	0.220	4.096**	2.242	0.506	1.492	0.398	-0.184
AWMSI	-3.137	-4.328	-2.887	-7.699**	-2.921’	0.615	-2.991	-3.418’
AWMFD	24.539	27.858	16.672	75.484**	26.224	-2.285	21.415	37.322’
Worship area ratio	26.561**	24.064*	19.307*	8.175	11.293’	23.481***	20.410*	14.326’
Sport area ratio	-0.607	6.86***	2.471’	0.279	2.014	3.697*	-0.063	4.113*
Waterbody area ratio	-0.372	-0.142	-7.953**	-9.099**	0.699	-4.292’	3.184’	-4.709’
Reserve land area ratio	1.022	-0.348	0.689	1.761	2.106*	-0.216	3.814***	1.640
Commercial area ratio	-1.054	3.094’	3.241**	-3.047’	-2.655’	-0.434	5.402***	0.366

LU richness	-0.033	-0.008	-0.036	-0.023	0.067’	0.022	-0.053	-0.064
Land patches #	-0.002	-0.001	0.003	-0.010**	-0.006*	0.000	0.001	-0.005’
HDB unit density	0.083**	0.054	0.197***	0.171***	0.111***	0.061*	0.045	0.161***
Nearest neighbor Index - building	3.690’	1.020	4.593**	2.666	2.661	0.683	-2.222	0.115
Centrality index	-17.249’	-6.315	-20.080*	-15.396	-10.343	-11.057	13.388	-4.408
Compactness index	-5.433	-39.428	-66.812	20.721	-60.498	-14.305	-18.589	-122.200
Building footprint area ratio	-2.200	1.540	-5.130***	0.377	-0.739	-1.111	-2.893’	-2.367
Residential building ratio	2.404**	0.460	2.617***	2.532*	0.278	1.037	0.120	2.033*
School building ratio	4.912***	-2.595’	0.493	2.542’	-0.853	2.160’	-3.963*	0.923
Commercial building ratio	-0.087	0.772	4.026**	-0.272	-4.456’	-0.518	-7.786**	-0.345
Mixed use building ratio	-2.374**	-0.062	0.978*	0.394	0.038	-0.362	-0.660	-0.678
Community building ratio	-4.173*	-8.537**	-4.198*	-7.269*	-1.725	1.319	-2.690	-0.236
Normalized bld footprint size	0.125	-0.027	0.395***	0.136	0.013	0.052	-0.180	0.006
Road density	-50.524’	-54.736	0.676	0.006	-0.488	-41.578	1.792*	-1.627**
Pedestrian mall length ratio	14.698*	7.690	-0.446	8.897	4.336	11.106*	-6.740	14.355**
Expressway length ratio	-0.237	-1.043’	-0.797’	-0.110	0.303	-0.498	-0.729	-0.153
Four-way intersection ratio	1.021*	0.339	0.311	0.264	0.787*	0.027	0.210	1.278**
Intersection per road length (km)	0.282**	-0.082	-0.102	0.213’	0.166*	0.045	-0.043	0.141’
Nearest neighbor Index - Establishment	1.160*	1.404’	-0.669	0.018	0.425	-0.294	-1.891*	1.679**
Distance to station (std.dev)	2.725’	2.815	-0.014	-1.557	-1.664	1.172	0.994	-1.929
Establishment entropy index	4.825*	1.293	1.884	7.506	3.977’	13.088**	2.779	11.611***
Local share of establishment	-2.286*	-0.315	0.470	0.185	-0.435	0.278	-2.028*	0.268
Locational quotient - food	0.314*	-0.220	-0.171	0.025	-0.138	-0.191	-0.022	-0.267’
Locational quotient - Social	0.055	-0.030	-0.044	-0.189	0.032	0.129	0.047	0.007
Locational quotient - Retail	0.943’	1.150’	1.170’	0.720	0.447	0.181	0.744	0.696
Manufacturing establishments	-0.660	-4.492	-0.531	1.055	-5.008’	-2.318	-0.748	-9.348’
Professional service establishments	-0.342	1.816	-7.717***	0.015	-3.222	0.313	-0.726	1.074
Cultural service establishments	-0.121	0.069	0.063	0.293	-0.344’	-0.785***	-0.012	0.093
Outdoor POIs	0.020	-0.119’	0.102*	-0.003	-0.050	-0.055	0.085	0.056
Transfer links (from other stations)	0.001	-0.052’	0.016	0.010	-0.045*	-0.010	-0.017	-0.031’

hub score (transfer network)	3.156*	-1.316	2.157	0.338	1.100	-8.141*	2.740	0.987
Strength (transfer network)	-0.008	0.038	-0.038	0.001	-0.011	0.064′	-0.001	0.009
Authority score (transfer network)	-0.671	-0.223	-0.788	0.040	-2.093	2.159	-2.989	-1.543
Locational quotient - Entertainment	0.069	0.068	-0.024	0.010	0.110	-0.019	0.116	-0.083
Other service establishments	6.938	6.446	0.787	1.574	3.408	-0.799	5.926	1.501
Sport establishments	0.016	-0.029	0.018	-0.002	-0.039	-0.086*	-0.061	-0.105*
Government establishments	11.522	0.102	8.133*	-5.368	8.312	-3.426	13.787*	5.857
Other establishments	-4.029	-1.552	-17.518***	0.251	-5.812	2.444	2.622	-5.896
Average distance to station	0.152	1.678	2.330′	2.978	0.470	3.905*	-2.174	0.100
Average pair distance	-0.186	-1.626′	-0.775	-1.171	-0.413	-0.376	0.611	-0.413
Average distance to station - retail	2.021′	1.084	-1.802′	-1.850	-0.923	-2.023′	0.862	-0.193
Average distance to station - food	0.647	-1.219	-1.242*	-1.400′	0.423	-0.415	1.041	0.839
Average distance to station - entertain	-0.503	-0.164	-0.194	-0.673	-0.153	-0.383	-0.262	-0.620′
Transit service frequency	-0.029	-0.004	-0.002	-0.025	0.003	-0.037	-0.014	-0.032
Financial establishments	-0.001	0.002	0.007***	0.003′	0.003′	-0.001	0.000	0.000
Information establishments	0.024***	-0.003	0.012**	-0.004	0.013*	0.000	0.004	0.002
Diversity - food	-0.467	1.114	0.629	1.583	0.805	0.414	0.143	1.766′
Diversity - social	0.056	-2.088	0.702	-3.458	-4.226*	-0.521	-1.529	-0.914
Diversity - entertain	1.101	2.343	-0.968	0.347	2.436	-1.889	1.673	-1.568
Diversity-retail	-0.868	-1.447′	-0.461	0.136	0.129	-0.534	-0.622	-1.737*
Friday	-0.309*	-0.316*	0.045	0.161	-0.156	0.062	-0.265′	-0.197′
Distance to home	-0.009	-0.002	-0.001	-0.014	0.026*	0.011	0.032*	-0.049***

Note. Log-Likelihood: -11272; Adjusted McFadden R2: 0.484; Significance codes: *** 0.001, ** 0.01, * 0.05, ‘ 0.1;

Table A5. Estimation results of MNL5

MNL5	Education	Leisure	Shopping	Eating	Social	Personal	Work-related	Other
Intercept	-21.6250	-10.2630	-0.8466	-74.516**	35.745*	-15.4900	-25.0660	-35.818’
Card holder types								
Student/Child	7.400***	3.079***	2.014***	2.486***	2.468***	2.088***	0.7043*	1.4085***
Adult (base)								
Senior	-0.5225	2.5114***	2.3418***	2.374***	2.489***	2.5337***	0.808***	1.422***
Time of day dummy variables								
Early morning (5am-7am)	0.5814	-3.337***	-4.242***	-3.292***	-6.733***	-2.576***	-1.0056’	-2.9845***
Morning peak (7am-9am)	0.2619	-3.6334***	-3.5373***	-3.5601***	-3.3831***	-3.1685***	-1.8089***	-2.4475***
Morning work (9am-11am)	0.0074	-2.2928***	-1.5931***	-1.8121***	-1.9911***	-1.5697***	-1.532***	-2.2963***
Noon (1am-1pm)	0.792***	-0.8098***	-0.3615*	0.3009	-0.2308	-0.3119’	-0.6785**	-0.3181’
Afternoon work (1pm – 5pm) (base)								
Afternoon peak (5pm – 7pm)	0.7103*	0.7133*	0.2332	1.3409***	1.1874***	0.2920	0.3506	1.3827***
Night (7pm-9pm)	0.5557’	1.5195***	0.5121*	2.0943***	1.3413***	1.245***	0.4372	0.6004*
Late Night (9pm-1am)	-19.0280	-0.7928	-1.2715**	0.1506	0.8809**	-20.1430	-0.4436	-0.5865
Friday (dummy)	-0.319*	-0.3986*	0.0579	0.1293	-0.1975’	0.0448	-0.2882*	-0.2074
Transit trips								
In-vehicle Time (min)	-0.0073’	-0.0180***	-0.0262***	-0.0311***	-0.0092**	-0.0182***	-0.0088’	-0.020***
MRT (dummy)	0.1297	0.3607	-1.1855***	-0.1094	-0.2618	-0.6621*	0.2138	-0.1006
#Transfers	0.1095	0.1174	-0.6313***	-0.4272**	0.1494’	-0.2108’	0.0123	-0.1710
Dist: Alighting station to home (km)	-0.0133	-0.0020	0.0013	-0.0087	0.0266*	0.0167	0.0367*	-0.039**
Weather								
Rain	-0.507*	0.670**	-0.0231	0.5414*	0.2027	0.4444*	0.3500’	-0.1608
Temperature	-0.0215	0.1041***	0.0433**	0.0439*	0.0523**	0.0671***	0.0327	0.0111
Land use measures								
Contagion Index	-0.0469’	-0.1060**	-0.0638**	-0.0601*	-0.0590**	-0.0416’	-0.0300	-0.1038***
LU Entropy	0.5780	-2.2780’	0.3490	0.8272	-1.3603	0.0445	-2.1896’	0.4809
Park area ratio	1.4129	3.8780**	3.1158***	2.6548’	3.3630**	0.7911	-0.3792	1.9898’
Open space area ratio	-0.4130	9.1668**	9.9345***	4.8259	2.8864	-1.0849	8.2786**	2.9640

Mixed use area ratio	-0.7553	0.4304	3.1933*	2.6195	-0.2031	0.5541	1.0654	-0.0720
Worship area ratio	28.742***	20.0710’	27.362***	13.5730	15.8050*	23.293***	22.623**	18.113*
Sport area ratio	-0.4975	5.5922***	0.1940	-2.3463	0.4484	2.3099	1.1253	3.1875*
Waterbody area ratio	-2.1258	0.5263	-6.1581*	-7.5196*	-0.4933	-5.8229*	3.7152*	-4.3828’
Reserve land area ratio	-0.0242	-0.1501	-0.0443	1.2244	1.3569	0.1202	3.2682**	1.7885’
Commercial area ratio	-1.7862	3.3238*	3.9661***	-1.4214	-3.3383*	-1.7692	5.044***	0.3503
LU Richness	-0.0200	-0.0443	-0.0636’	-0.0447	0.0487	0.0497	-0.0132	-0.0388
Number of land patches	-0.0010	0.0011	0.0026	-0.0114***	-0.0058*	-0.0014	-0.0012	-0.0069*
AWMSI	-2.8621	-2.8690	-2.3179	-8.3883**	-4.4028*	-1.2534	-3.5583	-4.4321*
AWMFD	23.4920	12.6160	8.7664	79.7910**	41.063*	13.7010	28.4410	45.431*
Building and street layout								
HDB unit density	0.0762**	0.0354	0.1777***	0.1592***	0.1113***	0.0506’	0.0667*	0.1313***
Nearest neighbor Index - Building	3.1336’	-1.4026	3.7010*	-0.4854	1.5540	-0.8015	-2.1097	-0.7715
Centrality	-14.332’	5.3175	-12.856’	-0.5175	-4.6420	-1.7775	11.0580	2.8315
Compactness	-0.035’	-0.052	-0.061	0.004	-0.098*	-0.102’	-0.021	-0.233
Building footprint area ratio	-1.5385	-0.2038	-7.195***	-0.5253	-2.2357’	-2.1391	-1.8740	-4.8094**
Residential building ratio	2.8707***	0.3096	2.230***	2.3382*	0.8075	0.8710	0.6856	1.5412’
School building ratio	5.0211***	-3.1218*	-0.5708	1.6035	-0.9370	1.6063’	-3.8270*	1.2054
Commercial building ratio	-0.4438	-0.5134	2.0349’	-2.9761	-4.6037*	0.1807	-8.3142**	-2.8918
Mixed use building ratio	-1.4955*	-0.4634	0.7094’	0.1132	-0.0538	-0.3559	-0.9075	-0.9320’
Community building ratio	-4.9021*	-7.0439**	-5.2292***	-7.0352*	-2.488’	0.4121	-2.7229	-1.3082
Normalized bld footprint size stdev	0.0353	-0.0958	0.392***	0.1172	-0.0376	0.0187	-0.2285’	-0.0487
Road density (m/area2)	-66.6740*	-52.6750	26.4660	-16.1030	9.4005	-26.5460	26.6830	11.4490
Pedestrian mall length ratio	8.9602	7.8174	-2.5985	6.2013	1.4636	9.2375*	-10.050’	12.2540*
Expressway length ratio	-0.3010	-1.4634*	-1.1168**	-0.3327	-0.1941	-0.7827’	-1.1966*	-0.4349
Four-way intersection ratio	0.9871*	0.3056	0.1449	0.1504	0.6846	-0.1442	0.1248	1.1021**
Intersection per km road length	0.2520**	-0.1112	-0.1130	0.2196*	0.0602	0.0159	-0.1213	0.1357’
Distribution of establishment								
Nearest neighbor Index	0.4751	0.7370	-0.4771	-0.8406	0.4007	-0.0171	-1.2036*	1.1003**
Entropy index	0.3772	3.8071’	-0.4169	3.0475’	-0.4470	1.5255	0.5701	3.5244*

Normalized mean distance to station	1.6104**	-0.4355	-0.6882	-1.1878’	-0.4978	0.8971	0.5363	0.6525
Stdev distance to station (km)	1.6282	-0.7688	-1.0860	-3.5213*	-2.4927*	-0.6562	0.4791	-2.7425*
Normalized mean distance - retail	-0.0350	0.8379’	-0.3973	-0.0618	0.0904	-0.4373	-0.1527	-0.0265
Normalized mean distance - entertain	-0.0569	0.2442	-0.1203	-0.0201	-0.0554	0.2073	-0.3324	-0.4216’
Local share of establishments	-0.1013	0.2067	-0.5878*	-0.0750	0.535**	-0.1031	-0.0925	-0.2643
Location quotient - food	0.0051	-0.0489	-0.0231	0.0659	0.0101	-0.0210	0.0188	-0.0363
Location quotient - social	-0.0248	-0.0396	-0.0229	-0.1678’	-0.1522’	0.0811	-0.0185	-0.0144
Location quotient - retail	-0.1442’	0.0064	-0.0976	0.1110	0.0877	-0.0329	0.0261	0.0984
Location quotient - entertainment	0.0309	0.0045	-0.0972’	-0.0062	0.0708’	-0.0115	-0.0559	-0.0060
# Manufacturing	-5.0001	0.6605	-0.0521	-1.2430	-2.7458	-2.3634	-4.0799	-9.6214’
# Professional service establishments	0.2780	-3.6622	-10.672***	-6.972’	-7.0851*	-4.4745’	-6.0854’	-2.3925
# Cultural establishments	0.1131	-0.413’	-0.250’	-0.449*	-0.0710	-0.624**	-0.416’	-0.2061
# Financial establishment	-3.7185’	0.7184	2.4756’	3.1747’	1.7193	-0.5176	-0.0033	-2.5946
# Information establishment	13.494’	0.2067	10.428*	5.4460	7.6847’	11.076*	6.6484	1.9171
# Outdoor POIs	0.081	-0.017	0.191***	0.061	-0.019	0.070	-0.080	0.097’
# Other services	-6.4898’	-4.5391	1.2719	-3.1494	-5.3853’	0.0230	-3.4560	-1.5356
# Sport establishment	0.068	0.064	0.065*	0.057	0.0079	0.032	0.030	0.095*
# Government establishment	0.6208	5.9658	13.231***	7.5174’	7.0731’	-1.3202	9.9358*	8.451’
# Other establishment	1.4771	-2.4840	-0.0331	1.5644	-0.4165	4.4004*	2.8745	6.1081*
Diversity index - food	0.0589	-0.2334	-0.3902	0.0427	0.2145	0.0881	-0.2626	-0.2141
Diversity index - social	0.1402	-0.0732	-0.0988	0.3600	-0.4936’	-0.1523	0.2707	0.2733
Transit Network Measures- Station Level								
# Transfer links (from other stations)	-0.0007	-0.0347	0.0167	0.0134	-0.0372*	-0.0220	-0.0259	-0.037’
# Transfer links (to other stations)	0.0126	0.0399	0.0032	0.0242	0.0497*	0.0142	0.0116	0.0080
Hub score index	5.349***	-0.7245	1.4767	-2.1486	0.2701	-3.1214	-2.3380	-1.8170
Strength	0.0113	0.0156	-0.0073	0.0086	-0.0035	-0.0001	0.0487	-0.0011
Transit trip frequency	0.3397	-0.1731	0.3446’	0.2066	-0.1716	-0.4543	-0.0543	-0.5184’

Note: Log-Likelihood: 11146.438, McFadden R^2: 0.4924, Likelihood ratio test: chisq = 21200 (p.value = < 2.22e-16)

Table A6. Estimation results of MNL6

MNL6	Working	Education	Leisure	Shopping	Eating	Social	Personal	Work-related	Other
Intercept	2.128	-1.935	-0.756	0.636	-0.334	0.355	0.015	-0.203	0.093
Card Types									
Student/Child	-2.430	4.310	0.544	-0.290	0.052	0.057	-0.278	-1.148	-0.817
Adult (base)									
Senior	-1.380	-0.549	0.440	0.536	0.363	0.554	0.689	-0.462	-0.190
Time of day									
Early morning (5am-7am)	0.997	0.545	-0.103	-0.475	-0.181	-0.604	-0.131	0.125	-0.173
Morning peak (7am-9am)	1.899	1.631	-0.579	-0.715	-0.839	-0.886	-0.696	0.290	-0.104
Morning work (9am-11am)	1.069	0.493	-0.281	-0.025	-0.357	-0.412	0.042	-0.045	-0.485
Afternoon work (1pm – 5pm) (base)									
Afternoon peak (5pm – 7pm)	-0.467	-0.123	0.012	-0.020	0.135	0.232	-0.190	-0.068	0.488
Night (7pm-9pm)	-0.642	-0.267	0.291	-0.010	0.474	0.221	0.155	-0.137	-0.087
Late Night (9pm-1am)	-0.017	-0.090	-0.013	-0.071	0.007	0.277	-0.066	-0.014	-0.012
Friday (dummy)	0.040	-0.063	-0.020	0.042	0.040	-0.014	0.023	-0.025	-0.023
Transit Trip									
In-vehicle Time (min)	0.213	0.080	-0.005	-0.201	-0.206	0.133	-0.030	0.109	-0.093
MRT (dummy)	0.108	0.025	0.026	-0.154	0.014	0.003	-0.054	0.061	-0.027
#Transfers	0.092	0.128	0.048	-0.248	-0.134	0.154	-0.039	0.053	-0.054
Dist: Alighting station to home (km)	0.013	-0.025	-0.014	-0.016	-0.042	0.087	0.025	0.112	-0.138
Weather									
Rain	-0.046	-0.055	0.025	-0.001	0.031	0.011	0.030	0.013	-0.007
Temperature	-0.198	-0.254	0.190	0.113	0.009	0.066	0.110	-0.045	0.009
Land use patches									
Contagion Index	0.120	0.027	-0.003	-0.090	-0.018	-0.011	-0.022	0.070	-0.074
Park area ratio	-0.101	-0.042	0.076	0.077	0.044	0.021	-0.047	-0.066	0.038
Open space area ratio	-0.073	-0.020	0.047	0.074	0.023	0.006	-0.070	0.007	0.005
Mixed use area ratio	-0.021	-0.027	0.001	0.044	0.051	-0.017	0.010	-0.026	-0.015
Worship area ratio	-0.155	0.075	-0.020	0.021	-0.025	0.003	0.079	-0.006	0.028

Sport area ratio	-0.016	-0.050	0.124	-0.061	-0.060	-0.014	0.036	-0.007	0.049
Waterbody area ratio	0.091	0.002	0.099	-0.156	-0.071	0.056	-0.008	0.057	-0.068
Reserve area ratio	-0.026	0.000	-0.019	-0.006	0.022	0.003	-0.002	0.029	0.000
Commercial area ratio	0.024	-0.084	0.063	0.178	-0.029	-0.143	-0.037	0.077	-0.051
LU Richness	0.007	0.004	-0.010	-0.014	-0.005	0.018	0.013	-0.010	-0.003
AWMSI	0.033	-0.026	-0.044	-0.091	0.012	0.024	0.063	-0.026	0.055
Building Layout									
HDB unit density	-0.298	-0.061	-0.075	0.153	0.096	0.145	-0.063	-0.031	0.134
Nearest neighbor Index	0.006	-0.010	-0.001	0.022	0.008	0.021	-0.046	0.009	-0.009
Compactness	0.062	-0.026	0.001	-0.002	0.020	-0.025	-0.018	0.011	-0.023
Building footprint area ratio	0.219	0.010	0.092	-0.201	0.010	-0.033	-0.060	0.075	-0.112
Residential building ratio	-0.057	0.053	-0.016	0.044	-0.019	-0.037	-0.012	0.018	0.025
School building ratio	-0.028	0.397	-0.140	-0.077	0.000	-0.079	0.063	-0.172	0.036
Commercial building ratio	0.063	-0.026	0.046	0.103	0.017	-0.091	-0.019	-0.050	-0.042
Mixed use building ratio	0.007	-0.024	0.004	0.035	-0.004	-0.005	0.012	-0.013	-0.011
Community building ratio	0.114	0.003	-0.051	-0.058	-0.059	-0.003	0.055	0.000	-0.002
Normalized bld footprint size stdev	0.010	0.006	-0.011	0.097	0.012	-0.071	0.040	-0.052	-0.031
Street Layout									
Road density	0.005	-0.060	-0.011	0.052	0.004	0.003	-0.008	-0.006	0.020
Pedestrian mall length ratio	-0.062	0.002	0.027	-0.003	0.020	0.002	0.018	-0.039	0.035
Expressway length ratio	0.071	-0.017	-0.037	-0.012	0.017	0.015	-0.022	-0.006	-0.008
Four-way intersection ratio	-0.020	0.015	0.001	-0.006	-0.021	0.034	-0.022	0.000	0.020
Intersection per km road	-0.030	0.042	0.001	-0.046	0.025	0.004	0.021	-0.021	0.004
Distribution of establishment									
Nearest neighbor Index	-0.009	0.171	0.010	-0.068	-0.126	0.038	-0.042	-0.079	0.105
Entropy index	-0.214	-0.124	0.068	0.073	0.006	-0.005	0.116	-0.031	0.111
Normalized mean dist. to station	0.003	0.080	0.004	-0.094	-0.066	0.002	0.049	-0.025	0.046
Normalized stdev dist. to station	0.032	0.094	0.025	-0.044	-0.052	-0.053	0.011	0.045	-0.059
Normalized mean distance - retail	0.010	0.022	0.009	-0.014	-0.010	0.001	-0.009	-0.007	-0.002
Normalized mean distance - food	-0.005	0.007	-0.005	-0.005	-0.009	0.004	0.006	-0.001	0.009

Normalized mean distance - leisure	0.071	0.106	-0.016	0.048	-0.048	-0.061	0.010	-0.016	-0.093
Establishment share	0.059	-0.017	0.006	-0.047	-0.009	0.013	0.008	0.004	-0.018
Location quotient -food	0.009	0.012	-0.003	-0.019	0.012	-0.002	-0.003	-0.001	-0.006
Location quotient -social	0.014	-0.007	-0.003	-0.009	-0.016	-0.011	0.022	0.000	0.010
Location quotient -retail	0.000	-0.005	-0.001	0.000	0.000	0.002	0.000	0.001	0.001
Location quotient -leisure	0.035	0.058	0.000	-0.048	-0.016	0.009	-0.023	-0.008	-0.006
# Manufacturing	0.099	-0.013	-0.027	-0.050	-0.022	-0.034	0.064	0.025	-0.042
# Professional service establishment	0.033	0.002	0.007	-0.131	-0.014	0.048	0.035	0.029	-0.010
# Cultural establishments	0.052	0.027	0.000	0.046	-0.021	0.005	-0.077	-0.024	-0.007
# Outdoor POIs	-0.017	0.013	-0.015	0.073	0.012	-0.055	0.003	-0.031	0.017
# Other services	-0.088	-0.210	0.024	0.260	0.059	-0.053	0.010	-0.011	0.009
# Sport establishment	-0.115	0.006	0.101	0.036	0.056	0.012	-0.071	-0.039	0.014
# Information establishment	-0.009	0.004	-0.004	-0.028	0.001	0.020	0.014	0.008	-0.007
Diversity index - food	0.052	0.058	-0.020	-0.095	0.008	-0.010	0.024	0.003	-0.020
Diversity index - social	0.065	0.097	-0.021	-0.036	-0.039	-0.099	-0.004	0.042	-0.004
Diversity index - leisure	0.016	0.008	0.000	-0.007	-0.004	-0.010	0.000	-0.001	-0.002
Transit Network Measures- Station Level									
Transfer degree	0.056	-0.025	-0.018	0.017	0.079	0.036	-0.046	0.020	-0.119
Hub centrality score	-0.019	0.214	-0.012	-0.024	0.006	-0.013	-0.113	-0.049	0.009
Authority centrality score	0.0005	0.0003	0.0001	0.0005	0.0001	-0.0004	-0.0004	-0.0003	-0.0003
Frequency	0.030	-0.028	0.026	0.112	-0.007	-0.051	-0.015	-0.023	-0.045