Title
Recurrent Neural Network Models of Human Mobility

Author
Lin, Ziheng

Publication Date
2018

Peer reviewed|Thesis/dissertation

Recurrent Neural Network Models of Human Mobility

by

Ziheng Lin

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

in

Engineering - Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alexei Pozdnukhov, Chair
Professor Joan Walker
Professor Paul Grigas

Spring 2018

Recurrent Neural Network Models of Human Mobility

Abstract

Recurrent Neural Network Models of Human Mobility

by

Ziheng Lin

Doctor of Philosophy in Engineering - Civil and Environmental Engineering

University of California, Berkeley

Professor Alexei Pozdnukhov, Chair

Locational data generated by mobile devices present an opportunity to substantially simplify methodologies and reduce analysis latencies in short-term transportation planning applications. Short-term transportation planning, such as traffic flow management or traffic demand management, requires accurate prediction of daily network congestion levels and the congestion contributors. The existing human mobility models using locational data have focused on predicting next activities, and many models limited the prediction to only temporal features or only spatial features that they cannot be directly applied to such applications. In this dissertation, we propose Long Short Term Memory (LSTM) models for learning and predicting human mobility sequences using mobile locational data. The major contributions of this dissertation include the following: first, we developed the LSTM mobility models that are capable of learning and predicting the entire mobility sequences within a time window of interest; second, we developed the LSTM mobility models that are able to predict activity sequences with activity type choices and explicit spatial-temporal choices; third, the LSTM mobility models are able to capture long-term activity dependencies. The LSTM models can be applied for transportation demand forecasting problems, including typical-day activity prediction, medium-term activity prediction, and activity prediction with social-demographic information. We performed validation through micro-simulation and compared the simulation results to real-world traffic counts. The results showed high similarities between generated traffic volumes and observed traffic volumes. The performance of LSTM models was also compared against baseline sequence models including Hidden Markov Models and a nearest neighbour model. Using daily activity structure and daily travel distance as metrics, we observed better performance of LSTM models due to the capability of learning long-term activity dependencies. Lastly, we extended the LSTM mobility models for learning activity sequences with contextual information. We demonstrated the capability of the LSTM models to handle both discrete and continuous contextual information.

To my parents.

Contents

Contents **ii**

List of Figures **iv**

List of Tables **vi**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Problem Statement . . . 2
1.3 Contribution . . . 2
1.4 Dissertation Outline . . . 3

2 Background **4**
2.1 Human Mobility Data Sources . . . 4
2.2 Predictability of Human Mobility . . . 5
2.3 Travel Demand Models . . . 5
2.4 Models of Sequences . . . 6
2.5 Fully-Connected and Convolutional Neural Network for Sequences . . . 7
2.6 Vanilla Recurrent Neural Network and Gradient Vanishing Problem . . . 7
2.7 Long-Short Term Memory . . . 8
2.8 Mixture Density Networks . . . 10

3 LSTM Mobility Model **12**
3.1 Introduction . . . 12
3.2 Existing Work . . . 13
3.3 Model Framework . . . 15
3.4 Model Structure . . . 16
3.5 Learning Long-term Dependencies . . . 23
3.6 Experiment Setup . . . 23
3.7 Results . . . 24
3.8 Discussion . . . 26

4 LSTM Models for Medium-term Human Activity Prediction **31**

4.1 Introduction . . . 31
4.2 Existing Work . . . 32
4.3 Model Structure . . . 33
4.4 Evaluation Method . . . 37
4.5 Experiment Setup . . . 37
4.6 Results . . . 40
4.7 Discussion . . . 42

5 LSTM Models with Contextual Information **44**
5.1 Introduction . . . 44
5.2 Existing Work . . . 45
5.3 Model Structure . . . 47
5.4 Experiment Setup . . . 51
5.5 Results . . . 53
5.6 Discussion . . . 58

6 Conclusion **60**
6.1 Summary . . . 60
6.2 Future Research Direction . . . 61

Bibliography **62**

List of Figures

2.1 LSTM Cell [34] 10

3.1 Model framework 16
3.2 LSTM Model Structures 17
3.3 Mixture density outputs of LSTM model when generating a day sequence with 5 activities. The model is trained on a made-up data set for illustration purpose only. (a) and (b) show the output probability distributions for activities labeled as 0 to 4 in latitude-longitude and starting-time-duration space, respectively. (c) shows the probabilities of activity types and end-of-day probabilities, which are averaged over all the mixture components. 21
3.4 An example of LSTM model learning long-term dependencies in sequences. . . . 27
3.5 Number of activities (labeled per highest posterior probability) by their respective start time within a course of 5 weekdays. 28
3.6 Joint distribution plot of duration and start hour of labeled activities. 28
3.7 Actual and simulated boarding and alighting counts on 2 major BART stations. 29
3.8 Micro-simulation validation. 29
3.9 A fragment of the multi-modal SF Bay Area network with sample traffic volume detectors and transit stations used for validation. Inset graphs illustrate three sample hourly vehicle volume profiles for observed (orange) and modeled (blue) flows on a typical weekday in June 2015. Sample transit counts histograms are shown in Figure 3.7. 30

4.1 LSTM Mobility Model Two-Layer Structure 34
4.2 Two layer model sampling a work related activity (top) and a secondary activity (bottom). 38
4.3 An example of minor differences in 2 activity sequences (top) and their corresponding string representations (bottom). 39
4.4 An example shows different location choices that affects the total traveling distance in a day. 39

4.5 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. 40
4.6 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. 42
4.7 Comparing individual level LSTM model to nearest neighbour model in terms of median hamming distance (left) and median travel distance (right). 43

5.1 LSTM Mobility Model with Contextual Information 47
5.2 Histogram of radius of gyration, travel distance, and activity start time for both "Low Radius" and "High Radius" groups. 52
5.3 Number of individuals in each numerical annual income level. 54
5.4 Comparing cumulative probability of radius of gyration (left) and traveling distance (right) of generated and observed daily traveling activities. 55
5.5 Comparing cumulative probability of activity start time for "Low Radius" group (left) and "High Radius" group (right). 55
5.6 Comparing place visit frequency between generated activity sequences and observed activity sequences for "Low Radius" group (left) and "High Radius" group (right) 57
5.7 Radius of gyration (left) and travel distance (right) of generated trajectories vs. radius of gyration input to the LSTM model as contextual information. Solid black lines represent the median. Dashed black lines represent the 25 and 75 percentile. The red line on the left plot represents $y = x$. The red line on the right plot represents $y = 2.9x + 6$. 58
5.8 Radius of gyration of generated trajectories (left) and observed trajectories (right) vs. numerical income level. 58
5.9 Travel distance of generated trajectories (left) and observed trajectories (right) vs. numerical income level. 59

List of Tables

5.1 Annual income level label 54
5.2 Jensen–Shannon divergence for radius of gyration 56
5.3 Jensen–Shannon divergence for travel distance. 56
5.4 Jensen–Shannon divergence for activity start time. 56

Chapter 1

Introduction

1.1 Motivation

From the exponential growth of on-demand shared ride services like Uber and Lyft to an increased reliance on public transit in the face of environmental concerns over emissions and uncertainty in gas prices, today's transportation policy makers and planners must deal with a rapidly changing urban mobility landscape. Despite methodological and technical advances in survey distribution and processing over the last 20 years [1], improvements in the accuracy of transportation demand prediction models have remained flat [29]. As a consequence, project overruns due to inaccurate economic value appraisals continue to cost taxpayers billions of dollars [75, 16]. Non-invasive, automated, continuous data collection mechanisms are increasingly being used to complement manual survey techniques by improving their statistical representative accuracy [21]. In particular, large volumes of call detail records (CDRs) from mobile phones have been used to construct individual daily itineraries and train travel activity models using weeks and months of data rather than several days' worth [80]. While applications of mining mobility trajectories from crowd-sourced locational data are common, existing models often trade in the ability to interpret data that transportation practitioners require in practice for the computational tractability necessary to deal with large volumes of data [18].

In practice, the activity-based travel models used by practitioners are incredibly rich in describing the intricacies of human activities and contexts of decision-making in travel-related choices. The appeal of econometric models estimated from these stated preference data of transportation system users rests on empirically-validated utility maximization axioms and the estimation of parameters for linear models that explain choices based on the characteristics of individual decision-makers and the attributes of the alternatives under consideration [8]. One key research challenge for the machine-learning community in this domain thus lies in developing a framework that reduces reliance on domain experts to specify robust models of travel decision-making while maintaining the flexibility to include covariates that rationalize observed behavior and responses to the applied system-level policies.

1.2 Problem Statement

To a large extent, human mobility is structured by highly regular daily/weekly schedules, demonstrating high predictability across diverse populations [32]. Longitudinal analysis also suggests that human mobility has high temporal and spatial regularities [65]. With the right tools, these patterns can be employed in predictive models. However, most of predictive models in the literature focus on prediction for only the next step activity. Few studies built models that were able to predict sequences of activities for an entire day or longer.

Another observation from the abundance of existing models is that very few include activity-type choices and spatial-temporal choices altogether for a complete prediction for human mobility. Eagle and Pendland [27], Farrahi and Gatica-Perez [28], and Zheng et al. [89] used unsupervised techniques such as PCA and topic models for primary activity type choices, which includes "home," "work," and "others." Liao et al. developed a hierarchical conditional random field (CRF) model for activity type and spatial choices [49]. Widhalm et al. [80] used an undirected relational Markov network for modeling spatial and temporal choices. Until very recently, a Markov model framework has been developed to include activity type, temporal, and implicit spatial choices [1] [84].

Furthermore, almost all human mobility sequence models make predictions by assuming Markov properties[66, 4, 82, 26, 84]. However, long-term activity dependencies can be observed from human mobility. For example, if an individual picks up a rental or shared vehicle implies s/he has to return it later. Another example is the responsibility of dropping off children to school and picking them up later in the day. Little research has been done to identify or quantify such an effect of activity dependencies. Nonetheless, statistics of vehicle-sharing and bike-sharing have shown strong behavioral patterns of picking up and returning transportation tools [64, 78], which are evidences of long-term activity dependencies.

To study the influence of social-demographic information towards traveling behavior, regression models are often used. It is usually difficult for regression models to generalize for modeling sequential choice involving multiple decisions simultaneously. In computer science research, on the other hand, large amounts of work focus on modeling and predicting human mobility on the individual level, such that the characteristic of individuals are often ignored [66, 4, 82, 26]. Generative models for sequences with contextual information have been well studied [76, 86, 81] for natural language problems. However, few such models have been developed in the domain of transportation research.

1.3 Contribution

To summarize, the existing human mobility models have focused on predicting next activities, and many models limited the prediction to only temporal features or only spatial features. The existing activity sequence models are developed by extending Markov Models or Hidden

[1] Implicit spatial choices are associated with each individual. The location choices are specified as distances to an individual's primary home location and primary work location

Markov Models that ignored long-term activity dependencies. Few works in transportation research have developed generative models that include social-demographic information. In this dissertation, we introduce Recurrent Neural Network models that applied to locational data. The major contributions of this dissertation include:

- We developed mobility sequence models that are capable of learning and predicting the entire mobility sequences within the time window of interest.
- We developed mobility sequence models that are able to predict activity sequences with activity-type choices and explicit spatial-temporal choices.
- We developed mobility sequence models that are able to capture long-term activity dependencies.
- We developed mobility sequence models that are capable of learning and generating sequences with contextual information.

1.4 Dissertation Outline

In Chapter 2, we introduce the background of our research and the mobility models that we developed. In Chapter 3, we introduce our first model structure for modeling daily activity sequences. We validate the model via agent-based traffic simulation and compare the simulation results to observed traffic volume. In Chapter 4, we introduced our second model structure for the application of medium-term activity sequence prediction. We evaluate the models against baseline models via partial sequence prediction. In Chapter 5, we introduce a model structure that includes the contextual information for sequence generation.

Chapter 2

Background

2.1 Human Mobility Data Sources

Collecting human mobility data via survey for aggregated analysis can be dated back in 1969, when the first National Household Traveling Survey (NHTS) was conducted. Since then, national surveys have been conducted every few years, and statistics of traveling behavior are documented [63]. Traveling surveys usually consist of questions regarding the destinations, start time, travel mode, purpose of the travel, and other information of each traveling activity during a single day. Surveys are usually conducted in person [11], by mail[63], and via applications on smartphones [42]. Survey data includes information that is often not accessible from non-invasive methods, and the collected information is often treated as ground truth. However, according to studies, surveys usually suffer from medium to low completion rate and inaccurate information, which is known as recall bias [22, 33].

GPS data is granular in both spatial and temporal resolution. Early work in extracting significant places took advantage of this relatively rich data source for activity inference models [90, 50]. Recent work uses GPS data to classify individual's mobility patterns [61] to measure the statistics in urban mobility [6] and to measure traffic flow [48]. GPS devices provides steady sampling rate with high accuracy. However, collecting GPS data often requires active user participation and permissions for physical devices, as well as careful battery management, thus limiting its practicality in terms of collecting continuous and representative samples from traveler populations.

Locational-based social network (LBSN) data usually contains locations of users of a service at the time of interaction (a check-in). Based on LBSN data, researches have developed methods for classifying activities into distinct categories [82, 45], for separating social trips from commute trips [19], promoting products and services, [92] and inferring mobility of grouped individuals [87]. Despite the increase of the popularity, LBSN is limited by the frequency of users' check-ins.

The anonymized Call Detail Records (CDRs) from cellular network operators provide a compromise between spatial-temporal resolution and ubiquity. Due to its relatively poor

resolution in space, CDR data has mainly been used to derive spatially aggregated results such as mass movements of population [25], aggregated origin-destination (OD) estimation [79], stylized mobility laws [32, 65], and disaster response [51].

2.2 Predictability of Human Mobility

Enabled by petabyte-scale data processing techniques, urban computing innovations and challenges have been drawing increasing social, commercial, and academic attention in the past decade [91]. Human activity prediction from spatial-temporal streams, an application of urban computing research, has been explored extensively by practitioners across disciplines[32, 18]. Early studies model human mobility as random walks or Levy flights, so the distribution of the leg length follows a heavy-tailed distribution [60]. However, studies have proven that human mobility is far from random walks because of high probabilities that individuals return to a few highly frequented locations [32, 65]. In fact, there is high predictability in human mobility regardless of the significant difference in individuals' travel patterns [65]. Predictability is defined as the probability that an algorithm can predict correctly the user's future whereabouts; predictability in human mobility can be as high as 93% [65]. Based on this discovery, recent studies have shifted the attention to discover patterns in human mobility from locational data sources. Some studies focused on the spatial patterns [37, 38], while others discovered daily and weekly mobility patterns [40, 71]. Correlations of spatial and temporal patterns have been found and quantified from aggregated mobility data [31].

2.3 Travel Demand Models

Trip Based Model

To understand transportation behavior from a global perspective, traditional trip-based models were applied in early research [55].Trip-based models are applied on the level of Traffic Analysis Zones (TAZs), which is defined according to the modeling needs. The characteristics of each TAZ, such as socio-economic, demographic, and land use information are defined as the Activity System. The graph links connecting TAZs are defined as the Transportation System. Trip-based models give a snapshot of the transportation network by applying the following four steps. The first step is trip generation, which determine the number of trips for different purposes, e.g. home-based and work-based trip. The second step is trip distribution, which assigns the origin and destination of trips to each TAZ. The third step is travel-mode assignment, which assigns modes to each trip based on the statistics of the mode share. The last step is route choice; vehicle trips will be assigned to links in the network while trips via public transportation methods will be assigned to the according routes [10].

While trip-based modeling frameworks can be calibrated to real-life traffic volume, the main drawback of the framework is that it considers each trip as an independent unit. Such

an approach ignores the sequential dependency of trips. For example, evening commute trips are consequences of morning commute trips. Another example is a multi-stop trip where an individual intends to visit multiple places within a time window. Those trips will be likely to share the same mode and some location choices. Again, the trip based models will fail to catch the dependencies.

Activity Based Model

Unlike trip-based models, which usually lack presentations of underlying travel behavior, activity-based modeling frameworks directly model stationary activities and transitions among them. Compared to trip-based models, activity based models are more natural frameworks for modeling transportation behavior. According to travel demand theory, demand of travel is derived from the need of transition among stationary activities, such as home, work, and shopping. Individuals pay the economic and time expenses of travel because of the spatial and temporal constrains of their stationary activities [15]. Thus, to model the travel demand on the network, the first step is to model the stationary activity trajectories. The trajectories include the spatial and temporal transitions and the purpose of each station activity. The second step is to select travel mode choice and route choice between consecutive stationary activities. This modeling framework extends the modeling power in the time horizon while keep the sequential dependencies among trips.

Compared with trip-based modeling, activity-based modeling requires more information to be collected. Modeling trajectories of stationary activities requires start time, duration, location, and activity purposes of each stationary activity along the trajectory. This information used to be available only by collecting surveys from individuals and households. But now it can be derived from locational data collected passively from mobile devices.

2.4 Models of Sequences

Hidden (semi-) Markov Models are generative models that can not only be used to analyze patterns in sequences, but also to generate new sequences [36]. Many recent works have been focused on predicting next-step activity, location, and purpose[66, 4, 82, 26]. Ye et al. developed HMM for predicting the activity location and purpose with the locations and purposes already labeled as categories [82]. Another work that uses HMM for predicting the next activity location defined location choices among the users' historical location clusters [53]. Similar approaches of next activity prediction have also been done by [30]. To advance HMM, Yin et al. have added temporal features to the Input-Output Hidden Markov Models (IO-HMMs) to reveal complete spatial-temporal activity sequences of urban mobility [85].

Recurrent neural networks (RNNs) have become the state of the art for sequence modeling and generation, especially in the language and translation domain [69]. Long short term memory (LSTM), which was first introduced by [41], is one of the most popular variations of RNNs, with proven ability to generate sequences in various domains, such as text [68],

images [35], and hand-writing [34, 20]. Despite the success of many applications of LSTM, very few of those applications in transpiration can be found. Some work has been done on predicting short-term traffic flow and traffic speed [52, 72]. Song et al. showed the capability of LSTM to predict individual locations and transportation mode with GPS data [67]. Alahi et al. Another work uses LSTM with shared hidden states to predict pedestrians' motion by considering their interaction [2]. Our work has been greatly inspired by the idea of mixture density network[13]. Mixture density network was extended from fully connected neural network to LSTM by Graves [34].

2.5 Fully-Connected and Convolutional Neural Network for Sequences

Long before recurrent neural network became popular, researchers applied fully-connected neural networks for learning sequences. Sequences, such as time series [24] and protein structures [62], are passed into neural networks as vectors with fixed lengths. Assuming the input sequence has a single dimension with length of l, we denote the input, the weights, the bias, and the output before activation and the output after activation at i'th layer as x_{i-1}, W_i, b_i, y_i and z_i. Since the input layer has dimension of $lx1$, the total number of trainable parameters is proportional to the sequence length l. The large number of trainable parameters make the fully-connected neural network inefficient in learning long sequences.

$$y_i = W_i x_{i-1} + b_i \tag{2.1}$$

$$z_i = \sigma(y_i) \tag{2.2}$$

Recent success of convolutional neural networks has demonstrated the ability of learning sequences [46] with convolutional neural networks. By using convolutional and pooling layers, patterns in the temporal dimension can be learned more efficiently. However, the total number of trainable parameters is still in proportion to the sequence length, which still makes learning long sequences inefficient.

2.6 Vanilla Recurrent Neural Network and Gradient Vanishing Problem

Vanilla recurrent network is a feed-forward neural network with memories for learning sequences. The parameters in a recurrent neural network are shared at different time steps. Compared with a fully-connected neural network, a recurrent neural network has much less parameters when learning sequences. The shared parameters also make recurrent neural networks flexible in terms of sequence lengths. Vanilla recurrent network was applied to time series data, such as stock price [43] and speech modeling [58].

At each time step, the output, y_t, from the recurrent neural networks is influenced by both the input, x_t, and memory, h_t, the at current time step. The next time step memory, y_{t+1}, is updated by passing the current output, y_t, through a non-linear activation function. The forward path of a vanilla recurrent neural network can be expressed as below.

$$y_t = W_x x_t + W_h h_{t-1} + b \tag{2.3}$$

$$h_{t+1} = \tanh(y_t) \tag{2.4}$$

We derive the gradient of parameter W_x and W_h with respect to the error, E, which is known as back propagation through time (BPTT). The gradient $\frac{\partial E_T}{\partial h_t}$ can be expended, and it increases or decreases exponentially with respect to the sequence lengths according to the gradient chain rule. This is known as the gradient vanishing problem, which leads to significant inefficiency in BPTT when training long sequences.

$$\frac{\partial E}{\partial W_x} = \sum_{t=0}^{T} \frac{\partial E_t}{\partial h_t} x_t \tag{2.5}$$

$$\frac{\partial E}{\partial W_h} = \sum_{t=0}^{T} \frac{\partial E_t}{\partial h_t} h_{t-1} \tag{2.6}$$

2.7 Long-Short Term Memory

Long-Short Term Memory (LSTM) was first described in [41] for solving the gradient vanishing problem in BPTT. The LSTM cell introduces gates for regulating flow of information. The input, output, and forget gates are denoted below as i, o, and f. Fig. 2.1 illustrates the micro-structure of a single LSTM cell. The gates are controlled by the input, the hidden state, and the hidden cell state at each time step using weights W_x, W_h and W_c. Sigmoid functions are used for gate activation and tanh is used as output activation function from the cell. The cell memory and the cell memory after activation are represented by c and h, which are usually known as hidden state and hidden cell state. The implementation of the gates and memory update are as following:

$$i_t = \sigma(W_{xi} x_t + W_{hi} h_{t-1} + b_i) \tag{2.7}$$

$$f_t = \sigma(W_{xf} x_t + W_{hf} h_{t-1} + b_f) \tag{2.8}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc} x_t + b_c) \tag{2.9}$$

$$o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + b_o) \tag{2.10}$$

$$h_t = o_t \tanh(c_t) \tag{2.11}$$

Here we derive the gradient of the weights in an LSTM cell with respect to the upstream error. First, the gradient of hidden cell state c_{t+1} with respect to the hidden state h_{t+1} is shown as the following:

$$\frac{\partial h_{t+1}}{\partial c_{t+1}} = o(1 - \tanh(c_{t+1})^2) \tag{2.12}$$

The gradient of gates i, g, o, f, with respect to hidden cell state c_{t+1} and hidden state h_{t+1} are formulated as the following:

$$\frac{\partial c_{t+1}}{\partial i} = g \tag{2.13}$$

$$\frac{\partial h_{t+1}}{\partial i} = o(1 - \tanh(c_{t+1})^2)g \tag{2.14}$$

$$\frac{\partial c_{t+1}}{\partial f} = c_t \tag{2.15}$$

$$\frac{\partial h_{t+1}}{\partial f} = o(1 - \tanh(c_{t+1})^2)c_t \tag{2.16}$$

$$\frac{\partial c_{t+1}}{\partial o} = \tanh(c_{t+1}) \tag{2.17}$$

$$\frac{\partial c_{t+1}}{\partial g} = i \tag{2.18}$$

$$\frac{\partial h_{t+1}}{\partial g} = o(1 - \tanh(c_{t+1})^2)i \tag{2.19}$$

By defining $a_\bullet = W_{x\bullet} x_t + W_{h\bullet} h_{t-1} + b_\bullet$, we derivative of gradient of $a_\bullet$ with respect to gates i, g, o, f.

$$\frac{\partial i}{\partial a_i} = i(1 - i) \tag{2.20}$$

$$\frac{\partial f}{\partial a_f} = f(1 - f) \tag{2.21}$$

$$\frac{\partial g}{\partial a_o} = o(1 - o) \tag{2.22}$$

$$\frac{\partial g}{\partial a_g} = (1 - \tanh(g)^2) \tag{2.23}$$

We further define a as concatenation of all $a_\bullet$. We derivative of gradient of weights.

$$\frac{\partial a}{\partial W_x} = x \tag{2.24}$$

$$\frac{\partial a}{\partial W_h} = h_t \tag{2.25}$$

$$\frac{\partial a}{\partial b} = \text{vector}(1) \tag{2.26}$$

Finally, we derivative of the gradient of previous memory c_t, h_t with respect to a.

$$\frac{\partial a}{\partial c_t} = c_{t+1} f + o(1 - \tanh(c_{t+1})^2) f \tag{2.27}$$

$$\frac{\partial a}{\partial h_t} = W_h \tag{2.28}$$

The implementation given above is for LSTM cells without peepholes, which means the gates are controlled only by the input x and the hidden state h. The version with peepholes allows the gates to also be controlled by the hidden cell state c.

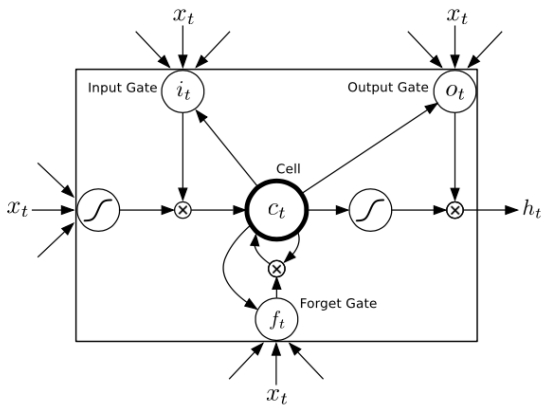


Figure 2.1: LSTM Cell [34]

2.8 Mixture Density Networks

Mixture density networks (MDN) were introduced by Bishop in 1994 [13]. The idea was later extended to LSTM that have mixture density outputs at each time step. We first recall the convention least square loss function, which is given as the following:

$$E = \frac{1}{2N} \sum_{n}^{N} [f(x_n; W) - t_n]^2 \tag{2.29}$$

where $f(x_n; W)$ represents output from a neural network with weights W given input x_n. The corresponding target is t_n. By applying the least square loss, the neural network minimizes the bias and variance at the same time.

$$
\begin{aligned}
E &= \lim_{N\to\infty} \frac{1}{2N} \sum_{n}^{N} [f(x_n; W) - t_n]^2 \\
&= \int \int \sum_{n}^{N} [f(x_n; W) - t_n]^2 p(x, t) dx dt
\end{aligned}
\tag{2.30}
$$

A model $f(x_n; W)$ that minimizes the least square loss is equivalent as it predicts the mean of target t given observation x. Thus, such a model cannot perform a reasonable prediction for an arbitrary conditional density function $p(t|x)$.

The idea of mixture density network is to have the neural network output serve as parameters of a mixture distribution. Part of the density network output will perform as the mixture weights α_i.

$$
\sum_{i} \alpha_i = 1 \tag{2.31}
$$

Other part of the neural network outputs are used as the rest of the parameters for the mixture distribution. For example, for a Gaussian mixture distribution, those parameters include mean u_i and standard distribution σ_i for each mixture.

Chapter 3

LSTM Mobility Model

3.1 Introduction

From the exponential growth of on-demand shared ride services like Uber and Lyft to an increased reliance on public transit in the face of environmental concerns over From the exponential growth of on-demand shared ride services like Uber and Lyft to an increased reliance on public transit in the face of environmental concerns over emissions and uncertainty in gas prices, today's transportation policy makers and planners must deal with a rapidly changing urban mobility landscape. Despite methodological and technical advances in survey distribution and processing over the last 20 years [1], improvements in the accuracy of transportation demand prediction models have remained flat [29]. As a consequence, project overruns due to inaccurate economic value appraisals continue to cost taxpayers billions of dollars [75, 16]. Non-invasive, automated, continuous data collection mechanisms are increasingly being used to complement manual survey techniques by improving their statistical representativeness [21]. In particular, large volumes of call detail records (CDRs) from mobile phones have been used to construct individual daily itineraries and train travel activity models using weeks and months of data rather than several days' worth [80]. While applications of mining mobility trajectories from crowd-sourced locational data are common (these are thoroughly reviewed in Section 3.2), existing models often trade off the interpretability that transportation practitioners require in practice for the computational tractability necessary to deal with large volumes of data [18].

In practice, the 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. The appeal of econometric models estimated from these stated preference data of transportation system users rests on empirically-validated utility maximization axioms and the estimation of parameters for linear models that explain choices based on the characteristics of individual decision-makers and the attributes of the alternatives under consideration [8]. One key research challenge for the machine-learning community in this domain thus lies in developing a framework that reduces reliance on domain experts to specify robust

models of travel decision-making while maintaining the flexibility to include covariates that rationalize observed behavior and responses to the applied system-level policies.

In this chapter, we describe the developed two-step generative modeling framework that is capable of learning activity sequences from large volumes of call data. By combining recurrent neural networks with probabilistic modeling techniques specific to the task of generating human activity chains from cellular data, we thus expand the scope of state-of-the-art human mobility modeling techniques to the realm of decision-support for transportation policy analysis.

This paper details the following contributions:

- We implement an extensible end-to-end processing and inference pipeline; using raw cellular data as input, our data processing framework provides transportation planners, policy-makers and related stakeholders with detailed and timely travel demand models.
- To the best of the author's knowledge, this is the first work using deep recurrent neural network architecture to generate human activity chains from cellular data. Thanks to its flexibility in modeling long temporal dependencies and explicit location choices, it was found to be the key modeling component to be deployed in practice.
- Leveraging rich activity-based inferences of observed travel, we connect our framework with the state-of-the-art discrete choice modeling to include socio-economic variables into explicit choice models that inform scenario micro-simulation and support policy evaluation by practitioners.

To validate the framework, we sample trained generative models to produce synthetic travel plans for the population in the region, and use them 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 and transit lines within the region of study. As the framework is currently being implemented for a test deployment in the San Francisco Bay Area, we discuss some additional metrics that we have set to validate its usability in practice.

3.2 Existing Work

Enabled by petabyte-scale data processing techniques, urban computing innovations and challenges have been drawing increasing social, commercial, and academic attention in the past decade [91]. Human activity recognition from spatiotemporal microdata streams, an application of urban computing research has been explored extensively by practitioners across disciplines[32, 18]. A summary of relevant developments in urban activity modelling is given below with respect to the main data sources and the properties of the explored algorithms.

Locational Data Sources

GPS data is granular in both spatial and temporal resolution. Early work in extracting significant places took advantage of this relatively rich data source for activity inference models [90, 50]. However, collecting GPS data often requires active user participation and permissions for physical devices as well as careful battery management, thus limiting its practicality in terms of collecting continuous and representative samples from traveler populations.

Locational-based social network (LBSN) data usually contains locations of users of a service at the time of interaction (a check-in), and is limited by the frequency of users' check-ins. Based on LBSN data, researches have developed methods for classifying activities into distinct categories [82, 45], for separating social trips from commute trips [19], promoting products and services [92], and inferring mobility of grouped individuals [87].

The anonymized Call Detail Records (CDRs) from cellular network operators provide a compromise between spatial-temporal resolution and ubiquity. Due to its relatively poor resolution in space, CDR data has been mainly used to derive spatially aggregated results such as mass movements of population [25], aggregated origin-destination (OD) estimation [79], stylized mobility laws [32, 65], and disaster response [51].

Generative Models for Sequential Data

To a large extent, human mobility is structured by highly regular daily/weekly schedules, demonstrating high predictability across diverse populations [32]. Longitudinal analysis also suggests that human mobility has high temporal and spatial regularities [65]. With the right tools, these patterns can be employed in predictive models.

Recurrent neural networks (RNNs) have become the state of the art for sequence modeling and generation, especially in the language and translation domain [69]. Long short term memory (LSTM) [41] is one of the most popular variations of RNNs with proven ability to generate sequences in various domains, such as text [68], images [35], and hand-writing [34, 20]. Recent works showed the capability of LSTM to predict individual mobilities using GPS data [67] and to predict pedestrians' motion from video data [2]. Another recent work showed the great performance of labeling acoustic and accelerometer sequences using a combination of Hidden Semi-Markov Model and Recurrent Neural Network [23]. The success of RNNs in these domains motivated our work of applying RNNs for human activity sequence modeling and generation.

Hidden (semi-) Markov Models are generative models that can not only be used to analyze patterns in sequences, but also to generate new sequences [36]. Recent work used Input-Output Hidden Markov Models (IO-HMMs) to reveal urban activity patterns (including the spatial-temporal profiles of urban activity and heterogeneous transition probabilities) and to generate synthetic activity sequences to inform microscopic traffic simulations [85].

The advantages of IO-HMM over standard HMM is that it relaxes the assumption that the transition probabilities and emission probabilities are homogeneous. Instead, the transitions

and emissions in IO-HMM can depend on the contextual variables such as time of day and day of week. While HMMs benefit from a simple interpretable structure and an ability to model dynamic latent variables, they suffer from a few limitations. For one, when training the model, the number of hidden nodes (each hidden node corresponds to an activity) needs to be pre-specified. Although one can use the number of conventional urban activities (home, work, food, shop, recreation, etc.), the learned activities do not necessarily correspond to each pre-defined activity type.

LSTMs, on the other hand, are more flexible as they use many continuous hidden units with implicit embedding corresponding to the discrete activities. LSTMs also hold an advantage over HMMs in modeling long-term dependencies thanks to the automatically learned "input," "output," and "forget" gates. Moreover, the flexibility and the tremendous learning power allows LSTMs to parametrize complex joint distributions required to reproduce the variability of destination location choice for urban travelers with heterogeneous preferences and lifestyles.

3.3 Model Framework

Our framework builds upon cell phone data processing and activity-based inferences of travel purposes with an IO-HMM, followed by a LSTM network that learns travelers' mobility sequences. Previous work showed the activity recognition by IO-HMM has an accuracy of 88% for home, work, and all secondary activities[85]. However, one missing piece in the work is explicit and privacy preserving location choices of activities. A common approach is to sample a random location within a traffic analysis zone (TAZ), which is chosen based on the learned spatial coefficients (distance to home and distance to work) of the activity. In this work, we are interested in an activity sequence modeling framework for producing realistic synthetic activity plans including explicit location choices. This objective leads us to the following two-step framework of generating urban activities. We show the two-step framework in Fig 3.1.

1. In the first step, raw CDR data containing a time-stamped record for each communication of anonymous user's device served by the cellular network go through a k-anonymity check [70] and optional differential privacy (DP) filters [56, 39] as required by a data provider. The masked CDR data are then pre-processed to a sequence of stay location clusters that may correspond to distinct yet unlabeled activities. 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. IO-HMMs are then used to label each activity and uncover the activity patterns [85].

2. In the second step, the activity sequences, together with the recognized activity labels, are sent to a generative recurrent neural network with LSTM cells for training. The trained model is able to learn explicit location choice with mixture density outputs for

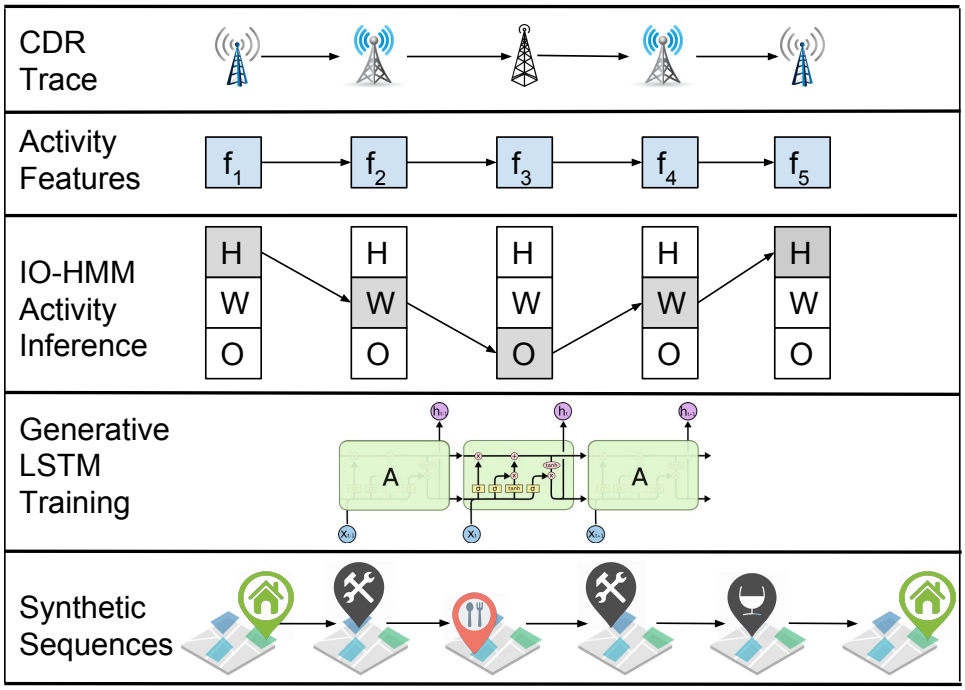


Figure 3.1: Model framework

each type of activity, and thus capable of generating realistic activity chains. LSTM is designed with a locational privacy bound parameter that controls the variance of sampled locations, making it suitable for placing additional privacy filters [39, 12].

Finally, travel plans are generated for the synthetic population in the region from the generative LSTM model. Synthetic travel plans do not correspond to any real user trajectories. To further protect privacy, traces that are too close to real trajectories can be filtered using existing techniques [39, 12]. The filtered synthetic travel plans, together with the travelers' travel mode choice parameters, are fed to a microscopic transport simulator.

The output of the simulator provides:

- detailed daily travel itineraries of the synthetic population,
- traffic volumes and transit passenger counts for comparison against real counts from highway sensors and transit agencies data,
- range of metrics for a given scenario including its environmental impact.

3.4 Model Structure

In this section, we describe the LSTM model structures that generate activity sequences with explicit spatial-temporal choice and activity types, which is shown in Fig 3.2. The LSTM

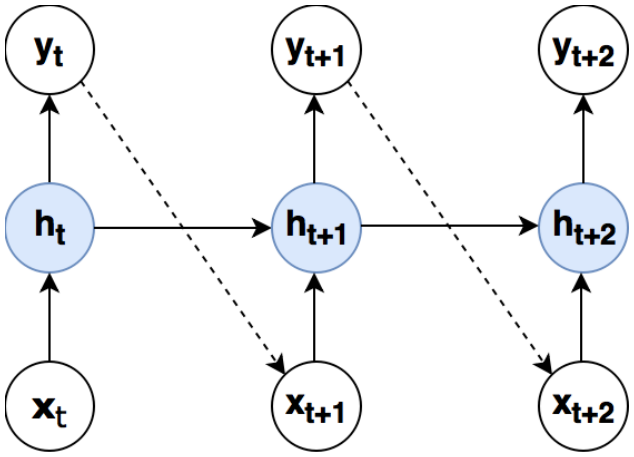


Figure 3.2: LSTM Model Structures

model takes features of current activity as an input and generates a mixture of distributions that spatial-temporal choice and activity type choice of next activity can be sampled from such distributions.

We employ techniques of mixture density networks (MDNs) [13, 34], for the spatial-temporal distributions. The idea of mixture density output is that a portion of the LSTM output is used as weights of mixtures, and the rest is used to parameterize each mixture component, which is described below in detail. In Fig. 3.2, we illustrate our model structure. We use x_t, h_t, and y_t to denote input variables, hidden states of LSTM layer(s), and outputs from LSTM layer(s) of t'th activity of the day, respectively.

Our model generates activity sequences according to the following three steps:

1. At every step, t, the LSTM layer(s) receive a set of inputs x_t.
2. The LSTM layer(s) then produce a set of outputs y_t, which is used to parametrize a mixture distribution $p(x_{t+1}|y_t)$.
3. A new activity x_{t+1} is sampled from distribution $p(x_{t+1}|y_t)$.

Input Variables

We design the input variables, x_t, such that it contains the features that affect the choices of next activity. Those features include the starting time and the type of the activity. Starting time is a continuous variable, while activity type is a categorical variable. Activity types consist of the labels of each activity including "home," "work," and "others," which are encoded as a one-hot vector. Thus, x_t can be expressed as the following.

$$x_t = \{\text{start\_time}_t, \text{activity\_type}_{t-1}, \text{day\_of\_week}_t\} \tag{3.1}$$

Output Variables

We design the mixture density output from the LSTM that is used for sampling the activity. Output variables y_t are decomposed and transformed into coefficients of mixture distribution $p(x_{t+1}|y_t, c_{t+1})$, which is used for generating the next x_t.

The output distribution from the one-layer structure is a joint distribution of activity starting time, duration, latitude, longitude, and probabilities of activity types; each component is weighted by π_i. The joint distribution of starting time, duration, latitude, and longitude is a 4-dimensional Gaussian distribution with correlation only between activity starting time and duration. p_{home}, p_{work}, and p_{other} describe a distribution for activity types with 3 categories: home-related, work-related, and other activities. In short, each mixture component indicates location, duration, purpose choice of next activity, and $p(x_{t+1}|y_t, c_{t+1})$ contains many of such choices by having multiple mixture components. Hence, according to the descriptions above, y_t is split as the following:

$$\begin{aligned} \boldsymbol{y_t} = \{&\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}_{\text{lat}}, \hat{\boldsymbol{\mu}}_{\text{lon}}, \hat{\boldsymbol{\mu}}_{\text{st}}, \hat{\boldsymbol{\mu}}_{\text{dur}}, \\ &\hat{\boldsymbol{\sigma}}_{\text{lat}}, \hat{\boldsymbol{\sigma}}_{\text{lon}}, \hat{\boldsymbol{\sigma}}_{\text{st}}, \hat{\boldsymbol{\sigma}}_{\text{dur}}, \hat{\boldsymbol{\rho}}_{\text{st, dur}}, \\ &\hat{\boldsymbol{p}}_{\text{home}}, \hat{\boldsymbol{p}}_{\text{work}}, \hat{\boldsymbol{p}}_{\text{other}}\} \end{aligned} \tag{3.2}$$

Output transformation Those raw outputs from LSTM are properly transformed before serving as mixture distribution parameters. The component weights $\hat{\boldsymbol{\pi}}$ and probability of activity type within each component are normalized using softmax function. Standard deviations $\hat{\boldsymbol{\sigma}}_\bullet$ are constrained to be non-negative using an exponential function with correlation coefficients, $\hat{\boldsymbol{\rho}}_\bullet$, scaled between -1 and 1 using tanh activation functions. The following equations summarize our model:

$$\pi^i = \frac{\exp(\hat{\pi}^i)}{\sum_j^N \exp(\hat{\pi}^j)}; i \in \{1...N\} \tag{3.3}$$

$$\mu_\bullet = \hat{\mu}_\bullet \tag{3.4}$$

$$\sigma_\bullet = \exp(\hat{\sigma}_\bullet) \tag{3.5}$$

$$\rho_\bullet = \tanh(\hat{\rho}_\bullet) \tag{3.6}$$

$$p_i = \frac{\exp(\hat{p}_i)}{\sum_j \exp(\hat{p}_j)}; i, j \in \{\text{home}, \text{work}, \text{other}\} \tag{3.7}$$

Where b is the sampling bias and can be added when generating sequences. Higher bias will increase the consistency and lower the variety in the generated sequences.

Mixture Distributions

For one-layer structure, the spatial-temporal variables and activity type variables are all jointly distributed in one mixture distribution. We write the mixture distribution as follows:

$$\begin{aligned} &p(\text{lat}_t,\text{lon}_t, \text{st}_t, \text{dur}_t, \text{type}_t|y_{t-1}) = \\ &\sum_i^N \pi^i p^i(\text{lat}_t,\text{lon}_t, \text{st}_t, \text{dur}_t, \text{type}_t|\mu_\bullet, \sigma_\bullet, \rho_\bullet, p_\bullet) \end{aligned} \tag{3.8}$$

We define the following decomposition of the joint distribution. Here we split a single spatiotemporal four-dimensional Gaussian distribution into a two two-dimensional Gaussian distributions for better computation efficiency, with an assumption of independence between the spatial (latitude, longitude) and temporal variables.

$$\begin{aligned} &p^i(\text{lat}_t, \text{lon}_t, \text{st}_t, \text{dur}_t, \text{type}_t|\mu_\bullet, \sigma_\bullet, \rho_\bullet, p_\bullet) \\ &\stackrel{\text{def}}{=} p^i(\text{lat}_t, \text{lon}_t)p^i(\text{st}_t, \text{dur}_t)p^i(\text{type}_t) \end{aligned} \tag{3.9}$$

$$p^i(\text{lat}_t, \text{lon}_t) = \mathcal{N}\left(\text{lat}_t, \text{lon}_t \middle| \begin{bmatrix}\mu^i_{\text{lat}} \\ \mu^i_{\text{lon}}\end{bmatrix}, \Sigma^i_{\text{lat,lon}}\right) \tag{3.10}$$

$$p^i(\text{st}_t, \text{dur}_t) = \mathcal{N}\left(\text{st}_t, \text{dur}_t \middle| \begin{bmatrix}\mu^i_{\text{st}} \\ \mu^i_{\text{dur}}\end{bmatrix}, \Sigma^i_{\text{st,dur}}\right) \tag{3.11}$$

$$\begin{aligned} p^i(\text{type}_t) = \sum_j j_t p_j; \\ j \in \{\text{home}, \text{work}, \text{others}\} \\ j_t \in \{0, 1\} \end{aligned} \tag{3.12}$$

where

$$\Sigma^i_{\text{lat,lon}} = \begin{bmatrix}\sigma^2_{\text{lat}} & 0 \\ 0 & \sigma^2_{\text{lon}}\end{bmatrix} \tag{3.13}$$

$$\Sigma^i_{\text{st,dur}} = \begin{bmatrix}\sigma^2_{\text{st}} & \sigma_{\text{st}}\sigma_{\text{dur}}\rho_{\text{st, dur}} \\ \sigma_{\text{st}}\sigma_{\text{dur}}\rho_{\text{st, dur}} & \sigma^2_{\text{dur}}\end{bmatrix} \tag{3.14}$$

Sequence generation

Sequence generation is initialized by feeding the LSTM model with x_0. The initial input x_0 is a vector of constants. Starting from $t = 1$, we sample location, duration, and activity type from $p(x_t|y_{t-1})$. The starting time of the next activity is updated by adding the sampled duration to the current time that the start time of each activity is observed prior to the sampling. The conditional mixture weights, mean, and variance of activity duration are calculated as below:

$$w^i(\mathrm{st}_t) = \frac{\pi^i \mathcal{N}(\mathrm{st}_t|\mu_{\mathrm{st}}, \sigma_{\mathrm{st}})^i}{\sum_j^N \pi^j \mathcal{N}(\mathrm{st}_t|\mu_{\mathrm{st}}, \sigma_{\mathrm{st}})^j} \tag{3.15}$$

Because of the correlation between activity starting time and duration, the mean and standard deviation of activity duration conditioned on the observed starting time, st_t, is expressed as follows:

$$\mu_{\mathrm{dur}|\mathrm{st}_t} = \mu_{\mathrm{dur}} + \frac{\sigma_{\mathrm{dur}}}{\sigma_{\mathrm{st}}}\rho_{\mathrm{st,dur}}(\mathrm{st}_t - \mu_{\mathrm{st}}) \tag{3.16}$$

$$\sigma_{\mathrm{dur}|\mathrm{st}_t} = \sqrt{(1 - \rho^2_{\mathrm{st,dur}})}\sigma_{\mathrm{dur}} \tag{3.17}$$

Now we can sample a new activity x_t from the mixture distribution $p(x_t|y_{t-1})$ following Eq. 3.18 through Eq. 3.21. First, a mixture component is sampled from the multinomial distribution of mixture weights (Eq. 3.15). Then, dur_t, lat_t, lon_t, $type_t$, end_t can be further sampled from the selected component, yielding a joint distribution of all those variables. Once a new activity is sampled, the time of day is incremented by dur_t.

$$\mathrm{k} \sim \mathrm{Multinomial}\Big(w(\mathrm{st}_t)_1, ..., w(\mathrm{st}_t)_N; n = 1\Big) \tag{3.18}$$

$$\mathrm{lat}_t, \mathrm{lon}_t \sim \mathcal{N}\left(\begin{bmatrix} \mu^k_{\mathrm{lat}} \\ \mu^k_{\mathrm{lon}} \end{bmatrix}, \Sigma^k_{\mathrm{lat,lon}} \right) \tag{3.19}$$

$$\mathrm{dur}_t \sim \mathcal{N}\Big(\mu^k_{\mathrm{dur}|\mathrm{st}_t}, \sigma^k_{\mathrm{dur}|\mathrm{st}_t}\Big) \tag{3.20}$$

$$\mathrm{type}_t \sim \mathrm{Multinomial}\Big(p^k_{\mathrm{home}}, p^k_{\mathrm{work}}, p^k_{\mathrm{others}}; n = 1\Big) \tag{3.21}$$

In Fig 3.3, we show an example of the LSTM model generating an activity sequence. We train the LSTM model using a made-up data set and show the spatial-temporal distribution for illustration purpose only. The LSTM model generates a sequence that consists of 5 activities, which are labeled as 0 through 4. Fig 3.3a and Fig 3.3b show the spatial and temporal distribution of each activity. Fig 3.3c shows the distribution of activity type.

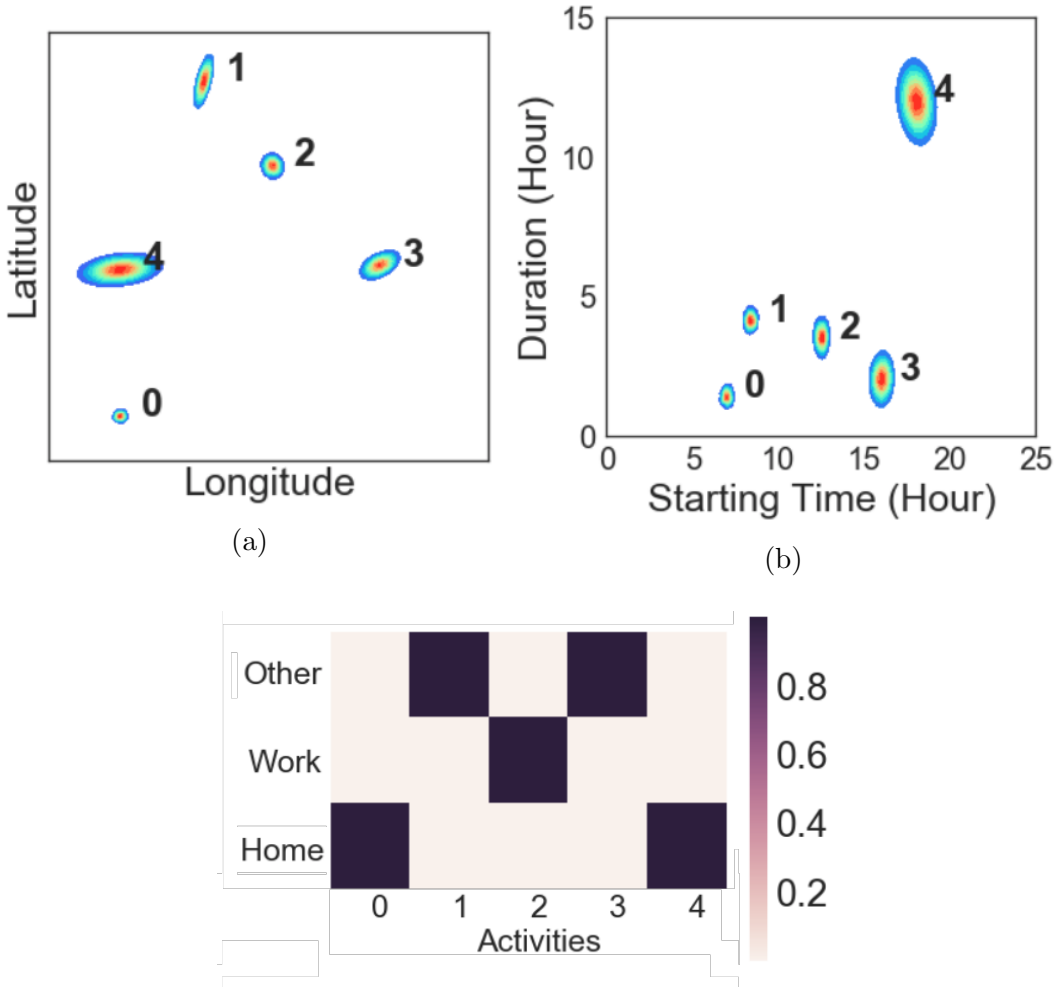


Figure 3.3: Mixture density outputs of LSTM model when generating a day sequence with 5 activities. The model is trained on a made-up data set for illustration purpose only. (a) and (b) show the output probability distributions for activities labeled as 0 to 4 in latitude-longitude and starting-time-duration space, respectively. (c) shows the probabilities of activity types and end-of-day probabilities, which are averaged over all the mixture components.

Loss Function and Model Estimation

We use negative log-likelihood as the loss of the model. Given an activity chain, the loss is calculated as the sum of negative log-likelihood of each observed activity. In our implementation, to deal with activities with different lengths, we pad the sequences to the same length and mask the padded portion when we calculate the loss. Adam optimizer [44] is used for

parameter estimation. Here we provide the derivatives of the loss with respect to the output of the LSTM model:

$$\begin{aligned}\boldsymbol{\ell} &= \sum_{t=1}^{T} -\log p(x_t|y_{t-1}) \\ &= \sum_{t=1}^{T} -\log \sum_{i}^{N} \pi_t^i p^i(\text{lat}_t, \text{lon}_t, \text{st}_t, \text{dur}_t, \text{type}_t|\mu_\bullet, \sigma_\bullet, \rho_\bullet, p_\bullet) \\ &= \sum_{t=1}^{T} -\log \left[\sum_{i}^{N} \pi_t^i \mathcal{N}\left(\text{lat}_t, \text{lon}_t \middle| \begin{bmatrix} \mu_{\text{lat}}^i \\ \mu_{\text{lon}}^i \end{bmatrix}, \Sigma_{\text{lat,lon}}^i \right) \right. \\ &\qquad \left. \mathcal{N}\left(\text{st}_t, \text{dur}_t \middle| \begin{bmatrix} \mu_{\text{st}}^i \\ \mu_{\text{dur}}^i \end{bmatrix}, \Sigma_{\text{st,dur}}^i \right) \sum_j j_t p_{t,j}^i \right]\end{aligned} \tag{3.22}$$

Here we show in detail the parameter update of neural network output y_t with respect to error $\boldsymbol{\ell}$. We start by defining the terms $\hat{\gamma}_t^i$ and γ_t^i as following expressions to simply the expression in later derivations.

$$\hat{\gamma}_t^i = \pi_t^i \mathcal{N}\left(\text{lat}_t, \text{lon}_t \middle| \begin{bmatrix} \mu_{\text{lat}}^i \\ \mu_{\text{lon}}^i \end{bmatrix}, \Sigma_{\text{lat,lon}}^i \right) \mathcal{N}\left(\text{st}_t, \text{dur}_t \middle| \begin{bmatrix} \mu_{\text{st}}^i \\ \mu_{\text{dur}}^i \end{bmatrix}, \Sigma_{\text{st,dur}}^i \right) \sum_j j_t p_{t,j}^i$$

$$\gamma_t^i = \frac{\hat{\gamma}_t^i}{\sum_i \hat{\gamma}_t^i}$$

Then, the derivatives of the loss function with respect to parameters are as follows:

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\pi}_t^i} = \pi_t^i \gamma_t^i - \gamma_t^i$$

$$\frac{\partial \boldsymbol{\ell}}{\partial(\hat{\mu}_{\text{lat}}^i, \hat{\mu}_{\text{lon}}^i, \hat{\sigma}_{\text{lat}}^i, \hat{\sigma}_{\text{lon}}^i)} = -\gamma_t^i \frac{\partial \log \mathcal{N}\left(\text{lat}_t, \text{lon}_t \middle| \begin{bmatrix} \mu_{\text{lat}}^i \\ \mu_{\text{lon}}^i \end{bmatrix}, \Sigma_{\text{lat,lon}}^i \right)}{\partial(\hat{\mu}_{\text{lat}}^i, \hat{\mu}_{\text{lon}}^i, \hat{\sigma}_{\text{lat}}^i, \hat{\sigma}_{\text{lon}}^i)} \tag{3.23}$$

$$\frac{\partial \boldsymbol{\ell}}{\partial(\hat{\mu}_{\text{st}}^i, \hat{\mu}_{\text{dur}}^i, \hat{\sigma}_{\text{st}}^i, \hat{\sigma}_{\text{dur}}^i, \hat{\rho}_{\text{st,dur}})} = -\gamma_t^i \frac{\partial \log \mathcal{N}\left(\text{st}_t, \text{dur}_t \middle| \begin{bmatrix} \mu_{\text{st}}^i \\ \mu_{\text{dur}}^i \end{bmatrix}, \Sigma_{\text{st,dur}}^i \right)}{\partial(\hat{\mu}_{\text{st}}^i, \hat{\mu}_{\text{dur}}^i, \hat{\sigma}_{\text{st}}^i, \hat{\sigma}_{\text{dur}}^i, \hat{\rho}_{\text{st,dur}})} \tag{3.24}$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{p}_{t,j}^i} = p_{t,j}^i \gamma_t^i - \gamma_t^i \quad \text{if} \quad j_t = 1$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{p}_{t,j}^i} = 0 \quad \text{if} \quad j_t = 0$$

where in Eq.3.23 and Eq.3.24

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\mu}_1^i} = \frac{C}{\sigma_1^i}\left(\frac{x_1^i - \mu_1^i}{\sigma_1^i} - \frac{\rho^i(x_2^i - \mu_2^i)}{\sigma_2^i}\right)$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\mu}_2^i} = \frac{C}{\sigma_2^i}\left(\frac{x_2^i - \mu_2^i}{\sigma_2^i} - \frac{\rho^i(x_1^i - \mu_1^i)}{\sigma_1^i}\right)$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\sigma}_1^i} = \frac{C(x_1^i - \mu_1^i)}{\sigma_1^i}\left(\frac{x_1^i - \mu_1^i}{\sigma_1^i} - \frac{\rho(x_2^i - \mu_2^i)}{\sigma_2^i} - 1\right)$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\sigma}_2^i} = \frac{C(x_2^i - \mu_2^i)}{\sigma_2^i}\left(\frac{x_2^i - \mu_2^i}{\sigma_2^i} - \frac{\rho(x_1^i - \mu_1^i)}{\sigma_1^i} - 1\right)$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\rho}^i} = \frac{(x_1^i - \mu_1^i)(x_2^i - \mu_2^i)}{\sigma_1^i \sigma_2^i} + \rho^i(1 - C^i Z^i)$$

where

$$C^i = \frac{1}{1 - \rho^{i2}}$$

$$Z^i = \frac{(x_1^i - \mu_1^i)^2}{\sigma_1^{i2}} + \frac{(x_2^i - \mu_2^i)^2}{\sigma_2^{i2}} - \frac{2\rho^i(x_1^i - \mu_1^i)(x_2^i - \mu_2^i)}{\sigma_1^i \sigma_2^i}$$

3.5 Learning Long-term Dependencies

We show examples of how the LSTM model learns long-term dependencies in sequences. We design the activity sequences with 8 timestamps, and the sequences consist of 3 activity types (type 0, 1 and 2). As shown in Fig 3.4a, in each sequence, the activity type at timestamp 1 will be repeated at timestamp 6. Thus, according to the activity type pairs at timestamp 1 and 6, the sequences form 3 patterns. We train the LSTM model using the sequences designed without extra features; the LSTM model learns those pattern in sequences. We performed two experiments. In the first experiment, we filled the rest of the timestamps in sequences with activity type 0. In the second experiment, we filled the different activity types between timestamps 2 and 5. Fig 3.4b and Fig 3.4c show the sequence generation results from the two experiments. We use jittered lines to show each sequence generated with different colors to identify different patterns. As we observe from the generated results, the LSTM model reproduces the 3 patterns in sequence successfully.

3.6 Experiment Setup

This section describes the steps in a full-scale regional experiment where we train LSTM model for commuters from each of the 34 super-districts in the San Francisco Bay Area in order to develop an actionable mobility model for a typical weekday.

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.

Data Pre-processing

We pre-process the data following the steps in [85]. The home and work locations are identified during the pre-processing step. We take cell phone users who 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, and have home and work at different locations. These criteria identify regular working commuters with a day structure containing both distinct Home and Work.

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]. Overall, the aggregated statistics of activity labeling by IO-HMM match with the travel surveys. The percentage of US employed person who go to work on an average weekday is 82.9% [74]; this number is 83.7% for IO-HMM labeling. 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 [11], and these numbers are 2.9%, 15.2%, and 43.7% in our labeled data.

We tested our models' generative power in the Bay Area context. As travel patterns vary greatly over the region, we trained 34 LSTM models, 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). We simulate $463,000$ agents in the Bay Area (15% sample of the commuters) for each weekday: Monday through Friday.

We use a single layer of LSTM with 128 units. The number of output mixture components, N, is chosen as 80. We use 10% dropout rate for the LSTM units to prevent over fitting of the data. For Adam optimizer, the learning rate, β_1, and β_2 are chosen as 0.001, 0.9, and 0.999, respectively. The sampling bias b is tuned as 2.0 in order to reduce number of outliers in the generated sequences.

3.7 Results

We present the temporal characteristics of the generated sequences in Fig. 3.5 and Fig. 3.6. From Monday to Friday, we observe a decreasing number of work and home activities while there is a slight increase in secondary (other) activities. Fig. 3.6 shows the joint distributions of starting time and duration of each activity type. The home activities starting around noon

are relatively short. Night-time home activities have strong correlations between starting time and duration. Work activities show typical working patterns of commuters, while some last the entire day starting in the morning and others last only half of the day. Secondary (other) activities peak around the morning and afternoon commute hour, and they usually last within 1.5 hours. The temporal distributions are very similar to what is reported in other studies [85, 80].

Scenario micro-simulation

Micro-simulation of a typical weekday traffic is performed using the MATSim[1] platform [5]. MATSim is a state-of-the-art agent-based multi-modal mobility micro-simulation tool that performs more choice and traffic assignments 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 an agent's daily utility, parametrically defined with several parameters, including income, the value of time, and mode preference parameters. The simulation is run on the SF Bay Area network containing all major transit routes, freeways, and primary and secondary roads (network fragment is visualized in Figure 3.9).

We have compared the results of the flows simulated from the generated activity sequences with the observed traffic and transit passenger volumes provided by the California DOT Performance Management System (PeMS) and the Metropolitan Transportation Commission respectively. The simulation is run at 15% of the total population with activity plans generated from the LSTM model, and the road capacities as well as total resulting counts are scaled accordingly.

Note that observed traffic and transit passenger counts are not used for model calibration. They are used as independent data to evaluate the validity of the synthetic travel sequences produced with the LSTM model. Fig. 3.9 demonstrates examples of the three characteristic hourly volume profiles, comparing the modeled and observed counts on freeways. Figure 3.7 shows examples of transit passenger counts entering and exiting 2 major rapid transit stations. Validation results for the full set of sensors are presented in Fig. 3.8. Fig. 3.8a show a comparison of the volumes for three distinct time periods. Fig. 3.8b summarizes the validation results over 300 freeway and transit 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 the 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 ($r^2 = 0.76, p < 10^{-3}$ in Fig. 3.8a) as compared to the ground truth data.

[1] MATSim code available at `https://github.com/matsim-org`

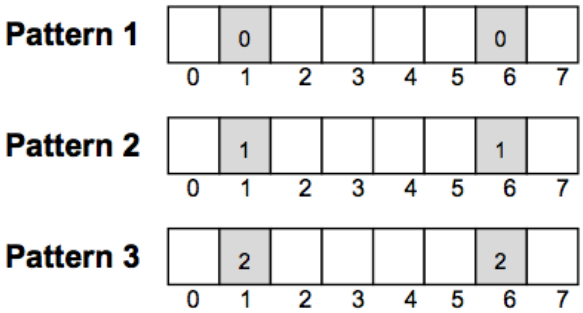
3.8 Discussion

In this chapter, we presented an end-to-end pipeline of processing, modeling, and simulating urban mobility from cellular data. We introduced a two-step generative model framework for learning urban activities and mobility. An IO-HMM model is used for labeling activity types of the pre-processed and anonymized cellular data in San Francisco Bay Area. We proposed an LSTM model that is capable of learning explicit location choices that is applied on the labeled activity sequences.

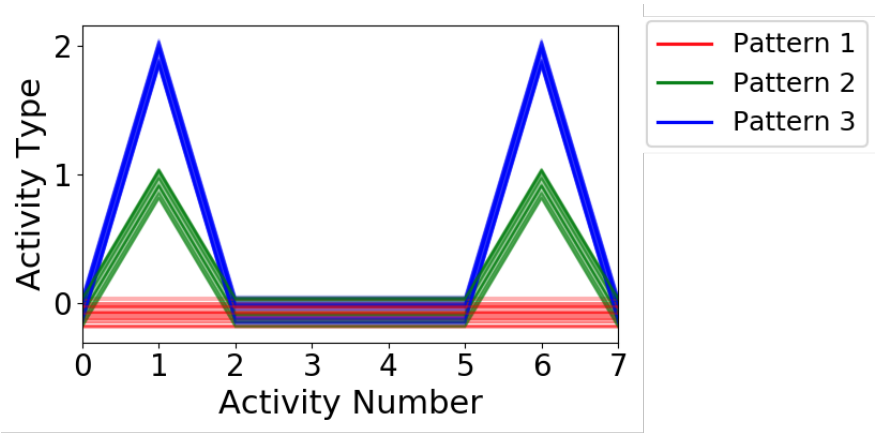
Our LSTM model is evaluated with either survey or real data collected from the transportation network. The activities labeled by IO-HMM were validated by comparing the aggregated activity statistics with 2015 travel, which showed high similarity to the survey results. To examine the generative power of the LSTM model, we synthesized urban mobility plans using trained models. An agent-based micro-simulation of travel with multiple travel modes was run using the synthesized plans. The vehicle traffic counts and public transit boarding and alighting counts from the simulation result were compared with real traffic and transit data. A reasonable fit accuracy was observed.

With privacy concerns in mind, we will also work on improving performance and modeling accuracy by partitioning a population into finer sub-groups (whether socially or spatially) to take advantage of parameter sharing between the IO-HMM/LSTM models. Along with conducting performance evaluations and enhancing practical usability, we plan to study the privacy/utility trade-offs of the overall system.

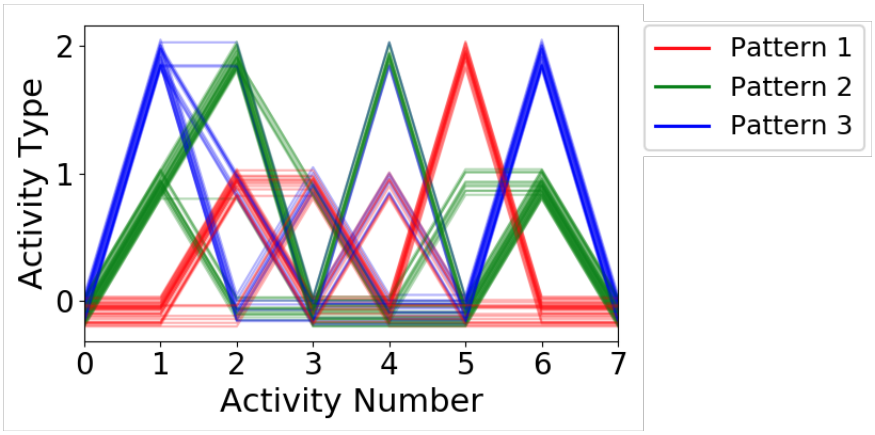
We identify the long-term dependency resolution of RNNs and the expressiveness of probabilistic modeling techniques as two components that working in concert have the potential to drive the future state of practice in transportation demand modeling. In particular, this research anticipates the rapidly growing development of novel techniques in learning the parameters of recurrent neural networks and training generative models that specify parameters necessary to model socioeconomic correlates of travel behavior. We look forward to developing additional innovations using this template, exploring its promising future to help mitigate the significant costs and delays associated with traditional practices of transportation planning and operations.



(a) Number of activities (labeled per highest posterior probability) by their respective start time within a course of 5 weekdays.



(b) Number of activities (labeled per highest posterior probability) by their respective start time within a course of 5 weekdays.



(c) Number of activities (labeled per highest posterior probability) by their respective start time within a course of 5 weekdays.

Figure 3.4: An example of LSTM model learning long-term dependencies in sequences.

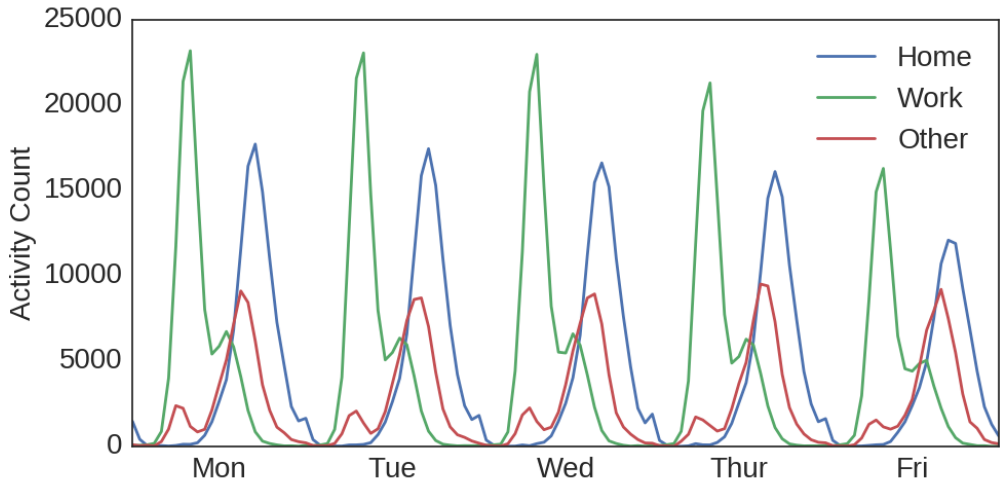
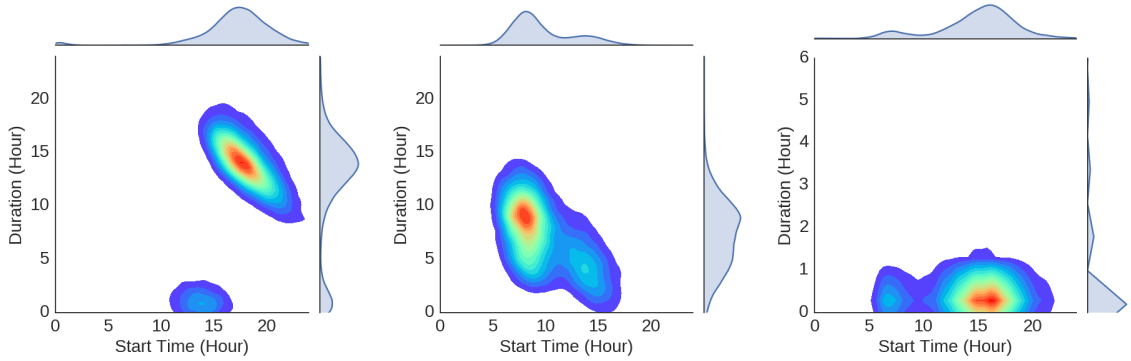
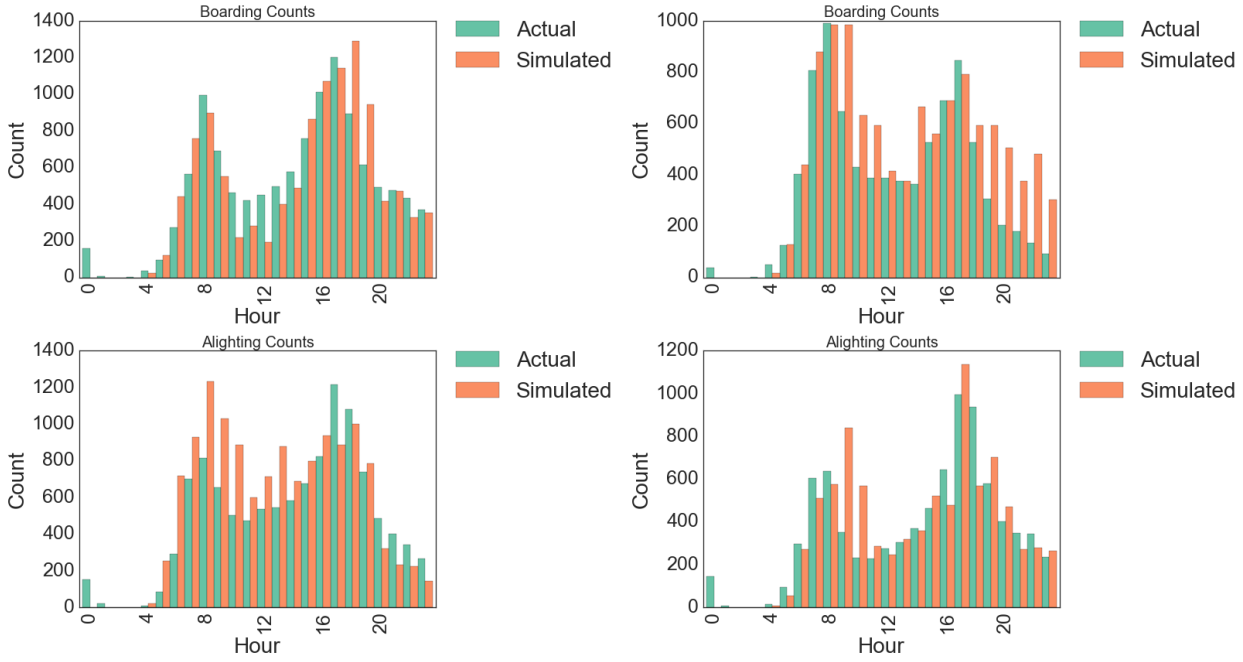


Figure 3.5: Number of activities (labeled per highest posterior probability) by their respective start time within a course of 5 weekdays.



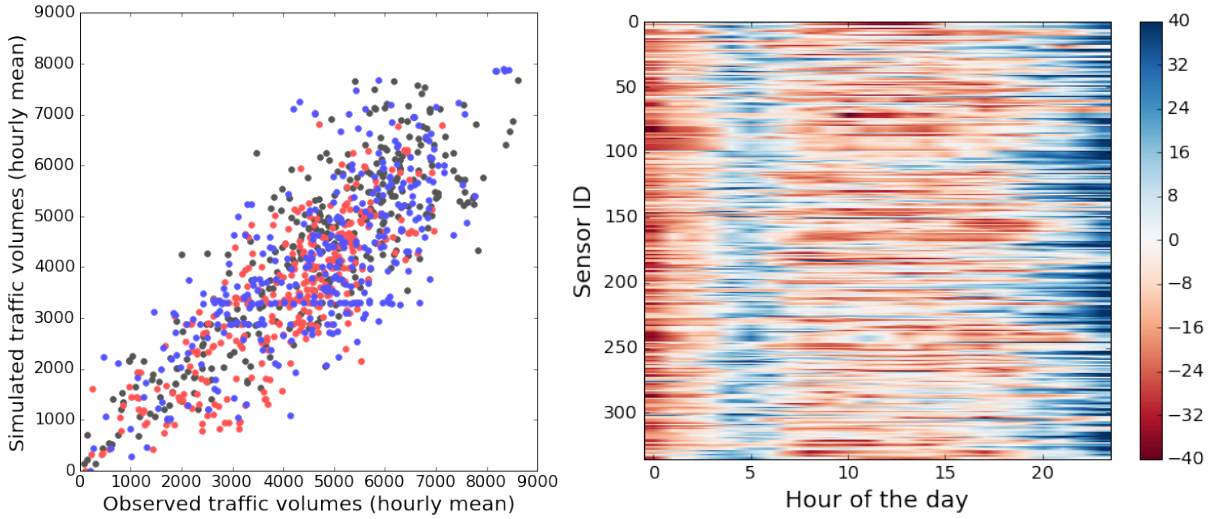
(a) Home (b) Work (c) Other Activities

Figure 3.6: Joint distribution plot of duration and start hour of labeled activities.



(a) 16th St. and Mission Station (b) MacArthur Station

Figure 3.7: Actual and simulated boarding and alighting counts on 2 major BART stations.



(a) Modeled vs observed volumes at 8am (black), 1pm (red) and 6pm (blue).

(b) Mean relative error (%) of modelled vs observed traffic volumes during the day.

Figure 3.8: Micro-simulation validation.

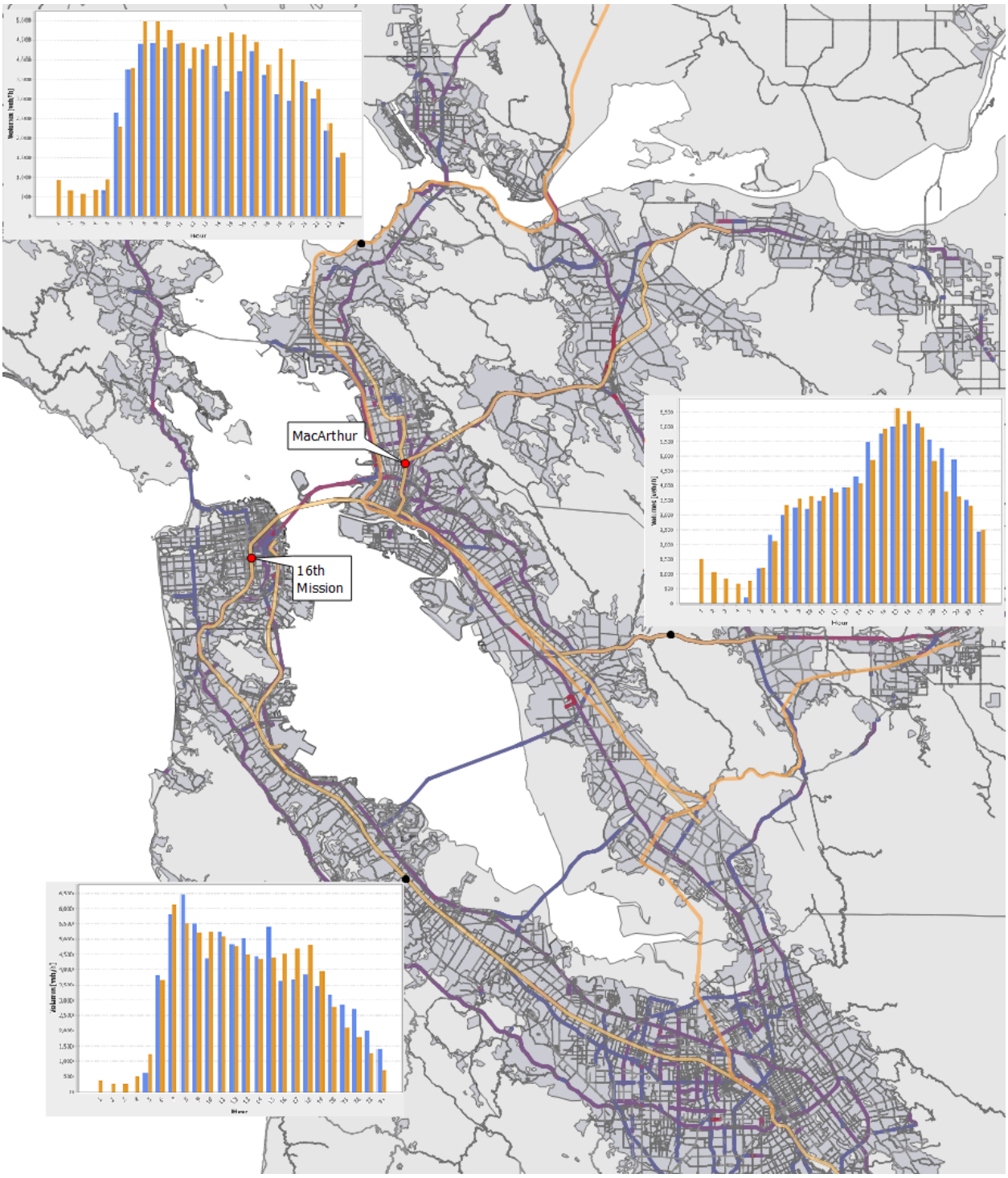


Figure 3.9: A fragment of the multi-modal SF Bay Area network with sample traffic volume detectors and transit stations used for validation. Inset graphs illustrate three sample hourly vehicle volume profiles for observed (orange) and modeled (blue) flows on a typical weekday in June 2015. Sample transit counts histograms are shown in Figure 3.7.

Chapter 4

LSTM Models for Medium-term Human Activity Prediction

4.1 Introduction

Travel demand forecast has been an integral part of most Intelligent Transportation Systems research and applications [77]. Long term forecast provides the basis for transportation planning and scenario evaluation. For example, transportation planners may need to answer many questions. 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 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 frameworks is medium term forecasting, 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 framework 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 mobility models. Cellular data are collected non-invasively for a large proportion of the population and continuously in time. It is also spatially rich compared to detector/camera data, the main data source for short term pre-

diction, as cellular data is ubiquitous and not limited to freeways or arterials. [1]

To validate the medium-term predictions from models, we provided the models with partially observed activity trajectories; the models predict the rest of the activity trajectories. We analyzed the discrepancies between predicted sequences and ground truth sequences at an individual level. We introduced our evaluation method for discrepancies, which included hamming distance between string representation of activity trajectories and difference in travel distances.

The main contributions of this chapter lie in three aspects:

- We proposed and solved a medium term travel demand forecast framework which fills the gap between mainstreams of long term travel demand forecasts 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.

4.2 Existing 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 [17]. 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?

In the recent decades, such forecasts are performed by activity-based models for demographic projections of a population. Activity-based models show how people make decisions about activity participation in the presence of constraints, including decisions about what activity to participate in, where to participate, when to participate, how to get there, and with whom.

Activity scheduling is the central task of an activity-based model. Three main approaches for activity scheduling (constrains-based, utility-based, and rule-based) all require detailed activity diaries data (activity start time, duration, location, transportation mode, etc.) as input [3]. However, the data collection is usually performed through travel surveys that are infrequent, expensive, and reflect the changes in transportation with significant delays. For

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

example, the National Household Travel Survey (NHTS), the data source that is typically the crux of travel demand models, is conducted every 5 years, and carries a total cost of millions of dollars. Thus, travel demand models are mainly targeted at "typical day" travel demand forecasts 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 [77]. Vlahogianni thoroughly summarized the available literature and categorized 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); and (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 areas where detectors/AVI data are 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 [88]. Ye et al. further introduced acceleration information and information from adjacent segments to improve the prediction of the travel speed of current forecasting segment [83]. 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.

4.3 Model Structure

We design a 2-layer LSTM model structure for modeling activity sequences as shown in Fig. 4.1. The first 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). $\boldsymbol{x_t}$ represents the input features, which include current time, day of week and the previous activity type. h_t^1 represents the first layer of LSTM cells and y_t represents the output from the first layer of LSTM cells. The loss function for this first

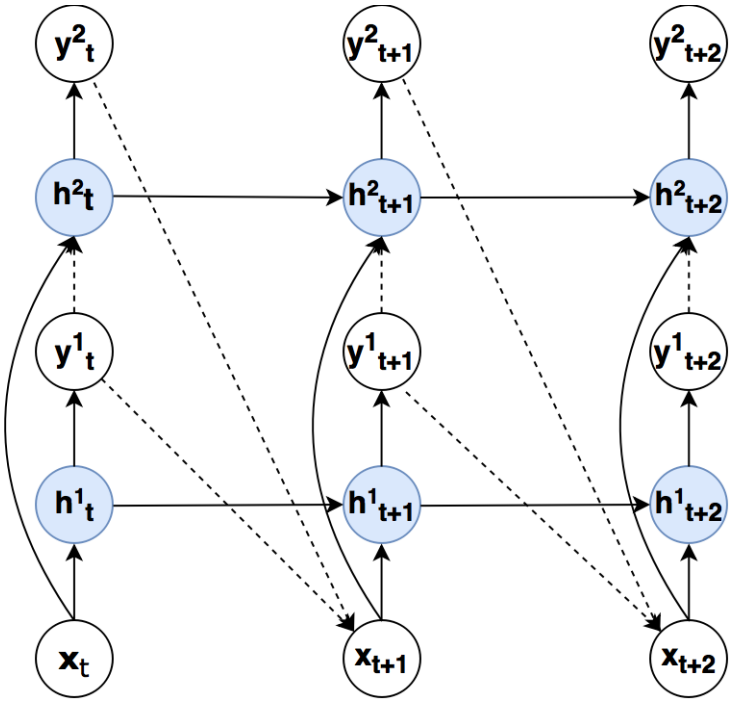


Figure 4.1: LSTM Mobility Model Two-Layer Structure

layer is:

$$\begin{aligned} \boldsymbol{\ell}_1(\boldsymbol{\theta_1}) &= -\sum_{t=1}^{T}\sum_{j}(z_t = j)\cdot\log\phi\left(h_t^1;\boldsymbol{\theta_1}\right)_j \\ &= -\sum_{t=1}^{T}\sum_{j}(z_t = j)\cdot\log p_{t,j} \end{aligned} \tag{4.1}$$

where z_t represents the activity type at t. 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 second 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 [13] and was further developed for handwriting synthesis tasks [34]. The input vector $\boldsymbol{x_t}$, 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. We model the activity start time, duration, latitude, and longitude as mixture Gaussian distributions. We define the decompostion of the LSTM output y_t as

$$\boldsymbol{y}_t = \{\hat{\boldsymbol{\pi}}_t, \hat{\boldsymbol{\mu}}_{lat}, \hat{\boldsymbol{\mu}}_{lon}, \hat{\boldsymbol{\mu}}_{\text{st}}, \hat{\boldsymbol{\mu}}_{\text{dur}}, \hat{\boldsymbol{\sigma}}_{lat}, \hat{\boldsymbol{\sigma}}_{lon}, \hat{\boldsymbol{\sigma}}_{\text{st}}, \hat{\boldsymbol{\sigma}}_{\text{dur}}, \hat{\boldsymbol{\rho}}_{\text{st, dur}}\} \tag{4.2}$$

We provide the transformation from the neural network output to the parameters of mixture distributions.

$$\pi^i = \frac{\exp(\hat{\pi}^i)}{\sum_j^N \exp(\hat{\pi}^j)}; i \in \{1...N\} \tag{4.3}$$

$$\mu_\bullet = \hat{\mu}_\bullet \tag{4.4}$$

$$\sigma_\bullet = \exp(\hat{\sigma}_\bullet) \tag{4.5}$$

$$\rho_\bullet = \tanh(\hat{\rho}_\bullet) \tag{4.6}$$

$$p_i = \frac{\exp(\hat{p}_i)}{\sum_j \exp(\hat{p}_j)}; i, j \in \{\text{home}, \text{work}, \text{other}\} \tag{4.7}$$

At each timestamp t, $\boldsymbol{\pi}_t$ is an M by 1 array representing the mixture component weights; M is the number of mixture components. $\boldsymbol{\mu}_{\text{st},t}$, $\hat{\boldsymbol{\mu}}_{\text{dur},t}$, $\boldsymbol{\mu}_{lat,t}$, and $\hat{\boldsymbol{\mu}}_{lon,t}$ are M by 1 array representing the component means of the activity start time, duration, latitude, and longitude. $\boldsymbol{\sigma}_{\text{st},t}$, $\boldsymbol{\sigma}_{\text{dur},t}$, $\boldsymbol{\sigma}_{lat,t}$, and $\boldsymbol{\sigma}_{lon,t}$, are M by 1 array representing the component standard deviations of the activity start time, duration, and latitude and longitude. $\boldsymbol{\rho}_{\text{st, dur},t}$ 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 the second layer is:

$$\boldsymbol{\ell}_2(\boldsymbol{\theta_2}) = \sum_{t=1}^{T} -\log \sum_i^M \pi_t^i \mathcal{N}(\boldsymbol{x_t} | \hat{\boldsymbol{\mu}}_t^i, \hat{\boldsymbol{\sigma}}_t^i, \hat{\boldsymbol{\rho}}_t^i) \tag{4.8}$$

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. We express the total loss of the LSTM model by combining the loss from both layers.

$$\begin{aligned}
\boldsymbol{\ell} &= \sum_{t=1}^{T} \left[-\log \sum_i^M \pi_t^i p^i(\text{lat}_t, \text{lon}_t, \text{st}_t, \text{dur}_t | \mu_\bullet, \sigma_\bullet, \rho_\bullet) - \log p(\text{type}_t | p_\bullet) \right] \\
&= \sum_{t=1}^{T} \left[-\log \sum_i^M \pi_t^i \mathcal{N}\left(\text{lat}_t, \text{lon}_t \middle| \begin{bmatrix} \mu_{\text{lat}}^i \\ \mu_{\text{lon}}^i \end{bmatrix}, \Sigma_{\text{lat,lon}}^i \right) \mathcal{N}\left(\text{st}_t, \text{dur}_t \middle| \begin{bmatrix} \mu_{\text{st}}^i \\ \mu_{\text{dur}}^i \end{bmatrix}, \Sigma_{\text{st,dur}}^i \right) \right. \\
&\qquad \left. -\log \sum_j j_t p_{t,j} \right]
\end{aligned} \tag{4.9}$$

We derive the derivatives of $\hat{p}_{t,j}$ first.

$$\begin{aligned}\frac{\partial \boldsymbol{\ell}}{\partial \hat{p}_{t,j}} &= -\frac{j_t}{\sum_{j'} p_{t,j'}} p_{t,j}(1 - p_{t,j'}) \\ &= -\frac{j_t p_{t,j'} - j_t p_{t,j'}^2}{\sum_{j'} p_{t,j'}}\end{aligned} \tag{4.10}$$

We then define the terms $\hat{\gamma}_t^i$ and γ_t^i as the following. We can interpret γ_t^i as the "responsitibility" of each mixture component.

$$\hat{\gamma}_t^i = \pi_t^i \mathcal{N}\left(\text{lat}_t, \text{lon}_t \middle| \begin{bmatrix}\mu_{\text{lat}}^i \\ \mu_{\text{lon}}^i\end{bmatrix}, \Sigma_{\text{lat,lon}}^i\right) \mathcal{N}\left(\text{st}_t, \text{dur}_t \middle| \begin{bmatrix}\mu_{\text{st}}^i \\ \mu_{\text{dur}}^i\end{bmatrix}, \Sigma_{\text{st,dur}}^i\right) \tag{4.11}$$

$$\gamma_t^i = \frac{\hat{\gamma}_t^i}{\sum_i \hat{\gamma}_t^i} \tag{4.12}$$

Then, the derivatives are of the loss function with respect to parameters are as follows:

$$\frac{\partial \boldsymbol{\ell}}{\partial(\hat{\mu}_{\text{lat}}^i, \hat{\mu}_{\text{lon}}^i, \hat{\sigma}_{\text{lat}}^i, \hat{\sigma}_{\text{lon}}^i)} = -\gamma_t^i \frac{\partial \log \mathcal{N}\left(\text{lat}_t, \text{lon}_t \middle| \begin{bmatrix}\mu_{\text{lat}}^i \\ \mu_{\text{lon}}^i\end{bmatrix}, \Sigma_{\text{lat,lon}}^i\right)}{\partial(\hat{\mu}_{\text{lat}}^i, \hat{\mu}_{\text{lon}}^i, \hat{\sigma}_{\text{lat}}^i, \hat{\sigma}_{\text{lon}}^i)} \tag{4.13}$$

$$\frac{\partial \boldsymbol{\ell}}{\partial(\hat{\mu}_{\text{st}}^i, \hat{\mu}_{\text{dur}}^i, \hat{\sigma}_{\text{st}}^i, \hat{\sigma}_{\text{dur}}^i, \hat{\rho}_{\text{st,dur}})} = -\gamma_t^i \frac{\partial \log \mathcal{N}\left(\text{st}_t, \text{dur}_t \middle| \begin{bmatrix}\mu_{\text{st}}^i \\ \mu_{\text{dur}}^i\end{bmatrix}, \Sigma_{\text{st,dur}}^i\right)}{\partial(\hat{\mu}_{\text{st}}^i, \hat{\mu}_{\text{dur}}^i, \hat{\sigma}_{\text{st}}^i, \hat{\sigma}_{\text{dur}}^i, \hat{\rho}_{\text{st,dur}})} \tag{4.14}$$

where in Eq.4.13 and Eq.4.14

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\mu}_1^i} = \frac{C}{\sigma_1^i}\left(\frac{x_1^i - \mu_1^i}{\sigma_1^i} - \frac{\rho^i(x_2^i - \mu_2^i)}{\sigma_2^i}\right) \tag{4.15}$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\mu}_2^i} = \frac{C}{\sigma_2^i}\left(\frac{x_2^i - \mu_2^i}{\sigma_2^i} - \frac{\rho^i(x_1^i - \mu_1^i)}{\sigma_1^i}\right) \tag{4.16}$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\sigma}_1^i} = \frac{C(x_1^i - \mu_1^i)}{\sigma_1^i}\left(\frac{x_1^i - \mu_1^i}{\sigma_1^i} - \frac{\rho(x_2^i - \mu_2^i)}{\sigma_2^i} - 1\right) \tag{4.17}$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\sigma}_2^i} = \frac{C(x_2^i - \mu_2^i)}{\sigma_2^i}\left(\frac{x_2^i - \mu_2^i}{\sigma_2^i} - \frac{\rho(x_1^i - \mu_1^i)}{\sigma_1^i} - 1\right) \tag{4.18}$$

$$\frac{\partial \boldsymbol{\ell}}{\partial \hat{\rho}^i} = \frac{(x_1^i - \mu_1^i)(x_2^i - \mu_2^i)}{\sigma_1^i \sigma_2^i} + \rho^i(1 - C^i Z^i) \tag{4.19}$$

where

$$C^i = \frac{1}{1-\rho^{i2}} \tag{4.20}$$

$$Z^i = \frac{(x_1^i - \mu_1^i)^2}{\sigma_1^{i2}} + \frac{(x_2^i - \mu_2^i)^2}{\sigma_2^{i2}} - \frac{2\rho^i(x_1^i - \mu_1^i)(x_2^i - \mu_2^i)}{\sigma_1^i \sigma_2^i} \tag{4.21}$$

Fig 4.2 shows an example of sequence generation. The LSTM model was trained on an individual's data, given a scenario that at 10am and a previous "home" activity. The first layer predicts the next activity type with "other" being the most probable type and "work" being the second most probable type. If "work" type is sampled, the second layer outputs the typical work hours and work location for that individual. If "other" type is sample, the second layer would predict the corresponding location and duration.

4.4 Evaluation Method

We introduce the evaluation method. For human activities, discrepancies between two trajectories consist of differences in spatial-temporal characteristics and activity types along the trajectories. A question one might ask is that how to compare human activity trajectories with different number of activities (i.e. different in length)? Fig 4.3, we show an example of comparing two similar daily activity trajectories with the only difference being the additional shopping activity in the second trajectory. If we align the activities based on their start time and compare the activity types, there is only 50% match between the 2 trajectories. However, the majority of the daily structures for the two trajectories are the same. A better measurement is to compute the hamming distance between the string representation of two trajectories. As shown in Fig 4.3, activity types along a trajectory can be binned the by fixed time interval so that the trajectory can be represented as a string. In our experiments, the model inputs are daily trajectories such that the string representations all have the same length. Hamming distances between two strings are computed as the minimum number of characters to be replaced such that the two strings becomes the same.

The hamming distance gives measurement of temporal characteristics and activity types between trajectories. Evaluation of the spatial characteristics between two trajectories is designed as comparing the total traveling distances of each trajectory. The travel distances between two stationary points are approximated using Euclidean distance. In Fig 4.4, we show an example that the same activity sequence "Home-Work-Shopping-Home" results different total travel distances because of different location choices.

4.5 Experiment Setup

We performed 2 experiments to evaluate the performance of the LSTM model. In the first experiment, we evaluated the model prediction on a single given day, and the prediction

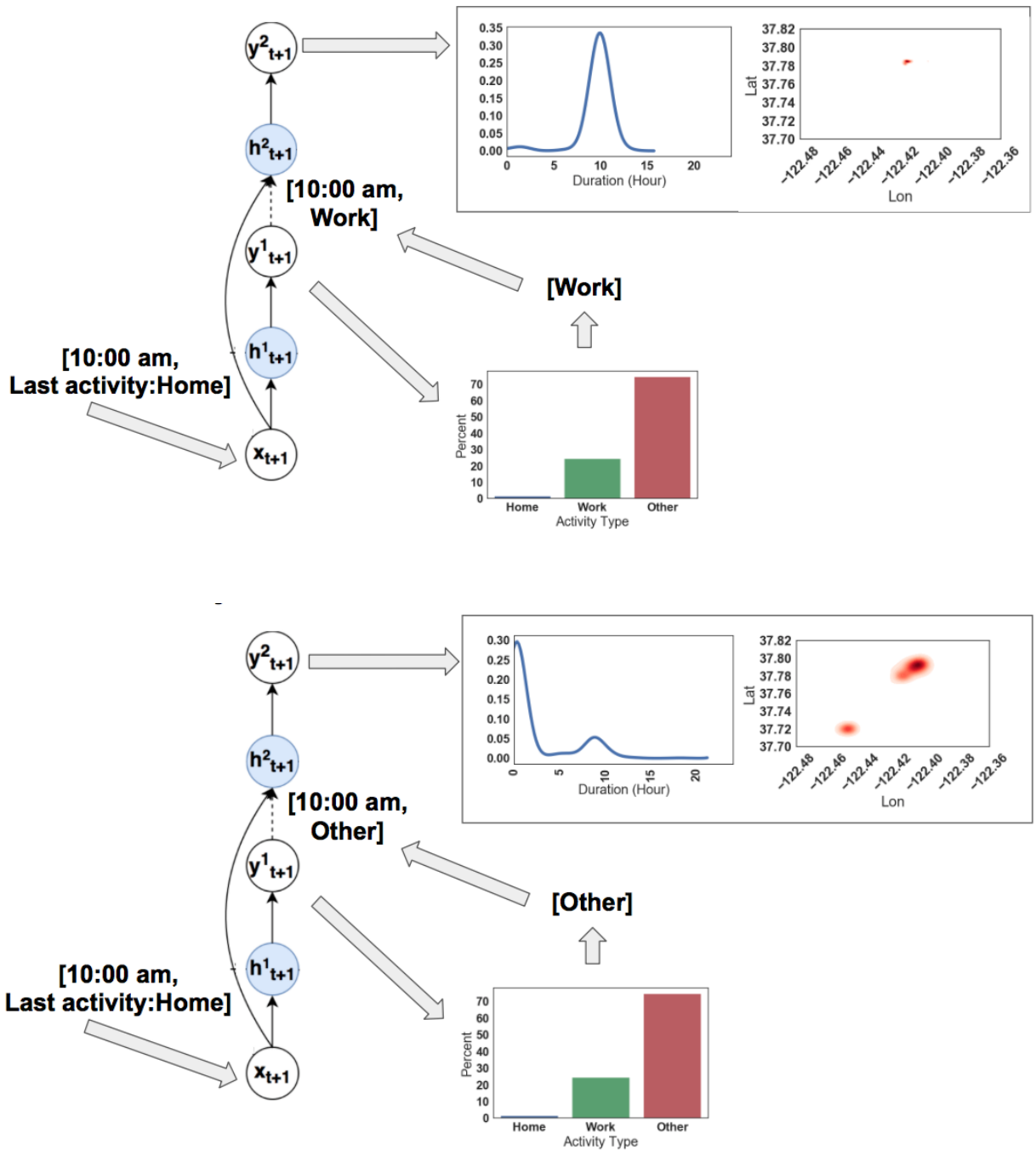


Figure 4.2: Two layer model sampling a work related activity (top) and a secondary activity (bottom).

error is reported. In the second experiment, we used cross-validation for evaluation, and the cross-validation error is reported.

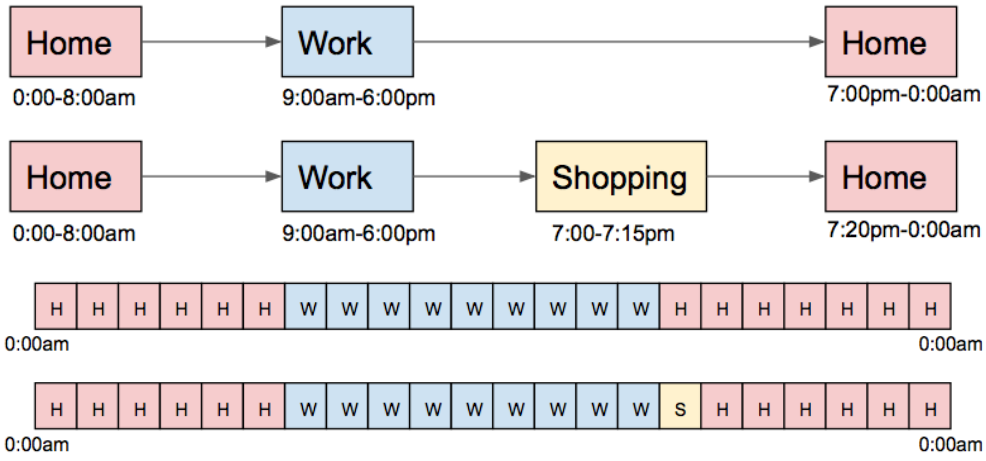


Figure 4.3: An example of minor differences in 2 activity sequences (top) and their corresponding string representations (bottom).

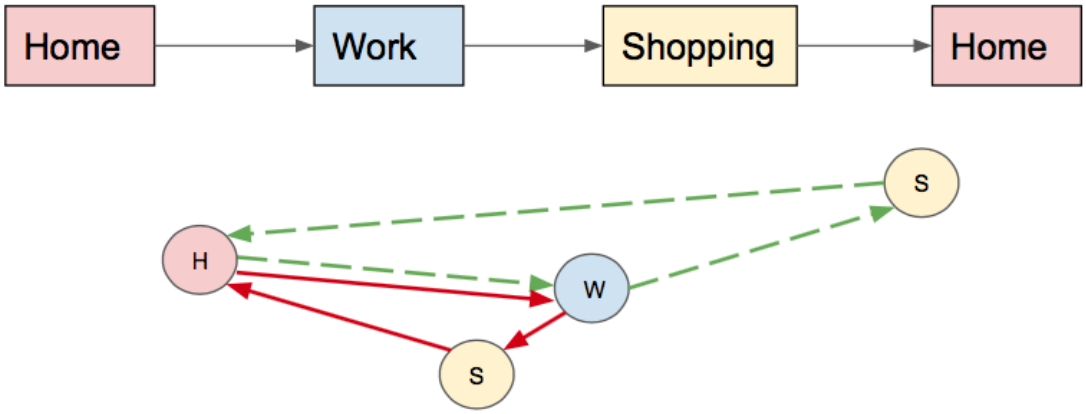


Figure 4.4: An example shows different location choices that affects the total traveling distance in a day.

In the first experiment, we performed prediction on a particular day. We used one month of CDR data in the San Francisco Bay Area. The population covered mainly consisted of commuters and travelers and are excluded. We assume the model receives streaming CDR data at different time of day (e.g. 3:00 am, 9:00 am, 3:00 pm, etc.), which were then processed to partially observed activity sequences. These partially observed sequences, along with the pre-trained parametric urban mobility models, were sent to the sequence predictor. The sequence predictor predicted and completed the activity sequences for the rest of the day based on the observed information. By the end of the day, full day CDR was observed and the ground truth activity sequences were validated against the predicted activity sequences

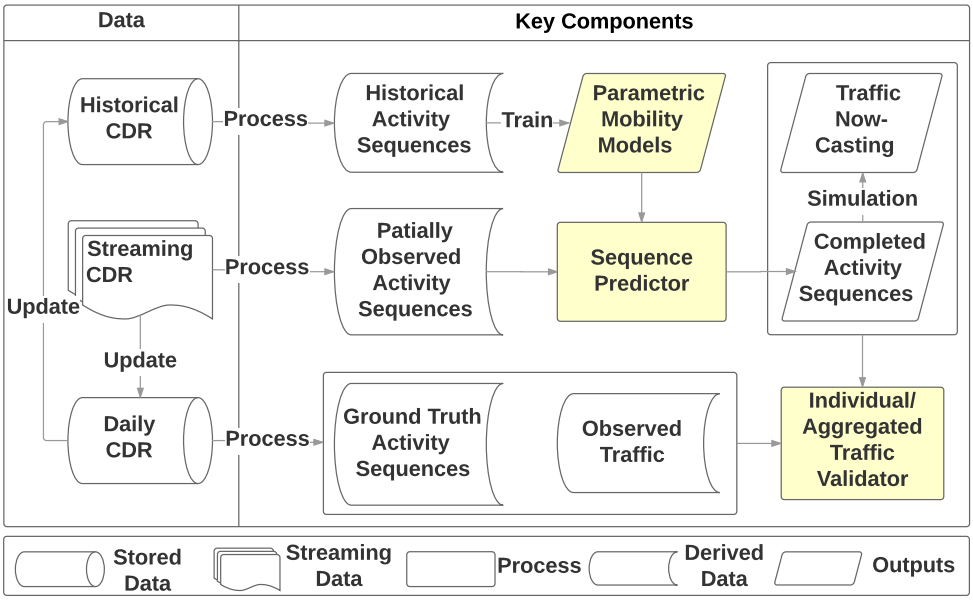


Figure 4.5: Modeling framework diagram. The left column represents the input to the algorithms and the right column represents the model components. Our key contribution of improved deep urban mobility models, sequence predictor, and validation are shown in shaded yellow.

at both individual level. The developed data processing and modeling pipeline is presented in Fig. 4.5. Anonymized historical CDR data are processed to unlabeled historical activity sequences [84]. We evaluated LSTM models with different model setting against baseline models including IOHMM models and nearest neighbour models.

In the second experiment, we used a different locational data set in the San Francisco Bay Area. This data set included 2 month location data with travelers. We performed 5-fold cross-validation and compared the prediction performance between different models.

4.6 Results

In this section, we compare the prediction power of different models with partially observed sequences. In the first experiment, the models were trained with one-month location data with the typical day of interest excluded. The validation process is to feed the models with observed activities up to the "cut hour," and the models generated the corresponding rest of the sequences. The validation metrics by comparing the generated parts with the original sequences were reported.

We chose a typical day to be June 10, 2015 for validation. We experimented "cut hours" that incremented from 3:00 am to 11:00 pm with intervals of 1 hour. The LSTM model used both layers of LSTM cells with 64 units. Number of output mixture components, N, was

chosen as 40. We used 10% dropout rate for the LSTM units to prevent over fitting of the data. For Adam optimizer, the learning rate, β_1, and β_2 were choose as 0.001, 0.9, and 0.999, respectively. The sampling bias b was tuned as 2.0.

The evaluation included the following models.

1. **LSTM-3**: The **two-layer model** structure with 3 activity type labels "home", "work" and "other".

2. **LSTM-7**: The **two-layer model** structure with 7 activity type labels which includes 2 types of "home" labels and 4 types of "work" labels and 1 type of "other" label.

3. **NN**: Nearest Neighbor model, the benchmark model, and the expected upper bound of the performance. NN is a fully personalized model that matches the observed trajectory with the trajectory history of the user, and uses the matched trajectory as prediction for the rest of day. The distance features we used were (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 calculated 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 were the same (in most of the 15 minutes) and 1 if not. The total Hamming distance is the sum of each segment. We gave the day type feature a high weight (in this case 100) so that NN would 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.

4. **IOHMM-unsupervised-7**: The IOHMM model with 7 hidden states, with the input and output features specified in [84].

5. **IOHMM-co-training-7**: The co-training IOHMM model specified in [84] with 7 activity type labels.

6. **IOHMM-co-training-11**: The co-training IOHMM model specified in [84] with 11 activity type labels.

In Fig. 4.6, 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 is 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. 4.6, we observe that the NN model performs best among all models because it is a fully personalized non-parametric model. The LSTM models are better at capturing the

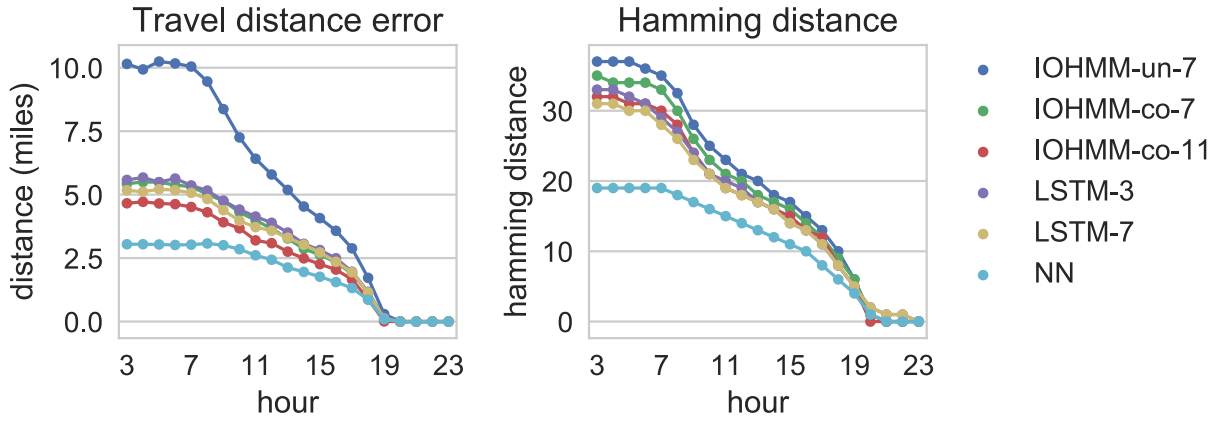


Figure 4.6: Models comparison. Two validation metrics are used: median travel distance error (left) and median Hamming distance (right). The x-axis is the prediction hour (cut hour) and the y-axis is the validation error. Each series of points represents the performance of a model.

day structures. Hamming error captures the performance of day structures such as "home," "work," and important secondary activities. Comparing LSTM-7 and IOHMM-co-7 with the same activity labels, LSTM models slightly outperforms IOHMM models because it is able to capture deeper activity structures and long-term activity dependencies.

In our second experiment, we used a different location data set in the San Francisco Bay Area and 5-fold cross-validation was performed. We randomly selected 1000 individuals for comparing performance of individual level models between LSTM and NN. We used the same range of "cut hours," which was from 3:00 am to 11:00 pm. Fig 4.7 shows the cross-validation error of individual level LSTM and NN models. In general, the difference in performance between the two individual level models is small. Compared to the first experiment, we observe similar hamming distance error for daily activity structure. The traveling distance error increases significantly due to the included distant travelers in the data set. The NN model shows lower prediction error at "cut hours" of 3:00 am while the LSTM model shows better performance when "cut hours" increases, i.e. greater parts of the sequences are observed.

4.7 Discussion

In this chapter, we proposed a medium term travel demand nowcasting framework. It predicts daily travel demand 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.

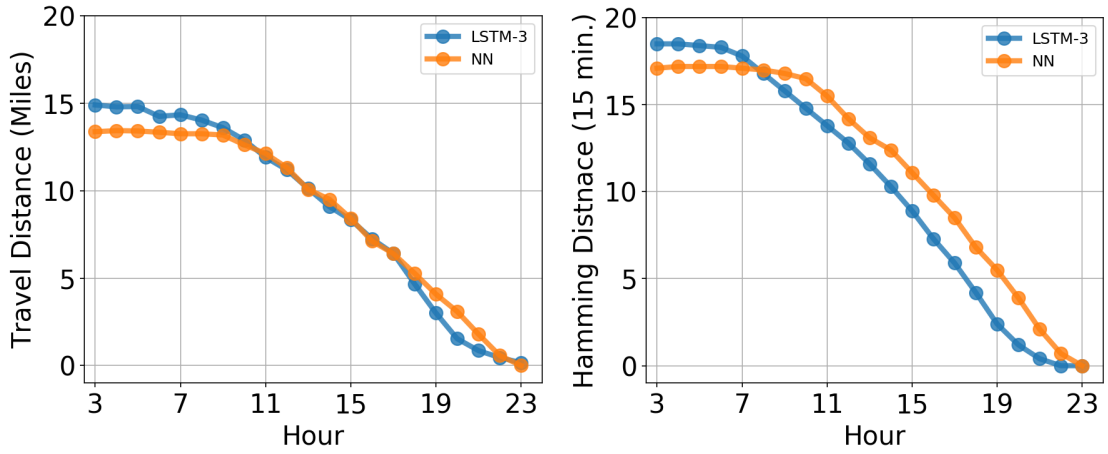


Figure 4.7: Comparing individual level LSTM model to nearest neighbour model in terms of median hamming distance (left) and median travel distance (right).

We presented the state-of-the-art LSTM model and its performance against other baseline models. 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 individual level discrepancies. 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 an individual level can approach this limit in terms of prediction accuracy. Results from aggregated level models showed that by increasing the number of activity types, there was an increase in the prediction accuracy in terms of both spatial and temporal features. Among the generative models we compared, IOHMM models are interpretable and have the power of activity recognition as a range of travel choices might depend on the activity types. LSTM models on both aggregated and individual level are better at learning day structures since they use continuous hidden state space and are expected to be better at learning long term dependencies.

Chapter 5

LSTM Models with Contextual Information

5.1 Introduction

Mobility models characterize multiple aspects of individuals' behavior patterns. In transportation research, one important focus is to understand and explain how social-demographic characteristics affect individuals' decisions on choices of transportation, such as traveling modes and traveling distances, etc. Discrete choice models are often used for inferring the influence of social-demographic characteristics to choices under specific scenarios. It is usually difficult for discrete choice model framework to be generalized for modeling sequential choice involving multiple decisions simultaneously. In computer science research, on the other hand, large amounts of work focus on modeling and predicting human mobility on the individual level, such that the characteristic of individuals are often ignored. Other research on aggregated level focus on summarizing human mobility statistics and mobility law without including social-demographic information. Many Markov and recurrent neural network models of sequences have been developed, but they have not been applied and evaluated on the human mobility domain. LSTM models, for example, have been successfully applied on image captioning problems. Convolutional neural networks (CNN) extract features from images, which are provided to the LSTM models as contextual information so that the LSTM models generate precise language to describe the input image. In transportation research, we face similar problems of learning human mobility sequences with characteristics of individuals. Thus, similar approaches can be applied that characteristics of individuals can be included as features into the initial input of the LSTM models. LSTM models will be trained with aggregated mobility data with characteristics of individuals. Thus, the model learns correlation between the given contextual information and the characteristic of the corresponding trajectories. This leads us to research opportunities of developing a generalized model for sequential mobility decisions that include population characteristics.

One major difficulty for this research is the limitation of data. Survey data usually

includes individual characteristics, as well as self-reported mobility sequences. But survey data usually has limitations that the number of activities an individual is able to recall is minimal. For example, in the National Household Traveling Survey (NHTS), participants report only one day of traveling history. Therefore, the spatial-temporal richness of survey data is significantly lower compared to other data sources such as locational data. On the other hand, locational data, such as CDR data, is intentionally anonymized to protect the privacy of the device holder; thus, it lacks contextual information.

In this chapter, we introduce an extension of the LSTM mobility model by including characteristic information of individuals. We demonstrate the model structure, and the decomposition of input and output from the model. To illustrate the learning power of the LSTM model, we performed 3 experiments using discrete contextual information and continuous contextual information. In the first experiment, we showed the performance of our model using binary labels by diving individuals based on the travel habits. In the second and the third experiments, we showed the capability of the models to handle continuous contextual information. We use radius of gyration, travel distance and activity starting time as measurements to compare the generated sequences to the observed sequences.

The main contributions of this chapter lie in two aspects:

- We proposed an end-to-end activity sequence model that includes contextual information.
- We demonstrate the capability of the LSTM models to generate activity sequences using either discrete or continuous contextual information.

5.2 Existing Work

Human Mobility and Contextual Information

Human mobility is significantly influenced by external environment, such as weather and hazard, and internal characteristics, such as age and income. Many existing works are limited on describing and developing statistics on these influences. Documents and researches reported statistics of human mobility affected by climate change [14] and hazard related disasters [54] in developing countries. Even weather change can affect the volume and distance of the movements in urban areas [57]. The effect of internal characteristics, such as age and gender, towards mobility is studied and quantified using credit card transaction data [47].

Discrete Choice Models

Multinomial logit (MNL) models were widely applied for inferring influence of social-demographic information toward traveling behavior for decades. Researches are mainly focused on modeling individual's traveling model choices under specific conditions[7]. The advantage of MNL models is that the high interpretability of parameters about the relative importance

of social-demographic features, and the parameter's statistical significance, can be easily estimated. By relying on the assumption of linear dependency between dependent and independent variables, MNL is able to predict the probability of each proposed choice even when the scenario is not observed. The disadvantage of the MNL models, apart from the linearity assumption, is that they are mainly suitable for modeling single choices. Although MNL models have been extended for sequential choices of daily activity patterns, it requires manually design aggregation of daily activity trajectories in to a few patterns so the data format is suitable for the model framework [11].

Markov Models with Contextual Information

Markov Models (MM's) and Hidden Markov Models (HMM's) are commonly used for modeling sequences such as speech and natural language and sequential human decisions. HMM's use a number of discrete hidden states and the transition and emission probabilities are learned from data. Two versions of HMM's that utilize context information have been developed. The first version is call Contextual Hidden Markov Models; the transition probability $p(x_t|xt-1, c_t)$ depends on both the previous hidden state x_{t-1} and the current context information c_t [59]. The second version is called Input Output Hidden Markov Model (IOHMM). In IOHMM, both state transition probability $p(x_t|xt-1, c_t)$ and emission probability $p(y_t|xt, c_t)$ are functions of contextual information c_t [9]. The advantage for HMM's is that they are suitable for sequence of arbitrary length, and they are capable for unsupervied learning of sequences. The disadvantage of the HMM's is that they need the manual selection of the number of hidden states. The contextual variables often need to be carefully crafted in order for the HMM's to perform well.

LSTM Models with Contextual Information

Generating sequence with influences of contextual information has been developed in image caption generation [76, 86, 81]. For image captioning problems, convolutional neural networks (CNN) extract features from images and provide them to the LSTM models as contextual information so that the LSTM models generate precise language to describe the input image. Unlike HMM's, which needs contextual information in every time step, the LSTM models need only a single time input of the contextual information, which usually happens at the beginning of sequence generation. Back propagation through time (BPTT) would help the LSTM models learn the joint distribution between the contextual distribution and the entire sequence. The advantage of LSTM models is the continuous hidden states that gives the flexibility of modeling arbitrary distributions with long-term dependencies. In our human mobility sequence learning problem, similar approaches can be applied that the social-demographic information can be included as features into the initial input of the LSTM models. LSTM models are going to be trained with aggregated mobility data with contextual information (e.g. social-demographic information); the models learn the correla-

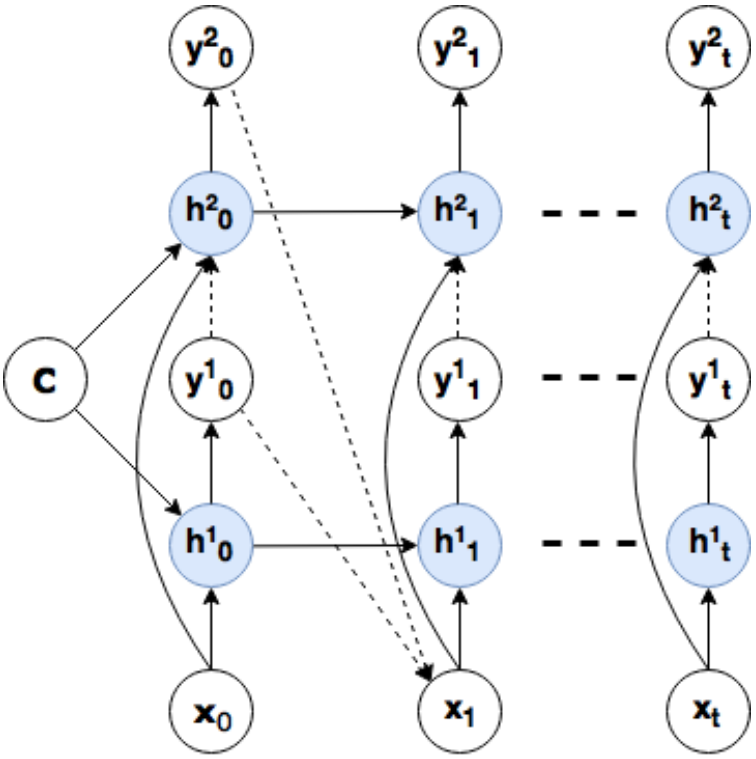


Figure 5.1: LSTM Mobility Model with Contextual Information

tion between the given contextual information and the characteristics of the corresponding trajectories.

5.3 Model Structure

In this section, we describe the LSTM mobility model with contextual information. The model structure is shown in Fig 5.1. We denote the target activity sequences as S and contextual information for the sequences as c. There are two layers in the model responsible for learning activity types and spatial-temporal distributions for the next activity given input x_t. The contextual information c is passed into both layers only at timestamp 0, while features of previous activity x_{t-1} is passed into both layers at each step, t.

To summarize, our model generates activity sequences according to the following four steps:

1. At timestamp 0, include contextual information c into model input.
2. At every step, t, the LSTM layer(s) receives a set of inputs x_t.
3. The LSTM layer(s) then produce a set of outputs y_t, which is used to parametrize a mixture distribution $p(x_{t+1}|y_t)$.

4. A new activity x_{t+1} is sampled from distribution $p(x_{t+1}|y_t)$.

Input variables, context variables, and output variables

We design the input variables, x_t, such that it contains the features that affect the choices of the next activity. Those features include the current time and the type of the previous activity. Current time is a continuous variable, while activity type is a categorical variable. Activity types consist of the labels of each activity including "home" "work," and "others," which are encoded as a one-hot vector. Thus, x_t can be decomposed as the following:

$$x_t = \{\text{start\_time}_t, \text{activity\_type}_{t-1}, \text{day\_of\_week}_t\} \tag{5.1}$$

The context variable c contains characteristics of individuals. We denote it as the following:

$$c = \{\text{Individual characteristics}\} \tag{5.2}$$

Output variables y_t are decomposed and transformed into coefficients of mixture distribution $p(x_{t+1}|y_t, c_{t+1})$, which is used for generating the next x_t. The first layer outputs the categorical distribution which is used for sampling activity types. And the second layer outputs the 2-dimensional mixture Gaussian distribution for sampling activity duration, and a categorical distribution for sampling the location ID of the activity. Hence, the outputs from the two LSTM layers, y_t^1 and y_t^2, can be split as the following:

$$y_t^1 = \{\hat{p}_{\text{activity types}}\} \tag{5.3}$$

$$\boldsymbol{y_t^2} = \{\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}_{\text{st}}, \hat{\boldsymbol{\mu}}_{\text{dur}}, \hat{\boldsymbol{\sigma}}_{\text{st}}, \hat{\boldsymbol{\sigma}}_{\text{dur}}, \hat{\boldsymbol{\rho}}_{\text{st, dur}}, \hat{\boldsymbol{p}}_{\text{activity type}}, \hat{\boldsymbol{p}}_{\text{location ID}}\} \tag{5.4}$$

Output transformation Those raw outputs from LSTM are properly transformed before serving as mixture distribution parameters. The component weights $\hat{\boldsymbol{\pi}}$ and probability of activity type within each component are normalized using softmax function. Standard deviations $\hat{\boldsymbol{\sigma}}_{\bullet}$ are constrained to be non-negative using an exponential function with correlation coefficients: $\hat{\boldsymbol{\rho}}_{\bullet}$, scaled beween -1 and 1 using tanh activation functions. The following equations summarize our model:

$$\pi^i = \frac{\exp(\hat{\pi}^i)}{\sum_j^N \exp(\hat{\pi}^j)}; i \in \{1...N\} \tag{5.5}$$

$$\mu_{\bullet} = \hat{\mu}_{\bullet} \tag{5.6}$$

$$\sigma_{\bullet} = \exp(\hat{\sigma}_{\bullet}) \tag{5.7}$$

$$\rho_{\bullet} = \tanh(\hat{\rho}_{\bullet}) \tag{5.8}$$

$$p_i = \frac{\exp(\hat{p}_i)}{\sum_j \exp(\hat{p}_j)}; i, j \in \{\text{activity types}\} \tag{5.9}$$

$$p_i = \frac{\exp(\hat{p}_i)}{\sum_j \exp(\hat{p}_j)}; i, j \in \{\text{location categories}\} \tag{5.10}$$

Output Distributions

We define the distributions output from the neural network that models activity types, location, start time, and duration. The activity type is modeled by a categorical distribution that is directly parameterized by the first layer output.

$$\begin{aligned} &p(\text{type}_t|y^1_{t-1}) = p_j; \\ &j \in \{\text{home}, \text{work}, \text{others}\} \end{aligned} \tag{5.11}$$

Start time and duration of activities are modeled by a two-dimensional Gaussian distribution. Location ID's are also modeled by a categorical distribution. We define the spatial and temporal joint distribution as the following:

$$\begin{aligned} &p(\text{location ID}_t, \text{st}_t, \text{dur}_t|y^2_{t-1}, \text{type}_t) = \\ &p(\text{location ID}_t) \sum_i^N \pi^i p^i(\text{st}_t, \text{dur}_t|\mu_\bullet, \sigma_\bullet, \rho_\bullet, \text{type}_t) \end{aligned} \tag{5.12}$$

$$p^i(\text{st}_t, \text{dur}_t) = \mathcal{N}\left(\text{st}_t, \text{dur}_t \middle| \begin{bmatrix} \mu^i_{\text{st}} \\ \mu^i_{\text{dur}} \end{bmatrix}, \Sigma^i_{\text{st,dur}}\right) \tag{5.13}$$

where

$$\Sigma^i_{\text{st,dur}} = \begin{bmatrix} \sigma^2_{\text{st}} & \sigma_{\text{st}}\sigma_{\text{dur}}\rho_{\text{st, dur}} \\ \sigma_{\text{st}}\sigma_{\text{dur}}\rho_{\text{st, dur}} & \sigma^2_{\text{dur}} \end{bmatrix} \tag{5.14}$$

Sequence generation

Sequence generation is initialized by feeding the LSTM model with the context information c and a default vector x_0. At each timestamp with the neural network output y_t, we sample location, duration, activity type from $p(x_{t+1}|y_t)$. The start time and duration of an activity is jointly distributed and is modeled as a two-dimensional Gaussian distribution. Since the start time of an activity is observed before the activity is sampled, we need to compute the distribution of next activity duration conditional on the observed start time. We show the parameters of the conditional mixture distribution as the following. The conditional mixture weights are expressed as follows:

$$w^i(\text{st}_t) = \frac{\pi^i \mathcal{N}(\text{st}_t|\mu_{\text{st}}, \sigma_{\text{st}})^i}{\sum_j^N \pi^j \mathcal{N}(\text{st}_t|\mu_{\text{st}}, \sigma_{\text{st}})^j} \tag{5.15}$$

The mean and the standard deviation of activity duration conditioned on the observed starting time, st_t, are expressed as follows:

$$\mu_{\text{dur}|\text{st}_t} = \mu_{\text{dur}} + \frac{\sigma_{\text{dur}}}{\sigma_{\text{st}}}\rho_{\text{st,dur}}(\text{st}_t - \mu_{\text{st}}) \tag{5.16}$$

$$\sigma_{\text{dur}|\text{st}_t} = \sqrt{(1 - \rho^2_{\text{st,dur}})}\sigma_{\text{dur}} \tag{5.17}$$

Now we can sample a new activity x_t from the mixture distribution $p(x_t|y_{t-1})$ following Eq. 5.18 through Eq. 5.21. First, a component is sampled from the multinomial distribution of mixture weights (Eq. 5.15). Then, dur_t, lat_t, lon_t, $type_t$, end_t can be further sampled from the selected component, yielding a joint distribution of all those variables. Once a new activity is sampled, the time of day is incremented by dur_t.

$$\mathrm{k} \sim \text{Multinomial}\Big(w(\mathrm{st}_t)_1, ..., w(\mathrm{st}_t)_N; n = 1\Big) \tag{5.18}$$

$$\mathrm{dur}_t \sim \mathcal{N}\Big(\mu^k_{\mathrm{dur}|\mathrm{st}_t}, \sigma^k_{\mathrm{dur}|\mathrm{st}_t}\Big) \tag{5.19}$$

$$\mathrm{type}_t \sim \text{Multinomial}\Big(p^k_{\mathrm{home}}, p^k_{\mathrm{work}}, p^k_{\mathrm{others}}; n = 1\Big) \tag{5.20}$$

$$\text{location ID}_t \sim \text{Multinomial}\Big(p^k_{\mathrm{ID}_1}, ..., p^k_{\mathrm{ID}_M}; n = 1\Big) \tag{5.21}$$

Model Estimation and Loss Function

We use negative log-likelihood as the loss of the model. The LSTM model is parameterized by θ. Thus, we propose to directly maximize the probability of the activity sequences given the contextual information by using the following:

$$\theta* = \arg\max_{\theta} \sum_{S,c} \log p(S|c;\theta) \tag{5.22}$$

We denote the activity at timestamp t along the activity sequence S as S_t. We limited the maximum number of activities in a sequence to be T. The joint probability of over S_1 ... S_T can be expressed as following:

$$\log p(S|c) = \sum_{t=0}^{T} \log p(S_t|c, S_1...S_{t-1}) \tag{5.23}$$

The output from the neural network y_t carries information accumulated in the neural network memory from all previous steps. We can further substitute the joint probability by the output distributions that we defined above:

$$\begin{aligned} &\log p(S_t|c, S_1...S_{t-1}) = \\ &- \log p(\mathrm{types}_t|y^1_t) - \log p(\text{location ID}_t, \mathrm{st}_t, \mathrm{dur}_t|y^2_t) \end{aligned} \tag{5.24}$$

Put it together, we write the loss function give sequence and contextual information pair (S, c) as the following:

$$\boldsymbol{\ell} = \sum_{t=1}^{T} \Bigg[- \log p(\mathrm{types}_t|y^1_t) - \log p(\text{location ID}_t, \mathrm{st}_t, \mathrm{dur}_t|y^2_t) \Bigg] \tag{5.25}$$

5.4 Experiment Setup

We present the experiments using discrete contextual information and continuous contextual in the following subsections. We use smartphone location data collected from individuals in the metropolitan area of Toronto, Canada. The data set contains two months (60 days) of data. To extract the place ID for location choices, we perform DBScan clustering on the aggregated location history from all individuals. The DBScan radius ϵ is chosen as 100 meters. For the LSTM model tunings, we use 64 units in both LSTM layers. The number of mixtures for the temporal mixture distribution is selected as 100. We use Adam optimizer [44] for optimization with learning rate of 0.001.

Discrete Contextual Information

In the first experiment, we used discrete contextual information. We labelled 50 individuals using binary labels based on the radius of gyration of their traveling trajectories. We selected the first group of 25 individuals with criteria that each individual had more than 40 days of activity sequences with less then 3 km of radius of gyration. This group was labeled as "Low Radius." The second group of 25 individuals were selected with criteria that each individual had more than 40 days of activity sequences with radius of gyration between 20 and 30 km. This group was labeled as "High Radius." In Fig 5.2, we show the spatial and temporal characteristics of both groups. For the "Low Radius" group, the daily travel distances were concentrated around 10 km, while for "High Radius" group, the daily travel distances concentrated around 60 km and 120 km. In the activity start time histogram, the "High Radius" group showed patterns of morning and evening commute while the "Low Radius" group's activity was more evenly distributed with some concentration in the middle of the day.

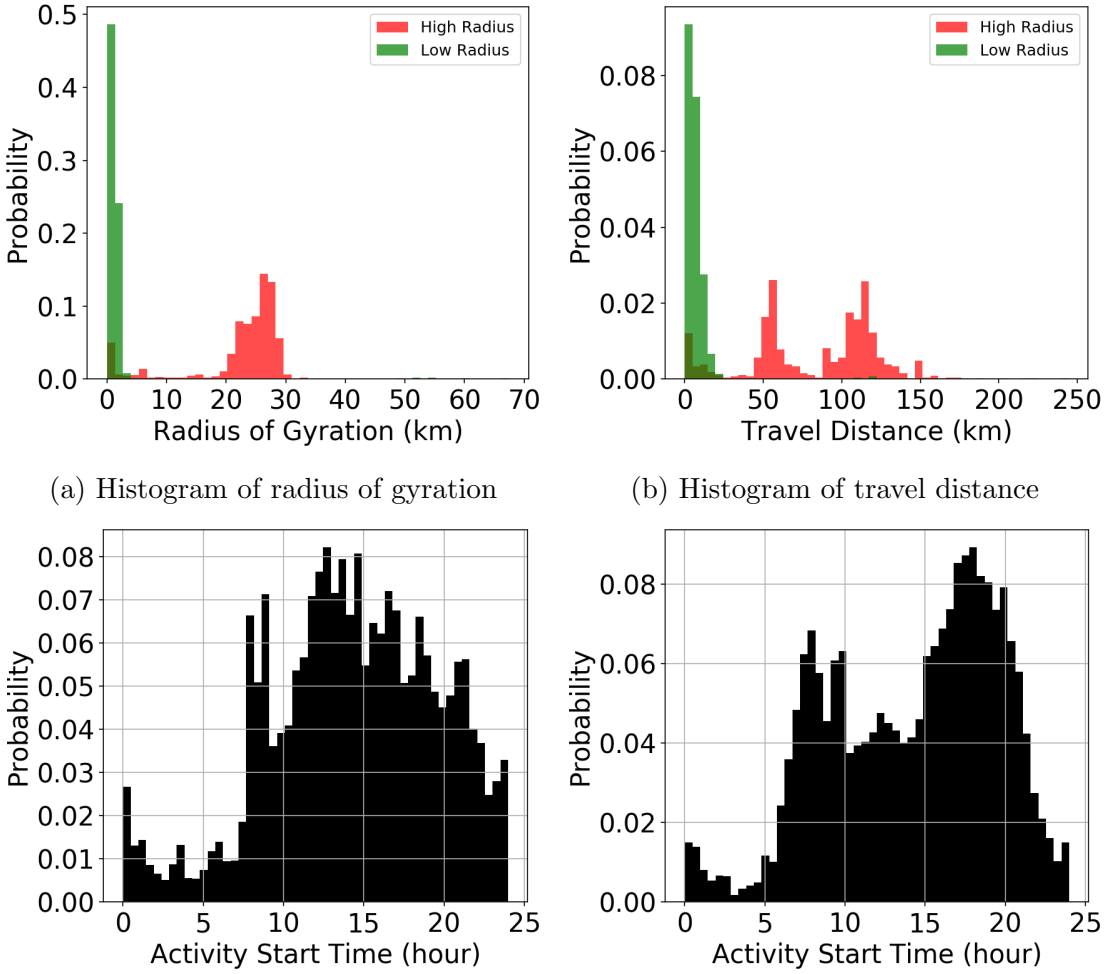
We evaluated similarities between spatial-temporal distributions within each contextual category c_i. We used a variation of Kullback–Leibler, which is called Jensen–Shannon divergence to measure the discrepancies between distributions. Jensen–Shannon divergence is a symmetrical distance measurement between two distributions, which is formulated as below:

$$D_{JS}(P||Q) = \frac{1}{2}D_{KS}(P||M) + \frac{1}{2}D_{KS}(Q||M) \tag{5.26}$$

where $M = \frac{1}{2}(P + Q)$ and D_{KS} is the Kullback–Leibler divergence, which is formulated as below.

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{Q(i)}{P(i)} \tag{5.27}$$

For temporal features, we compared the distributions of activity start time since the distribution of activity start time showed important characteristics of the population, such as morning and evening commute and mid-day travels. For spatial features, radius of gyration, and travel distance of daily activity trajectories were used. Travel distance captured the length of the trajectory. Radius of gyration is a measurement of distribution of mass in a



(a) Histogram of radius of gyration

(b) Histogram of travel distance

(c) Histogram of activity start time for "Low Radius" group

(d) Histogram of activity start time for "High Radius" group

Figure 5.2: Histogram of radius of gyration, travel distance, and activity start time for both "Low Radius" and "High Radius" groups.

physical object. Along human mobility trajectories, each stationary activity is regard as a point mass. Thus, radius of gyration measures the spatial distribution of the stationary activity along the trajectory. Radius of gyration is calculated using the formulation below.

$$r_g = \sqrt{\frac{1}{N}\sum_{i}^{N}(\vec{r_i} - \vec{r_c})^2} \tag{5.28}$$

where $\vec{r_c}$ represents the center of mass of the activity sequence with length N. $\vec{r_i}$ represents the location of stationary activity at timestamp i. Traveling distance is approximated using

sum of the Euclidean distance between two consecutive stationary points. It is formulated as

$$d = \sum_{i=1}^{N} |\vec{r_i} - \vec{r}_{i-1}| \tag{5.29}$$

where $\vec{r_o}$ is defined as location of the first activity along the sequence. For comparison, we show multiple Jensen–Shannon divergence of spatial-temporal distributions. We compare observed trajectories to generated trajectories. We also compare observed trajectories to observed trajectories with different level of Gaussian noise added on spatial and temporal dimensions.

Continuous Contextual Information

We experimented continuous contextual information. In the second experiment, we used radius of gyration, which is calculated using Eq.5.28, as the continuous contextual information to label each activity trajectory. To be specific, we selected daily activity trajectories from 100 individuals with radius of gyration no more than 25km. Duration sequence generation, we fed the model using radius of gyration with fine intervals from 1km to 25km. Radius of gyration and travel distance of generated trajectories were compared with the radius of gyration values that were input to the model. Radius of gyration and travel distance calculated from generated trajectories should be linearly dependent on the input radius of gyration.

In the third experiment, we used numerical income level as the continuous contextual information. The income levels were inferred by matching home locations detected from individual's location data to census tract data. The income levels in the census tract are reported by the city of Toronto [73]. We used 16 numerical income levels as input contextual variables. The mapping between numerical income level and the income amount is listed in Table 5.1. We selected a total of 112 individuals living in Toronto, Canada such that their income level was evenly distributed. The distribution of income level is shown in Fig.5.3 Duration sequence generation, we fed the model using numerical income level from 0 to 15 with fine intervals. We compared generated trajectories to observed trajectories in terms of distributions of radius of gyration and distribution of travel distance.

5.5 Results

Discrete Contextual Information

Here we present the results by comparing the distributions of the generated sequences with observed sequences. We generated 500 daily activity sequences within each group. In Fig 5.5, we compare the cumulative distribution of radius of gyration, travel distance, and activity start time in both groups. From observation, the cumulative distributions from generation and observation match closely. From the cumulative distributions, we are able to identify the peaks of density from the slope of the curves. The generated sequences show a density

Table 5.1: Annual income level label

Annual income level	**Annual income (CAD)**
0	<10k
1	10k - 15k
2	15k - 20k
3	20k - 25k
4	25k - 30k
5	30k - 35k
6	35k - 40k
7	40k - 45k
8	45k - 50k
9	50k - 60k
10	60k - 75k
11	75k - 100k
12	100k - 125k
13	125k - 150k
14	150k - 200k
15	>200k

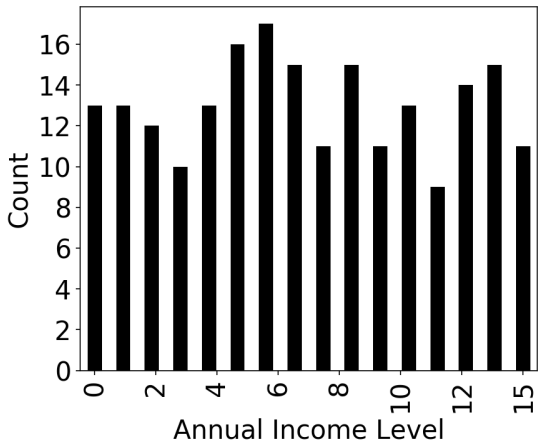


Figure 5.3: Number of individuals in each numerical annual income level.

peak at less than 3 km radius of gyration and a density peak at less than 10 km of travel distance for the "Low Radius" group. For the "High Radius" group, the generated sequences show a density peak between 20 to 30 km of radius of gyration and two peaks at 60 km and 120 km of traveling distance. For start time of activities, the generated sequences showed morning and evening density peaks for the "High Radius" group.

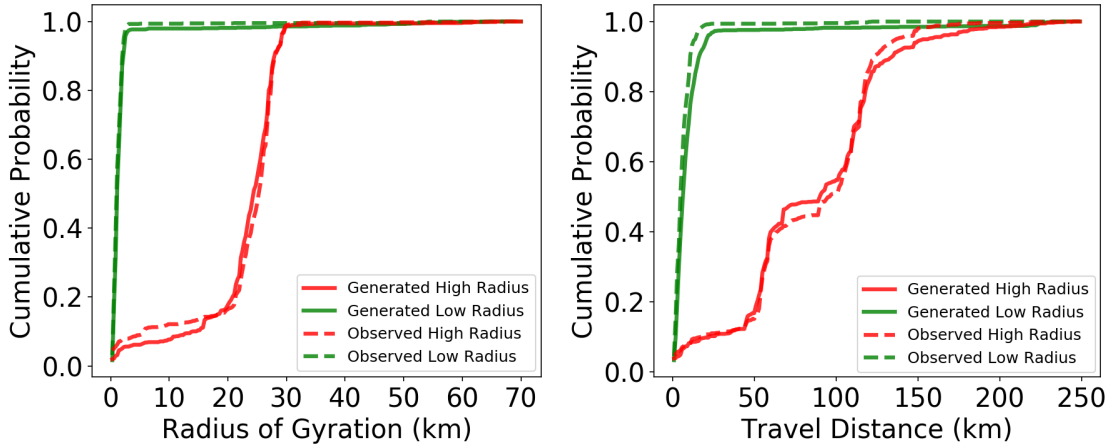


Figure 5.4: Comparing cumulative probability of radius of gyration (left) and traveling distance (right) of generated and observed daily traveling activities.

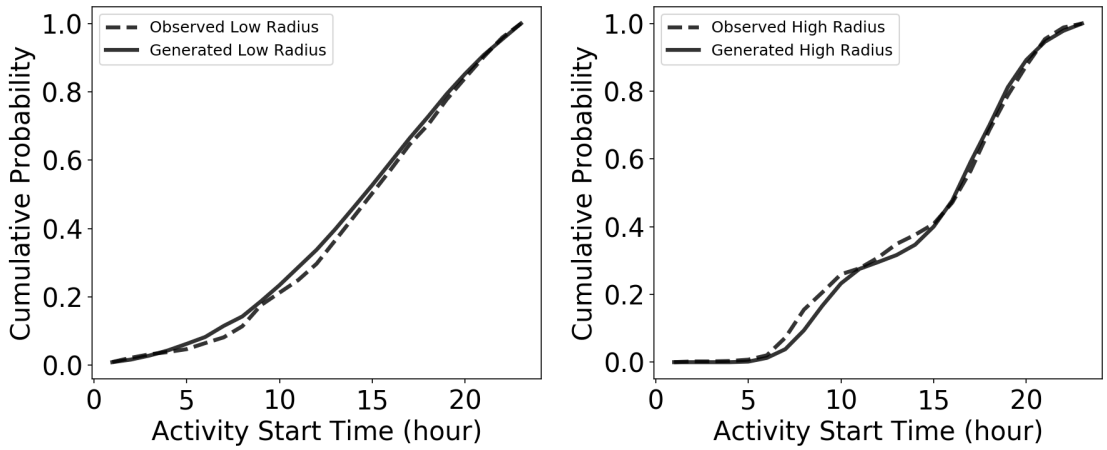


Figure 5.5: Comparing cumulative probability of activity start time for "Low Radius" group (left) and "High Radius" group (right).

We compute Jensen–Shannon divergence between the generated spatial and temporal distributions to the observed distribution. The discrete Jensen–Shannon divergence is computed using Equation 5.26 and 5.27. Discrete spatial and temporal densities are computed using spatial bin size of 1km and temporal bin size of 15 minutes. The results are reported in Table 5.2 to Table 5.4. For reference, we also computed the Jensen–Shannon divergence between observed trajectories and observed trajectories with added zero-mean Gaussian noise. Different noise levels are experimented. Results show that, by comparison, the spatial and temporal distributions generated are within reasonable range of error. Distributions of radius of gyration have better fit to the original distributions comparing to the ones with noise level of $\sigma = 1km$. For travel distance, the generated distributions have close divergence score

to the ones with noise level of $\sigma = 5km$. For activity start time, the generated distributions have better fit to the original distributions compared to the ones with noise level of $\sigma = 2hour$.

Table 5.2: Jensen–Shannon divergence for radius of gyration

	Observed vs. Generated	Observed vs. Added Noise ($\sigma = 1km$)	Observed vs. Added Noise ($\sigma = 2km$)
Radius of Gyration (Low Radius)	3.921×10^{-5}	2.125×10^{-4}	7.184×10^{-4}
Radius of Gyration (High Radius)	1.296×10^{-4}	3.235×10^{-4}	9.690×10^{-4}

Table 5.3: Jensen–Shannon divergence for travel distance.

	Observed vs. Generated	Observed vs. Added Noise ($\sigma = 5km$)	Observed vs. Added Noise ($\sigma = 10km$)
Travel Distance (Low Radius)	2.390×10^{-4}	2.367×10^{-4}	1.019×10^{-3}
Travel Distance (High Radius)	4.244×10^{-4}	5.891×10^{-4}	9.304×10^{-4}

Table 5.4: Jensen–Shannon divergence for activity start time.

	Observed vs. Generated	Observed vs. Added Noise ($\sigma = 2hour$)	Observed vs. Added Noise ($\sigma = 4hour$)
Activity Start Time (Low Radius)	2.096×10^{-4}	2.113×10^{-4}	1.112×10^{-3}
Activity End Time (High Radius)	1.444×10^{-3}	2.137×10^{-3}	6.632×10^{-3}

We further investigated the place visit frequency. Since locations of places are clustered location ID's, we were able to compute and compare place visit frequencies between generation and observation within each group. The frequencies were first normalized such that they sum to 1 within each group. We performed the comparison by plotting the generated frequencies and the observed frequencies as shown in Fig 5.6. Perfect generation should have all dots aligned on the 45 degree line, while our generation results the dots closely clustered around the 45 degree line. The "Low Radius" group show a R^2 value of 0.855 and the "High Radius" group show a R^2 value of 0.876.

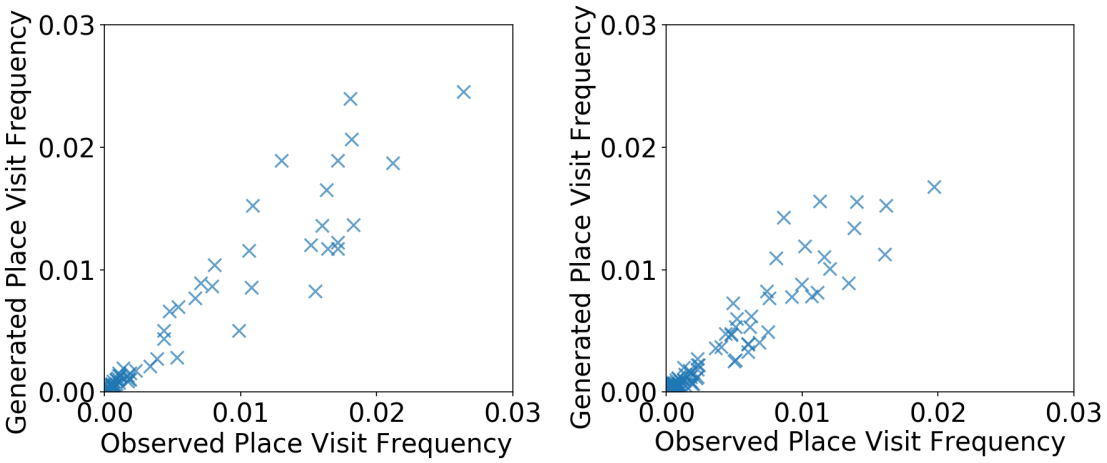


Figure 5.6: Comparing place visit frequency between generated activity sequences and observed activity sequences for "Low Radius" group (left) and "High Radius" group (right)

Continuous Contextual Information

Here we present the results from sequence generation using continuous contextual information. In the second experiment, we used radius of gyration as continuous contextual information. Duration sequence generation, the LSTM model took input radius of gyration from 1km to 25 km with interval of 0.1 km. 500 trajectories were generated at each input radius of gyration value to obtain the 25, 50 and 75 percentile radius of gyration and travel distances from generated trajectories. Result is presented in Fig.5.7. Results show the the radius of gyration and travel distances from generated trajectories have linear correlation to the input radius of gyration. The zigzag shape on the generated travel distance plot is due to the geography in the metropolitan region of Toronto. We also noticed that the variances of generated radius of gyration and travel distance are relatively higher at input radius of gyration of 1km due to the training sample density being low at low radius of gyration.

In the third experiment, we used numerical income level as continuous contextual information. Duration sequence generation, the LSTM model took input income level from 0 to 15 with interval of 0.1. 500 trajectories were generated at each input income value in order to obtain the 25, 50 and 75 percentile radius of gyration and travel distances from generated trajectories. The comparison between generated distributions and observed distribution was presented in Fig.5.8 and Fig.5.9. We observed high similarity between generated and observed distributions. We also noticed that the generated trajectories had higher radius of gyration and travel distance compared with observed trajectories at income level around 5 and 10. This experiment is more challenging since the numerical income level has non-linear relationship towards the radius of gyration and travel distance. Secondly, the training data contains only integer numerical income level, which is sparse in the income range of interest.

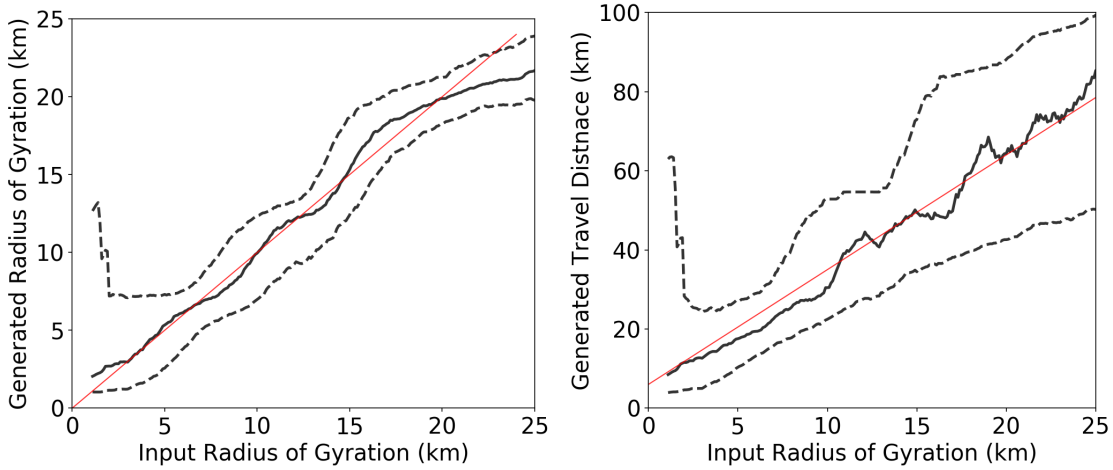


Figure 5.7: Radius of gyration (left) and travel distance (right) of generated trajectories vs. radius of gyration input to the LSTM model as contextual information. Solid black lines represent the median. Dashed black lines represent the 25 and 75 percentile. The red line on the left plot represents $y = x$. The red line on the right plot represents $y = 2.9x + 6$.

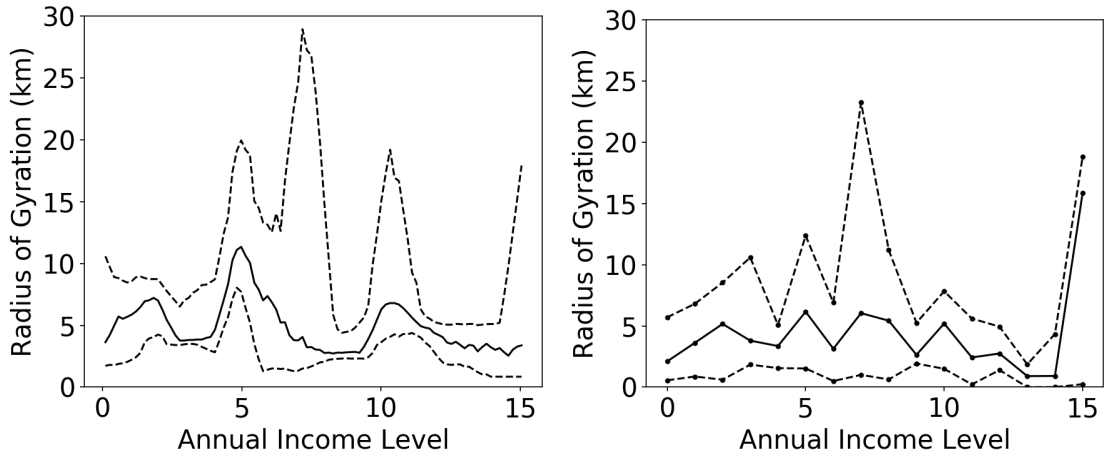


Figure 5.8: Radius of gyration of generated trajectories (left) and observed trajectories (right) vs. numerical income level.

5.6 Discussion

Generative models with contextual information have been developed for decades and were successfully applied to other fields, such as natural language. There is a lack of research on applying those model on transportation research. In this chapter, we introduce an extension of the LSTM mobility model by including characteristic information of individuals. We demonstrate the model structure, decomposition of input and output from the model.

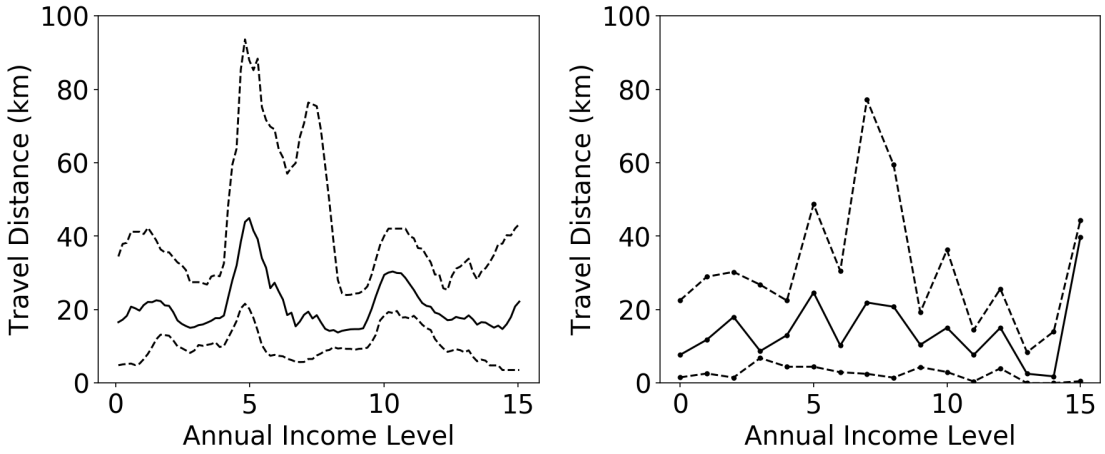


Figure 5.9: Travel distance of generated trajectories (left) and observed trajectories (right) vs. numerical income level.

We perform the model evaluation using 2 months of location data collected from individuals living in Toronto, Canada. In the first experiment, we experimented with discrete contextual information. Individuals were divided into two groups based on the radius of gyration of daily activity trajectories. The LSTM model captured the spatial and temporal patterns in activity sequences in both groups given contextual information. By comparing the generated and the observed spatial-temporal distributions Jensen–Shannon divergence scores, we found the generated activity trajectories are within reasonable range of error. The comparison of generated place visit frequencies and observed place visit frequencies showed reasonable similarity with R^2 score of 0.855 and 0.876 for both groups.

In the second and third experiment, we illustrated the results of continuous contextual information. In the second experiment, we used radius of gyration as continuous contextual information. LSTM models successfully learned the correlation between radius of gyration values and the corresponding trajectories. We observed linear relationship between the contextual input and the generated radius of gyration. The generated travel distances also showed such linear relationship to the contextual input. In the third experiment, the LSTM learned non-linear relationship between numerical income level and the activity trajectories. We used income level inferred from individuals' detected home locations and census data. The generated distributions showed high similarity to the observed distributions in terms of radius of gyration and travel distance despite some discrepancies that can be observed. Overall, the results show that the LSTM mobility model effectively learned the joint distribution of mobility trajectories and the corresponding contextual information even when the contextual information was sparsely distributed. The LSTM models are able to give reasonable prediction of trajectories with contextual information that never appeared in the training data set, which suggested the models can be generalized for transportation demand forecasting problem with locational data and individual characteristics.

Chapter 6

Conclusion

6.1 Summary

In this dissertation, we presented the LSTM mobility model structures for modeling urban mobility sequences. Limitations in the existing mobility models prevented them from being directly applied to transportation practices, which motivated this dissertation. The major contributions of this dissertation include the following: First, we developed the LSTM mobility models that are capable of learning and predicting the entire mobility sequences within a time windows of interest; Second, we developed the LSTM mobility models that were able to predict activity sequences with activity type choices and explicit spatial-temporal choices; Third, the LSTM mobility models are able to capture long-term activity dependencies; and lastly, the LSTM mobility models can be easily extended, which are capable of learning and generating sequences with contextual information.

We presented the one-layer LSTM model structure in Chapter 3, and the two other structures presented in Chapter 4 and Chapter 5 are extensions to the one-layer model. The LSTM models presented are capable of learning activity type and spatial-temporal features simultaneously using mixture density output from the LSTM networks since the flexibility and learning power of LSTM neural network allows its output to formulate arbitrary parametric distributions. In Chapter 5, the model structure was extended for learning activity sequences with contextual information.

We first evaluated the LSTM model using agent-based traffic simulation. We present the framework as a state of art end-to-end pipeline from processing locational data to traffic simulation. The multimodal simulation results showed high similarity to the observed freeway traffic volume and transit boarding counts. In Chapter 4, we applied the LSTM models on the medium term partial sequence prediction. The LSTM models show better performance in aggregated level against baseline models due to the capability in learning long-term dependencies. Further experiments on an individual level showed comparable results between LSTM models and NN models. LSTM models show better performance in partial sequence prediction when a greater part of the sequences are observed. Finally, in Chapter 5, we

evaluate the LSTM models using sequences with contextual information and by comparing the generated and observed distributions of spatial-temporal features.

6.2 Future Research Direction

The LSTM models or neural network models in general are supervised models as they rely on labeled data. Despite the spatial and temporal richness of the locational data streaming from smartphones, there are few sources of locational data available for research purposes that include labeled activity types and contextual information about the device user due to privacy concerns. In our experiments, we combined all secondary activity types into "others" due to lack of labels. However, the richness of urban activities greatly exits in the secondary activities. Collecting labeled data would require manual input and privacy consent from the device holder; however, it is beneficial to keep the data collection method non-invasive. Thus, one future research direction is to discover data sources with labeled activity types.

Additionally, another future research direction is to understand the hidden states and gate activation in LSTM models. Hidden Markov Models use discrete hidden states so that HMM's can be used as unsupervised models for sequences since the hidden states can often be interpreted from the emission probabilities. LSTM models, on the other hand, use continuous hidden states to control the outputs: gates and updates of the hidden states. It is usually difficult to interpret the hidden states of LSTM models because of the dimensionality and sensitivity of the states that easily affect the behavior of the network. By developing a method to categorize hidden states and gate activation, LSTM models could potentially be used as an unsupervised model for sequences.

Traveling behavior is driven by complex decisions of individuals, which include attributes such as interpersonal interactions, in-home activities, and various other constraints. In this dissertation, the LSTM models we introduced focus on the modeling habits and patterns in stationary activities and trajectories of individuals. Part of the reason for our focus is the data sources. Locational data being the common data generated from mobile devices generally lacks information about the device owner and the decisions and attributes mentioned above. When further information becomes available, such as calendar information and social network connections, better modeling frameworks for more coherent traveling decisions can be developed with this additional information.

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] Alexandre Alahi et al. "Social lstm: Human trajectory prediction in crowded spaces". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 961–971.

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

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

[5] M Balmer et al. *Agent-based simulation of travel demand: Structure and computational performance of MATSim-T*. ETH, Eidgenössische Technische Hochschule Zürich, IVT Institut für Verkehrsplanung und Transportsysteme, 2008.

[6] Armando Bazzani et al. "Statistical laws in urban mobility from microscopic GPS data in the area of Florence". In: *Journal of Statistical Mechanics: Theory and Experiment* 2010.05 (2010), P05001.

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

[8] 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.

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

[10] Chandra R Bhat and Frank S Koppelman. "Activity-based modeling of travel demand". In: *Handbook of transportation Science*. Springer, 1999, pp. 35–61.

[11] 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.

[12] Vincent Bindschaedler and Reza Shokri. "Synthesizing plausible privacy-preserving location traces". In: *Security and Privacy, 2016 IEEE*. IEEE. 2016, pp. 546–563.

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

[14] Richard Black et al. "Climate change: Migration as adaptation". In: *Nature* 478.7370 (2011), p. 447.

[15] 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.

[16] Chantal C Cantarelli et al. "Cost overruns in large-scale transportation infrastructure projects: explanations and their theoretical embeddedness". In: (2010).

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

[18] Cynthia Chen et al. "The promises of big data and small data for travel behavior (aka human mobility) analysis". In: *Transportation Research Part C: Emerging Technologies* 68 (2016), pp. 285–299. ISSN: 0968090X. DOI: 10.1016/j.trc.2016.04.005.

[19] 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.

[20] Junyoung Chung et al. "A recurrent latent variable model for sequential data". In: *Advances in neural information processing systems*. 2015, pp. 2980–2988.

[21] Caitlin Cottrill et al. "Future mobility survey: Experience in developing a smartphone-based travel survey in Singapore". In: *Transportation Research Record: Journal of the Transportation Research Board* 2354 (2013), pp. 59–67.

[22] Steven S Coughlin. "Recall bias in epidemiologic studies". In: *Journal of clinical epidemiology* 43.1 (1990), pp. 87–91.

[23] Hanjun Dai et al. "Recurrent Hidden Semi-Markov Model". In: (2016).

[24] Stanislas Dehaene, Jean-Pierre Changeux, and Jean-Pierre Nadal. "Neural networks that learn temporal sequences by selection". In: *Proceedings of the National Academy of Sciences* 84.9 (1987), pp. 2727–2731.

[25] 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.

[26] 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.

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

[28] 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.

[29] Bent Flyvbjerg, Nils Bruzelius, and Werner Rothengatter. *Megaprojects and risk: An anatomy of ambition*. Cambridge University Press, 2003.

[30] 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.

[31] Wei Geng and Guang Yang. "Partial Correlation between Spatial and Temporal Regularities of Human Mobility". In: *Scientific Reports* 7 (2017).

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

[33] Ursula Grandcolas, Ruth Rettie, and Kira Marusenko. "Web survey bias: sample or mode effect?" In: *Journal of marketing management* 19.5-6 (2003), pp. 541–561.

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

[35] Karol Gregor et al. "DRAW: A recurrent neural network for image generation". In: *arXiv:1502.04623* (2015).

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

[37] Samiul Hasan, Xianyuan Zhan, and Satish V Ukkusuri. "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media". In: *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*. ACM. 2013, p. 6.

[38] Samiul Hasan et al. "Spatiotemporal patterns of urban human mobility". In: *Journal of Statistical Physics* 151.1-2 (2013), pp. 304–318.

[39] Xi He et al. "Dpt: Differentially private trajectory synthesis using hierarchical reference systems". In: *Proceedings of the VLDB Endowment* 8.11 (2015), pp. 1154–1165.

[40] Eelco Herder and Patrick Siehndel. "Daily and weekly patterns in human mobility." In: *UMAP Workshops*. 2012.

[41] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[42] 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.

[43] Ken-ichi Kamijo and Tetsuji Tanigawa. "Stock price pattern recognition-a recurrent neural network approach". In: *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*. IEEE. 1990, pp. 215–221.

[44] Diederik Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[45] Felix Kling and Alexei Pozdnoukhov. "When a city tells a story: urban topic analysis". In: *Proceedings of the 20th international conference on advances in geographic information systems*. ACM. 2012, pp. 482–485.

[46] Yann LeCun, Yoshua Bengio, et al. "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.

[47] Maxime Lenormand et al. "Influence of sociodemographic characteristics on human mobility". In: *Scientific reports* 5 (2015), p. 10075.

[48] Ian XY Leung et al. "Intra-city urban network and traffic flow analysis from GPS mobility trace". In: *arXiv preprint arXiv:1105.5839* (2011).

[49] 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.

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

[51] 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.

[52] Xiaolei Ma et al. "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data". In: *Transportation Research Part C: Emerging Technologies* 54 (2015), pp. 187–197.

[53] 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.

[54] Gordon McBean and Idowu Ajibade. "Climate change, related hazards and human settlements". In: *Current Opinion in Environmental Sustainability* 1.2 (2009), pp. 179–186.

[55] Michael G McNally. "The four-step model". In: *Handbook of Transport Modelling: 2nd Edition*. Emerald Group Publishing Limited, 2007, pp. 35–53.

[56] Darakhshan J Mir et al. "Dp-where: Differentially private modeling of human mobility". In: *Big Data, 2013 IEEE International Conference on*. IEEE. 2013, pp. 580–588.

[57] Jun Pang, Polina Zablotskaia, and Yang Zhang. "On Impact of Weather on Human Mobility in Cities". In: *International Conference on Web Information Systems Engineering*. Springer. 2016, pp. 247–256.

[58] Barak A Pearlmutter. "Learning state space trajectories in recurrent neural networks". In: *Neural Computation* 1.2 (1989), pp. 263–269.

[59] Mathieu Radenen and Thierry Artières. "Contextual hidden markov models". In: *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE. 2012, pp. 2113–2116.

[60] Injong Rhee et al. "On the levy-walk nature of human mobility". In: *IEEE/ACM transactions on networking (TON)* 19.3 (2011), pp. 630–643.

[61] Luca Rossi, James Walker, and Mirco Musolesi. "Spatio-temporal techniques for user identification by means of GPS mobility data". In: *EPJ Data Science* 4.1 (2015), p. 11.

[62] Burkhard Rost and Chris Sander. "Improved prediction of protein secondary structure by use of sequence profiles and neural networks". In: *Proceedings of the National Academy of Sciences* 90.16 (1993), pp. 7558–7562.

[63] Adella Santos et al. *Summary of travel trends: 2009 national household travel survey*. Tech. rep. 2011.

[64] Stefan Schmöller and Klaus Bogenberger. "Analyzing external factors on the spatial and temporal demand of car sharing systems". In: *Procedia-Social and Behavioral Sciences* 111 (2014), pp. 8–17.

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

[66] 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.

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

[68] Ilya Sutskever, James Martens, and Geoffrey E Hinton. "Generating text with recurrent neural networks". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, pp. 1017–1024.

[69] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *NIPS*. 2014, pp. 3104–3112.

[70] Latanya Sweeney. "k-anonymity: A model for protecting privacy". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 557–570.

[71] Etienne Thuillier et al. "Clustering weekly patterns of human mobility through mobile phone data". In: *IEEE Transactions on Mobile Computing* (2017).

[72] Yongxue Tian and Li Pan. "Predicting Short-Term Traffic Flow by Long Short-Term Memory Recurrent Neural Network". In: *Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on*. IEEE. 2015, pp. 153–158.

[73] City of Toronto. "2016 Census: Income". In: (2016).

[74] US Bureau of Labor Statistics. *American Time Use Survey: 2015 Results*. Tech. rep. June 2016.

[75] US Department of Transportation: Federal Transit. *The Predicted and Actual Impacts of New Starts Projects - 2007: Capital Cost and Ridership*. Tech. rep. 2008, p. 9.

[76] Oriol Vinyals et al. "Show and tell: A neural image caption generator". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3156–3164.

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

[78] Patrick Vogel, Torsten Greiser, and Dirk Christian Mattfeld. "Understanding bike-sharing systems using data mining: Exploring activity patterns". In: *Procedia-Social and Behavioral Sciences* 20 (2011), pp. 514–523.

[79] Pu Wang et al. "Understanding road usage patterns in urban areas". In: *Scientific reports* 2 (2012).

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

[81] Kelvin Xu et al. "Show, attend and tell: Neural image caption generation with visual attention". In: *International Conference on Machine Learning*. 2015, pp. 2048–2057.

[82] 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.

[83] 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.

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

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

[86] Quanzeng You et al. "Image captioning with semantic attention". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4651–4659.

[87] Chao Zhang et al. "GMove: Group-Level Mobility Modeling Using Geo-Tagged Social Media". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 1305–1314.

[88] 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.

[89] 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.

[90] 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.

[91] Yu Zheng et al. "Urban computing: concepts, methodologies, and applications". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 5.3 (2014), p. 38.

[92] Wen-Yuan Zhu et al. "Modeling user mobility for location promotion in location-based social networks". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 1573–1582.