STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT DOCUMENTATION PAGE

DRISI-2011 (REV 10/1998)

1. REPORT NUMBER CA18-3137	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER

4. TITLE AND SUBTITLE Travel Demand Nowcasting	5. REPORT DATE 04/30/2018
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR Alexei Pozdnoukhov	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Civil and Environmental Engineering University of California, Berkeley Berkeley, California 94720	10. WORK UNIT NUMBER
	11. CONTRACT OR GRANT NUMBER 65A0529
12. SPONSORING AGENCY AND ADDRESS California Department of Transportation Division of Research, Innovation and System Information PO Box 942873, MS 83 Sacramento, CA 94273-0001	13. TYPE OF REPORT AND PERIOD COVERED Final Report 3/1/2017 - 2/28/2018
	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

Predictive models of urban mobility can help alleviate traffic congestion problems in future cities. State-of-the-art in travel demand forecasting is mainly concerned with long (months to years ahead) and very short term (seconds to minutes ahead) models. Long term forecasts aim at urban infrastructure planning, while short term predictions typically use high-resolution freeway detector/camera data to project traffic conditions in the near future. In this report, we present a medium term (hours to days ahead) travel demand forecast system. Our approach is designed to use cellular data that is collected passively, continuously and in real time to predict the intended travel plans of anonymized and aggregated individual travelers. The traffic conditions derived through traffic simulation can overcome the data sparsity for short term prediction. The data resolution, prediction tolerance and accuracy for medium term travel demand forecast are compromises between those of long term forecast and short term prediction.

17. KEYWORDS Nowcasting, Forecast, Travel Demand	18. DISTRIBUTION STATEMENT

19. SECURITY CLASSIFICATION (*of this report*)	20. NUMBER OF PAGES 63	21. COST OF REPORT CHARGED

Reproduction of completed page authorized.

DISCLAIMER STATEMENT

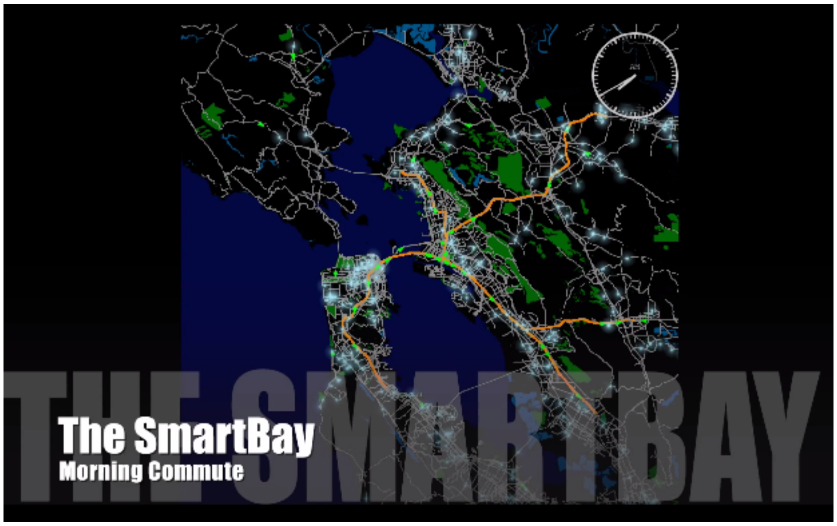
This document is disseminated in the interest of information exchange. The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California or the Federal Highway Administration. This publication does not constitute a standard, specification or regulation. This report does not constitute an endorsement by the Department of any product described herein.

For individuals with sensory disabilities, this document is available in alternate formats. For information, call (916) 654-8899, TTY 711, or write to California Department of Transportation, Division of Research, Innovation and System Information, MS-83, P.O. Box 942873, Sacramento, CA 94273-0001.

AM-PM: Travel Demand Nowcasting

Final report submitted to Caltrans

Contract NO. 65A0529, Task Order 52.6
Principal Investigator: Prof Alexei Pozdnoukhov



Department of Civil and Environmental Engineering
University of California, Berkeley
Berkeley, California 94720
(510) 984 8696
alexeip@berkeley.edu

Disclaimer Statement

This document is disseminated in the interest of information exchange. The contents of this report reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California or the Federal Highway Administration. This report does not constitute an endorsement by the Department of any product described herein.

For individuals with sensory disabilities, this document is available in Braille, large print, audiocassette, or compact disk. To obtain a copy of this document in one of these alternative formats, please contact: the Division of Research and Innovation, MS-83, California Department of Transportation, Division of Research, Innovation, and System Information, P.O. Box 942873, Sacramento, CA 94273-0001.

Contents

List of Figures

List of Tables

Acknowledgments

This work was partially funded by the State of California Department of Transportation (CalTrans) through UCCONNECT faculty research grant program, agreement 65A0529. Partial support from AT&T is also acknowledged.

Chapter 1

Introduction

1.1 Overview of Report Contents

Predictive models of urban mobility can help alleviate traffic congestion problems in future cities. State-of-the-art in travel demand forecasting is mainly concerned with long (months to years ahead) and very short term (seconds to minutes ahead) models. Long term forecasts aim at urban infrastructure planning, while short term predictions typically use high-resolution freeway detector/camera data to project traffic conditions in the near future. In this report, we present a medium term (hours to days ahead) travel demand forecast system. Our approach is designed to use cellular data that is collected passively, continuously and in real time to predict the intended travel plans of anonymized and aggregated individual travelers. The traffic conditions derived through traffic simulation can overcome the data sparsity for short term prediction. The data resolution, prediction tolerance and accuracy for medium term travel demand forecast are compromises between those of long term forecast and short term prediction.

To do so we developed a variety of generative sequence learning methods to train activity models from cellular data. Experimental results show that input-output hidden Markov models (IOHMMs) used in a semi-supervised manner perform well for location prediction while long short term memory models (LSTMs) are better at predicting temporal day structure patterns thanks to their continuous hidden state space and ability to learn long term dependencies. We validated our predictions by comparing predicted versus observed (1) individual activity sequences; (2) aggregated activity and travel demand; and (3) resulting traffic flows on road networks via a hyper-realistic microsimulation of the predicted travel itineraries. This report covers research results in developing machine learning based methods used to produce activity-based travel demand models from locational data available to cellular telecommunication operators in a form of Call Detail Records (CDRs).

The report is organized as follows. Chapter 2 reviews related work on long term travel demand forecast models (Chapter 2.1) and short term traffic prediction models (Chapter 2.2). Urban mobility models are also reviewed in Chapter 2.3. Chapter 3 depicts the framework of medium term travel demand forecast. Chapter 4 presents our processing pipeline of the raw cellular data . Chapter 5 improves the state-of-the-art deep generative urban mobility models using co-training input-output hidden Markov models (IOHMM) (Chapter 5.2) and long short term memory (LSTM) (Chapter 5.3). Technical details on sequence completion from partially observed sequence are presented in Chapter 6. In Chapter 7, we report on experiments, model selection, and validation results. We conclude the present work and offers discussions in Chapter 8.

1.2 Problem Statement

Travel demand forecast has been an integral part of most Intelligent Transportation Systems research and applications [50]. Long term forecast (days, months, or even years ahead) provides the basis for transportation planning and scenario evaluation. For example, transportation planers may need to answer the question of: how many people will be affected if a new subway line is introduced? How will travel patterns be changed if a major bridge is upgraded? The tolerance to these forecasts is usually high due to the long forecast horizon - days, months, or even years ahead. These studies typically use data collected from travel surveys that are infrequent, expensive, and reflect changes in transportation only after significant delays, which are not able to provide accurate predictions for short term or real time.

On the other hand, short term prediction (seconds to hours ahead) studies traffic conditions in a transportation network based on its past behavior, which is critical for many applications such as travel time estimation, real time routing, etc. The tolerance to these predictions is usually low due to the short forecast horizon - seconds to hours ahead. These studies use high-resolution data, usually collected from sensors and detectors on freeways.

However, one main concern is that these studies are limited to regions where high-resolution data is available. Moreover, such forecasts can only inform local operations such as adapting traffic light timing in response to growing queues.

One missing element of comprehensive transportation systems optimization systems is medium term forecasting (hours to days ahead), which, for example, could answer the question: based on observations of early morning or noon traffic, what will traffic be like during the evening commute? This could be a critical piece of knowledge used in the design of demand-responsive congestion mitigation interventions.

1.3 Research Objectives

In this report, we propose a medium term travel demand forecasting system to fill this gap. The idea is that given a large volume of partially observed user traces derived from cellular data [1] available at different times of day (e.g., 3:00 am, 9:00 am, 3:00 pm, etc.), we complete the individual daily activity sequences for the remaining period with pre-trained generative mobility models. The spatial-temporal resolution of cellular data makes it a perfect source for medium term travel demand forecast, whose target tolerance is in between of the long term and short term forecasts. This framework can also be used as an objective testbed for comparing the performance of different urban mobility models.

To validate the predictions, we can compare (1) at individual level: the discrepancies (e.g. differences in number of activities, travel distance, Hamming distance, etc.) between predicted sequences and ground truth sequences (observed by the end of a day) per individual; (2) at aggregated level: the hourly travel demand - number of activities, travel distances from all users; and (3) the resulting traffic volumes on all the major freeways within the region of study from predicted sequences and ground truth sequences.

[1] We emphasize that no personally identifiable information (PII) was gathered or used in conducting this study. The mobility data that was analyzed was anonymous and aggregated in strict compliance with the carrier's privacy policy. CDR (call detailed record) raw locations are converted into highly aggregated location features before any modeling takes places.

1.4 Significance of the Study

The main contributions of this report lie in three aspects:

- We proposed and solved a medium term travel demand forecast system which fills the gap between mainstreams of long term travel demand forecast and short term traffic state prediction. This system uses cellular data which will not suffer from the expensive data collection cycle for manual survey, or the data availability problems for short term traffic prediction.
- We improved and compared the state-of-the-art deep generative urban mobility models. Lessons learned from training different types of urban mobility models are summarized for future researchers.
- We explored the predictability of human mobility with parametric sequence learning models as related to using individualized non-parametric "nearest neighbor" approach.

Chapter 2

Related Work

2.1 Long Term Travel Demand Forecast

Long term travel demand models are the main tools for evaluating how travel demand changes in response to different input assumptions, scenarios and policies [14]. For example, how will the national, regional, or even local transportation system perform 30 years into the future? What policies or investments could influence this performance?

Earlier efforts on travel demand models has focused on trip-based approaches which comprises of four steps: trip generation, trip distribution, mode split, and route assignment [26, 6]. In the recent decades, such forecasts performed by activity-based models for demographic projections of a population have drawn more attention.

Activity-based travel model derives travel demand from people's needs and desires to participate in activities [14]. It models how people make decisions about activity participation in the presence of constraints, including decisions about what activity to participate, where to participate, when to participate, how to get there and with whom. Agents can adapt and change their decisions by learning from their behavior [38].The major advantages of the activity-based model over traditional trip-based models lie in four aspects: (1) consistency and integrity among sub-models; (2) behavioural realism; (3) disaggregated nature; and (4) more detailed performance metrics [41].

Activity scheduling is the central task of an activity-based model. Three main approaches for activity scheduling (constrains-based, utility-based, and rule-based) all require detailed activity diaries data (activity start time, duration, location, transportation mode, etc.) as input [3]. However, the data collection is usually performed through travel surveys that are infrequent, expensive, and reflect the changes in transportation with significant delays. For example, the National Household Travel Survey (NHTS), the data source that is typically the crux of travel demand models, is conducted every 5 years, and carries a total cost of millions of dollars. Thus travel demand models are mainly targeted at "typical day" travel demand forecast in the long term future. The tolerance to the forecast error is also high. As smart phone data become ubiquitous, developing a conceptual framework using alternative data, to frequently update activity-based models provides a new opportunity to make the near-term travel demand "nowcasting" more accurate.

2.2 Short Term Traffic Forecasting

With growing availability of data, short-term traffic forecasting became a very developed research area. It concerns predictions of traffic conditions made from seconds to hours into the future based on current and past traffic information. Most of the effort has focused on modeling traffic characteristics such as volume, density, speed, and travel times [50]. Vlahogianni thoroughly summarized the available literature and categorize reports mainly based on (1) What is the study area (motorway or arterial); (2) What is the study predicting (traffic volume, speed, density, or travel time); (3) What is the prediction algorithm (statistical time series model, machine learning model or hybrid).

However, there are certain limitations in short term traffic prediction. First, most of the studies use detectors or camera video (AVI) data. These data are mainly available on freeways and arterials, but not on the whole network. Thus, traffic predictions are mainly available for area where detectors/AVI data is available. To enrich the source of data, GPS of probe vehicles has been used in travel time and speed prediction. Zheng and Van Zuylen predicted complete link travel times based on the information collected by probe vehicles using three-layer neural network model [56]. Ye et, al. further introduced acceleration information and information from adjacent segments to improve the prediction of the travel speed of current forecasting segment [53]. Second, the prediction horizon usually ranges from a few seconds to a few hours. This will limit the use cases for the traffic prediction. For example, people may plan their afternoon trips in the morning based on traffic predictions more than a few hours ahead.

2.3 Human Mobility Modeling and Prediction

Urban mobility models characterize multiple aspects of individuals' travel patterns. Large amount of works focus on the activities (trip purpose), such as the spatial (location, [47, 4, 17, 39]) - temporal (start time and duration, [44]) choices of a single activity, or activity patterns (daily/weekly activity scheduling, [19, 21, 35, 57, 45, 5, 9, 12, 51, 54]). Another branch of research considers trips linking these activities, studying trajectories [32, 48, 37], travel mode [60, 11, 42, 58, 24, 7], by applying map matching and route choice [49, 15].

State-of-the-art can also be classified by the data sources used to model individual urban mobility. Early studies mainly used travel surveys [9, 12, 7]. In the recent decade, with the mobile phone data more available, passively collected data such as GPS [36, 35, 4, 60, 55, 34, 57, 39, 44, 49, 32, 5, 11, 31, 42, 58, 15, 24, 48], CDR (call detailed record) [25, 46, 19, 21, 18, 17, 40, 45, 51, 13, 54, 37] and location-based social networks (LBSN) data [16, 52] has provided grounds for new approaches in urban mobility studies. GPS data is granular in both spatial and temporal resolution. However, the availability of such granular data is usually limited to hundreds of travelers. LBSN data is exact in locations, and may provide additional social relation, comments and reviews on the venues for larger samples of travelers. However, LBSN data is limited by its discontinuity and sparsity in time. CDR data provides a trade-off between spatial-temporal resolution and ubiquity, while covering millions of travelers.

Studies that are not concerned with predictive or generative methods fall into two categories: first category tends to purely understand generic human mobility laws using descriptive statistics [25, 46, 34, 13], the other category focuses on the problem of recognition (activity, travel mode, [36, 40, 49, 45, 11, 31, 15]) rather than prediction. The studies of second category are mainly conducted on mobile phone data since activity type and travel mode are not explicitly observed from the data itself. For studies that do focus on predictive (generative) power, most works focus on predicting only next location (or duration) since it is a well formulated task that is also easier to validate. Some researchers make prediction by assuming Markov properties [47, 4, 52, 18]; other researchers treat prediction of next location as a classification (regression) problem using supervised learning [17]; and some researchers use trajectory matching techniques to make the prediction [55, 39]. However, not much research has been done on models that are capable of predicting a sequence of activities with locations and durations for the full day or longer.

Another observation is that most of the previous studies focus on only one aspect of urban mobility (such as location, duration, travel mode), or model these several aspects separately. Not many studies have focused on modeling daily activity patterns and scheduling that fuse activity type, location and duration together, which enables the model to generate a sequence of samples. Eagle and Pendland [19], Farrahi and Gatica-Perez [21], and Zheng et al. [57] used unsupervised techniques such as PCA and topic models to cluster daily activity patterns. However, they only included primary activity types such as "home" and "work", all other activities are categorized as "other". Liao et al. unified the process of map matching, place detection, and significant activity inference through a hierarchical conditional random field (CRF) using GPS data [35]. However, their model is discriminative in nature and is most suitable for recognition, rather than generating new sequences. Widhalm et al. [51] used an undirected relational Markov network to infer urban activities with CDR data. However, they did not model activity transitions due to the lack of cliques for consecutive activities.

To summarize, existing literature has focused on long term travel demand and short term traffic state forecasts, while current methods of urban mobility modeling have got limitations that make them only partly useful for medium-term forecasting. In this report, we fill this gap with sequence learning methods applied to build generative urban mobility models from cellular data.

Chapter 3

Modeling Framework

The developed data processing and modeling pipeline is presented in Fig. 3.1. Anonymized historical CDR data are processed to unlabeled historical activity sequences [54]. Urban mobility models are built upon these historical activity sequences. In this report, we improved the state-of-the-art urban mobility models including interpretable IOHMM models, as detailed in Chapter 5.2, and deep LSTM models, as detailed in Chapter 5.3.

On a target day, we receive streaming CDR data at different time of day (e.g. 3:00 am, 9:00 am, 3:00 pm, etc.), which are then processed to partially observed activity sequences. These partially observed sequences, along with the pre-trained parametric urban mobility models, are sent to the sequence predictor. The sequence predictor predicts and completes the activity sequences for the rest of the day based on the observed information, as detailed in Chapter 6. The completed activity sequences are sent to MATSim, a state-of-the-art agent-based traffic micro-simulation tool that performs traffic assignment. MATSim generates the predicted traffic conditions for the day.

By the end of the day, full day CDR are observed and processed to ground truth activity sequences. These ground truth activity sequences are validated against the predicted activity sequences at both individual level and aggregated level at different times of day. We also validate the resulting traffic from predicted activity sequences versus ground truth sequences, as detailed in Chapter 7. Finally, historical CDR database is updated with the new day's CDR, and urban mobility models can be updated and re-trained overnight.

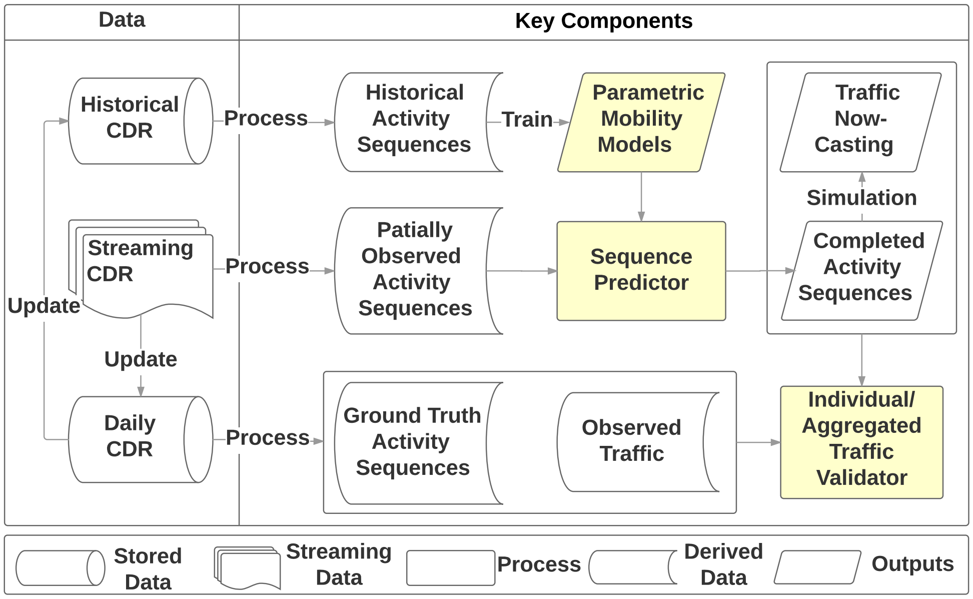


Figure 3.1: Modeling framework diagram. The left column represents the input to the algorithms and the right column represents the model components. Our key contribution of improved deep urban mobility models, sequence predictor, and validation are shown in shaded yellow.

Chapter 4

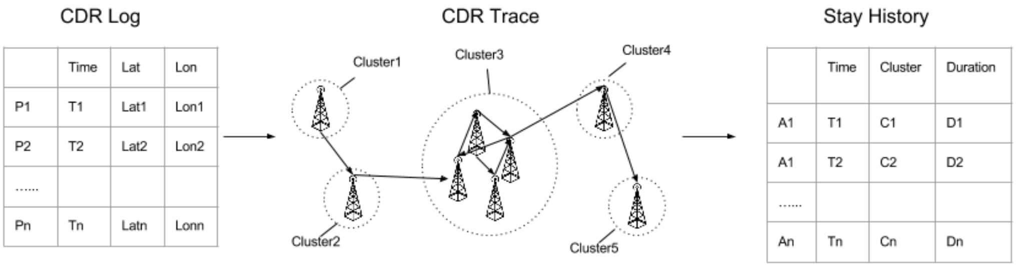
Data Processing

4.1 Introduction

Cellular Data such as CDR logs does not give us information about activities directly. Raw CDR data contains a timestamped record for each communication of anonymous user's devices served by the cellular network. Due to positioning errors and connection oscillations, it is not straightforward to extract features to perform activity recognition from raw CDR sequences. A pre-processing step is first performed to convert the records to a sequence of stay location clusters that may correspond to distinct yet unlabeled activities, as shown in Fig. 4.1. The clustering can be seen as a first layer of hashing locations, which preserves privacy. Attributes of each activity, such as the start time, duration, location features, and the context of the activity (whether this activity happens during a home-based trip, work-based trip, or a commute trip), is also extracted as a result of this processing.

From the activity sequences, primary activities such as home and work can be inferred[1]. Detecting home and work location features are useful in many respects: first, this allows us to perform dynamic population estimation. Second, with home and work inferred, we can identify specific groups of users by a set of predefined decision rules. One of the most simple rules is to group users by their geographical area. This makes it possible to train

[1] Note that once the pre-processing and home/work inference steps are applied, only features associated with location clusters are used for modeling, such as distances to home and work. This can be seen as a second layer of anonymization of user's locations, since no specific location cluster IDs are associated with any user at any time in the modeling process itself.


CDR Log

	Time	Lat	Lon
P1	T1	Lat1	Lon1
P2	T2	Lat2	Lon2
......			
Pn	Tn	Latn	Lonn

Stay History

	Time	Cluster	Duration
A1	T1	C1	D1
A1	T2	C2	D2
......			
An	Tn	Cn	Dn

Figure 4.1: Call Detail Records (CDR) data processing. The table at left represents the raw CDR format, i.e., time stamped record of communications. A stay points detection algorithm is used to convert the raw CDR data to a sequence of stay locations with start time, duration and location ID, as represented in the table at right.

separate models for users residing in a specific neighborhood or a Transportation Analysis Zone (TAZ) since people living in different geographical zones might show different travel behaviors. Moreover, we can train separate models for regular commuters/part-time/unemployed groups of residents within a community. The model structures are expected to be significantly different within each group. Finally, home and work inference for anonymized cellular users adjusted to the full population provides daytime/nighttime population density estimates, as shown in Fig. 4.3.

With the activity sequences (including home and work anchor activities) identified, we can understand the daily activity structure of travelers that are traditionally available solely via manual surveying. They include: (1) the distribution of number of tours before going to work, during work and after getting back home; (2) the distribution of number of stops during each type of tour (home-based, work-based and commute tours); and (3) the interactions in stop-making across different times of day (e.g. how making an evening commute stop will affect the decision in making a post-home stop) [9].

4.2 Processing Pipeline

4.2.1 Stay points detection in CDR

The goal of stay location recognition is to turn CDR logs into a list of sequential stay location identifiers with start time and duration for each user, as illustrated in Fig. 4.1. Each record of raw CDR logs contains the timestamp and the approximated latitude and longitude of events recorded by the data provider. This is a CDR-specific step that requires fine-tuning of several threshold parameters. Note that once the pre-processing steps and the following are applied, only features associated with clusters locations are used, such as distances to home and work. This can be seen as a layer of anonymization of user's locations, since no specific location cluster IDs are further associated with any user at any time in the activity modeling process itself. The main steps of the algorithm are as follows:

(1) Cluster CDR records. The first step in stay location detection is filtering out positioning errors. This is achieved by spatial clustering. For GPS data, accuracy ranges of 10-100m are used in many studies that use GPS to detect stay locations [20]. The distance thresholds for GPS stay-location clustering is much smaller than the thresholds for CDR records. For example, a roaming distance of 300 meters [29] and 1000 meters [51] was used to cluster points to reflect the spatial measurement accuracy of the CDRs. For our stay-location detection, we use a density based clustering with similar parameters. At the end of the clustering step, consecutive data points with the same cluster ID are combined into a single record with start time equal to the timestamp of the first of the consecutive events at that cluster, and end time equal to the time stamp of the last of the consecutive events at that location cluster.

(2) Construct and process an oscillation graph. Consecutive CDR records may have nearly identical timestamps, but different location IDs. Such oscillations occur because the cell phone is communicating with multiple cell towers. These instantaneous location jumps may occur because of traveling users whose cell phone have just come in contact with a new cell tower along the way, but often such location jumps are observed even though users are standing still. In the latter case a user's location appears to oscillate back and forth between two clusters.

When a user's location is simultaneously reported in two location clusters, an edge between these two clusters is added to the oscillation graph. Edges in the oscillation graph connect clusters that are suspicious for oscillations. An example oscillation graph described in that section is shown in Figure 4.2. Each node in the graph represents a location cluster. There is an edge if oscillation has been observed between two clusters. The thicker the edge, the more oscillations have been observed.

(3) Filter oscillation points. With cluster-pairs transformed into an oscil-

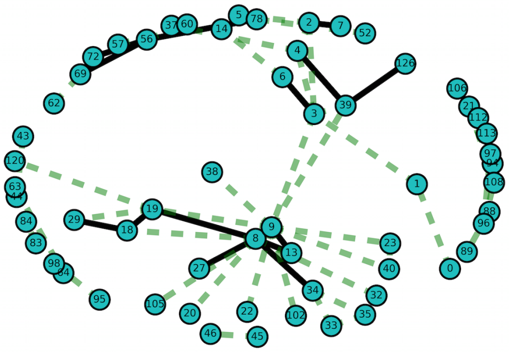


Figure 4.2: Sample oscillation graph

lation graph, one can discern oscillations from travel based on the pattern of location cluster sequences. Suppose the locations of two consecutive records are location cluster A and location cluster B, respectively. If edge (A, B) exists in the oscillation graph, and if the user visits cluster A, then B, back and forth, the visit to B is determined to be an oscillation - the points are combined into a single record with a duration determined by the combined time spent in A and B. We assign the location of these records to cluster A if the user spends more time in A than B, else it is assigned to cluster B.

(4) Filter locations with short durations. At this point, positioning noise and oscillation noise are removed. Now we have a sequential list of location cluster visits, each with a start and end time. Some of these cluster visits are stay locations, and others are pass-by points. The accepted threshold for stay locations varies widely. The threshold was set to 20 minutes in [59], 15 minutes in [51] and 10 minutes in [29]. Several GPS applications use stay durations ranging from 90 seconds to 10 minutes. We chose a threshold of 5 minutes, because in the activity based modeling context, 5 minutes is an appropriate threshold for an activity location, as opposed to a way-point.

4.2.2 Home and Work Location Inference

We recognize the importance of long-term recurrent stay points such as “home” and “work” that enforce a structure in the users’ daily mobility. Various strategies have been used for home and work location detection. A mixture of Gaussians is a popular method to model locations centered on home and work [16]. Another suggested definition of “home” was the location where the user spends more than 50% of time during night hours with night hours defined as 8pm to 8am [33]. Similarly, work hours can be defined as the area where the user spends more than 50% of time during day hours.

We adopt accepted methods in order to simplify processing and, most importantly, infer “anchor” points in the daily sequences that provide space-time context that is crucial to build a generative model of secondary activities. A range of travel choices, such as mode of transportation and destination choice, depend on the overall structure of the day. Moreover, early identification of home and work allows pre-clustering users into groups with similar behaviors by using heuristic decision rules (employed/unemployed/part-time worker, etc).

Our detection of the home and work locations is similar to the method of [33]. We identify home as the location where the user spends the most stay hours during home hours, and we identify work as the location where the user spends the most hours during the work hours. However, we define home and work hours to be much narrower time windows than the 8am-8pm criteria used in [33]. Borrowing from [29], the hours from midnight to 6am are defined as home activity hours, and 1pm to 5pm on weekdays are defined as working hours because they capture the core set of working hours for both early and late workers [28].

4.3 Description of Data

The data used in these studies comprise a month of anonymized and aggregated CDR logs collected in Summer 2015 by a major mobile carrier in the US, serving millions of customers in the San Francisco Bay Area. No personally identifiable information (PII) was gathered or used for this study. As described previously, CDR raw locations are converted into highly aggregated location features before any actual modeling takes places.

4.4 Data Processing Results

We pre-process the data following the aforementioned steps. The home and work locations are identified during the pre-processing step. For further modeling purpose, we focus on regular commuters that:

- showed up for more than 21 days a month at their identified “home” place;
- showed up for more than 14 days a month at their identified “work” place;
- have home and work **not** at the same location.

These criteria identify regular working commuters with a day structure containing both distinct Home and Work.

4.4.1 Home/Work Inference results

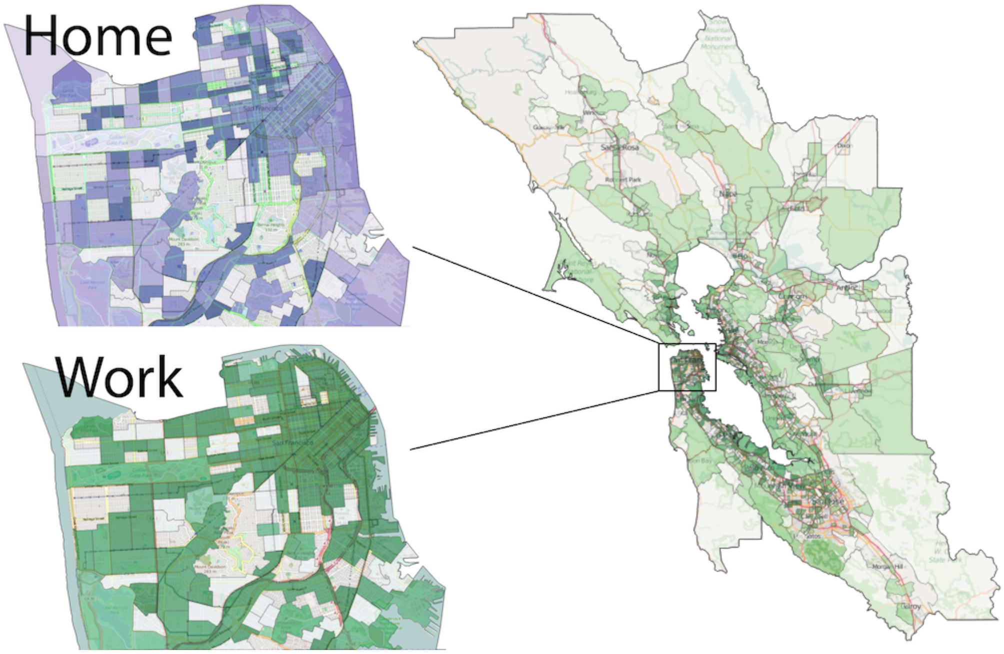
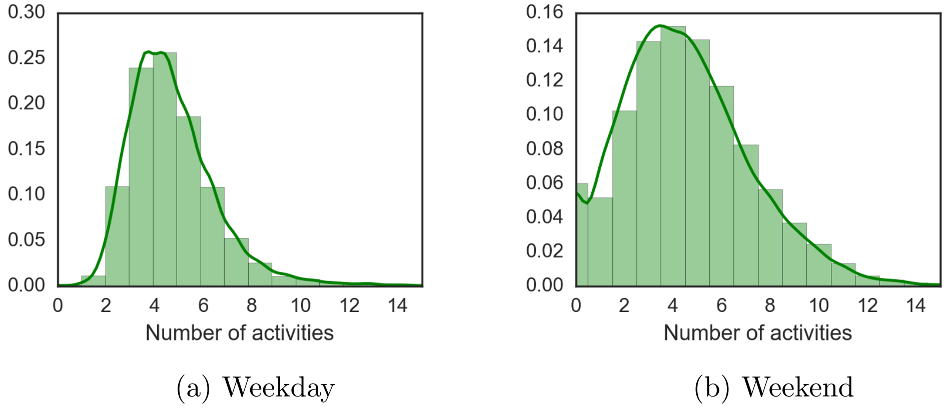


Figure 4.3: Density map of inferred home and work locations for San Francisco residents, aggregated at the census tract level (left), and an overall geographical scope of analysis with work locations density (right).

Fig. 4.3 shows the density map of inferred home and work locations for San Francisco residents (individuals with home in San Francisco city), aggregated at the census tract level. As shown in the right of Fig. 4.3, the work locations are spread in the SF Bay Area. The highest density occurs in San Francisco, Oakland, and some South Bay cities. Focusing on work locations in San Francisco, many of the inferred work locations are in Downtown San Francisco, the Financial District, and SoMA - three San Francisco neighborhoods with high employment density [30]. As expected, the home locations are more spread out throughout the city.

4.4.2 Number of Daily Activities



(a) Weekday

(b) Weekend

Figure 4.4: Empirical distributions of the average number of daily activities of San Francisco subscribers on a weekday (left) and on a weekend (right), after pre-processing.

Empirical distributions of the average number of daily activities for this population is shown in Fig. 4.4. The median number of activities is 4.4 per weekday and 4.0 per weekend. This is consistent with the California Household Travel Survey, reporting a number of 4 activities per day [1].

Chapter 5

Urban Mobility Models from Cellular Data

With the processed activity sequences and inferred primary activities from previous chapter, we can train the urban mobility models that we can use to predict the user activities. To validate the recognition results and to direct the learning process, we collected a small set of ground truth activities based on short range antennas which have relatively high spatial resolution. Point of interests (POI) data are joined with these short range antennas to identify the possible activities performed there and a set of rules are used to help us collect labeled activities. With the model coefficients and a set of sampled home and work locations of the total population, we can generate activity sequences and produce synthetic travel plans required by a microscopic traffic simulator.

5.1 Collection of Ground Truth Activities

Considering the choices for activity types, one would like to set a high number that encompasses a wide variety of travel purposes, however, data quality and availability limits the number of feasibly identifiable activities. Moreover, an ambiguity in semantic meaning of activity types (consider “leisure” vs “recreation”) asks for limiting the number of hidden states that show useful in practical applications. We describe here an empirical procedure for collecting ground truth data on activity types that provide useful insights on these modeling choices. The number of hidden states of the IOHMM model are set according to the labels of these ground truth activities. For CDR, it is usually hard to collect ground truth activities due to its low spatial resolution. However, there is a set of short range antennas that serve only a small range of area, which have relatively high spatial resolution. These short range antennas provide us the opportunity to collect “ground truth” activities.

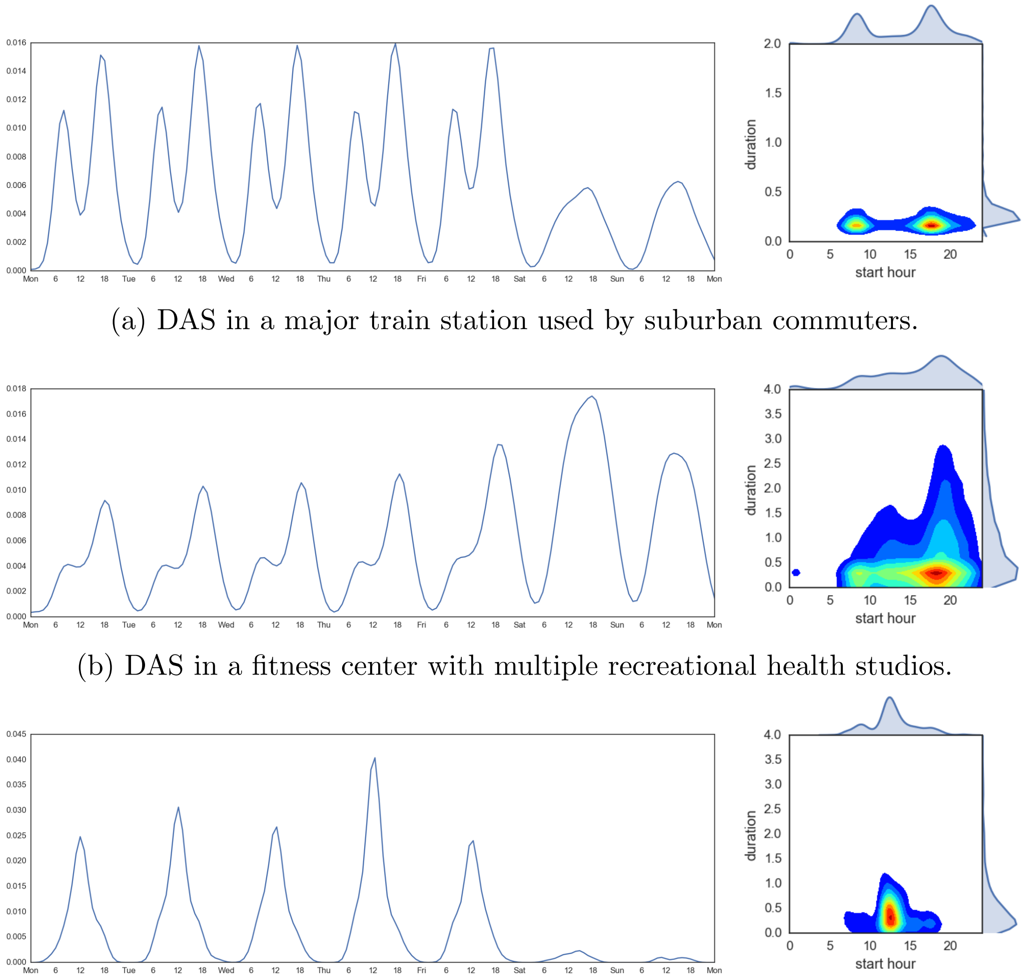
Short Range Distributed Antenna Systems (DASs)

A common component of a cellular networks is a set of distributed antenna systems (DASs) that are short ranged, including Indoor DASs (IDASs) and Outdoor DASs (ODASs). IDASs are usually installed in large commercial buildings such as shopping malls to ensure better signal coverage. And ODASs are usually installed at high occupancy outdoor venues such as stadiums or concert arenas. These antennas are set up to maximize signal strength for the users located in the building or stadium served by a given DAS, ensuring more precise localization. Fig. 5.1 illustrates the times and durations of connections established by users served by three particular DASs. The patterns are structured in time, indicating the activities performed there are quite regular and their purpose can be inferred from domain knowledge with high confidence.

Designation of Rules for Ground Truth

IDASs are often installed in large mixed-use commercial buildings. For example, one commercial building with IDAS installed could have bakeries, restaurants, taxi stands, gym and fitness centers, retail stores, as well as other businesses such as accounting and financial services. We designed a set of spatial-temporal decision rules to label a set of activities that can be considered as the ground truth. For instance, if a user is connected to a DAS in a food court at noon for one hour, this is most likely to be indicative of a lunch activity. Although we do not have complete certainty that this is indeed the activity type, the event is indistinguishable from a lunch break in terms of its mobility footprint, and with high likelihood we interpret this as a food activity.

We first acquired place information from POI databases such as Google places API and Factual Global Places API. Then, we joined this information with the locations of the DASs in order to extract activities that could be performed at each DAS. The place information provides listings of local



(a) DAS in a major train station used by suburban commuters.

(b) DAS in a fitness center with multiple recreational health studios.

(c) DAS in a business district building with a large food court.

Figure 5.1: Structural patterns of empirical data collected at short range DASs well explain the activity performed around the DASs: the number of activities start times within a course of a week (left) and an empirical joint distribution plot of the visit duration vs start times (right).

Table 5.1: Rules of labeling secondary activities based on activity spatial-temporal features

Activity	Duration (hours)	Start hour	Context	Location category
Lunch	0.25 - 1	11-12		Food
Dinner	0.25 - 2	17-18		Food
Shop	0.25 - 1	7-9 14-15 20-21	Home based or during evening commute	Shop
Transport	< 0.25		Commute	Transport
Recreation	1-4	7-21	Home based or during evening commute	Recreation
Personal	any	7-21		Personal
Travel	any	any		Out of the region

business and point of interest (POI) at most given locations. Since multiple activities can happen at the same location, we need some additional rules based on the spatial-temporal features of activities, as shown in Table 5.1. The "location category" column of the table indicates that the category is among the category labels returned from the APIs.

Note that the rules used to label activities as reported in Table 5.1 are restrictive. Given that the main purpose of these labels is to validate the proposed models, our goal is to be very confident in the activities we label. Thus, these rules are designed to pursue high precision rather than high coverage.

5.2 IOHMM Based Urban Mobility Models

5.2.1 IOHMM Architecture

This section introduces main parametric mobility module shown in Fig. 3.1.

Given the user stay history, that is, a list of stay location features with start times and durations, we would like to convert it into a sequence of activities enriched with semantic labels ("shopping", "leisure", etc.). We would also like to understand the activity pattern of the users, which can then be used to generate new sequences. To be more specific, the activity patterns can be defined by: (1) Spatial and temporal profiles such as location choice, start time, and duration. (2) A heterogeneous context-dependent probability model for transitions between activities.

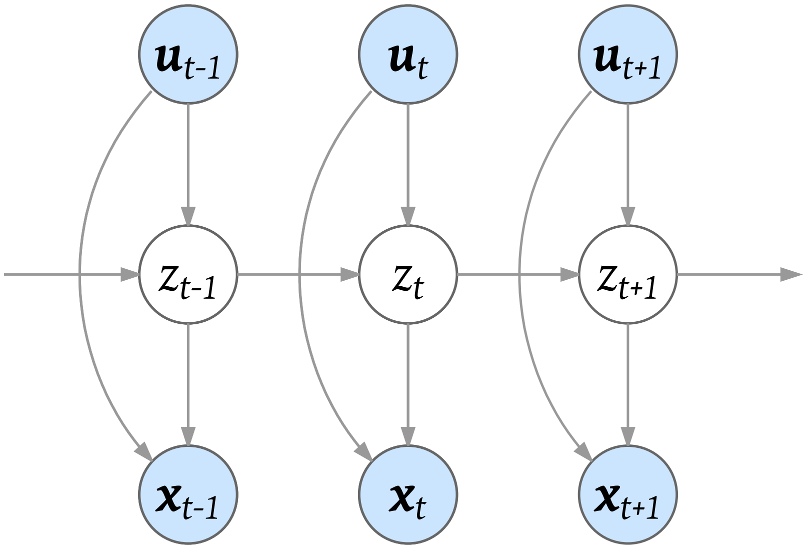


Figure 5.2: IOHMM Architecture. The solid nodes represent observed information, while the transparent (white) nodes represent latent random variables. The top layer contains the *observed* input variables $\boldsymbol{u_t}$; the middle layer contains *latent* categorical variables z_t; and the bottom layer contains observed output variables $\boldsymbol{x_t}$.

Hidden Markov Models (HMMs) have been extensively used in the context of action recognition and signal processing. However, standard HMMs assume homogeneous transition and emission probabilities. This assumption is overly restrictive. For instance, if a user engages in a home activity on a weekday, and departs for the next activity in the morning, she is likely going to work. If she departs in the evening, the trip purpose is likely to be recreation or shopping. Therefore, we propose to use the IOHMM architecture that incorporates contextual information to overcome the drawbacks of the standard HMM. In Fig. 5.2, the solid (blue) nodes represent observed information, while the transparent (white) nodes represent latent random variables. The top layer contains the *observed* contextual variables $\boldsymbol{u_t}$, such as time of day, day of the week, and information about activities in the past (such as the number of hours worked on that day). Note that the values of the input variables $\boldsymbol{u_t}$ used to represent the context have to be known prior to a transition. The middle layer contains *latent* categorical variables z_t corresponding to unobserved activity types. The bottom layer contains observed variables $\boldsymbol{x_t}$ that are available during training of the models (but not when generating activity sequences), such as location features and duration of the stay.

Likelihood of a data sequence under this model is given by:

$$\begin{aligned} L\left(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{u}\right) &= \sum_{\boldsymbol{z}} \Big(\Pr\left(z_1 \mid \boldsymbol{u_1}; \boldsymbol{\theta_{in}}\right) \cdot \\ & \prod_{t=2}^{T} \Pr\left(z_t \mid z_{t-1}, \boldsymbol{u_t}; \boldsymbol{\theta_{tr}}\right) \cdot \\ & \prod_{t=1}^{T} \Pr\left(\boldsymbol{x_t} \mid z_t, \boldsymbol{u_t}; \boldsymbol{\theta_{em}}\right) \Big). \end{aligned} \tag{5.1}$$

IOHMM architecture has been well described in [8]. The difference between IOHMM and semi-supervised IOHMM lies in the forward-backward algorithms. If we have ground truth activity (hidden states z) for timestamp t, then we will use $I_{j,t}$ to replace $\varphi_{ij,t}$ where $I_{j,t}$ is 1 if the hidden state $z_t = j$ at timestamp t in the labeled data, 0 otherwise, since $\Pr\left(z_t = j \mid z_{t-1} = i\right)$ reduces to $\Pr\left(z_t = j\right)$ with observed information. A summary of the differences between HMM, IOHMM and semi-supervised IOHMM is presented in TABLE 5.2.

Parameter Estimation

IOHMM includes three groups of unknown parameters: initial probability parameters ($\boldsymbol{\theta_{in}}$), transition model parameters ($\boldsymbol{\theta_{tr}}$), and emission model parameters ($\boldsymbol{\theta_{em}}$). Expectation-Maximization (EM) is a widely used approach to estimate the parameters of IOHMM. The EM algorithm consists of two steps.

E step: Compute the expected value of the complete data-log likelihood, given the observed data and parameters estimated at the previous step.

M step: Update the parameters to maximize the *expected* data likelihood given by:

Table 5.2: Highlights of comparison between an HMM vs. IOHMM vs.semi-supervised IOHMM ($\boldsymbol{u_t}$, z_t, $\boldsymbol{x_t}$ denote input, hidden and output variables respectively, i is an index of a hidden state, t is a sequence timestamp index, $I_{j,t}$ is 1 if the hidden state $z_t = j$ at timestamp t in the labeled data, 0 otherwise).

	HMM	IOHMM	semi-supervised IOHMM
initial state probability π_i	$\Pr(z_1 = i)$	$\Pr(z_1 = i \mid \boldsymbol{u_1})$	
transition probability $\varphi_{ij,t}$	$\Pr(z_t = j \mid z_{t-1} = i)$	$\Pr(z_t = j \mid z_{t-1} = i, \boldsymbol{u_t})$	
emission probability $\delta_{i,t}$	$\Pr(\boldsymbol{x_t} \mid z_t = i)$	$\Pr(\boldsymbol{x_t} \mid z_t = i, \boldsymbol{u_t})$	
forward variable $\alpha_{i,t}$	$\delta_{i,t} \sum_l \varphi_{li,t}\alpha_{l,t-1}$, with $\alpha_{i,1} = \pi_i\delta_{i,1}$		$\delta_{i,t} I_{i,t} \sum_l \alpha_{l,t-1}$, with $\alpha_{i,1} = I_{i,t}\delta_{i,1}$, if t is observed
backward variable $\beta_{i,t}$	$\sum_l \varphi_{il,t}\beta_{l,t+1}\delta_{l,t+1}$, with $\beta_{i,T} = 1$		$\sum_l I_{l,t+1}\beta_{l,t+1}\delta_{l,t+1}$, with $\beta_{i,T} = 1$, if $t+1$ is observed
complete data likelihood L_c	$\sum_i \alpha_{i,T}$		
posterior transition probability $\xi_{ij,t}$	$\varphi_{ij,t}\alpha_{i,t_1}\beta_{j,t}\delta_{j,t} \,/\, L_c$		
posterior state probability $\gamma_{i,t}$	$\alpha_{i,t}\beta_{i,t} \,/\, L_c$		

$$
\begin{aligned}
Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\boldsymbol{k}}\right) = & \sum_{i=1} \gamma_{i,1} \log \Pr\left(z_1 = i \mid \boldsymbol{u_1}; \boldsymbol{\theta_{in}}\right) \\
+ & \sum_{t=2}^{T} \sum_{i} \sum_{j} \xi_{ij,t} \log \Pr\left(z_t = j \mid z_{t-1} = i, \boldsymbol{u_t}; \boldsymbol{\theta_{tr}}\right) \\
+ & \sum_{t=1}^{T} \sum_{i} \gamma_{i,t} \log \Pr\left(\boldsymbol{x_t} \mid z_t = i, \boldsymbol{u_t}; \boldsymbol{\theta_{em}}\right). \qquad (5.2)
\end{aligned}
$$

In the above, $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\boldsymbol{k}}\right)$ is the expected value of the complete data log likelihood; k represents the EM iteration; T is the total number of timestamps in each sequence; $\boldsymbol{u_t}$, z_t and $\boldsymbol{x_t}$ are the inputs, hidden states, and observations at step t; and $\boldsymbol{\theta}$ are the model parameters to be estimated. The meaning of other variables is given in the first column of Table 5.2.

Transition and Emission models

The parameter estimation procedure of IOHMM described above implies that any supervised learning model that supports gradient ascent on the log probability can be integrated into the IOHMM. For example, in Equation 5.2, each of the model parameters ($\boldsymbol{\theta}$) can be estimated with neural networks. A neural network with a softmax layer can be used to learn the initial probability parameters ($\boldsymbol{\theta_{in}}$) through back-propagation, another neural network with a softmax layer for learning the transition probability parameters ($\boldsymbol{\theta_{tr}}$), and a third with customized layers for estimating emission model parameters ($\boldsymbol{\theta_{em}}$).

Note that the EM algorithm can be naturally implemented in a MapReduce framework, a programming model and an associated implementation for processing large data sets on computing clusters. The Expectation step can be fit into the Map step, calculating the posterior state probability γ and posterior transition probability ξ in parallel for each training sequence. The estimated posterior probabilities γ and ξ are collected in the Reduce step. The source code of an implementation developed as a part of this research is available from `https://github.com/Mogeng/IOHMM`.

5.2.2 Semi-Supervised Co-Training

Supervised learning of activity types requires data with labeled ground truth. In urban mobility, the ground truth activities are derived by either manually labeled [35], or collected for a small group of participants from a survey accompanying GPS data [31]. Privacy concerns and spatial resolution of CDR data precludes us from obtaining extensive ground truth labels. While fully unsupervised models can be used to cluster activities with similar temporal and spatial profiles, the recognized activities may not correspond to conventional activity types. In this subsection, we propose to use semi-supervised learning to reach a compromise – we use a small set of ground truth activities based on short range distributed antenna systems (DASs) to direct the learning process. As we have mentioned, short range antennas usually serve only a small range of area, which have relatively high spatial resolution. These short range antennas provide us the opportunity to label "ground truth" activities with Point of Interest (POI) information and domain knowledge.

Traditionally, semi-supervised learning is used to improve classifier performance, that is, to use "cheap" unlabeled data to assist training of labeled data. In our work, we adopt another view of semi-supervised approach, that is, we use labeled data to help direct the pattern recognition from unlabeled data. Zhu [61] did a thorough literature review on semi-supervised learning methods, including self-training, co-training, graph-based methods and Expectation-Maximization (EM) in generative models. In our work, we took the advantage of EM in generative models and co-training to improve the activity pattern recognition performance.

The idea behind co-training is that one uses two views of a sample that inform the learning algorithms by teaching one another. Ideally each sample is represented by two independent sets of features, which is however unlikely to exist [22]. Co-training can also be applied by using the same set of features but two different classifiers, which has been proven to perform well by [23]. It is expected to be less sensitive to mistakes than self-training.

In this work, we choose to use a semi-supervised IOHMM with EM algorithm as the generative classifier, and a decision tree (DT) classifier as its discriminative counterpart. With this combination, we have both the classification power of discriminative model and the generative power of IOHMM models.

The difference between IOHMM and semi-supervised IOHMM lies in the forward-backward algorithms. If we have ground truth activity (hidden states

Algorithm 1 Co-training of urban activities

```
Input: Labeled data $L$, unlabeled sequences $S$, confidence thresholds $\theta_1$ and $\theta_2$.
Output: IOHMM model $m_1$ and DT model $m_2$.
    Initialization: $L_1 = L_2 = L$
while $L_1$, $L_2$ changes do
    Train semi-supervised IOHMM $m_1$ from $S$ and $L_1$.
    Train DT model $m_2$ from $L_2$.
    Classify the unlabeled data with $m_1$ and $m_2$ separately.
    Add data labeled by $m_1$ with confidence $\geq \theta_1$ to $L_2$.
    Add data labeled by $m_2$ with confidence $\geq \theta_2$ to $L_1$.
end while
return $m_1$, $m_2$.
```

z) for timestamp t, then we will use $I_{j,t}$ to replace $\varphi_{ij,t}$ where $I_{j,t}$ is 1 if the hidden state $z_t = j$ at timestamp t in the labeled data, 0 otherwise, since $\Pr(z_t = j \mid z_{t-1} = i)$ reduces to $\Pr(z_t = j)$ with observed information. A summary of the differences between HMM, IOHMM and semi-supervised IOHMM is presented in TABLE 5.2.

5.2.3 Model Specifications

As we have mentioned, there are two components in the co-training process, one is the semi-supervised IOHMM with EM, and the other is the decision tree classifier. We will present our specifications (features) in this section.

Semi-Supervised IOHMM Model with EM

Input-Output Variables

In practice, models of simple structure (linear, multinomial logistic, Gaussian) with interpretable variables and parameters are preferred. For example, in an application below, we include the following input variables $\boldsymbol{u_t}$: (1) a binary variable indicating whether the day is a weekend; (2) five binary variables indicating the time of day that the activity starts, morning (5 to 10am), lunch (10am to 2pm), afternoon (12 to 2pm), dinner (4 to 8pm) or night (5pm to midnight); and (3) for the users with identified work location, the number of hours the user has spent at work this day. This variable contains accumulated knowledge on the past activities.

The IOHMM model also includes the following outputs $\boldsymbol{x_t}$ at each timestamp t: (1) $x^{(1)}$, the distance between the current stay location and the user's home; (2) $x^{(2)}$, the distance between the current stay location and the user's work place; (3) $x^{(3)}$, the duration of the activity; and (4) $x^{(4)}$, whether the user has visited this stay location cluster previously.

The selection of the inputs and outputs is guided by common knowledge. The activity start time is relevant for differentiating activity types. The number of hours worked in a day is a strong indicator of a person's likelihood to return to work (after a midday activity, for example). The model inputs contain information that is known at the start of the transition to a new activity. In contrast, the output features contain information that is not available at the transition to a new activity. For example the duration and the location or land-use in the vicinity of a new activity is unknown at the time of the transition. In other words, output variables can be observed when training the models, but must be inferred when sampling sequences of activities from the model.

The model outputs have a strong dependence on the activity type. For example, the distance that a person is willing to travel from home for a leisure trip may be longer than the distance that a person is willing to travel for a shopping trip. The duration depends both on the activity type, activity start time, and on the previous activities in the day. e.g., the expected duration of a work activity will decrease if a person has already worked in the day.

Initial, Transition and Emission Models

Multinomial logistic regression models are used as the initial probability model and transition probability models. Note that for succinctness, we use $\boldsymbol{\theta}$ in each of the following equations to represent the $\boldsymbol{\theta_{in,tr,em}}$ in Equation 5.2. The first term of Equation 5.2 can be written as:

$$\Pr\left(z_1 = i \mid \boldsymbol{u_1}; \boldsymbol{\theta}\right) = \frac{e^{\boldsymbol{\theta^i u_t}}}{\sum_k e^{\boldsymbol{\theta^k u_t}}}. \tag{5.3}$$

The $\boldsymbol{\theta}$ for initial probability model is a matrix with the i^{th} row ($\boldsymbol{\theta^i}$) being the coefficients for the initial state being in state i. The second term of Equation 5.2 can be written as:

$$\Pr\left(z_t = j \mid z_{t-1} = i, ; \boldsymbol{\theta}\right) = \frac{e^{\boldsymbol{\theta_i^j u_t}}}{\sum_k e^{\boldsymbol{\theta_i^k u_t}}}. \tag{5.4}$$

The $\boldsymbol{\theta}$ for transition probability models is a set of matrices with the j^{th} row of the i^{th} matrix ($\boldsymbol{\theta_i^j}$) being the coefficients for the next state being in state j given the current state being in state i.

To gain interpretability, we use linear models for the outputs represented as continuous random variables. We assume a Gaussian distribution for the distance to home and work variables $x^{(1)}$ and $x^{(2)}$ and the activity duration variable $x^{(3)}$. Where $x^{(1)}$ and $x^{(2)}$ depend only on the hidden activity type, the duration variable $x^{(3)}$ depends on the hidden activity and also the contextual input variables. The third term of Equation 5.2 can be written as:

$$\Pr\left(x_t \mid z_t = i, \boldsymbol{u_t}; \boldsymbol{\theta_i}\right) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_t - \boldsymbol{\theta_i} \cdot \boldsymbol{u_t})^2}{2\sigma_i^2}}, \tag{5.5}$$

The $\boldsymbol{\theta}$ for one such output emission model is a set of arrays where $\boldsymbol{\theta_i}$ and σ_i denote the coefficients and the standard deviation of the linear model when the hidden state is i. While we chose to represent outputs $x^{(1),(2),(3)}$ as Gaussian random variables, Gamma regression could be applied to duration $x^{(3)}$ to capture the non-negative, continuous, and right-skewed nature of these response variables. Moreover, response variables $x^{(1)}$ and $x^{(2)}$ could be modeled simultaneously using multivariate linear regression to capture the correlations between distance to home and distance to work.

Output $x^{(4)}$ is a binary variable, and we used logistic regression model as the output model. The probability in the third term of Equation 5.2 can be written as:

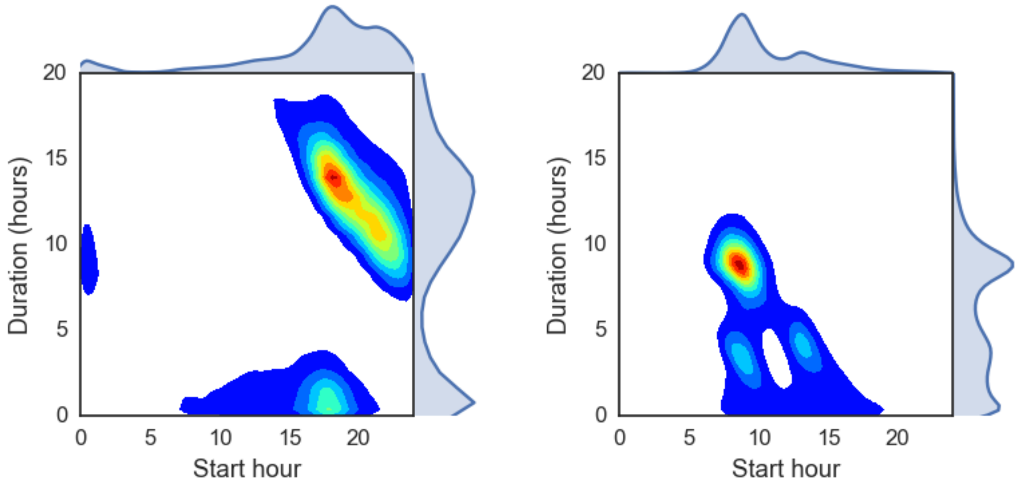
$$\Pr\left(x_t = 1 \mid z_t = i, \boldsymbol{u_t}; \boldsymbol{\theta_i}\right) = \frac{1}{1 + e^{-\boldsymbol{\theta_i} \cdot \boldsymbol{u_t}}}. \tag{5.6}$$

Decision Tree Counterpart

Decision trees are interpretable classifiers that are capable of generating arbitrarily complex decision boundaries. They have been used successfully in many diverse areas [43]. In this work, we use CART (Classification and Regression Trees) classifier. The features we include are the combination of input and output features in IOHMM.

5.2.4 Model Selection

Model selection for co-training includes the choice of hidden states. The choice should come directly from the collection of ground truth activities.



(a) Home

(b) Work

Figure 5.3: Joint distribution plot of duration and start hour for home (left) and work (right).

As we collected ground truth activities for “Food/Shop”, “Stop in Transit”, “Recreation”, “Personal Business”, and “Travel”, we include these five secondary activities in the hidden states.

We further noticed a significant heterogeneity within home and work activities. Temporal profile of home activities in Fig. 5.3a has two major clusters. The upper cluster indicates regular overnight home activities (H_1) and the lower cluster indicates short stay at home before going to some other activities (H_2). The temporal profile of work activities in Fig. 5.3b has three clusters. The upper cluster indicates regular “9 to 5” work activities without a break (W_1). The lower left cluster represents the morning work activities (W_2) and the lower right cluster represents the afternoon work activities (W_3). Considerably, the transition probability from H_2 to work is lower, and the transition probability from W_2 to “Food/Shop” should be higher but to “Recreation” should be lower than the transition probability from W_1 or W_3. By separating home and work activities into sub-activities, we expect to get better contextual-dependent transition probabilities. A more rigorous definition of sub-activities is:

1. H_1: cross day home activity that starts before 3:00 am and end after 3:00 am.

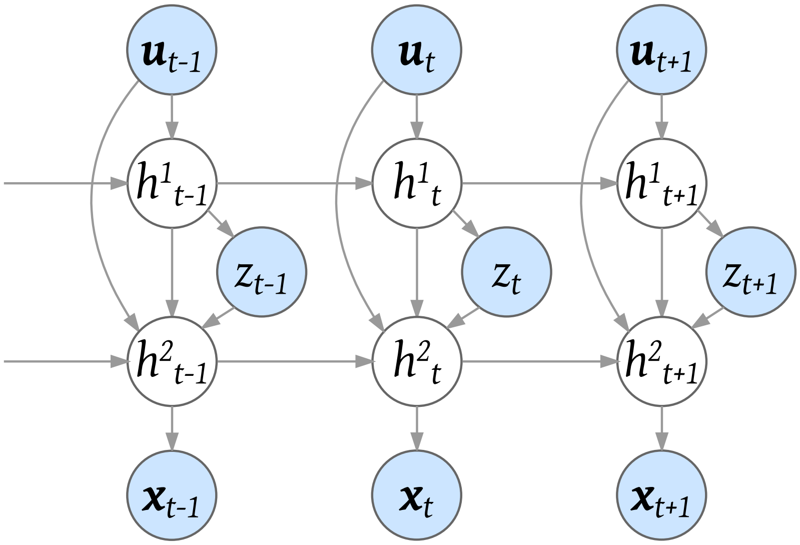


Figure 5.4: Deep LSTM Urban Mobility Architectures. The solid nodes represent observed information, while the transparent (white) nodes represent latent random variables. The top layer contains the *observed* input variables $\boldsymbol{u_t}$; the middle layer contains categorical variables z_t (latent in IOHMM since we include secondary activities while observed in LSTM since we only include "home", "work", and "other"); and the bottom layer contains observed output variables $\boldsymbol{x_t}$. h_t are LSTM cells in the LSTM architecture.

2. H_2: other home activities.

3. W_1: work activity if it is the only work activity in a day.

4. W_2: first work activity if there are more than one.

5. W_3: second work activity if there are more than one.

6. W_4: other work activities.

We compare experimentally the basic and extended specifications (one with 7 activities and the other with 11 activities) in Chapter 7.1.

5.3 LSTM Based Urban Mobility Models

LSTM models have been extensively used for modeling complex sequences, including natural language, videos and handwriting trajectories. We design a 2-layer LSTM model structure for modeling activity sequences as shown in Fig. 5.4.

The top layer models activity transitions between “home”, “work”, and “other” (we treat all secondary activities as “other” since we do not have full ground truth labels for all secondary activities). $\boldsymbol{u_t}$ represents the input contextual features similar to the ones specified in IOHMM models. The only difference is that we include the observed previous activity (one of “home”, “work”, and “other”) in this feature vector. The reasons are (1) in LSTM models, the previous activity type is observed prior to transition to a new activity, and (2) for generating new activity based on the previous activity, we need to include this previous activity in the training phase. Note that in IOHMM models, we use dynamic programming to get the probabilities of previous activity, as detailed in Chapter 6.1. h_t^1 represent the first layer of LSTM cells and z_t represents the observed current activity type. The loss function for this top layer is:

$$L_1\left(\boldsymbol{\theta_1}\right) \;\; = \;\; -\sum_{t=1}^{T}\sum_{j}\left(z_t = j\right)\cdot\log\phi\left(h_t^1;\boldsymbol{\theta_1}\right)_j$$

where ϕ is the softmax function, $\boldsymbol{\theta_1}$ is the collection of parameters for this LSTM neural network, and j belongs to one of the activity types “home”, “work” and “other”.

The bottom layer is a mixture density network (MDN) which models the **distributions** of spatial (location) and temporal (duration) variables $\boldsymbol{x_t}$ associated with each activity type z_t. MDN was first described in [10] and was further developed for handwriting synthesis tasks [27]. The contextual vector $\boldsymbol{u_t}$, first layer LSTM cells h_t^1, second layer LSTM cells from previous timestamp h_{t-1}^2, and the current activity type z_t are the inputs to the second layer LSTM cells h_t^2, which generates the coefficients of the mixture distributions (in our task we assume Gaussian distribution for each output feature) $\{\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}_{d_h}, \hat{\boldsymbol{\mu}}_{d_w}, \hat{\boldsymbol{\mu}}_{\text{st}}, \hat{\boldsymbol{\mu}}_{\text{dur}}, \hat{\boldsymbol{\sigma}}_{d_h}, \hat{\boldsymbol{\sigma}}_{d_w}, \hat{\boldsymbol{\sigma}}_{\text{st}}, \hat{\boldsymbol{\sigma}}_{\text{dur}}, \hat{\boldsymbol{\rho}}_{\text{st, dur}}\}$. At each timestamp t, $\hat{\boldsymbol{\pi}}_t$ is an M by 1 array representing the mixture component weights, M is the number of mixture components. $\hat{\boldsymbol{\mu}}_{d_h,t}$, $\hat{\boldsymbol{\mu}}_{d_w,t}$, $\hat{\boldsymbol{\mu}}_{\text{st},t}$, and $\hat{\boldsymbol{\mu}}_{\text{dur},t}$ are M by 1 arrays representing the component means of the distance to home, distance to work, start time, and duration. $\hat{\boldsymbol{\sigma}}_{d_h,t}$, $\hat{\boldsymbol{\sigma}}_{d_w,t}$, $\hat{\boldsymbol{\sigma}}_{\text{st},t}$, and $\hat{\boldsymbol{\sigma}}_{\text{dur},t}$ are M by 1 arrays representing the component standard deviations of the distance to home, distance to work, start time, and duration. $\hat{\boldsymbol{\rho}}_{\text{st, dur},t}$ represents the correlation between start time and duration. This second layer mixture networks is meant to divide “home”, “work”, and “other” activities into smaller and finer components, each has its local spatial-temporal distributions. The loss function for this bottom layer is:

$$L_2\left(\boldsymbol{\theta_2}\right) \quad = \quad \sum_{t=1}^{T} - \log \sum_{i}^{M} \pi_t^i \mathcal{N}(\boldsymbol{x_t} | \hat{\boldsymbol{\mu}}_t^i, \hat{\boldsymbol{\sigma}}_t^i, \hat{\boldsymbol{\rho}}_t^i)$$

where $\boldsymbol{\theta_2}$ is the collection of parameters of the neural network used to generate the mixture density distribution coefficients $\{\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\rho}}\}$, i is the index of the mixture component. $\mathcal{N}$ is the Gaussian probability density function.

This two-layer structure extends Lin et al. [37] as we moved the modeling of activity types into the first layer. Otherwise we keep the same model specifications and loss functions as in that report.

Chapter 6

Urban Mobility Prediction

The problem we are solving in this section is to predict the activity sequence of the rest of day, given partially observed sequences at a cut time (e.g. 9:00 am). This problem can be tackled by breaking it into two inferential sub-problems: (1) what an individual has done; and (2) what he/she is likely to do. We will show how these two sub-problems are tackled using IOHMM model and LSTM model, respectively.

6.1 Prediction using IOHMM models

6.1.1 Filtering

The first step is calculating $\Pr\left(z_{t-1} = i \mid \boldsymbol{u_{1,\ldots,t-1}}, \boldsymbol{x_{1,\ldots,t-1}}\right)$. Since the next activity to be generated depend on the contextual variables such as time of day and day of week information, as well as the previous hidden activity, we need to understand what is the last observed activity. There are two cases:

1. By the cut time, the last observed activity is completed. That is, the person is traveling to the next activity location. This case is simple since we can use standard forward algorithm to estimate the posterior probability $\Pr\left(z_{t-1} = i \mid \boldsymbol{u_{1,\ldots,t-1}}, \boldsymbol{x_{1,\ldots,t-1}}\right)$ of the last observed activity. One thing to note is that we need to sample a travel time that are longer than the observed travel time from the complete of the last activity to respect the fact that no new activities happened before the cut time.

2. By the cut time, the last observed activity is not completed. In this case, we apply a modification to the forward algorithm: the emission probability of duration of last activity is a survival function: $\Pr\left(x_t > d_t^o \mid z_t = i, \boldsymbol{u_t}\right)$, where d_t^o is the observed duration of the last activity until the cut time. After the filtering, we sample a new duration with the truncated distribution whose lower bound is d_t^o to respect the fact that the activity ends after the cut time.

6.1.2 Activity generation

With the last activity inferred, the activity generation algorithm is same as follows: at the end of this activity the relevant context information $\boldsymbol{u_t}$ is updated and the next activity is selected given the newly obtained transition probabilities. Next, the activity duration is sampled from the conditional distribution given the activity type and the start time. Next, the activity location is selected - if the activity is a home or work activity, the exercise is trivial. If not, we calculate the probability of choosing each cluster in the user's historical location clusters based on the conditional distribution of $x^{(1)}$ distance to home and $x^{(2)}$ distance to work given the activity type. By adopting the historical location clusters of the user, we reduce the variance of the location choice. The process continues until the full daily sequence of activities is generated.

Due to the nature of IOHMM, we must filter out and discard unrealistic activity chains generated in this process. We determine unrealistic activity chains to be chains that do not end the day at home and activity chains where 3 or more of the same activity type occur in a row. These filters constrain the overall structure of the day to be aligned with a feasible/conventional day structure. For simulation purposes we also filter activity chains that include long-distance travel out of the Bay Area.

6.2 Prediction using LSTM models

The procedure is straightforward based on Fig. 5.4. The LSTM model first calculates $h^1_{1,..,t-1}$, $h^2_{1,..,t-1}$ based on observed $\boldsymbol{u_{1,..,t-1}}$ and $z_{1,..,t-1}$. To generate the next activity at timestamp t, we first update the contextual vector $\boldsymbol{u_t}$ and top LSTM layer h^1_t. The softmax outputs of the top layer is used for sampling the new activity type z_t. z_t, along with $\boldsymbol{u_t}$, h^1_t, h^2_{t-1} are used in the bottom layer of the model. The sampling of the output variables distance to home, distance to work, and duration from the distributions of mixture density network (MDN) is similar to the ones described in [37, 27]. The rest of the generation process is similar to the generation process of IOHMM model.

Chapter 7

Experimental Results

In this section, we describe two regional experiments of medium term travel demand forecast at different times of day. The master data used in these studies comprise a month of anonymized and aggregated CDR logs collected in Summer 2015 by a major mobile carrier in the US, serving millions of customers in the San Francisco Bay Area. No personally identifiable information (PII) was gathered or used for this study. As described previously, CDR raw locations are converted into highly aggregated location features before any actual modeling takes places.

The first experiment use the City of San Francisco for model selection. We evaluate the prediction performance of different models and validate the predictions at individual and aggregated level. The second experiment scales to whole San Francisco Bay Area where we predict the traffic conditions based on trained models for commuters from each of the 34 super-districts. We evaluate the resulting traffic from micro-simulation and validate it against the resulting traffic of observed ground truth data.

We choose a typical weekday June 10, 2015 as the target day. For each regular commuter with available data on that day, we slice the data by different cut time (e.g. 3:00 am, 4:00 am, ..., 11:00 pm) and predict the activities for the rest of the day based on the observed information by the cut time.

7.1 Model Comparison

In this subsection, we evaluate the performance of different models and methods.

1. **NN**: Nearest Neighbor model, the benchmark model and the expected upper bound of the performance. NN is a fully personalized model that match the observed trajectory with the trajectory history of the user, and use the matched trajectory as prediction for the rest of day. The distance features we used are (1) difference in day type (weekday or weekend, 0 if equal and 1 if not), and (2) the Hamming distance between observed partial sequence and each historical sequence by cut time. We calculate the Hamming distance by segmenting each sequence into 15-minutes segments. For each 15-minutes segment, we set the distance as 0 if the location clusters in two sequences are same (in most of the 15 minutes) and 1 if not. The total Hamming distance is the sum of each segment. We give the day type feature a high weight (in this case 100) so that NN will search the matching sequence within the same day type. Note that NN model is only used for trajectory matching and does not provide insights and interpretability as other activity models.

2. **IOHMM-unsupervised-7**: The IOHMM model with 7 hidden states, with the input and output features specified in Chapter 5.2.3.

3. **IOHMM-co-training-7**: The co-training IOHMM model specified in Chapter 5.2.2. In this model we treat home and work as two activities, thus with 5 secondary activities there are 7 states in total. The threshold parameters for both semi-supervised IOHMM model with EM (θ_1) and Decision Tree (θ_2) are 0.9. This threshold is chosen based on literature and validation accuracies on secondary activity recognition.

4. **IOHMM-co-training-11**: In this model we separate “home” and “work” to 6 sub-activities defined in Chapter. 5.2.4. Thus there are 11 states in total.

5. **LSTM-3**: The LSTM model specified in Chapter 5.3. We used 64 hidden units in each LSTM cell and 40 mixture components in the mixture density network (MDN).

6. **LSTM-7**: In this model we separate “home” and “work” to 6 sub-activities thus there are 7 activity types including “other”.

In Fig. 7.1, we plot how the two validation metrics, (1) median travel distance error (left), and, (2) median Hamming distance (right) change for

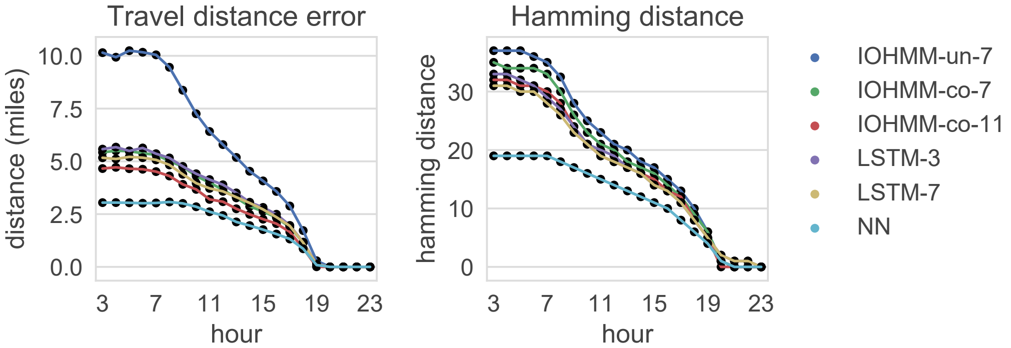


Figure 7.1: Models comparison. Two validation metrics are used: median travel distance error (left) and median Hamming distance (right). The x-axis is the prediction hour (cut hour) and the y-axis is the validation error. Each series of points represents the performance of a model.

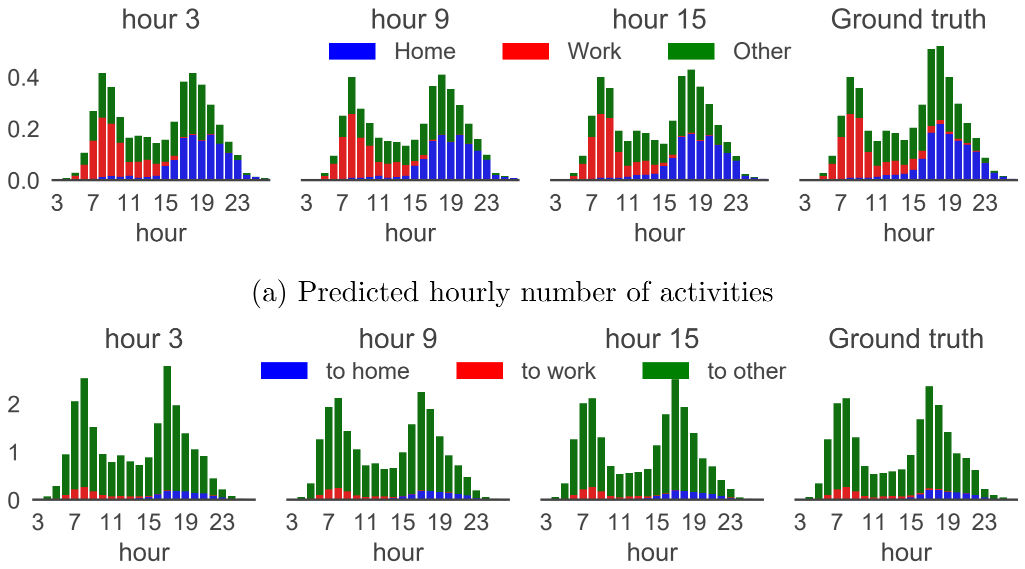
different cut hours using different models. The travel distance error is calculated as the difference between the observed daily travel distance and predicted daily travel distance. The median error of all users are used in the plot. The travel distance error mainly captures the spatial location choice performance of models. The Hamming distance is calculated as in NN models by segmenting the daily sequence into 96 discrete 15-minutes segments. The median error of all users are used in the plot. The Hamming error mainly captures the temporal day structure performance of models. From Fig. 7.1, we can see that: (1) NN models performs best among all models because it is a fully personalized non-parametric model; (2) IOHMM models are better at spatial performance than LSTM models since we used co-training to direct the learning of secondary activity profiles. This is also proven by comparing the unsupervised model performance with the co-training results; (3) LSTM models are better at capturing the day structures. Hamming error captures the performance of day structures such as “home”, “work”, and important secondary activities. LSTM models slightly outperforms IOHMM models in this metric because it is more flexible and deeper in modeling activity transitions and long term dependencies; (4) By separating “Home” and “Work” into smaller sub-activities, we get better spatial-temporal performance in both IOHMM models and LSTM-models. This proved our assumption that by separating these primary activities, we can better learn the activity transitions between primary activities and between primary activities and secondary activities; (5) We can explore the limit of the predictability of human mobility. The median travel distance error at the beginning of the day using fully personalized model is about 3 miles, and this number is about 5 miles using non-parametric group models. The median Hamming error is 20 at the beginning of the day using fully personalized model, that is, 5 hours of wrongly predicted activities within a day. This error is mainly due to the shift in home and work hours. Since different people has different start hour of work and preferences on the time of going back home, fully personalized model is better at capturing this based on the individual's history.

7.2 Aggregated Level Evaluation

We validate the predicted versus observed hourly aggregated travel behavior in this subsection. We adopt the IOHMM-co-training-11 as our urban mobility model. The aggregated pattern is very similar between the best performed IOHMM and LSTM models.

Fig. 7.2a shows the average number of activities (y-axis) starting in each hour (x-axis). To make it more informative, we decompose the total number of activities into "home", "work" and "other". We can see that the predicted number of activities of each type is quite comparable to the ground truth observed at the end of the day. The same peak of work activities in the morning and home activities in the evening are observed in all predictions and ground truth. The main difference between our predictions and the ground truth is that we tend to under-predict the number of "other" activities.

Fig. 7.2b shows the average travel distance in miles (y-axis) in each hour (x-axis). One observation is that the travel distance of "to work" in the morning peak and "to home" in the evening peak are low compared to "to other". This is because some people go for secondary activities before arriving at work and home, as shown in Fig. 7.2a. The other observation is that though the predicted number of secondary activities is lower, the travel distances to these locations are higher in our predictions. This indicates some inefficiencies in our secondary location choice - people select most convenient locations for secondary activities, and points towards possible improvements in location choice model for secondary activities.



(a) Predicted hourly number of activities

(b) Predicted hourly travel distance in miles

Figure 7.2: Predicted aggregated travel demand. The average number of activities (top) and travel distance in miles (bottom) (y-axis) starting in each hour (x-axis). Each of the four subplot represents the prediction at hour 3:00 am, 9:00 am, 3:00 pm, and the observed ground truth.

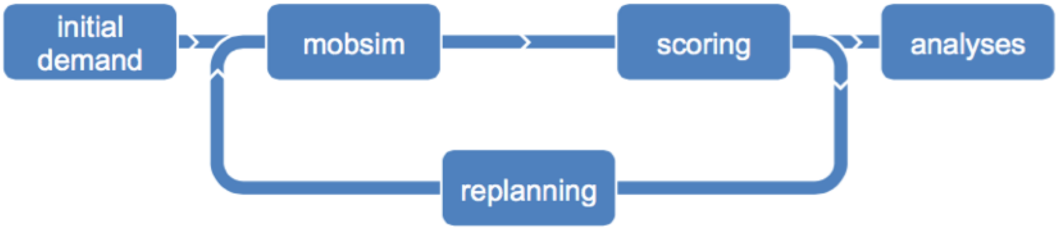


Figure 7.3: The MATSim Cycle [2]

7.3 Evaluation via Traffic Micro-simulation

In this subsection, we span the scope of the study to the 34 super-districts as defined by the San Francisco Metropolitan Transportation Commission (MTC) to validate the predicted resulting traffic in a region with 7.5M citizens. Since most of the short range DASs are located in urban area such as the City of San Francisco, the ground truth secondary activities are rarely available for other super-districts in Bay Area. Thus we train 34 semi-supervised IOHMM model with “home” and “work” as ground truth, one for each super-district. For each regular commuter with data available on June 10, 2015, we predict his/her activities for the rest of day based on the activities observed by a cut time. Traffic micro-simulation is a conventional approach in studying performance and evaluating transportation scenarios. The MATSim (Multi-Agent Transport Simulation) platform is an agent-based activity model that performs microscopic modeling of traffic (using link performance functions) and agent decision making [2]. The MATSim run cycle, as shown in Figure 7.3, is an iterative process whereby agents make adaptations to routing, activity timing, and other optional choices until convergence is reached. As input, each agent is assigned an activity chain (initial demand), complete with activity types, timing and location. During the mobility simulation (mobsim), the agents travel the network, interact, and experience congestion which lowers their overall utility scores for the day. During the replanning phase, a subset of agents may adapt their routes and activity timings. For our simulations, we restricted replanning adaptation to random selection of 10% of the population during each iteration. Many other forms of adaption are possible with MATSim, but for this project we have restricted adaptation to timing and routing. Agents incur a negative penalty for deviating from their original activity timings, so dramatic shifts in activity start and end times are not possible. Rerouting agents are allowed to update their routes to the new shortest path, based on the loaded network conditions in the most recent mobility simulation.

The MATSim road network was created using OpenStreetMap (OSM) road network data, downloaded in July, 2015. The user-generated OSM data offers very complete coverage in major metropolitan regions as well as rich feature sets including: link distance, number of lanes, speed limit, and hierarchical road classification. A manual inspection of dozens of freeway links in California found the OSM features to be accurate.

The data was clipped and filtered using Osmosis, an open source Java application for editing OSM data. The OpenStreetMap Standards and Conventions define tags for classifying roads hierarchically. There are 14 tags which encompass nearly all road links in the dataset. These range from "motorway" and "trunk" down to "residential" and smaller hierarchical classes. We found that for the Bay Area, the "residential" links constitute 74% of all links in the network. By filtering out the "residential" links, we were able to greatly improve the computational running time of MATSim without compromising regional-scale demand patterns. It is possible to maintain "residential" links for a localized area for future studies which require accurate neighborhood-level traffic reproduction. However, other limiting factors, such as the realism of MATSim's queueing, traffic signal, and physics engines call into question the efficacy of including the lowest hierarchy links in the network.

Once filtered, the geometry was simplified to a straight-line network to improve simulation speeds. Each intersection is a node, and a straight edge represents the road link connecting two intersections. To maintain realistic travel time skims, attributes of the original geometry network are preserved as attributes of link objects: length and free-flow travel speed. The final network used in the Smart Bay studies includes 564,368 links, and 352,011 nodes.

Our experiment is as follows: For each cut time (e.g. 3:00 am, 9:00 am, 3:00 pm, 9:00 pm), we compared the results of the flows produced on the Bay Area network containing all freeways and primary and secondary roads (a total of 24'654 links) from the predicted activity sequences with the ground truth activity sequences. The fit score (1) adjusted R^2; (2) mean absolute percentage error (MAPE, %) are summarized in TABLE 7.1. Fig. 7.4 plots the volume profiles of two freeway locations, one near the entrance of bay bridge in the eastbound and the other near the crossing of I-880 and US-101. For each location, 4 subplots shows the predictions (in blue) at 3:00 am, 9:00 am, 3:00 pm and 9:00 pm vs the ground truth profiles (in orange). We can see that (1) the predictions get closer to the ground truth volumes with more observed data in the day and (2) our predictions tend to generate slightly higher traffic volumes than ground truth traffic. This is consistent with our previous discussion on the inefficiencies in secondary location choices.

TABLE 7.1 proves that we can use observed information of the day to improve traffic volume prediction. The coefficient of determination increase and the MAPE decrease with the prediction hour. When we make prediction

Table 7.1: The coefficient of determination (R^2) and mean absolute percentage error (MAPE, %, in the parenthesis) of the predicted versus ground truth resulting traffic counts on 600 locations on the Bay Area road network. The row index is the prediction hour and the column index is the predicted hour. No scores are reported under diagonal because the traffic in the predicted hour is already observed by the prediction hour.

	3	6	9	12	15	18	21	24
3	1	0.864	0.881	0.876	0.890	0.891	0.924	0.896
	(0)	(38.1)	(16.2)	(18.0)	(19.1)	(14.2)	(14.5)	(19.7)
9	-	-	0.997	0.977	0.947	0.931	0.934	0.937
	-	-	(2.9)	(9.0)	(14.1)	(10.8)	(13.4)	(15.1)
15	-	-	-	-	0.995	0.962	0.960	0.955
	-	-	-	-	(4.4)	(8.8)	(11.1)	(13.0)
21	-	-	-	-	-	-	0.999	0.998
	-	-	-	-	-	-	(2.1)	(3.8)

at the beginning of each hour, we can improve the coefficients of determination in that hour to be greater than 0.99 and the MAPE less than 5%. The artifact of perfect prediction of 3:00 am is because we defined the start of the day as 3am, there should be few traffic occurring during that hour. If we predict three hours ahead (e.g. prediction of 6:00 pm traffic at 3:00 pm), the coefficients of determination are greater than 0.96 and the MAPEs are less than 10% (except for the prediction for 6:00 am). The lower predictability at off-peak hours (e.g. 6:00 am and 12:00 am) is consistent with the observations in [54] of higher variability in travel choices for secondary activities.

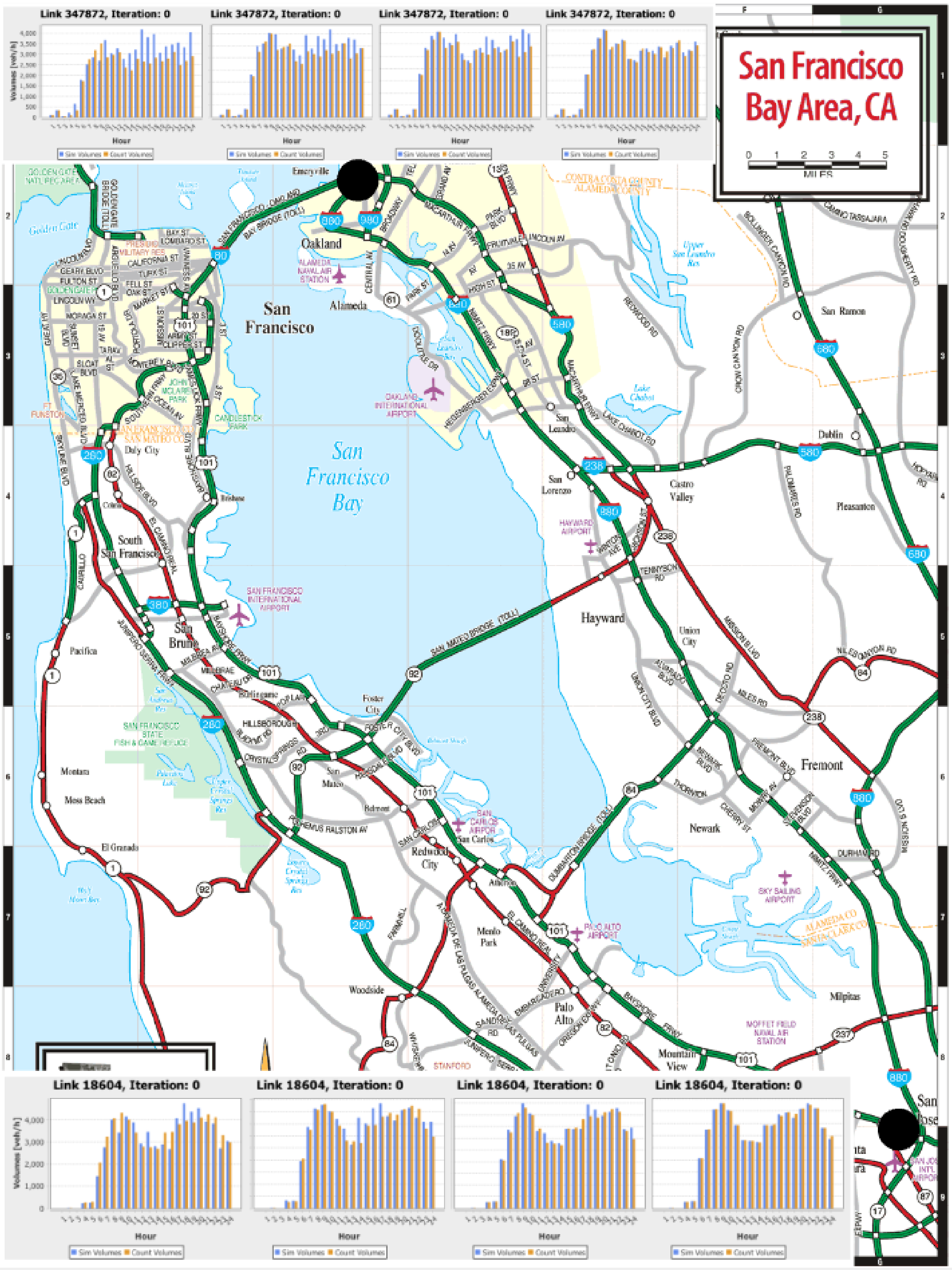


Figure 7.4: A fragment of the SF Bay Area road network. Inlet graphs illustrate two sample hourly vehicle volume profiles for observed (orange) and predicted (blue) at 3am, 9am, 3pm, and 9pm.

Chapter 8

Conclusion and Recommendations

In this report, we proposed a medium term travel demand nowcasting system. It predicts daily travel demand and traffic conditions at different times of day with partially observed user traces from cellular data and pre-trained urban mobility models. This solution bridges the gap between long term forecast (days, months to years ahead) and short term prediction (seconds to hours ahead), which are the two mainstreams of literature in travel demand forecasting.

We improved the state-of-the-art deep generative parametric mobility models using co-training in IOHMM and LSTMs. We provided partially observed user traces at different times of day to these models and generated the complete daily sequences. We validated the results with the ground truth sequences based on (1) individual level discrepancies; (2) aggregated level hourly travel demand; and (3) the resulting traffic through micro-simulation. A non-parametric individualized nearest neighbor model was explored as the practical limit of predictability of individual's daily travel. We demonstrated that parametric models trained at aggregated group level (due to privacy concern) can approach this limit in terms of prediction accuracy. Among the generative models we compared, IOHMM models are interpretable and has the power of activity recognition as a range of travel choices might depend on the activity types. Co-training applied to IOHMM models performs better at secondary activity location choices since we used the ground truth activities to direct the learning process. LSTM models are better at learning day structures since they use continuous hidden state space and are expected to be better at learning long term dependencies. Future research will focus on incorporating activity types in LSTM models and using existing ground truth labels to direct the learning process of LSTM models.

We consider San Francisco residents as a group in the first experiment and each super-district as a group in the second experiment. We trained one urban mobility model for each group. However, certain heterogeneity in activity patterns exists among different sub-groups. Correctly partitioning the population into sub-groups should help us better approach the limit of the predictability in human mobility. We acknowledge it as a current limitation of the report.

In terms of traffic volumes, our experiments show promising results of medium term forecast. We have reached a MAPE of less than 5% one hour ahead and 10% three hour ahead. Results also show that we can improve the prediction accuracy by incorporating more of the observed data by the time of prediction. Our prediction of traffic conditions is available not only for freeways and arterial where high-resolution detectors data are available from direct observations. Our system provides accurate prediction for the whole network, detailed in terms of activities and travel itineraries of citizens, providing an actionable model to improve performance of regional transportation systems and inform interventions towards reducing negative impacts of congestion.

References

[1] *2010-2012 California Household Travel Survey Final Report Appendix.* http://www.dot.ca.gov/hq/tpp/offices/omsp/statewide_travel_analysis/files/CHTS_Final_Report_June_2013.pdf.

[2] Andreas Horni, Kai Nagel, and Kay Axhausen, eds. *The Multi-Agent Transport Simulation: MATsim.* Apr. 2016. URL: http://www.matsim.org/docs/userguide.

[3] Theo Arentze et al. "Data needs, data collection, and data quality requirements of activity-based transport demand models". In: *Transportation research circular* E-C008 (2000), 30–p.

[4] Daniel Ashbrook and Thad Starner. "Using GPS to learn significant locations and predict movement across multiple users". In: *Personal and Ubiquitous computing* 7.5 (2003), pp. 275–286.

[5] Mitra Baratchi et al. "A hierarchical hidden semi-Markov model for modeling mobility data". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM. 2014, pp. 401–412.

[6] Gary Barnes and Gary A Davis. "Understanding Urban Travel Demand: Problems, Solutions, and the Role of Forecasting". In: (1999).

[7] Moshe E Ben-Akiva and Steven R Lerman. *Discrete choice analysis: theory and application to travel demand.* Vol. 9. MIT press, 1985.

[8] Yoshua Bengio and Paolo Frasconi. "An input output HMM architecture". In: (1995).

[9] Chandra R Bhat and Sujit K Singh. "A comprehensive daily activity-travel generation model system for workers". In: *Transportation Research Part A: Policy and Practice* 34.1 (2000), pp. 1–22.

[10] Christopher M Bishop. "Mixture density networks". In: (1994).

[11] Wendy Bohte and Kees Maat. “Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands”. In: *Transportation Research Part C: Emerging Technologies* 17.3 (2009), pp. 285–297.

[12] John L Bowman and Moshe E Ben-Akiva. “Activity-based disaggregate travel demand model system with activity schedules”. In: *Transportation Research Part A: Policy and Practice* 35.1 (2001), pp. 1–28.

[13] Francesco Calabrese et al. “Understanding individual mobility patterns from urban sensing data: A mobile phone trace example”. In: *Transportation research part C: emerging technologies* 26 (2013), pp. 301–313.

[14] Joe Castiglione, Mark Bradley, and John Gliebe. *Activity-based travel demand models: a primer*. Tech. rep. 2014.

[15] Jingmin Chen and Michel Bierlaire. “Probabilistic multimodal map matching with rich smartphone data”. In: *Journal of Intelligent Transportation Systems* 19.2 (2015), pp. 134–148.

[16] Eunjoon Cho, Seth A Myers, and Jure Leskovec. “Friendship and mobility: user movement in location-based social networks”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, pp. 1082–1090.

[17] Trinh Minh Tri Do and Daniel Gatica-Perez. “Where and what: Using smartphones to predict next locations and applications in daily life”. In: *Pervasive and Mobile Computing* 12 (2014), pp. 79–91.

[18] Nathan Eagle, Aaron Clauset, and John A Quinn. “Location Segmentation, Inference and Prediction for Anticipatory Computing.” In: *AAAI Spring Symposium: Technosocial Predictive Analytics*. 2009, pp. 20–25.

[19] Nathan Eagle and Alex Sandy Pentland. “Eigenbehaviors: Identifying structure in routine”. In: *Behavioral Ecology and Sociobiology* 63.7 (2009), pp. 1057–1066.

[20] Yingling Fan et al. “SmarTrAC: A smartphone solution for context-aware travel and activity capturing”. In: (2015).

[21] Katayoun Farrahi and Daniel Gatica-Perez. “Discovering routines from large-scale human locations using probabilistic topic models”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.1 (2011), p. 3.

[22] Volkmar Frinken et al. “Co-training for handwritten word recognition”. In: *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE. 2011, pp. 314–318.

[23] Sally Goldman and Yan Zhou. “Enhancing supervised learning with unlabeled data”. In: *ICML*. 2000, pp. 327–334.

[24] Hongmian Gong et al. “A GPS/GIS method for travel mode detection in New York City”. In: *Computers, Environment and Urban Systems* 36.2 (2012), pp. 131–139.

[25] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. “Understanding individual human mobility patterns”. In: *Nature* 453.7196 (2008), pp. 779–782.

[26] KoNSTADINos G GouuAs and Ryuichi Kitamura. “Travel demand forecasting with dynamic microsimulation”. In: (1992).

[27] Alex Graves. “Generating sequences with recurrent neural networks”. In: *arXiv preprint arXiv:1308.0850* (2013).

[28] Sibren Isaacman et al. “Identifying important places in peoplefffdfffdfffds lives from cellular network data”. In: *Pervasive computing*. Springer, 2011, pp. 133–151.

[29] Shan Jiang et al. “A review of urban computing for mobile phone traces: current methods, challenges and opportunities”. In: *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM. 2013, p. 2.

[30] *Jobs per Square Mile*. `http://http://www.sustainablecommunitiesindex.org/indicators/view/209`.

[31] Youngsung Kim et al. “Activity recognition for a smartphone based travel survey based on cross-user history data”. In: *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE. 2014, pp. 432–437.

[32] John Krumm and Eric Horvitz. “Predestination: Inferring destinations from partial trajectories”. In: *International Conference on Ubiquitous Computing*. Springer. 2006, pp. 243–260.

[33] Kevin S Kung et al. “Exploring universal patterns in human home-work commuting from mobile phone data”. In: (2014).

[34] Kyunghan Lee et al. “Slaw: A new mobility model for human walks”. In: *INFOCOM 2009, IEEE*. IEEE. 2009, pp. 855–863.

[35] Lin Liao, Dieter Fox, and Henry Kautz. “Extracting places and activities from gps traces using hierarchical conditional random fields”. In: *The International Journal of Robotics Research* 26.1 (2007), pp. 119–134.

[36] Lin Liao, Dieter Fox, and Henry Kautz. “Location-based activity recognition”. In: *Advances in Neural Information Processing Systems* 18 (2006), p. 787.

[37] Ziheng Lin et al. “Deep Generative Models of Urban Mobility”. In: *Submitted to ICDM 2017.* (2017).

[38] Charles M Macal and Michael J North. “Tutorial on agent-based modeling and simulation”. In: *Simulation Conference, 2005 Proceedings of the Winter*. IEEE. 2005, 14–pp.

[39] Anna Monreale et al. “Wherenext: a location predictor on trajectory pattern mining”. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2009, pp. 637–646.

[40] Santi Phithakkitnukoon et al. “Activity-aware map: Identifying human daily activity pattern using mobile phone data”. In: *International Workshop on Human Behavior Understanding*. Springer. 2010, pp. 14–25.

[41] Soora Rasouli and Harry Timmermans. “Activity-based models of travel demand: promises, progress and prospects”. In: *International Journal of Urban Sciences* 18.1 (2014), pp. 31–60.

[42] Sasank Reddy et al. “Using mobile phones to determine transportation modes”. In: *ACM Transactions on Sensor Networks (TOSN)* 6.2 (2010), p. 13.

[43] S Rasoul Safavian and David Landgrebe. “A survey of decision tree classifier methodology”. In: *IEEE transactions on systems, man, and cybernetics* 21.3 (1991), pp. 660–674.

[44] Salvatore Scellato et al. “NextPlace: a spatio-temporal prediction framework for pervasive systems”. In: *International Conference on Pervasive Computing*. Springer. 2011, pp. 152–169.

[45] Christian M Schneider et al. “Unravelling daily human mobility motifs”. In: *Journal of The Royal Society Interface* 10.84 (2013), p. 20130246.

[46] Chaoming Song et al. “Limits of predictability in human mobility”. In: *Science* 327.5968 (2010), pp. 1018–1021.

[47] Libo Song et al. “Evaluating location predictors with extensive Wi-Fi mobility data”. In: *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*. Vol. 2. IEEE. 2004, pp. 1414–1424.

[48] Xuan Song, Hiroshi Kanasugi, and Ryosuke Shibasaki. “Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level”. In: IJCAI. 2016.

[49] Arvind Thiagarajan et al. “VTrack: accurate, energy-aware road traffic delay estimation using mobile phones”. In: *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*. ACM. 2009, pp. 85–98.

[50] Eleni I Vlahogianni, Matthew G Karlaftis, and John C Golias. “Short-term traffic forecasting: Where we are and where wefffdfffdfffdre going”. In: *Transportation Research Part C: Emerging Technologies* 43 (2014), pp. 3–19.

[51] Peter Widhalm et al. “Discovering urban activity patterns in cell phone data”. In: *Transportation* 42.4 (2015), pp. 597–623.

[52] Jihang Ye, Zhe Zhu, and Hong Cheng. “What’s your next move: User activity prediction in location-based social networks”. In: *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM. 2013, pp. 171–179.

[53] Qing Ye, Wai Yuen Szeto, and Sze Chun Wong. “Short-term traffic speed forecasting based on data recorded at irregular intervals”. In: *IEEE Transactions on Intelligent Transportation Systems* 13.4 (2012), pp. 1727–1737.

[54] Mogeng Yin et al. “A generative model of urban activities from cellular Data”. In: *IEEE Transactions in ITS* (2017).

[55] Josh Jia-Ching Ying et al. “Semantic trajectory mining for location prediction”. In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM. 2011, pp. 34–43.

[56] Fangfang Zheng and Henk Van Zuylen. "Urban link travel time estimation based on sparse probe vehicle data". In: *Transportation Research Part C: Emerging Technologies* 31 (2013), pp. 145–157.

[57] Jiangchuan Zheng and Lionel M Ni. "An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM. 2012, pp. 153–162.

[58] Yu Zheng et al. "Learning transportation mode from raw gps data for geographic applications on the web". In: *Proceedings of the 17th international conference on World Wide Web*. ACM. 2008, pp. 247–256.

[59] Yu Zheng et al. "Mining interesting locations and travel sequences from GPS trajectories". In: *Proceedings of the 18th international conference on World wide web*. ACM. 2009, pp. 791–800.

[60] Yu Zheng et al. "Understanding mobility based on GPS data". In: *Proceedings of the 10th international conference on Ubiquitous computing*. ACM. 2008, pp. 312–321.

[61] Xiaojin Zhu. "Semi-supervised learning literature survey". In: (2005).