Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

# Dynamic Origin-Destination Demand Estimation From Link Counts, Cellular Data And Travel Time Data

Sudatta Mohanty[a], Alexey Pozdnukhov[a]*

[a] *University of California, Berkeley, CA, 94720*

## Abstract

Accurate real-time Origin-Destination (O-D) demand estimation is critical for development of Advanced Traveller Information Systems (ATIS) as well as Advanced Traffic Management Systems (ATMS). Traditional O-D estimation techniques typically pose it as optimization problems involving offline calibration of O-D priors from historical data and updating these prior based on real-time data sources such as link counts, link travel times and Global Positioning System (GPS) probe data. However, two major drawbacks have emerged from this approach. Firstly, the generation of effective priors requires extensive population data with high temporal sampling frequency. Such data is generally hard to obtain and process with low memory and compute power. Secondly, the optimization problems proposed are often non-convex in nature. Therefore, convergence to an optimal solution is not guaranteed. Moreover, the algorithms suggested don't scale well with increase in the size of the transportation network. In order to tackle these challenges, we first generate priors through a convex optimization framework by Wu et al. (2015) for O-D estimation in quasi-static settings from cell phone Call Detail Records (CDRs). We extend this approach to a completely dynamic setting by utilizing real-time link counts and link travel times through an Entropy Maximization framework proposed by Janson et al. (1992). Approximate solution to the traffic assignment sub-problem is also achieved from Wu et al. (2015). We test this procedure for estimating commute demand for a simplified freeway network representing nine counties of the San Francisco Bay Area. We estimate demands for 2,916 O-D pairs for three hours at five-minute intervals, making it a 104,976 dimensional problem. We analyze the spatio-temporal distribution of errors, the effects of wider cell-phone coverage and updating estimates over time. Results indicate that an increase in cell-phone coverage from 0% to 100% leads to a reduction in average Root Mean Squared Error (RMSE) from 15.7 to 7.2 while updating estimates also leads to reductions in RMSE when O-D demand is high.

*Keywords:* Entropy Maximization; Cellular Data; Traffic Assignment

* Corresponding author. Tel.: +1-510-717-5461.
*E-mail address:* sudatta.mohanty@berkeley.edu

## 1. Introduction

Accurate knowledge of travel demand is essential for most transportation planning as well as traffic control operations. Origin-Destination (O-D) demand may contain clues towards critical high-level information such as peak hour times, spatial distribution of activities, likely congestion patterns among several other traffic characteristics. As a result, O-D demand estimation has received considerable attention over the last 30 years. Reviews of various techniques that have been employed for O-D matrix estimation can be found in Abrahamsson (1998), Bera and Rao (2011), F. M. O. Neto et al. (2016) and A. R. P. Neto, F. M. O. Neto, and Loureiro (2017). Two broad categories of O-D demand estimation are *static*, where the estimates for the entire time frame are derived offline with no updates due to incoming data (Zhang, Osorio, and Flötteröd, 2017), and *dynamic*, where the estimates are updated with incoming data (Okutani and Stephanedes, 1984; Ashok and Ben-Akiva, 2000). In this paper, we propose a technique to dynamically update O-D estimates with incoming data and also demonstrate how the error in estimates reduces as updates are made.

The key challenge for O-D estimation is that the problem is highly *underspecified* (i.e. the number of variables far exceeds the number of constraints or equations available) (A. R. P. Neto, F. M. O. Neto, and Loureiro, 2017). As a result, solving for the O-D estimates closest to the true values typically involves minimizing metrics which capture deviation from certain observations on a transportation network as well as explain the traffic flow phenomena. The most commonly used observation is link level traffic counts. Seminal work by Cascetta and Nguyen (1988) proposed a general optimization problem formulation for such an approach which was later formalized by Yang, Lu, and Hao (2017) as follows:

$$\begin{aligned} Minimize_D \quad & F_1\left(D,\widehat{D}\right)+F_2\left(C,\hat{C}\right) \\ s.t. \quad & C = A(D) \\ & D \in \Omega \end{aligned} \tag{1}$$

where, $D$ and $\widehat{D}$ are the estimated and prior O-D demand matrices respectively,
$C$ and $\hat{C}$ are the true and estimated link flow vectors,
$A$ is an operator that relates the estimated O-D demand and true link counts, commonly through traffic assignment,
$F_1$ and $F_2$ are distance operators and
$\Omega$ is the feasible region for $D$.

Based on the formulation of the problem, O-D estimation can be *deterministic*, such as Entropy Maximization (Zuylen and Luis G. Willumsen, 1980; Janson and Southworth, 1992) or *probabilistic*. Common probabilistic models used in the past include Generalized Least Square (GLS) such as Cascetta (1984) and Yang, Lu, and Hao (2017), Maximum Likelihood estimation such as Spiess (1987) and Bayesian inference such as Maher(1983). The common criticism faced by purely deterministic models is that they don't account for the possibility of noise in observations (L. Willumsen, 1984) and that such a formulation assumes consistency between the equations developed even when the data is very sparse e (F. M. O. Neto et al., 2016). Another criticism is that of fixed route choice probability matrices which don't account for heterogeneity in users or change in traffic conditions during a trip. With probabilistic models, the criticism is the common assumption of O-D demand and traffic volumes following Poisson distribution, which is susceptible to *overdispersion* (i.e. variance exceeding the mean) (Hazelton, 2003). Some papers propose a Gaussian distribution assumption which lead to a very high number of covariance parameters, especially in dynamic settings (Pitombeira-Neto, Loureiro,and Carvalho, 2016). Common route choice models in this procedure involve Multinomial Logit (Ben-Akiva,Lerman, and Lerman, 1985). But since the choice set for individuals is often unknown, the *Independence of Irrelevant Alternatives* (IIA) assumption can get violated (Cheng and Long, 2007). In this study, we propose a hybrid formulation which contains an Entropy Maximization part and a probabilistic part and aim to mitigate the limitations of past models. The Entropy Maximization formulation is based on the assumption that in the absence of any constraints imposed due to daily activity schedules, individuals would like to spread their trips over space and time as much as possible in order to face minimum congestion impacts. In order to reduce noisy observations, we include observations from multiple data sources and also conduct sensitivity analysis with regards to noise in the observed data. We update the proposed model and route choice estimates dynamically at short time intervals in order to incorporate possibilities of new routes emerging over time as well as the possibility of change in route choice probabilities for an individual during a trip. Another modification is the use of Kullback–Leibler divergence (KL-divergence) to model the deviation from prior O-D estimates instead of the common probabilistic methods such as GLS, Maximum Likelihood and Bayesian Inference. This is because GLS, similar to $L_2$ norm based loss functions, is well-known to be heavily influenced by outliers whereas Maximum Likelihood and Bayesian Inference force a predefined probability distribution on the O-D flow or traffic volumes, which may not be realistic. K-L divergence also represents the *relative entropy* or the *information gain* between the prior estimate and the current estimate. This gives more intuitive meaning to the formulation.

Finally, we discuss the various data sources used to perform O-D estimation. Link level traffic counts and speed estimates can be obtained from sensors buried under the road. However, such data is usually very sparse and typically only available on freeways. Therefore, recent studies for O-D estimation use several other data sources such as split ratios at intersections (Veeraraghavan, Masoud, and Papanikolopoulos, 2003), vehicle plate scanning (Castillo, Menendez, and Jimenez, 2008; Castillo, Jimenez, et al., 2013), probe trajectories (Yang, Lu, and Hao, 2017). In this study, we depend on three main data sources for O-D estimation - (i) cell-phone *Call Detail Records* (CDRs) (Becker et al., 2011) (ii) real-time travel time Application Program Interfaces (APIs) (Wang and Xu, 2011) and (iii) link counts (Zuylen and Luis G.Willumsen, 1980; Janson and Southworth, 1992). We build on the Entropy Maximization framework developed by Janson and Southworth (1992) for estimating static O-D matrices from link count data. We extend the formulation to estimate O-Ds dynamically along with a prior O-D estimates in order to ensure that a “fall-back” exists in case sufficient data is unavailable. In past studies, travel surveys were used to determine prior estimates of O-D matrices. This data collection was often expensive and had low sampling frequency. Instead, we use cellular data to obtain prior O-D estimates which are updated at very high frequency (in the order of a few minutes). Cellular data can be partitioned into cells based on the nearest cell towers (Wu et al., 2015; Baert and Seme, 2004; Yin et al., 2017). CDRs contain information about cells through which each data transmission is performed and their corresponding times. As a result, by tracking CDR data while an individual is in motion, we can obtain a proxy for the possible route taken up to the granularity of the nearest cell tower. This data has been used to determine route flows in Wu et al. (2015) as well as complete activity chains in Yin et al. (2017). We utilize CDR data as well as real-time travel time APIs in order to develop the route choice probability matrix which may be updated as more data becomes available. We develop a framework which is scalable and guarantees convergence by ensuring that the optimization problems solved are always convex in nature. We demonstrate the scalability by solving 104,976 dimensional problem at five-minute intervals and show the accuracy gain due to the dynamic updating of the model.

In the following sections, we first provide a mathematical overview of the problem. Next, we describe the static O-D estimation problem posed by Janson and Southworth (1992) and how the proposed approach overcomes some of its shortcomings. We provide the algorithm employed with route flow estimation sub-routine (Wu et al., 2015). We test the framework for estimating home-work and work-home O-Ds for a simplified freeway network representing the San Francisco Bay Area. Analysis is performed regarding the spatio-temporal distribution of errors, the effects of wider cell-phone coverage and updating estimates over time.

## 2. Problem Overview

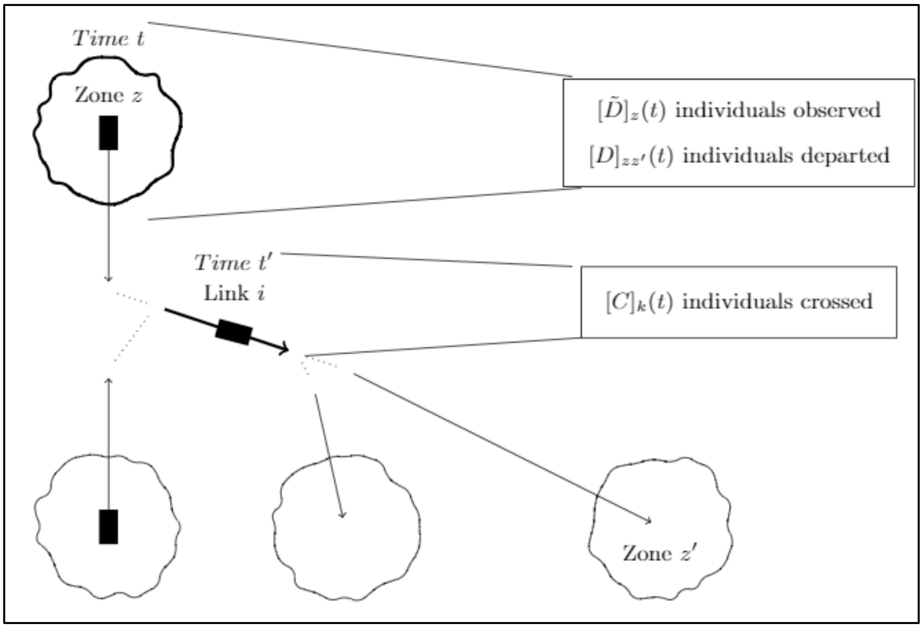
Time $t$
Zone $z$
$[\tilde{D}]_z(t)$ individuals observed
$[D]_{zz'}(t)$ individuals departed
Time $t'$
Link $i$
$[C]_k(t)$ individuals crossed
Zone $z'$

Figure 1: Problem Overview For Dynamic Origin-Destination Demand Estimation - The goal is to estimate the number of individuals departing zone *z* and travelling towards zone *z'* during time interval *t*, defined as $[D]_{zz'}(t)$

Assume that a road network graph $G = (\mathcal{V}, \mathcal{A})$ is divided in *Z* zones such that each each $i^{th}$ street link lies exactly within its (single) zone *z*. The objective here is to estimate the Origin-Destination (O-D) demand matrix in real-time. The O-D demand values during a time interval *t* are represented as entries in a $|Z|^2$ dimensional vector *D(t)* where the element labelled $[D]_{zz'(t)}$ represents the number of individuals leaving zone *z* and travelling towards zone *z'* during time interval *t*.

The following data sources are assumed to be available:

i. At the link level, complete information is assumed about the total number of individuals crossing link *i* during each time interval *t'*. These values are represented through elements of vector$[C]_i(t)$. This is an aggregation of individuals whose trips could have originated in any zone $\tilde{z}$ headed towards any zone $\widetilde{z'}$ and at any time interval $\widetilde{t'} \leq t'$. The dimensionality of such a vector is $|\mathcal{A}|$. Typically, this information is available from sensors installed inside the roads which detect the passage of any vehicle.

ii. Partial knowledge is assumed about the number of individuals in zone *z* at the end of each time interval *t*. These values are represented as elements of a $|Z|$ dimensional vector $[\tilde{D}]_z(t)$. This is a lower bound of the total number of individuals present in zone *z* at the end of time interval *t*. Typically, this information is available from aggregated cell phone *Call Detail Records* (CDRs) up to the granularity of the nearest cell phone tower.

iii. It is also assumed that perfect knowledge is available regarding link travel times, *TT(t)*, across the whole network at any instance. The dimensionality of such a vector is $|\mathcal{A}|$. Typically, this information can be extracted from travel time APIs for navigation apps.

## 3. OD Estimation Problem Formulation And Proposed Approach

### *3.1. Traditional Approach*

Janson and Southworth (1992) proposed an *Entropy-Maximization* approach for estimating the O-D matrix from link counts. The general hypothesis behind this approach is that in the absence of any constraints regarding desired activity schedules, the trip departure choice of individuals is likely to be “spread-out” over space and time, thus minimizing traffic congestion. Janson and Southworth (1992) claim that this hypothesis holds even in the presence of constraints imposed by desired activity schedules with a minor modification. In presence of such constraints, individuals are likely to choose departure times that are as "spread-out" as possible while ensuring that the desired activity schedules of all individuals are met. The state of departure time choices may be represented with the help of the entropy of the distribution of number of departures over space and time. As a result, the problem of estimating the number of departing individuals from any zone *z* towards any other zone *z'* can be expressed as the problem of maximizing this entropy of the system in the presence of constraints imposed by desired activity schedules. However, since activity schedules of individuals are not directly observable, they propose the use of a proxy which captures such behavior. The proxy chosen here is that of link counts on certain links in the network. These link counts combined with the structure of the road network provide a rough estimate of the times of the day and locations with high/low activity concentrations. The overall problem formulated by Janson and Southworth (1992) is as follows:

$$\begin{aligned} &Minimize_{\{D(t)\}_{t\in T}} \sum\nolimits_{t\in T} -H(D(t)) \\ &s.t. \quad \sum\nolimits_{\{t\in T\}} P(t',t)D(t) \quad = C(t), \forall \;\; t' \in T \end{aligned} \tag{2}$$

where, *H* represents the Entropy function defined by:

$$\mathrm{H(x)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \{\mathrm{x_i} * \log(\mathrm{x_i}) - \mathrm{x_i}\}, \quad \mathrm{n = card(x)} \tag{3}$$

$[\mathrm{D}]_{\mathrm{zz'}}(\mathrm{t})$ represents the number of departing individuals from zone *z* to zone *z'* during time interval *t*

$[\mathrm{P}]_{\mathrm{i,zz'}}(\mathrm{t'},\mathrm{t})$ represents the probability that a trip departing zone *z* to zone *z'* during time interval *t* uses link *i* during time interval *t'*.

In order to solve the optimization problem stated above, the entries of matrix *P(t',t)* must be first estimated. Janson and Southworth (1992) suggested solving a *Dynamic Traffic Assignment* (Merchant and Nemhauser,1978) as a sub-routine for estimating the link-choice probability matrix *P(t',t).* This involves approximately solving the a non-convex optimization problem described below to estimate the number of individuals departing zone *z* during time interval *t*, heading towards zone *z'* and arriving at link *i* at time interval *t'* represented as entries of matrix $[\mathcal{D}]_{\mathrm{i,zz'}}(\mathrm{t'},\mathrm{t})$. The algorithm proposed for solving DTA aims to find an equilibrium condition wherein no individual can change his/her path and reduce the time that he/she takes to reach the desired destination. The problem is formulated as:

$$\begin{aligned} &Minimize_{Dj(t',t)} \sum\nolimits_{(t',t)\epsilon TxT} \sum\nolimits_{j\,\epsilon\, Z\,x\,Z} \varphi(ADj(t',t)) \\ &\mathrm{s.t.} \sum\nolimits_{(t',t)\epsilon TxT} \sum\nolimits_{j\,\epsilon\, Z\,x\,Z} ADj(t',t) = C(t), \forall t' \in T \end{aligned} \tag{4}$$

where,

$$[A]_{i,zz'} = \begin{cases} 1, & if\ link\ i\ lies\ along\ shortest\ path\ (or\ least\ disutility\ path)\ between\ zone\ z\ and\ z', \\ 0, & otherwise \end{cases}$$

$Dj(t',t)$ represents the $\mathrm{j}^{\mathrm{th}}$ column of *D(t',t)*

$\varphi$ is a non-convex function dependent on the travel utility of individuals.

After estimating the entries of the matrix *D(t',t)*, the entries of the matrix *P(t',t)* may be estimated by calculating frequency ratios as follows:

$$[\mathrm{P}]_{\mathrm{i,zz'}}(\mathrm{t}',\mathrm{t}) = \frac{[\mathrm{D}(\mathrm{t}',\mathrm{t})]_{\mathrm{i,zz'}}}{\sum_{\mathrm{i}\in\mathcal{A}}[\mathrm{D}(\mathrm{t}',\mathrm{t})]_{\mathrm{i,zz'}}} \tag{5}$$

However, the traditional approach suffers from the following shortcomings:

i. Since the DTA problem is non-convex, the solutions are only guaranteed to be local minima. There is no approximation guarantee on the values of $[\mathrm{D}]_{\mathrm{i,zz'}}(\mathrm{t}',\mathrm{t})$ estimated by the DTA solver as compared to the corresponding optimal values $[\mathrm{D}^*]_{\mathrm{i,zz'}}(\mathrm{t}',\mathrm{t})$. As a result, the estimated values of $[\mathrm{P}]_{\mathrm{i,zz'}}(\mathrm{t}',\mathrm{t})$ may not be accurate.

ii. Moreover, there is also no approximation guarantee on the distribution of estimated values of $[\mathrm{D}]_{\mathrm{i,zz'}}(\mathrm{t}',\mathrm{t})$ over space and time. This may significantly affect quality of the estimated matrix *P(t',t)* and in turn may significantly affect the estimated O-D values *D(t)*.

iii. It assumes that the individuals maximize their utility based on complete information about the network. But it has since been shown that this condition doesn't hold in practice (Vlahogianni, Karlaftis, and Golias, 2014).

iv. The decisions taken by individuals cannot be modified once their trip has begun. However, in practice, this is not true. Due to the use of Advanced Travel Information Systems (ATIS), there is dynamic decision making regarding the route choice of individuals (Pillac et al., 2013). Also, location choice for secondary trips such as shopping might also be dynamic (Horni et al., 2009).

### *3.2. Proposed Approach*

The traditional problem formulation for static O-D demand estimation is extended with the help of real-time data sources, namely Call Detail Records (CDRs) and travel time APIs. The following extensions are proposed:

1. Adding partially observed departure constraints:

The total number of trips departing from zone *z* in time interval *t* is greater than or equal to the difference of the number of observed individuals in zone *z* at the end of time interval *(t-1)* and the number of observed individuals at the end of time interval *t*. This data is known from Call Data Records (CDRs). It may be represented by the following constraint:

$$[D]_{zz'}(t) \geq \max\{\big[\widetilde{\mathrm{D}}\big]_{\mathrm{z}}(\mathrm{t}-1) - \big[\widetilde{\mathrm{D}}\big]_{\mathrm{z}}(\mathrm{t}), 0\} \tag{6}$$

where, $\big[\widetilde{D}\big]_{z}(t)$ represents the number of observed individuals in zone *z* during time interval *t*

2. Estimation *P(t',t)* via data-driven approach:

A data-driven estimation of the value of *P(t',t)* is proposed via formulation of Quadratic Program (QP) and using network-wide link travel times. The approach is described as follows:

- QP Formulation:

First, it involves formulating a QP problem based on Wu et al. (2015). This optimization problem aims to estimate the number of individuals traveling along any given route *r* in the network during a specified time interval *t*. This may be represented as elements of a vector $\Gamma(\mathrm{t})$. Typically, the dimensionality of this matrix might be very large. However, Wu et al. (2015) showed through experiments in downtown Los Angeles that nearly 95% of all trips were covered if only the top 50 routes between each O-D pair were considered. Such a distribution of route choices between each O-D pair is hypothesized to be true in most large cities. Therefore, we may only consider a constant *M* number of trips between each O-D pair. Thus, the dimensionality of vector $\Gamma(\mathrm{t})$ reduces to$\mathrm{M}|\mathrm{Z}|^2$. The proposed approach assumes the availability of individual-level data in a form of a cellpath, *c*, defined as a time-stamped sequence of discrete regions within which a user can be located during a trip. It is a common format of mobility data available from cellular network carriers and IT service providers. For this approach, we assume that the area covered by the road network is divided in *C* such cell paths. The problem formulation is as follows:

$$\begin{aligned} &Minimize_{\Gamma(t)} && \frac{1}{2}\left|\left|\tilde{A}\Gamma(t) - C(t)\right|\right|_2^2 \\ &\mathrm{s.t.} && \mathrm{U}\Gamma(\mathrm{t}) = \mathrm{F}(\mathrm{t}) \\ &&& \Gamma(\mathrm{t}) \geq 0 \end{aligned} \tag{7}$$

where,

$$[\tilde{A}]_{i,r} = \begin{cases} 1, & if\ link\ i\ lies\ along\ route\ r \\ 0, & otherwise \end{cases}$$

$$[\mathrm{U}]_{\mathrm{c,r}} = \begin{cases} 1, & if\ cellpath\ c\ is\ covered\ by\ route\ r \\ 0, & otherwise \end{cases}$$

$[\mathrm{F}]_{\mathrm{c}}(\mathrm{t}) = \text{total number of observed individuals along cellpath c during time interval t}$

- Reduction To Least Squares:

The QP formulation can be further reduced to a least squares formulation. The constraint in equation (7) can be re-written as:

$$\begin{aligned} &\mathrm{U}\Gamma(\mathrm{t}) = \mathrm{F}(\mathrm{t}) \\ \Leftrightarrow &\sum_{\mathrm{r^*} \in \mathrm{R^c}} [\Gamma(\mathrm{t})]_{\widehat{\mathrm{r^*}}} = [\mathrm{F}]_{\mathrm{c}}(\mathrm{t}), \forall \mathrm{c} \in \mathrm{C} \end{aligned} \tag{8}$$

where,
$\mathrm{R^c}$ = subset of routes along cellpath *c*

Here, note that $\mathrm{R^c}$ is a disjoint set since each route has at most 1 cellpath associated with it. Suppose route $\widehat{\mathrm{r^*}}$ has associated cellpath $\hat{\mathrm{c}}$. Then, the following change of variables can be applied:

$$[\tilde{\Gamma}]_{\hat{\mathrm{c}}}(\mathrm{t}) = [\mathrm{F}]_{\hat{\mathrm{c}}}(\mathrm{t}) \tag{9}$$

Therefore, the problem can be reduced to least squares as follows:

$$Minimize_{\tilde{\Gamma}(t)} \frac{1}{2}\left\|\tilde{\tilde{A}}(t)\tilde{\Gamma}(t) - C(t)\right\|_2^2$$
$$\text{s.t. } \mathbf{1}^{\mathrm{T}}\tilde{\Gamma}(\mathrm{t}) = 1$$
$$\tilde{\Gamma}(\mathrm{t}) \geq 0 \qquad (10)$$

$$\left[\tilde{\tilde{\mathrm{A}}}\right]_{\mathrm{i,c}}(\mathrm{t}) = \begin{cases} [F]_c(t), if\ link\ i\ lies\ along\ any\ route\ r\ which\ covers\ cellpath\ c \\ 0,\ \ otherwise \end{cases}$$

Wu et al. (2015) then propose solving the least squares problem by eliminating equality constraints by a technique discussed in Boyd and Vandenberghe (2004), Section 4.2.4 and then applying accelerated gradient descent to find the optimal value $\Gamma^*(\mathrm{t})$.

- Estimating route choice probability matrix $\tilde{\mathrm{P}}(\mathrm{t})$:

It may be noted that each route *r* has at most one associated departing zone *z* and one associated arrival zone *z'*. Therefore, the solution to the least squares problem can be used to calculate the route choice probability matrix (i.e. probability of taking route *r* for an individual departing zone *z* and traveling to zone *z'*) as follows:

$$\left[\tilde{\mathrm{P}}\right]_{\mathrm{zz',r}} = \frac{[\Gamma^*(\mathrm{t})]_{\mathrm{r}}}{\sum_{\tilde{\mathrm{r}}' \in \mathrm{R}^{\mathrm{r}}} [\Gamma^*(\mathrm{t})]_{\tilde{\mathrm{r}}'}} \qquad (11)$$

where,
$\mathrm{R}^{\mathrm{r}}$ = subset of routes departing zone *z* and arriving at zone *z'*.

- Estimating link departure time incidence matrix $\hat{\mathrm{A}}(\mathrm{t}',\mathrm{t})$:

Real-time information about network-wide link travel time can be used to derive a matrix storing the probability of a trip departing along route *r* during time interval *t* to reach link *i* at time interval *t'*. Therefore, we have:

$$\left[\hat{\mathrm{A}}\right]_{\mathrm{r,i}}(\mathrm{t}',\mathrm{t}) = \begin{cases} 1,\ \ if\ link\ i\ lies\ along\ route\ r\ and\ travel\ time\ lies\ in\ the\ interval\ (t'-t) \\ 0,\ \ otherwise \end{cases} \qquad (12)$$

- Representing *P(t',t)* as a product of two estimated matrices:

The probability matrix *P(t',t)* can be estimated as a product of the two matrices estimated in the previous two steps as follows:

$$[P]_{\{i,zz'\}(t',t)} = \sum\nolimits_{r\,\in R} \tilde{P}_{zz',r}(t) * \hat{A}_{r,i}(t',t)$$
$$\Rightarrow \mathrm{P}(\mathrm{t}',\mathrm{t}) = \tilde{\mathrm{P}}(\mathrm{t}) * \hat{\mathrm{A}}(\mathrm{t}',\mathrm{t})$$

where, *R* = set of all routes that are analyzed in the network

3. Addition of prior OD estimates from Wu et al. (2015):

Wu et al. (2015) assume a quasi-static setting where flows are constant along each route and propose a convex optimization framework which makes use of cell phone CDRs in order to estimate route flow. While the quasi-static assumption may not always hold true for time-varying networks, it does provide a prior estimate for ODs which may be utilized as part of the Entropy Maximization framework. A KL-divergence term between the prior and estimated ODs is introduced in the objective function in order to penalize heavy discrepancy from prior estimates. A hyperparameter term ν is introduced to determine the relative weight given to the KL-divergence term. This term depends on the accuracy of the prior OD estimates which in turn depends on the validity of the quasi-static assumption. The value of this term for a given scenario is determined through cross-validation. The new objective function is as shown below:

$$Minimize_{\{D(t)\}} \sum\nolimits_t \nu D_{KL}\left(D(t), \widehat{D}(t)\right) - H\big(D(t)\big) \tag{13}$$

where, $\mathrm{D_{KL}(a, b)}$ = KL-divergence between the distributions of vectors a and b
$\widehat{\mathrm{D}}(\mathrm{t}) = \sum_{\widetilde{r'} \in R^r} [\Gamma^*(t)]_{\widetilde{r'}}$ (where $\Gamma^*(\mathrm{t})$ and $\mathrm{R^r}$ are as described in equation (10))
ν = hyperparameter for relative weight between the KL-divergence term and Entropy term

### *3.3. Overall Problem Formulation*

The overall optimization problem is formulated as follows:

$$Minimize_{D(t)} \sum\nolimits_{t \in \mathrm{T}} \nu D_{KL}\left(D(t), \widehat{D}(t)\right) - H\big(D(t)\big)$$
$$\mathrm{s.t.} \sum_{\mathrm{t \in T}} \mathrm{P(t', t)D(t) = C(t'), \forall t' \in T}$$
$$[D]_{zz'}(t) \geq max\big[\widetilde{D}\big]_z(t) - \big[\widetilde{D}\big]_z(t-1), 0$$

where, $\nu$, $\widehat{D}(t)$ and *P(t',t)* are estimated using the data-driven approach described above

### *3.4. Proposed Algorithm*

The overall algorithm proposed for dynamic OD estimation is as follows:

**Algorithm For Dynamic OD Estimation**

**Require:**

- Set of all OD pairs *zz'*
- Set of all relevant cellpaths *c*
- Set of all relevant routes *r*
- Road network graph *G*
- Link counts at uniform time intervals *C(t)*

```
At each time step $t$:
    1. For each route $r$, find corresponding cellpath $c$
    2. For each route $r$, find set of links $i$ along route and populate matrix $\tilde{\hat{A}}(t)$
    3. Solve LS problem (9) and apply change of variables from (8) to obtain route flow $\Gamma(t)$
    4. Estimate departure time link incidence matrix $\hat{A}(t', t)$ from equation (11)
    5. For each OD pair $zz'$:
        a. Estimate route choice probability $[\tilde{P}]_{zz',r}(t)$ using equation (10)
        b. Estimate link use probability $[P]_{i,zz'}(t', t)$ using equation (12)
        c. Obtain partially observed departures from CDR data $\sum_{z' \in Z}[D]_{zz'}(t)$
    6. Solve convex optimization problem in equation (14) to obtain OD estimate $D(t)$
Repeat 1-6 for $\hat{t} = t - h$ and update$D(\hat{t})$; if no significant change observed, then break.
```

## 4. Experimental Results

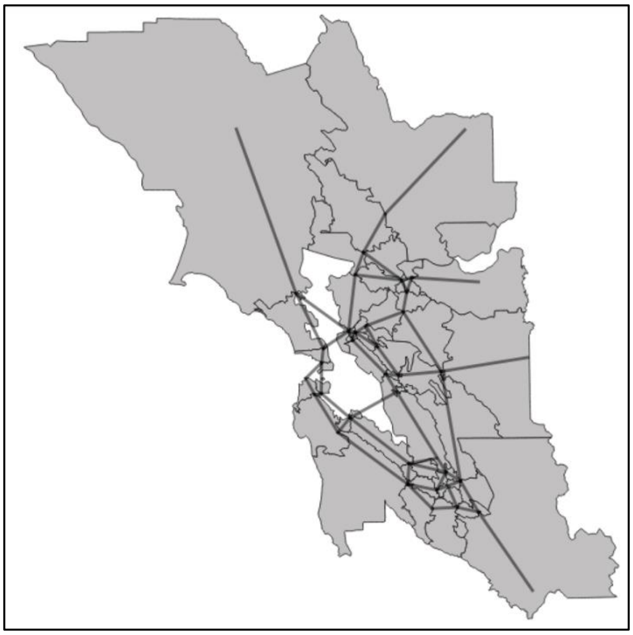

Figure 2: Test network for Proposed Dynamic O-D Estimation Algorithm

Experiments were conducted for the test network described in Figure (2) with simulated data generated for home-work and work-home trips. The activity initial plans were generated using the mean start time and duration of trips was 8.5 hrs and 1 hr respectively. Then, equilibrium demand patterns were determined by running simulations on agent based traffic simulation software *MATSim*. The ground truth O-D patterns were recorded and compared to O-D estimation using the proposed algorithm (see section 3.4) with updates made every 5 minutes. The accuracy of estimates was determined using RMSE values between ground truth O-Ds and estimated O-Ds. In order to formulate the partially observed departure constraints in simulated setting, it is first assumed that partially observations follow a uniform distribution in the range of $\left[0, \frac{CDR\ coverage}{100} * True\ departure\ count\right]$. Then, the value of partial departure for each constraint is determined by sampling from this distribution.

The following section describes the effect of variation in parameter $\nu$, the effect of quality of priors, the effect of partially observed departures through cell phone CDRs and the effect of updating predictions over time on the accuracy of OD estimates (Equation (14)).

### *4.1. Tuning Of Hyperparameter $\nu$ Based On Quality Of Priors*

The first step is to tune hyperparameter $\nu$ in Equation (14) based on the validity of assumptions made by Wu et al. (2015), which in turn affects the quality of priors, $\widehat{D}(t)$.

For this analysis, the CDR coverage is assumed to 100 percent. Next, $\nu$ is tuned based on average RMSE values of demand across all O-Ds but only during peak demand hours. Finally, the amount of potential noise in the prior estimate, $\widehat{D}(t)$, is varied by assuming a Gaussian noise term. Thus, the value of a noisy prior may be represented as:

$$\widehat{D_N}(t) = \left(\widehat{D}(t) + \widehat{D}(\mathrm{t}) \odot \epsilon\right)_+,\ \epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma),\ \Sigma = \sigma * I \qquad (15)$$

where, $\odot$ represents element-wise multiplication of vectors

Here, we also impose a constraint that the prior OD estimates cannot be negative. Now, the value of $\sigma$ is varied and the corresponding optimal $\nu$ is determined by minimizing the RMSE. The results are summarized in Figure (3).

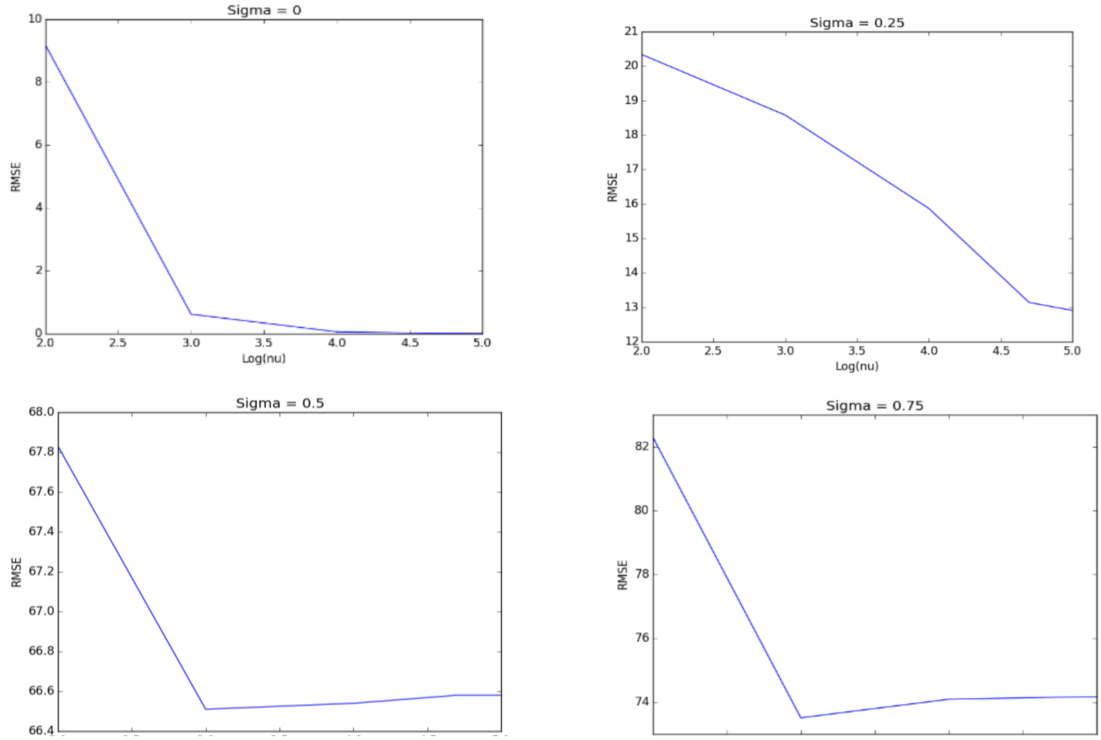

Figure 3: Variation of RMSE vs log($\nu$) for various values of $\sigma$, $\nu$ as defined in Equation (16)

As per intuition, with more noisy prior estimates (i.e. higher value of $\sigma$), the optimal value of $\nu$ for best estimates decreases since we have lower confidence on the priors. For $\sigma$ = 0 and $\sigma$ = 0.25, the RMSE is a decreasing function of $\nu$. This indicates that we don't require any further correction in prior estimates through link counts and partially observed departures. However, for $\sigma$ = 0.5 and $\sigma$ = 0.75, the optimal value for log($\nu$) is 3, which implies that the optimal value of $\nu$ is 1000.

In real-world scenarios, the quality of prior estimates may be first evaluated by determining the accuracy percentage based on previously observed demand during peak hours. In case the accuracy is high (error <= 25%), there is no need to re-evaluate ODs. In case the accuracy is low, the optimal value of parameter $\nu$ may be determined using plots similar to those in Figure (3).

### 4.2. Temporal Distribution Of Errors

The temporal distribution of RMSE across a day is plotted in Figure (4) for $\nu$ = 1000 and $\sigma$ = 0.5 for two scenarios:

- With partially observed departure constraints
- Without partially observed departure constraints

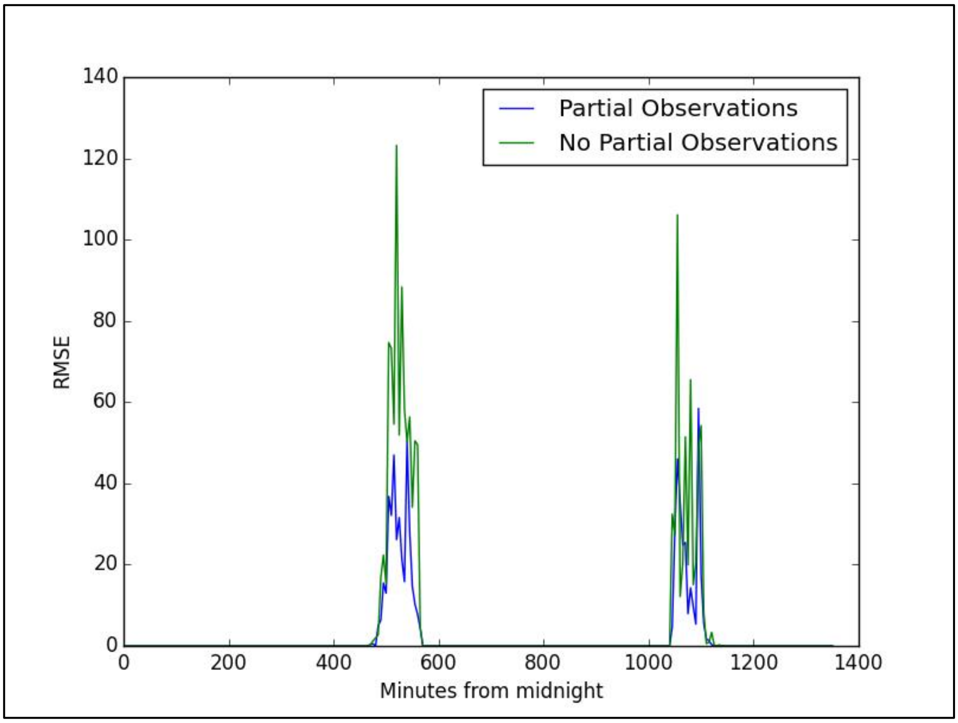

Figure 4: Variation of RMSE vs time of day for $\nu$ = 1000 and $\sigma$ = 0.5

The plot shows that addition of partially observed departure constraints helps in reducing the overall RMSE across all OD pairs during peak hours. Note that the feasible region for the case when partially observed departure constraints are added is smaller than the case when these constraints are not added. Therefore, the optimal objective value with partially observed departure constraints is at least as large as that without these constraints. However, the optimal value is closer to the ground truth. This is because ground truth observations are influenced by daily activity plans which may make the observed OD counts deviate from the entropy maximization estimate as well as the prior estimate.

### *4.3. Spatial Distribution Of Errors*

Next, the spatial distribution of RMSE across all OD pairs averaged over the course of the day is plotted in Figure (5) for $\nu$ = 1000 and $\sigma$ = 0.5 for two scenarios:

- With partially observed departure constraints
- Without partially observed departure constraints

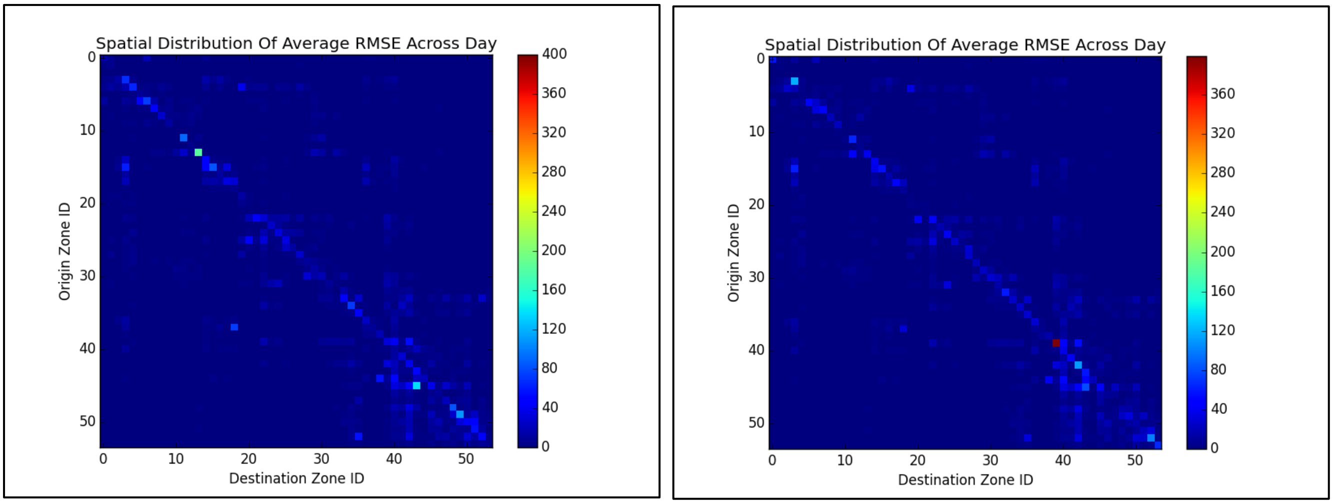

Figure 5: Variation of RMSE over OD pairs averaged over entire day for $\nu$ = 1000 and $\sigma$ = 0.5

From both figures, it can be noted that since most errors lie along the diagonal, trips which originate and terminate in the same zone are not predicted accurately in either framework. This might be because the starting position of all trips is assumed to be at the centroid of each zone. Therefore, the probability matrix, *P(t',t)*, does not incorporate the time required to reach the centroid of a particular zone as well as the possibility of trips never travelling through the origin centroid. These errors start to dissipate in trips involving separate origins and destinations since most of the travel time is spent on the freeway part of the trip and the inaccuracies during the trip start/end don't play as crucial a role. This problem may be rectified by updating the map to include a more detailed network. However, it must be noted that the number of constraints increases linearly with the number of links in the network where link counts are observed.

It may also be observed that both frameworks tend to have higher errors when estimating ODs for trips with origins and destinations in the 38-44. These zones represent Santa Clara county where trip lengths and commute times are typically shorter than average in the Bay Area because most individuals residing here work in the nearby Silicon Valley (see http://www.vitalsigns.mtc.ca.gov/). As a result, it is harder to judge the origin/destination probability of vehicles observed on each link.

Finally, it can be observed that there are more “light blue” regions in the heatmap for RMSE with partially observed departures. This indicates that errors are more evenly spread out across various zones when partially observed departure constraints are added. Without these constraints, the error is heavily concentrated on trips with origin and destination zone 39. This effect suggests that the addition of partially observed departure constraints prevent extremely high errors, but may lead to small errors across multiple zones. This is a desirable property since congestion patterns typically depend on whether aggregate demand is greater than aggregate capacity in a particular zone. Therefore, large errors in even a few demands may lead to higher variation in congestion estimates as compared to small errors in a larger number of demands.

### *4.4. Effect Of Updating Estimates Over Time*

In order to measure the effect of updating estimates of O-D as more information gets revealed over time, the change in average RMSE across all O-D estimates over time was studied. The value of $\nu$ and $\sigma$ are set as 1000 and 0.5 respectively and partially observed departure constraints are also included in the optimization framework. In Figure (6), the change in the RMSE for the O-D estimate across all zones at *t = 540 mins* and *t = 545 mins* (start of morning peak) is studied between *t = 540* mins and *t = 720 mins*. Updates for estimates are made every *35 mins*. As per intuition, the RMSE decreases with incoming information. The effect is more pronounced when the RMSE values are higher (as in the case of O-D estimates at *t=545 mins*).

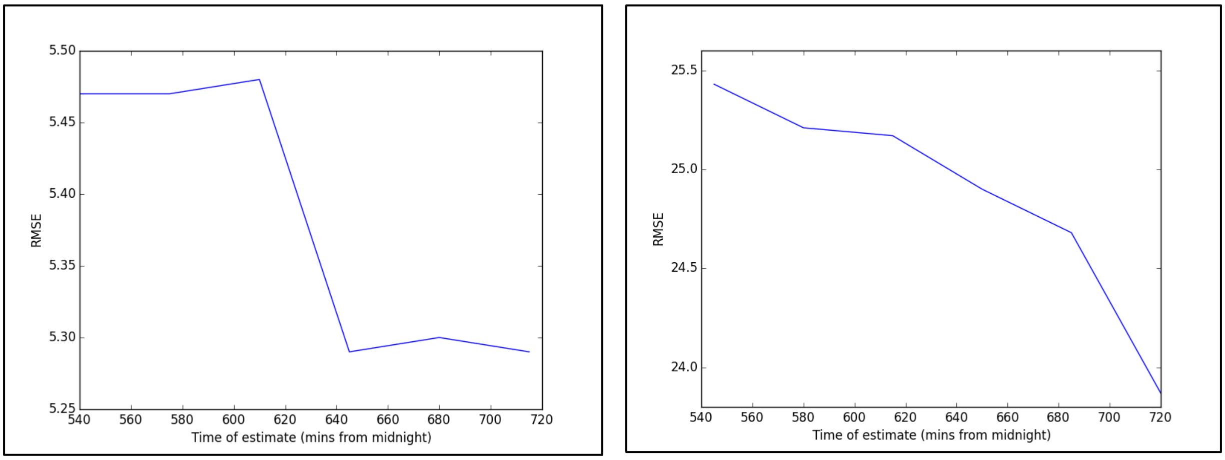

Figure 6: Change in the RMSE for the O-D estimate across all zones with incoming information for $\nu$ = 1000 and $\sigma$ = 0.5 –

(a) RMSE at *t = 540 mins* with updates between *t = 540 mins* and *t = 715 mins* every *35 mins*,

(b) RMSE at *t = 545 mins* with updates

### *4.5. Effect Of Cell Phone Data Coverage*

The next experiment is to test the effect of cell phone data coverage on the quality of OD estimates. To mimic the effect of various amounts of cell phone coverage, it is assumed that partially observations follow a uniform distribution in the range of $\left[0, \frac{CDR\ coverage}{100} * True\ departure\ count\right]$. The value of $\nu$ and $\sigma$ are set as 1000 and 0.5 respectively. The RMSE for OD estimates across all ODs is calculated for peak hours only. The variation of RMSE with cell phone coverage is displayed in Figure (7).

As per intuition, the addition of more accurate partially observed departure constraints significantly reduces overall RMSE. It is also acknowledged that better CDR coverage may also improve the quality of priors, thus further improving OD estimates. Therefore, the plot below gives a lower bound on the potential improvements due to higher cell phone CDR coverage.

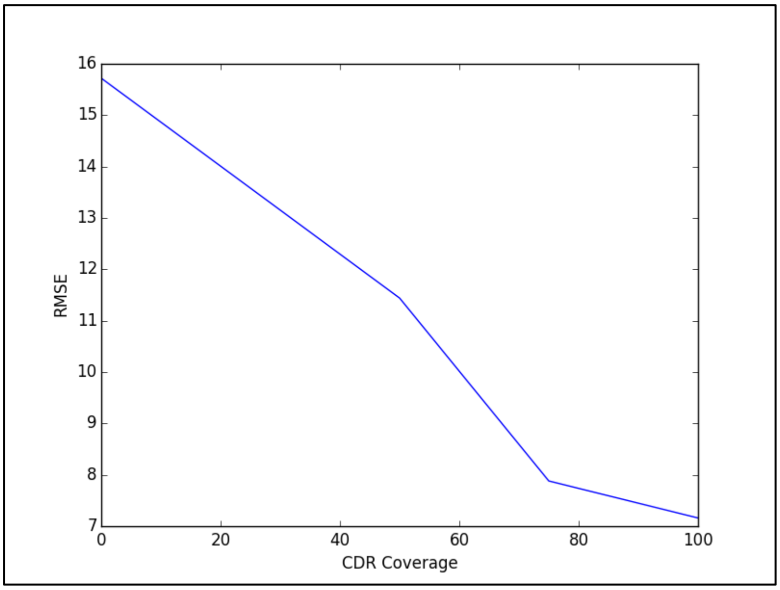

Figure 7: Change in the RMSE for the O-D estimate across all zones during peak hours with varying cell phone data coverage

### *4.6. Effect Of Noise In Link Choice Probability Matrix*

The final experiment is designed to test the impact of noise in link choice probability matrix which may arise because of inaccurate predictions from travel time APIs, invalidity of the quasi-static setting assumption made by Wu et al. (2015) or insufficient sample of trips used for estimating *P(t',t)*. In order to approximate the impact of noise in the data, we corrupt the estimated *P(t',t)* matrix using a Gaussian noise term. The noisy link choice probability matrix may be represented as:

$$P(t',t)_N(t) = (P(t',t) + P(t',t) \otimes \epsilon)_+,\ \epsilon \sim \mathcal{N}(\mathbf{0},\Sigma),\ \Sigma = \sigma * I$$

where, $\otimes$ represents Hadamard product of matrices

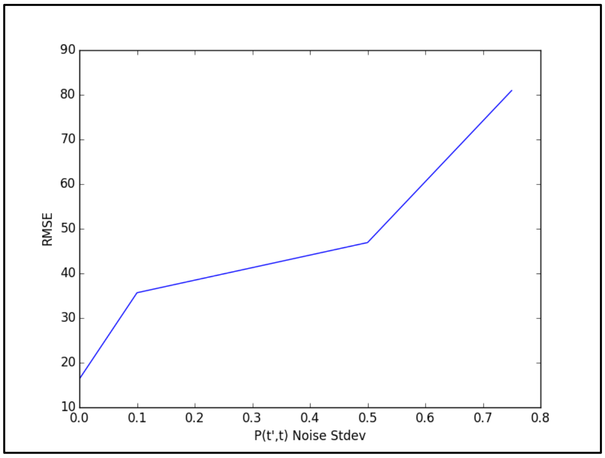

Figure 8: Change in the average RMSE for the O-D estimate across all zones during peak hours with varying noise in link choice probability matrix *P(t',t)*

Figure (8) displays the change in average RMSE across all O-D estimates during peak hours with the increase in $\sigma$. We can notice that the RMSE is quite sensitive to noise in the link choice probability matrix. However, with improvement in quality of the data provided through travel time APIs, we can expect to see large improvements in the quality of the O-D estimates. It is important to first analyze the quality of travel time data available and also the validity of the quasi-static assumption before making O-D estimates. Based on Figure (8), we can approximately judge the impact this is likely to have on the quality of estimated O-Ds.

## 5. Conclusion

This study provides a novel approach to the dynamic Origin-Destination (O-D) demand estimation problem through the incorporation of modern data sources such as cellular data and travel time APIs. The approach is aimed at solving two major limitations of existing techniques, namely the requirement of extensive high frequency population level data for generation of effective priors and the non-convex nature of the problem formulation which leads to lack of guarantees on the nature of the solutions achieved. Prior O-D estimates are derived from recent cell phone Call Detail Records (CDRs) through a convex optimization framework proposed by Wu et al. (2015). This approach has been shown to scale well with the size of the network and demands. Then, this framework is extended to a dynamic setting with the help of an Entropy Maximization approach proposed by Janson et al. (1992). Travel time APIs and CDRs are utilized for solving the traffic assignment problem and updating the route choice probability matrix dynamically. The overall framework thus involves only convex optimization problems which ensure convergence guarantees. Tests are conducted on an agent-based simulation of home-work and work-home trips on a simplified freeway network in the nine counties in the San Francisco Bay Area. The spatio-temporal distribution of errors reveals that errors are generally when there are a higher number of very short trips. This may be analyzed further in future work with modifications for modeling these trips better. Tests also show that updating estimates over time and having higher cell phone data coverage helps prediction accuracy. Finally, the impact of noise in the data obtained through travel time APIs is estimated. It is shown that the quality of estimates is sensitive to such noise which motivates research in developing better algorithms for travel time prediction.

## References

Wu, Cathy et al. (2015). “Cellpath: Fusion of cellular and traffic sensor data for route flow estimation via convex optimization”. In:Transportation Research Part C: Emerging Technologies.

Janson, Bruce N and Frank Southworth (1992). “Estimating departure times from traffic counts using dynamic assignment”. In:Transportation Research Part B: Methodological26.1, pp. 3–16.

Abrahamsson, Torgil (1998). “Estimation of origin-destination matrices using traffic counts-a literature survey”.

Bera, Sharminda and KV Rao (2011). “Estimation of origin-destination matrix from traffic counts: the state of the art”.

Neto, Francisco Moraes Oliveira et al. (2016). “Discussão conceitual sobre métodos de reconstruçãode matrizes origem-destino estáticas em redes de transportes”. In:Transportes24.1, pp. 107–119.

Neto, Anselmo Ramalho Pitombeira, Francisco Moraes Oliveira Neto, and Carlos Felipe GrangeiroLoureiro (2017). “Statistical models for the estimation of the origin-destination matrix from traffic counts”. In:TRANSPORTES25.4, pp. 1–13.

Zhang, Chao, Carolina Osorio, and Gunnar Flötteröd (2017). “Efficient calibration techniques for large-scale traffic simulators”. In:Transportation Research Part B: Methodological97, pp. 214–239.

Okutani, Iwao and Yorgos J Stephanedes (1984). “Dynamic prediction of traffic volume through Kalman filtering theory”. In:Transportation Research Part B: Methodological18.1, pp. 1–11.

Ashok, Kalidas and Moshe E Ben-Akiva (2000). “Alternative approaches for real-time estimation and prediction of time-dependent origin–destination flows”. In:Transportation Science34.1, pp. 21–36.

Cascetta, Ennio and Sang Nguyen (1988). “A unified framework for estimating or updating origin/destination matrices from traffic counts”. In:Transportation Research Part B: Methodological22.6,pp. 437–455.

Yang, Xianfeng, Yang Lu, and Wei Hao (2017). “Origin-Destination Estimation Using Probe Vehicle Trajectory and Link Counts”. In:Journal of Advanced Transportation2017.

Zuylen, Henk J. Van and Luis G. Willumsen (1980). “The most likely trip matrix estimated from traffic counts”. In:Transportation Research Part B: Methodological14.3, pp. 281–293.issn: 0191-2615.doi:http://dx.doi.org/10.1016/0191-2615(80)90008-9.url:http://www.sciencedirect.com/science/article/pii/0191261580900089.

Cascetta, Ennio (1984). “Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator”. In:Transportation Research Part B: Methodological18.4-5, pp. 289–299.

Spiess, Heinz (1987). “A maximum likelihood model for estimating origin-destination matrices”. In:Transportation Research Part B: Methodological21.5, pp. 395–412.

Maher, MJ (1983). “Inferences on trip matrices from observations on link volumes: a Bayesian statistical approach”. In:Transportation Research Part B: Methodological17.6, pp. 435–447.

Willumsen, LG (1984). “Estimating time-dependent trip matrices from traffic counts”. In:Ninth Inter-national Symposium on Transportation and Traffic Theory. VNU Science Press Utrecht, pp. 397–411.

Hazelton, Martin L (2003). “Some comments on origin–destination matrix estimation”. In:Transportation Research Part A: Policy and Practice37.10, pp. 811–822.

Pitombeira-Neto, Anselmo Ramalho, Carlos Felipe Grangeiro Loureiro, and Luis Eduardo Carvalho (2016). “Bayesian inference on dynamic linear models of day-to-day origin-destination flows in transportation networks”. In:arXiv preprint arXiv:1608.06682.

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

Cheng, Simon and J Scott Long (2007). “Testing for IIA in the multinomial logit model”. In:Sociological methods & research35.4, pp. 583–600.

Veeraraghavan, Harini, Osama Masoud, and Nikolaos P Papanikolopoulos (2003). “Computer vision algorithms for intersection monitoring”. In:IEEE Transactions on Intelligent Transportation Sys-tems4.2, pp. 78–89.

Castillo, Enrique, Jose Maria Menendez, and Pilar Jimenez (2008). “Trip matrix and path flow re-construction and estimation based on plate scanning and link observations”. In:Transportation Research Part B: Methodological42.5, pp. 455–481.

Castillo, Enrique, Pilar Jimenez, et al. (2013). “A Bayesian method for estimating traffic flows based on plate scanning”. In:Transportation 40.1, pp. 173–201.

Becker, Richard A et al. (2011). “Route classification using cellular handoff patterns”. In:Proceedings of the 13th international conference on Ubiquitous computing. ACM, pp. 123–132.

Wang, Fahui and Yanqing Xu (2011). “Estimating O–D travel time matrix by Google Maps API: implementation, advantages, and implications”. In:Annals of GIS17.4, pp. 199–209.

Baert, A-E and David Seme (2004). “Voronoi mobile cellular networks: topological properties”. In:Parallel and Distributed Computing, 2004. Third International Symposium on/Algorithms, Models and Tools for Parallel Computing on Heterogeneous Networks, 2004. Third International Workshopon. IEEE, pp. 29–35.

Yin, Mogeng et al. (2017). “A generative model of urban activities from cellular data”. In:IEEE Transactions on Intelligent Transportation Systems.

Merchant, Deepak K and George L Nemhauser (1978). “A model and an algorithm for the dynamic traffic assignment problems”. In:Transportation science12.3, pp. 183–199.

Vlahogianni, Eleni I, Matthew G Karlaftis, and John C Golias (2014). “Short-term traffic forecasting: Where we are and where we’re going”. In:Transportation Research Part C: Emerging Technologies43, pp. 3–19.

Pillac, Victor et al. (2013). “A review of dynamic vehicle routing problems”. In:European Journal of Operational Research225.1, pp. 1–11.

Horni, Andreas et al. (2009). “Location choice modeling for shopping and leisure activities with MAT-Sim: combining microsimulation and time geography”. In:Transportation Research Record: Journal of the Transportation Research Board2135, pp. 87–95.

Boyd, Stephen and Lieven Vandenberghe (2004). Convex optimization. Cambridge university press.