Deep Generative Models of Urban Mobility

Ziheng Lin∗
Civil Systems, CEE, UC Berkeley
zihenglin@berkeley.edu

Mogeng Yin†
Transportation, CEE, UC Berkeley
mogengyin@berkeley.edu

Sidney Feygin‡
Civil Systems, CEE, UC Berkeley
sid.feygin@berkeley.edu

Madeleine Sheehan
Transportation, CEE, UC Berkeley

Jean-Francois Paiement
AT&T Research

Alexei Pozdnoukhov
CEE, UC Berkeley

ABSTRACT

Locational data generated by mobile devices present an opportunity to substantially simplify methodologies and reduce analysis latencies in transportation planning applications. In this paper, we describe a modeling framework that supports most common transportation planning tasks, delivering actionable solutions at a fraction of time and cost as compared to the state of practice. The framework builds up on cell phone data processing and activity-based inferences of travel purposes with an Input-Output Hidden Markov Model (IO-HMM), followed by a Long Short Term Memory (LSTM) network that learns travelers’ mobility sequences. It combines the desired interpretability due to the parametric specification of an IO-HMM with flexibility and predictive power of deep neural models. We describe our target use case for the synthesized activity chains: delivering decision support and transportation scenario evaluation to practitioners. We outline domain-driven operational objectives and verify that our framework meets these criteria by illustrating its usability in typical transportation demand planning applications. It is currently being deployed for testing by a major network carrier serving millions of users in the San Francisco Bay Area.

CCS CONCEPTS

• Computing methodologies → Modeling methodologies;

KEYWORDS

Cellular data, generative models, recurrent neural network, demand forecasting, activity-based models

ACM Reference format:
DOI: 10.475/123_4

∗equally contributing author
†equally contributing author
‡equally contributing author

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

KDD’17, Halifax, Nova Scotia, Canada
© 2017 Copyright held by the owner/author(s). 123-4567-24-567/08/06…$15.00
DOI 10.475/123_4

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 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 [15]. As a consequence, project overruns due to inaccurate economic value appraisals continue to cost taxpayers billions of dollars [9, 34]. Non-invasive, automated, continuous data collection mechanisms are increasingly being used to complement manual survey techniques by improving their statistical representativeness [12]. 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 [36]. While applications of mining mobility trajectories from crowd-sourced locational data are common (these are thoroughly reviewed in Section 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 [10].

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 [4]. 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.

1.1 Contribution of this work

In this paper, we describe the developed two-step generative modeling framework that is capable of learning activity sequences from large volumes of call data while incorporating behavioral parameters that are sensitive to policy levers. 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 authors' 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 discourse 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 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 ground truth of observed travel, which took advantage of this rich activity-based inference of observed travel, thus expanding the scope of state-of-the-art human mobility modeling techniques to the realm of decision support for transportation policy analysis.

2 RELATED 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 [40]. Human activity recognition from spatiotemporal microdata streams, an application of urban computing research, has been explored extensively by practitioners across disciplines [10, 16]. 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.

2.1 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 [24, 41]. 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, researchers have developed methods for classifying activities into distinct categories [23, 37], for separating social trips from commute trips [11], promoting products and services [42] and inferring mobility of grouped individuals [39].

Figure 1: Modeling framework diagram.

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 [13], aggregated origin-destination (OD) estimation [35], stylized mobility laws [16, 29], and disaster response [25]. Until recently, there was limited work using cellular data for urban activity recognition [36], especially in the context of high-fidelity travel modelling [38].

2.2 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 [16]. Longitudinal analysis also suggests that human mobility has high temporal and spatial regularities [29]. 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 [31]. Long short term memory (LSTM) [21] is one of the most popular variations of RNNs with proven ability to generate sequences in various domains, such as text [30], images [18] and handwriting [17]. A recent work showed the capability of LSTM to predict pedestrians’ motion from video data [2]. 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 activity patterns, but also to generate new sequences [19]. Our 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 [38].

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
In the first step, raw CDR data containing a timestamped record for each communication of anonymous user’s device served by the cellular network go through a k-anonymity check [32] and optional differential privacy (DP) filters [20, 27] 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 [38].

(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 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 [7, 20].

The full framework (shown in Figure 1) has some additional important components. Current practice of understanding traveler’s mode choice preference is based on travel surveys that are usually outdated to reflect the rapid changes in people’s behavior. A discrete choice modeler (DCM) allows us to obtain up-to-date travelers’ preference using CDR data. Its importance is two-fold: (1) The calibrated DCM is one of the direct outputs to transportation decision makers, which can be used for scenario evaluation and urban planning. (2) It is a way to calibrate the utility functions of agents in a microscopic travel simulation. For enhanced realism in extrapolating scenarios towards unseen situations, we use a simulation where agents are independent decision makers with logic of travel choices (such as travel mode choice given the trip purpose) being governed by a parametric utility function. To this extent we also use social-demographic information from census data (such as the median income in the district a traveler lives) as contextual variables to make the discrete choice model more accurate and practical in extrapolating scenarios [4], as detailed in Section 5.3.

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 [7, 20]. 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,
- aggregated travel demand volumes and performance metrics accessible through basic interactive elements implemented within the travel demand explorer (Figure 9) to evaluate the proposed policy and infrastructure changes in the transportation system.
4 LEARNING ACTIVITY SEQUENCES WITH LONG SHORT-TERM MEMORY RNNs

In this section, we describe the LSTM model architecture that generates activity sequences with explicit location choice and activity type labeled by IO-HMM. We introduce the design of input and output variables, the sequence generation process, and the loss function. For a detailed discussion of the implementation of LSTM cells in the context of recurrent neural networks, we direct reader to [17].

We employ mixture density networks (MDNs), first described in [8] and used with success in a handwriting synthesis task incorporating LSTM RNNs [17]. In Fig. 3, we illustrate our model structure. We use $x_t$, $c_t$ and $y_t$ to denote input variables, context variables, and output variables, respectively, at time $t$. The hidden state of the LSTM layer(s) at time $t$ is $h_t$.

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

1. At every time step, the LSTM layer receives a set of inputs from both input variables $x_t$ and the context variables $c_t$.
2. The LSTM layer(s) then produce a set of outputs $y_t$, which is used to parametrize distribution $p(x_{t+1}|y_t, c_{t+1})$ for sampling a new set of input variables $x_{t+1}$.

4.1 Input and Context Variables

We design the input such that it contains the encoding of previous activity, $x_t$, conditional on additional context information, $c_t$, which helps the model to choose the next activity. We encode an activity using its starting time and day of week, duration, location (latitude and longitude), activity types and a variable indicating the end of the day. Latitude, longitude, duration of an activity are represented as continuous real-valued numbers. Activity types consist of the labels of each activity including "home", "work" and "others", which are encoded as a one-hot vector. The binary-valued end-of-day probability is used as an indicator function to trim the generated daily activity sequences. The context variables, $c_t$, include attributes that are not sampled from $p(x_{t+1}|y_t, c_{t+1})$ (further detailed in Section 4.2), but that significantly affect the next activity choice. These include activity staring time, day of week, home and work locations. We include an individual’s home and work locations since secondary activity locations are often affected by those. Thus, $x_t$ and $c_t$ can be expressed as the following.

$$x_t = \{\text{lat}_t, \text{lon}_t, \text{dur}_t, \text{type}_t, \text{end}_t\}$$

$$c_t = \{\text{st}_t, \text{dayOfWeek}_t, \text{lat}_{\text{home}}, \text{lon}_{\text{home}}, \text{lat}_{\text{work}}, \text{lon}_{\text{work}}\}$$

where the abbreviations "lat", "lon", "st", "dur" refer to latitude, longitude, starting time, and duration of an activity, respectively. The variables "type", "end" and "dayOfWeek" represent the activity types, end of day indicator, and day of week values, respectively.

4.2 Output Variables and Mixture Distributions

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 $x_t$. In particular, it is a joint distribution of activity starting time, duration, latitude, longitude, probabilities of activity types and end-of-day probability; and each component is weighted by $\pi_i$. The joint distribution of starting time, duration, latitude, longitude is a 4-dimensional Gaussian distribution with correlation only between activity starting time and duration. $p_{\text{home}}, p_{\text{work}}$ and $p_{\text{other}}$ describe a distribution for activity types with 3 categories: home-related, work-related, and other activities. The end-of-day probability is shown as a single component, $p_{\text{end}}$. 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.

$$y_t = \{\hat{\pi}, \hat{\mu}_{\text{lat}}, \hat{\mu}_{\text{lon}}, \hat{\mu}_{\text{st}}, \hat{\mu}_{\text{dur}}, \hat{\sigma}_{\text{lat}}, \hat{\sigma}_{\text{lon}}, \hat{\sigma}_{\text{st}}, \hat{\sigma}_{\text{dur}}, \hat{\rho}_{\text{lat}, \text{dur}}, p_{\text{home}}, p_{\text{work}}, p_{\text{other}}, p_{\text{end}}\}$$

Those raw outputs are properly scaled before serving as mixture distribution parameters. The component weights $\hat{\pi}$ and probability of activity type within each component are normalized using softmax function. Standard deviations $\hat{\sigma} \cdot$ are constrained to be non-negative using an exponential function with correlation coefficients, $\hat{\rho} \cdot$, scaled between $-1$ and $1$ using tanh activation functions. The end-of-day probabilities $p_{\text{end}}$ are scaled to lie between 0 and 1 using a logistic sigmoid function. The following equations summarize our model:

$$\pi^i = \frac{\exp(\hat{\pi}^i - b)}{\sum^N_j \exp(\hat{\pi}^j - b)}; i \in \{1...N\}$$

$$\mu^i = \hat{\mu}^i \cdot$$

$$\sigma^i = \exp(\hat{\sigma}^i \cdot)$$

$$\rho^i = \tanh(\hat{\rho}^i \cdot)$$

$$p_{\text{type}}^i = \frac{\exp(\hat{\pi}^i_{\text{type}})}{\sum_{\text{type}} \exp(\hat{\pi}^i_{\text{type}})}; i, j \in \{\text{home, work, other}\}$$

$$p_{\text{end}} = \frac{1}{1 + \exp(\hat{\pi}_{\text{end}})}.$$
We define the following decomposition of the joint distribution so that the loss (negative log-likelihood) of the model can be calculated conveniently.

\[
p^i(x_{t-1}, y_{t-1}, c_t) \propto \sum_{i} \pi^i \cdot \mathcal{N}(x_t; \mu^i, \sigma^i, \text{lat}_{t}, \text{lon}_{t})
\]

(10)

We define the following decomposition of the joint distribution so that the loss (negative log-likelihood) of the model can be calculated conveniently.

\[
p^i(st_{t}, dur_{t}) \cdot p^i(type_{t}) \cdot p^i(end_{t})
\]

(11)

\[
p^i(lat_{t}, lon_{t}) = \mathcal{N}(\text{lat}_{t}, \text{lon}_{t}; \mu^i_{lat}, \mu^i_{lon}, \Sigma_{lat,lon})
\]

(12)

\[
p^i(st_{t}, dur_{t}) = \mathcal{N}(st_{t}, dur_{t}; \mu^i_{st}, \mu^i_{dur}, \Sigma_{st,dur})
\]

(13)

\[
p^i(type_{t}) = \sum_{j} p_{j} \cdot (1 - p_{j});
\]

(14)

\[
\pi_i = \text{end}_{t} \cdot p_{end} + (1 - \text{end}_{t}) \cdot (1 - p_{end})
\]

(15)

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

4.3 Sequence generation

Sequence generation is initialized by feeding the LSTM model with \(x_0\) and \(c_0\). We specify the initial input vector \(x_0\) as constant since we do not give the model any information about the first activity and the model should generate the first activity from its learnt weights. The initialization of \(c_0\) is described in Section 5.2. Starting from \(t = 1\), we sample location, duration, activity type and end-of-day probability from \(p(x_{t}, y_{t-1}, c_t)\), which is a mixture distribution conditioned on the observed starting time of the activity. The mixture weights are calculated as the following.

\[
w^i(st_{t}) = \frac{\pi^i \cdot N(st_{t}; \mu^i_{st}, \sigma^i_{st})^i}{\sum_{i} \pi^i \cdot N(st_{t}; \mu^i_{st}, \sigma^i_{st})^i}
\]

(16)

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

\[
\mu_{dur|st_{t}} = \mu_{dur} + \frac{\sigma_{dur}}{\sigma_{st}} \cdot \rho_{st,dur} \cdot (st_{t} - \mu_{st})
\]

(17)

\[
\sigma_{dur|st_{t}} = \sqrt{(1 - \rho_{st,dur}^2)} \cdot \sigma_{dur}
\]

(18)

Now we can sample a new activity \(x_{t}\) from the mixture distribution \(p(x_{t}, y_{t-1}, c_t)\) following Eq. 19 through Eq. 23. First, a component is sampled from the multinomial distribution of mixture weights (Eq. 16). 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 in \(c_t\) is incremented by \(dur_{t}\).

4.4 Loss Function and Model Estimation

We use negative log-likelihood as the loss of the model. The loss function can be decomposed using Eq. 10 through Eq. 15.

\[
\text{loss} = \sum_{t=1}^{T} - \log \sum_{i} \pi^i \cdot p^i(x_{t}, y_{t-1}, c_t)
\]

(24)

Given an activity chain, the loss is calculated as sum of negative log-likelihood of each observed activity. Thus, it is cumulative over the entire sequence. The final gradients are backpropagated through the model weights at each time step. We use the Adam optimizer [22] for the stochastic gradient descent.

Individuals may have different numbers of daily activities such that the LSTM model has to handle sequences with different length. Most implementations of LSTM networks do not directly handle sequences with different length because the number of unrolling steps is fixed when the model is compiled. We use zero padding to pad activity sequences to the same length; and the loss due to the padding is masked when calculating the sequence loss.

5 EXPERIMENTAL RESULTS

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 place.

5.1 Data Pre-processing

We pre-process the data following the steps in [38]. The home and work locations are identified during the pre-processing step. We take cell phone users that showed up for more than 21 days a month at their identified “home” place; showed up for more than 14 days a month at their identified “work” place; have home and work not at the same location. These criteria identify regular working commuters with a day structure containing both distinct Home and Work.
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% [33], 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 [6] and these numbers are 2.9%, 15.2% and 43.7% in our labeled data.

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

As travel patterns vary greatly over the region, we trained 34 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). Using standard procedures to fit the population marginals with the census data [14], we sample residents home and work locations to create synthetic commuters with a predetermined home TAZ and work TAZ. The precise home and work locations (lat/lon coordinates) are sampled uniformly within the home and work TAZs.

We use single layer of LSTM with 128 units. 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. The context variables \( \phi \) are initialized with starting time of first activity drawn from the observations within the super-district with added Gaussian noise and home and work locations are determined as mentioned above. The sampling bias \( b \) is tuned as 2.0 in order to reduce number of outliers in the generated sequences. We also apply filters on the activity sequences generated to delete unrealistic ones: we filtered the sequences that don’t end with home activities and those containing multiple consecutive home or work activities.

We present the temporal characteristics of the generated sequences in Fig. 4 and Fig. 5. In Fig. 4, we observe a decreasing number of work and home activities while there is a slight increase in secondary (other) activities. Fig. 5 shows the joint distributions of starting time and duration of each activity type. The home activities started around noon are relatively short. Night-time home activities have strong correlations between staring time and duration. Work activities show typical working pattern of commuters that some last the entire day staring in the morning while 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. Finally, the last activities of the day determined by the LSTM model is consistent with the pattern of night-time home activities. We refer a reader to [38] for a similar but much more detailed analysis of inferred activities.

5.3 Behavioral parameters and DCM
While providing accurate spatio-temporal representation of travel, the IO-HMM/LSTM component of our framework lacks an important property: it does not relate the observed choices to socio-demographic characteristics of travelers, nor does it allow implicit forecasting of travel choices beyond previously observed conditions (changing tolls, transit fares, or increasing travel delays due to growing congestion). To provide this capability, we developed an additional modeling component based on discrete choice models (DCMs).

Popularized by a Nobel prize winning economic school of thought [26], DCMs are a de facto standard in practical evaluation of the travelers’ population response to system parameters and policy interventions. In practice, planners operate with parametric DCMs to learn traveler’s choice preferences and determine how travelers trade off various attributes of a given set of travel choice alternatives [5]. DCMs are typically fit globally to an entire metropolitan region, but, if used in combination with the IO-HMM/LSTM component, DCMs allow for rich, activity and context-dependent travel choice model.

A common assumption in behavioral modeling is that there exists a utility \( U \), that person \( n \) obtains from choosing a travel alternative \( i \), is a linear form over attributes of each travel alternative in the choice set: \( U_{ni} = \beta z_{ni} + \epsilon_{ni} \), where \( z_{ni} \) represents a vector of observed variables of trip \( i \) and traveler \( n \). \( \beta \) is a vector of the corresponding coefficients that are interacted with each of the observable variable, and \( \epsilon_{ni} \) represents the utility of unobservable factors that contribute to travel choices.
The probability of traveler \( n \) selecting travel alternative \( i \) is the probability that the utility of alternative \( i \) is greater than the utility of all other alternatives. With an assumption that \( e^\beta z_{ni} \) follows a generalized extreme value Type I distribution and a traveler is considering \( J \) total alternatives with known attributes, the probability of choosing mode \( i \) is given by:

\[
P_{ni} = \frac{e^{\beta z_{ni}}}{\sum_{j=1}^{J} e^{\beta z_{nj}}}
\]

With a powerful set of techniques to infer travel choices within a context of daily tours described above, the presented framework enables estimation of DCMs.

Here we illustrate it on a simple case of a travel mode choice model, noting that more sophisticated model specifications can be considered if required. We selected a subset of several hundred trans-bay trips with an origin or destination of downtown San Francisco, where both driving and public transit alternatives are available. For this set of observed trips, we used a discriminative model to detect trip mode [28], and queried an in-house routing service that provides travel times and costs for a set of possible travel alternatives between the observed origin and destination zones. Eq. (25) shows a DCM specification that accounts for the time and cost of travel, a traveler’s anticipated income (the median income of the traveler’s home census tract is used as a proxy for the traveler’s income), and unobserved systematic mode preferences captured in an offset term \( \beta_{drive} \) (the so-called alternative-specific constant).

\[
V_{drive} = \beta_{drive} + \beta_{income} \cdot Income \\
+ \beta_{TT} \cdot TravelTime_{drive} \\
+ \beta_{TC} \cdot TravelCost_{drive}
\]

\[
V_{public\_transit} = \beta_{TT} \cdot TravelTime_{public\_transit} \\
+ \beta_{TC} \cdot TravelCost_{public\_transit}
\]

The maximum likelihood estimation of this model resulted in the parameters listed in Table 1. The estimated parameters indicate that the utility of a travel option decreases as travel time and travel cost increase; that, all else equal, for trips to and from downtown San Francisco, travelers have an affinity for public transit over driving (likely because driving downtown can be a hassle and parking can be expensive); that higher income travelers are more likely to drive; and finally, comparing the travel time and cost coefficients, travelers value their time at a rate of about $30/hour. These result allow evaluating the response of the city commuters to interventions such as changing transit fares, improving the level of service (travel time by transit), and evaluating the price at which travelers will choose Uber over existing public transit options. The inferred values can also inform the agents’ logic in the micro-simulation of travel.

5.4 Scenario micro-simulation

Traffic micro-simulation is a conventional approach in studying performance and evaluating transportation planning and development scenarios, including the ones where the travel system conditions are changed and the response is extrapolated with behavioral models.

Table 1: Discrete choice model parameters for travel mode

| Variable     | Coeff | Std. error | Z   | P > |Z| |
|--------------|-------|------------|-----|-----|----|
| \( \beta_{drive} \) | -1.048 | 0.525 | -1.966 | 0.046 |
| \( \beta_{income} \) | 0.0156 | 0.006 | 2.557 | 0.011 |
| \( \beta_{TT} \) | -1.9495 | 0.413 | -4.72 | 0.000 |
| \( \beta_{TC} \) | -0.0653 | 0.061 | -1.071 | 0.284 |

Micro-simulation of a typical weekday traffic is performed using the MATSim\(^2\) platform [3]. MATSim is a state-of-the-art agent based multi-modal mobility micro-simulation tool that performs mode choice and traffic assignment for the set of agents with predefined activity plans. It varies departure times and routing of each agent depending on the congestion generated on the network, in order to maximize agent’s daily utility, 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, primary and secondary roads (network fragment is visualized in Figure 6).

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, 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. 6 demonstrates examples of the three characteristic hourly volume profiles comparing the modeled and observed counts on freeways. Figure 7 shows examples of transit passengers counts entering and exiting 2 major rapid transit stations. Validation results for the full set of sensors are presented in Fig. 8. Fig. 8a shows a comparison of the volumes for three distinct time periods. Fig. 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 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. 8a) as compared to the ground truth data.

6 CONCLUSIONS AND FUTURE WORK

In this paper, 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, developed in our previous work, is used for labeling activity types of the preprocessed and anonymized cellular data in San Francisco Bay Area.

---

\(^1\)This example results in parameter values within a range of common practice, but it serves as an illustration only. Policy analysis conducted by planning authorities involves a more thorough and detailed investigation.

\(^2\)MATSim code available at https://github.com/matsim-org
16th St. and Mission Station  MacArthur Station

(a) Modeled vs observed volumes at 8am (black), 1pm (red) and 6pm (blue).
(b) Mean relative error (%) of modeled vs observed traffic volumes during the day.

Figure 6: 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 7.

Figure 7: Actual and simulated boarding and alighting counts on 2 major BART stations.

Figure 8: Micro-simulation validation.

We proposed an LSTM model that is capable of learning explicit location choices that is applied on the labeled activity sequences.

Both IO-HMM\(^3\) and LSTM\(^4\) models are 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.

We further demonstrated the capabilities of the framework to contribute to the transportation planning practice by applying a discrete choice model on traveler’s mode choice. Travelers’ behavior can be interpreted from the estimated parameters, and the parameters such as the value of time and mode preferences used in micro-simulation or in other analysis methodologies of evaluating scenarios in transportation planning and policy making.

6.1 Travel Demand Explorer

Several improvements are planned to further build on the work presented herein. We have prototyped and will evaluate the usability of our system with a graphical travel demand explorer dashboard (Figure 9), which provides a system-level visualization of alternative mobility scenarios. It shows aggregated travel volumes and routes as well as basic interactive elements to evaluate the impacts of proposed policies and changes in transportation infrastructure on travel impact and level of service metrics.

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.

A range of issues remain where limitations of the current design present a challenge. One limitation involves DCM calibration from indirectly assigned socio-demographic variables based on the aggregated values within Census tracts. Other limitations are concerned with the nature of cellular data and inherent difficulties in the identification of the number of car-pools, on-demand vs. private vehicle trips, and modeling short-range and non-motorized travel to name a few.

\(^3\) IO-HMM code is available at https://github.com/Mogeng/IO-HMM
\(^4\) LSTM code is available at https://github.com/zihenglin/LSTM-Mobility-Model
Figure 9: Travel demand explorer interface.

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.

ACKNOWLEDGMENTS

This work was partially funded by a gift from AT&T. Support from the California DOT through UCCONNECT program, agreement 65A0529, is also acknowledged. We thank Dr. Rashid Waraich and Andrew Campbell for their help in micro-simulation set up.

REFERENCES