University of California
Center for Economic
Competitiveness in
Transportation

UCONNECT

Demand Forecasting and Activity-based Mobility Modelling from Cell Phone Data

Final Report UCCONNECT 2016 - TO 012 - 65A0529

Alexey Pozdnukhov; Assistant Professor; University of California, Berkeley

Sponsored by

Caltrans

2016

Berkeley UC Irvine UCLA UCR UCSB CAL POLY POMONA

STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT DOCUMENTATION PAGE

TR0003 (REV 10/98)

ADA Notice

For individuals with sensory disabilities, this document is available in alternate formats. For information call (916) 654-6410 or TDD (916) 654-3880 or write Records and Forms Management, 1120 N Street, MS-89, Sacramento, CA 95814.

1. REPORT NUMBER	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER
2016 – TO 012 – 65A0529		

4. TITLE AND SUBTITLE

Demand Forecasting and Activity-based Mobility Modeling from Cell Phone Data

Machine learning methods were developed to transform anonymized cellular data collected by carriers into regional activity-based travel demand models.

5. REPORT DATE

March 31, 2016

6. PERFORMING ORGANIZATION CODE

N/A

7. AUTHOR

Alexey Pozdnukhov

8. PERFORMING ORGANIZATION REPORT NO.

N/A

9. PERFORMING ORGANIZATION NAME AND ADDRESS

University of California at Berkeley
Institute of Transportation Studies
Berkeley, CA 94720

10. WORK UNIT NUMBER

N/A

11. CONTRACT OR GRANT NUMBER

65A0529

12. SPONSORING AGENCY AND ADDRESS

California Department of Transportation (Caltrans)
1227 O Street Sacramento, CA 95814

University of California Center on Economic Competitiveness in Transportation (UCCONNECT)
2616 Dwight Way, Berkeley, CA 94720-1782

13. TYPE OF REPORT AND PERIOD COVERED

Final Report, March 1, 2015- March 31, 2016

14. SPONSORING AGENCY CODE

N/A

15. SUPPLEMENTARY NOTES

16. ABSTRACT

This project develops machine learning algorithms and methods for processing of cell phone location logs to generate travel behavior data. The project initially focuses on bias correction and activity inference for generating activity-based travel demand models. Inferred activity chains are used to calibrate an agent-based traffic micro-simulation for the SF Bay Area, and validated on loop detector counts.

17. KEY WORDS

activity-based travel demand models, cellular data, machine learning, agent-based simulation

18. DISTRIBUTION STATEMENT

The readers can freely refer to and distribute this report. If there are any questions, please contact one of the authors.

19. SECURITY CLASSIFICATION *(of this report)*	20. NUMBER OF PAGES	21. COST OF REPORT CHARGED
Unclassified	53 (45+8)	Free for E-copy

Disclaimer Statement

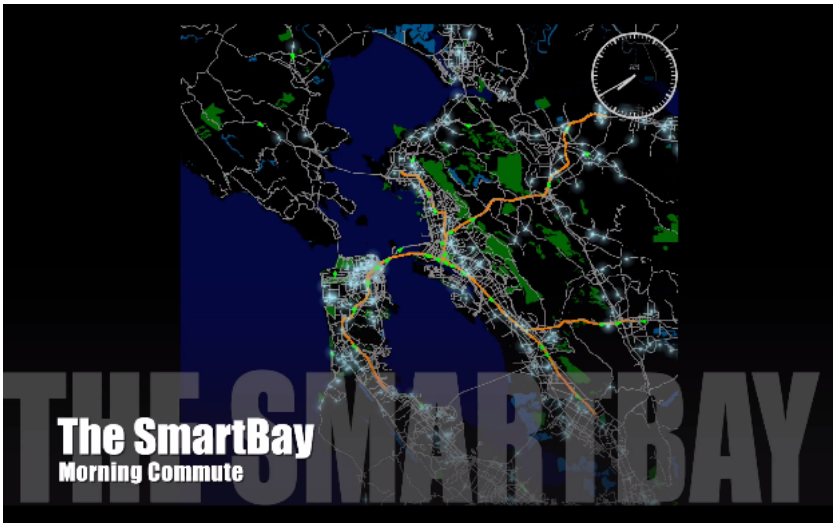
This document is disseminated in the interest of information exchange. The contents of this report reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California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

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

Demand Forecasting and Activity-based Mobility Modelling from Cell Phone Data

Final report submitted to Caltrans

Contract NO. 65A0529
Principal Investigator: Prof Alexey Pozdnukhov



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

Contents

List of Figures

List of Tables

Acknowledgments

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

Chapter 1

Introduction

1.1 Overview of Report Contents

This report covers research results in developing machine learning based methods used to produce activity-based travel demand models from locational data available to cellular telecommunications operators in a form of Call Detail Records (CDRs).

It covers main steps in model development, gives technical details on each of the steps, and describes an application of demand modelling within an agent-based simulation platform useful for planning and scenario evaluation.

1.2 Problem Statement

The consequences of population growth and increasing rates of urbanization are multi-faceted, but the associated pressures on constrained resources are becoming a major issue worldwide. This growing pressure on ageing infrastructure is already affecting the quality of citizens lives and limiting economic growth. The transportation field is responding to these global trends and evolving at an ever increasing pace. Novel mobility paradigms such as increasing multi-modality, on-demand transportation and car/ride sharing contribute to a possible solution, but they also change the transportation landscape quicker than traditional data sources, such as travel surveys, are able to reflect. Volatility of job markets, evolving demographics, internal migration and influx of citizens and businesses to cities further increase the intrinsic variability of the evolution of travel demand patterns.

These dynamics of changing demand are shaped by individual choices that are made in different contexts, as the heterogeneity of preferences influences collective outcomes such as mode shares in transportation, or Vehicle Miles Travelled, or the degree of urban sprawl. It is the decisions of individual persons, households and businesses that ultimately shape cities, interacting within a political economy of development. Quantifiable models based on bottom-up individual-level data are required to guide top-down policy regulation and governance in the face of these fast changes.

It is therefore more important than ever to be able to measure and realistically model travel demand in near real-time and at the level of individual travelers, and link it to the macro-scale level for policy analysis and decision support. Demand prediction inaccuracies alone reduce efficiency of infrastructure investments in all but few projects. Fortunately, as sensors and localization technologies have become ubiquitous over the past decade, mobility data has increasingly grown in volume giving rise to new opportunities for high accuracy data-driven modelling. With the emergence of these technologies, and their rapid growth, it has become possible in the US to perform traffic monitoring at unprecedented scales. Billions of locations are generated every day from mobile devices. In the context of this project, we focus on cell phone location data as signal- derived geographic location data collected by telecom operators. These data collection is non-invasive and does not require applications to be installed on users mobile devices. While some research has already been conducted to show usefulness of cell phone data in transportation, including the previous work of the PI, much remains to be done. This reports presents recent developments of the methods for processing cell phone location data to generate travel behavior information with the focus on demand modelling and activity-based traffic micro-simulation.

1.3 Research Objectives

Cell phone location data show potential to provide robust information about activity location, frequency of repeated travel, small area origin-destination data, and much more. This information can be used to model, evaluate, and analyze the movement and flow patterns of a study area. But the opportunity comes with significant scientific and technical challenges. It has yet to be studied if cell phone data is representative of population groups and their travel behavior. No guidelines exist on procedures for reasonableness

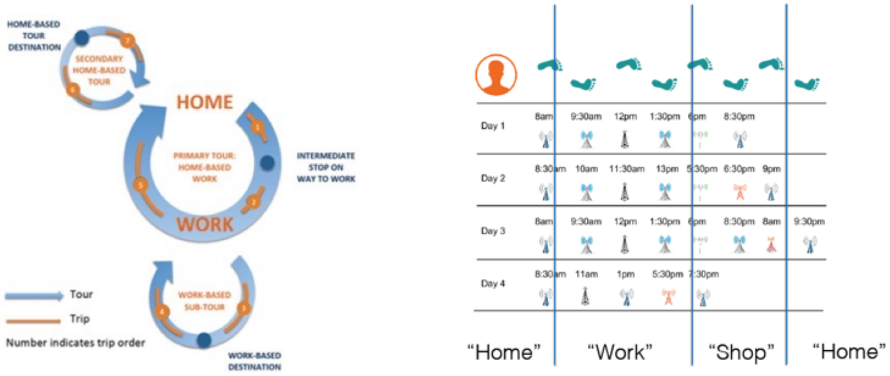


Figure 1.1: Left: a typical activity-based travel demand model structure. Home and Work states present the framework for secondary tours in the day sequence of activities and associated set of tours and trips. Right: the project's goal is to convert sequences of cellular traces into sequences of activities, and validate the resulting travel demand model within an agent-based microsimulation.

checks and quantitative comparisons of cell phone derived data with other data sources. Theres no established methodology for applying statistical indicators and benchmarks to assess data quality. The issues of sampling bias, geographic level of detail, availability and resolution of data for urban and rural areas, coverage among both cell phone users and cell phone carriers have to be resolved. Moreover, processing huge volumes of these data requires scalable methods and distributed computing implementations. The above-mentioned challenges are at the core of the initial effort completed within the project. This report focuses on an application in the San Francisco Bay Area (the SmartBay project) for which the experimental results are presented.

The project focuses on particularly promising applications for the use of cell phone data in the development, estimation, and calibration and validation of a travel demand models. We accept a common structure of the models, centered on Home- and Work- based tours (Figure 1.1).

Non-parametric scalable methods of machine learning can be used to infer detailed home- and work-based trips and origin-destination matrices, tour composition, individual and joint travel, and activity type and duration. These methods enable cell phone location data to be used in various types of travel behavior analyses, such as mode choice, trip purpose, trip chaining, analysis of anomalies and disruptions during special events. A step-by- step

methodology to prepare, process, validate, and apply algorithms to cell phone location data for different travel behavior analyses have been explored and presented in Chapters 2,3,4 and 5.

Due to the need to protect privacy, it is not expected that cellular providers will make raw data directly available to transportation agencies. Moreover, data alone are only partly useful for policy analysis and decision making. It is necessary to develop models for what-if scenario evaluation that are able to extrapolate and forecast the response of dynamics of transportation systems with respect to changes in infrastructure and evolving demographics. This project therefore concentrated on the methods that can be implemented internally on secure data provider's infrastructure where data are kept secure. The project investigated the use of anonymized and aggregated cell phone data within the two stages of foreseen applications:

- Demand forecasting grounded in activity-based models, performed internally on data provider infrastructure.
- Agent-based traffic modelling with virtual (simulated) travel itineraries derived from the calibrated activity-based models.

This two stage approach of using travel itineraries generated from activity-based travel models allows accurate representation of observed travel and provides a working scheme for real-world deployment within a public-private partnership.

1.4 Significance of the Study

The methods developed in this project allow drastically reducing timescales at which demand models become available to MPOs and regional and state transportation agencies. While traditional survey-based approaches require intensive labor to collect and process data, and may take several years to complete, the developed methods provide most up-to-date travel demand models with a latency of several days. Preliminary results on applying the developed models to simulate a typical weekday travel in the SF Bay Area show promising performance and accuracy.

Chapter 2

Data Handling and Bias Correction

2.1 Data Handling

Activity based travel models are the main tools used to evaluate traffic conditions in the context of rapidly changing travel demand. However, data collection for activity based models is performed through travel surveys that are infrequent, expensive, and reflect the changes in transportation with significant delays. Thanks to the ubiquitous handheld smartphone devices, we see an opportunity to amend these surveys with data extracted from either devices (GPS), service providers and smartphone apps operators (LBSN) or network-side carrier mobile phone usage logs, such as call detail records (CDRs).

2.1.1 Travel Behavior Data Sources

Below we outline the differences between several main types of data sources usually considered as alternatives to the traditional travel surveys, as well as related state-of-the-art methods used for data processing.

GPS. GPS data is granular in both spatial and temporal resolution, but is usually available for a very limited sample of the population. GPS data are usually collected for a small sample of a household travel survey participants. With no manual annotation of important location, GPS data need to be processed with state-space models for travel itineraries recognition. Using GPS data, Hidden Markov Models (HMMs) were extended to model

clustered historical locations[25]. A discriminative version of the state-space model was proposed by [22]. It unified the process of map matching, place detection, and significant activity inference through a hierarchical conditional random field (CRF). However, as a supervised and discriminative model, it still requires manually labeled data for training.

LBSN. Locational-based social network (LBSN) data is usually exact in locations, and may provide additional social relation, comments and reviews of the locations. It is available through service (apps) providers such as Twitter, Foursquare, and others. However its temporal resolution is limited by the discontinuity and often large gaps between subsequent 'check-ins'. Recent work on using these data for travel modelling have shown promise but highlight the need for further research. Cho et al. developed a period and social mobility model (PSMM) to separate social trips from commute trips [6]. Ye et al. created an extended HMM model that incorporated spatial and temporal covariates to classify activities into one of 9 distinct categories [32]. Kling applied a probabilistic topic model to obtain a decomposition of the stream of digital traces into a set of urban topics related to dominant activities within neighborhoods [20].

CDR. The anonymized Call Detail Records (CDRs) from cellular network operators provide a compromise between spatial-temporal resolution and ubiquity. Despite its poorer resolution in space, CDR data provides ubiquitous coverage for millions of people within any given large metropolitan area. Latent Dirichlet Allocation (LDA) and Author Topic Models (ATM) were used to cluster the daily CDR trajectories [12, 13]. The use of auxiliary land use data and geographical information database has emerged to mine possible activities around a certain cell tower [27]. [33] analyzed the semantic meaning of historical trajectories. Whenever the next location is to be determined, historical semantic trajectories are matched and the location with the highest matching score is predicted. Zheng et al. used a simple graphical model without state transition to explore the hidden activities with respect to temporal features[34].

This report deals with CDR data. Below it gives specifics on the processing and general spatial coverage bias correction procedure for Home- and Work inference. Building on state-of-the-art methods, next chapters describe methods for secondary activity inference.

2.2 Data Processing and Home- and Work Detection

The first step to inferring both primary and secondary activities is to extract activities locations from the raw CDR traces. A common way of extracting activity locations is by spatial clustering followed by a filter based on dwell time [17, 9, 35, 31]. However, these methods have two main drawbacks. First, most of these approaches do not tackle a common feature in the cell phone data of connection oscillations between cell towers. An oscillation occurs when a user is stationary, but his/her cell phone switches to communicate with another antenna nearby, due to the signal propagation conditions at the time of the connection or as a result of the load balancing mechanism in data transmission over the network. These oscillations can occur between antennas located as far as several miles apart. Second, the spatial resolution itself is limited to the area covered by a single antenna. These oscillations, along with the low spatial resolution, result in two potential problematic situations: (1) A user might be standing still but seen as moving (oscillations); or (2) a user has moved within the range of a cell tower but is seen as standing still. In this report, we describe a stay location extraction algorithm that filters the obvious oscillations while not over-filtering short-term travel.

2.2.1 Data processing

In this section, we introduce terminology used in the report, as illustrated in Figure 2.1.

Definition 1. ***CDR log****: In addition to call, data, and short message records the CDRs used in the experiments below are enriched with cell tower handover data. Each record contains the start time, end time, duration, and the latitude and longitude of events provided by the data collector.*

Definition 2. ***CDR trace****: A CDR trace is the sorted list of relevant CDRs by start time, for a single user.*

Definition 3. ***Stay history****: CDR traces exhibit positioning noise and oscillation noise. A filtering algorithm is applied to infer true location clusters and turn the raw traces to sequences stay points. Each stay is represented by the location cluster, start time, end time and duration. The stay history is the time sorted list of stay points.*

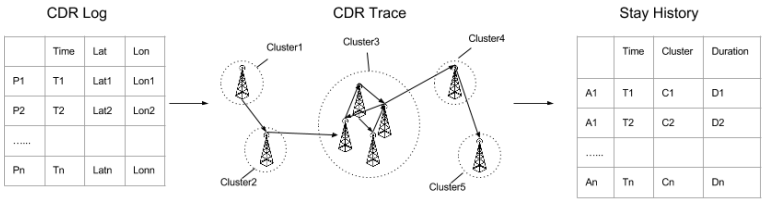


Figure 2.1: Preliminaries and terms definitions

Definition 4. ***Activity****: An activity is a stay point with a semantic label of trip purpose, such as "home", "work", or "shopping".*

2.2.2 Activity Locations Inferences

As we are interested in the activities associated with travel, our approach starts with a conversion of the CDR traces of every user into a sequence of stay points. In brief, the main conversion steps are as follows. For each cell phone user: (1) Cluster CDR records with similar locations using a density-based clustering algorithm, (2) construct an oscillation graph to identify potential oscillation sites, (3) filter out oscillations using rules inferred from data collection statistics, and, (4) filter out stay locations where the stay duration is less than a defined threshold. Recognizing the importance of long-term recurrent stay points such as "home" and "work" that enforce a structure in the users' daily mobility, we treat them in the next pre-processing step.

2.2.3 Home and Work Inference

The key step is to infer locations of primary highly regular activities such as home and work. This problem has received considerable attention in location sequences analysis [17, 21, 16], providing solutions based on ranking places by duration and number of visits within different periods of time.

Various strategies have been used for home and work location detection. We adopt accepted methods in order to simplify processing and, most importantly, infer "anchor" points in the daily sequences that provide space-time context that is crucial to build a generative model of secondary activities.

A mixture of Gaussians is a popular method to model locations centered on home and work [6]. Another suggested definition of "home" was the

location where the user spends more than 50% of time during night hours with night hours defined as 8pm to 8am [21]. Similarly, work hours can be defined as the area where the user spends more than 50% of time during day hours. Our detection of the home and work locations is similar to the method in [21], we identify home as the location where the user spends the most stay hours during home hours, and we identify work as the location where the user spends the most hours during the work hours. However, we define home and work hours to be much narrower time windows than the 8am-8pm criteria used in [21]. Borrowing from [17], the hours midnight to 6am are defined as home activity hours, and from [16] 1pm to 5pm on weekdays are defined as working hours. According to [16] 1pm to 5pm captures the core set of working hours for both early and late workers. If there is a tie in number of home hours spent at multiple stay locations, the stay location with the most visits during the home hours is considered the home location. The same is true for the work location.

2.3 Spatial Coverage Biases

Detected home and work locations are aggregated spatially at the level of cellular network resolution (typically an antenna or a sector level). Spatial coverage of the cellular network varies according to the constraints of the built environment, as well as the anticipated demand for cellular communications, FCC regulatory directives, carrier's internal planning and availability of resources in providing best possible coverage for their customers. Customer base (the penetration level of any given carrier) can not be assumed to be homogeneous in space nor proportional to the total population numbers or represent the socio-demographics. While the latter problem can not be accurately solved with no personal data on the carrier's customers, the spatial coverage biases can be corrected.

Assuming the Census data on housing and employment in the region as the ground truth for night- and day- time population distributions, the CDR-derived home and work numbers can be re-calibrated accordingly with spatially-varying linear adjustment coefficients. Then, an Iterative Proportional Fitting (IPF) is applied to adjust the Home-based Work trips OD tables in a way that matches the marginal night- and day-time population distribution from Census data.

Adjusted Home-based-Work trips can then be extended with a set of sec-

ondary trips derived with a generative machine learning model as described in the next Chapter.

Chapter 3

Activity Inferences with Input Output Hidden Markov Models

Inference of activity types beyond "home" and "work" (such as "dining", "shopping", or "leisure", etc.) from cellular data is a non-trivial task. Cellular data, while collected at scale, suffers from limited spatial resolution dictated by the spatial siting of cellular antennas, as well as from lacking the ground truth observations.

This chapter describes the last step in the three-step approach to annotating user activities with machine learning methods. This last step is to understand secondary, non-mandatory activities and activity transition patterns. Choices of secondary activities are less constrained in daily schedules and are more flexible and variable throughout the days. A generative model is required to capture this inherent variability. We apply Input-Output Hidden Markov Models (IO-HMMs) to learn the activity pattern across multiple users in an unsupervised manner. These patterns include: (1) Heterogeneous transition probabilities between activities given different contexts; and (2) Spatial and temporal profile of activities. This part of the model can answer questions such as:

- If it is evening and a user has been working for 8 hours, what is his/her most likely next activity?
- If a user is going shopping, how far away from his/her home will it likely located?
- If a user is going for a recreational activity, how much time will he/she most likely spend there?

3.1 Modelling Travel Choices with Machine Learning Methods

We start by explaining the range of machine learning models available for the task, and justify our choice towards the IO-HMM.

3.1.1 Supervised & Unsupervised Models

Supervised models. Supervised learning methods require data with labeled ground truth. The ground truth are either manually labeled [8, 14], or collected for a small group of participants from a survey accompanying GPS data [19]. Using data collected from natural mobile phone communication patterns of 80 users over a year with labeled ground truth, different supervised learning models including SVM and decision trees were compared to learn the activity pattern of users [24]. Logistic regression model was used for identifying semantically meaningful places from 60 user's CDR data with ground truth [16]. However, the activity categories are mainly limited to home and work, excluding secondary activities. Liao et al. manually labeled ground truths to extract places and activities [22, 23]. However, this model was only applied to 4 people and is not scalable to large population.
Unsupervised models. On the other hand, unsupervised models (including topic models and state-space models) are used to cluster activities with similar temporal and spatial profiles. Most of the inferred routines are interpretable, including"going to work late", "going home early" and "working non-stop" [12, 11, 13, 10, 34]. "Eigenbehavior" decomposition found that communities within a population's social network tend to be clustered within the same behavior space [7].

3.1.2 Discriminative & Generative Models

Discriminative. Discriminative state-space models such as CRFs [22, 23, 31] are more flexible when modelling the relationship between input, output and state variables. However, due to their undirected nature, discriminative state-space models cannot be used for activity generation directly.
Generative: Hidden (Semi-) Markov Models (HMM/ HSMM) are generative models that can not only be used to analyze activity patterns, but also generate new sequences [28, 25, 15]. Recently, Baratchi et al. developed a hi-

erarchical hidden semi-Markov-based model that could capture both frequent and rare mobility patterns in the movement of mobile objects [3].

In this work, not only are we interested in understanding the activity patterns themselves, but we also aim to fit these patterns into probabilistic analysis and activity based travel micro-simulation. Thus, we require generative models. At the same time, privacy considerations preclude us from collecting ground truth data, restricting us to the use of unsupervised models. Moreover, in order to produce the activity pattern of a large population of users, we build models that shares parameters across user groups.

3.1.3 IO-HMM for Secondary Activity Recognition

Given the user stay history, that is, a list of stay points with start time, duration and location, we want to convert the list into a sequence of activities with semantic labels and identify interpretable activity patterns. To be more specific, the activity semantics can be defined by: (1) Spatial and temporal context such as land use type in the area, start time, and duration. (2) A heterogeneous context-dependent probability model for transitions between activities.

Creating a model that carry aforementioned information is significant for two reasons. First, it provides a backbone for modular travel demand modelling required by transportation practitioners. A range of travel choices (for example mode of transportation, activity location, etc.) depend on the structure of daily activity plan, and the durations and start/end times of activities. Second, such generative model is essential for representing variability of travel choices and consequent network flows in probabilistic or computational "what-if" scenario assessment in transportation systems analysis.

IO-HMM Architecture

Hidden Markov Models (HMMs) have been extensively used in the context of action recognition, and signal processing. However, standard HMMs assume homogeneous transition and emission probabilities. This assumption does not hold for our problem. For instance, if a user engages in a home activity on a weekday, if they depart the home activity in the morning, they are likely going to work. If they depart the home activity in the evening, the trip purpose is likely to be leisure or shopping. Therefore, we propose to use the IO-HMM architecture, by incorporating rich context information to

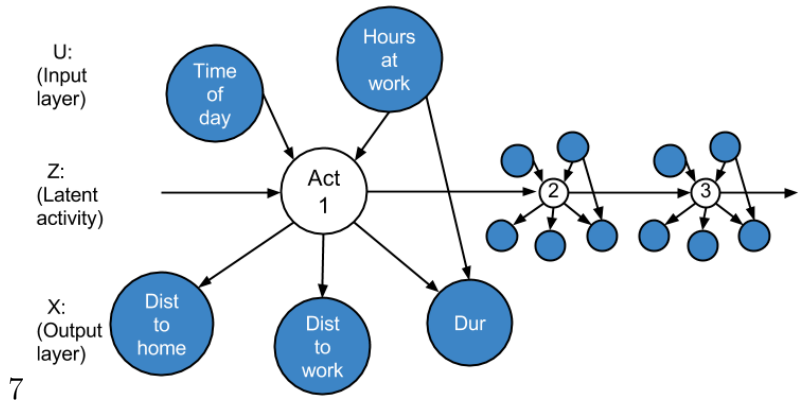


Figure 3.1: IO-HMM Specification

overcome the drawbacks of the standard HMM. In Figure 3.1, the solid nodes are observed information while the transparent nodes are latent variables. The top layer U is the observed contextual variable layer, such as time of day, previous location information. The middle layer Z is the latent activity layer. The bottom layers X are other observed information, such as spatial choice of the activity and duration of the stay. It is worth noting that each output can be associated with different context information that best explains the observations for each output.

By adopting the IO-HMM architecture, we assume that our latent state (activity) will not only depend on the previous state, but also some contextual information such as the time of day. We also assume that some of the outputs will not only depend on the state, but also some of the context information.

Formulation

IO-HMM architecture has been well described in [4, 30]. Here we will only highlight the main difference between IO-HMM and standard HMM, as summarized in Table 3.1.

Parameter Estimation

Similar to standard HMM, Expectation-Maximization (EM) has been widely used to estimate the parameters of IO-HMM.

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

Table 3.1: Highlights of HMM vs IO-HMM

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

is filling in the hidden variables using the knowledge of observed data and previous parameters.

M step: Update the parameters to maximize the expected data likelihood, which is given by:

$$\begin{aligned} Q\left(\theta, \theta^k\right) &= \sum_{i=1} \gamma_{i,1} \log \Pr\left(z_1 = i \mid u_1; \theta_1\right) \\ &+ \sum_{t}^{T} \sum_{i} \gamma_{i,t} \log \Pr\left(x_t \mid z_t = i, u_t; \theta_2\right) \\ &+ \sum_{t=2}^{T} \sum_{i} \sum_{j} \xi_{ij,t} \Pr\left(z_t = j \mid z_{t-1} = i, u_t; \theta_3\right) \end{aligned} \tag{3.1}$$

The estimation of the initial probability, transition probability, and all output parameters can be done in parallel. It is also worth noting that θ_1 and θ_2 can be estimated with any supervised model that allows for probabilistically interpreted outputs, such as a neural network. θ_3 can be estimated with any supervised model, such as generalized linear models or neural networks. The only criterion is that the implementation of these supervised models must support sample weights. The sample weights are given by the posterior transition probabilities ξ and posterior state probabilities γ, as in Equation 3.1. We implemented an easy to use Python IO-HMM code that supports many output types, including generalized linear models, multinomial logistic regression, and neural networks with categorical or probabilistic outputs. Moreover, the EM algorithm can be naturally adapted to the MapReduce

framework, a programming model and associated implementation for processing and generating large data sets with a parallel, distributed algorithm on clusters. The Expectation step can be fit into the Map step, calculating the posterior state probability γ and posterior transition probability ξ in parallel for each sequence. The estimated posterior probabilities γ and ξ will be collected in the Reduce step. Then, standard supervised learning model implementations that supports sample weights will be used to update the parameters. Most of these standard supervised models can be trained in the MapReduce framework as well.

Input Output Specification

For initial probabilities and transition probabilities, we include the following model inputs (1) a binary variable indicating whether the day is a weekend; (2) three binary variables indicating the time of day that the activity starts, morning (5 to 10 am), afternoon (12 to 2 pm) or night (5 pm to midnight); and (3) the number of hours the user has been at work this day.

The IO-HMM model also includes the following outputs: (1) The euclidean distance between the stay location and the user's home; (2) the euclidean distance between the stay location and the user's work; (3) the duration of the activity. Linear models are used as the output models for output (1), (2), and (3). This selection of the inputs and outputs is not arbitrary. The activity start time is relevant to differentiating activity types. The number of hours worked that day is also a strong indicator of a person's likelihood to return to work (after a midday activity for example).

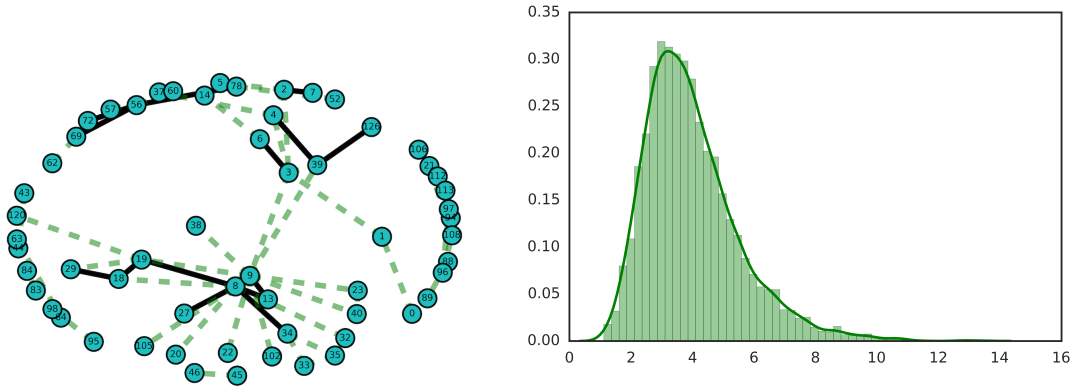
The model input features contain information that is known at the start of the transition to a new activity. In contrast, the output features contain information that is not available at the transition to a new activity. For example the duration and location or land-use at the vicinity of a new activity is unknown at the time of the transition. The model outputs have a strong dependence on the activity type. For example the distance that a person is willing to travel from home for a leisure trip may be longer than the distance that a person is willing to travel for a shopping trip. In this sense, the output is explained by the activity type. The duration depends both on the activity type and on whether the location has already been visited in the day. For example, if a person works for a few hours, leaves for lunch, and then returns to work, we expect the afternoon work duration to be shorter. This relationship is depicted graphically in Figure 3.1.

Finally, we emphasize that an activity label is just a latent variable. It is usually interpreted based on its spatial-temporal profile. These spatial-temporal profile are just the input and output observations associated with each activity. For instance, if we see an "activity" is prone to happen at noon in an area with many restaurants, we may label this activity as "lunch".

Chapter 4

Agent-Based Traffic Micro-Simulation

The final objective of this project is to demonstrate the efficacy of using cellular data and our activity and mobility inference algorithms for travel demand modeling through the creation of an agent-based microsimulation on the MATSim platform (http://www.matsim.org, [2]). Using cellular data for building travel demand models offers many benefits to transportation agencies. Travel demand models are traditionally estimated using either stated or revealed preference travel survey data. These data collection methods provide a rich set of features for survey participants, but are limited in four critical ways. 1) One-off travel surveys are expensive, both in terms of monetary and time costs. 2) Sample sizes are very small compared to the populations they are supposed to represent. 3) Travel surveys are not longitudinal. In most instances, a travel diary is collected for only one or two weeks, making it difficult to capture seasonal effects. 4) Travel surveys are conducted infrequently. Given the accelerating pace of global urbanization, metropolitan economies and land-use evolve too rapidly to use traditional travel survey methods for building demand models. Transportation practitioners need more temporally relevant data sources.



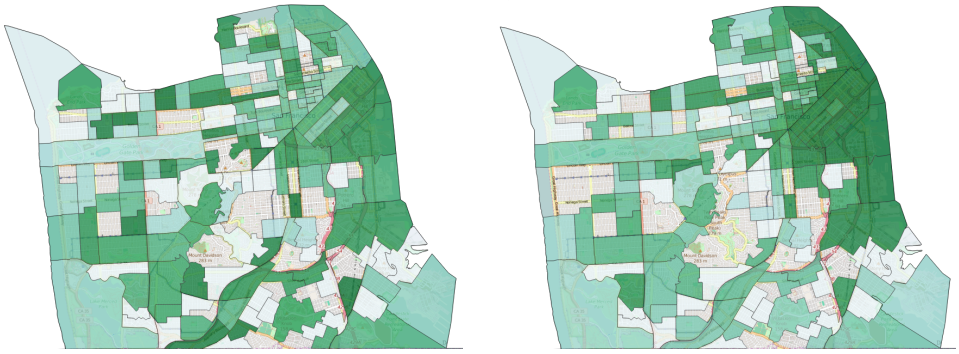
(a) Sample cellular connections oscillation graph

(b) Average number of daily activities within a selected user group

Figure 4.1: Activity inference algorithm outputs

4.1 Demand

This sections presents a detailed description of the results derived from applying the processing steps described in Chapters 2 and 3 to a CDR dataset. The data used in this study comprise anonymized and aggregated CDR logs collected by a major mobile carrier in the US, in the San Francisco Bay Area. No personally identifiable information (PII) was gathered or used in conducting this study. Moreover, the level of spatial aggregation of the data records makes it impossible to recover the identity of specific individuals. After initial anonymization of any customer personally identifiable information in the data, all locations are uniformly randomized within analysis zones that contain at least several hundred customers, such that customers cannot be distinguished from hundreds of others within the same zone. Moreover, note that the actual output of the models presented in this report can be seen as an additional level of aggregation of each customers' location profiles. We randomly selected 10% of the available anonymous subscribers in the area of interest for the experimental study presented in this section of the report.



(a) Home locations (b) Work locations

Figure 4.2: Home and work inference results

4.1.1 Home and Work Inference Results

4.1.2 Pre-processing of Data

We pre-process the data following the oscillation removal heuristics. An example oscillation graph described in that section is shown in Figure 4.1a. Each node in the graph represents a location cluster. There is an edge if oscillation has been observed between two clusters. The thicker the edge, the more oscillations have been observed. After filtering, the distribution of the average number of activities is shown in Figure 4.1b. The average number of activities per user per day is 3.8. This is consistent with the California national household travel survey reporting numbers between 3.6 and 4 [1].

Figure 4.2 shows an example of the density of inferred home and work locations for the sample we describe, aggregated at the census tract level within the City of San Francisco. Many of the inferred work locations are in Downtown San Francisco, the Financial District, and the SoMA, three San Francisco neighborhoods with high employment density [18]. The home locations, as expected, are more spread out throughout the city. Spatial coverage correction was applied to the data, and adjustment coefficients computed using the 2015 projections of population and employment based on the recent results from the ACS as published as a part of MTC's Travel Model One.

4.1.3 Activity Inference Results

Activity Profiles

For a descriptive example below we randomly selected a set of anonymous commuters with identified approximate home and work locations. Since weekday mobility is of most interest for most applications, we included only data on weekdays. The analysis of different activity patterns between weekday and weekends is left for future studies.

Figure 4.4 depicts the distribution of start times of activities. The y-axis gives the number of users who started the activity at a given hour. By evaluating these weekly activity start-time patterns in combination with the output coefficients in Table A.1, we can assign semantic labels for activity type to each of latent activity states. Note that identification of "home" and "work" activities benefits from the previously detected home and work locations.

For example, we have identified activity state 1, shown in blue in Figure 4.4 as the "work" activity. The activity has a peak start time of about 9 am each day, the distance to work is approximately 0 miles, the average duration is 7.44 hours and will decrease by 0.88 hours for each hour a person has worked on that day. This activity has highest peaks in Figure 4.4, signifying that this is a very regular activity with concentrated start times. The positive visited coefficient from Table A.1 associated with State 1 represents that it is visited frequently.

Latent activity state 6 is also easy to label, one can identify it as the "home" activity. The typical start time of this activity ranges from 3pm to midnight. The distance to home is approximately 0 miles, and the average duration at this activity is about 14 hours. The activity is also quite regular, though the start time has more variation than the start time of the "work" activity. This can be explained by the fact that people have more variation on the secondary activity selection before going back to home. It is also regularly visited, as indicated by the highest "visited" coefficient in Table A.1

The remaining states are slightly harder to label, but we do see some clear trends that suggest what the activity type is. Activity 0 and 3 show similarities in start time. There are peaks in start time around noon and in the evening. As shown in Table A.1, activity state 0 has an average duration of about 0.7 hours, and usually happens closer to home than work. Activity 3

has an average duration of 1.2 hours and usually occurs close to work. Based on these properties, we assign activity 0 the label "shopping" and activity 3 the label "food near work".

We have assigned activity state 4 a label of "leisure/ personal". The state has an average duration of almost 4 hours, much longer than the durations of activity state 0 and 3. While the 3.5 hour duration could suggest a leisure or a work-related activity, we see in Table A.1 that the activity is more likely to occur near leisure land-use or restaurant and shopping land uses than personal business. As shown in 4.4, it often starts in the evening hours, and tellingly, more users engage in this activity on Fridays than the other weekdays. Also as indicated by Table A.1, state 4 has higher probability of having a leisure land use than other states.

State 2 is hard to categorize. We have assigned it the label "medium distance trip". The distance from home and work are 12 and 18 miles, respectively, but the average duration of this trip is relatively small (1.5 hours). This state could encompass both off-site work related trips and/or longer-distance dining or leisure activities. As shown in 4.4, this activity type is not very common compared to the "home" and "work" activities, or even compared to the "food/shopping" activities.

Activity state 5 is the most irregular and infrequent. The average distance from home and work is quite high (about 140 miles). This activity type seems to occur predominantly on Fridays and likely encompasses longer-distance weekend leisure trips. Its irregularity is also suggested by the negative visited coefficient in Table A.1.

If we focus on the the daily start time and duration distribution, as in Figure 4.3, we can be more convinced of our labeling decisions. The downward slope of the start hour vs. duration plot shown in Figure 4.3c signifies that if a user arrives at home later in the day, they are likely to spend fewer hours at home. This matches our intuition - people who go out to other activities after work likely arrive home later and spend fewer hours at home before leaving the next morning. There is also a small cluster of users who spend a few hours at home in the afternoon before leaving for another activity.

Figure 4.3e shows the start time vs. duration for the "work" state. There is a large concentration of people who start the work activity early in the morning and stay for 8 - 10 hours. There are also people who visit the work state for shorter durations in the morning or afternoon. These clusters likely correspond to people who work for a few hours, leave during the lunch hour, and then return to work.

Activity Transition

Finally, one of the strengths of our generative model is that it can generate the expected next activity based on the heterogeneous transition probabilities. Figure 4.5a shows the transition matrix associated with mornings. The labels on the left indicate the state the user is transitioning from, and the labels on the top indicate the state the user is transitioning to. The most significant transition is from "home" to "work". Figure 4.5b shows the transition matrix associated with evenings. The most significant transition is from all other states to "home". However, if the user's transition from activity is "home", then he/she is more likely to transitions to "food" or "leisure". Figure 4.5c shows the transition matrix in the afternoon, for users who have not yet visited the "work" state in the day. For these users, there is a high probability of going to work. Figure 4.5d shows the afternoon transition for users who have already spent 5 or more hours in the "work" state. For these users, the probability of returning to work is significantly reduced compared to Figure 4.5c. By keeping all the input context information equal as in the previous case, and only specifying that the simulated user has previously worked for 5 hours on that day, one can see that the probability of going to work is significantly reduced.

4.2 Agent-based travel micro-simulation

The cases study encompasses the San Francisco Bay Area: a 7,000 sq-mi region spanning the nine counties under the jurisdiction of the Association of Bay Area Governments. The 2014 population of the Bay Area is estimated to be 7.5M residents. Data sources for the road network creation are discussed below.

Network Data

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

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

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

Demand Data

One of the essential input datasets for MATSim is a file describing the daily activity sequence for every agent in the population. The combination of methods described in Chpaters 2 and 3 allows generating a set of travel itineraries for the population in the area for further analysis within an agent-based simulation. We generated these demand files using a hybrid approach combing CDR-inferred activities with the Metropolitan Transportation Commission's (MTC) Travel Model.

MTC Travel Model Data

The MTC Travel Model is an activity-based demand model developed by Parsons Brinckerhoff, Inc. under contract for the MTC [26]. It is a member of the Coordinated Travel - Regional Activity Modeling Program (CT-RAMP) family of models. The model development, calibration and validation process

is described in a 2012 report. Agent populations were synthesized using historical and forecasted census and socio-economic distributions. The 2000 US Census Public Use Microdata Sample (PUMS) was used for generating empirical distributions of eight person types and four household types employed by the model. Aggregated TAZ-level socio-economic distributions from the year 2000 were provided by the Association of Bay Area Governments (ABAG). The baseline model used population distributions from the year 2000. Future scenario populations were drawn using IPF with forecasted distributions of TAZ-level person and household categories and socio-economic variables. The activity segmentation was based on the 2000 Bay Area Travel Survey (BATS). The 16 original activity categories from BATS were aggregated into 10 types for the Travel Model. All major agent decision making, excepting traffic assignment were modeled using a sequence of multinomial logit choices ranging in scope from work and school location to intra-tour mode choices. MTC Travel One model was calibrated and validated against the Caltrans State Highway traffic count databases.

In demand and activity generation, we used the MTC Travel Model (detailed above) and cellular data in the following way.

- **Primary activities.** Home and Work locations detected from cellular data were randomized within each TAZ and adjusted (scaled up) according to the total population estimates available from ABAG as published within the MTC Travel Model One.

- **Secondary activities.** While the developed methods allow us to generate secondary activities for individuals (see section 4.1 above), further research is required to guarantee that samples generated from the model will provide guarantees in protecting users locations privacy. Particularly, in case of the IO-HMM model that is trained on historical movement data of a single individual, it may over-fit a set of activities and locations in a way that in its entirety it will be unique to the given user (a set of locations and transitions may represent a mobility "fingerprint" for an individual). While this risk is minimal, in the following model that we describe, a set of secondary activities was replaced by the fully synthetic tours generated with the MTC model. Further research is required to understands privacy guarantees and aggregation trade-offs in modelling detailed travel itineraries of travellers.

Interested readers are encouraged to contact the authors regarding the

results obtained for the demand model with secondary activities that are fully based on cellular data.

Performance data

Validation of the traffic volumes was conducted using freeway traffic counts from the Caltrans Performance Monitoring Systems (PeMS). We developed tools in the Python programming language for conducting spatial matching in order to assign PeMS sensor stations to the appropriate links in the virtual MATSim network. At a total of 1,166 sensors were successfully matched and passed data quality filtering. We collected the 5-minute rollup traffic count data from June - August, 2015. For each sensor, we generated a typical weekday profile by taking the hourly mean volume for weekdays. These profiles were used for calibration and validation.

4.3 Microsimulation Creation

4.3.1 MATSim Platform

The MATSim (Multi-Agent Transport Simulation) platform is an agent-based activity model that performs microscopic modeling of traffic (using link performance functions) and agent decision making [2]. The MATSim run cycle, Figure 4.6, is an iterative process whereby agents make adaptations to routing, activity timing, and other optional choices until convergence is reached. As input, each agent is assigned an activity chain (initial demand), complete with activity types, timing and location. During the mobility simulation (mobsim), the agents travel the network, interact, and experience congestion which lowers their overall utility scores for the day. During the replanning phase, a subset of agents may adapt their routes and activity timings. For our simulations, we restricted replanning adaptation to random selection of 10% of the population during each iteration. Many other forms of adaption are possible with MATSim, but for this project we have restricted adaptation to timing and routing. Agents incur a negative penalty for deviating from their original activity timings, so dramatic shifts in activity start and end times are not possible. Rerouting agents are allowed to update their routes to the new shortest path, based on the loaded network conditions in the most recent mobility simulation.

We used the hybrid MTC-CDR activity model to generate initial demand for a typical weekday. The scenarios simulate a single 24-hour day for 463,000 agents, scaled up to represent the total driving population. For this initial implementation, we cast all agent demand into private passenger-car equivalents. Thus each virtual car only carries a single agent, but it may represent more than one passenger trip.

4.3.2 Model Parameters Calibration

For calibration, we used the full set of 1,166 PeMS sensor stations that could be matched to the MATSim network. Calibration efforts were evaluated based on the change to total absolute error summed over all sensors for the whole day.

Rahka et al. define model calibration as the selection of "input parameter values that reflect the local study areas network, climactic, and driver characteristics" [29]. Driver characteristics calibration enters into the hybrid MTC-CDR demand generation described in previous sections. It also enters within MATSim. Agents evaluate their day of experienced activities and trips using the Charypar-Nagel scoring function; a utility function tailored for the co-evolutionary learning algorithm that the agents employ [2]. The key behavior of this scoring function is a trade-off between the positive score accrued performing activities, and the negative score from traveling. We tuned several agent scoring parameters during calibration: how sensitive agents are to starting activities later than scheduled, the penalty for ending early, and the disutility of travel. These parameters were found to be of second order importance compared to changes in network performance.

MATSim provides two main levers for calibrating network performance: the flow capacity factor and the queue storage capacity factor. The flow factor dictates how rapidly link travel speed decays with volume. The link storage factor controls the link density constraints, which determine the acceptance rate of a link for incoming vehicles. The role of the two factors is succinctly described by Nurhan et al. (2003): "link outflows are constrained both by the flow capacity of the link itself and by space limitations on the receiving link" [5]. For calibrating these factors, a good initial guess is to take use the agent-to-population ratio. So if we have a 10% sample, a good initial guess is 0.10. However, the complexity of the simulation prevents easy

prediction of the impacts of adjustments to these factors and an iterative guess-and-check approach is required. For our population of 463,000 agents, we found that 0.12 worked
best for both scale factors.

The final calibration parameter is the counts scale factor. This parameter does not actually impact the agent behavior or link performance. It simply scales the simulated the counts up to match the observed volumes. After a simulation runs to convergence, we adjust the counts scale factor such that the total simulated and observed counts match. We chose a final value of 13.35, meaning each agent vehicle represents 13.35 observed vehicles.

4.3.3 Validation

Since we used the MTC Travel Model in our hybrid demand model, we sought to reproduce same validation metrics employed by that model's creators. We have attempted to reproduce the the measures in Tables 68 - 70 in the MTC Travel Model calibration and validation report [26]. These tables describe daily, AM peak, and PM peak predicted and observed flows at 27 key screenlines located at county borders and bridges. The AM peak is defined as 6:00 - 09:59, and the PM peak is 16:00 - 18:59. In Tables 4.1, 4.2, and 4.3, we have used our PeMS typical weekday profiles for the observed values and the MATSim simulated volumes for the predicted counts. We have included 'NA' placeholders for locations that were included in the MTC report, but for which no PeMS data was available.

4.4 Results

We ran our simulations on a Windows 7 desktop computer with an eight-thread Intel i7 processor running at 4.00 GHz. The machine has 64-GB of RAM and RAID 00 SSD storage drives. In a typical run, we allocated all eight threads and 24-GB of RAM to the Java virtual machine running MATSim. Typical running time for a population of 463,000 agents was 8.5 hours for a total of 10 iterations of the MATSim cycle. This is a sufficient amount of iterations to judge the performance of a new calibration configuration.

Screenline Facility	Observed	Predicted	Predicted Less Obsv'd	Percent Difference
Bay Area Bridges				
US-101, Golden Gate Bridge (S)	40,871	40,633	-238	-0.6%
I-80, SF/Oakland Bay Bridge	256,878	267,841	10,963	4.3%
Cal-92, San Mateo/Hayward Bridge (W)	56,619	67,520	10,902	19.3%
Cal-84, Dumbarton Bridge (N)	34,444	71,075	36,631	106.3%
I-580, Richmond/San Rafael Bridge (E)	NA	NA	NA	NA
I-80, Carquinez Bridge (E)	54,886	56,269	1,383	2.5%
I-680, Benicia/Martinez Bridge	110,225	120,775	10,550	9.6%
Cal-160, Antioch Bridge	NA	NA	NA	NA
Bay Area Bridges Sub-Total	***553,924***	***624,113***	***70,189***	**12.7%**
San Francisco / San Mateo Line				
US-101, Bayshore Freeway (N)	102,273	86,623	-15,650	-15.3%
Cal-35, Skyline Blvd. (N)	NA	NA	NA	NA
Cal-1, Junipero Serra Blvd. (N)	92,537	96,130	3,593	3.9%
I-280, Foran Freeway	161,584	110,498	-51,086	-31.6%
SFM/SM County Line Sub-Total	***356,394***	***293,252***	***-63,142***	***-17.7%***
San Mateo / Santa Clara County Line				
Cal-82, El Camino Real (N)	NA	NA	NA	NA
US-101, Bayshore Freeway (N)	98,064	87,050	-11,014	-11.2%
I-280, Serra Freeway (N)	62,123	39,784	-22,339	-36.0%
SM / SC County Line Sub-Total	***160,187***	***126,834***	***-33,353***	***-20.8%***
Santa Clara / Alameda County Line				
I-680, at Scott Creek Road (N)	42,818	59,981	17,164	40.1%
I-880, Nimitz Freeway (N)	71,029	123,771	52,742	74.3%
SC / Ala Line Sub-Total	***113,847***	***183,752***	***69,905***	***61.4%***
Alameda / Contra Costa County Line				
I-580, Knox Freeway	85,286	93,353	8,067	9.5%
I-80, Eastshore Freeway	166,246	188,255	22,009	13.2%
Cal-24, Caldecott Tunnel (E)	171,213	202,601	31,388	18.3%
I-680, in Dublin/San Ramon	159,296	194,990	35,694	22.4%
Ala / CC County Line Sub-Total	***582,041***	***679,199***	***97,159***	***16.7%***
Solano / Napa County Line				
Route 29, nodatapa-Vallejo Highway (N)	NA	NA	NA	NA
Solano / Sonoma County Line				
Route 37, Sears Point Road	38,001	26,319	-11,682	-30.7%
Napa / Sonoma County Line				
Route 121, Carneros Highway (N)	NA	NA	NA	NA
Route 128, Calistoga-Healdsburd Rd. (E)	NA	NA	NA	NA
Sonoma / Marin County Line				
US-101, Redwood Highway (N)	70,644	87,240	16,596	23.5%
Screenline Totals	**1,875,036**	**2,020,708**	**145,672**	**-30.7%**

Table 4.1: Screen Line Validation for Complete Day

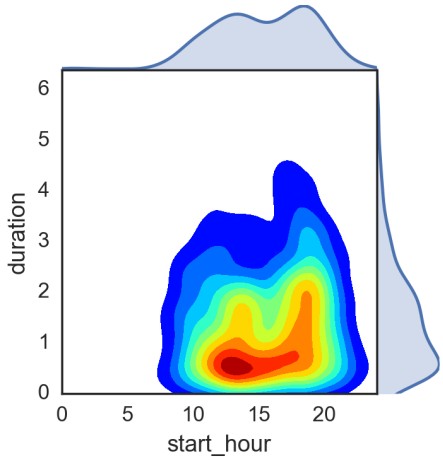
Screenline Facility	Observed	Predicted	Predicted Less Obsv'd	Percent Difference
Bay Area Bridges				
US-101, Golden Gate Bridge (S)	5,299	6,735	1,436	27.1%
I-80, SF/Oakland Bay Bridge	51,765	69,360	17,595	34.0%
Cal-92, San Mateo/Hayward Bridge (W)	18,382	20,502	2,120	11.5%
Cal-84, Dumbarton Bridge (N)	2,514	14,651	12,137	482.8%
I-580, Richmond/San Rafael Bridge (E)	NA	NA	NA	NA
I-80, Carquinez Bridge (E)	17,085	9,697	-7,388	-43.2%
I-680, Benicia/Martinez Bridge	19,444	37,973	18,529	95.3%
Cal-160, Antioch Bridge	NA	NA	NA	NA
Bay Area Bridges Sub-Total	***114,489***	***158,918***	***44,430***	**38.8%**
San Francisco / San Mateo Line				
US-101, Bayshore Freeway (N)	22,301	21,314	-987	-4.4%
Cal-35, Skyline Blvd. (N)	NA	NA	NA	NA
Cal-1, Junipero Serra Blvd. (N)	18,302	28,615	10,313	56.3%
I-280, Foran Freeway	32,897	25,819	-7,077	-21.5%
SFM/SM County Line Sub-Total	***73,500***	***75,748***	***2,249***	***3.1%***
San Mateo / Santa Clara County Line				
Cal-82, El Camino Real (N)	NA	NA	NA	NA
US-101, Bayshore Freeway (N)	20,496	27,355	6,859	33.5%
I-280, Serra Freeway (N)	13,905	11,620	-2,285	-16.4%
SM / SC County Line Sub-Total	***34,401***	***38,975***	***4,574***	***13.3%***
Santa Clara / Alameda County Line				
I-680, at Scott Creek Road (N)	11,871	17,727	5,856	49.3%
I-880, Nimitz Freeway (N)	17,640	34,234	16,594	94.1%
SC / Ala Line Sub-Total	***29,511***	***51,962***	***22,450***	***76.1%***
Alameda / Contra Costa County Line				
I-580, Knox Freeway	18,649	24,812	6,163	33.0%
I-80, Eastshore Freeway	32,003	47,924	15,920	49.7%
Cal-24, Caldecott Tunnel (E)	39,555	50,803	11,247	28.4%
I-680, in Dublin/San Ramon	35,358	46,118	10,761	30.4%
Ala / CC County Line Sub-Total	***125,565***	***169,657***	***44,092***	***35.1%***
Solano / Napa County Line				
Route 29, nodatapa-Vallejo Highway (N)	NA	NA	NA	NA
Solano / Sonoma County Line				
Route 37, Sears Point Road	8,935	9,350	415	4.6%
Napa / Sonoma County Line				
Route 121, Carneros Highway (N)	NA	NA	NA	NA
Route 128, Calistoga-Healdsburd Rd. (E)	NA	NA	NA	NA
Sonoma / Marin County Line				
US-101, Redwood Highway (N)	12,631	28,525	15,893	125.8%
Screenline Totals	**399,032**	**533,134**	**134,102**	**-30.7%**

Table 4.2: Screen Line Validation for AM Peak

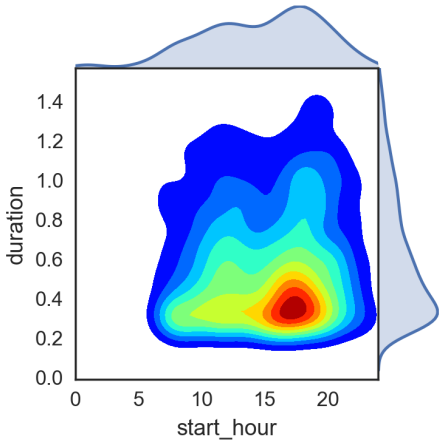
Screenline Facility	Observed	Predicted	Predicted Less Obsv'd	Percent Difference
Bay Area Bridges				
US-101, Golden Gate Bridge (S)	12,210	12,594	384	3.1%
I-80, SF/Oakland Bay Bridge	59,934	70,728	10,794	18.0%
Cal-92, San Mateo/Hayward Bridge (W)	10,441	10,885	444	4.3%
Cal-84, Dumbarton Bridge (N)	14,155	21,950	7,794	55.1%
I-580, Richmond/San Rafael Bridge (E)	NA	NA	NA	NA
I-80, Carquinez Bridge (E)	10,698	16,472	5,774	54.0%
I-680, Benicia/Martinez Bridge	34,388	28,236	-6,151	-17.9%
Cal-160, Antioch Bridge	NA	NA	NA	NA
Bay Area Bridges Sub-Total	***141,825***	***160,865***	***19,040***	**13.4%**
San Francisco / San Mateo Line				
US-101, Bayshore Freeway (N)	22,113	23,728	1,615	7.3%
Cal-35, Skyline Blvd. (N)	NA	NA	NA	NA
Cal-1, Junipero Serra Blvd. (N)	23,105	23,776	672	2.9%
I-280, Foran Freeway	42,084	28,957	-13,127	-31.2%
SFM/SM County Line Sub-Total	***87,302***	***76,462***	***-10,840***	***-12.4%***
San Mateo / Santa Clara County Line				
Cal-82, El Camino Real (N)	NA	NA	NA	NA
US-101, Bayshore Freeway (N)	23,436	23,296	-140	-0.6%
I-280, Serra Freeway (N)	18,934	11,032	-7,902	-41.7%
SM / SC County Line Sub-Total	***42,370***	***34,328***	***-8,042***	***-19.0%***
Santa Clara / Alameda County Line				
I-680, at Scott Creek Road (N)	10,970	16,467	5,497	50.1%
I-880, Nimitz Freeway (N)	16,312	31,972	15,660	96.0%
SC / Ala Line Sub-Total	***27,282***	***48,439***	***21,157***	***77.6%***
Alameda / Contra Costa County Line				
I-580, Knox Freeway	23,036	23,832	796	3.5%
I-80, Eastshore Freeway	39,495	49,929	10,434	26.4%
Cal-24, Caldecott Tunnel (E)	44,145	52,186	8,041	18.2%
I-680, in Dublin/San Ramon	43,924	50,891	6,966	15.9%
Ala / CC County Line Sub-Total	***150,601***	***176,838***	***26,238***	***17.4%***
Solano / Napa County Line				
Route 29, nodatapa-Vallejo Highway (N)	NA	NA	NA	NA
Solano / Sonoma County Line				
Route 37, Sears Point Road	11,694	5,937	-5,757	-49.2%
Napa / Sonoma County Line				
Route 121, Carneros Highway (N)	NA	NA	NA	NA
Route 128, Calistoga-Healdsburd Rd. (E)	NA	NA	NA	NA
Sonoma / Marin County Line				
US-101, Redwood Highway (N)	20,608	20,312	-296	-1.4%
Screenline Totals	**481,682**	**523,181**	**41,499**	**-30.7%**

Table 4.3: Screen Line Validation for PM Peak

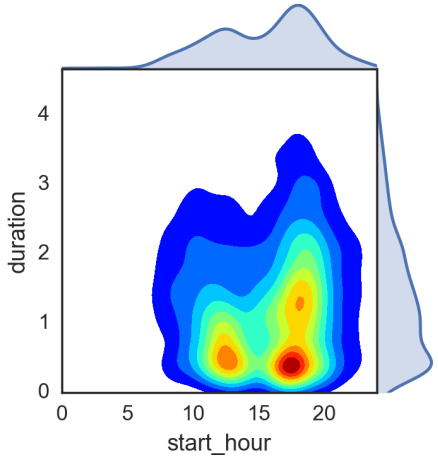
MATSim produces many useful images for calibration and validation based on sensor counts data. A histogram of observed and simulated counts is produced for every sensor location in the network. An example from the north bound direction of I-280, Figure 4.7, reveals that the current simulation does a good job of capturing peak flows, but underestimates mid-day volumes. Figure 4.8 shows that for the morning commute, most estimated flows fall within within one multiple of the observed values. MATSim can also be used to produce interactive visualizations of regional traffic flow and activities, such as the one shown on the front cover of this report.



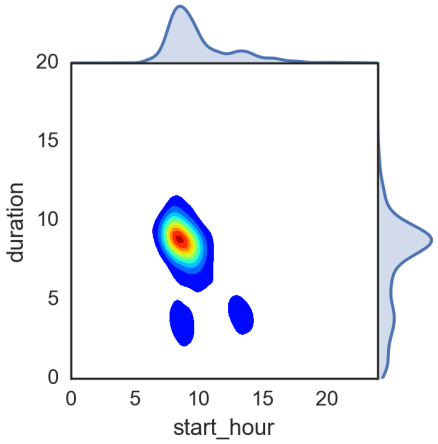
(a) Medium distance



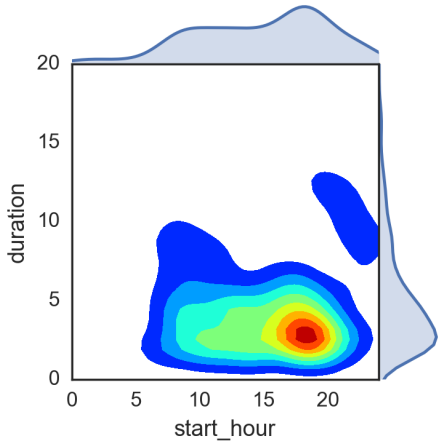
(b) Shop



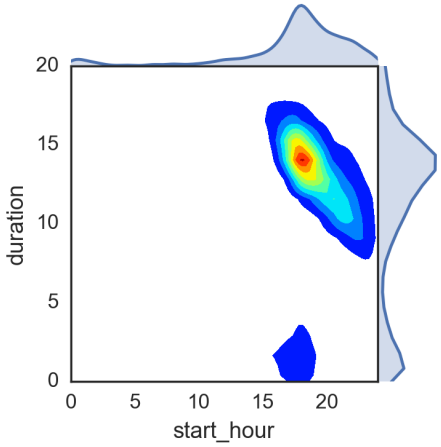
(c) Regular food



(d) Work



(e) Leisure



(f) Home

Figure 4.3: Temporal profile of inferred activities

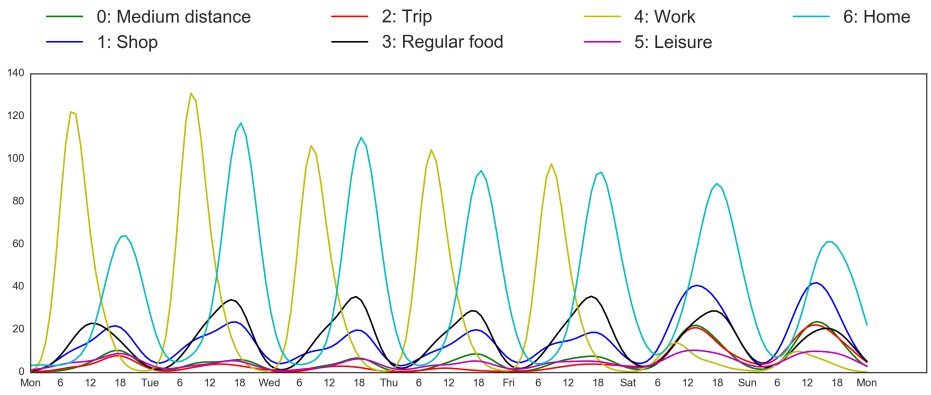
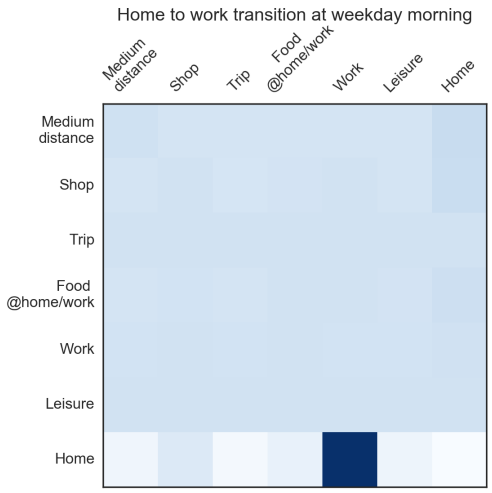
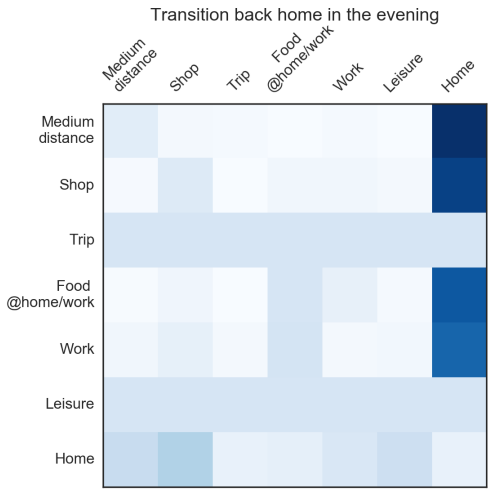


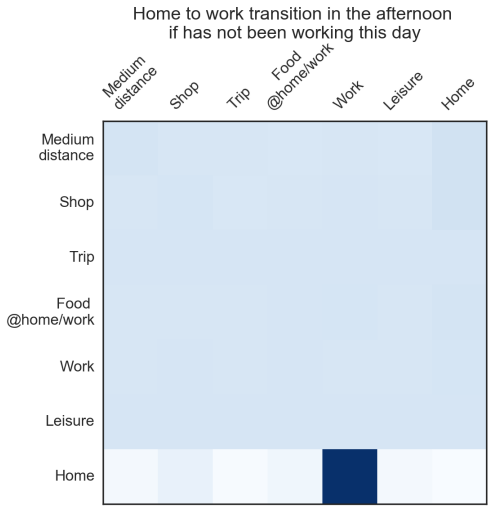
Figure 4.4: Start Time by Week



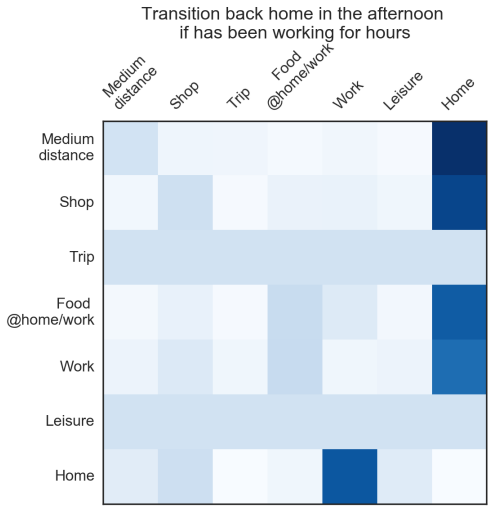
(a) Morning (6-10am)



(b) Night (5pm-midnight)



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



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

Figure 4.5: Heterogeneous activity transition probabilities

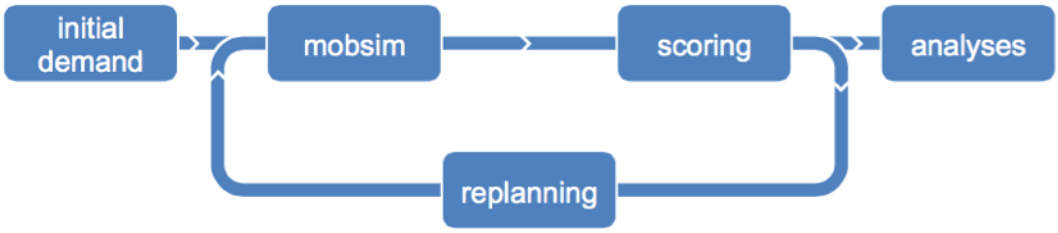


Figure 4.6: The MATSim Cycle [2]

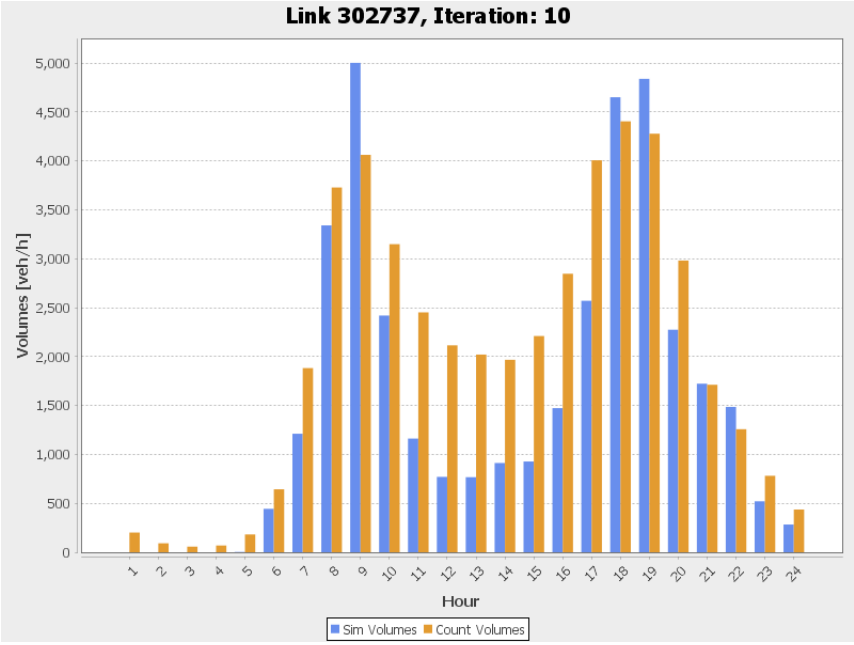


Figure 4.7: Observed (orange) vs Simulated (blue) Counts along the Dumbarton Bridge

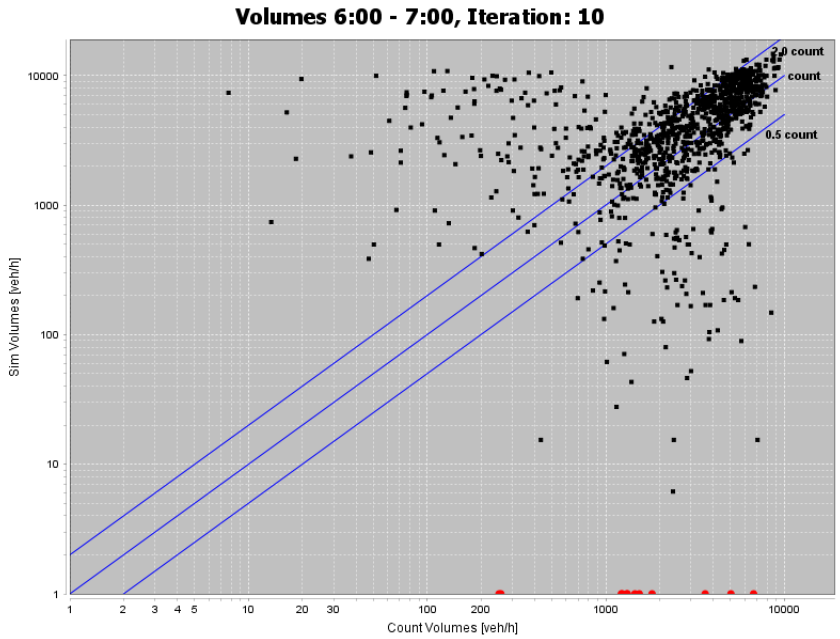


Figure 4.8: Observed vs Simulatd Counts for All Links

Chapter 5

Conclusion and Recommendations

In this project, we developed data processing methods to infer the types, locations, and durations of primary and secondary activities of travellers from anonymized cellular data. The method is based on an unsupervised, generative state-space machine learning model known as Input-Output Hidden Markov Model. The recovered activity patterns reveal the spatial and temporal profile of activities, as well as the heterogeneous transition probabilities given different context information. The generative nature of the model gives us directly producing travel itineraries of the population, i.e. providing an activity based travel demand model for the rewgion. We have also made a large step towards using automatically and continuously collected cellular data as an aid to traditional manual surveys, therefore reducing time and costs of the practice of travel demand modelling.

The presented approach consists of three simple steps. We first extracted stay point locations from raw traces that contain noise in the form of positioning error and oscillations between cell towers. We justified this algorithm by showing that the number of stay locations inferred from it are consistent with the number of reported trips in the California Household Travel Survey [1]. Second, we infered the primary activity locations, i.e. home and work. Third, we applied the IO-HMM model to further infer secondary activities and model the heterogeneous transition patterns and associated travel.

We then demonstrated the validity of the approach by building a travel micro-simulation based on the produced travel demand in the San Francisco

Bay Area, and observed a reasonable fit to the observed travel volumes.

We realize that there are certain limitation that require further work. We need further work to find out privacy-preserving ways of generating detailed travel itineraries for travellers. In terms of model specification and further accuracy improvement, we showed that even a few simple context variable can capture the spatial and temporal profile of each activity, and the heterogeneous activity transitions. However, more input or output variables, such as the weekly or daily periodicity of activities, might give us more information about the activity type. Finally, since our model is unsupervised, no direct ground truth is available. However, there are a variety of indirect sources that we can use for evaluation.

In general, the project has achieved its goal in demonstrating the feasibility of building activity-based travel models from cellular data, at a fraction of the time and cost as compared to the traditional manual surveying framework.

Appendices

Appendix A

IO-HMM Output Coefficients

A generative model for secondary activities calibrated for a sample set of users used linear models as the output models for (1) distance to home, (2) distance to work, and (3) duration of the activities. Since we did not specify inputs for (1) and (2), only the intercepts (constants) were fitted (one column for each output), which depend on the hidden activity. For duration, we specified that it does not only depend on activity type, but also an input variable "hours worked" and and indicator variable "is weekend". There are three coefficients (per hidden state) estimated for this output. The table below provides the values of the estimated parameters.

Table A.1: Model coefficients for the output variables

	Dist to home	Dist to work	Duration		
			constant	hours worked	weekend
Medium Distance	10.44	10.46	1.40	0.03	0.18
Shopping	2.17	3.41	0.62	0.00	0.01
Trip	127.39	126.07	3.32	0.59	0.89
Regular food	2.92	1.16	1.29	0.00	0.33
Work	3.46	0.00	7.53	-1.00	-0.21
Leisure	2.68	3.21	5.93	-0.14	-0.70
Home	0.00	3.46	11.57	0.12	5.47

References

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

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

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

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

[5] Nurhan Cetin, Adrian Burri, and Kai Nagel. In: *IN PROCEEDINGS OF SWISS TRANSPORT RESEARCH CONFERENCE (STRC), MONTE VERITA, CH*. 2003, pp. 3–4272.

[6] 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.

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

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

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

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

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

[12] 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.

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

[14] João Bártolo Gomes, Clifton Phua, and Shonali Krishnaswamy. "Where will you go? mobile data mining for next place prediction". In: *Data Warehousing and Knowledge Discovery*. Springer, 2013, pp. 146–158.

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

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

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

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

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

[20] 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.

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

[22] 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.

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

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

[25] 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.

[26] MTC. *Travel Model Development: Comparison to Legacy Model*. May 2012. URL: http://mtcgis.mtc.ca.gov/foswiki/pub/Main/Documents/2011_10_20_RELEASE_Comparison_to_Trip_Based_Model.pdf.

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

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

[29] Hesham Rakha et al. "Systematic verification, validation and calibration of traffic simulation models". In: *75th Annual Meeting of the Transportation Research Board, Washington, DC*. Citeseer, 1996. URL: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.1387&rep=rep1&type=pdf` (visited on 10/27/2015).

[30] Ingmar Visser, Maarten Speekenbrink, et al. "depmixS4: An R-package for hidden Markov models". In: *Journal of Statistical Software* 36.7 (2010), pp. 1–21.

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

[32] 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.

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

[34] 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.

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