View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Institutional Knowledge at Singapore Management University

Singapore Management University
Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

8-2015

On mining lifestyles from user trip data

Meng-Fen CHIANG
Singapore Management University, mfchiang@smu.edu.sg

Ee-peng LIM
Singapore Management University, eplim@smu.edu.sg

DOI: https://doi.org/10.1145/2808797.2808906

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

Part of the Databases and Information Systems Commons, and the Transportation Commons

Citation

CHIANG, Meng-Fen and Ee-peng LIM. On mining lifestyles from user trip data. (2015). *ASONAM '15: Proceedings of the 2015 IEEE/ ACM International Conference on Advances in Social Networks Analysis and Mining 2015: August 25-28, Paris.* 145-152. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/3079

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

On Mining Lifestyles from User Trip Data

Meng-Fen Chiang
Living Analytics Research Center
Singapore Management University
80 Stamford Road, Singapore
Email: mfchiang@smu.edu.sg

Ee-Peng Lim
Living Analytics Research Center
Singapore Management University
80 Stamford Road, Singapore
Email: eplim@smu.edu.sg

Jia-Wei Low
Living Analytics Research Center
Singapore Management University
80 Stamford Road, Singapore
Email: jwlow@smu.edu.sg

Abstract—**Large cities today are facing major challenges in planning and policy formulation to keep their growth sustainable. In this paper, we aim to gain useful insights about people living in a city by developing novel models to mine user lifestyles represented by the users' activity centers. Two models, namely ACMM and ACHMM, have been developed to learn the activity centers of each user using a large dataset of bus and subway train trips performed by passengers in Singapore. We show that ACHMM and ACMM yield similar accuracies in location prediction task. We also propose methods to automatically predict "home", "work" and "others" labels of locations visited by each user. Through validating with human-labeled home and work locations, we show that the accuracy of location label assignment is surprisingly very good even using an unsupervised method. With the location labels assigned, we further derive interesting insights of urban lifestyles at both individual and population levels.**

I. Introduction

Motivation. Urban cities nowadays are densely populated as they become the centers of both business and culture. To cope with large population of people working and living in these cities, it is pertinent for city planners to create various systems to meet the needs of city dwellers. These include a transportation system that provides the means for people to travel easily within and between cities, healthcare system that covers good medical and hospital care with affordable costs, etc..

Smart urban city's systems have to be continuously and upgraded so as to adapt to the evolving city lifestyle. In the past, this is non-trivial without conducting a large-scale user survey. With the increased use of digital sensors in the city, massive human data are recorded each day allowing us to discover insights about user lifestyles which can in turn be used to improve the design of public systems and services.

Objectives. In this paper, we focus on analyze human mobility from bus and subway transaction records for modeling lifestyles of users in a city. We define the lifestyle to consist of two components. The first component characterizes a user by clusters of stay intervals at locations where the user visits regularly, referred to as *activity centers*. The second component consists of labels assigned to activity centres to provide additional semantics. We call the combined components the *lifestyles*. By considering the two, we can derive the activity interests of users as well as assign their visited locations with one of the three semantic labels: HOME, WORK, and OTHERS. Our research objective is to develop models to automatically summarize the lifestyle patterns at both *spatial-temporal* and *semantic* levels without resorting to labor-intensive efforts.

Apart from acquiring the human mobility data, the above research task is challenging because the mobility data traces are raw transactional records and do not capture human trajectory at all times nor reveal the users' final destinations. We therefore do not have the trajectory data to derive specific locations that are homes or offices. It is however interesting to discover activity centers of the user where each activity center is represented by a set of nearby stations visited by the user on a regular basis. We introduce two probabilistic models called **A**ctivity **C**entre **M**ixture **M**odel (ACMM) and **A**ctivity **C**enter **H**idden **M**arkov **M**odel (ACHMM) to mine individual lifestyle patterns. We show that the user's lifestyle patterns can effectively predict their coarse grained locations and be used to study population trends.

The following summarizes our main contributions in the paper.

- Based on a real world transportation dataset, we develop ways to convert raw transactional records into user trips, and further define the stay durations of a user. The stay durations are subsequently clustered into activity center(s) that define the user's lifestyle pattern.
- We develop ACMM and ACHMM lifestyle models to learn the activity centers of a user based on the periodical mobility patterns observed in a very large subway and bus trip dataset. The former is based on Non-Bayesian Gaussian Mixture Model while the latter is based on Gaussian Hidden Markov Model. Each activity center is a set of locations the user spends time regularly. Unlike the previous models, ACMM and ACHMM are specially designed to model time intervals of user spending time at locations as opposed to time points of user locations often observed in mobile phone call detail records.
- We evaluate both ACMM and ACHMM using a future location prediction task. The two models are shown to perform better than the frequency based baseline model. The performance of ACMM and ACHMM are comparable although ACHMM offers a richer model to capture the transition between activity centers. We also study the behavior of ACHMM under different parameter settings, derive insights about users' home and work patterns, and profile the home and work regions in Singapore.

Paper Outline. The remainder of the paper is organized as follows. We first define several important concepts and terms before introducing the user lifestyle modeling problem in Section II. Our proposed lifestyle models are given in Section III and evaluated in Section IV. We cover the related works in Section V before concluding the paper in Section VI.

II. Problem Formulation

In this section, we first introduce the concept of activity center and then define the problem of user lifestyle modeling. We want to model a user's lifestyle by distilling his activity centers with spatial-temporal properties and the corresponding activity labels. Each activity center is represented by a cluster of similar time periods and stay locations. Furthermore, we assign each activity center with one of the three semantic labels: HOME, WORK and OTHERS. Finally, we demonstrate how to utilize the proposed model for predicting a user's locations.

A. *Definitions*

Our research assumes that in a public transportation dataset, there are subway stations and bus stops collectively called *stations*. Users make trips among stations using stored fare cards. To determine where users spend time on, we require their data at the trip level. Nevertheless, trip data are not always readily available as people may have to change between bus and MRT services in a single trip. For example, a user may start a journey from home to workplace by taking a bus from home to the nearest subway station before riding on a train to the workplace next to a downtown subway station.

We thus define a trip to consist of a series of leg records representing the different legs of the trip. Each *leg record* is represented by a pair of start and end stations, i.e. g_s and g_e respectively, the time departing from g_s and the time arriving at g_e. We denote the leg record as $r_i = \langle g_{s,i}, t_{s,i}, g_{e,i}, t_{e,i} \rangle$.

We construct a trip from multiple consecutive leg records as follows. We first extract a maximal sequence of leg records $r_1, r_2, \cdots, r_l$ such that $t_{s,i+1} - t_{e,i} < \eta$ $(1 \leq i < l)$ where η represents the inter-leg time gap threshold. This threshold applies to between train legs, between bus legs, as well as between train and bus legs. We then construct a *trip* as $\langle g_{s,1}, t_{s,1}, g_{e,l}, t_{e,l} \rangle$ with l denoting the trip length. Note that when a leg record cannot be combined with any other leg records to form a multi-leg trip, the leg record itself is then a trip of length one.

To determine an appropriate inter-leg time gap threshold η to combine legs into trips, we examine the inter-leg time gaps for the subway and bus transaction dataset to be described in Section II-B. Figure 1 shows the distribution of time gaps between two consecutive legs between 1 and 30 minutes. There are negligible inter-leg time gaps beyond 30 minutes. This distribution can be fitted by a Gamma distribution with shape = 2.24 and scale = 0.7. As the probability of inter-leg time gap less than 15 minutes according to this fitted distribution is larger than 90%, we set η = 15 minutes.

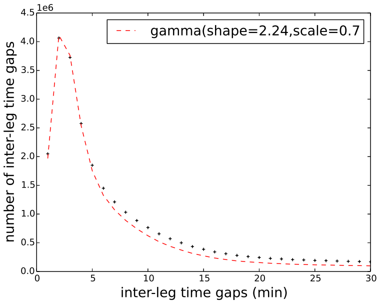


Fig. 1. Inter-leg time gap distribution

We denote the trips of a user u by $TR(u)$. From two consecutive trips of $TR(u)$, say $tr_i = \langle s_{s,i}, t_{s,i}, s, t_{e,i} \rangle$ and $tr_{i+1} = \langle s, t_{s,i+1}, s_{e,i+1}, t_{e,i+1} \rangle$, we define a *stay interval* of the user u as $(t_{e,i}, t_{s,i+1})$ which is the duration of time spent at some station g. Due to the definition of trip, every stay interval has a duration $\geq \eta$. We finally use $S(u)$ to denote the set of all stay intervals derived from $TR(u)$.

TABLE I. DATA STATISTICS

# bus stations	4,903
# subway stations	130
# users with $\geq$ 30 legs over 15 days	1,246,901
# selected users with stay intervals	338,637
# leg records by selected users	87,138,443
# trips by selected users	53,938,488
# stay intervals by selected users	12,559,660

Definition 1: (**Activity Center**) An activity center of user u is a cluster of stay intervals $S_k(u)$, $S_k(u) \subseteq S(u)$, that share similar start and end times.

Definition 2: (**Geo-Focus of Activity Center**) Given an activity center $S_k(u)$, the geo-focus $GF_k(u)$ is a set of stations the user u stays at during the respective stay intervals in $S_k(u)$. For example, the typical stations of the activity center from late night to early morning are near the user's home.

Definition 3: (**Lifestyle**) A lifestyle of a user u is a set of K tuples $\langle S_k(u), GF_k(u) \rangle$'s, where K denotes the number of activity centers.

We now define the problem of user lifestyle modeling as follows.

User Lifestyle Modeling Problem. Given a sequence of trips $TR(u)$ of user u, the problem of user lifestyle modeling is to design a probabilistic model that generates user's lifestyle, i.e., $\{\langle S_k(u),\, GF_k(u) \rangle\}$ for $1 \leq k \leq K$, which characterizes u's periodical stay patterns.

In the above problem formulation, K is an input parameter. Several applications could utilize the user lifestyle models in interesting ways. For example, a targeted marketing application could utilize the user's lifestyle model to determine the suitable time to serve ads or discount coupons to a target consumer. City planners can utilize the activity centers of the user population to estimate demand for municipal services (e.g., childcare, hospitals, etc.). With the regular lifestyle patterns of a user, it is also possible to predict the user's location during a given time interval. We formally define the problem of location prediction task as follows.

Location Prediction Task. Given the sequence of historical stay intervals of a user u, $S(u)$, and a query stay interval $s = \langle t_{start},\, t_{end} \rangle$, we want to predict the station the user will visit during s.

B. *EZ-Link Dataset*

In this research, we obtained a dataset consisting of 4 billion bus and subway leg records generated by 5 million passengers in Singapore's transportation system. The transactions are recorded by passengers tapping their EZ-Link cards at the entries and exits of the subway stations, and when boarding and alighting from buses. Each leg record consists of a card id (which uniquely identifies the user), transportation mode (bus or subway), entry station, entry date/time, exit station and exit date/time. All these records were generated in January 2012.

To focus on users who are residents of Singapore, we selected users with at least 30 leg records over at least 15 days. We call this selected dataset the EZ-Link Dataset. Table I summarizes the statistics of EZ-Link Dataset.

Table II shows three trip legs of a user and two derived stay intervals. For example, the first two trip legs suggest that the user stayed nearby Serangoon subway station from at 22:02 to 13:39 as the user exited and entered Serangoon station at 22:02 and 13:39 respectively. Similarly, the last two trip legs suggest that the user stayed nearby Harbourfront subway station from

TABLE II. AN EXAMPLE OF TRIP LEGS AND DERIVED STAY INTERVALS

	Entry Time	Exit Time	Entry Station	Exit Station
Trip Legs	21:25	22:02	City Hall	Serangoon
	13:39	14:09	Serangoon	Harbourfront
	18:47	19:15	Harbourfront	Serangoon
	Start Time	**End Time**		**Stay Station**
Stay Intervals	22:02	13:39		Serangoon
	14:09	18:47		Harbourfront

14:09 to 18:47 as the user exited and entered Harbourfront at 14:09 and 18:47 respectively

III. PROPOSED LIFESTYLE MODELS

In a lifestyle model associated with a user, each activity center is a cluster of similar stay intervals belonging to the user who has some stay patterns. It also represents the user's preference to perform some activities at some locations (or stations). In one extreme, each activity center consists of a single stay interval at a single station (when K is very large) but such an activity center does not capture any regularity of user's movement patterns. In another extreme, creating only one single activity center consisting of all stay intervals belong to the user will likely make the stay intervals of the activity center overly incoherent. Determining a suitable criteria for forming activity centers and assignment of semantic labels to the activity centers are interesting research questions. We present the activity semantic labeling and the station semantic labeling in Sections III-C and III-D respectively.

In the following, we introduce two proposed models to determine the activity centers. The two models, Activity Center Mixture Model and Activity Center Hidden Markov Model, are described in the following.

A. Activity Center Mixture Model

The **A**ctivity **C**enter **M**ixture **M**odel (ACMM) is derived from Non-Bayesian Gaussian Mixture Model (GMM) on stay intervals and their stations to discover a Gaussian mixture with K components that best describes the observed stay intervals. As shown in Figure 2(a), ACMM is determined by grouping the stay intervals in $S(u) = (s_i)_{i=1}^N$ and the stations $G(u) = (g_i)_{i=1}^N$ into K clusters. Each cluster indicates a representative group of stay intervals and stations. The distribution of stay intervals is described by a Gaussian component $N(\mu_{z_i}, \sigma_{z_i}^2)$. The distribution of respective stations in that group is described by a Categorical distribution $Categorical(M, \mathbf{p}_{z_i})$, where M denotes the number of stations. The following equation gives the formal definition:

$$p(z_i) = Categorical(\varphi) \tag{1}$$

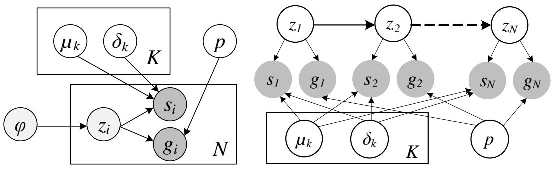
$$p(s_i|z_i) = N(\mu_{z_i}, \sigma_{z_i}^2) \tag{2}$$

$$p(g_i|z_i) = Categorical(M, \mathbf{p}_{z_i}) \tag{3}$$

where $|S(u)| = N$, z_i is the cluster assignment for the i-th stay interval s_i at station g_i, φ is a parameter of K outcomes such that $\varphi_k > 0$, for $k = 1, \cdots, K$, $\sum_1^K \varphi_k = 1$, and $\mathbf{p}_{z_i}$ is the parameter of Categorical distribution such that $\sum_{j=1}^M \mathbf{p}_k[j] = 1$ for every cluster k.

As $p(s_i|z_i)$'s and $p(g_i|z_i)$'s are independent of one another, the likelihood to generate a user u's trip data using ACMM can be expressed as:

$$\begin{gathered} p_{ACMM}(S(u), G(u)|\mu, \sigma, \mathbf{p}, \varphi) = \\ \prod_{i=1}^{N} \sum_{z_i=1}^{K} p(s_i|z_i, \mu_i, \sigma_i) \cdot p(g_i|z_i, \mathbf{p}) \cdot p(z_i|\varphi) \end{gathered} \tag{4}$$



(a) (b)

Fig. 2. (a) Activity Center Mixture Model; (b) Activity Center Hidden Markov Model

```
Input: $S(u), G(u)$: a set of trip events
Output:
$L^{(t)}$: log-likelihood at $t^{th}$ iteration;
$\theta^{(t)}$: $\mu$ and $\sigma$ at $t^{th}$ iteration;
$p^{(t)}$: $p$ at $t^{th}$ iteration;
$Z^{(t)}$: cluster assignments at $t^{th}$ iteration;
Initial assignment $\theta^{(0)}$, $p^{(0)}$;
$L^{(0)} \leftarrow computeLogLikelihood(p^{(0)}, Z^{(0)})$;
$t = 0$ ;
Repeat
$t++$ ;
$Z^{(t)} \leftarrow clusterAssignment(\theta^{(t-1)}, S(u), G(u))$;
$\theta^{(t)}, p^{(t)} \leftarrow updateParameters(Z^{(t)}, S(u), G(u))$;
$L^{(t)} \leftarrow computeLogLikelihood(p^{(t)}, Z^{(t)})$;
Until $abs(L^{(t)} - L^{(t-1)}) \leq \varepsilon$ ;
return $L^{(t)}$, ($\theta^{(t)}$, $p^{(t)}$), ($Z^{(t)}$);
```

Algorithm 1: Learning of Parameters (ACMM)

where μ and σ denote the set of μ_k's and σ_k's respectively. Note that ACMM models time intervals and stations, which is different from Periodic Mixture Model(PMM) [1] that models time points and spatial locations. Periodic Mixture Model(PMM) models human mobilities (check-in data) using separate spatial and temporal Gaussian components with social influence. The learning process of ACMM parameters is given in Algorithm 1.

B. Activity Center Hidden Markov Model

A user's lifestyle not only consists of the clusters of their activity centers, but also the underlying transitions between activity centers. Instead of considering the ordering of stay intervals, ACMM treats the data as i.i.d. As a result, ACMM fails to model the transitions. An intuitive example is that a user may go for happy hours after work in city center more often than go home immediately. Such transition probabilities between states will help to determine the next state based on current state. To capture such a factor in a lifestyle, we relax the i.i.d., assumption and explore the latest transitions in a stay interval sequence. In particular, we propose the **A**ctivity **C**enter **H**idden **M**arkov **M**odel (ACHMM), a variant of the Gaussian Hidden Markov Model that learns the clusters of stay intervals as hidden states (or activity centers), and parameters to generate the stay intervals and stations for each activity center.

The plate diagram of ACHMM is shown in Figure 2(b). Consider an observed sequence of stay intervals $S(u) = (s_i)_{i=1}^N$, we aim to derive the following probabilities: (1) the transition probabilities denoted as a $K \times K$ transition probability matrix M between the K clusters (or activity centers), (2) the parameters defining the emission probabilities of stay intervals, μ_k and σ_k for $k = 1, \cdots, K$, and the parameters defining the emission probabilities of stations $\mathbf{p}$. We explain this in greater detail as follows.

Transition Probabilities. Given a sequence of stay intervals $S(u) = (s_i)_{i=1}^N$, ACHMM assigns each stay interval to a state and forms a state sequence $Z(S(u)) = \langle z_1, z_2, ..., z_N \rangle$, where $p(z_i) = Categorical(\varphi)$. The Markov assumption allows the probability distribution of z_i to depend on the state of previous latent variable z_{i-1} through a conditional distribution $p(z_i|z_{i-1})$. The conditional distribution $p(z_i|z_{i-1})$ for $i = 1, \cdots, N$ forms a $K \times K$ transition probability matrix M, where $\sum_{k=1}^K p(z_i = k|z_{i-1}) = 1$. As as result, the probability of generating the sequence of stay intervals can be defined as follows:

$$p(z_1, z_2, \cdots, z_N) = p(z_1) \prod_{i=2}^N p(z_i|z_{i-1}) = p(z_1) \prod_{i=2}^N M(z_{i-1}, z_i) \tag{5}$$

where the initial latent node z_1 does not have a parent node and thus is represented by the initial probability $\pi_{1,k}$ and $\sum_{k=1}^K \pi_{1,k} = 1$. In this work, we default the initial probability over states as a uniform distribution.

Emission Probabilities. ACHMM determines the conditional distribution of observed stay intervals $p(s_i|z_i)$ with a Gaussian distribution governed by mean μ_{z_i} and covariance $\sigma^2_{z_i}$, $p(s_i|z_i) \sim \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i})$.

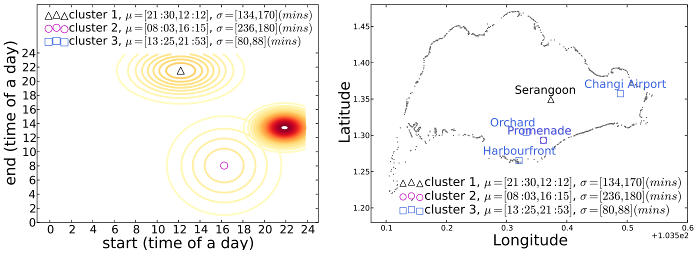
Once the hidden state of each stay interval is determined, we obtain the occurrences of stations that is associated with each stay interval in a hidden state. ACHMM determines the emission probabilities of stations for observed stations in each hidden state as a Categorical distribution, $p(g_i|z_i) \sim Categorical(\mathbf{p}_{z_i})$

The joint probability distribution over observed stay intervals and stations is then given by

$$\begin{aligned} p_{ACHMM}&(S(u), G(u)|\mu, \sigma, \mathbf{p}, \varphi, M) = \\ &(\sum_{z_1=1}^K p(s_1|z_1) \cdot p(g_1|z_1) \cdot p(z_1|\varphi)) \cdot \\ &\prod_{i=2}^N (\sum_{k=1}^K p(s_i|z_{i,k}) \cdot p(g_i|z_{i,k}) \cdot p(z_i|z_{i-1,k})) \end{aligned} \tag{6}$$

where μ and σ denote the set of μ_k's and σ_k's respectively.

Figure 3 shows three spatial and temporal patterns of activity centers derived by ACHMM (with $K = 3$) for a user (UID=4266653). Each hidden state indicates an activity center. Figure 3(a) reveals three activity centers from the user's stay intervals. For example, $cluster_1$ centers around the stay interval μ=[21:30,12:12) covering the stay intervals from evening to next mid-day. $cluster_2$ centers around the stay interval μ=[08:03,16:15) and covers the stay intervals from morning to afternoon. $cluster_3$ reflects the activities that center around the stay interval μ=[13:25,21:53) from early afternoon to late evening. Figure 3(b) illustrates the geo-focus of activity centers by stations. For example, the geo-focus in $cluster_1$, centered around the stay interval μ=[21:30,12:12), is located at Serangoon subway station, which is a typical residential area in Singapore. The geo-focus in $cluster_2$ is located at Promenade subway station, featuring shopping malls in Marina Center area. The geo-focus in $cluster_3$ is located nearby shopping centeres at Promenade, Orchard, Harbourfront subway stations, and Changi Airport subway stations.



(a) temporal-focus

(b) geo-focus

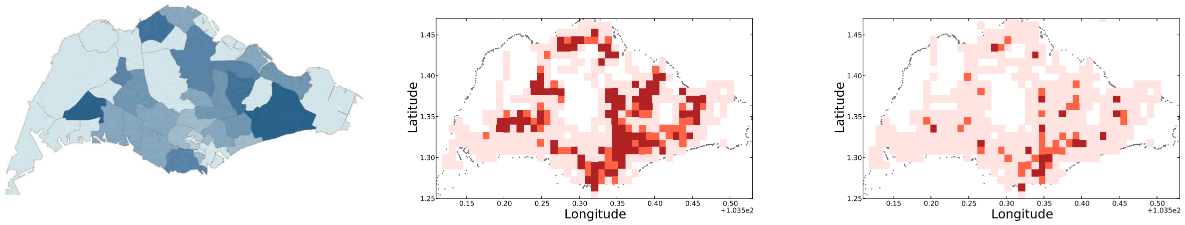
Fig. 3. Activity centers derived from ACHMM for user(UID=4266653).

C. State Label Assignment

To analyze the lifestyle patterns at both user and population levels we assign HOME, WORK and OTHERS to the states of ACHMM (with $K = 3$). As ACHMM+TH, which is introduced in detail in Section III-D, achieves high accuracy in station label assignment (90% in F-score), we thus further extend ACHMM+TH and apply to state label assignment so as to derive interesting popular trends. Essentially, among states with the state mean stay interval $[t^s, t^e]$ $t^e < t^s$ (eg., [18:00, 07:00]), the one with the longest duration is assigned the HOME label. The remaining stations are assigned the OTHERS label. Likewise, among states of state mean stay interval $[t^s, t^e]$ such that $t^s < t^e$ (eg., [07:00, 18:00]), the state with the longest duration is assigned the WORK label. The remaining states are assigned the OTHERS label.

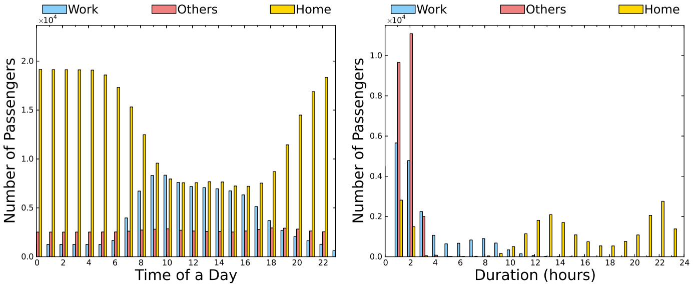
Heatmaps of Home and Work regions. We profile the residence, work and casual areas of 20K users by summing the emission probabilities of stations in their HOME and WORK states as shown in Figure 4(b) and 4(c) respectively. The intensity of a region in the heatmap is high (approach red) as the emission probabilities of the stations in the region obtain higher values. Figure 4(b) shows that most users live in the peripheral Singapore. The top HOME stations are located at: "Boon Lay"(West), "Woodlands"(North) and "Yishun"(East), "Ang Mo Kio"(East), "Toa Payoh"(East). This HOME state heatmap is consistent with the 2010 census report of 3.77 million Singapore residents as shown in Figure 4(a). The darker color in the figure represents highly populated residential area. Figure 4(c) however shows that most people work in the South Central area which covers the downtown and financial district of Singapore which sees a high concentration of offices and businesses. The top WORK stations are "Bugis"(South Central), "Harbour Front"(West Central), "Orchard"(Central), "City Hall"(South Central), and "Ang Mo Kio"(East). The distinction between HOME and WORK heatmaps also suggests that most users have to travel some distance to get to their work places.

Start and End times of Users' WORK and HOME States. The learned ACHMM model of each user also captures the stay intervals of the HOME and WORK states. We can therefore analyze the time intervals users are expected to stay home or work. We first divide a full day into 24 one-hour intervals, i.e., [00:00,00:59], [01:00, 01:59], $\cdots$, [23:00, 23:59]. We then count the number of users with HOME states covering each of these one-hour intervals. Figure 5(a) shows the distribution of time intervals for HOME, WORK and OTHERS states. The figure shows that most people begin their work state around 09:00 hrs and end their work around 17:00 hrs although a number of users may start their work day as early as 06:00 hrs.



(a) Home (b) Home (c) Work

Fig. 4. Labeling spatio-temporal activity centers: (a) residence population by June 2010, (b) HOME and (c) WORK



(a) Time Intervals (b) Durations

Fig. 5. Distributions of time intervals and durations.

On the other hand, most users stay at home from 18:00 hrs to 08:00 hrs. The home durations of users cover a longer stretch of time than work durations. As a result, there is no single time interval when *all* users are expected to be at simultaneously. Almost all users, in contrast, are expected to be at home during the stretch of time from 00:00 hrs to 05:00 hrs. Unlike HOME and WORK states, users in the OTHERS state do not share a common demarcated period as users can be in this state any time.

Length of State Intervals. Now, we want to determine the amount of time users spend in different states. Figure 5(b) shows that most people spend 9, 14 and one hours in the WORK, HOME and OTHERS states respectively. There are however significant number of people spending 10 or more hours in the WORK state.

D. Station Label Assignment

To further understand a user's lifestyle, we need to assign semantic labels to the stations the user visits. For simplicity, we consider three station labels: HOME, WORK and OTHERS. We focus on unsupervised methods as they are simple to implement and does not require human annotated labels. In this work, we propose the following four heuristics-based methods for assigning labels to the user's stations. Each method assigns, for a given user, one label (i.e., HOME, WORK or OTHERS) to each station of the user.

Frequency-based heuristics (FH): Among stations visited by user u, the most and second frequently visited stations are assigned with HOME and WORK label respectively, and the remaining stations are assigned with OTHERS label.

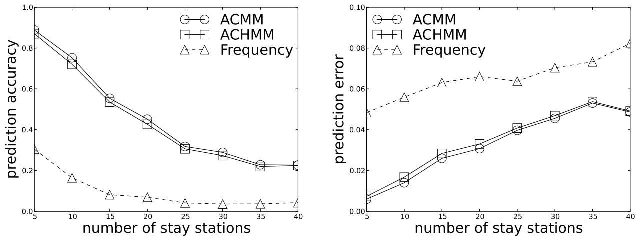
Duration-based heuristics (DH): The stations that the user spends longest average duration and the second longest average duration are assigned with HOME and WORK labels respectively. The remaining stations are assigned with OTHERS label.

Time-based heuristics (TH): Among stations of the user with at least one stay interval $[t^s, t^e]$ such that $t^e < t^s$ (e.g., [18:00, 07:00]), the one with the longest duration is assigned the HOME label. The remaining stations are assigned the OTHERS label. Likewise, among stations of the user with at least one stay interval $[t^s, t^e]$ such that $t^s < t^e$ (e.g., [07:00, 18:00]), the station with the longest duration is assigned the WORK label. The remaining stations are assigned the OTHERS label.

ACHMM with frequency-based heuristics (ACHMM+FH): We first mine ACHMM (with $K = 3$) states for user u from observed stay intervals. As a station may be associated with stay intervals that belong to different states, we further determine the dominant state c for the station g based on frequency. We denote frequency of the dominant state c of station g by $Frequency_c$(u,g). The station with the first and second highest $Frequency_c$ are assigned the HOME and WORK labels respectively. The remaining stations are labeled as OTHERS.

ACHMM with duration-based heuristics (ACHMM+DH): Once the dominant state c for each station g are determined, we derive the average duration of g based on the set of stay intervals of g associated with the dominant state c, and denote it as $aDuration_c$(u,g). The stations with the first and second longest $aDuration_c$'s are assigned the HOME and WORK labels respectively. The remaining stations are labeled as OTHERS.

ACHMM with time-based heuristic(ACHMM+TH) Once the dominant state c for each station g are determined, we derive the average time difference for each station g (in minutes) based on the set of stay intervals of g associated with dominant state c, denoted as $aTDiff_c$(u,g), using equation 7. Essentially, $aTDiff_c$(u,g) gives: (a) negative value for a stay interval if aST_c(u,g) $<$ aET_c(u,g), or (b) positive value if aET_c(u,g) $\leq$ aST_c(u,g), where the average start times in dominant cluster c is defined as aST_c(u,g)=$\frac{\sum_{(t_{s,i},t_{e,i})\in S_c(u,g)} t_{s,i}}{|S_c(u,g)|}$ and the average end times in dominate cluster c is defined as aET_c(u,g)=$\frac{\sum_{(t_{s,i},t_{e,i})\in S_c(u,g)} t_{e,i}}{|S_c(u,g)|}$. The intuition behind this heuristics is that we observe users stay at home before midnight or get out after midnight. For stay intervals covering the midnight, we derive their average time difference. The same is done for the other stay intervals. The two average time differences are then used to determine the label of the station g. For example, the $aTDiff_c$(u,g) of the interval [21:00,08:00] is 660, and the $aTDiff_c$(u,g) of the interval [08:00,19:00] is -660. Among all g visited by user u, we assign HOME label to the one with the largest positive $aTDiff_c$(u,g) and WORK label to the one with the smallest negative $aTDiff_c$(u,g). We assign OTHERS label to the remaining locations.



(a) Hit Rate

(b) Euclidean Distance

Fig. 6. Performance and Goodness-of-Fit Comparison

$$aTDiff_c(u,g) = \begin{cases} aST_c(u,g) - aET_c(u,g) & \text{(a)} \\ -aST_c(u,g) + aET_c(u,g) + 1,440 & \text{(b)} \end{cases} \qquad (7)$$

IV. Experiments and Results

We conducted two evaluation tasks using our models on EZ-Link dataset. The first task evaluates the effectiveness of both ACMM and ACHMM models through a location prediction task. The second task evaluates the effectiveness of station label assignments.

A. *Evaluation based on Location Prediction*

In this evaluation, we apply our models to predict the future location for individual user given the user's historical stay intervals and respective stations as training data. Due to the way we model the user lifestyle, the future location is one of the many stations the user has visited earlier. The stay intervals are constructed from the EZ-Link dataset using the inter-leg time gap threshold $\eta = 15$ minutes.

We compared the performance of both ACMM and ACHMM against a baseline frequency model (FREQ) that returns the station most visited by the user. For both ACMM and ACHMM, we set $K = 3$. Each model will be trained using first 80% of a user's stay intervals, and evaluated using the remaining 20% of the user's stay intervals. ACMM and ACHMM return the station with the highest likelihood value for a query stay interval. Since the focus is to evaluate the accuracy of predicted future location, we use the following two performance metrics, namely: (i) prediction accuracy in terms of *Hit Rate* and (ii) *Euclidean Distance* between the predicted station and actual station. Hit rate is defined by the fraction of test stay intervals with correctly predicted stations.

As shown in 6(a) and Figures 6(b), ACMM and ACHMM perform significantly better than FREQ. There is however very minor performance difference between ACMM and ACHMM. This could be attributed to the shorter series of stay intervals among the users. As ACHMM provides a more detailed modeling user lifestyle, we decide to use the model in the subsequent analysis.

B. *Evaluation of Station Label Assignment*

The station label assignment task assigns HOME, WORK or OTHERS label to each station visited by a user. In this evaluation task, our goal is to label user locations to answer two research questions: (a) *How accurate is automatic labeling of a user's station?* and (b) *What are the important features for determining the semantic of a station?* To evaluate the task accuracy, we first construct the ground truth labels.

Ground truth annotation. We first randomly selected 100 users and their trip data from our EZ-Link Dataset in January 2012. On average, each user has 35 stay intervals and 4.4 unique stay stations. There are altogether 441 user-station pairs as we pair each user with the stay stations they visited. We recruited six local residents who are familiar with the city to annotate these user-station pairs with HOME, WORK and OTHERS labels. We divided the 100 users into two equal-sized groups and assign three annotators to assign labels to the (user,station) pairs that belong each user group. Each annotator is required to annotate the labels of the stay stations of a user before moving on to the stay stations of the next user. In the end, each user-station pair obtains three labels from three different annotators. From the annotators' assigned labels, we derived the ground truth label of a user-station pair by majority vote nv. User-station pairs are assigned to a ground truth label if two or more annotators agreed on the label. Table III shows the results of annotation. Table III shows that all annotators have high level of agreement (88% in complete agreement) across labels. The numbers of user-station pairs assigned with HOME, WORK and OTHERS labels are 106, 105 and 230 respectively. This suggests that that OTHERS stations are the majority.

We next examine the distribution of number of home, work and others stations. As expected, most users attach with exactly one home station (94% users) or one work station (95% users). Very few users attach with two home (or work) stations and none attaches with three or more home (or work) stations. When a user has two home (or work) stations, they are usually very near each other suggesting that they are not far from the actual home (or work) location of the user. This is verified by the average distance of 199 meters between the HOME stations of the same user (or 522.3 meters between the WORK stations of the same user) . Most users attach with two or three OTHERS stations as expected as they may visit several places other than home and work stations. The average distance between OTHERS stations per user is 11 km. Interestingly, four users do not attach with OTHERS stations ($h = 0$). These users may use private transport to get to these stations.

Based on the ground truth labels, we evaluate the four station label assignment method described in Section III-D. We measure the accuracy of station label assignment methods using F-score. For each label, say HOME, we define:

$$\text{Precision} = \frac{\text{\#user-station pairs correctly assigned the label}}{\text{\#user-station pairs predicted the label}}$$

$$\text{Recall} = \frac{\text{\#user-station pairs correctly assigned the label}}{\text{\#user-station pairs with the ground truth label}}$$

$$\text{F-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

As shown in Table IV, the results vary with the different target station labels. The assignment of WORK appears to be harder than HOME. As we can see, DH works more effective for *O/Non-O* labels among all labels. This is because the stay patterns at OTHERS locations usually reflect short durations; whereas the stay pattern at HOME locations usually reflect regular start and end times. The results show that TH outperforms DH and FH as TH considers both duration and relative values of start and end times as well. FH yields inconsistent performances as it performs well for *O/Non-O* labels but poor for minority labels. With refined information augmented by ACHMM, ACHMM+DH and ACHMM+TH outperform DH and TH themselves respectively. As location label is highly correlated with stay intervals, the stations with similar stay intervals (which is clustered by ACHMM) thus more likely reflect a real-world

TABLE III. ANNOTATED LABEL DISTRIBUTION (nv: NUMBER OF ANNOTATORS IN AGREEMENT)

Label	**# user-station pairs**			**# users with h labeled stations**				
	nv=3	*nv=2*	total	*h=0*	*h=1*	*h=2*	$h \geq 3$	total
Home	93 (88%)	13 (12%)	106	0	94	6	0	100
Work	91 (87%)	14 (13%)	105	0	95	5	0	100
Others	205 (89%)	25 (11%)	230	4	9	40	47	100
Total	389 (88%)	52 (12%)	441					

activity center for a user. Given the same ACHMM cluster results, ACHMM+TH particularly outperforms ACHMM+DH for all labels because the time difference (i.e., $aTDiff_c$(u,g)) is more informative than duration (i.e., $aDuration_c$(u,g). Interestingly, unsupervised approaches (i.e., ACHMM-based) perform more stable across labels than heuristic approaches as shown by the standard deviation of average F-score.

Comparison with Supervised Station Label Assignment methods. To determine how the unsupervised methods perform compared with supervised methods, we conducted a comparison with Support Vector Machine (SVM) using 5-fold cross validation. To allow the importance of features to be analysed, we used SVM with linear kernel. We divided the annotated user-station pairs into five equal size folds using four of them for training a SVM classifier and one fold for obtaining the predicted labels of the trained classifier. This was repeated for every fold to be used for test. The predicted labels of all folds then combined together to derive the overall accuracy. We introduce two supervised station label assignment methods, one exploits temporal features (*SVM+T*) and the one exploits both temporal feature and dominant cluster features *SVM+TC*. The definition of each feature is listed in Table V. As shown in Table IV, SVM+TC outperforms SVM+T once cluster features are included in the training process. Particularly, the cluster features can slightly improve minority classes, HOME and WORK. Among all approaches, supervised approaches provide optimal and stable performance in terms of average F-score and standard deviation of average F-score.

Feature Analysis. To investigate fundamental factors to indicate labels, we show the feature coefficients learned from SVM with linear kernel. The feature coefficient gives useful interpretation of the importance of each feature and the absolute value of the coefficient relative to the others gives an indication of how important the feature is for separating data points into different classes. For example, aST_c(u,g) and $Frequency$(u,g) are the two most prominent features in both *H/Non-H* and *W/Non-W* classifiers. aST_c(u,g) is prominent because users tend to regularly start to stay at home (work) in late evening (in early morning) compared to OTHERS locations. $Frequency$(u,g) is prominent because users tend to stay at home (work) places more often than OTHERS locations. In particular, the dominant cluster-based feature, aST_c(u,g), is more prominent than the global aST(u).

To differentiate OTHERS locations from remaining labels, the global $Frequency$(u,g) and dominant cluster-based $Frequency_c$(u,g), are both essential. As reported in Table V, the coefficient of $Frequency$(u,g) (3.48) and $Frequency_c$(u,g) (2.42) are significantly higher than remaining features. This coincides with our intuition that users tend to repetitively stay at HOME/ WORK places, whereas users do not necessarily receptively stay at OTHERS locations.

V. RELATED WORKS

A. Urban Computing

Urban computing applies data analytics to wide range of spatial-temporal data from sensors, people, vehicles, transportation networks, and others to model the dynamics of urban cities for better modeling of city activities and planning/design of city facilities [13], [4], [12]. Based on travel survey data, Zhong et al. [14] proposed a centrality index for determining functional centers and proposed attractiveness indices for spatial impact analysis.

Wang et al. further demonstrated that travel time can be modeled and predicted using taxi sensor data [9]. Based upon GPS data tracking car movement, Giannotti et al. proposed a querying and mining framework to discover trajectory clusters, trajectory patterns as a sequence of regions and corresponding time intervals associated to each region [2].

Based on aggregated mobile call detail records, Toole et al. modeled the spatial-temporal changes of call detail records in each urban region and trained classifiers to predict the land use [7]. Using similar kind of data, Wang et al. [8] investigated the correlation between individual movements and social interactions. Such strong correlations can be useful in predicting future user movement and social interactions.

In the social media setting, Silva et al. [6] compared Instagram and Foursquare data to derive same users' movement patterns, popularity of regions in cities, and points of interests. Noulas et al. [5] analyzed urban human mobility in several metropolitan cities using Foursquare data. They found that users adopt uniform probabilities traveling in the first 100 meters but decreasing probabilities beyond that. They also discovered that the average distance of human movements is inversely proportional to the city's density. Hong et al. [3] proposed a topic model for modeling both location topics and user topical interests in geotagged Twitter data.

B. Location Prediction

Location prediction research focuses on learning the history of human movement for future location prediction. Cho et al. [1] provided several observations from cell phone location data and two online location-based social networks. Specifically, they identified three fundamental factors of users' mobility: geographical periodicity, temporal periodicity, and social network structure. Based on these observations, the authors proposed to model human mobilities using separate spatial and temporal Gaussian components with social influence. Xue et al. [10] addressed the data sparsity issue in location prediction and proposed a Sub-Trajectory Synthesis algorithm to predict destinations. The algorithm decomposes historical trajectories into sub-trajectories and synthesizes the sub-trajectories using a Markov model to increase the prediction coverage. Yang et al. [11] explored social spatial-temporal events to predict both time and location of user's next movements. They proposed to integrate the social and spatial-temporal information of human movements to reveal the regularity and the dynamics of social interactions of users. They addressed both time and

TABLE IV. ACCURACY OF STATION LABEL ASSIGNMENT

	H/Non-H			W/Non-W			O/Non-O			Average
Method	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	F-score
FH	0.53	0.53	0.53	0.52	0.50	0.51	0.92	0.96	0.93	0.66 ± 0.20
DH	0.60	0.58	0.59	0.55	0.53	0.54	0.84	0.88	0.86	0.66 ± 0.14
TH	0.84	0.82	0.82	0.95	0.93	0.94	0.86	0.89	0.87	0.88 ± 0.82
ACHMM+FH	0.49	0.59	0.52	0.39	0.37	0.37	0.81	0.81	0.80	0.56 ± 0.18
ACHMM+DH	0.58	0.57	0.57	0.55	0.53	0.54	0.85	0.88	0.86	0.66 ± 0.14
ACHMM+TH	0.86	0.95	0.89	0.95	0.93	0.94	0.93	0.87	0.89	0.90 ± 0.02
SVM+T	0.94	0.98	0.95	0.95	0.95	0.94	0.92	0.96	0.94	0.94 ± 0.005
SVM+TC	0.96	0.96	0.96	0.96	0.95	0.95	0.92	0.95	0.93	0.95 ± 0.01

TABLE V. FEATURE WEIGHTS OF SVM+TC.

Temporal Feature	**Definition**	H/Non-H	W/Non-W	O/Non-O
aST(u,g)	average start times associated with station g for user u	-1.15	1.25	-0.26
aET(u,g)	average end times associated with station g for user u	1.01	-1.16	-0.11
aDuration(u,g)	average durations associated with station g for user u	-0.19	-0.27	0.48
Frequency(u,g)	number of time that user u spent time at station g	**-1.95**	**-1.64**	**3.48**
Dominant Cluster Feature	**Definition**	H/Non-H	W/Non-W	O/Non-O
aST$_c$(u,g)	average start times associated with station g that belongs to dominate cluster c for user u	**-1.41**	**1.91**	-0.41
aET$_c$(u,g)	average end times associated with station g that belongs to dominate cluster c for user u	1.31	-1.45	-0.01
aDuration$_c$(u,g)	average durations associated with station g that belongs to dominate cluster c for user u	-0.002	-0.35	0.33
Frequency$_c$(u,g)	number of times that station g is associated with in its dominate cluster c for user u	-1.15	-1.30	**2.42**
aTDiff$_c$(u,g)	average time differences of station g that belongs to dominant cluster c for user u	0.10	-1.25	0.30
vStartTime$_c$(u,g)	variance of start times associated with station g that belongs to dominant cluster for user u	-0.75	-0.001	0.31
vEndTime$_c$(u,g)	variance of end times associated with station g that belongs to dominant cluster for user u	-0.04	-0.33	-0.04
nUniqStn$_c$(u,g)	number of unique stations in dominant cluster that station g belongs to for user u	0.22	0.31	0.22
mDuration$_c$(u,g)	the duration of mean interval μ_c=$[t_s, t_e]$ of dominant cluster c that station g belongs to for user u	-0.23	0.51	-0.0003
clusterSize$_c$(u,g)	number of observations in dominant cluster c that station g belongs to for user u	-0.24	-0.65	0.52

location prediction. The former aims to estimate how long it will take before the next social spatial-temporal event occurs to a given person. For location prediction, they proposed a ranking model which combines the periodicity and sociality of human movements

Our work differs from existing location prediction research in two aspects. Firstly, we focus on modelling spatial-temporal information of users' staying behaviour to predict stay locations of given time interval. Secondly, other than predicting locations as a way to evaluate our models, we further explore semantic aspects of users' staying behaviors and offer insights at both individual (personal activity centers) and population level (residential and work areas).

VI. CONCLUSION

Transportation data such as commuter bus and subway trip data capture much of the urban movement and activity data in an urban city. In this research, we learn from bus and subway trips the regularities of user movement that represent their activity centers. The two probabilistic lifestyle models we have developed are shown to work effectively on trip data represented by start and end locations and time points. We also show that the models can generate population density distribution similar to that of of a census report. This combined study leads to several novel insights about urban population and usage of urban regions.

Two directions for future work are of particular interest. Firstly, we could explore different number of states in ACMM and ACHMM. This is particularly appropriate for users who have more complex lifestyles spending time at more than three activity centers. Secondly, we plan to explore our proposal lifestyle models to study evolution of lifestyles among users and relate them to real world events.

ACKNOWLEDGMENT

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA).

REFERENCES

[1] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *SIGKDD*, 2011.

[2] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *VLDB Journal*, 2011.

[3] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *WWW*, 2012.

[4] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M. C. González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *SIGKDD International Workshop on Urban Computing*, 2013.

[5] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 2012.

[6] T. H. Silva, P. O. Vaz de Melo, J. M. Almeida, J. Salles, and A. A. Loureiro. A comparison of foursquare and instagram to the study of city dynamics and urban social behavior. In *SIGKDD International Workshop on Urban Computing*, 2013.

[7] J. L. Toole, M. Ulm, M. C. González, and D. Bauer. Inferring land use from mobile phone activity. In *SIGKDD International Workshop on Urban Computing*, 2012.

[8] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *ACM SIGKDD*, 2011.

[9] Y. Wang, Y. Zheng, and Y. Xue. Travel time estimation of a path using sparse trajectories. In *SIGKDD*, 2014.

[10] A. Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang, and Z. Xu. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In *ICDE*, 2013.

[11] N. Yang, X. Kong, F. Wang, and P. S. Yu. When and where: Predicting human movements based on social spatial-temporal events. In *SDM*, 2014.

[12] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and POIs. In *SIGKDD*, 2012.

[13] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: Concepts, methodologies, and applications. *ACM TIST*, 2014.

[14] C. Zhong, S. M. Arisona, X. Huang, and G. Schmitt. Identifying spatial structure of urban functional centers using travel survey data: a case study of singapore. In *SIGSPATIAL International Workshop on Computational Models of Place*, 2013.