Activity-Based Urban Mobility Modeling from Cellular Data

by

Mogeng Yin

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Alexei Pozdnoukhov, Chair
Professor Alexei (Alyosha) Efros
Professor Mark Hansen

Spring 2018

Activity-Based Urban Mobility Modeling from Cellular Data

Abstract

Activity-Based Urban Mobility Modeling from Cellular Data

by

Mogeng Yin

Doctor of Philosophy in Civil and Environmental Engineering

University of California, Berkeley

Assistant Professor Alexei Pozdnoukhov, Chair

Transportation has been one of the defining challenges of our age. Transportation decision makers are facing difficult questions in making informed decisions. Activity-based travel demand models are becoming essential tools used in transportation planning and regional development scenario evaluation. They describe travel itineraries of individual travelers, namely what activities they are participating in, when they perform these activities, and how they choose to travel to the activity locales. However, data collection for activity-based models is performed through travel surveys that are infrequent, expensive, and reflect changes in transportation with significant delays. Thanks to the ubiquitous cell phone data, we see an opportunity to substantially complement these surveys with data extracted from network carrier mobile phone usage logs, such as call detail records (CDRs). The large scale cellular data also opens up the opportunities for researchers to study urban mobility, population estimation, disaster response and social events, etc. However, most of the urban mobility models from cellular data focus on only one aspect of urban mobility (such as location, duration, or travel mode), or model several aspects separately. Moreover, most urban mobility studies ignore the activity types (trip purposes) since the information are not naturally available from the raw cellular traces. These trip purposes carry important information in activity-based travel demand modeling since many travel decisions depend on these activity types, such as travel mode and destination location.

In this dissertation, we explore a framework that develops the state-of-the-art generative activity-based urban mobility models from raw cellular data, with the capability of inferring activity types for complementing activity-based travel demand modeling.

To do so, we first present a method of extracting user stay locations from raw and noisy cellular data while not over-filtering short-term travel. Significant locations such as home and work places are inferred. Along this pre-processing pipeline, we also produce meaningful aggregated statistics about how people construct their daily lives and participate in activities. These statistics used to be available purely from traditional travel surveys, thus were updated very infrequently.

With the processed yet unlabeled activity sequences, we improve the state-of-the-art generative activity-based urban mobility models step by step. First, we designed a method of collecting ground truth activities with the help from short range distributed antenna system (DAS), which has high spatial resolution. As a vanilla model, we first developed Input-Output Hidden Markov Models (IO-HMMs) to infer travelers' activity patterns. The activity patterns include primary and secondary activities' spatial and temporal profiles and heterogeneous activity transitions depending on the context. To have a directed learning process, we explored several semi-supervised approaches, including self-training and co-training. The co-training model has both the generative power of IOHMM model and the discriminative nature of decision tree model.

We apply the models to the data collected by a major network carrier serving millions of users in the San Francisco Bay Area. Our activity-based urban mobility model is experimentally validated with three independent data sources: aggregated statistics from travel surveys, a set of collected ground truth activities, and the results of a traffic micro-simulation informed with the travel plans synthesized from the developed generative model. As a classification task, we found that our full IOHMM outperforms partial IOHMM which outperforms standard HMM since IOHMM can incorporate more contextual information. We also found that co-training outperforms self-training, which outperforms the unsupervised IOHMM, thanks to the guidance of ground truth samples. This work is our first effort in exploring an end-to-end actionable solution to the practitioners in the form of modular and interpretable activity-based urban mobility models.

One direct application of the urban mobility model is travel demand forecasting. 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 dissertation, we present a medium term (hours to days ahead) travel demand forecast system. Our approach is designed to use cellular data that are 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 long term forecast and short term prediction.

We further improved our urban mobility models in two directions. We first separated home and work activity into smaller sub-activities, expecting to get better activity transition probabilities. On the other hand, we made our IOHMM deeper and continuous in hidden state space, with the help of long short term memory units (LSTM). Experimental results show that IOHMMs used in a semi-supervised manner perform well for location prediction while 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. Results show that we can improve the prediction accuracy by incorporating more of the observed data by the time of prediction. We can reach a mean absolute percentage error (MAPE) of less than 5% one hour ahead and 10% three hours ahead.

To my mom, dad and grandparents.

Contents

List of Figures

List of Tables

Acknowledgments

Cal gave me everything I could ever wished. I would not be who I am and would not have what I have without my five years at Cal. In this magical campus, I had success and failure. I had laughter and tears. I had enthusiasm and weariness. I had inspirations and frustration. I learn and I teach. I receive and I give. There are so many things I will miss and I have started to miss. I will miss the McLaughlin Hall, which is cool in the summer and warm in the winter, and more importantly, it is the harbor for me. I will miss the saber-toothed tiger who always give me five when we meet. I will miss my offices 107D and 118 where I spent even more time than at home. I will miss the ITS library where I played board games and ping-pong with friends, where I also stayed up whole nights to finish projects. Most importantly, I will miss the people I met at Cal, without whom my life at Cal would not have been this rich.

I have my deepest gratitude for my advisor, Dr. Alexei Pozdnoukhov. You fight for all the best opportunities for me. Without you, I would not have such an interesting topic that I can dedicate myself to study. I would not have the opportunity to intern at AT&T with Jean-Francois and to intern at Sidewalk Labs. I would not have the opportunity to broaden my view on computer science and machine learning in many conferences and symposiums. I would also have no chance of getting a M.S degree in computer science without your recommendation. I will always remember the times when we sit together in your office to work on proposals. I will always remember the times we revise our papers at night and submit them a few minutes before the deadline. I will also always remember the times we play ping-pong at different locations as friends.

I am also extremely grateful to my committee chair Professor Mark Hansen. Thank you for always caring about me as your student, even if you are not obligated to do so. I still remember that you provided my with the funding as a GSI for CE93 during my second year, without which I would have no source to cover my expense. I also learned from you that it is the idea and direction that are the most important things in research, not the ways to complicate models.

I would also like to thank my committee member from the computer science department, Professor Alexei (Alyosha) Efros. Thank you so much for spending time reading my prospectus, rehearsing with me on my presentations, and asking priceless questions regarding my research. You gave me new ideas in trying different models from a computer science perspective. You also gave me the enlightenment on how to make a good presentation for people both in and outside the area, which is a life-long lesson for me.

Aside from my committee for my PhD, I would also like to thank other faculty members in our lovely department for their wonderful teaching, patient guidance and ongoing support, including Professor Joan Walker, Professor Max Shen, Professor Michael Cassidy Professor Carlos Daganzo, Professor Raja Sengupta, and Professor Alexander Skabardonis.

My special thanks go to K. Shankari. I joined your “e-mission” project group in my second semester at Cal. Back to that time, I had little knowledge about computer science and application program development. I still believe I did more damage than contribution to the project. It was your patience, trust, and ongoing teaching that made me realize that I can make it. It was you who made up my mind to step into the world of computer science and urban computing. I would not be who I am without the magical meet with you. I would also like to extend my gratitude to Shankari's advisor in the computer science department, Professor David Culler and Professor Randy Katz for relating me like their student.

I would like to thank Professor Alex Bayen for advising me and caring about me. Without you I would not have the chance to get the CS master at Cal. I thank my AT&T advisor Jean-Francois Paiement. You are both my mentor and my friend. You give me advice on my research and on my life. You always help me solve problems and wish the best for me. I thank my advisor at undergraduate school and also a Cal alumni, Professor Meng Li. Without your recommendation, I would not even have the opportunity to come to Cal. I thank Shelley Okimoto, Bernadette Edwards, Helen Bassham, Jeanne Marie and all the staff in our department for their immense help, without which my PhD journey would not have been so smooth.

My Journey at Cal would not have been so happy without my peers. From my seniors Haotian Liu, Yi Liu, Michael Seelhorst, Xiaofei Hu, Lu Hao, Akshay Vij, Jinwoo Li, Darren Reger, Haoyu Chen, to my class Yanqiao Wang, Dounan Tang, Han Cheng, Lei Kang, Sreeta Gorripaty, Timothy Brathwaite, to my junior Chao Mao, Wei Ni, Menqiao Yu. And special thanks go to my teammates in Alexei's group Andrew Campbell, Sid Feygin, Ziheng Lin, Max Gardner, Colin Sheppard, Sudatta Mohanty, Maddie Sheehan, Danqing Zhang and Sangjae Bae. I will always remember the times we chat, play, and discuss. You have been inspirations for my research and my life.

My best friends at Cal, who have also been my friends since we were teenager, Zeshi Zheng and Bingyao Zhu, always got my back when I need them. They taught me all the things I need to learn to live in a foreign country. My life would not have been so rich and colorful without them.

Most importantly, I am most grateful to my mom and dad for their unconditional love and support. I would not have a chance to pursue my PhD study without your time, money and patience on educating me. You taught me to be strong, tough, courageous and modest by practising so since I was a little boy. To my dearest grandma, my thoughts are always with you. My happy memories with you led me through ups and downs over these years. I hope I have made you proud of me.

Chapter 1

Introduction

1.1 Motivation

In the United States, transportation system is critical to meeting the mobility and economic needs of local communities, regions, and the nation. Major challenges that roadway transportation faces are increasing traffic congestion, accidents, transportation delays, and vehicle emissions. According to the 2012 urban mobility report [107], in 2011, the average delay per commuter due to the congestion was 38 hours, this number was 61 in Bay Area. Urban Americans suffered from a congestion cost of $120 billion, that was $820 per commuter. This number was $1300 in Bay Area. The excess CO_2 emission caused by congestion was 10 billion lbs, that was 160 lbs per commuter, and 500 lbs in Bay Area. To address the current problems and meet the growing travel demand, the solution is either expanding roadway infrastructure or efficiently and effectively using existing infrastructure [46]. It is widely recognized, however, that the opportunities for building new physical infrastructure are decreasing because of increasing cost, environmental impact, and space limitations. Developments in research and technology such as advanced materials, communications technology, new data collection technologies, and human factors science offer a new opportunity to improve the traffic management.

However, transportation decision makers confront difficult questions to make informed choices [25]. How will the national, regional, or even local transportation system perform 30 years into the future? What policies or investments could result in a desired mode shift and an alleviation of congestion? How will economic, demographic, or land use changes affect transportation system performance? Will travel demand management strategies or intelligent transportation systems alleviate congestion? Will a new transit investment attract riders? Given a set of desired outcomes, decision makers must identify capital investments and policies that will achieve these objectives.

Travel models are created to support the aforementioned decision makers by providing information about the impacts of different transportation and land use investments and policies, as well as demographic and economic trends. Travel models produce quantitative information about travel demand and transportation system performance that can be used to evaluate alternatives and make informed decisions. A variety of travel models has been used in transportation planning, from simple sketch planning models that produce rough "order of magnitude" information to trip-based travel models that use trips as the unit of analysis. Trip-based travel models, often referred to as 4-step models, have been used for decades to support regional, sub-regional, and project-level transportation analysis and decision making.

For decades, with the revolution of computation techniques, activity-based travel demand models (ABM) have become more widely used in practice. Activity-based models share some similarities to traditional 4-step models: activities are generated, destinations and travel modes are determined, and the specific network facilities or routes used for each trip are assigned. However, activity-based models incorporate some significant advances over 4-step trip-based models, such as the explicit representation of realistic constraints of time and space and the linkages among activities and travel.

However, one of the critical issues of activity-based travel demand model is its expensive data collection process. Activity-based demand modeling requires privacy sensitive disaggregated data related to individual activities and choices. In order to get such data, agencies carry out expensive manual targeted surveys that can only be completed and analyzed with significant delays. For instance, National Household Travel Survey (NHTS) happens only every 5 to 10 years. The cost of the 2001 NHTS was estimated to be approximately 10 million dollar. In the state of California alone, there is a potential $4 million savings for 40000 households at a cost of $100 per 1-2 day sample[59]. The latencies and deficiencies in urban data flow mechanisms create systematic risks for the cities of the future, endangering the very foundations of their functioning. It is critical for public agencies to receive timely and accurate information supporting their everyday decision making practices.

Thanks to the ubiquitous sensor networks and location-based services, people generate data while traveling (and even standing still). Therefore, a widely popular way to approach this data collection problem is through crowd-sourcing [64], which is a pervasive technology-driven way to substitute manual surveying. Navigation services, such as Google Map, use historical data and "crowd-sourcing" to estimate the traffic condition and to deliver real-time routing options. It analyzes the GPS-determined locations and calculates the speed of users along the road [129]. Location-based services and social networks (LBSN), such as Yelp, Twitter, and Facebook, rely on check-ins to keep track of the local businesses and points of interest (POI) you visit and keep your friends updated with your latest comings and goings. Every minute, Facebook users share nearly 2.5 million pieces of content. Twitter users tweet nearly 300K times, and Instagram users post nearly 220K new photos. The aforementioned collective sensing requires the access to GPS or Internet, which becomes unavailable if a user turns off locational services or internet. On the other hand, the pervasive sensing from telecommunications companies such as Verizon, AT&T rely solely on the cellular network. Timestamped locations (at the resolution of cell tower) are recorded whenever calls and short messages are made or data is used.

However, current usage of cellular data for transportation analysis had focused on aggregate level information, for instance, dynamic population estimation, traffic flow and OD estimation, etc. An integrated framework to activity-based travel demand models using large scale mobile phone data to characterize individual movements has not been seen.

1.2 Objective and Challenges

The objective of the dissertation is to improve or build low cost modular components that allow the development of timely activity-based urban mobility models with cellular data, stored in private repositories, such as AT&T.

Activity-based urban mobility models need information about the schedule of each individual and some knowledge about people's decision making process [8]. To be more specific, at the top level, we would like to identify the activity pattern of individuals, e.g. day structures, activity transitions, etc. Under that, a series of decisions about activity, location, travel mode, start time and duration is to be determined.

The challenges include:

- **Privacy and security issues**: Location traces might reveal the time and location of individuals' significant activities[109]. These traces are easy to mine and may cause re-identification of the individual even when the data is anonymized. While building the components, we ensure that no individual and disaggregated data come out of the private repository.

- **Information uncertainty**: Activities (trip purposes) and their contexts are usually reported in manual surveys thus are naturally available for modelers. However, due to the low temporal and spatial resolution of the cellular data, there are many ways uncertainties can be introduced to the data, such as the uncertainties in activity locations due to data triangulation. Extra steps are required to address the uncertainties and extract the information.

- **Population disaggregation**: An activity-based travel demand model is usually modeled and applied for population subgroups. A smaller group size might result in the data insufficiency and the model might overfit to the data. One extreme would be an individual model that is trained on an individual's data. This will not only cause the overfitting but also raise privacy concerns. On the other hand, a model trained at a larger group size might be too coarse to capture the heterogeneous activity patterns among sub-groups of the population. Finding a way to disaggregate the whole population into appropriate sub-populations is important in ensuring good uses of the models.

- **Model validations**: Considering the validation for activity-based travel demand models, Yasmin et, al. has summarized three methods of validating an activity-based travel demand model. The first two validation methods are testing the transportation performance and population behavior against the base-year data and future year data.

These two methods have been practiced by many researchers including Bowman et, al. [21]. The other method is by testing the model's spatial transferablity. That is testing the usefulness of the transferred model, information, or theory in the new context [132]. Yasmin et al. proposed a multi-level validation pipeline to test the model transferability at macro-, meso-, and micro-level. Certain applications requires specific methods of validation. In this dissertation, we validated our mobility models with multiple validation methods mentioned above.

1.3 Dissertation Outline

The dissertation is structured in the following manner:

- Chapter 2 reviews key concepts in the dissertation, including activity-based travel demand models, current practice of using large scale cellular data, and existing human urban mobility models. By reviewing the activity-based travel demand models, we identify the key components of the models such as data sources, modeling frameworks, and applications. We also recognize the inefficiencies in collecting the required data for modeling, which introduces our discussion about using cellular data as an alternative data source to complement the traditional manual survey. By reviewing the current application of large scale cellular data in transportation, we find that most research focuses on exploring the power of large cellular data at aggregate level. Its power in disaggregated activity-based travel models is still to be matured. Urban computing, as an interdisciplinary field, has drawn increasing attention in the recent decade. There has been many models using mobile phone traces to model human mobility. These works can be characterized by their data sources, some using GPS data, cellular data, and some using locational based social network data, check-in data, etc. These work can also be characterized by their applications: some mainly focus on understanding human mobility laws, some recognize daily activity patterns and some predict the timing and location of future activities. We find that most of the studies focus on only one aspect of urban mobility, a fused framework with equally strong power of recognition and prediction is yet to be proposed.

- Chapter 3 summarizes our lessons learned and success gained from processing noisy cellular data, and converting the raw data into activity sequences which are the input for modeling. We follow a common framework of extracting activity locations first by spatial clustering followed by a filter based on dwell time. Our pre-processing method is better in handling positioning error and oscillation error so that we can filter the obvious oscillations without over-filtering short-term travel. Along the pre-processing pipeline, we can gain many research product that was originally available purely from manual survey, such as dynamic population estimation, home and work location distribution, and a good understanding about how people construct their daily activities.

- Chapter 4 introduces our activity-based urban mobility models from cellular data. We present the thinking process of model selections, ground truth collections, step-by-step model improvements and illustrate the model through a case study with the activity sequences of San Francisco regular commuters extracted in previous chapter. We present the model estimation results, visualizations that help understand activity profiles and transitions. We also validate our models with three independent data sources: aggregated statistics from travel surveys, a set of collected ground truth activities, and the results of a traffic micro-simulation informed with the travel plans synthesized from the developed generative model.
- Chapter 5 studies an application scenario of the activity-based urban mobility models in Chapter 4. We propose a medium term (hours to days ahead) travel demand nowcasting problem that fills the gap in the literature, which mainly focuses on either long term (months to years ahead) or very short term travel (seconds to minutes ahead) demand forecast. Long term forecasts aim at urban infrastructure planning and policy evaluation, while short term predictions typically use high-resolution freeway detector/camera data to project traffic conditions in the near future and the main application is for real-time routing and travel time estimation. Our medium term problem addresses questions such as: based on observations of early morning or noon traffic, what will traffic be like during the evening commute? This could be critical in the design of demand-responsive congestion mitigation interventions. And it is a question we could answer with the mobility models we train in Chapter 4.
- Chapter 6 provides a comprehensive summary of the research motivation, objective, methodological frameworks, experimental results, applications, and corresponding findings. This chapter also focuses on identifying future research directions for more comprehensive and unified activity-based travel demand models with cellular data.

1.4 Contributions

This dissertation focuses on complementing activity-based travel demand models using cellular data. This dissertation presents a comprehensive review of the current problems of activity-based travel demand models and human mobility models with cellular data. This dissertation also provides the low cost building blocks that can be used directly in real-world applications. The contribution of the dissertation are six-fold.

- First, we implement an end-to-end processing and inference pipeline from raw cellular data to support travel demand models and traffic simulation tools used by transportation practitioners. The building blocks of the pipeline can be directly applied to any region with data of similar structure.
- Second, we propose a way of preprocessing raw cellular data that is better in handling positioning error and oscillation error so that we can filter the obvious oscillations

without over-filtering short-term travel. This approach can be applied to different data sources with different temporal-spatial resolution by adjusting only a few hyper-parameters.

- Third, to the best of the authors' knowledge, this is the first work using context dependent non-homogeneous generative models of the Input-Output Hidden Markov Model (IO-HMM) architecture to analyze activity patterns from cellular data. We empirically show that our generative model outperforms baseline approaches which ignore contextual information in modeling activity profiles and transitions. In addition, we further explored using semi-supervised co-training to direct the learning process and found that we can have both the generative power of IOHMM and discriminate power from its counter part decision tree model. The directed learning approach leads to better recognition accuracy and location choice modeling. A distributed implementation of the learning and inference methods in a MapReduce framework in pySpark is available at `https://github.com/Mogeng/IO-HMM`. It includes IO-HMM extended with multiple output models such as multinomial logistic regression, generalized linear models, and neural networks.

- Fourth, we propose to validate our models with independent information sources. We annotate secondary activities such as "recreation", "food", "stop in transit" with strong spatial-temporal evidence. We also estimate heterogeneous context-dependent transition probabilities. To validate the model, we compare our annotations to "ground-truth" land-use information of buildings with short range distributed antenna systems, compare the learned activity patterns with travel survey results, and finally compare ground truth traffic counts in the San Francisco Bay Area to a micro-simulation of travel plans derived from the generative model.

- Fifth, we propose an application scenario of the activity-based mobility model we study in Chapter 4. We solve a medium term travel demand forecast system which fills the gap between long term travel demand forecast and short term traffic state prediction.

- Sixth, we explore the predictability of human mobility with parametric sequence learning models as compared to an individualized non-parametric "nearest neighbor" approach. 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.

Chapter 2

Literature Review

2.1 Activity-Based Travel Demand Models

Introduction

Activity-based travel demand model derives travel demand from people's needs and desires to participate in activities [25]. In some cases these activities may occur within their homes, but in many cases these activities are located outside their homes, resulting in the need to travel. Activity-based models are based on behavioral theories about 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. Because they represent decisions and the resulting behavior more realistically, activity-based models are often better at representing how investments, policies, or other changes will affect people's travel behavior.

Activity-based models often provide much more robust capabilities and sensitivities for evaluating scenarios under different policies, because activity-based models typically function at individual level and represent how these persons travel across the entire day.

Another critical advantage of activity-based models is that they produce more detailed performance metrics, such as how travel benefits accrue to different populations, which can be used to support equity analyses. In addition, activity-based models can produce all of the trip-based model measures used to support regional planning, regional air quality, transit, and transportation demand management forecasting [25].

Modeling Framework

To simulate a typical day in an urban area, microsimulation tools need information about the schedule of each individual and some knowledge about people's decision making process.

The activity-based model is mainly composed of three modules.

- **Household and agents**: Agent-based models require agents, preferably grouped into

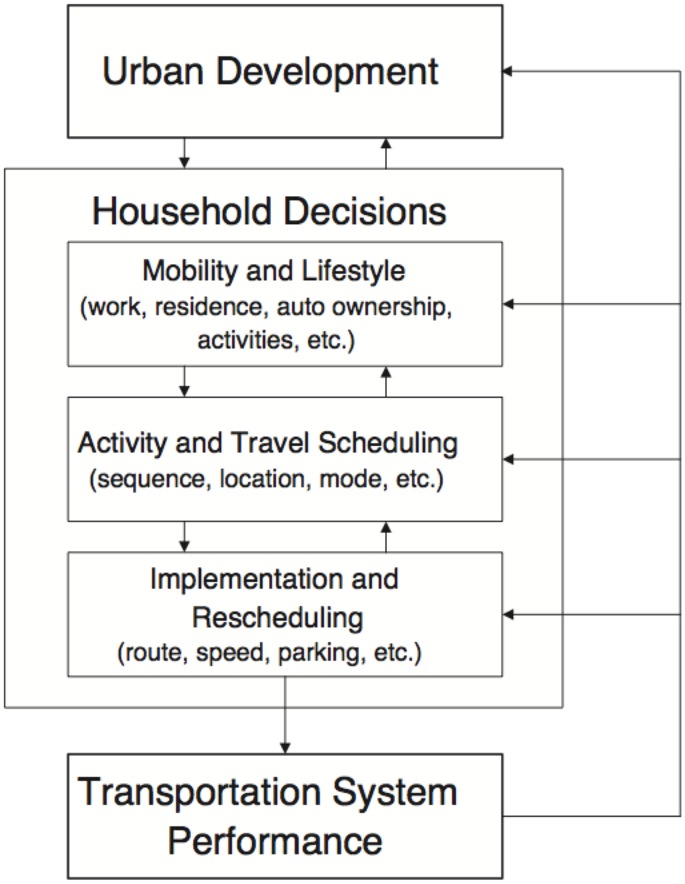

Urban Development
Household Decisions
Mobility and Lifestyle
(work, residence, auto ownership, activities, etc.)
Activity and Travel Scheduling
(sequence, location, mode, etc.)
Implementation and Rescheduling
(route, speed, parking, etc.)
Transportation System Performance

Figure 2.1: Activity-based model framework

households, and even better grouped into social networks. A large number of systems employ iterative proportional fitting to draw agents from the fitted multidimensional table in order to get their social demographics and activity patterns [9].

- **Activity and scheduling**: The modeling of the schedule is the central task of an activity-based modeling approach, realizing its vision of human behaviour as a coherent (daily) whole[9]. The primary question to be answered are:
  - Activity pattern
  - Sequence of primary tour including destination, mode, route choice, starting time, duration, and accompany
  - Sequence of secondary tour including destination, mode, route choice, starting time, duration and accompany

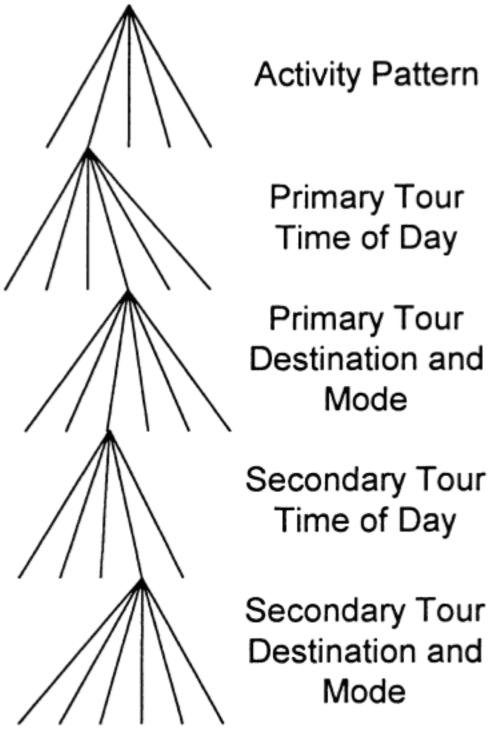

Activity Pattern
Primary Tour Time of Day
Primary Tour Destination and Mode
Secondary Tour Time of Day
Secondary Tour Destination and Mode

Figure 2.2: Utility maximizing agent-based modeling

Three different modeling approaches have been attempted to the development of activity-based models of travel demand:

– **Constraints-based models**: The primary purpose of constraints-based models is to check whether any given activity agenda is feasible in a specific spacetime context. Inputs to these models are activity programs, which describe a set of activities of certain duration that can be performed at certain times. A combinatorial algorithm is typically used to generate all possible activity sequences[5].

– **Utility-maximizing models**: These models extended the complexity of discrete choice models, in particular, the nested logit model [20]. This nested logit model is consist of five nests: 1) activity pattern, representing a choice of a pattern with and one without travel, plus a system of conditional tours defined by four tiers: 2) primary tour time of day 3) primary destination and mode 4) secondary tour time of day, and 5) secondary tour destination and mode. Model parameters were estimated simultaneously within each hierarchy and sequentially across hierarchies.

– **Rule-based models**: Rule-based models are used to depict decision heuristics, which relaxes the strict and behaviourally unrealistic assumption of utility-maximizing models. Individuals and households are assumed to conduct activities to attain certain goals. Certain rules (either learned from models or empirical rules) drive the choice of activity participation, jointly with prior commitments and constraints[100].

- **Traffic simulation and rescheduling**: Because of the complex and dynamic nature of the activity-based demand model, especially when we tend to generate the activity chain for an individual over the entire day, microsimulation as a method for implementing activity-based travel behavior models for forecasting and policy analysis purposes has received ever-increasing attention [93]. Given the system's complexity, closed-form analytical representations of the system are generally not possible, in which case numerical, computer-based algorithms are the only feasible method for generating estimates of future system states [92]. Microsimulation represents an effective method for generating policy-sensitive forecasts from disaggregate, activity-based models. The performance of an agent's plan is scored at the end of each iteration of the microsimulation, until a steady-state approximating a dynamic Nash-equilibrium is reached. For a predetermined share of the agents, new plans are generated by searching for new shortest-path or by optimizing the starting times and duration. The scoring function is mainly utility based [9].

Application and Future Trend

Activity-based models have become more widely used in practice. The domain of traffic and transportation systems is well suited for an activity-based approach because transportation systems are usually geographically distributed in dynamic changing environments. Techniques and methods resulting from the field of activity-based models have been applied to many aspects of traffic and transportation systems, including modeling and simulation, dynamic routing, congestion management, and intelligent traffic control[26].

However, several problems with the current practice of activity-based models need to be emphasized. First, most of the data comes from travel survey or travel diary that only includes activity patterns of one typical day. The "day of week" and long term patterns may not be discovered with current data scheme. This data collection process is also expensive and with significant delays. Second, individual decision-making may depends on the social network she belongs to. Current models that do incorporate family and friend decision-making are based on relatively simple extensions of models of individual choice behaviour and have been descriptive and analytical. With the ubiquity of smart phones, developing a conceptual framework, using these alternative data, to assist comprehensive activity-based models is necessary.

2.2 Current Practice of Using Large Scale Cellular Data

The increasing availability of the large scale cellular data has enabled transportation research at many levels.

- **Aggregated human mobility**: Call detailed record (CDR) data, although relatively low in spatial-temporal resolution, allows the study of aggregated human mobility patterns. For instance, Gonzalez et, al. studied the distribution of travel distance and simple reproducible patterns using CDR of 100K individuals over six months in a European data set [55]. Song et, al. discovered a 93% potential predictability in user mobility across the whole user base [112]. Kung et, al. showed that the home-work time distributions and average values within a single region are indeed largely independent of commute distance or country [71].

- **Dynamic population**: Deville et, al. estimated the population in France and Portugal using 5 months' and 10 months' phone call data [31]. They mapped the night users of each cell tower to the administrative unit proportional to the overlapping area between the Voronoi tessellation of the cell tower and the administrative unit. A log linear regression was then used to scale the mobile-phone based population to census-derived population. They received a R-value of 0.9. Their focus is on the dynamics of population rather than finding the population residential location. A similar work was done for the Ivory coast of Africa. Sterly et, al. used the call data of 500K callers collected by Orange Telecom over 14 days to estimate the population [116]. They simply assigned the callers to the administrative unit from where they placed the highest number of calls. Not surprisingly, their result leads to higher population in urban areas and lower ones in rural areas since many people tend to make more phone calls at work than home. Ahas et, al. developed a more rigorous but computational expensive model to estimate the home, work anchor points of 0.5 million users with 12 months' data collected by a major operator in Estonia [2]. They reached R-value of 0.99 between the number of modeled homes and the number of residents in the population register in Estonia's 227 municipalities. But in some major cities the R-value was lower, at 0.86.

- **Traffic and OD estimation**: Considering traffic analysis, two major approaches have been explored:

  - An OD matrix was first estimated from CDR [22, 91, 126, 62, 121]. A rescaling was performed to match the derived OD with total traffic count. Traffic flow on the road network was then assigned using iterative proportional fitting (ITA) [126, 121] or microsimulators [62]. This method was exactly the four-step model with the OD estimated from real CDR as an alternative to the traditional trip generator.

  - A direct mapping from cell tower to the road network [28, 110, 76, 131].

    It is worth mentioning that while the second approach has been extensively explored using GPS data (HMM-based model in [120], conditional random field model in [61, 60]), much complexity will be introduced when applied to CDR data which has low spatial-temporal resolution .

- **Land use and Urban planning**: CDR data can also assist urban planning through analyzing land use. Reades et. al. analyzed the spatial-temporal patterns in the city of Rome from aggregate mobile phone usage data collected over the course of three months in late 2006 and covered a region of 47 km^2 [101]. A classification algorithm was used to identify clusters of locations with similar zoned uses and mobile phone activity patterns from three weeks of CDR data for roughly 600K users in the Boston region.

- **Disaster response**: Bengtsson et, al. estimated the geographic distribution of population movements after the devastating Haiti earthquake in 2010 and found that the distribution corresponded well with results from a population-based survey [16]. They showed feasibility of rapid estimates and identification of areas at potentially increased risk of cholera outbreak within 12 hours of receiving data. Interestingly, the predictability of people's trajectories remained high and even increased slightly during the three-month period after the earthquake. Lu et, al. found that the duration of people's stays outside the city, as well as the time for their return, all followed a skewed, fat-tailed distribution [85]. These studies suggested that CDR data may be of great value in predicting population movement as a response to big disasters.

- **Disease spread**: Mobile phone data could provide valuable, complimentary and contemporary data on an ongoing basis in infectious disease control and elimination [118]. Wesolowski analyzed the regional travel patterns of nearly 15 million individuals over the course of a year in Kenya with mobile phone data. Combined with malaria prevalence information, they identified the dynamics of human carriers that drive parasite importation between regions. They also identified important routes that contribute to malaria epidemiology on regional spatial scales. The analysis of human movement patterns from Zanzibar to mainland Tanzania suggested a few people account for most of the risk for imported malaria [119].

- **Special social events**: CDR data also provides a special opportunity to characterize traffic flows generated by special social events. By analyzing about 1 million mobile traces, Calabrese et, al. concluded that people who live close to an event are preferentially attracted by it and events of the same type show similar spatial distribution of origins [23]. This study showed that CDR data could potentially be used to predict where people would come from for future events and take decisions about events management and congestion mitigation.

- **Inferring social network and demographics**: The underlying social network from cellular data has made a deep understanding of social interaction possible. Eagle et, al. found that self-reported friendship deviate from mobile phone records depending on the recency and salience of the interactions [38]. They accurately inferred 95% of friendships based on the observational data alone. Dong et, al. employed a conditional random field (CRF) model to jointly classify age and gender of users based on their calling profile [34]. Furletti et, al. used temporal calling patterns to identify four categories of users: residents, commuters, in transit and tourists/visitors using around 7.8 million CDR records collected in the city of Pisa, Italy, from January to February 2012 [48].

However, most of the aforementioned applications focused on exploring the power of cellular data at aggregate level. Models that allow analyzing CDR data at individual level and assisting activity-based demand modeling have yet to be developed

2.3 Urban Mobility Models

Urban mobility models study many aspects of individual travel. This section summarizes related works on state-of-the-art urban mobility models. We will organize our discussion of the literature by considering the data sources, modeling techniques, and modeling objectives.

Author(s) and Date	Data	Modeling	Prediction	Method
Gonzalez et al. (2008) [55]	CDR	Human mobility laws	NA	Statistical
Song et al. (2010) [112]	CDR	Human mobility laws	NA	Statistical
Eagle and Pendland (2009) [37]	CDR	daily activity pattern (primary)	Rest of day	PCA
Song et al. (2004) [113]	Wi-Fi	Location	Next location	Markov models
Akoush and Sameh (2007) [3]	Wi-Fi	Location	Next location	NN
Farrahi and Gatica-Perez (2011) [43]	CDR	daily activity pattern (primary)	NA	Topic models
Liao et al. (2006) [80, 78]	GPS	daily activity pattern	NA	CRF
Eagle et al. (2009) [36]	CDR	Location	Next Location	DBN
Gao et al. (2012) [50]	GPS	Location	Next Location	Markov models with context
Cho et al. (2011) [29]	LBSN	Location	Loctaion	Topic models
Ashbrook et al. (2003) [7]	GPS	Location	Next location	Markov models
Ye et al. (2013) [133]	LBSN	Activity and location	Next activity and location	HMM with context
Laasonen (2005) [72]	CDR	location	Next location	Trajectory matching
Do and Gatica-Perez (2014) [33]	CDR	location	Next location	Random forest
Zheng et al. (2008) [143]	GPS	Travel mode	NA	Decision tree
Sohn et al. (2006) [111]	CDR	Travel mode	NA	Boosted logistic regression
Phithakkitnukoon et al. (2010) [98]	CDR	Activity	NA	Rule based
Mathew et al. (2012) [90]	GPS	Location	Next location	HMM
Ying et al. (2011) [136]	GPS	Activity and Location	Next location	Trajectory matching
Lee et al. (2009) [75]	GPS	Human mobility laws	NA	Statistical
Zheng et al. (2012) [139]	GPS	daily activity pattern (primary)	Location	Topic models
Monreale et al. (2009) [94]	GPS	Location	Next location	Trajectory matching
Scellato et al. (2011) [105]	GPS	Duration	Duration	KNN
Thiagarajan et al. (2009) [120]	GPS	Map matching	NA	HMM
Bauman et al. (2013) [12]	GPS	Location	Next location	Markov models with context
Chon et al. (2012) [30]	GPS	Duration	Duration	Markov models with context
Krumm et al. (2006) [70]	GPS	Trajectory	Destination	Trajectory matching
Schneider et al. (2013) [106]	CDR	Daily activity pattern (primary)	NA	Statistical
Etter et al. (2013) [39]	GPS	Loctaion	Next location	DBN and NN and GBDT
Gambs and Killijian (2012) [49]	GPS	Loctaion	Next location	Markov models
Asahara et al. (2011) [6]	GPS	Loctaion	Next location	Mixed Markov models
Lu et al. (2012) [87]	GPS	Loctaion	Next location	Ensemble
Gomes et al. (2013) [53]	GPS	Loctaion	Next location	Multiple
Jeung et al. (2008) [65]	GPS	Trajectory	Next location	Trajectory matching
Gidofalvi and Dong (2012) [51]	GPS	Location and duration	Duration and next location	Semi Markov models
Baratchi et al. (2014) [10]	GPS	Daily activity patterns, location, duration	NA	Hierarchical HSMM
Bhat and Singh (2000) [17]	Travel survey	Activity and travel scheduling	daily travel plan	Discrete choice
Bowman and Ben-Akiva (2001) [20]	Travel survey	Activity and travel scheduling	daily travel plan	Discrete choice
Widhalm et al. (2015) [128]	CDR	Daily activity patterns	NA	DBN
Calabrese et al. (2013) [24]	CDR	Daily trip length	NA	Statistical
Liu et al. (2013) [83]	GPS	Activity	NA	Ensemble
Bohte and Maat (2009) [19]	GPS	Activity and travel mode	NA	Rule based
Stopher et al. (2008) [117]	GPS	Activity and travel mode	NA	Rule based
Wolf et al. (2001) [130]	GPS	Activity and travel mode	NA	Rule based
Kim et al. (2014) [69]	GPS	Activity	NA	Decision tree
Stenneth et al. (2011) [115]	GPS	Travel mode	NA	Random forest
Reddy et al. (2010) [102]	GPS	Travel mode	NA	DT + DHMM
Doyle et al. (2011) [35]	CDR	Travel mode	NA	Rule based
Zheng et al. (2008) [141]	GPS	Travel mode	NA	Decision tree
Wang et al. (2010) [124]	CDR	Travel mode	NA	K-means
Chen and Bierlaire. (2015) [27]	GPS	Map matching and travel mode	NA	HMM
Leontiadis et al. [76]	CDR	Map matching	NA	A*
Anderson and Muller (2006) [4]	CDR	Travel mode	NA	Clustering + HMM
Widhalm et al. (2012) [127]	GPS	Travel mode	NA	Ensemble + HMM
Gong et al. (2012) [54]	GPS	Travel mode	NA	Rule based
Schuessler and Axhausen [108]	GPS	Travel mode	NA	Rule based
Ben-Akiva and Lerman (1985) [13]	Travel survey	Travel mode	Travel mode	Discrete choice
Song et al. (2016) [114]	GPS	Trajectory	Trajectory	LSTM
Yin et al. (2017) [135]	CDR	Activity patterns	Activity sequences	IOHMM
Lin et al. (2017) [81]	CDR	Trajectory	Trajectory	LSTM

Table 2.1: Literature for urban mobility modeling and prediction

Data Sources

Early studies mainly used travel surveys [17, 20, 13]. In the recent decade, with the mobile phone data more available, passively collected data such as GPS [80, 78, 7, 143, 136, 75, 139, 94, 105, 120, 70, 10, 19, 69, 102, 141, 27, 54, 114], CDR (call detailed record) [55, 112, 37, 43, 36, 33, 98, 106, 128, 24, 135, 81] and location-based social networks (LBSN) data [29, 133] 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.

Modeling Techniques

There are two main streams of modeling techniques, one targeting at individualized applications such as mobility prediction and the other targeting at population applications such as clustering human daily activity patterns.

Individualized Models

- **Markov type models**: Simple Markov type models assumes that the current location depends on the previous location and some other contextual information. This type of model also includes the models considering contextual conditional probabilities, for instance, the work using GPS data [50, 7, 32, 70, 49, 6], and CDR data [113, 86, 84, 72]. However, this type of model is only interested in modeling next location but not when the next activity happens thus does not consider the duration at the same time. To overcome this limitation, a non-homogeneous semi Markov model was used to model the activity chain using a travel survey [89]. However, If observations are not accurate, simple Markov type models may be too naive to capture the error structure.

- **Trajectory matching models**: Trajectory matching models are similar to Markov type models, but not restricted to observations and context on the most recent observation. Another major difference between trajectory matching models and simple Markov type models is that Markov models are generative, while trajectory matching models are usually discriminative and are mainly used to make predictions. Because of its discriminative nature, it would not be a problem to predict location and start time of the next activity using conditional probabilities. Examples using GPS data include [140, 140, 136, 94, 70], and examples using CDR data includes [72, 137, 68, 95, 125, 65, 88, 65, 77]. However, since these models are discriminative and focus directly on the locations, no activity recognition is performed through these models.

- **HMM and HSMM models**: To account for data noise, a hidden layer is added to the Markov models. These include HMM model using CDR data [99], using GPS data [90, 58], and LBSN data [133]. These models are really generative models and respect the transition between activities. However, only Ye et, al.[133] incorporated contextual information but their focus is only predicting the category of the activity instead of locations.

- **DBN and CRF models**: To incorporate contextual information, complex graphical models have been applied to model mobility. As an example of a generative model, a dynamic Bayesian network (DBN) was adopted to detect abnormal or normal behavior [36]. On the other hand, Liao et, al. developed a discriminative version of the previous model [78, 79]. They used hierarchical conditional random field (HCRF) to extract places and activities. However, this approach needed ground truth obtained from manual labeling. Thus their model was only applied to four people and was not scalable to large population.

- **Classification models**: This class of models are pure machine learning models that focus on the predictions rather than interpreting the structure of daily routines. For instance, regression models [33], neural networks [82, 3, 74, 96], K-nearest neighborhood [12], decision tree [39, 53, 122], SVM [53], and ensembling [87] have been used to make predictions about next activity location or timing. It is worth noting that most of these models came from the Nokia Mobile Data Challenge. The data included rich context information including date, location of the user, cell tower id, phone calls and application usage collected from the smart phones of 80 users. The power of these models is bound to be reduced if not so much contextual information is available.

Group Models

- **Motif Models**: These models relate daily mobility patterns with trip chains extracted from travel diary surveys or mobile phone data [66, 106]. The authors expressed daily activity chains as daily networks with nodes representing locations and directed edges representing trips. The same distribution of trip configurations have been found in different cities, and measured by both travel surveys and mobile phone data. They found that only 17 unique networks are sufficient to capture the daily mobility pattern of 90% of the population in surveys and mobile phone datasets for different countries. The authors found that although most of the people visit less than five locations, a small fraction behave significantly differently because people report visits up to 17 different places within a day in their surveys. However, focusing on activity transitions, the authors discarded information about the purpose of the activity, the travel time and the activity duration as well as the distances and the number of trips between the visited locations.

- **Topic Models**: These models tend to use latent topics to identify structure in human daily routines. Popular Topic models include LDA models [43, 42, 44, 41, 139] and "eigenbehaviors" [37]. Eagle et, al. using "eigenbehavior" decomposition found that communities within a population's social network tend to be clustered within the same behavior space. Therefore, if strong behavioral homophily is present in the data, it should equally be possible to infer an individual's affiliations by quantifying the individual's distance from a community's behavior space. On the other hand, Farrahi et, al. found that most of the routines (topics) were quite interpretable, including "going to work late", "going home early", "working non-stop" and "having no reception" (phone off)" at different times over varying time-intervals. The advantages of topic model is that these are generative models and consider the periodic nature of human routines. However, the construction of daily routines need semantic meaning of each location. This limits the daily routine to include only home, work and "Other".

- **Nested Logit Discrete Choice Model**: Bowman et, al. modeled a hierarchical nested logit choice model to discover the activity pattern [20]. Under the level of activity pattern, they modeled the schedule of primary tour and then secondary tour. Tour models included the choice of time of day, destination and mode of travel, and were conditioned by the choice of activity pattern. Their framework is shown in Figure 2.2. The model was designed to capture individual's decisions throughout an entire day by explicitly representing tours and their interrelationships in an activity pattern. The model was targeted at assisting activity-based demand models. These features gave the model potential to improve travel forecasts by capturing activity-based policy responses. However, in their methodology, the activity pattern was pre-defined by 54 types, but not discovered from the data.

- **Hierarchical Hidden Semi Markov Models (HHSMM)**: Baratchi et, al. proposed a hierarchical hidden semi-Markov-based model which could capture both frequent and rare mobility patterns in the movement of mobile objects [10]. In the top layer, the authors used a super-state to indicate the hidden mobility pattern. Under the pattern layer, the activity chain was modeled as a hidden semi-Markov model again. In this case, they modeled the activity pattern transition and the activity transition simultaneously using multiple days' data. Their model outperformed other baseline models including standard HSMM in terms of next place prediction accuracy. However, a problem of their HHSMM model was that their HSMM does not depend on the contextual information such as time of day.

Modeling Objectives

Considering modeling objectives, large amount of works focus on activities, such as activity locations, [113, 7, 33, 94], start times and durations [105], and daily/weekly activity scheduling [37, 43, 78, 139, 106, 10, 17, 20, 128, 135]. Another branch of research considers trips linking these activities, studying trajectories [70, 114, 81], travel mode [143, 19, 102, 141, 54, 13], by applying map matching and route choice [120, 27].

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 [55, 112, 75, 24], the other category focuses on the problem of recognition (activity, travel mode, [80, 98, 120, 106, 19, 69, 27]) 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 [113, 7, 133, 36]; other researchers treat prediction of next location as a classification (regression) problem using supervised learning [33]; and some researchers used trajectory matching techniques to make the prediction [136, 94]. However, not much research has been done on models that are capable in generating a sequence of locations (duration) 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 focus 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 [37], Farrahi and Gatica-Perez [43], and Zheng et al. [139] 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 [78]. However, their model is discriminative in nature and is most suitable for recognition, rather than generating new sequences. Widhalm et al. [128] 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. In this dissertation, we improve the modeling of activity patterns (spatial-temporal profiles of primary and secondary activity) with explicit modeling of contextual dependent activity transition probabilities.

Chapter 3

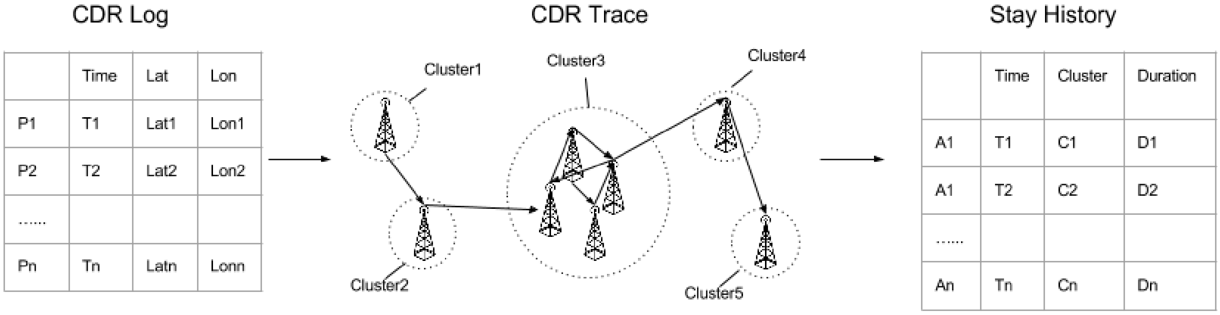
From Cellular Data to Urban Activities

3.1 Introduction

Cellular data does not give information about activities directly. Raw CDR data contains a timestamped record for each communication of anonymous users' devices served by the cellular network. Due to positioning errors and connection oscillations, it is not straightforward to extract features to model urban mobility 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. 3.1. 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, 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 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. 3.3.

[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.



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

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

Figure 3.1: 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.

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) [17].

3.2 Processing Pipeline

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. 3.1. Each record of raw CDR logs (such as a phone call, short message, or data usage) 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 [40]. The distance thresholds for GPS stay-location clustering is much smaller than the thresholds for CDR

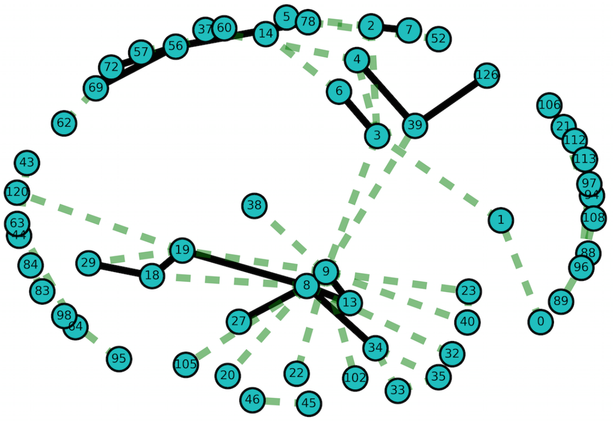


Figure 3.2: Sample oscillation graph. Each node in the graph represent a location cluster. Edges in the oscillation graph connect clusters that are suspicious for oscillations. The thicker the edge,the more oscillations have been observed.

records. For example, a roaming distance of 300 meters [66] and 1000 meters [128] 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 (with the same timestamp) 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 3.2. 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 oscillation 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. Note that though the thickness of the edges, as a hyperparameter for the pre-processing, does not matter in our method, it might be fine-tuned for the processing of other types of data.

(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 [142], 15 minutes in [128] and 10 minutes in [66]. Several GPS applications use stay durations ranging from 90 seconds to 10 minutes. We chose 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.

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 [29]. 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 [71]. 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 [71]. 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 [71]. Borrowing from [66], 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 [63].

3.3 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.

3.4 Experimental 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.

Home/Work Inference results

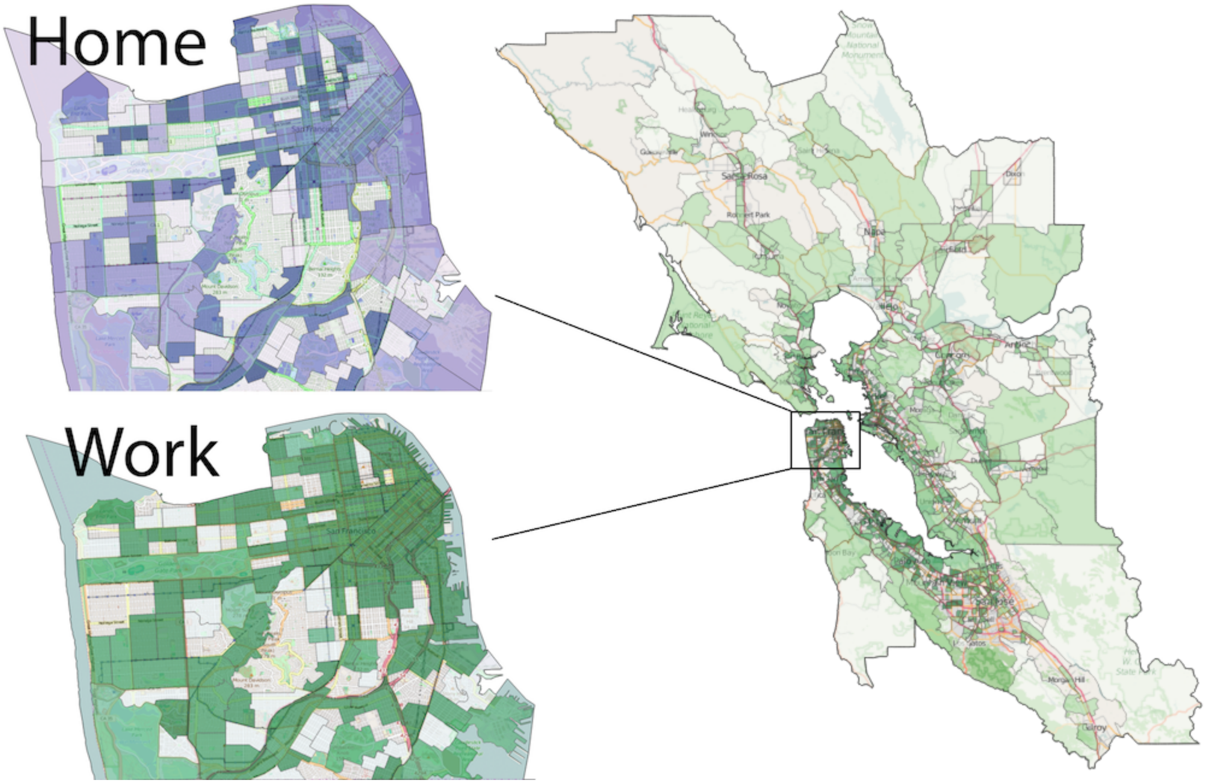
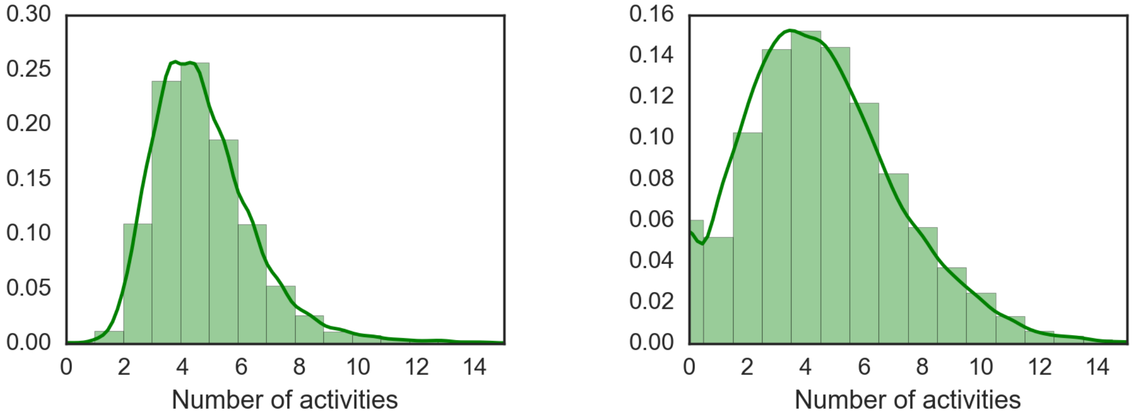


Figure 3.3: Density map of inferred home and work locations for San Francisco residents, aggregated at the census tract level (left), and an overall geographical scope of analysis with work locations density (right).

Fig. 3.3 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. 3.3, 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 [67]. As expected, the home locations are more spread out throughout the city.

Number of Daily Activities



(a) Weekday

(b) Weekend

Figure 3.4: 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. 3.4. 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].

Summary Statistics For Day Skeleton

Table 3.1: Summary Statistics for Day Skeleton

	Weekday	Weekend
H	9.5%	72.3%
HWH	88.1%	26.6%
HWHWH	2.4%	1.1%

For San Francisco regular commuters, Table 3.1 shows that on average 88.1% of them visit their work place once on a typical weekday, featuring a Home-Work-Home (HWH) day skeleton. Note that going out for a lunch from work place and returning to the work place is considered as HWH day skeleton because no home activities happen between the two work activities. 2.4% of the regular commuters have some home activities between two visits to their work place, featuring a HWHWH day skeleton. 9.5% of regular commuters do not go to work on a typical weekday, featuring a home-based day skeleton.

On a typical weekend, 72.3% of the regular commuters do not go to work at all. 26.6% of the commuters visit their work place once and 1.1% of the commuters have some home activities between two visits to their work place. These numbers are similar to the ones in the 2015 American Time Use Survey conducted by the Bureau of Labor Statistics [97].

Distribution of Tours for HWH Days

Table 3.2: Distribution of Tours for HWH Days

	Before-morning-commute	Work-based	Post-home
0	93.8%	72.5%	77.9%
1	5.6%	21.8%	19.7%
2	0.5%	4.3%	2.1%
3+	0.1%	1.4%	0.3%

Consider a Home-Work-Home day for a regular commuter, a user might have some home-based tours before-morning-commute (e.g. morning workout tour), a home-to-work commute tour, some work-based tours (e.g. lunch tour) and a work-to-home commute tour and some home-based tours after coming back from work (e.g. recreation tour).

Table 3.2 shows that 93.8% of the regular commuters go to work directly without any before-morning-commute tours. 6.2% of commuters have one or more tours before-morning-commute. 72.5% of people do not have any tours during work, which means they might have lunch at their work place. 77.9% of people do not have any post-home activities.

Bhat summarized a similar table based on 1990 Bay Area Household Travel Survey by the Metropolitan Transportation Commission (MTC) [17]. He showed that 96.9% of the people did not have any before-morning-commute tours, 74% of the people did not have any work-based tours and 79.7% of the people did not have any post-home tours. We can see that the proportions of no tours in 1990 Bay Area Household Travel Survey are consistently slightly higher than our numbers. There might be two reasons. First, people's behavior pattern might have shifted a little over 25 years. Second, users might tend to under-report their activities and tours in travel surveys for privacy concerns. It is also worth noting that the statistics on day structures are summarized based on the home and work locations identified with the criterion we proposed. We might miss a certain proportion of workers who do not have regular home or work hours, such as people who work at late night. This might cause a little deviation when comparing our results with Bhat's.

Distribution of Activities for HWH Days

Table 3.3: Distribution of Activities for HWH Days

	Before-morning-commute	Home-work commute	Work-based	Work-home commute	Post-home
0	93.8%	58.5%	72.5%	42.4%	77.9%
1	3.9%	27.9%	16.1%	30.5%	12.3%
2	1.3%	9.2%	6.3%	15.6%	5.7%
3	0.5%	2.7%	2.5%	6.6%	2.3%
4+	0.5%	1.7%	2.6%	4.9%	1.8%

Activities can happen during tours. Table 3.3 summarizes the distribution of number of activities during each type of tour for San Francisco regular commuters.

We can see that 41.5% of people make at least one stop during home-work commute tour, and 57.6% of people make at least one stop during work-home commute. This means more people choose to participate in activities after work rather than before going to work. The trend is also captured in the report by Bhat. However, in his report, the percentage of people who participate in activities during home-work commute is 14.8% and the percentage of people who participate in activities during work-home commute is 26%. These numbers are lower than our numbers. Again we suspect people tend to under-report their activities in manual surveys.

Interactions in Stop-Making Across Different Times of Day

Table 3.4: Interactions in Stop-Making Across Different Times of Day

Control Variable	Value	Percentage of individuals having an activity during		
		Mid-day	Evening commute	Post-home
Had a mid-day activity?	Yes	-	58.4	21.6
	No	-	57.3	22.3
Had an evening commute activity?	Yes	27.9	-	19.9
	No	27.0	-	25.1
Had a post-home activity?	Yes	26.9	51.8	-
	No	27.7	59.2	-

To understand how people construct their daily life, it also helps to summarize the interactions in activity participation across different times of day. From Table 3.4, we can see that there is little interactions between having a mid-day activity and having a work-home commute/post-home activity. However, people having an evening commute activity are less likely to have a post-home activity and vice versa, people who have a post-home activity are less likely to have an evening commute activity. This is also observed in Bhat's report [17]. They also found the reason for the interaction is that about half of the mid-day activities are for eating purposes and the other half for work-related businesses. On the contrary, around half of the evening commute activities and more than half of the post-home activities are for social-recreational or shopping purposes. There is a substantial substitution in activity-participation between the evening commute and post-home periods. In contrast, this substitution effect is minor between mid-day activities and evening commute/post-home activities.

Also we found that for people who attend at least one evening commute or post-home activities, they tend to leave their work place earlier, on average at around 4:30 pm. On the other hand, commuters who do not have any evening commute/post-home activities, leave work later, on average at about 6:10 pm.

3.5 Conclusion

In this chapter, we followed a common framework of extracting activity locations first by spatial clustering followed by a filter based on dwell time. Our preprocessing method is better in handling positioning errors and oscillation errors so that we can filter the obvious oscillations without over-filtering short-term travel. With the activity sequences identified, we can have a primary understanding about how people construct their daily activities. We found similar patterns to the findings in Bhat's report, which is based on 1990 Bay Area travel survey [17]. However, we found evidence that people may under-report their daily activities in manual surveys due to privacy concerns. This makes cellular data a better source for understanding the true activity patterns.

Chapter 4

Activity-Based Urban Mobility Models from Cellular Data

4.1 Introduction

As we have reviewed in Chapter 2.3, most urban mobility models from cellular data focus on a single aspect of urban mobility, such as activity location, duration or travel mode between two activity locations. Trip purposes (or activity types) are not the focus in these urban mobility studies. However, trip purpose is an important aspect in activity-based travel demand models, as we have reviewed in Chapter 2.1. Typical activity-based travel models used by practitioners are incredibly rich in describing the intricacies of human activities and context of decision making in travel-related choices. For years, trip purpose is included in discrete choice models of travel mode as context information [14]. It is a significant factor influencing decisions on mode and other attributes of travel.

Therefore, one key research challenge lies in developing urban mobility models with trip purposes ("home", "work", "dining", "shopping", "recreation", etc.) recognized from noisy locational data, such as anonymized mobile phone traces registered via cellular network, with a level of activity-chain detail that is comparable in richness to that of a specifically designed travel survey.

As we have reviewed main related works on urban mobility models, a summary of relevant developments in activity-based urban mobility models is given below with respect to the main methods and approaches.

Supervised models: Considering activity recognition, supervised learning methods require data with labeled ground truth. The ground truth is either manually labeled [39, 53], or collected for a small group of participants from a survey accompanying GPS data [69]. Liu et al. classified activities into "home", "work/school", "non-work obligatory", "social visit" and "leisure" using different supervised learning models including SVM and decision trees. Their data was collected from natural mobile phone communication patterns of 80 users over a year with labeled ground truth [83]. Liao et al. manually labeled ground truth to extract places and activities [78, 79]. However, this model was only applied to four people and is not scalable to large populations.

Unsupervised models: On the other hand, unsupervised models are used to cluster activities with similar temporal and spatial profiles. “Eigenbehavior” models by Eagle et al. [37] and previously mentioned LDA and ATM models by [43, 41] all fall into this category.

Discriminative models: Discriminative state-space models such as CRFs [78, 79] are more flexible when modeling the relationship between input, output and state variables. However, due to their undirected nature, discriminative state-space models cannot be used for activity generation directly.

Generative models: Hidden (semi-) Markov models are generative models that can not only be used to analyze activity patterns, but also to generate new sequences [58]. Using GPS data, Baratchi et al. developed a hierarchical hidden semi-Markov-based model that captures both frequent and rare mobility patterns in the movement of mobile objects [10].

The study of most direct relevance to our work is by Widhalm et, al. [128]. They used similar temporal-spatial features to infer urban activities with an undirected relational Markov network. However, one major drawback of their model is the lack of cliques for consecutive activities, i.e., the study did not model activity transitions. This is unfavorable for activity inference and new sample generation. Sampling consecutive activities independently without considering the dependencies of following activities to previous activities is not appropriate. To overcome this drawback, we explicitly model contextual dependent activity transition probabilities to improve the accuracy of activity inference and the reliability of new activity chain generation, as detailed in Chapter 4.4.

In this chapter, we develop generative activity-based urban mobility models from cellular data with user activities recognized along the model development phase. These urban mobility models reveal temporal activity profiles and the pattern of transitions between activities. We explore different ways of improving the urban mobility models, such as using input output hidden Markov model (IOHMM) instead of standard HMM to incorporate context variables in transition and emission models. We also explored using semi-supervised co-training to direct the learning process so that we can have the generative power of IOHMM and the discriminative model of its counterpart decision tree model at the same time. Validations of models using CDR data are usually difficult due to its low spatial resolution. In addition to validation through comparing aggregated statistics with travel survey by Widhalm et, al.[128], we provide a direct validation on activity recognition using a set of “ground truth” activities based on short range antennas. To validate the urban mobility model and to show its capability of generating realistic activity chains, we use the model to generate synthetic travel plans of individuals with home and work locations sampled from census data. We show that the generated activity chains are realistic and are consistent with the distribution reported in the travel surveys. The synthetic travel plans are used as inputs to an agent-based microscopic traffic simulator. We validate the resulting traffic volumes against an independent dataset of traffic counts collected on all the major freeways within the region of study.

4.2 Modeling Framework

For the activity-based urban mobility model, not only are we interested in understanding the activity patterns themselves, we also aim to model these patterns in a generative probabilistic framework suitable for generating inputs to activity-based travel micro-simulations. Thus, we require generative models. At the same time, privacy considerations and limited availability of ground truth location data preclude us from using fully discriminative supervised approaches, suggesting the choice of unsupervised and semi-supervised models. In order to produce activity patterns for large populations of users, we build models that can leverage distributed implementation and that can share parameters across multiple user groups. These objectives led us to an IOHMM approach with modular heterogeneous transitions/emissions components with interpretable parameters, as detailed in Chapter 4.4.

The developed data processing and modeling pipeline is presented in Fig. 4.1. The left column shows the primary data sources. This includes the cellular call detail data (CDR), a comprehensive point of interest (POI) database within the region of interest, and the traffic data (vehicle counts, volumes) to calibrate and validate the microscopic traffic simulation. POI databases are usually available from open source maps such as OpenStreetMap, or comercial APIs such as Google Places API and Factual Places API. These POI databases provide a list of POIs and their category labels around a location upon query. These POI information is useful in constructing the labeled activities as “ground truth”. The middle column contains the key modules to develop the urban mobility models and the right column shows the resulting products. Our key contribution is the activity recognition and generation module outlined with the red dashed rectangle, and in particular the components shown in shaded yellow.

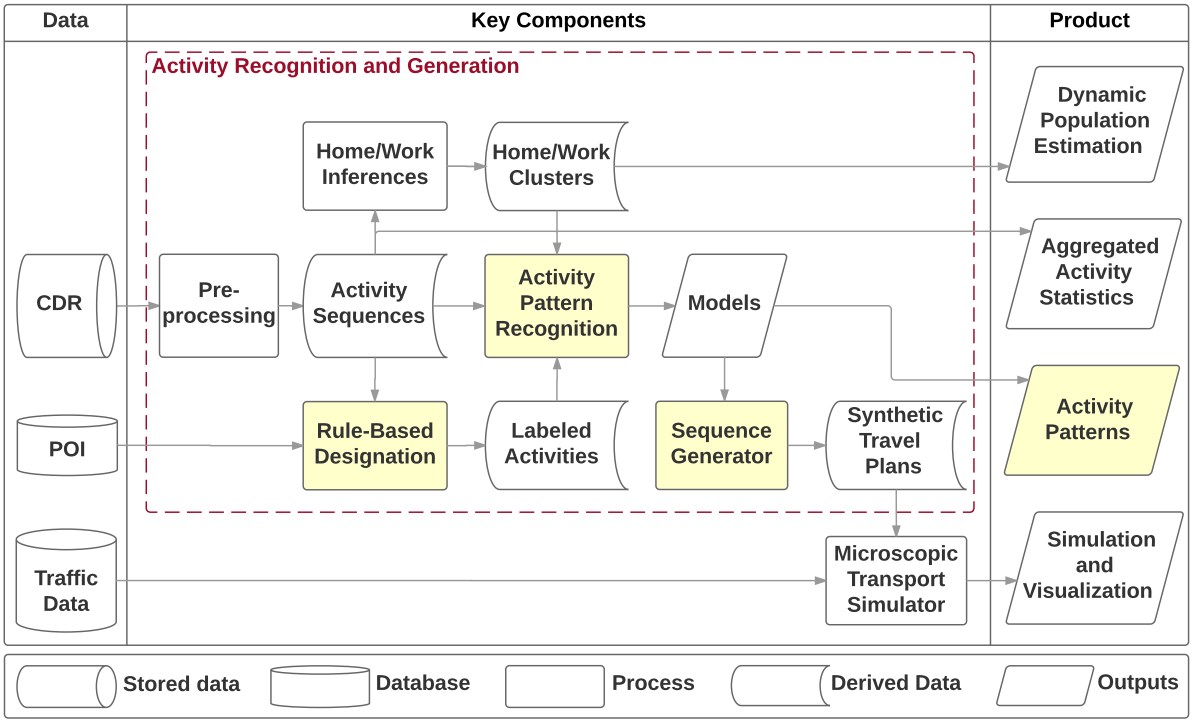
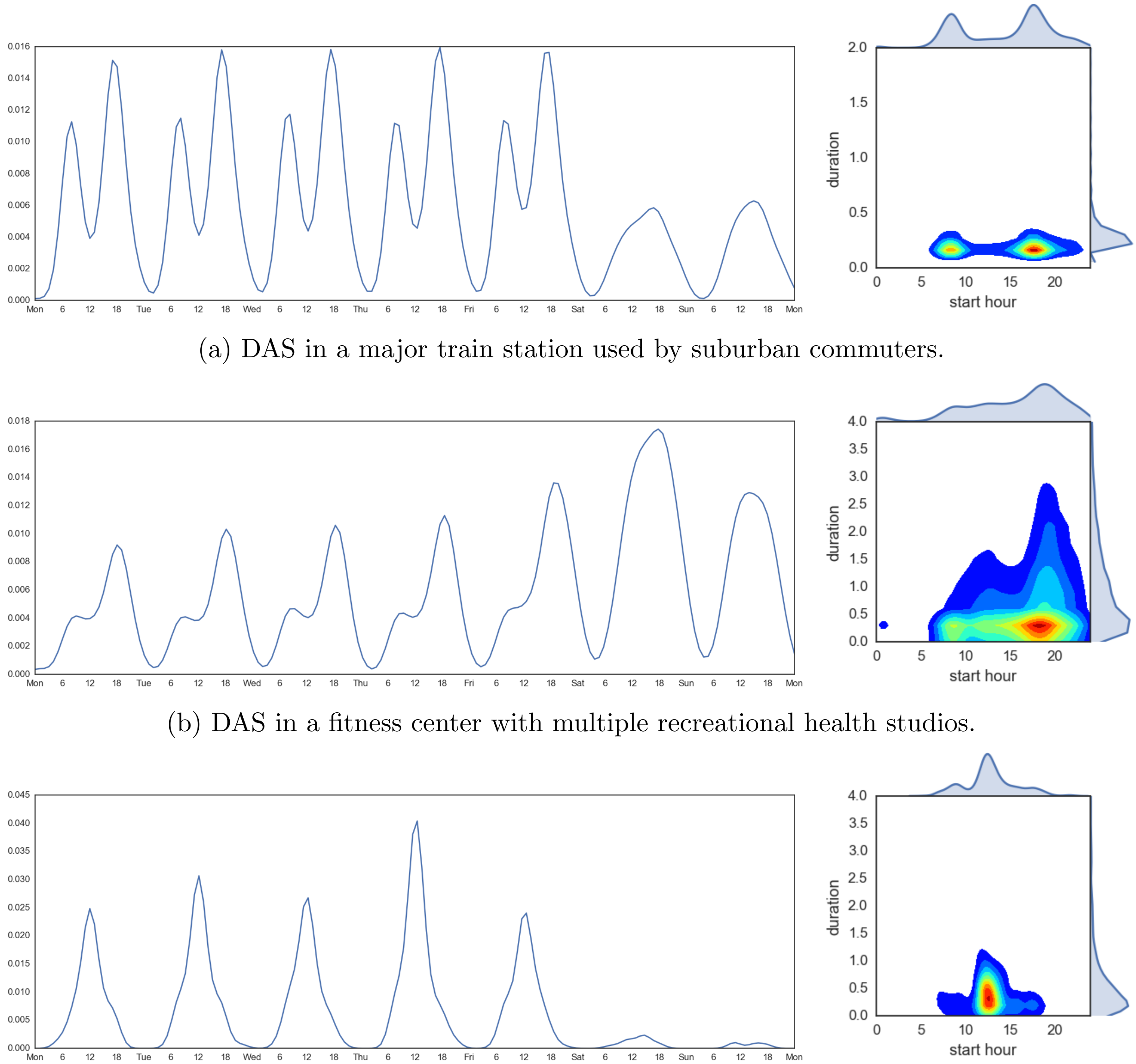


Figure 4.1: Modeling framework diagram. The left column represents the input to the research; the middle column represents the key modeling components; and the right column represents the products of the research. Our key contribution of activity recognition and generation module are outlined with the red dashed rectangle, and the key components are shown in shaded yellow.

With the processed activity sequences and inferred primary activities from previous chapter, we can perform the secondary activity recognition and analyze the activity patterns, including spatial-temporal profiles of activities and activity transition probabilities. These are the cores of our urban mobility models. The resulting models and analysis will be the third product of the research. To validate the activity recognition results, we collect 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, as detailed in Chapter 4.3. 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. Ground truth traffic counts data is used to validate the simulation results and showcase the validity of the presented work for transportation planning and operations practice. This is the fourth product in Fig. 4.1.

4.3 Collection of Ground Truth Activities

Model selection for the activity-based urban mobility models includes the choice of hidden states (activity types). One would like to set a high number of hidden states 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") suggests limiting the number of hidden states to mitigate confusion 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 is 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 small areas, which have relatively high spatial resolution. These short range antennas provide us the opportunity to collect "ground truth" activities.



(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 4.2: 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).

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

Table 4.1: 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

signal coverage. 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. 4.2 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 design 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 acquire place information from POI databases such as Google places API and Factual Global Places API. Then, we join 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 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 4.1. The "location category" column of the table indicates that the category is among the category labels returned from the APIs.

4.4 Semi-Supervised IOHMM for Secondary Activity Modeling

Given the user stay history, that is, a list of stay location features with start times and duration, we would like to convert it into a sequence of activities enriched with semantic labels ("shopping", "leisure", etc.), and a heterogeneous context-dependent probability model of transitions between the activities.

IOHMM Architecture

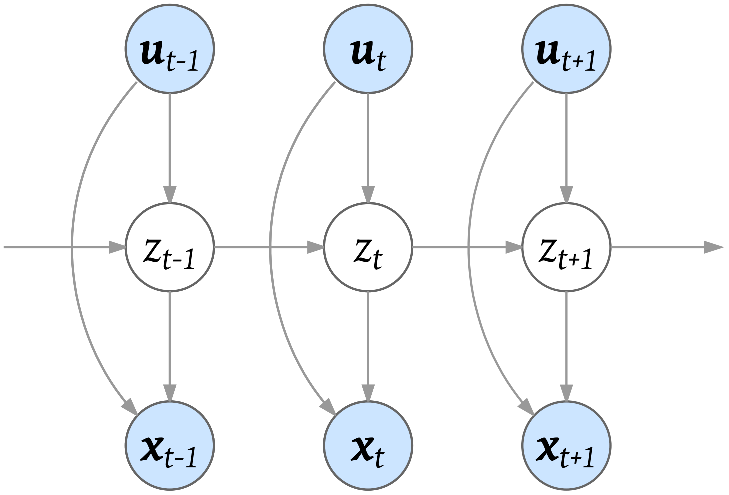


Figure 4.3: IOHMM Architecture. The solid nodes represent observed information, while the transparent (white) nodes represent latent random variables. The top layer contains the *observed* input variables $\boldsymbol{u_t}$; the middle layer contains *latent* categorical variables z_t; and the bottom 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. 4.3, 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:

$$\begin{aligned} L\left(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{u}\right) &= \sum_{\boldsymbol{z}} \Big(\Pr\left(z_1 \mid \boldsymbol{u_1}; \boldsymbol{\theta_{in}}\right) \cdot \\ & \prod_{t=2}^{T} \Pr\left(z_t \mid z_{t-1}, \boldsymbol{u_t}; \boldsymbol{\theta_{tr}}\right) \cdot \\ & \prod_{t=1}^{T} \Pr\left(\boldsymbol{x_t} \mid z_t, \boldsymbol{u_t}; \boldsymbol{\theta_{em}}\right) \Big). \end{aligned} \tag{4.1}$$

IO-HMM architecture has been well described in [15]. Variable notation and important differences between IO-HMM and standard HMM are summarized in Table 4.2.

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:

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta^k}\right) &= \sum_{i=1} \gamma_{i,1} \log \Pr\left(z_1 = i \mid \boldsymbol{u_1}; \boldsymbol{\theta_{in}}\right) \\ &+ \sum_{t=2}^{T} \sum_{i} \sum_{j} \xi_{ij,t} \log \Pr\left(z_t = j \mid z_{t-1} = i, \boldsymbol{u_t}; \boldsymbol{\theta_{tr}}\right) \\ &+ \sum_{t=1}^{T} \sum_{i} \gamma_{i,t} \log \Pr\left(\boldsymbol{x_t} \mid z_t = i, \boldsymbol{u_t}; \boldsymbol{\theta_{em}}\right). \end{aligned} \tag{4.2}$$

In the above, $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta^k}\right)$ 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 4.2.

Table 4.2: 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 $\gamma_{i,t}$	$\alpha_{i,t} \beta_{i,t} \,/\, L_c$		

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 4.2, 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`.

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 manual label [78], or collected for a small group of participants from a survey accompanying GPS data [69]. 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.

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 [144] 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 [47]. Co-training can also be applied by using the same set of features but two different classifiers, which has been proven to perform well [52]. 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. We will also include a self-training experiment using only the semi-supervised IOHMM with EM algorithm as the baseline.

Algorithm 1 Self-training of urban activities

```
Input: Labeled data $L$, unlabeled sequences $S$, confidence thresholds $\theta$
Output: IOHMM model $m$.
while $L$ changes do
    Train semi-supervised IOHMM $m$ from $S$ and $L$.
    Classify the unlabeled data with $m$ and
    Add data labeled by $m$ with confidence $\geq \theta$ to $L$.
end while
return $m$.
```

Algorithm 2 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 $m_1$, $m_2$.
```

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 $\Pr(z_t = j \mid z_{t-1} = i)$ 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 4.2.

4.5 Model Specifications

As we have mentioned, there are two components in the co-training process, one is the generative IOHMM, and the other is the decision tree classifier. We will present our specifications (features) in this section. Note that we use the same IOHMM specification for the unsupervised IOHMM and HMM model where it applies.

IOHMM Specification

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 4.2. The first term of Equation 4.2 can be written as:

$$\Pr\left(z_1 = i \mid \boldsymbol{u_1}; \boldsymbol{\theta}\right) = \frac{e^{\boldsymbol{\theta^i u_t}}}{\sum_k e^{\boldsymbol{\theta^k u_t}}}. \tag{4.3}$$

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 4.2 can be written as:

$$\Pr\left(z_t = j \mid z_{t-1} = i, ; \boldsymbol{\theta}\right) = \frac{e^{\boldsymbol{\theta_i^j u_t}}}{\sum_k e^{\boldsymbol{\theta_i^k u_t}}}. \tag{4.4}$$

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 4.2 can be written as:

$$\Pr\left(x_t \mid z_t = i, \boldsymbol{u_t}; \boldsymbol{\theta_i}\right) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_t - \boldsymbol{\theta_i} \cdot \boldsymbol{u_t})^2}{2\sigma_i^2}}, \tag{4.5}$$

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 4.2 can be written as:

$$\Pr\left(x_t = 1 \mid z_t = i, \boldsymbol{u_t}; \boldsymbol{\theta_i}\right) = \frac{1}{1 + e^{-\boldsymbol{\theta_i} \cdot \boldsymbol{u_t}}}. \tag{4.6}$$

Finally, we emphasize that an activity label is just a latent categorical variable. A semantic label can be associated to it following an in-depth analysis the we present in Section 4.7 below.

Decision Tree Specification

Decision trees are interpretable classifiers that are capable of generating arbitrarily complex decision boundaries. They have been used successfully in many diverse areas [103]. 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.

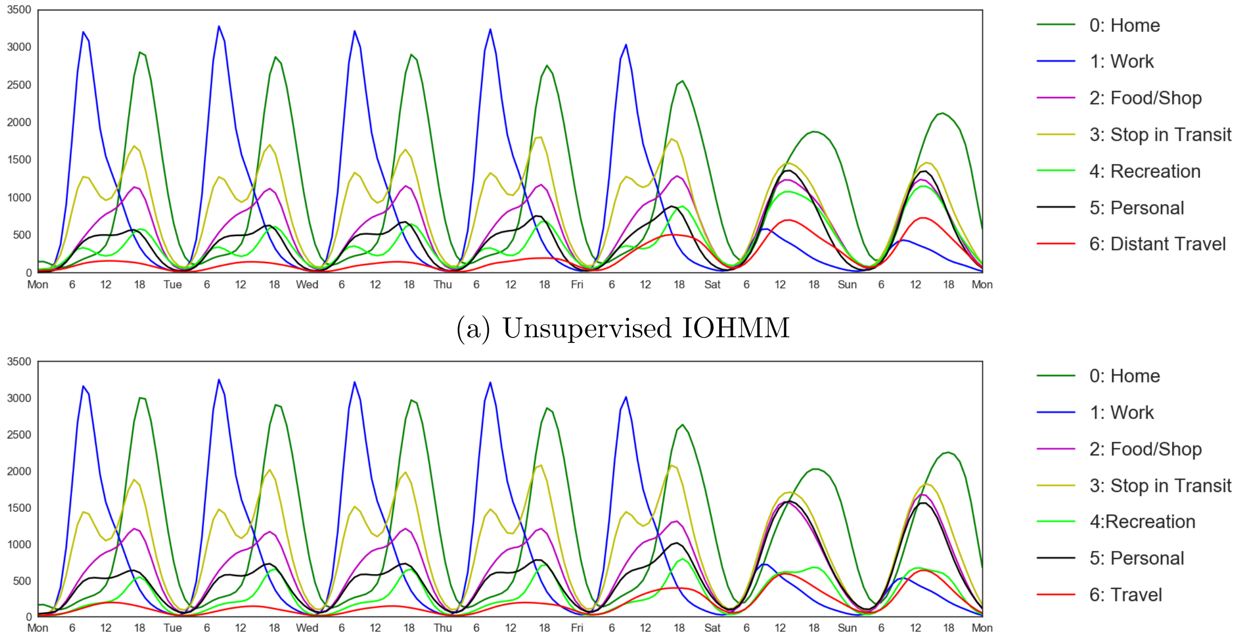
4.6 Description of Data

The data we used in this chapter are the processed activity (yet unlabeled) sequences from the previous chapter. Each sequence contains the one month activities from a San Francisco regular commuter. As we mentioned, the median number of activities is 4.4 per weekday and 4.0 per weekend. The model was trained on a group of 20,000 anonymous San Francisco residents (about 2% of the population).

4.7 Experimental Results

In this section we present the results of the unsupervised IOHMM and co-training IOHMM that have been fit to the four super-districts that make up the city of San Francisco.

Two temporal representations help identify the latent semantics of the hidden states (i.e. activities). Fig. 4.4a depicts the distribution of start times of activities using unsupervised IOHMM model and Fig. 4.4b depicts the distribution of start times of activities using co-training. The y-axis gives the number of activities started at a given hour. For the unsupervised model, by evaluating these weekly activity start-time patterns in combination with the output coefficients in Table 4.3, and the joint distribution of start time and duration in Fig. 4.5 we can assign semantic labels for activity type to each of latent activity states. From the two figures we can see that the activity profiles recognized from co-training are very similar to the ones recognized by unsupervised IOHMM. We will discuss the results from unsupervised IOHMM in the following sections since the unsupervised model can be applied to areas where high resolution data (such as short range antennas) is not available thus ground truth data cannot be collected to assist semi-supervised models.



(a) Unsupervised IOHMM

(b) Co-training

Figure 4.4: Number of activities (labeled per highest posterior probability) by their respective start time within a course of a week.

The coefficients of the unsupervised IOHMM emission models are reported in Table 4.3. Recall that we use linear models as the output models for $x^{(1)}$, distance to home, $x^{(2)}$, distance to work, and $x^{(3)}$, duration of the activities. Logistic regression was used as the output model for $x^{(4)}$, cluster has been visited before. Since $x^{(1)}$ and $x^{(2)}$ depend only on the hidden activity, only the intercepts are estimated. For $x^{(3)}$, we specify that the duration depends on activity type and also on the "day of week", "time of day" and "hours worked" input variables, there are 8 coefficients estimated per hidden state for this output. Since $x^{(4)}$ "has visited" is a binary variable, only one parameter per hidden state is identifiable.

Primary Activities: Home and Work

Activity state 0, shown in green in Fig. 4.4a is the "home" activity. The typical start time ranges from 3pm to midnight. The home activity exhibits greater variation in start time on Friday and weekends than on other weekdays. The positive "weekend" coefficient on the duration of this activity indicates that people stay at home longer during weekends.

The temporal profile of home activities in Fig. 4.5a has two major clusters. The upper cluster indicates regular overnight home activities. This cluster can be further separated into two clusters. One peaks at 6pm, representing the home activity directly after work. The other peaks at 9pm, representing the home activity after some secondary activities in the evening. Since the home activity duration is generally set by the regular work start hour,

Table 4.3: Unsupervised IOHMM Model coefficients for the output variables per hidden activity (see interpretation in the text).

State: latent activity	Dist to home	Dist to work	Duration								Visited	
			constant	weekend	morning	lunch	afternoon	dinner	evening	hours worked	no	yes
0: Home	0.00	7.22	9.45	2.17	-6.29	-2.57	-0.94	0.20	1.29	-0.03	0	2.19
1: Work	7.22	0.00	4.00	-0.02	2.98	0.76	0.19	-0.64	-0.10	-0.26	0	1.76
2: Food/Shop	2.37	1.90	0.84	0.18	0.00	-0.01	-0.04	-0.01	0.25	0.00	0	-0.53
3: Stop in Transit	3.21	3.63	0.16	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0	-0.46
4: Recreation	2.36	15.03	2.76	0.17	-0.42	-0.64	-0.45	-0.68	0.37	0.04	0	-0.44
5: Personal	18.79	16.94	0.93	0.46	0.17	0.12	-0.05	-0.03	-0.05	0.01	0	-1.35
6: Distant Travel	787.94	784.71	4.26	0.78	-0.75	-0.39	-0.76	-1.27	1.11	0.29	0	-1.17

the downward slope of the upper cluster signifies that if a user arrives at home later in the day, they are likely to spend fewer hours at home.

Activity state 1, shown in blue in Fig. 4.4a is the "work" activity. It has highest peaks in Fig. 4.4a, signifying that it is a very regular activity with concentrated start times.

According to Table 4.3, a work activity has a base duration of 4 hours, if it starts in the morning, the user is likely to stay 2.98 hours longer, that is 6.98 hours in total; if it begins in the afternoon or evening the average duration is shorter. As a compounding effect of returning to work in the afternoon or evening, the "hours worked" column indicates that the expected duration will decrease by 0.26 hours for every hour that the user already spent at work in the day. The "is weekend" column indicates that if a user chose to work on weekend, the average work activity duration is not significantly different from that on weekdays; note that (from Fig. 4.4a) the probability of visiting the work activity is much lower on the weekend. The "visited" column indicates the propensity of the location being frequently revisited. For the work activity, the coefficient 1.76 indicates a very high likelihood of returning to the same location to perform the same activity.

From Fig. 4.5b, we can see that the temporal profile of work activities has three clusters. The upper cluster indicates regular "9 to 5" work activities without a break. The lower left cluster represents the morning work activities and the lower right cluster represents the afternoon work activities. All three clusters are tilted at -45 degrees. This is due to the usually fixed lunch hour at noon and end of work at about 5pm.

Secondary Activities

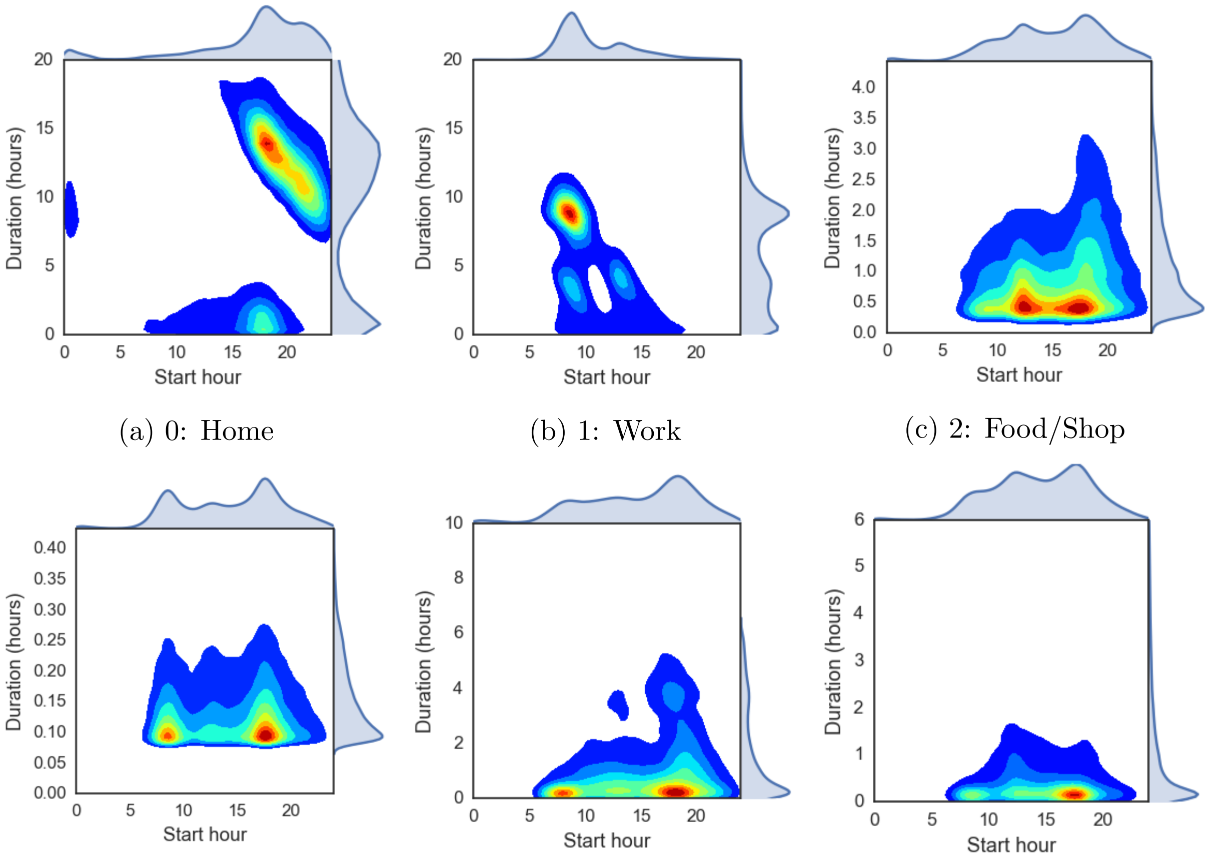
The remaining states are secondary activities. "Food/shop" activity peaks in start time around noon and in the evening. As shown in Table 4.3, "food/shop" activity has an average duration of about 0.84 hours, and is close to both home and work place. As shown in Fig. 4.5c, the duration of this activity is slightly longer in the evening. From Fig. 4.4a we see that, on weekends, this activity peaks at noon. The weekend activity duration, according to Table 4.3, is about 0.2 hours longer than it is on weekdays.

"Short stop in transit" is located close to home and work, and has an average duration of about 10 minutes, according to Table 4.3. From Fig. 4.5e, we can see that this activity peaks in the early morning and late afternoon right before home activity. Fig. 4.5d and Table 4.3 also indicate that the duration is not affected by time of day or day type (weekend vs. weekday). From Fig. 4.4a, we can see that this activity is visited more frequently on weekdays than weekends. It is worth noting that although short stop in transit is less revisited than home and work activities, it is more likely to be revisited compared to other activities.

As seen in Table 4.3, the "recreation" activity is quite close to home but far from the work place. The state has an average duration of 2.7 hours, much longer than the durations of food/shop activity and short stop in transit. This activity last longer in evening hours or weekends. As shown in Fig. 4.4a, this activity often starts in the early morning or evening hours on weekdays, and tellingly, more users engage in this activity on Fridays and weekends.

Activity state 5 is "personal" business. The distances from home and work are 19 and 17 miles, respectively, and the average duration of this activity is 0.93 hours. This state could encompass both off-site work related trips and/or longer-distance dining or leisure activities. As shown in Fig. 4.4a. Due to the distance of this activity, more users engage in this activity on weekends and this activity is least likely to be revisited.

Activity state 6, "distant travel", or more accurately activities that occur while traveling, is the most irregular and infrequent. The average distances from home and work are quite high (average 800 miles). This activity type seems to occur predominantly on Fridays and weekends according to Fig. 4.4a.



(a) 0: Home

(b) 1: Work

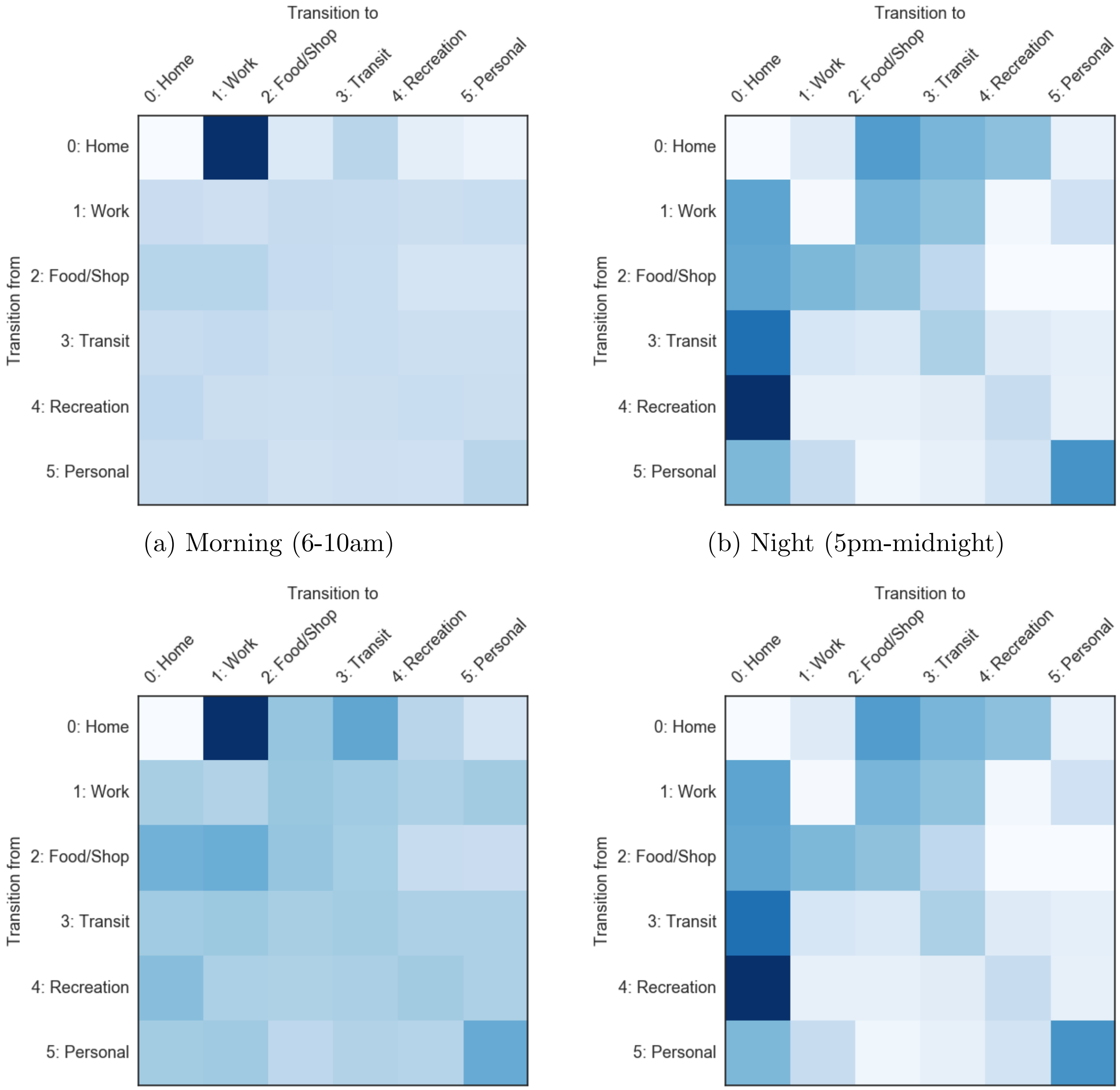
(c) 2: Food/Shop

(d) 3: Stop in Transit

(e) 4: Recreation

(f) 5: Personal

Figure 4.5: Joint distribution plot of duration and start hour per activity type. The labels are gained by assigning the activity to the one with the highest posterior probability after training.



(a) Morning (6-10am)

(b) Night (5pm-midnight)

(c) Afternoon (12-2pm), users who have not visited work

(d) Afternoon (12-2pm), users who have worked 5 hours

Figure 4.6: Heterogeneous activity transition matrices under different contextual variables.

Activity Transitions

We omitted "distant travel" activity from the transition matrix since if a person is traveling a long distance, the next activity is also most likely to be categorized as "distant travel"; the distance dominates the state. Fig. 4.6a shows the transition matrix associated with mornings. The labels on the left indicate the state the user is transitioning from, and the labels on the top indicate the state the user is transitioning to. The most significant transition is from "home" to "work." Fig. 4.6b shows the transition matrix associated with evenings. The transitions from all other states to "home" are significant. However, if the user's transition from activity is "home", then she is more likely to transition to "food" or "recreation" activities. Fig. 4.6c shows the transition matrix in the afternoon, for users who have not yet visited the "work" state in the day. For these users, there is a high probability of going to work. As in Fig. 4.6d, by keeping all the input context information equal as in the previous case, and only specifying that the simulated user has previously worked for 5 hours on that day, one can see that the probability of going to work is significantly reduced.

4.8 Model Validations

Recognition Accuracy

The distribution of collected ground truth activities are biased and do not correspond to the true distribution of urban activities. To reasonably evaluate performance of IOHMM, we need to sample a subset of ground truth activities so that the sample weight is consistent with the true distribution of urban activities. According to the the distribution given by the 2015 Travel Decisions Surveys (TDS), conducted by San Francisco Municipal Transportation Agency (SFMTA)[104], we sampled (scaled) 10000 home activities, 7500 work activities, 5000 Food/Shop activities, 7500 Stop in Transit activities, 3000 recreation activities, 4000 personal activities and 1000 Travel activities.

Table 4.4: Confusion matrix of inferred activities vs "groun truth" activities

Ground Truth	Annotations								
	Home	Work	Food/Shop	Transit	Recreation	Personal	Travel		
Home	9994	0	0	0	1	1	4	0.999	Recall
Work	0	7495	0	0	0	2	3	0.999	
Food/Shop	0	0	3013	413	1307	267	0	0.603	
Transit	0	0	31	6980	359	130	0	0.931	
Recreation	0	0	1519	0	1403	78	0	0.468	
Personal	0	0	321	17	84	3426	152	0.857	
Travel	0	0	0	0	0	11	989	0.989	
	1.000	1.000	0.617	0.942	0.445	0.875	0.862	0.876	
	Precision								

For the unsupervised IOHMM model, we get 87.6% accuracy on all activities, with a macro-precision of 82%, a macro-recall of 83.5% and a macro-f1 score of 0.827. From the confusion matrix in Table 4.4, we can see that most confusion happens between "food/shop" and "recreation" activities. This is natural because "food/shop" and "recreation" activities are similar in time and space. We also notice that some "food/shop" activities are mistaken as

Table 4.5: Comparison of model accuracy

Model	All Activities		Secondary Activities	
	Accuracy	F1	Accuracy	F1
HMM	0.859	0.783	0.739	0.698
Partial IOHMM	0.866	0.824	0.752	0.754
Full IOHMM	**0.876**	**0.827**	**0.771**	**0.758**
Self-training	**0.930**	**0.915**	**0.875**	**0.883**
Co-training	**0.961**	**0.949**	**0.930**	**0.930**

a “short stop in transit”, this is because some “food/shop” activities and “stop in transit” are close in space, thus some short “food/shop” activities are taken as “stop in transit” because of the duration. Since the activities that we labeled as “personal” are mainly medium distance activities that could encompass longer-distance dining, some “food/shop” activities could also be confused as “personal”.

To compare the performance of different models, we also report the accuracy of (1) Hidden Markov Models (HMM) with the same output as IOHMM but with no inputs; (2) Partial IOHMM with transition probabilities dependent on inputs while all emissions are only conditioned on hidden states; (3) Full IOHMM as described; (4) Self-training with EM; and (5) Semi-supervised co-training, in Table 4.5.

We report the accuracy and macro-f1 score as metrics of success for our models. F1 score can be interpreted as a weighted average of the precision and recall. For multi-class tasks, macro-f1 score calculates the average per-class precision and recall and then perform the f1 score calculation.

Comparing full IOHMM, partial IOHMM, and standard HMM, we can see that the full IOHMM has the best performance. Since “home” and “work” are rather easy to infer, we also report the performance for secondary activities only. For the five class classification task, we get 77.1% accuracy. Another observation is that the macro-f1 score of the partial and full IOHMMs do not differ too much, but all outperform the pure HMM. These results exhibit the benefits of the context-dependent transition models. Since “home” and “work” have high accuracy, the improved performance is mainly in secondary activity recognition. In all cases, f1 score is smaller than the accuracy. This is because the class that has higher support also has higher accuracy. Since accuracy score is a weighted average with support while macro-f1 score is an unweighted average, f1 score is lower than the accuracy.

On the other hand, comparing the semi-supervised approaches with the unsupervised approach, we can see that semi-supervised co-training has the best recognition accuracy and f1 score, which outperforms the self-training results, which outperforms the unsupervised IOHMM. As expected, this improvement on the activity recognition is due to the learning direction given by the ground truth activities. We acknowledge that the ground truth activities used in the semi-supervised methods have the same temporal-spatial distribution as the ground truth activities used for validation. Since they were collected strictly, there might be overfitting issues which results in the very high recognition accuracy of semi-supervised models. We show in Chapter 5 that semi-supervised models are also better in predicting activity sequences than unsupervised models in a totally different task, which shows the advantage of the directed learning process.

Survey-derived statistics

Another way to evaluate the method is to compare our model with aggregated statistics from surveys. We consider the Travel Decisions Survey (TDS), which contains 1000 random digit dial and cell phone samplings in the area of interest. Overall, the activity proportions of our model match with TDS. If we split our Food/Shop activities into half food and half shop, food and recreation is 20% in our model versus 21% in TDS; shopping and errand (personal) is 21% in our model versus 20% in TDS. Work/school activity is 22.5% in our model vs 23% in TDS. The main difference is with the “Home” activity, for which TDS report a proportion of 35%, which is a little higher than the proportion of 30% reported by our model. This discrepancy is likely due to under-reporting of secondary activities in TDS.

4.9 Activity Sequence Generation from an IOHMM

One of our goals is to enable activity-based travel demand models that use cellular data to create synthetic agent travel patterns without compromising the privacy of cell phone users. As such, we test our models’ generative power in the Bay Area context — we simulate $463,000$ agents in the Bay Area (15% sample of the commuters) and create a day-long activity plan for all agents with anticipated start-times, locations, and durations of all activities in the day.

As travel patterns vary greatly over the region, we trained 34 IOHMMs, each for a subset of cell phone users residing within each of the 34 super-districts as defined by the San Francisco Metropolitan Transportation Commission (MTC). Using the Iterative Proportional Fitting [45] procedure to fit the population marginals with the census data, we sample residents home and work locations to create synthetic driver with a predetermined home TAZ and work TAZ. The numbers were further adjusted according to occupancy statistics from CHTS (single driver, two and multi-person carpool). The precise home and work locations (lat/lon coordinates) are sampled uniformly within the home and work TAZs.

Each simulated user is assumed to start her day at home. The home departure time and the transition time are drawn from their respective distributions to determine the start time of the first activity. Home departure times for the first non-home activity of the day are modeled as Gaussian random variables with super-district dependent mean departure time and standard deviation calibrated from CDR records. As IOHMM is trained on the observed travel sequences with *revealed* departures times, we assume that it captures the dependencies of transition times on the origin and destination, travel mode and traffic conditions.

Generation continues until the activity start time reaches midnight. At every step, previous activity state and context information are used to obtain transition probabilities from the IOHMM and sample the next activity state according to the transition probabilities. After the activity type has been selected, the activity duration is sampled from a truncated normal distribution with mean and standard deviation coming from output $x^{(3)}$ of the IOHMM. Next, the activity location is selected - if the activity is a home-activity or work-activity, the exercise is trivial. If not, we use IOHMM outputs $x^{(1)}$ and $x^{(2)}$ - the distance between the stay location and the user's home output and distance from the stay location to the user's work output from the IOHMM to generate a new destination TAZ from the choice set of TAZs within matching distances. The precise location of the activity is sampled uniformly from the selected TAZ. Note that future research on destination location choice models could improve the location selection process for secondary activities.

Due to the nature of IOHMM, we must filter out and discard unrealistic activity chains generated in this process. One reason of the unrealistic chain is due to the softmax approximation of the transition probabilities - though there is strictly no transitions from home to home or from work to work (thus the probabilities should be strictly zero), the softmax function cannot be exactly zero. The second reason of unrealistic chain is due to the coarse representation of the contextual variables, especially the temporal variables - different start times may end up with the same contextual variables. This problem can be solved by increasing the dimension of the contextual variables. However, this may cause the overfitting issue thus is not presented in this work without sufficient amount of data. 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. Fig. 4.7 presents 4 common and interesting (among top 20) activity patterns generated from IOHMM model.

Overall, the aggregated statistics of activity patterns match with the travel surveys. For example, the percentage of US employed person who go to work on an average weekday is 82.9% [73], this number is 83.7% for our simulated population. Considering the summary statistics for people who go to work, we compare the percentage of people who participate in activities at different times of day. The percentage of people participating in at least one activity before morning commute, during morning commute and after work is 3.1%, 14.8% and 46.3% in the Bay Area Travel Survey [17] and these numbers are 2.9%, 15.2% and 43.7% in our simulated population.

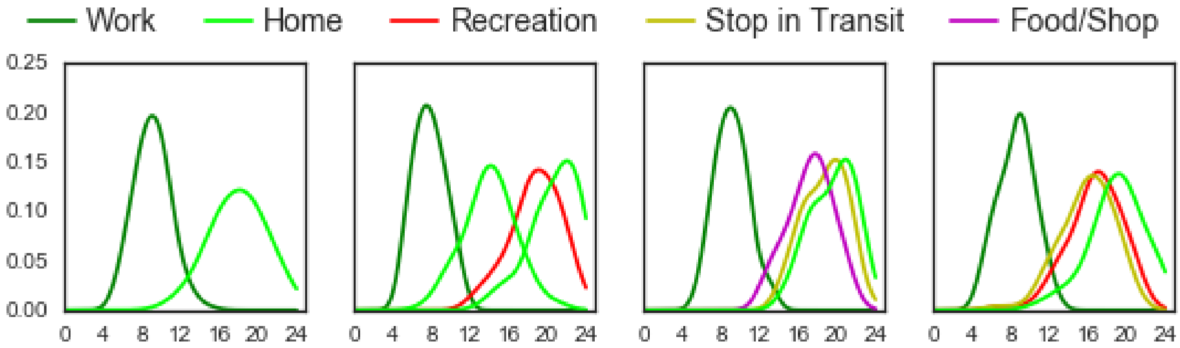


Figure 4.7: Distribution of activity start times over a course of a day of four example common activity patterns generated from the Bay Area IOHMMs. Note that all simulated activitity patterns start at home, so (a) designates the Home-Work-Home travel pattern. The x-axis designates the start time of the activity, the y-axis represents the proportion of trips (for users with this activity pattern) starting at this time.

4.10 Evaluation via Traffic Micro-simulation

Traffic micro-simulation is a conventional approach in studying performance and evaluating transportation planning and development scenarios. Ground truth observations of the flows at sections of the road network provide an independent data source that can be used to evaluate the accuracy of the activity generation model. We present here a summary of the validation results based on the traffic volume data collected by the California DOT freeway Performance Management System (PeMS) in the 9 counties of the Bay Area (see Fig. 4.8). Micro-simulation of a typical weekday traffic is performed using the MATSim platform [9]. MATSim is a state-of-the-art agent based traffic micro-simulation tool that performs traffic assignment for the set of agents with pre-defined activity plans. It varies departure times and routing of each agent depending on the congestion generated on the network, in order to maximize agent's daily utility score. We have 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 generated activity sequences with the observed traffic volumes. As the model is trained to reproduce average weekday, hourly traffic volumes are taken as averages over all weekdays (except for Mondays and Fridays) of Summer 2015. The simulation is run at 15% of the total population, and the road capacities as well as total resulting counts are scaled accordingly.

Note that observed traffic counts are not used for model calibration. They are used as independent data to evaluate the validity of the synthetic travel sequences produced with IOHMM. The locations of the sensors on the road network are presented in Fig. 4.8. It also demonstrates examples of the three characteristic hourly volume profiles comparing the modeled and observed counts. The results for the full set of sensors are presented in Fig. 4.9. Fig. 4.9a shows a comparison of the volumes for three distinct time periods.

Fig. 4.9b summarizes the validation results over all 600 sensors in terms of the relative error (% volume) over-/under-estimated by the model as compared to the ground truth. One can notice lower accuracy at night and early morning hours explained by the fact that the model was developed and applied on a subset of daily commuters and did not include a large portion of trips performed by unemployed population and people working from home, besides multiple other traffic components (commercial fleets, taxis, visitors) that are out of scope of the model. Despite it's relative simplicity, the model has demonstrated a reasonable accuracy ($r^2 = 0.81, p < 10^{-3}$ in Fig. 4.9a) as compared to the ground truth data.

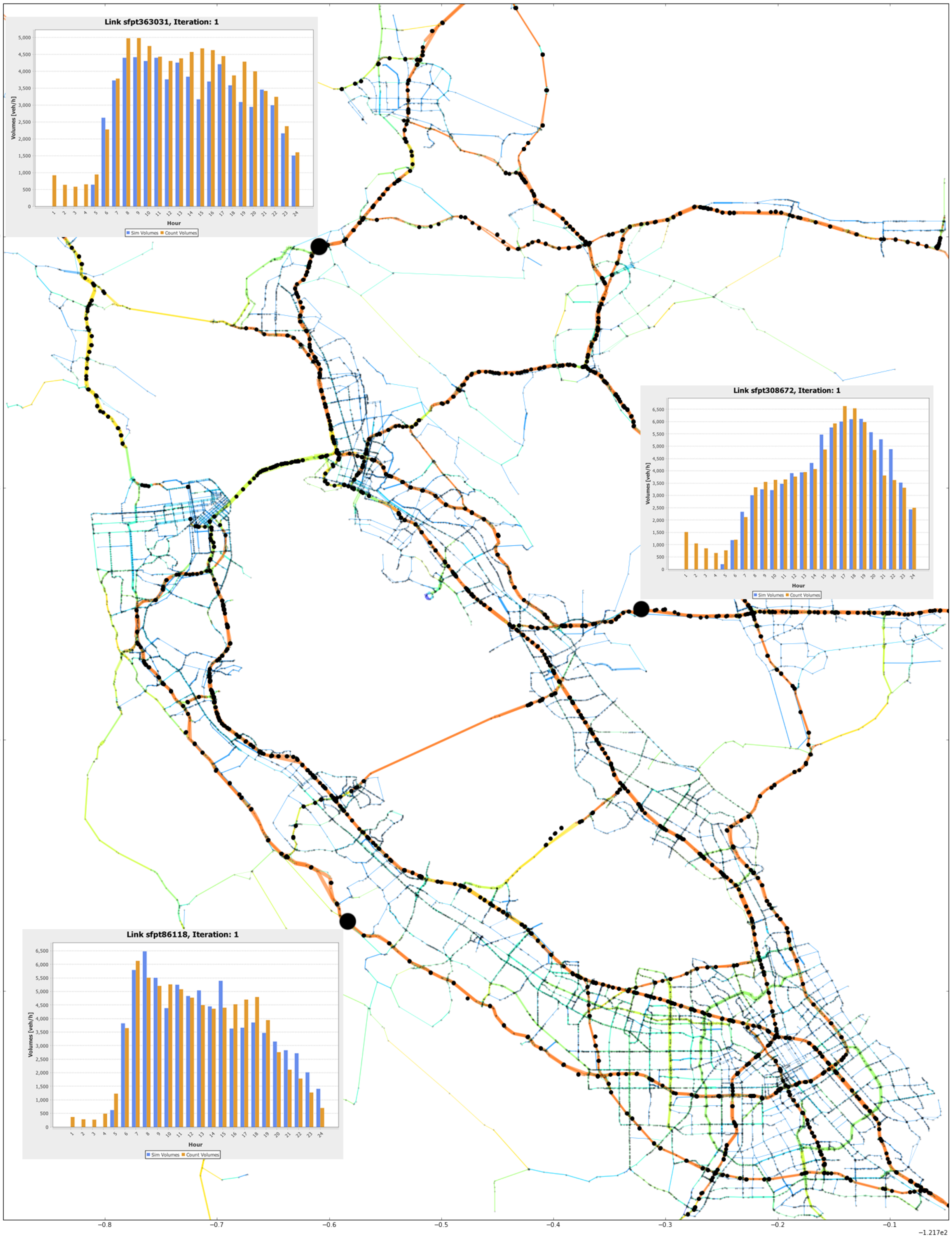
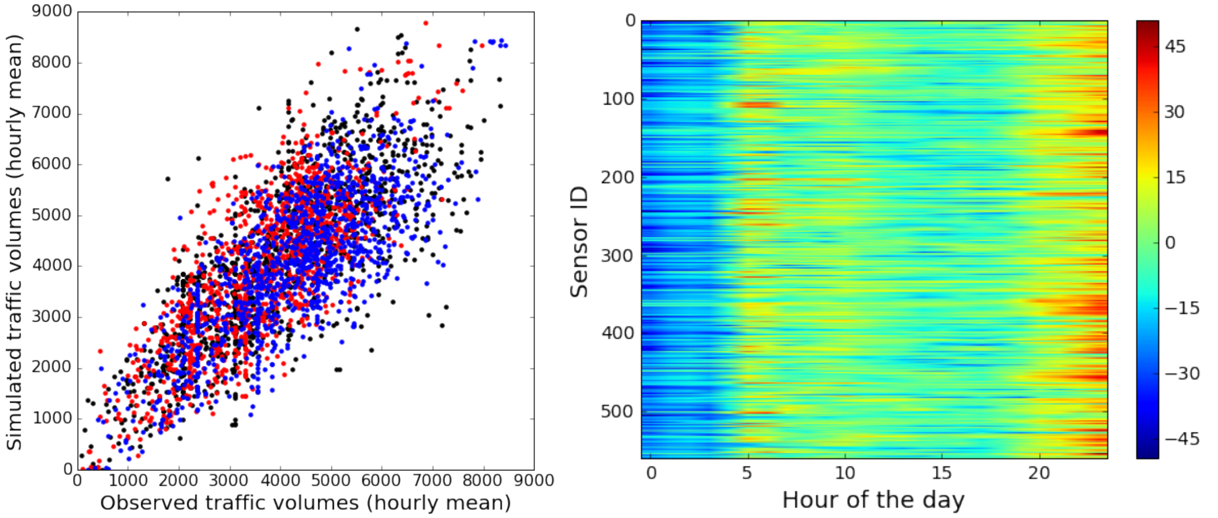


Figure 4.8: A fragment of the SF Bay Area road network with the location of 600 traffic volume detectors used for validation (shown with small black dots). Inlet graphs illustrate three sample hourly vehicle volume profiles for observed (orange) and modeled (blue) flows on a typical weekday in Summer 2015.



(a) Modeled vs observed volumes at 8am (black),1pm (red) and 6pm (blue) ($r^2 = 0.81, p < 10^{-3}$).

(b) Mean relative error (%) over all 600 sensors of modeled vs observed traffic volumes during the day over all 600 sensors.

Figure 4.9: Micro-simulation validation with the observed freeway traffic volumes

4.11 Conclusion

In this chapter, we developed scalable and interpretable generative activity-based urban mobility models for regional mobility analysis from cellular data. As an illustration, we inferred the activity patterns including primary, secondary activities and heterogeneous activity transitions of a set of anonymized San Francisco Bay Area commuters using unsupervised and semi-supervised generative state-space models. We validated this inference by comparing it with (1) 2015 Travel Decisions Surveys (TDS) on the aggregated activity statistics; and (2) a set of ground truth activities based on short range distributed antenna system (DAS); (3) observed volumes of vehicular traffic flow in the regional road network on an average weekday. To examine the generative power of the model, we synthesized travel plans for each agent with home and work locations sampled from census data. An agent-based microscopic traffic simulation was conducted to compare the resulting traffic with real traffic, and a reasonable fit accuracy was observed. An interesting extension to this work is to compare the activity sequence generation power of different techniques, from baseline models with only home and work activities to more advanced IOHMM models and recurrent neural network such as long short term memory (LSTM) models.

Several improvements can be built upon the presented work. Partitioning a population into sub-groups (whether socially or spatially) for shared parameter modeling is a partly open problem. Currently we approached it by defining rules to identify groups of a similar day structure, and applying geographic constraints. This step will be compared to an alternative specification that involves a mixture of urban mobility models.

With privacy concerns and data limitations in mind, the location choice model implemented in this chapter is relatively simple. Future work may incorporate a discrete choice model on a set of TAZs so that locations can be directly sampled when generating activity sequences.

Activity patterns inferred and analyzed in this chapter reveal the spatial and temporal profile of activities of regular commuters, as well as the heterogeneous transition probabilities dependent on contextual information. The generative nature of our proposed model allows to sample accurate travel scenario inputs needed by activity-based travel micro-simulation models. A range of issues remain where the advantages of using cellular data alone are not straightforward. This includes travel mode detection, identification of the number of car-pools, modeling short-range and non-motorized travel to name a few. Nevertheless, such methods derived from automatically and continuously collected cell phone data are bound to make a substantial impact on urban and transportation planning, and represent a significant improvement upon the state-of-the-art.

Chapter 5

AM-PM: Travel Demand Nowcasting

5.1 Introduction

Travel demand forecasting has been an integral part of most Intelligent Transportation Systems research and applications [123]. Long term forecasts (days, months, or even years ahead) provide 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? These studies typically use data collected from travel surveys that are infrequent, expensive, and reflect changes in transportation only after significant delays.

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. 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. In this chapter, 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 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 urban mobility models.

To validate the predictions, we 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. Results prove that we can improve the medium term travel demand forecast by incorporating observed information by the time of prediction. The mean absolute percentage error can be less than 5% one hour ahead and around 10% three hours ahead for the regional road network.

The main contributions of this chapter 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.
- 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 5.2 reviews related work on long term travel demand forecast models and short term traffic prediction models. Chapter 5.3 depicts the framework of medium term travel demand forecast. Chapter 5.4 improves the state-of-the-art deep generative urban mobility models using long short term memory (LSTM). Technical details on sequence completion from partially observed sequence are presented in Chapter 5.6. We describe the data in Chapter 5.7. In Chapter 5.8, we report on experiments, model selection, and validation results. We conclude the present work and offers discussions in Chapter 5.9.

5.2 Related Work

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 [25]. 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 [56, 11]. In the recent decades, such forecasts are performed by activity-based models for demographic projections of a population. 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 [5]. However, the data collection is usually performed through travel surveys that are infrequent, expensive, and reflect the changes in transportation with significant delays. 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.

Short Term Traffic Forecasting

With growing availability of data, short-term traffic forecasting became a very developed research area. It concerns predictions of traffic parameters 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 [123]. Vlahogianni thoroughly summarized the available literature and categorize papers 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. However, 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 [138]. Ye et, al. further introduced acceleration information and information from adjacent segments to improve the prediction of the travel speed of current forecasting segment [134]. 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.

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

5.3 Modeling framework

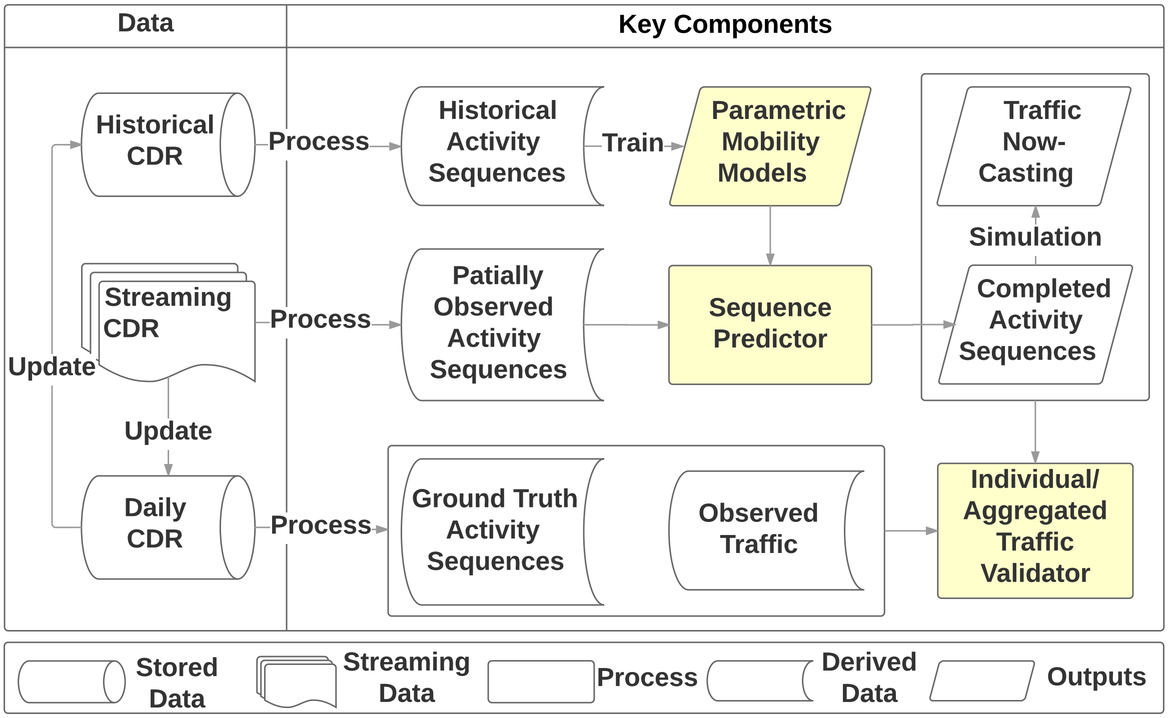
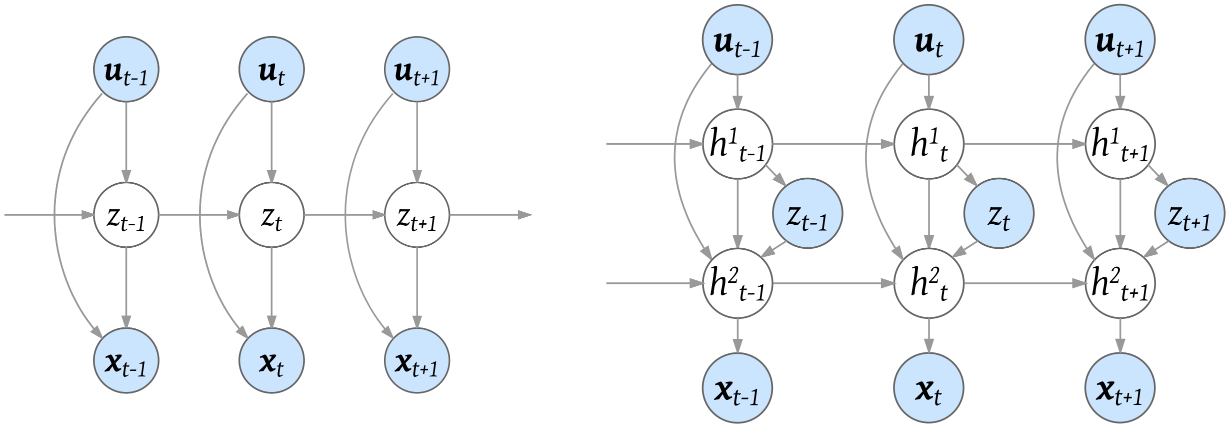


Figure 5.1: AM-PM Modeling framework diagram. The left column represents the input to the algorithms and the right column represents the model components. Our key contribution of improved deep urban mobility models, sequence predictor, and validation are shown in shaded yellow.

The developed data processing and modeling pipeline is presented in Fig. 5.1. Anonymized historical CDR data are processed to unlabeled historical activity sequences [135]. Urban mobility models are built upon these historical activity sequences. In this chapter, we improved the state-of-the-art urban mobility models by using deep LSTM models, as detailed in Chapter 5.4, and improving the model selection process by separating home and work activities into smaller sub-activities.

On a target day, we receive streaming CDR data at different times 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 5.6. 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.



(a) IOHMM

(b) LSTM

Figure 5.2: Comparison of Deep Urban Mobility Architectures, IOHMM (left) and LSTM (right). 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.

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 5.8. Finally, historical CDR database is updated with the new day's CDR, and urban mobility models can be updated and re-trained overnight.

5.4 Long Short Term Memory (LSTM) 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. 5.2b.

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 5.6. 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:

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

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 [18] and was further developed for handwriting synthesis tasks [57]. 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, $\hat{\boldsymbol{\pi}}_{\boldsymbol{t}}$ 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:

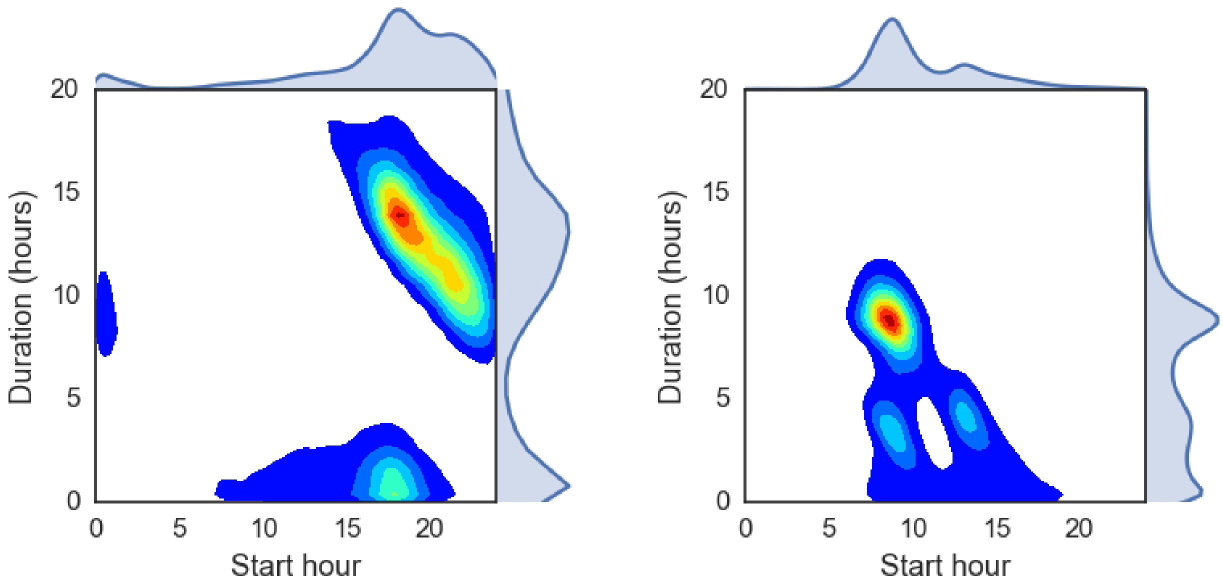
$$L_2\left(\boldsymbol{\theta_2}\right) \quad = \quad \sum_{t=1}^{T} -\log\sum_{i}^{M}\pi_t^i\mathcal{N}(\boldsymbol{x_t}|\hat{\boldsymbol{\mu}}_t^i, \hat{\boldsymbol{\sigma}}_t^i, \hat{\boldsymbol{\rho}}_t^i)$$

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. [81] 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 paper.

5.5 Model Specifications

Model selection for the IOHMM models includes the choice of hidden states. The choice should come directly from the collection of ground truth activities (Recall that we collected ground truth activities for “Food/Shop”, “Stop in Transit”, “Recreation”, “Personal Business”, and “Travel”.)



(a) Home

(b) Work

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

We further noticed a significant heterogeneity within home and work activities. The 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). It is easy to imagine that 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.
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 5.8.

5.6 Predictive Methodology with Cellular Data

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.

Prediction using IOHMM models

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 is longer than the observed travel time from the complete of the last activity to respect the fact that no new activities happen 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.

Activity generation

With the last activity inferred, the activity generation algorithm is same as what we have described in Chapter 4.9: 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. This is different from Chapter 4.9: in that chapter, the population is synthetic, we do not have location history of the user and thus can only generate a new destination from the choice set of TAZs and a random point within the TAZ. 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.

Prediction using LSTM models

The procedure is straightforward based on Fig. 5.2b. 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 [81, 57]. The rest of the generation process is similar to the generation process of IOHMM model.

5.7 Description of Data

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.

5.8 Experimental Results

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 Section 4.4.

3. **IOHMM-co-training-7**: The co-training IOHMM model specified in Section 4.4. 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 Section. 5.5. Thus there are 11 states in total.

5. **LSTM-3**: The LSTM model specified in Section 5.4. 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”.

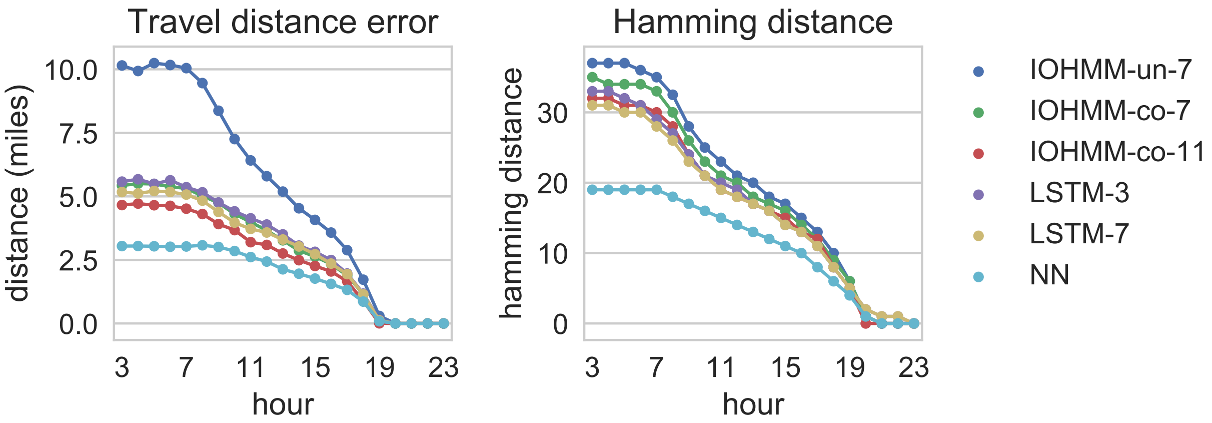


Figure 5.4: Models comparison. Two validation metrics are used: median travel distance error (left) and median Hamming distance (right). The x-axis is the prediction hour (cut hour) and the y-axis is the validation error. Each series of points represents the performance of a model.

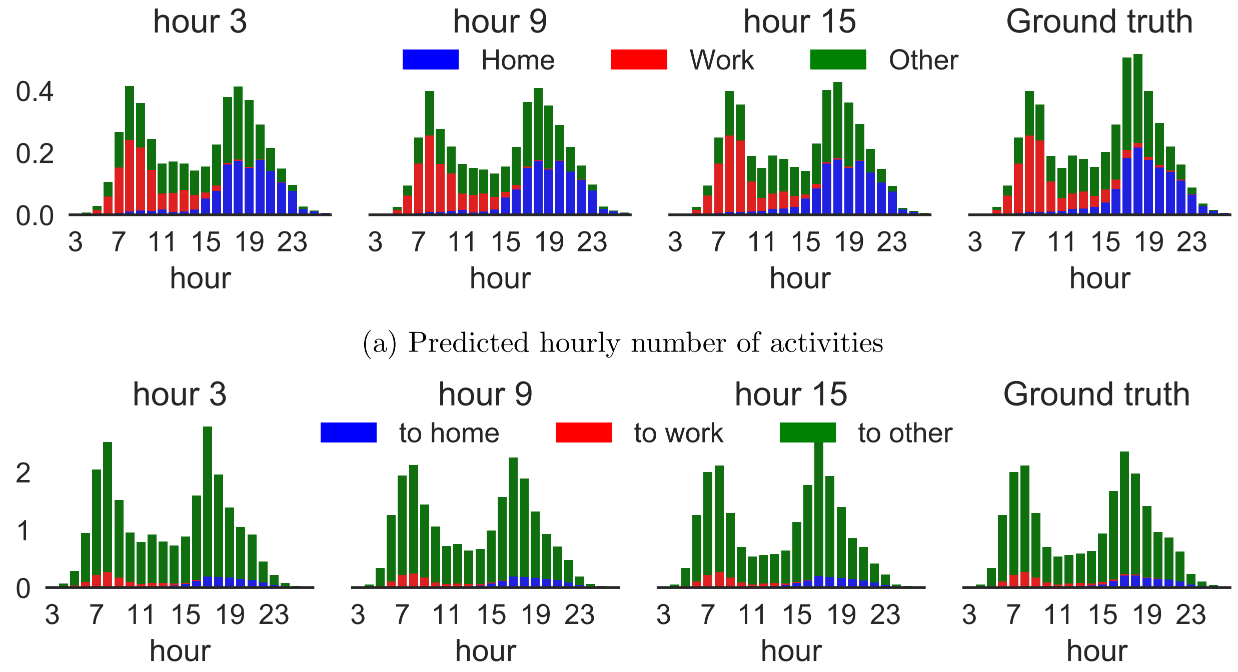
In Fig. 5.4, we plot how the two validation metrics, (1) median travel distance error (left), and, (2) median Hamming distance (right) change for different cut hours using different models. The travel distance error is calculated as the difference between the observed daily travel distance and predicted 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. 5.4, 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 proves 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.

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. 5.5a 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. 5.5b 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. 5.5a. 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 5.5: 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.

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. MATSim is a state-of-the-art agent based traffic micro-simulation tool that performs traffic assignment for the set of agents with pre-defined activity plans. 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. TABLE 5.1 summarizes the fit score (1) adjusted R^2; (2) mean absolute percentage error

(MAPE, %). Fig. 5.6 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 5.1: 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)

TABLE 5.1 proves that we can use observed information of the day to improve traffic volume prediction. The coefficient of determination increases and the MAPE decreases with the prediction hour. When we make prediction 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 [135] of higher variability in travel choices for secondary activities.

5.9 Conclusion

In this chapter, 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 LSTMs and finer model selection process. 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 chapter.

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.

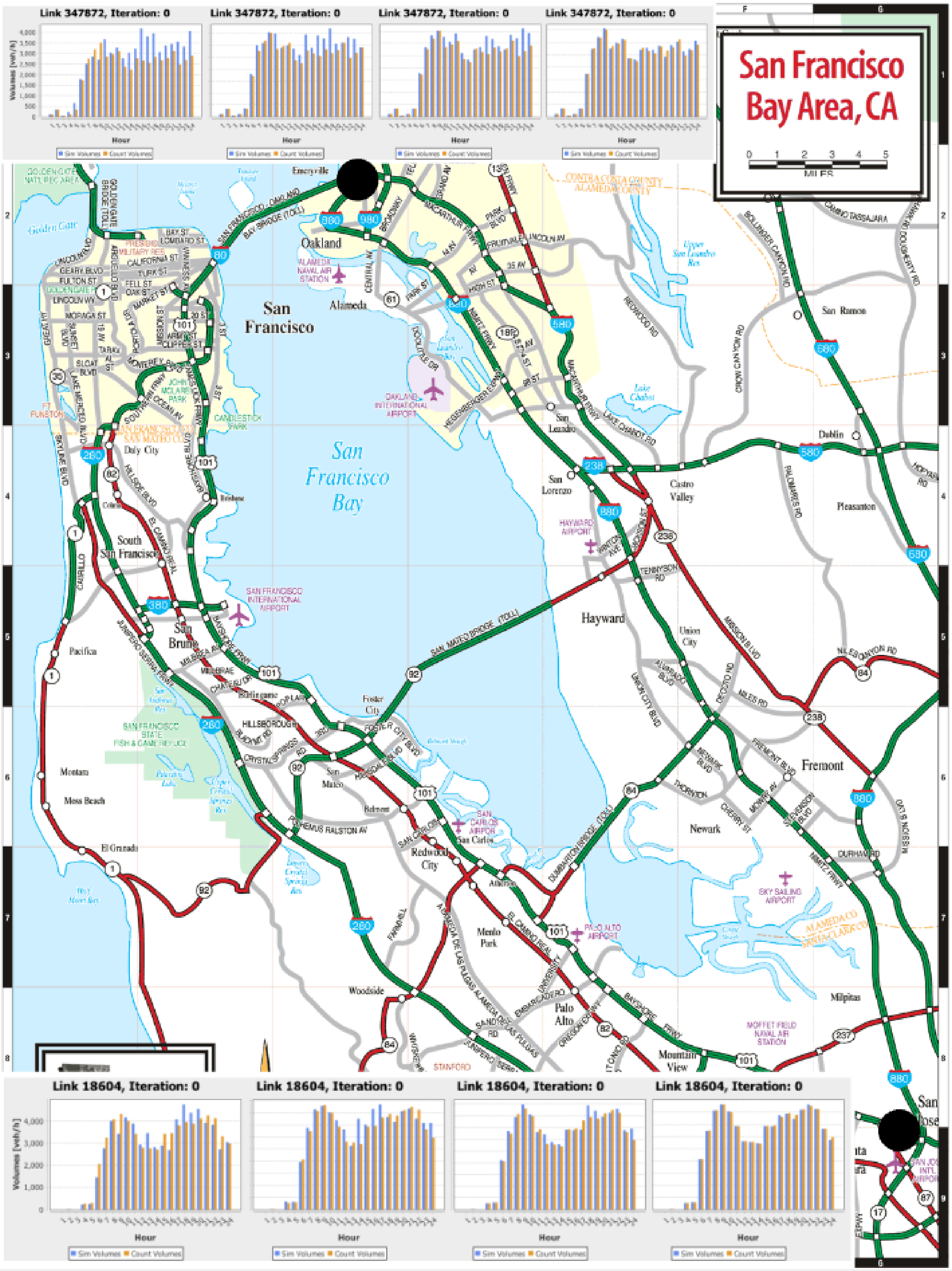


Figure 5.6: A fragment of the SF Bay Area road network. Inlet graphs illustrate two sample hourly vehicle volume profiles for observed (orange) and predicted (blue) at 3am, 9am, 3pm, and 9pm.

Chapter 6

Conclusions

Transportation is one of the defining challenges of our age. Our cities are getting more congested and less dependable. People are spending more time, paying more money, and causing more environmental externalities living their daily lives. To address the current problems and meet growing travel demand, one solution is improving the efficiency and effectiveness of the existing infrastructure. On the other hand, IT-reliant innovations have already made an important impact on transportation industry across cities and countries, such as navigation services such as Google maps and on-demand ridesharing services such as Lyft and Uber. These novel mobility paradigms change the transportation landscape quicker than traditional data sources, such as travel surveys, are able to reflect. Public agencies charged with a mandate to manage critical transportation infrastructures are slow to react to these changes, as they are reliant on out-dated information, tools and models.

The current manually collected travel surveys, such as The National Household Travel Survey (NHTS), usually happens every a few years. The cycle takes a long time due to the long data collection, cleaning, analytic, and modeling process. It also takes millions of dollars for a regional travel survey to take place. However, the quality of the data cannot be guaranteed. Based on our analysis, people may tend to under-report their daily travel diary due to privacy concerns. On the other hand, typical travel surveys covers only one percent of households in a metropolitan area, and typically only records a few day of travel per household. Therefore, no weekly travel pattern or long term effect can be captured.

Thanks to the ubiquitous sensor networks and location-based services, people generate data while traveling, just by carrying and using a mobile phone. Therefore, it is appealing to use data from such services as a substitute for manual surveying. Cellular data vastly increase the population coverage. Since the data are continuously collected, weekly travel patterns and long term changes can also be analyzed. The data processing and modeling pipeline can be automated so the delay normally associated with administering and processing travel surveys is eliminated. The spatially and temporally rich cell phone traces can support models of the locational and temporal activity choices.

In this dissertation, we investigated the use of cellular data to complement traditional travel surveys for activity-based travel demand models. We developed scalable and interpretable generative activity-based urban mobility models for regional mobility analysis from cellular data. We also presented a direct application of the trained activity-based urban mobility model for medium term travel demand nowcasting, which is a missing piece in the literature. The main findings and contributions are summarized as follows.

6.1 Findings and Contributions

Inference of activity types and analysis of activity patterns from cellular data are non-trivial tasks. Cellular data, while collected at large scale, suffers from low spatial and temporal resolutions. Temporally, cellular data might contain gaps where important information may be missing. Spatially, cellular data is triangulated so the locations are not exact. In addition the cell tower might be switched due to capacity constraints so the user is observed at multiple cell towers even when she is standing still. In chapter 3, we summarized our lessons learned from processing noisy cellular data. After the initial spatial clustering, we examined the cases where the user was standing still but was observed to have moved. To do so, we constructed an oscillation graph of the clusters for each user where the edges indicates the probability of oscillation between the clusters. In this case, some short activities due to tower oscillations can be merged as a single activity so that we will not over-filter short-term travel and miss many activities. Along the processing pipeline, we also found that the daily skeleton we processed from the cellular data match with the statistics from a travel survey. These statistics are important in helping us understand how people construct their daily lives. For example, there is a strong substitution effect between evening-commute activities and post-home activities. If a person pursues an evening-commute activity, she is much less like to have a post-home activity and vise versa. This substitution effect is not observed between work-based activities and evening-commute/post-home activities because they are usually for different purposes. Another finding is that people who have evening-commute or post-home activities tend to leave work earlier, at around 4:30pm, while people who do not have evening-commute or post-home activities tend to leave work later, at around 6:10pm.

In Chapter 4, we found a way to collect ground truth activities using data from short range distributed antenna systems (DAS). We found that DASs are very helpful in identifying activity types thanks to their high spatial resolution. The visits to these DASs are very structured in the temporal dimension too. We developed several scalable and interpretable activity-based models for regional mobility analysis from cellular data. As an illustration, we inferred the activity patterns including primary, secondary activities and heterogeneous activity transitions of a set of anonymized San Francisco Bay Area commuters using unsupervised and semi-supervised generative state-space models. We validated the inferred activities with a set of ground truth activities based on the DASs. We found that the semi-supervised co-training model has the best classification performance. This semi-supervised co-training model preserves the generative power of the IOHMM model and the classification power of the decision tree model. We also confirmed the advantages of IOHMM over standard HMM where IOHMM can incoporate more contextual information. For example, if a user is observed to go to some activity from home, if it is a weekday morning, she is likely to go to work but if it is a weekend evening, it is likely to be for food or recreation. We also found that the marginal proportion of each activity type is consistent with the 2015 Travel Decisions Surveys. To test the generative power of the IOHMM model, we synthesized travel plans for each agent with home and work locations sampled from census data. An agent-based microscopic traffic simulation was conducted to compare the resulting traffic with the observed volumes of vehicular traffic flow in the regional road network on an average weekday. We found that the fit was quite reasonable, with a coefficient of determination of 0.81.

In Chapter 5, we presented a direct application of the activity-based urban mobility model. We proposed a medium term travel demand nowcasting problem which predicts the traffic conditions hours to days ahead. This medium term travel demand nowcasting problem fills a gap in the literature, which mainly studies long term travel demand forecast (months to years ahead) and extremely short term traffic prediction (seconds to minutes ahead). In this chapter, we discussed several ways to improve the urban mobility models specifically for this application. One improvement we made was to separate home and work activities into smaller sub-activities, such as overnight home activities and short stay at home before going for another activity. By separating these home and work activities, we got better context-dependent transition probabilities. For example, the transition from morning work activity to recreation should not be as strong as the transition from an afternoon work activity to recreation. Another improvement we made was to make the IOHMM model deeper and continuous in hidden state space. We developed a LSTM urban mobility model and compared it with the IOHMM model. A non-parametric individualized nearest neighbor model was explored as the practical limit of predictability of individuals daily travel. This nearest neighbor model should have the best predictive performance since it is fully personalized. 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, 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. In terms of validation of traffic volumes, our experiments show promising results of medium term forecast. We have reached a mean absolute percentage error (MAPE) of less than 5% one hour ahead and 10% three hours ahead. Results also show that we can improve the prediction accuracy by incorporating more of the observed data by the time of prediction. It is worth mentioning that our prediction of traffic conditions is not restricted to area where high-resolution freeway detectors data are available since our main data source is the cellular data which is more ubiquitous.

6.2 Directions for Future Research

This dissertation took the first step in using cellular data to support activity-based travel demand models. This approach should be further developed in the following directions:

Travel modes

This dissertation using cellular data to inform activity-based travel demand models have achieved a good understanding of the activity (trip purpose) patterns. However, the missing piece is travel mode and route inferences. We have left the mode and route choice to our micro-simulator MATSim after the activity types, timing and locations are determined. Ideally, these mode and route choices should also be informed from the cellular data. Moreover, a discrete choice model (DCM) based on the inferred travel mode could be trained so that such model can be directly used for transportation planning. However, this mode detection and travel mode classification task is non-trivial. To develop a discriminative classifier that detects the mode of the observed trips or a sequence of modes in a multiple leg journey, we need a considerable amount of ground truth data with known modes to be available for training. Such a classifier also requires a k-shortest path algorithm that generates plausible alternatives routes for the journey. On the other hand, these discriminative travel mode classifiers might not be spatially transferable since certain location features might occur in the input vector. Therefore, it would be better to utilize the discriminative recognition step of the observed mode in order to build a behaviorally grounded model that predicts the chosen mode within a set of available alternatives as a function of user characteristics and transportation system variables. It would be based on the discrete choice modeling paradigm and results in a set of parameters calibrated for distinct neighborhoods and/or segments of population.

Location choice model

With privacy concerns and data limitations in mind, the location choice model implemented in this dissertation is relatively simple. We use only the distance to home and distance to work as the features people consider when choosing activity locations. This is a compromise when we train a model across a group of users. If we do not use these location features but the exact location/cluster instead, we will face privacy concerns as well as feature sparsity issues. Though these two features are sufficient in helping us recognize activities, they are relatively simple when sampling locations in new sequences. Certain improvements can be made to the location choice model without sacrificing privacy. Future work may incorporate a discrete choice model on a set of TAZs so that locations can be directly sampled when generating activity sequences.

Population heterogeneity

Partitioning a population into sub-groups (whether socially or spatially) for shared parameter modeling is a partly open problem. Currently we approached it by defining rules to identify groups of a similar day structure, and applying geographic constraints. The underlying assumption is that people within the same group tend to have similar activity patterns. However, if we define a group that is too small, we have to train too many models to cover the whole population and this is likely to sacrifice user privacy. On the other hand, if we define a group that is too large, we may not be able to capture heterogeneity of activity patterns that 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 dissertation. Future works may replace these geographic constrains by an alternative specification that involves a mixture of IOHMM models. This mixture IOHMM has the following architecture: each user belongs to a lifestyle (activity pattern) which is a hidden state; each life style has a corresponding IOHMM model that represents its activity profiles. After a training process, we will be able to understand what is the lifestyle of a user (or what is the probability the user belongs to each lifestyle) and what is the IOHMM model for each lifestyle. The training process might incorporate a hierarchical EM inference.

Better urban mobility model

In this dissertation, we have improved the state-of-the-art activity-based urban mobility model. We have shown how IOHMM outperforms the partial IOHMM which outperforms the vanilla HMM model. We further improved the recognition accuracy by using semi-supervised co-training which adds a discriminative decision tree to the training process. In this way, the learning process is directed and will not free flow too far. To learn better activity transitions for better activity generation, we separated home and work activities into smaller sub-activities. We showed that the activity prediction under this finer schema has lower error. In addition, we proposed a LSTM-based urban mobility model, which goes deeper and continuous in hidden state spaces. This urban mobility model can better capture the temporal structure and long term dependencies thus predicts the day structure better. 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. On the other hand, this LSTM model has lower interpretability than the IOHMM model; future research may improve the interpretability of the LSTM based model while preserving its deeper architecture.

Utility-based urban mobility model

In this dissertation, we make our first attempt to complement traditional travel surveys with cellular data. The mainstream of the research using traditional travel surveys are utility-based models. These models have the capability to model activity types and travel modes all together. These utility-based models can also be used to generate activities and trips directly. However, these models are usually based on the assumptions that there are no uncertainties in information about activity types and travel modes, as is true for travel surveys. Future research may study an end-to-end utility based urban mobility models that is capable of modeling activity types and travel modes in a unified framework, and is capable of dealing with the uncertainties in cellular data.

6.3 Concluding Remarks

The development of travel demand models requires the increasing availability and better quality of travel data. With data from travel surveys, researchers develop activity-based travel demand models, which outperforms traditional tour-based travel demand models with more consistent sub-models and more detailed evaluation metrics. However, these traditional travel surveys suffer from expensive processes of data collection and cleaning which make it difficult to keep models up to date. Thanks to the ubiquitous mobile phone data, we see an opportunity to improve the travel demand models into a new stage. Cellular data is rich in time and space, and is collected continuously and pervasively. The process of cleaning and analyzing it can be automated, making the time and cost of using cellular data less expensive.

In this study, we present an end-to-end research pipeline from processing raw cellular data, building activity-based urban mobility models, to validating the model performances with independent data sources. Though cellular data is the main data source in this study, traditional travel surveys play an important role in the following ways: first, statistics from travel survey such as the number of activities per person per day are the main reference points for tuning/validation of hyper-parameters in the pre-processing step. Second, the user population distribution of a certain mobile carrier may not be the same as the true population distribution. A rescaling/resampling step based on travel surveys and population data is required as part of any travel study based on cell phone data. Finally, we validate our activity recognition results with travel survey data to ensure that the marginal distribution of activities corresponds well within the two data sources. It is worth noting that on the one hand, our study cannot be solid without traditional travel surveys. On the other hand, our main modeling modules are independent with the data from traditional surveys: it is mainly used in hyper-parameter tuning, rescaling and validation.

For future works, we imagine two branches can emerge from this study. One branch could propose a unified utility-based urban mobility model that benefit from data fusion of traditional travel survey and cellular data. Traditional travel surveys are not only used for rescaling/validation purposes, but used in the actual modeling phase. The other branch could explore ways of reducing the sample size from the traditional travel survey (or totally getting rid of travel surveys). These potential improvements are bound to make substantial impacts on urban and transportation planing, and improve the travel demand models into a new stage.

Bibliography

[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] Rein Ahas et al. "Using mobile positioning data to model locations meaningful to users of mobile phones". In: *Journal of urban technology* 17.1 (2010), pp. 3–27.

[3] Sherif Akoush and Ahmed Sameh. "Mobile user movement prediction using bayesian learning for neural networks". In: *Proceedings of the 2007 international conference on Wireless communications and mobile computing*. ACM. 2007, pp. 191–196.

[4] Ian Anderson and Henk Muller. "Practical Activity Recognition using GSM Datafffdfffdfffd". In: ().

[5] 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.

[6] Akinori Asahara et al. "Pedestrian-movement prediction based on mixed Markov-chain model". In: *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*. ACM. 2011, pp. 25–33.

[7] 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.

[8] Michael Balmer, Kay Axhausen, and Kai Nagel. "Agent-based demand-modeling framework for large-scale microsimulations". In: *Transportation Research Record: Journal of the Transportation Research Board* 1985 (2006), pp. 125–134.

[9] Michael Balmer and K Meister. *Agent-based simulation of travel demand: Structure and computational performance of MATSim-T*. 2008.

[10] 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.

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

[12] Paul Baumann, Wilhelm Kleiminger, and Silvia Santini. “The influence of temporal and spatial features on the performance of next-place prediction algorithms”. In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM. 2013, pp. 449–458.

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

[14] Moshe Ben-Akiva and Bruno Boccara. “Discrete choice models with latent choice sets”. In: *International Journal of Research in Marketing* 12.1 (1995), pp. 9–24.

[15] Yoshua Bengio and Paolo Frasconi. “An input output HMM architecture”. In: (1995).

[16] Linus Bengtsson et al. “Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti”. In: *PLoS medicine* 8.8 (2011), p. 1128.

[17] 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.

[18] Christopher M Bishop. “Mixture density networks”. In: (1994).

[19] 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.

[20] 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.

[21] John L Bowman, Mark A Bradley, and J Gibb. “The Sacramento activity-based travel demand model: estimation and validation results”. In: *European Transport Conference*. 2006.

[22] N Caceres, JP Wideberg, and FG Benitez. “Deriving origin destination data from a mobile phone network”. In: *Intelligent Transport Systems, IET* 1.1 (2007), pp. 15–26.

[23] Francesco Calabrese et al. “The geography of taste: analyzing cell-phone mobility and social events”. In: *International conference on pervasive computing*. Springer. 2010, pp. 22–37.

[24] 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.

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

[26] Bo Chen and Harry H Cheng. “A review of the applications of agent technology in traffic and transportation systems”. In: *Intelligent Transportation Systems, IEEE Transactions on* 11.2 (2010), pp. 485–497.

[27] 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.

[28] Peng Cheng, Zhijun Qiu, and Bin Ran. “Particle filter based traffic state estimation using cell phone network data”. In: *Intelligent Transportation Systems Conference, 2006. ITSC’06. IEEE*. IEEE. 2006, pp. 1047–1052.

[29] 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.

[30] Yohan Chon et al. “Evaluating mobility models for temporal prediction with high-granularity mobility data”. In: *Pervasive computing and communications (PerCom), 2012 IEEE international conference on*. IEEE. 2012, pp. 206–212.

[31] Pierre Deville et al. “Dynamic population mapping using mobile phone data”. In: *Proceedings of the National Academy of Sciences* 111.45 (2014), pp. 15888–15893.

[32] Trinh Minh Tri Do and Daniel Gatica-Perez. “Contextual conditional models for smartphone-based human mobility prediction”. In: *Proceedings of the 2012 ACM conference on ubiquitous computing*. ACM. 2012, pp. 163–172.

[33] 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.

[34] Yuxiao Dong et al. “Inferring user demographics and social strategies in mobile social networks”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 15–24.

[35] John Doyle et al. “Utilising mobile phone billing records for travel mode discovery”. In: (2011).

[36] 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.

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

[38] Nathan Eagle, Alex Sandy Pentland, and David Lazer. “Inferring friendship network structure by using mobile phone data”. In: *Proceedings of the National Academy of Sciences* 106.36 (2009), pp. 15274–15278.

[39] Vincent Etter et al. “Where to go from here? Mobility prediction from instantaneous information”. In: *Pervasive and Mobile Computing* 9.6 (2013), pp. 784–797.

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

[41] Katayoun Farrahi and Daniel Gatica-Perez. “A probabilistic approach to mining mobile phone data sequences”. In: *Personal and ubiquitous computing* 18.1 (2014), pp. 223–238.

[42] Katayoun Farrahi and Daniel Gatica-Perez. “Discovering human routines from cell phone data with topic models”. In: *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on.* IEEE. 2008, pp. 29–32.

[43] 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.

[44] Katayoun Farrahi and Daniel Gatica-Perez. “What did you do today?: discovering daily routines from large-scale mobile data”. In: *Proceedings of the 16th ACM international conference on Multimedia.* ACM. 2008, pp. 849–852.

[45] Stephen E. Fienberg. “An Iterative Procedure for Estimation in Contingency Tables”. In: *The Annals of Mathematical Statistics* 41.3 (1970), pp. 907–917. ISSN: 00034851. DOI: 10.2307/2239244. URL: http://dx.doi.org/10.2307/2239244.

[46] Lino Figueiredo et al. “Towards the development of intelligent transportation systems”. In: *Intelligent transportation systems.* Vol. 88. 2001, pp. 1206–1211.

[47] 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.

[48] Barbara Furletti et al. “Identifying users profiles from mobile calls habits”. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing.* ACM. 2012, pp. 17–24.

[49] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. “Next place prediction using mobility markov chains”. In: *Proceedings of the First Workshop on Measurement, Privacy, and Mobility.* ACM. 2012, p. 3.

[50] Huiji Gao, Jiliang Tang, and Huan Liu. “Mobile location prediction in spatio-temporal context”. In: *Nokia mobile data challenge workshop.* Vol. 41. 2012, p. 44.

[51] Győző Gidófalvi and Fang Dong. “When and where next: individual mobility prediction”. In: *Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems.* ACM. 2012, pp. 57–64.

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

[53] Joao Bártolo Gomes, Clifton Phua, and Shonali Krishnaswamy. “Where will you go? mobile data mining for next place prediction”. In: *International Conference on Data Warehousing and Knowledge Discovery*. Springer. 2013, pp. 146–158.

[54] 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.

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

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

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

[58] Ramaswamy Hariharan and Kentaro Toyama. “Project Lachesis: parsing and modeling location histories”. In: *Geographic Information Science*. Springer, 2004, pp. 106–124.

[59] David T Hartgen and Elizabeth San Jose. “Costs and Trip Rates of Recent Household Travel Surveys”. In: (2009).

[60] Tom Hunter, Pieter Abbeel, and Alexandre Bayen. “The path inference filter: model-based low-latency map matching of probe vehicle data”. In: *Intelligent Transportation Systems, IEEE Transactions on* 15.2 (2014), pp. 507–529.

[61] T Hunter et al. “Trajectory reconstruction of noisy GPS probe vehicles in arterial traffic”. In: *preparation for IEEE Transactions on Intelligent Transport Systems* (2011).

[62] Md Shahadat Iqbal et al. “Development of origin–destination matrices using mobile phone call data”. In: *Transportation Research Part C: Emerging Technologies* 40 (2014), pp. 63–74.

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

[64] Jerald Jariyasunant et al. “Quantified traveler: Travel feedback meets the cloud to change behavior”. In: *Journal of Intelligent Transportation Systems* 19.2 (2015), pp. 109–124.

[65] Hoyoung Jeung et al. “A hybrid prediction model for moving objects”. In: *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. IEEE. 2008, pp. 70–79.

[66] 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.

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

[68] Ilkcan Keles et al. "Location prediction of mobile phone users using apriori-based sequence mining with multiple support thresholds". In: *New Frontiers in Mining Complex Patterns*. Springer, 2014, pp. 179–193.

[69] 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.

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

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

[72] Kari Laasonen. "Clustering and prediction of mobile user routes from cellular data". In: *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer. 2005, pp. 569–576.

[73] US Bureau of Labor Statistics. *AMERICAN TIME USE SURVEY fffdfffdfffd 2015 RESULTS*. Tech. rep. June 2016.

[74] Cristian-Liviu Leca, Ioan Nicolaescu, and Cristian-Iulian Rîncu. "Significant location detection & prediction in cellular networks using artificial neural networks". In: *Computer Science and Information Technology* 3.3 (2015), pp. 81–89.

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

[76] Ilias Leontiadis et al. "From cells to streets: Estimating mobile paths with cellular-side data". In: *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*. ACM. 2014, pp. 121–132.

[77] Hongjun Li et al. "Hotspot district trajectory prediction". In: *Web-Age Information Management*. Springer, 2010, pp. 74–84.

[78] 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.

[79] Lin Liao, Dieter Fox, and Henry Kautz. "Hierarchical conditional random fields for GPS-based activity recognition". In: *Robotics Research*. Springer, 2007, pp. 487–506.

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

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

[82] Shiang-Chun Liou and Hsuan-Chia Lu. “Applied neural network for location prediction and resources reservation scheme in wireless networks”. In: *Communication Technology Proceedings, 2003. ICCT 2003. International Conference on*. Vol. 2. IEEE. 2003, pp. 958–961.

[83] Feng Liu et al. “Annotating mobile phone location data with activity purposes using machine learning algorithms”. In: *Expert Systems with Applications* 40.8 (2013), pp. 3299–3311.

[84] George Liu and Gerald Maguire Jr. “A class of mobile motion prediction algorithms for wireless mobile computing and communication”. In: *Mobile Networks and Applications* 1.2 (1996), pp. 113–121.

[85] Xin Lu, Linus Bengtsson, and Petter Holme. “Predictability of population displacement after the 2010 Haiti earthquake”. In: *Proceedings of the National Academy of Sciences* 109.29 (2012), pp. 11576–11581.

[86] Xin Lu et al. “Approaching the limit of predictability in human mobility”. In: *Scientific reports* 3 (2013).

[87] Zhongqi Lu et al. “Next Place Prediction by Learning with Multiple Models”. In: *Proceedings of the Mobile Data Challenge Workshop*. 2012.

[88] Mingqi Lv, Ling Chen, and Gencai Chen. “Mining user similarity based on routine activities”. In: *Information Sciences* 236 (2013), pp. 17–32.

[89] Tai-Yu Ma et al. “Multistate nonhomogeneous semi-markov model of daily activity type, timing, and duration sequence”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2134 (2009), pp. 123–134.

[90] Wesley Mathew, Ruben Raposo, and Bruno Martins. “Predicting future locations with hidden Markov models”. In: *Proceedings of the 2012 ACM conference on ubiquitous computing*. ACM. 2012, pp. 911–918.

[91] Erik Mellegard, Simon Moritz, and Mohamed Zahoor. “Origin/Destination-estimation using cellular network data”. In: *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE. 2011, pp. 891–896.

[92] Eric J Miller. “Microsimulation and activity-based forecasting”. In: *Activity-Based Travel Forecasting Conference*. 1997.

[93] Eric J Miller and Paul A Salvini. “Activity-based travel behavior modeling in a microsimulation framework”. In:

[94] 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.

[95] Mert Ozer et al. “Predicting the change of location of mobile phone users”. In: *Proceedings of the Second ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*. ACM. 2013, pp. 43–50.

[96] Smita Parija et al. “Location Prediction of Mobility Management Using Soft Computing Techniques in Cellular Network”. In: *International Journal of Computer Network and Information Security (IJCNIS)* 5.6 (2013), p. 27.

[97] *Percent of population who worked on weekdays and weekend days.* `https://www.bls.gov/tus/charts/chart11.pdf`.

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

[99] Pratap S Prasad and Prathima Agrawal. “Movement prediction in wireless networks using mobility traces”. In: *Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE.* IEEE. 2010, pp. 1–5.

[100] 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.

[101] Jonathan Reades, Francesco Calabrese, and Carlo Ratti. “Eigenplaces: analysing cities using the space-time structure of the mobile phone network”. In: *Environment and Planning B: Planning and Design* 36.5 (2009), pp. 824–836.

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

[103] 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.

[104] San Francisco Municipal Transportation Agency (SFMTA). *Travel Decisions Survey 2015.* Tech. rep. 2015.

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

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

[107] David Schrank, Bill Eisele, and Tim Lomax. “TTI’s 2012 urban mobility report”. In: *Texas A&M Transportation Institute. The Texas A&M University System* (2012).

[108] Nadine Schuessler and Kay Axhausen. “Processing raw data from global positioning systems without additional information”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2105 (2009), pp. 28–36.

[109] Katie Shilton. “Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection”. In: *Communications of the ACM* 52.11 (2009), pp. 48–53.

[110] Keemin Sohn and Daehyun Kim. “Dynamic origin–destination flow estimation using cellular communication system”. In: *Vehicular Technology, IEEE Transactions on* 57.5 (2008), pp. 2703–2713.

[111] Timothy Sohn et al. “Mobility detection using everyday gsm traces”. In: *International Conference on Ubiquitous Computing*. Springer. 2006, pp. 212–224.

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

[113] 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.

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

[115] Leon Stenneth et al. “Transportation mode detection using mobile phones and GIS information”. In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM. 2011, pp. 54–63.

[116] Harald Sterly, Benjamin Hennig, and Kouassi Dongo. “”Calling Abidjan”–Improving Population Estimations with Mobile Communication Data (IPEMCODA)”. In: *Mobile Phone Data for Development-Analysis of mobile phone datasets for the development of Ivory Coast*. 2013, pp. 108–114.

[117] Peter Stopher et al. “Deducing mode and purpose from GPS data”. In: *Institute of Transport and Logistics Studies* (2008), pp. 1–13.

[118] Andrew J Tatem et al. “Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning”. In: *Malaria journal* 13.1 (2014), p. 52.

[119] Andrew J Tatem et al. “The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents”. In: *Malar J* 8.287 (2009), pp. 10–1186.

[120] 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.

[121] Jameson L Toole et al. “The path most travelled: Mining road usage patterns from massive call data”. In: *arXiv preprint arXiv:1403.0636* (2014).

[122] Le Hung Tran et al. “Next place prediction using mobile data”. In: *Proceedings of the Mobile Data Challenge Workshop (MDC 2012)*. EPFL-CONF-182131. 2012.

[123] 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.

[124] Huayong Wang et al. “Transportation mode inference from anonymized and aggregated mobile phone call detail records”. In: *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. IEEE. 2010, pp. 318–323.

[125] Jingwei Wang et al. “User Travelling Pattern Prediction via Indistinct Cellular Data Mining”. In: *Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC)*. IEEE. 2013, pp. 17–24.

[126] Pu Wang et al. “Understanding road usage patterns in urban areas”. In: *Scientific reports* 2 (2012), p. 1001.

[127] Peter Widhalm, Philippe Nitsche, and Norbert Brändie. “Transport mode detection with realistic smartphone sensor data”. In: *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE. 2012, pp. 573–576.

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

[129] Wikipedia. *Google Traffic — Wikipedia, The Free Encyclopedia*. [Online; accessed 7-August-2015]. 2015. URL: `https://en.wikipedia.org/w/index.php?title=Google_Traffic&oldid=673924809`.

[130] Jean Wolf, Randall Guensler, and William Bachman. “Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data”. In: *Transportation Research Record: Journal of the Transportation Research Board* 1768 (2001), pp. 125–134.

[131] Steve Yadlowsky et al. *Link Density Inference from Cellular Infrastructure*. Tech. rep. 2015.

[132] Farhana Yasmin, Catherine Morency, and Matthew J Roorda. “Macro-, meso-, and micro-level validation of an activity-based travel demand model”. In: *Transportmetrica A: Transport Science* 13.3 (2017), pp. 222–249.

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

[134] 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.

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

[136] 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.

[137] Daqiang Zhang et al. “NextCell: Predicting location using social interplay from cell phone traces”. In: *Computers, IEEE Transactions on* 64.2 (2015), pp. 452–463.

[138] 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.

[139] 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.

[140] Yu Zheng and Xing Xie. “Location-based social networks: Locations”. In: *Computing with Spatial Trajectories*. Springer, 2011, pp. 277–308.

[141] 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.

[142] 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.

[143] 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.

[144] Xiaojin Zhu. “Semi-supervised learning literature survey”. In: (2005).