Computers, Environment and Urban Systems 48 (2014) 124–137

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Inferring building functions from a probabilistic model using public transportation data

Chen Zhong [a,*], Xianfeng Huang [a,b], Stefan Müller Arisona [c], Gerhard Schmitt [a], Michael Batty [d]

[a] Future Cities Laboratory, Department of Architecture, ETH Zurich, 8092 Zurich, Switzerland
[b] State Key Lab of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 430079 Wuhan, China
[c] Institute of 4D Technologies, University of Applied Sciences and Arts Northwestern Switzerland FHNW, 5210 Windisch, Switzerland
[d] Centre for Advanced Spatial Analysis, University College London, 90 Tottenham Court Road, W1N 6TR London, England, United Kingdom

ARTICLE INFO

Article history:
Received 19 March 2013
Received in revised form 14 July 2014
Accepted 16 July 2014
Available online 24 August 2014

Keywords:
Bayesian model
Spatial statistics
Building function
Activity
Smart card data

ABSTRACT

Cities are complex systems. They contain different functional areas originally defined by planning and then reshaped by actual needs and use by the inhabitants. Estimating the functions of urban space is of significant importance for detecting urban problems, evaluating planning strategies, and supporting policy making. In light of the potential of data mining and spatial analysis techniques for urban analysis, this paper proposes a method to infer urban functions at the building level using transportation data obtained from surveys and smart card systems. Specifically, we establish a two-step framework making use of the spatial relationships between trips, stops, and buildings. Firstly, information about the travel purposes for daily activities is deduced using passengers' mobility patterns based on a probabilistic Bayesian model. Secondly, building functions are inferred by linking daily activities to the buildings surrounding the stops based on spatial statistics. We demonstrate the proposed method using large-scale public transportation data from two areas of Singapore. Our method is applied to identify building functions at building level. The result is verified with master plan, street view, and investigated data, and limitations are identified. Our work shows that the presented method is applicable in practice with a good accuracy. In a broader context, it shows the effectiveness of applying integrated techniques to combine multi-source data in order to make insights about social activities and complex urban space.

1. Introduction

Urban systems are composed of many different forms of functional areas, which interact with one another to generate the complexity that defines a city. These functional areas are historically associated with many urban processes, some related to the institutions that are used to support planning but most being shaped by individuals' actual needs through processes of bottom-up change. In this spirit, Jane (1961) described cities as 'problems of organized complexity'. Taking a small park as an example, she argued that "... even this partial influence of the park's design upon the park's use depends, in turn, on who is around to use the park and when, and this in turn depends on uses of the city outside the park itself...". Similarly, in the book by Rodrigue (2013), land uses are defined in two ways. Formal land use refers to its form, pattern, and aspect, while functional land use refers to its socioeconomic description in space. The latter aspect is likely to imply higher levels of dynamic temporal change compared to the former as activities change faster than the physical locations and land uses that contain them. As discussed in Green (2007), functional changes in cities are not tied to morphological changes. In places such as Singapore, it is crucial to understand urban functions and their compatibility with the original Master Plan, which is very important to the development of the urban system, and the current push in understanding the dynamics of urban areas requires costly cross-sectional survey data, which in principle should be used to dynamically update information. As a potential solution to these problems, only recently has the availability of multiple location data sources, such as GSM traces, Wi-Fi data, GPS traces from taxis and smart-card data, emerged, and this is, for the first time, greatly stimulating the use of these "big" data sets for urban analysis. As it implies in Yuan, Zheng, and Xie (2012) that regions of different functions in a city can be detected using human mobility data and points of interest data. In Roth and et al. (2011), the characteristics of a polycentric urban form are defined from the analysis of large-scale, real-time smart-card data from which individuals' movement patterns can be inferred.

* Corresponding author.

http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.004
0198-9715/

The motivation for our study is to measure the structure and form of urban spaces in terms of real functions associated with urban land use using newly available 'big data' and, in this way, to explore the potential of using smart-card data to infer urban functions. To do this, we deduce information about individuals' travel purposes for daily activities from mobility patterns so that we can link these activities to specific locations to detect building functions. A two-step framework that makes use of the spatial relations between trips, bus stops, and building plots is presented. In this framework, we first analyze the survey data to find the mobility patterns of typical travel purposes based on travel time, activity time, and travel frequency. The analyzed results are then used for trip classification of the smart-card data using a probabilistic model. By analyzing the distribution of bus trips from each stop to the surrounding buildings, we can infer the most likely function of the building using a standard inverse distance weight function. The data used in this study are the Household Interview Travel Survey (HITS) and seven-day smart-card data obtained in Singapore from the Land Transport Authority, which pertain to all public transport usage. In an exploratory manner, we apply the method at the building level as the main focus of this paper. However, a more reasonable result is achieved at the block level due to the volume and resolution of the data, and we discuss the implications of this issue here. The inferred functions reflect the real use of urban space, which can be used to verify independent observations from various original plans. Moreover, corresponding results from different time series can be further used to detect the changes in activity location choice.

This paper develops three main contributions. First, we apply a probabilistic model to infer the travel purpose of a trip using spatiotemporal as well as socially related information. This enables us to explore a method of studying urban spaces through peoples' mobility patterns. Second, we propose a framework to infer building functions by combining multi-source data, namely, survey data and smart-card data, while integrating data mining techniques using a spatial statistical method. Third, the proposed method is demonstrated with real data collected in Singapore. Consequently, this study is focused on investigating the potential of using big data to infer the dynamics of various space functions, and we believe that this is one of the first attempts anywhere to extract activity and land usage data in this manner. The remainder of the paper is organized as follows. In the next section, related works are reviewed. In Section 3, the proposed methodology is presented, including the terminology, framework, and details of the probabilistic model. Section 4 discusses the experiments demonstrating the proposed method using the Household Interview Transportation Study (HITS) and the smart-card data. Section 5 concludes the paper and discusses further research.

2. Related work: Discovering functional areas in cities from movement data

Assessing the functions of urban spaces in terms of land use types is of significant importance for understanding urban problems (Taleai et al., 2007) and for evaluating planning strategies (Kajtazi, 2010). However, assessing urban functionality requires costly survey methods such as field investigation and interview questionnaires. In addition to the amount of manpower and time involved, the reliability of information is heavily influenced by subjective factors such as the time, place, and investigator's personal experience. The development of techniques based on Geographic Information Systems (GIS) and the availability of multiple data sources, such as GSM traces on cars, trains and taxis; mobile phone calls; Wi-Fi data; and smart-card systems, provide us with alternative solutions and change the way we can approach urban analysis.

Valuable insights have been gained regarding social activity and the complexity of urban space through analyses of movement data because urban travel is a good proxy for the transfer of urban flows, such as people, material products, and energy, thus revealing the importance of temporal dynamics in cities. Findings have been achieved from exploring such dynamics, for instance, using mobile telephone position data to analyze daily activity patterns (Phithakkitnukoon et al., 2010; Ratti et al., 2006), comparing the differences in temporal patterns with respect to the consumption of space (Ahas et al., 2010), studying spatiotemporal structures of urban mobility at a large scale (Sun et al., 2011) and classifying land uses based on aggregated data (Pei et al., 2014). The GPS trajectories of taxi cabs traveling in urban areas provide detailed location information, and in (Qi et al., 2011), the on-entrance/off-exit frequencies of taxi passengers were used to depict social activities in a region. Similarly, temporal patterns of pick-ups and drop-offs have been analyzed and associated with different land-use features in Liu and et al. (2012). In Yuan et al. (2012), taxi data combined with points of interests (POIs) were used to discover regions with different functionalities in a city. Some discussions regarding the opportunities and challenges of using various location data can be found in Lu and Liu (2012).

As smart-card payment systems are rapidly being adopted in cities around the world, they have also become an important data source that produces large quantities of very detailed data about an individual's daily travel (Pelletier, Trépanier, & Morency, 2009). In Quebec, data mining methods and public transport planning models have been used to obtain an improved portrait of users' travel behavior, and this has been tested using twelve one-week records (Agard, Morency, & Trépanier, 2006). In Seoul, a study was conducted by Park, Kim, and Lim (2008), in which boarding times and disembarking times were mapped and analyzed to prove the reliability of smart-card data. Liang and et al. (2009) investigated spatiotemporal human mobility patterns using smart-card data in Shenzhen, China, while (Munizaga & Palma, 2012) estimated a public transport O-D matrix from smart-card and GPS data in Santiago, Chile for transport systems analysis. In a study by Roth et al. (2011), data were collected from the smart-card system in the London tube, which is based on the Oyster card system, and were used to infer the statistical properties of individual movement patterns and to identify the polycentric nature of the various transport hubs in central London.

A clear trend that is revealed from this brief survey is in the exploration of the potential of using 'big' positional (geospatial) data for the analysis of urban forms, as proposed in Ratti and et al. (2006). In line with such trend, there are new methods of urban analysis emerging. For instance, discrete choice models can be used to estimate dynamic workplace capacities (Ordóñez Medina & Erath, 2013), identify urban activities from a synthesis of smart-card and survey data (Chakirov & Erath, 2012) and discover different functional regions within a city using floating car and point of interest data (Yuan et al., 2012). Machine learning methods are also being introduced to infer land use from mobile phone activity records and from zoning regulations (Toole et al., 2012). Spatial network analysis has been applied to the same set of smart-card data used in this paper to trace the urban transformation of decentralization in Singapore (Zhong et al., 2014).

In light of these potential uses of new data sources and analysis methods, this paper proposes a two-step integrated spatial data mining method aiming at inferring building functions using smart-card and survey data. A probabilistic model based on a Bayesian classifier and its related spatial statistics is used (1) to integrate considerably more attributes compared to when simply using spatiotemporal information and (2) to infer additional types of activity places instead of detecting only residential and workplaces, which has occurred in most case studies to date. Bayesian

models have been chosen because they are the most fundamental and stable for data mining and information retrieval (Mosegaard & Tarantola, 2002). This type of probability analysis is a mature technique in data-mining applications (Jensen, 2009) and exhibits good performance in other analogous applications that process events with multiple variables and known priori probabilities such as what we have in our own case. For instance, land use classification using radar terrain images (Decatur, 1989) is a typical example in image processing. Bayesian belief networks are used to integrate multiple data layers to estimate potential compatibilities and conflicts between development and landscape conservation as a decision-making tool (Mccloskey, Lilieholm, & Cronan, 2011). Integrating Bayesian belief networks with other modeling approaches for simulating future population and land-use change also provides good examples for prediction (Kocabas & Dragicevic, 2012). In sum, the characteristics of the Bayesian model make it a powerful tool for addressing sequential events in cities for events with complex network relations. In our research, we apply the Bayesian model to build relations between the mobility patterns of individuals' trips and their daily activities so that we might infer building functions specifically from the survey data and smart-card data for bus travel in Singapore, thus making this a comparatively new and innovative application.

3. An outline of the methodology

In this section, we first define key concepts in the context of this paper. Next, based on these definitions, we introduce the framework of the proposed method. Finally, a probabilistic model, which is the core part of the method, is presented.

3.1. Basic concepts

Land use, building function, daily activity, and mobility pattern are four basic concepts used throughout this work. Briefly stated, ***land use*** is the planned or naturally emerging usage that constrains the functions of buildings located on the land, which is usually enclosed by a plot. However, these constraints are compromised by the actual needs of people. Consequently, the real ***building function*** is reflected by actual usage, which reflects ***daily activities*** performed in the building. Information about the activities that motivated trips to a location can then be deduced from the ***mobility patterns*** of the travel behavior of people. These four concepts are generally used with ambiguous meaning; therefore, we must redefine them in the context of this paper, as follows.

Land use is the human use of land. It has been defined as "the human use of land involving the management and modification of the natural environment or wilderness into built environment such as fields, pastures, and settlements" (Watson et al., 2000). It informs the original planned usage and restricts the practical usage of the land.

Building function refers to the actual use of a building. In contrast to any preplanned zoning, this is how a building is used in reality. Building function refers to information at a smaller spatial scale than land use *per se* and is thus not fully compliant with land use. For instance, a grocery store may be located in a residential use area, thus distorting the real usage to one of mixed use, not simply residential. This paper takes a building as a basic unit that describes such a function. The function is determined by what type of daily activities actually occurs inside the building.

Daily activity refers to the routine and institutional activities that are followed usually over the 24 h day (Rodrigue, Comtois, & Slack, 2013), such as working, shopping, and eating, which are common social activities associated with any individual in the population, apart perhaps from the very young, very old and infirmed. This type of activity, which occurs regularly, is reported in our survey data and manifests as predictable patterns. Travel purpose and daily activity are used interchangeably in this paper.

Relations between the above three concepts are illustrated in Fig. 1. The figure shows that daily activities and building functions remain as question marks because this is the information that we are generating from movement patterns in this research.

Mobility pattern refers to travel behaviors, such as starting time and travel frequency, and past research has shown that an individual usually has very stable mobility patterns that can be analyzed and used as travel behavior to make predictions (Agard et al., 2006; Bagchi & White, 2005; Liang et al., 2009; Park et al., 2008). We use such mobility patterns to distinguish trips for different activities.

3.2. The framework

We will introduce a scenario to enhance the explanation of the above four concepts before moving on to the framework of our method. As shown in Fig. 2, people travel for specific purposes with respect to their daily activities. They arrive at one bus stop and subsequently travel to their final activity places, which are likely to be in surrounding functional buildings or public spaces. Two research questions remain for us to answer in this scenario: How do we deduce information about the daily activities that motivated

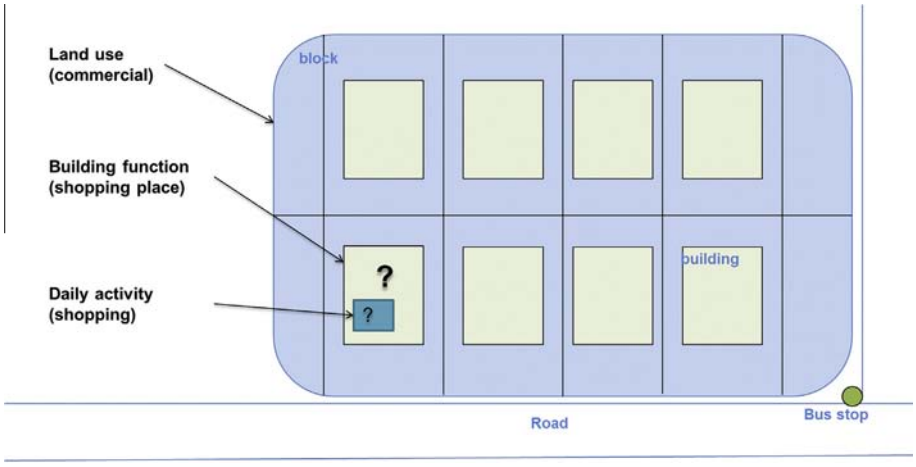


Fig. 1. An example of land use, building function, and daily activity.

the trip from the mobility patterns of people? How do we infer the building functions that meet daily activity requirements considering the spatial relation between trip, stops, and buildings? To answer these questions, we establish a framework based on a probabilistic model that links travel, daily activities, and building functions.

This proposed framework is broken down into two steps, which are coordinated with the two research questions, as implied in Fig. 3. After preliminary data processing, we first deduce information about the daily activities that motivated the trips using mobility patterns. This is performed using a Bayesian classifier. The result of this first step is a probability distribution of daily activities linked to each bus stop. Next, we make use of the spatial relations between bus stops and buildings to find the possible final destinations of trips, i.e., where people will perform their daily activities. These daily activities are reflected, of course, in the building functions.

3.3. Preliminary data processing

Specific input data, in our case, are introduced here for a better explanation of our method. Four types of input data are used: survey data, which is used for the statistical analysis of mobility patterns; smart-card data, which provides information that reflects peoples' daily activity; geo-referenced bus stop location points; and geo-referenced building footprints.

Preliminary data processing of these data sets is conducted. Firstly, a statistical analysis is applied to the survey data to find mobility patterns, which are subsequently used as our prior knowledge of peoples' travel behaviors. Secondly, smart-card data is processed. The original records provide information regarding trip ID, passenger ID, boarding bus stop ID, alighting bus stop ID, trip transfer time, starting time, traveling time, fare, and distance. We estimate the staying time by calculating the time between two trips, which is based on the tap in/tap out from a selected area for the same passenger ID. The frequency is a statistic of how many times a passenger ID appears on different dates associated with going to the same area. The newly generated record consists of six parameters: *passenger ID*, *passenger age*, *arrival time*, *staying time*, *frequency*, and *ID of the arrival stop*. The statistical results as well as the processed data structures using real sample data are show in Sections 4.2 and 4.3. Finally, the bus stops and building footprints are then stored in the Shapefile format, which are imported into ArcGIS and manipulated by ArcGIS functions such as redefining projections and calculating distances.

3.4. A probabilistic model for trip classification

The objective of this step is to deduce the most likely travel purposes for daily activities. By comparing possible data mining techniques, we found that using prior knowledge from survey data

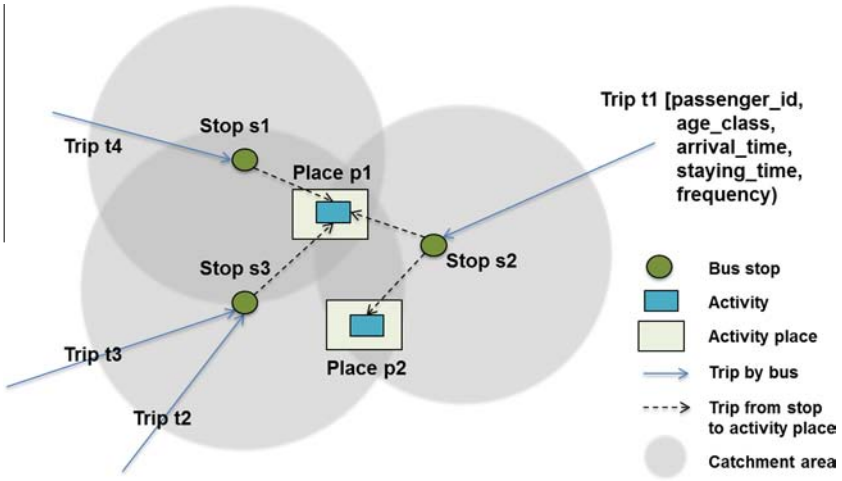


Fig. 2. A schematic scenario of people traveling from bus stops to buildings for daily activities.

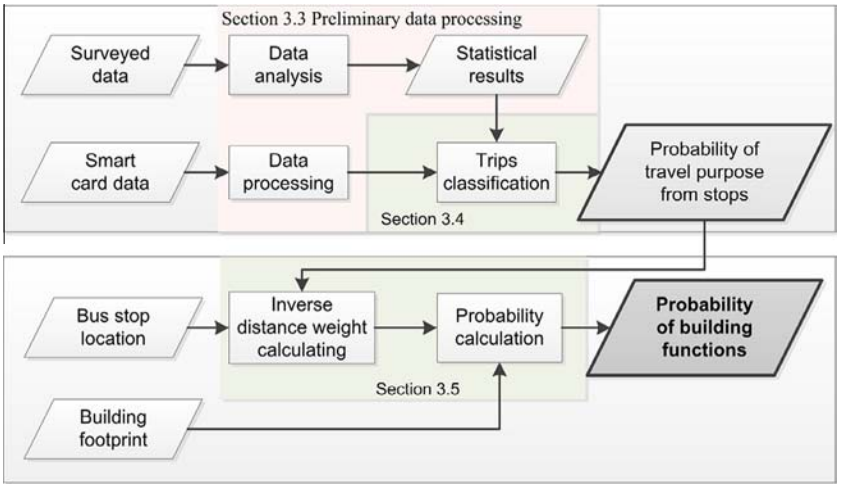


Fig. 3. Overview of the two-step framework for inferring building functions using transportation data.

to classify new records of smart-card data is a typical application of the Bayesian classifier.

A Naive Bayes classifier is a probabilistic classifier based on Bayes' theorem. Bayes' theorem expresses the relation between conditional probabilities when some events are contingent on other events (Carlin & Louis, 1997). Given sample input data, the Bayesian classifier assigns the most likely class label to a sample by evaluating its feature vector and its prior probability. The Naive Bayes model has been shown to be effective in many practical applications (Rish, 2001). Because the events of trips and their feature attributes satisfy conditioned independence, inferring information about daily activities can be formulated as an application of the Bayesian classifier. In this section, we will define the parameters that we have already defined in the Bayesian classifier.

Definition 1. Trip T: a trip is a generated record. A record is generated by a set of time-ordered points recording how one passenger arrives and leaves one place to engage in a certain urban activity. Each trip reveals mobility patterns, which are expressed by multiple attributes. For instance, in our case, trip t is denoted as $t = [a_a, a_t, a_d, a_f]$, where the attribute a_a stands for passenger age, a_t stands for arrival time, a_d stands for duration, and a_f stands for frequency. These attributes are mobility patterns that reveal people's travel purposes and that are linked to a certain activity created by a passenger after making the trip.

Definition 2. Activity class C: this is the set of possible urban activities that motivate a trip. It is also the information we want to deduce. In our case study, six activity classes are used, i.e., $C = \{c_{working}, c_{go\_home}, c_{shopping}, c_{studying}, c_{eating}, c_{social\_related}\}$.

For each activity candidate c, there is a prior probability $P(c)$. For each attribution $a_i (a_i \in \{a_a, a_t, a_d, a_f\})$ of a trip instance $t = [a_a, a_t, a_d, a_f]$ belonging to activity class $c (c \in C)$, there is a prior probability $P(a_a|c)$. This prior probability is our *a priori* knowledge that was learned from a statistical analysis of the survey data. As shown in Eq. (1) below, given a new trip instance $t = [a_a, a_t, a_d, a_f]$, the question can be formulated as follows: What is the most likely activity c that motivates the travel based on the prior known probability? The answer is found by calculating the maximum $P((a_a, a_t, a_d, a_f)|c)$. Therefore, the likelihood of trip $t = [a_a, a_t, a_d, a_f]$ belonging to $c (c \in C)$ is

$$\begin{aligned} P(c|(a_a, a_t, a_d, a_f)) &= P(a_a, a_t, a_d, a_f, c)/P(a_a, a_t, a_d, a_f) \\ &= P(c)P(a_a|c)P(a_t|c)P(a_d|c)P(a_f|c)/P(a_a, a_t, a_d, a_f) \end{aligned} \quad (1)$$

$t = [a_a, a_t, a_d, a_f]$ belongs to the activity class C_{MAP}, which has a maximum likelihood given by (2)

$$C_{MAP} = \max_{c_j \in C} P(c_j) \prod_i P(a_i|c_j) \quad (2)$$

The result of this step is a probability distribution of the travel purpose of each trip. Summing all trips by arrival stop, this result reflects the required functions provided by the buildings surrounding a stop. An example of the intermediate result of this first step is shown in Fig. 6. Trips concluding at 136 bus stops in one area are classified into six groups. The x-axis shows the bus stop ID, while the y-axis shows the probability distribution of the travel purposes for daily activities for each bus stop using six colors.

3.5. A spatial statistical model for extracting building functions

We consider that building functions can be derived from probability distributions of daily activities linked to arrival bus stops. To do this, we have to determine the source stops, where humans flow into surrounding buildings, and because these vary in quantity, different weights for each surrounding stop to one building should be used. Here, we generate a probability landscape of activities. Such a landscape portrays the probability of a certain activity happening at each area according to some sample points. The sample points in our case are the stops surrounding the area in question, where we assume that people choose the stop nearest to their destinations. We thus apply interpolation to the nearest neighbors of each stop. There are then two steps in generating the landscape: calculating weights that neighboring stops contribute to one area and calculating the final probabilities of each activity in one area, which is a theoretical problem.

To establish spatial relations between buildings and stops is to find the most likely buildings where people are heading to from their alighting stops. Logically, people will disembark at the stop nearest to their destination. People will sometimes go to bus stops further away for better bus service, but the stop will be relatively close to the destination. That is to say, distance is the most important factor influencing peoples' decisions. We demonstrate the statistical results of the survey data, where Fig. 4 (left) is the frequency distribution of the number of people and their walking time to the bus stops. This shows that most people manage to walk to their destinations from the bus stops within 10 min and that the number of people decreases as the distance increases. Fig. 4 (right) is a fitted curve that follows an exponential function of walking time and the probability of people choosing this bus stop to disembark. We conclude that the further the distance, the less likely that people are coming from this stop, which perfectly matches the inverse distance weight (IDW) function.

Although there are many variants of interpolation, as a tentative approach to the problem, we use inverse distance weighting (IDW), where each measured point has a local influence that diminishes with distance. The method weights the points closer to a particular location more highly than those further away, and the weights are defined generically for each point as

$$W_i(x, y) = 1/d_{ij}(x, y)^{\lambda} \quad (3)$$

where (x, y) is the geographical coordinates of a location point i, $W_i(x, y)$ is the weight of point i contributing to its neighbor point, j and $d_{ij}(x, y)$ is the distance from point i to point j. Note that the weights are normalized about a particular point to sum to 1, that is, $\sum_{\forall x,y} W_i(x, y) = 1$, and λ is a parameter set here as 1, which gives the coefficient of the population distribution and walking distance shown in Fig. 4. The weight is later used in the total probability theorem in definition 3, which follows.

Definition 3. Stop-activity-subspace S: this definition views a stop as a source of human flow. It distributes the flows traveling to different activities. We consider a stop as an experimental sample space to identify building functions, and we assume that these experimental sample spaces are independent of each other. Therefore, let $\{s_1, s_2, \ldots, s_n\}$ be a subset of the sample space s of an experiment. For $s \in S$, $P(s) > 0$, where $P(s)$ is the probability of a human flow from a bus stop s. For an activity c in the space s, $P(c|s) > 0$. $P(c|s)$ is the probability of an activity c arriving at stop s. Therefore, the probability of one building having a type of activity c conducted inside is

$$P(c) = \sum_{k=1}^{n} P(s)P(c_k|s) \quad (4)$$

The result of this step is the final probability distribution of the building functions. The process of calculating IDW and the results of the probability distribution of building functions are illustrated in Figs. 8–11.

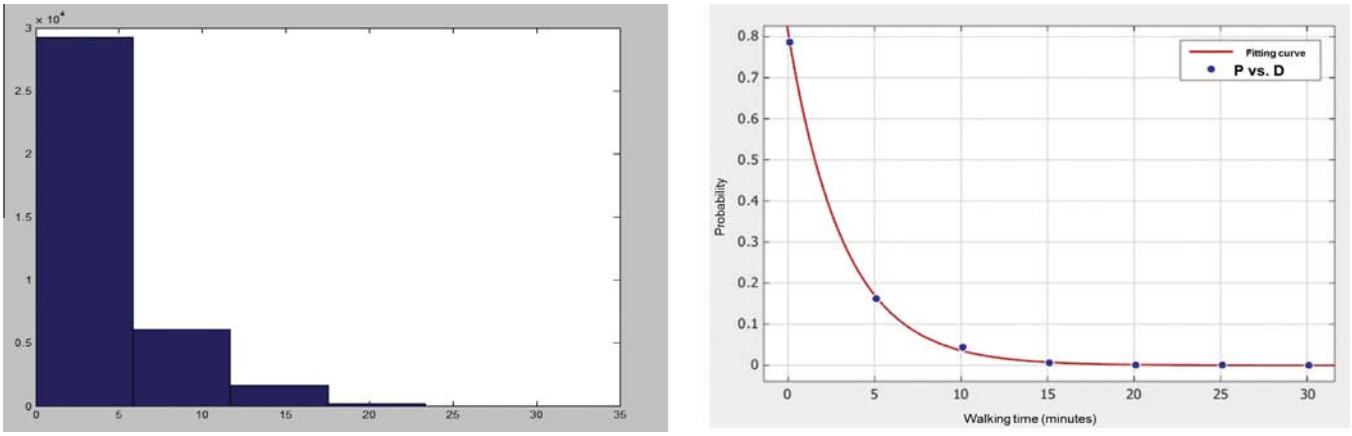


Fig. 4. Probability distribution of walking time from bus stop to destination.

Fig. 5. Two case study areas: Jurong East and Rochor.

4. Experiments and analysis

4.1. Data description

Two types of data are used as inputs. Survey data based on a Household Interview Travel Survey (HITS) is gathered by the Singapore Land Transport Authority (LTA) every four to five years to give transport planners and policy makers insights into residential travel behavior. Approximately 1% of households in Singapore are surveyed, with household members answering detailed questions about their trips. The HITS results provide very detailed information, including age, occupation, travel purpose, travel destination, walking time, waiting time, travelling time, and travel frequency for an activity. This paper uses the 2008 HITS results, which contain 88,601 records, in which 37,228 records are associated with the public transportation mode (Cheong & Toh, 2010). The smart-card data used in this study were collected by a fare collection system and kindly provided by the Singapore LTA (Land Transport Authority). This present study was conducted based on smart-card records from an entire week in April 2011.

4.2. Statistical analysis of survey data

Two essential elements are determined based on the statistical results. First, what are the mobility patterns ($[a_a, a_t, a_d, a_f]$) that can be used to classify trips? Second, what are the predictable daily activities (Activity C) as well as their corresponding building functions? The answers to these two questions are developed in the following sections.

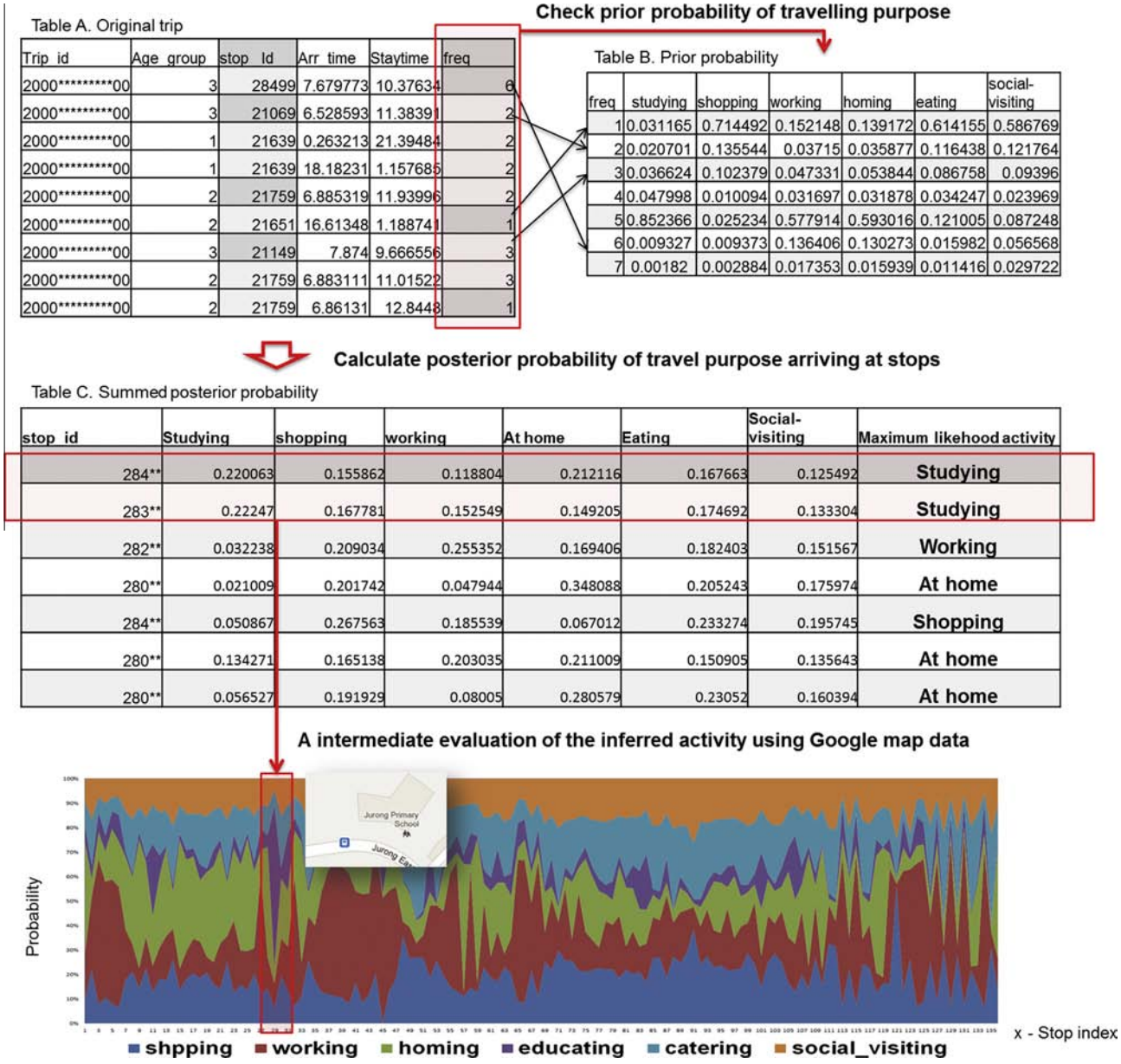


Table A. Original trip

Trip_id	Age_group	stop_Id	Arr_time	Staytime	freq
2000*********00	3	28499	7.679773	10.37634	6
2000*********00	3	21069	6.528593	11.38391	2
2000*********00	1	21639	0.263213	21.39484	2
2000*********00	1	21639	18.18231	1.157685	2
2000*********00	2	21759	6.885319	11.93996	2
2000*********00	2	21651	16.61348	1.188741	1
2000*********00	3	21149	7.874	9.666556	3
2000*********00	2	21759	6.883111	11.01522	3
2000*********00	2	21759	6.86131	12.8448	1

Table B. Prior probability

freq	studying	shopping	working	homing	eating	social-visiting
1	0.031165	0.714492	0.152148	0.139172	0.614155	0.586769
2	0.020701	0.135544	0.03715	0.035877	0.116438	0.121764
3	0.036624	0.102379	0.047331	0.053844	0.086758	0.09396
4	0.047998	0.010094	0.031697	0.031878	0.034247	0.023969
5	0.852366	0.025234	0.577914	0.593016	0.121005	0.087248
6	0.009327	0.009373	0.136406	0.130273	0.015982	0.056568
7	0.00182	0.002884	0.017353	0.015939	0.011416	0.029722

Table C. Summed posterior probability

stop_id	Studying	shopping	working	At home	Eating	Social-visiting	Maximum likehood activity
284**	0.220063	0.155862	0.118804	0.212116	0.167663	0.125492	Studying
283**	0.22247	0.167781	0.152549	0.149205	0.174692	0.133304	Studying
282**	0.032238	0.209034	0.255352	0.169406	0.182403	0.151567	Working
280**	0.021009	0.201742	0.047944	0.348088	0.205243	0.175974	At home
284**	0.050867	0.267563	0.185539	0.067012	0.233274	0.195745	Shopping
280**	0.134271	0.165138	0.203035	0.211009	0.150905	0.135643	At home
280**	0.056527	0.191929	0.08005	0.280579	0.23052	0.160394	At home

Fig. 6. Trip classification. The input data of trips (top left), statistical prior probability (top right), calculated posterior probability (bottom left), an intermediate evaluation of the probability distribution f. Daily activities at 136 bus stops (bottom, x – stop index, y – probability of activities).

4.2.1. Statistical results of mobility patterns

The statistical results are used as mobility patterns to classify different trips, where a set of mobility patterns ($[a_a, a_t, a_d, a_f]$) is found using a statistical analysis of the surveyed data, HITS, to infer travel purposes. In Table 1, we compare six mobility patterns: alighting time, age distribution, activity frequency, time use of activity, walking time from stops to buildings and activity locations of daily activities. To match attributes of the surveyed data with those of the smart-card data, we aggregate the surveyed data into discrete categories. The categories of arrival time, staying time and travel frequency are the same as those shown in Table 1. Because smart-card data only represents three age groups (Children and student card (4–20), adult card (20–50) and senior citizen's card (50 up)), which are fewer than what is represented by the survey data, we aggregate the survey data into these three categories of age groups. The priori probability distribution used as input into the Bayesian classifier is generated from the aggregated data.

4.2.2. Sectors of daily activities

Further explanations are given here regarding the different sectors of daily activities (C) and of building functions, which are used in this paper for demonstrating our method. We demonstrate the method using a classification of the most widely used daily activities and building functions in the urban analysis.

Originally, our classification was derived from survey data in Singapore. As a pilot study, we selected the experimental activities from the given list of travel purposes according to three criteria. The first criterion was that they should be the options (travel purposes) that account for a large proportion of the survey results. Second, the travel purposes should show distinct mobility patterns and thus support the validity of the classification. Third, the selected activities as a whole should cover all representative daily activities in urban analysis, including necessary activities, such as going home and working; optional activities, such as dining in a restaurant; and social activities, such as social visiting, which are all relevant to the urban designer's point of view (Gehl, 1987).

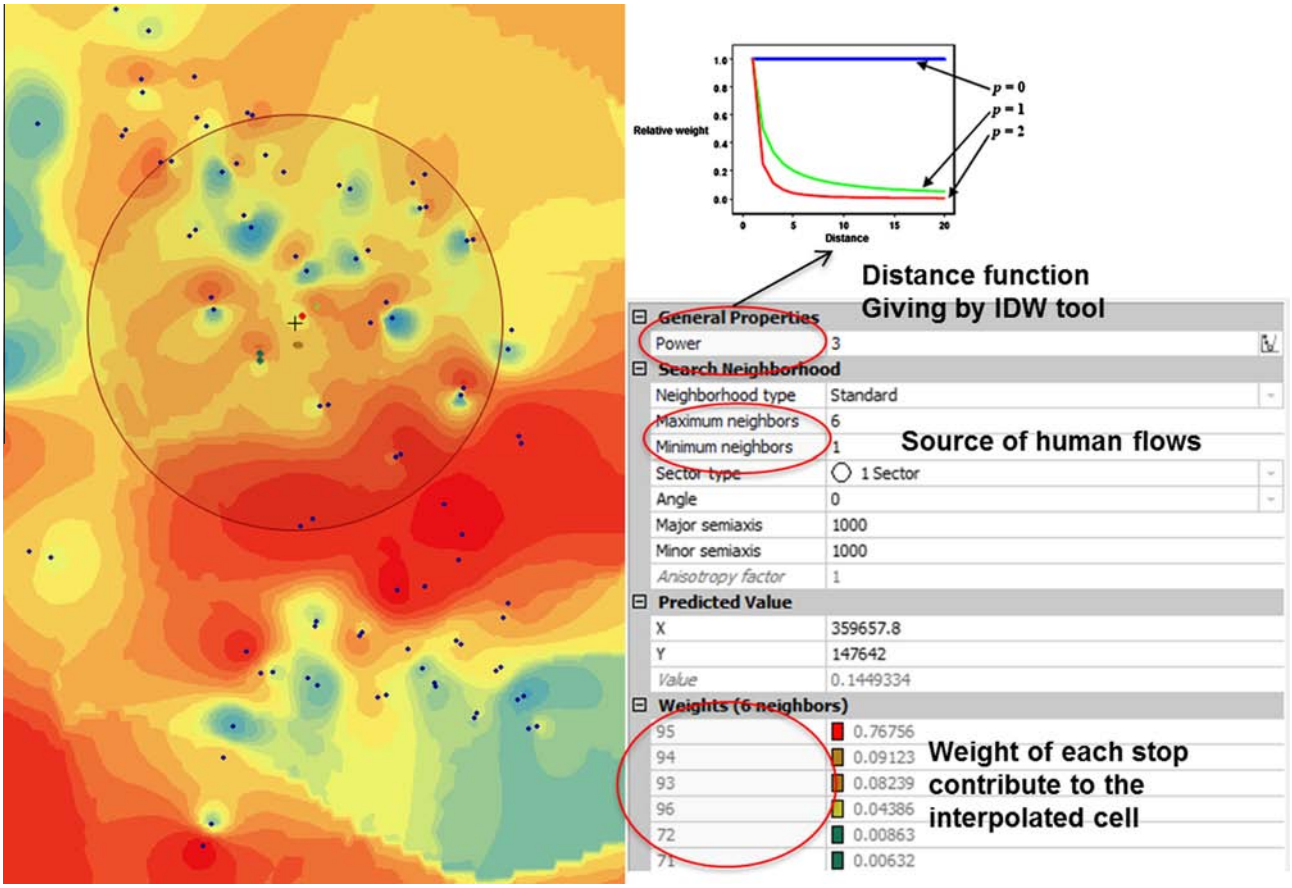


Fig. 7. Calculated weight of each bus stop contributing to the final probability of cells from the IDW interpolation tool in ArcGIS.

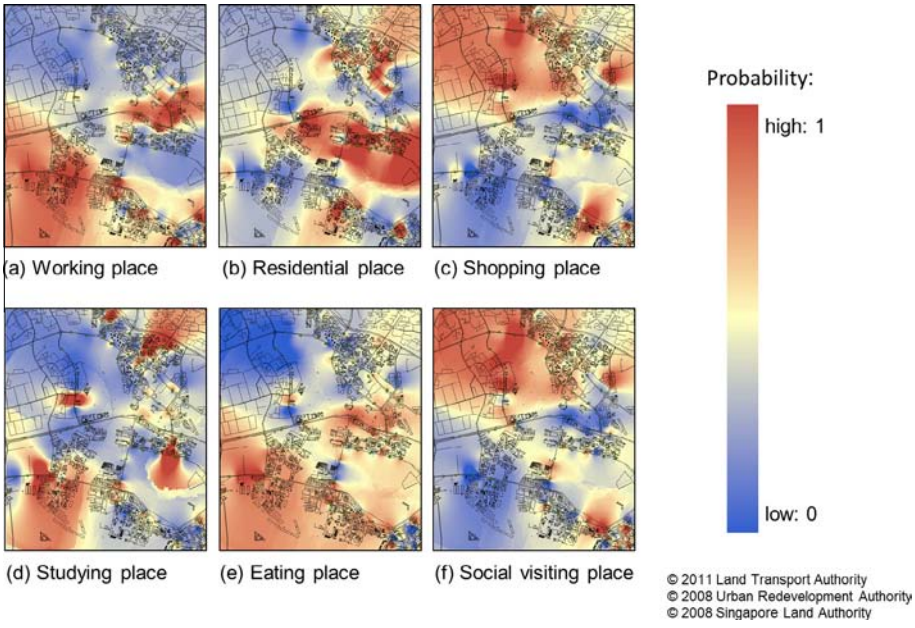


Fig. 8. The results of the interpolated probability landscape of each activity class in the Jurong East area.

Based on these three criteria, six travel purposes are selected: shopping, working, staying at home, studying, eating, and social visiting. It should be noted that social visiting combines a variety of activities, including entertainment, accompanying someone as a colleague or friend, and religious matters. These activities are performed relatively infrequently, take a comparatively shorter time, occur irregularly, and present similar mobility patterns. The transfer mode occurs when people are heading to the next travel service. This is excluded from our analysis and filtered out during the preliminary data processing. These original travel purposes

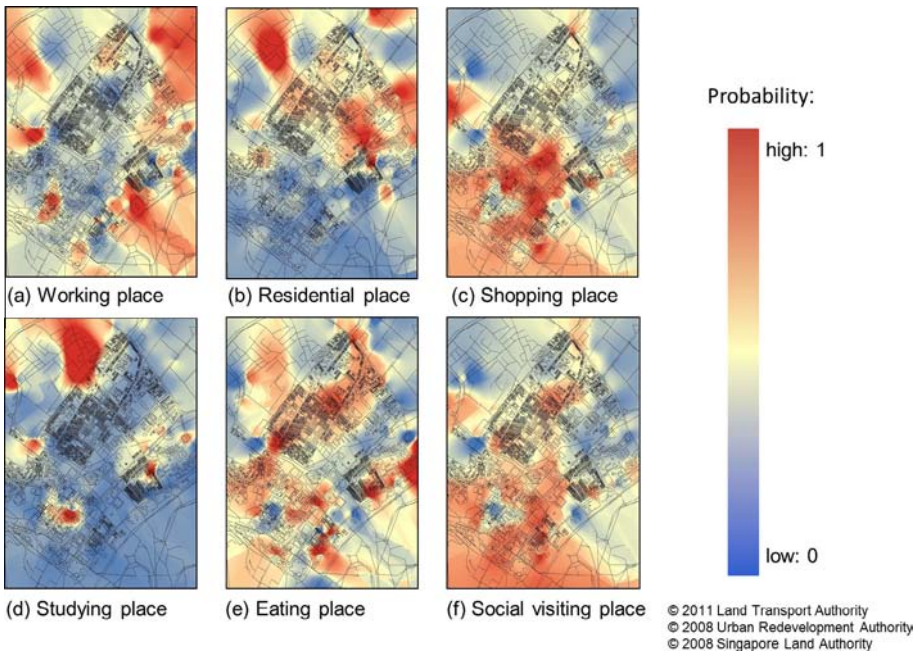


Fig. 9. The results of the interpolated probability landscape of each activity class in the Rochor area.

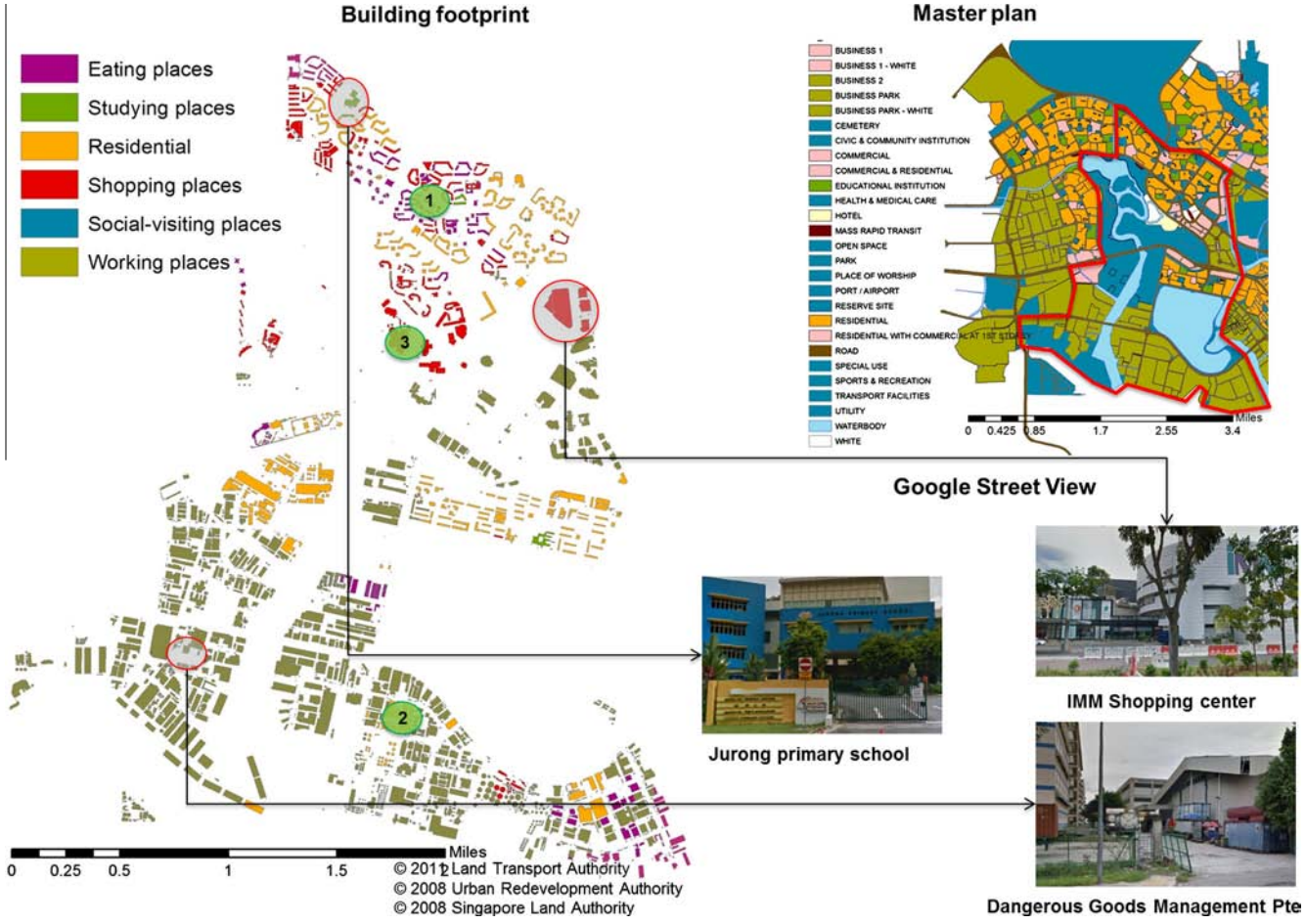


Fig. 10. The results for Jurong East: inferred building functions (top left) compared to the Master Plan (top right) and to Google Street View (bottom right). Note that land use types are aggregated into five categories for a better comparison. Shopping areas are linked to wider categories of land use for commercial and business places in the Master Plan. The same rules apply to Fig. 11.

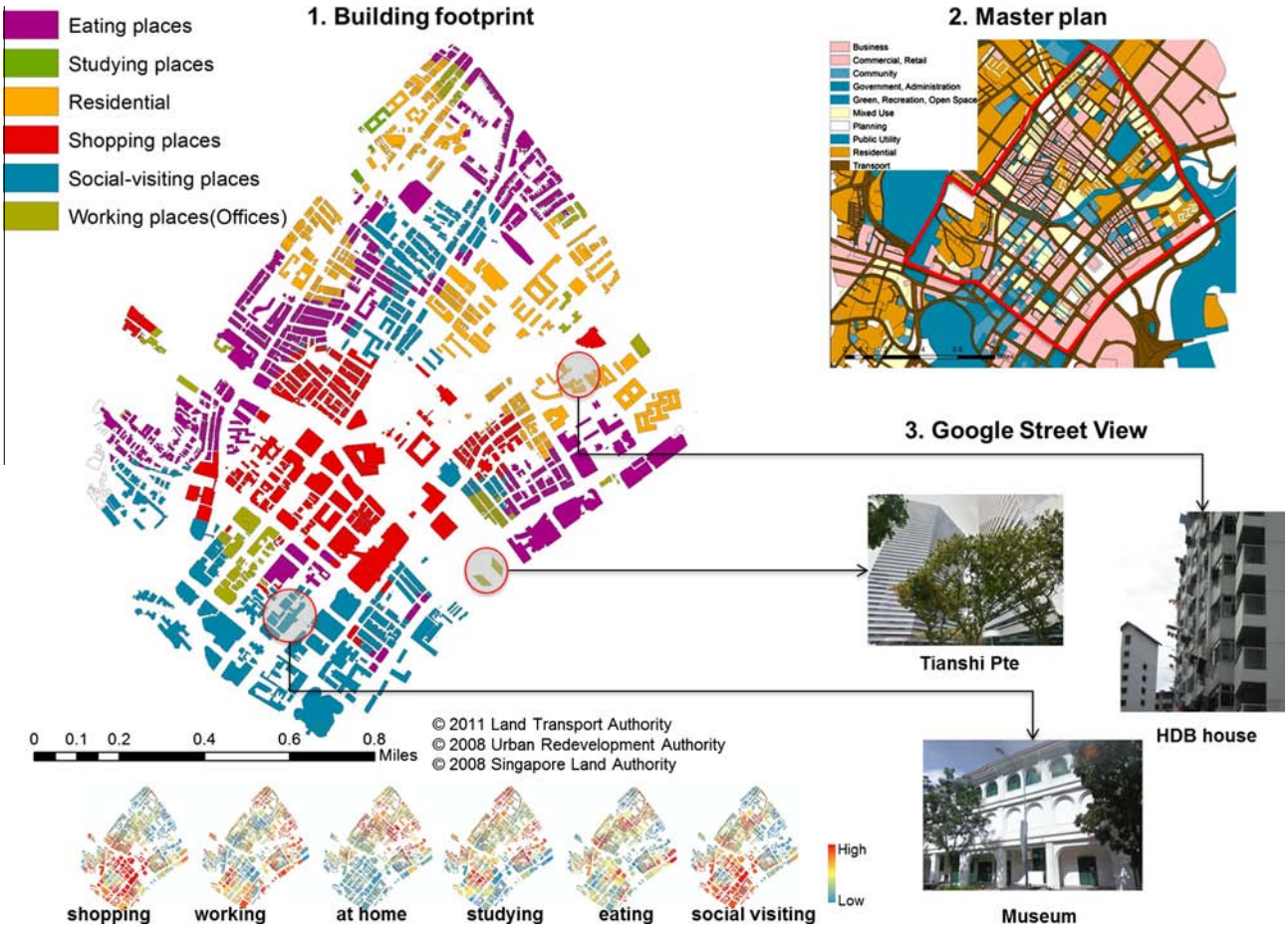


Fig. 11. The results for Rochor: inferred building functions (top left) compared to the master plan (top right) and to Google Street View (bottom right). Probability distribution of the six daily activities at the building level in the Rochor area (bottom left).

(right) have been renamed according to their corresponding daily activities (left), as shown in Table 2. The selected activities take place in functional areas, which are shown in Table 1 (6), and cover the main functions of urban space.

4.3. An experiment: The case study

We illustrate our method step by step with a case study from Jurong East in Singapore, the location of which is shown in Fig. 5. Jurong East is part of Jurong, the largest town in Singapore, which has the second largest resident population and which contains multiple land uses such as education, commercial, residential, and industrial. We chose an area approximately 1500 * 2000 m, totaling approximately 3.214 million square meters, and our statistical data covers seven days' worth of trips from 136 bus stops located inside or on the border of the selected area. After the preliminary data processing, we extracted an average of 128,000 valid trip records per day. The results were mapped to 2737 buildings, and where we used the directly footprint, some buildings were decomposed into several smaller ones.

To evaluate the feasibility of our method, another case study was conducted in Rochor, Singapore (also shown in Fig. 5). Rochor is located in the central region of Singapore and contains many commercial buildings, a few residential houses, and other functional services. We chose an area approximately 5000 * 3000 m, totaling approximately 17.857 million square meters, and the statistical data cover trips from 188 bus stops located inside or on the border of the selected area. In this case, after the preliminary data processing, we obtained an average of 189,000 valid trip records per day, and the results were mapped to 3909 buildings.

To emphasize the spatial resolution, in the experiment, the number of travel records is much denser than the number of buildings. On average, each building will contain people from approximately 10 bus stops, and each bus stop has an average 5000 travel records per day.

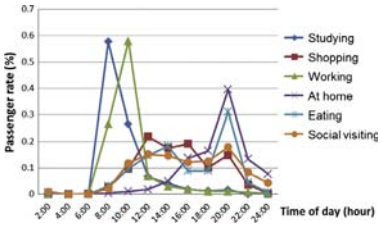
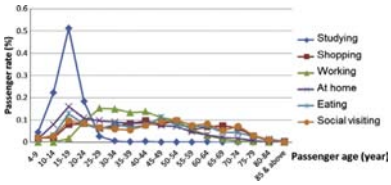
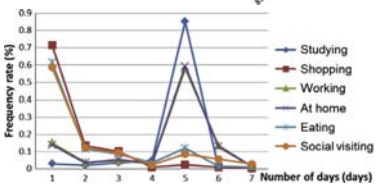
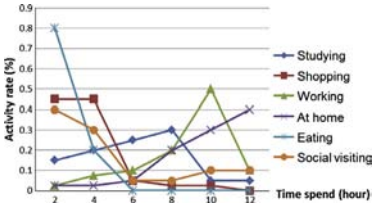
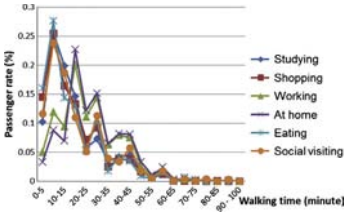
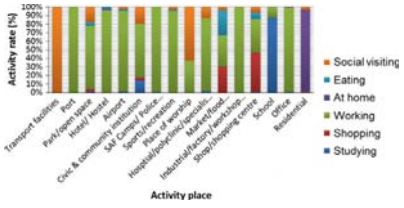
4.3.1. Trip classification

The original smart-card data provide information about trip ID, passenger ID, trip transfer time, starting time, travel time, fare, and distance. We estimate the staying time by calculating the interval time between two trips, generated in/out of a selected area, with the same passenger ID. Frequency is a count of the time that the same passenger ID appears in the selected data sets on different dates. Fig. 6, which includes Table A (top left), shows examples of the generated data structure, which is the result of the preliminary data processing.

As shown in the framework, after the preliminary data processing, we perform a trip classification using the Bayesian classifier with input from the analyzed results. Fig. 6 shows example tables, including the generated trip records shown in Table A (top left), the prior probabilities shown in Table B (top right) and Table C (middle), which are the results of the classification showing the inferred probability distributions of daily activities linked to each bus stop.

In the first step, the value of the prior probability $P(a_i|c)$ is read from the prior probability table. To provide a clearer interpretation, we marked the attribute "activity frequency" as an example.

Table 1
Six patterns of travel behavior found in the survey data and used to build clustering prototypes for urban activities.

Pattern	Statistical results	Explanation
Alighting time		When do people start their activities? Two hours is used as the basic unit in this analysis. This shows that there are different peak hours for different travel purposes: going to work and to school mainly occur in the morning; going home occurs in the evening; eating occurs at lunch and dinner times; and social visiting and shopping are evenly distributed throughout the day
Age distribution		Who is the major group traveling for a certain purpose? Age is used to divide people into different groups in this analysis. Going to school occurs mainly among teenagers, and working covers all age groups but is concentrated for young people. The other activities have comparatively similar distributions
Activity frequency		How many times did people engage in the same activity in the past seven days? The frequency of activities shows how often people made the same trip during the past seven days. As shown in the left figure, going to work, going home and going to school occur comparatively regularly, whereas the others occur occasionally
Staying time		How long do people spend on certain activities? Staying time is estimated as the period between two trips used to perform the activities. There is no direct information in the survey; therefore, we summarized this from the literature, including statistical data on working hours obtained from official websites (Government of Singapore, 2012) and other survey data about time use (U.S.Statistics, 2011) as well as from some ad hoc interview data
Walking distance		If taking the bus, how far is the bus stop from the destination by walking? The walking distance shows how long it takes to travel from the bus stop to the destination by walking. It is used for defining the spatial-correlation coefficients used in Section 3.5, where the distribution of activities types is drawn
Location		Where do people perform the activities? The activity place shows what type of place people travel to for certain activities. It is used to define the correlations between travel purposes, activities, and building functions

A different frequency refers to a different value of a prior probability. The prior probability is read from Table B. As such, there are tables of prior probability distributions for the other attributes. In the second step, after checking all of the individual attributes' prior probabilities, we use Eq. (1) in Section 3.4 to calculate the probability of activities, thus finding the most likely activity that motivated this trip. Table C is the posterior probability distribution of the six daily activities linked to one stop, e.g., bus stop "284**", which has the highest probability of education, abbreviated as "e", in the table. This means that the majority of people disembarking at this bus stop are traveling for education, which implies that there might be an educational institute nearby. The chart figure (bottom) in Table C shows the probability distributions of the six activities at 136 bus stops in Jurong East. The probability distributions of the six daily activities are labeled in six different colors. The *x*-axis shows the bus stop ID, while the y-axis shows the proportion of activities at each stop. We have highlighted stop "284**" from the chart figure. The color purple, which we use to represent studying, accounts for the largest proportion. An intermediate evaluation of the results is performed to check the general effectiveness of our estimation. We checked the buildings surrounding stop "284**" on Google Maps and determined that the closest building is a school, which explains why the main activity of going to stop "284**" is studying.

4.3.2. Spatial statistics of trips from bus stops to destinations

We used the IDW spatial analysis tool in ArcGIS to interpolate the probability distributions of certain functions that correspond to the daily activities performed in the Jurong East area (shown in Fig. 7). We set the number of neighborhood stops to a minimum of 1, assuming that everywhere can be reached, and a maximum of 10, assuming that people may come from 10 nearby stops. We noticed that 10 stops is more than what is observed in real situations, but our experiments show that changing the maximum

Table 2
Statistical data of daily travel purposes from HITS (only counting included public transport modes).

Urban activity (travel purpose)	Number of trip records
Social visiting (return from another home)	27
Transfer mode	31
Social visiting (entertainment)	68
Social visiting (sports/exercise)	111
Social visiting (religion-related matters)	166
Social visiting (medical/dental(self))	187
Social visiting (recreation)	216
Social visiting (to accompany someone)	219
Social visiting (personal errand/task(pay bill/ banking))	301
Social visiting (to drop-off/pick-up someone)	339
Social visiting (work-related business)	422
Eating (meal/eating break)	438
Social visiting(social visit/gathering)	1043
Shopping	1385
Studying (education)	4383
Working (go to work)	10,151
Staying at home(return home)	17,727

number only slightly changes the interpolated results, and the overall accuracy remains almost the same.

The results of IDW are presented in Figs. 8 and 9 and show the probability distribution of certain functions in Jurong East. For example, the top left image shows the areas mapped in red having a higher probability of being a working place than those in blue. Compared with the Master Plan, the overall distributions appear correct: working places are mostly located in the southern area, commercial places are located in the southern area, and residential places are located in the southern and middle areas; in contrast, comparatively few studying places exist. In the next step, we perform a spatial union of the overall distributions of the six activities and building footprints. An example of this mapping to building footprints is shown in Fig. 11 (bottom left).

4.4. Analysis of results and discussion

The final results for the Jurong East area are shown in Fig. 10 on the left side; the 2008 Master Plan[1] is shown at the top right side and shows the planned land use of this area. The 2320 building footprints are marked with different colors representing the most likely functions of the buildings. Some buildings are randomly selected for comparison using Google Street View. We show three of them in Fig. 10 (right bottom). Similarly, the results for the Rochor area are shown in Fig. 11 using the same layout. In addition, the probability distributions of each function, which are the intermediate results of our method, are shown in Fig. 11 (left bottom).

In the following section, we discuss in detail the accuracy of our results, causes of any errors and corresponding solutions for future work. We verified our method by comparing our results to the Master Plan, Google Street View and to survey data directly collected, which we contend is a form of ground trothing. We catalogue these results below under several points, and this provides the reader with a summary of the relative success of the method and its predictive analysis. These points are as follows:

1. We have compared our results inferred from data from 2011 to the Master Plan 2008 to estimate their compatibility and to obtain a sense of the overall distribution of the inferred functions. As defined in Section 3.1, the Master Plan gives the global constraints on land use and building functions. In Fig. 10 (right), we map the Master Plans with an almost similar color code to facilitate an easy and clear comparison. We say 'almost similar' because we aggregated the land use categories given by the Master Plan. For instance, business use (pink) could be restaurants as eating places (purple) and small shops as shopping places (red). The rest, as we show, includes residential buildings (orange), which are located on top of the residential area (orange) and mixed use area (light yellow); working places (yellow green), which are located on top of industrial and office areas (yellow green); and schools (bright green), which are located on top of educational land use areas (bright green). Note that the land use plan and the inferred building functions are not at the same spatial scale, validating our method only at the block level. The accuracy was previously demonstrated in Figs. 8 and 9.
2. Selected landmarks are compared to information from Google Street View at the building scale. As shown in Fig. 10 (bottom right), we selected various landmarks, including a school surrounded by residential houses, a shopping mall surrounded by industrial and residential areas and a typical industrial building. We show a similar picture in Fig. 11 (bottom right). All of these were matched with the information from Google Street View. This demonstrates that our method can effectively infer functions at the building scale with a high percentage of correct cases. It can detect more detailed building use than that given by the Master Plan, with the results distinguishing industry and office buildings from shopping malls.
3. We have compared our experimental results with ground truth data. The result of our method is a probability distribution of each function of a building, and we map each building with the function that has the maximum value. The results indicate the dominant use of the building instead of the only use of the building. We visited parts of the site to investigate whether a building has the dominant function that we calculated. The determination of the function relies on facility types based on the statistical results in Table 1. For instance, working space mainly refers to office buildings and industrial buildings, and social visiting can occur in buildings that provide mainly non-commercial services, such as libraries, community centers, and churches. The results are shown in Table 3.

However, there were errors, although not substantial, in our results concerning building functions, and this is to be expected because it is impossible to have complete accuracy using this statistical approach, although it is heavily driven by observed data. One reason for these errors is that the data sets were not synchronized. The survey data and the Master Plan were made in 2008, whereas the smart-card data were collected in 2011, and the

Table 3
Comparing our predicted results with a sample of ground truth data.

Site	Size (approximate)	Number of counted buildings	Number of buildings mapped with an incorrect function	Percentage of correctness (%)
	500 * 500 m	140	24	82.85
	1200 * 1200 m	114	15	86.84

[1] The Master Plan in Singapore is the statutory land use plan that guides Singapore's development in the medium term over the next 10 to 15 years. http://www.ura.gov.sg/uol/master-plan.aspx?p1=View-Master-Plan accessed in 2014.

Table 4
A sample of travel survey data with selected relevant attributes shown.

ID	Age	Occupation	Origin postcode	Destination postcode	Start time	Arrival time	Activity place	Trip purpose	Travel mode	Walking time	Frequency
1	40	Manager	5****6	5****3	6:25	9:15	Clinic	Work	Public bus	10	1
2	25	Retired	5****3	5****6	9:30	12:15	Home	Go home	Public bus	10	4
3	69	Retired	5****6	5****9	12:30	14:00	Shops	Shopping	Walk	15	5

Google Street View data are from recent internet updates (within the last year, 2013). Although this is a small difference in time, this may still cause some errors. In addition to the unsynchronized data sets, we should note other causes of the errors, and four typical errors are marked in Fig. 10 with green circles. The errors, their causes, and possible solutions will now be discussed.

1. **Border effects**. Stops have more influence on buildings located in nearby blocks than on buildings located on the other side of the street, as shown by the green circles in Fig. 10. However, this difference cannot be easily quantified. In the micro view, to solve this problem, streets that are geographic borders of districts should be considered as barriers with respect to the inverse distance weight functions used in Section 3.5. In the macro view, the accessibility of the global street network should be measured. In fact, how street types influence urban activities is another difficult topic. In this paper, to focus on the data mining method, we have neglected this.
2. **Scale issues**. Because building functions are calculated using the types of trips ending at surrounding bus stops, the results are closely related to the density of the bus stops. For areas with a dense bus stop distribution, the results will be more stable. Conversely, the results will be heavily influenced by small sample spaces, as shown in Error 2. To some degree, we can say that this method exhibits a better performance at the block scale (as shown in Fig. 8) than at the building scale (as shown in Fig. 10). A possible solution is that of fusing multi-source data, such as GPS traces on taxis, and this may increase the resolution of the data.
3. **Mobility patterns inefficiency**. Peoples' travel purposes are motivated by daily activities and result in different mobility patterns. The mobility patterns analyzed from the survey data in this paper are insufficient to distinguish many activity classes. For instance, social visiting and eating, as shown in Fig. 8, have very similar probability distributions with respect to mobility. Reflecting on the reality of the situation, problems exist when we attempt to further distinguish service buildings, such as community centers, libraries, and churches, which we note as Error 3. From a technical point of view, more features are needed to achieve more precise identification. This can be enhanced by adding additional features derived from integrated knowledge about the city as a social system.
4. **Influence from other travel modes**. The functions of buildings were inferred from travel via the public transportation system, meaning that travel by private car, bike or walking were excluded. For some cities, such as those in North America, private cars are the most popular choice of travel mode for the majority of the population. Applying our method using only smart-card data will not achieve satisfactory results for these populations. However, in our case studies in Singapore, where the public transportation system is the major travel mode, smart-card data are rich enough to represent the functions of the main urban spaces. We consider that these methods have promise for dense and large urban areas, such as London and Tokyo, where a large proportion of travelers use buses and trains.

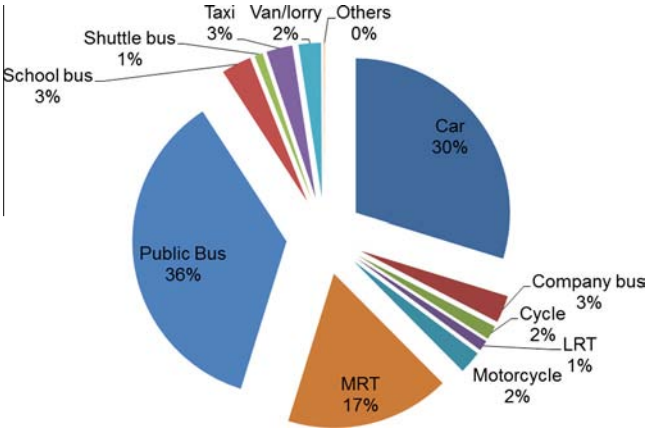


Fig. 12. Travel mode share in 2008.

5. Conclusions

In this paper, we have introduced a method of combining survey and smart-card data to infer information about social activities and to generate insights into urban building functions. Specifically, we proposed a probabilistic model to infer information about daily activities from individuals' travel and infer building functions from corresponding daily activities. The model can also be used as an alternative way of driving data acquisition and analysis of functional urban spaces. Moreover, this is a comparative approach that provides not only a quantitative estimation method but also insights into how people use urban space in reality. Because urban movement data are becoming increasingly more accessible from smart cards and related GPS capture, applying our method to historical data may help us obtain a better understanding of the dynamics of urban spaces. The information we infer in this way is thus extremely useful to planners in obtaining efficient and updated information about urban functionality, which enables them to better manage their short-term and long-term plans. Note that the input parameters of our probabilistic model generated from the smart-card data are simple and only require boarding and disembarking times and locations. This information could also be obtained from other sources of urban mobility data such as taxi data. Therefore, we believe that our method has a wider applicability and can be applied to other types of data with equal or higher spatiotemporal information quality than that collected using smart cards.

There is much work to do and many problems to study to counter the limitations of our approach. In addition to the issues discussed previously, namely, border effects, scale issues, pattern insufficiency and transportation data insufficiency, there are additional questions that we have not discussed here but that must be addressed to make further progress. First, the effect of long-distance travel using the metro system should be considered in the future. This is excluded from our paper because the density of metro stations is much lower than that of bus stops, and this would cause serious scale problems if we were to carry out the same analysis using these data. Second, we conducted our experiment in Singapore, which has a strong Master Plan and where land use is not highly mixed. Whether our method can be adapted for substan-

tially more complex and mixed land use areas where people live in different areas and use transit systems in very heterogeneous ways is a topic we need to explore in the future. Third, the two essential parts of our method, namely, the Bayesian classifier and the IDW function, are used as a first attempt at exploring the possibility of using these types of big data to determine the dynamics of spatial functionality. We do not disaggregate this analysis by time of day, only by spatial area, and there are interesting experiments still to be carried out in these directions. Last but not least, other algorithms can be used to try and improve the accuracy of the results, and in future work, these will be important issues to explore.

In the future, we will further investigate and apply our method to cities in different countries. We will also advance our method by integrating multi-disciplinary knowledge, addressing the scale issues by introducing crowd-sourced data and by avoiding the border effects using a network analysis. In general, we consider integrated techniques and combined information as a way to make progress.

Acknowledgments

This work was established at the Singapore-ETH Centre for Global Environmental Sustainability (SEC), co-funded by the Singapore National Research Foundation (NRF) and ETH Zurich. The authors would like to express their sincere gratitude to the Singapore Land Transport Authority forsupporting this research and providing the required data.

Appendix A

See Fig. 12 and Table 4.

References

Agard, B., Morency, C., & Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *12th IFAC Symposium on Information Control Problems in Manufacturing-INCOM* (pp. 17–19).

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.

Bagchi, M., & White, P. R. (2005). The potential of public transport smart card data. *Transport Policy, 12*(5), 464–474. <http://dx.doi.org/10.1016/j.tranpol.2005.06.008>.

Carlin, B. P., & Louis, T. A. (1997). Bayes and empirical bayes methods for data analysis. *Statistics and Computing, 7*(2), 153–154. http://dx.doi.org/10.1023/a:1018577817064.

Chakirov, A., & Erath, A. (2012). *Activity identification and primary location modelling based on smart card payment data for public transport*. Zurich, Switzerland: Eidgenössische Technische Hochschule Zürich, IVT, Institute for Transport Planning and Systems, Zurich, Switzerland.

Cheong, C. C., & Toh, R. (2010). Household Interview Surveys from 1997 to 2008 – A Decade of Changing Travel Behaviours.

Decatur, S. E. (1989). Application of neural networks to terrain classification. Neural Networks, *IJCNN, International Joint Conference* (pp. 283–288).

Gehl, J. (1987). *Life between buildings: Using public space*. New York: Van Nostrand Reinhold.

Government of Singapore (2012). Yearbook of Statistics Singapore. <http://www.singstat.gov.sg/> Accessed 15.03.13.

Green, N. (2007). Functional polycentricity: A formal definition in terms of social network analysis. *Urban Studies, 44*(11), 2077–2103.

Jane, J. (1961). *The death and life of great American cities*. New York: Eandom House.

Jensen, F. V. (2009). Bayesian networks. *Wiley Interdisciplinary Reviews: Computational Statistics, 1*(3), 307–315.

Kajtazi, B. (2010). *Measuring Multifunctionality of Urban Area: Advanced GIS Analysis for Measuring Distance, Density, Diversity and Time of Urban Services*. Saarbrücken: LAP Lambert Academic Publishing.

Kocabas, V., & Dragicevic, S. (2012). Bayesian networks and agent-based modeling approach for urban land-use and population density change: A BNAS model. *Journal of Geographical Systems*, 1–24.

Liang, L., Anyang, H., Biderman, A., Ratti, C., Jun, C. (2009). Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. *Intelligent Transportation Systems, 2009. ITSC '09. 12th International IEEE Conference* (pp. 1–6), 4–7 October.

Liu, Y., Wang, F., Xiao, Y., & Gao, S. (2012). Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning, 106*(1), 73–87.

Lu, Y., & Liu, Y. (2012). Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems, 36*(2), 105–108.

Mccloskey, J. T., Lilieholm, R. J., & Cronan, C. (2011). Using Bayesian belief networks to identify potential compatibilities and conflicts between development and landscape conservation. *Landscape and Urban Planning, 101*(2), 190–203.

Mosegaard, K., & Tarantola, A. (2002). 16 Probabilistic approach to inverse problems. In H. K. William, H. K. P. C. J. Lee, & K. Carl (Eds.), *International geophysics*. Academic Press, pp. 237–265.

Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies, 24*, 9–18.

Ordóñez Medina, A., Erath, A. (2013). Estimating dynamic workplace capacities using public transport smart card data and a household travel survey. *Presented at Transportation Research Board (TRB) 92nd Annual Meeting*. Washington, D.C.

Park, J. Y., Kim, D. J., & Lim, Y. (2008). Use of smart card data to define public transit use in Seoul∗ South Korea. *Transportation Research Record: Journal of the Transportation Research Board, 2063*(-1), 3–9.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 1–20 (ahead-of-print).

Pelletier, M. P., Trépanier, M. and Morency, C. (2009). *Smart card data in public transit planning: A review*. CIRRELT.

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 A. Salah, T. Gevers, N. Sebe, & A. Vinciarelli (Eds.), *Human behavior understanding* (pp. 4–25). Berlin Heidelberg: Springer.

Qi, G., Li, X., Li, S., Pan, G., Wang, Z., Zhang, D. (2011). Measuring social functions of city regions from large-scale taxi behaviors. *Pervasive Computing and Communications Workshops (PERCOM Workshops)* (pp. 384–388).

Ratti, C., Williams, S., Frenchman, D., & Pulselli, R. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B Planning and Design, 33*(5), 727.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence* (pp. 41–46).

Rodrigue, J.-P., Comtois, C., & Slack, B. (2013). *The geography of transport systems*. London: Routledge.

Roth, C., Kang, S., Batty, M., & Barthélemy, M. (2011). Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PloS one, 6*(1), e15923.

Sun, J., Yuan, J., Wang, Y., Si, H., & Shan, X. (2011). Exploring space–time structure of human mobility in urban space. *Physica A: Statistical Mechanics and Its Applications, 390*(5), 929–942.

Taleai, M., Sharifi, A., Sliuzas, R., & Mesgari, M. (2007). Evaluating the compatibility of multi-functional and intensive urban land uses. *International Journal of Applied Earth Observation and Geoinformation, 9*(4), 375–391. http://dx.doi.org/10.1016/j.jag.2006.12.002.

Toole, J. L., Ulm, M., González, M. C., Bauer, D. (2012). Inferring land use from mobile phone activity. *Proceedings of the ACM SIGKDD International Workshop on Urban Computing* (pp. 1–8).

U.S.Statistics (2011). *American Time Use Survey*. <http://www.bls.gov/tus/>.

Watson, R. T., Noble, I. R., Bolin, B., Ravindranath, N., Verardo, D. J., & Dokken, D. J. (2000). *Land use, land-use change, and forestry: A special report of the intergovernmental panel on climate change*. Cambridge University Press.

Yuan, J., Zheng, Y., Xie, X. (2012). Discovering regions of different functions in a city using human mobility and POIs. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 186–194). Beijing, China: ACM.

Zhong, C., Arisona, S. M., Huang, X., Batty, M., & Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, 1–22 (ahead-of-print).