Link Density Inference from Cellular Infrastructure

Steve Yadlowsky
steve.yadlowsky@berkeley.edu
652 Sutardja Dai Hall
Berkeley, CA 94720
Corresponding author

Jérôme Thai
Electrical Engineering and Computer Sciences
jerome.thai@berkeley.edu
652 Sutardja Dai Hall
Berkeley, CA 94720

Cathy Wu
Electrical Engineering and Computer Sciences
cathywu@berkeley.edu
652 Sutardja Dai Hall
Berkeley, CA 94720

Alexey Pozdnukhov
Civil and Environmental Engineering
alexeip@berkeley.edu
115 McLaughlin Hall
Berkeley, CA 94720

Alexandre Bayen
Electical Engineering and Computer Sciences
Civil and Environmental Engineering
bayen@berkeley.edu
642 Sutardja Dai Hall
Berkeley, CA 94720

4991 words + 9 figures + 0 tables
January 7, 2015

ABSTRACT
This work explores the problem of estimating road link densities from cellular tower signals by mobile subscribers in urban areas. We pose the estimation problem as a quadratic program, and present a robust framework that produces vehicle density estimates and is suitable for large-scale problems. We demonstrate that both simple and sophisticated models of cellular network connections can be handled robustly by the framework, without sacrificing efficiency or scalability. We present a numerical experiment on the I-15 corridor in San Diego based on a calibrated Aimsun microsimulation and a simulated cell network, demonstrating the framework can practically be implemented as part of an integrated corridor management system. The numerical results demonstrate that when the cell phone connection model is chosen appropriately, the estimates are consistent with those observed in a microsimulation.

INTRODUCTION

Vehicle density estimation is a critical component of future traffic management in urban areas. Vehicle density on freeways and arterial roads alike can be inferred from instrumentation added to the road network, but installing and maintaining such sensors is time consuming and costly, (Fontaine and Smith *(1)*). Heavily used freeways are usually instrumented, but most arterial roads are not. Thus, estimates of traffic conditions can only be found on a small fraction of the entire road network in an urban area. The framework presented in this article uses data from cellular network infrastructure to estimate current link densities on the road network. This not intended for traffic information systems, due to the granularity of the data, however better estimates of counts of cars on arterial roads on a 10-to-15 minute interval would be a useful input to traffic management systems that could provide richer information for decision making. For example, it could be used as data for improving demand estimation and management, and would allow traffic managers to adjust signal timing schedules to account for changes in current conditions, since most signal timing strategies are adjusted on a similar time resolution, (Lee and Williams *(2)*).

Cell tower usage data has become an increasingly popular source of data for traffic demand estimation, as mobile phone network coverage is generally ubiquitous in urban areas, however one of the challenges of using cell phone tower data is the coarse granularity of the sensors, both spatially and temporally, (Cheng et al. *(3)*). For this reason, cellular infrastructure data is not as useful as GPS or Bluetooth for traffic information systems, (Herrera et al. *(4)* and Work et al. *(5)*). On the other hand, they are a pervasive source of data, where penetration rates in the population are exceptionally high compared to other data sources such as GPS or other wireless probes, as noted by Calabrese et al. *(6)*. This makes them valuable for traffic management applications, where congestion and demand information are useful for making more data-driven management decisions.

Studies of cell phone data have focused on numerous important areas of traffic modeling. Because of extensive coverage and appropriate level of accuracy and precision in cellular infrastructure data, applications to demand modeling have shown impressive results, allowing researchers to shift from census based models to more sophisticated models of data-driven origin-destination inference, allowing higher temporal resolution, as presented by Toole et al. *(7)*. In this article, we investigate the usage of this data to estimate current counts of vehicles on links in the road network, a more localized quantity of significant importance to traffic modeling. The map shown in Figure 1, created from the numerical work presented later in the article, shows how solving this problem is similar to projecting cell tower connection density onto the road network.

In this article, we present a convex optimization framework for estimating link flows from cell tower usage data. Our model is based on two main ideas: connections to cell towers can be modeled by a probability distribution that gives the probability of being on each link given that one is connected to a given cell tower, and a link similarity model can encode the relationship in density between links in the road network. The framework presented poses the problem as a quadratic program that can be solved in $O(n^3)$ time, where n is the number of cell towers in the region, which is usually less than 1000. Numerical results establish consistency with microsimulation results, and discuss how this framework can be extended to further improve results.

RELATED WORK

Location data from mobile phones has been embraced by the urban planning community as a powerful and pervasive source of spatiotemporal data about urban communities, (Ratti et al. *(8)*).

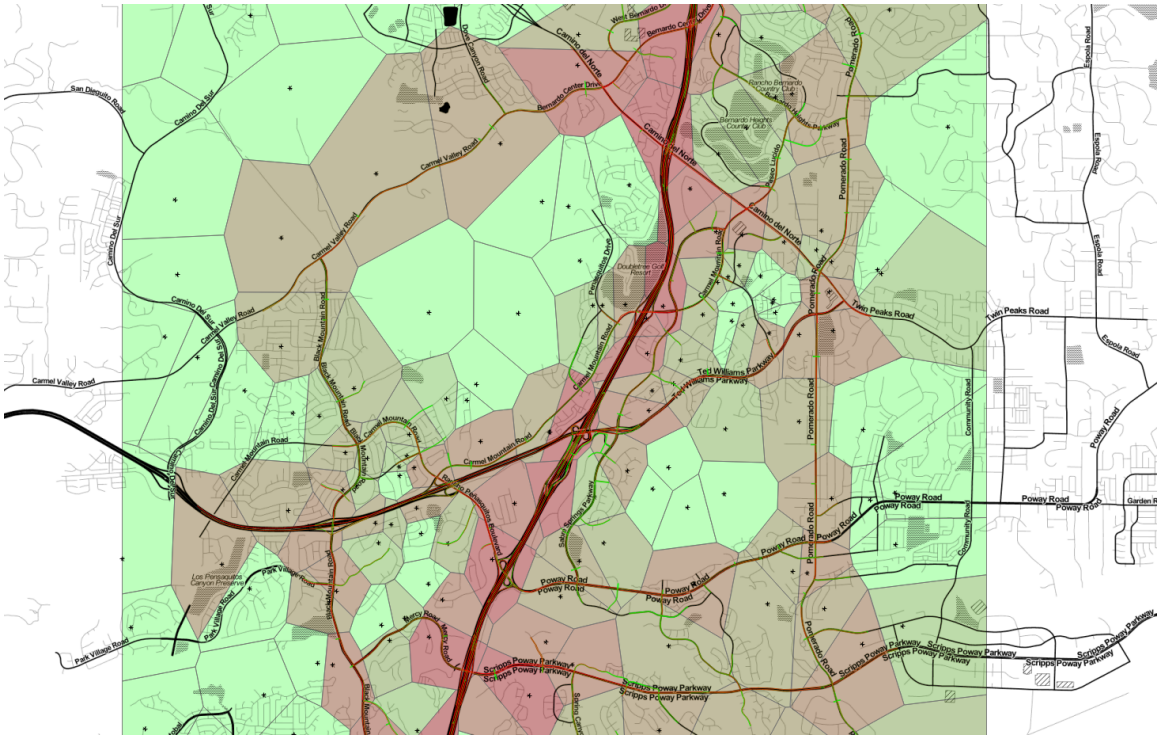


FIGURE 1 : Example of density projection from the data used for the present study. Regions (resp. links) colored in red have high cell tower connection (resp. vehicular) density. Best viewed in color. A time-lapse video is available on `http://connected-corridors.berkeley.edu/gallery/videos`

Estimating traffic demand from cell phone data has become a popular area of research in the transportation community in recent years, (Wang et al. *(9)*, Yuan and Raubal *(10)* and Steenbruggen et al. *(11)*). As noted before, cell phone data presents very different challenges from traditional sensors for traffic engineering, such as loop detectors and cameras, (Calabrese et al. *(12)*). There are various forms of cell phone data studied in the literature that all have specificities and challenges, (Rose *(13)*, Wang et al. *(9)* and Iqbal et al. *(14)*). For instance, location data acquired from triangulation has been used to get a fine position estimate of the phone. Jiang et al. *(15)* used this to model density and track users throughout an urban network.

Call detail records (CDR), indicating which tower a call, text, or data transmission came through, is another data source studied from cell tower infrastructure. While it is not as fine grained, it is a commonly logged data source among cell phone network operators, and has good penetration rates. Moreover, it is a good compromise between user privacy and usefulness. One of the most common applications of CDR data is origin-destination matrix estimation, (Iqbal et al. *(14)*). Wang et al. *(9)* demonstrates a method of using time-dependent origin-destination information collected from cell phone data to estimate link flows through the traffic network using a user equilibrium model.

Pozdnoukhov and Kaiser *(16)* demonstrate that continuous space models via area-to-point kriging can be used to project CDR data onto an urban area. Our work is an extension to this idea, where the probability distribution is constrained to the road network as opposed to the entire urban area, allowing us to focus on the nuances of a road network to get better estimates of vehicle

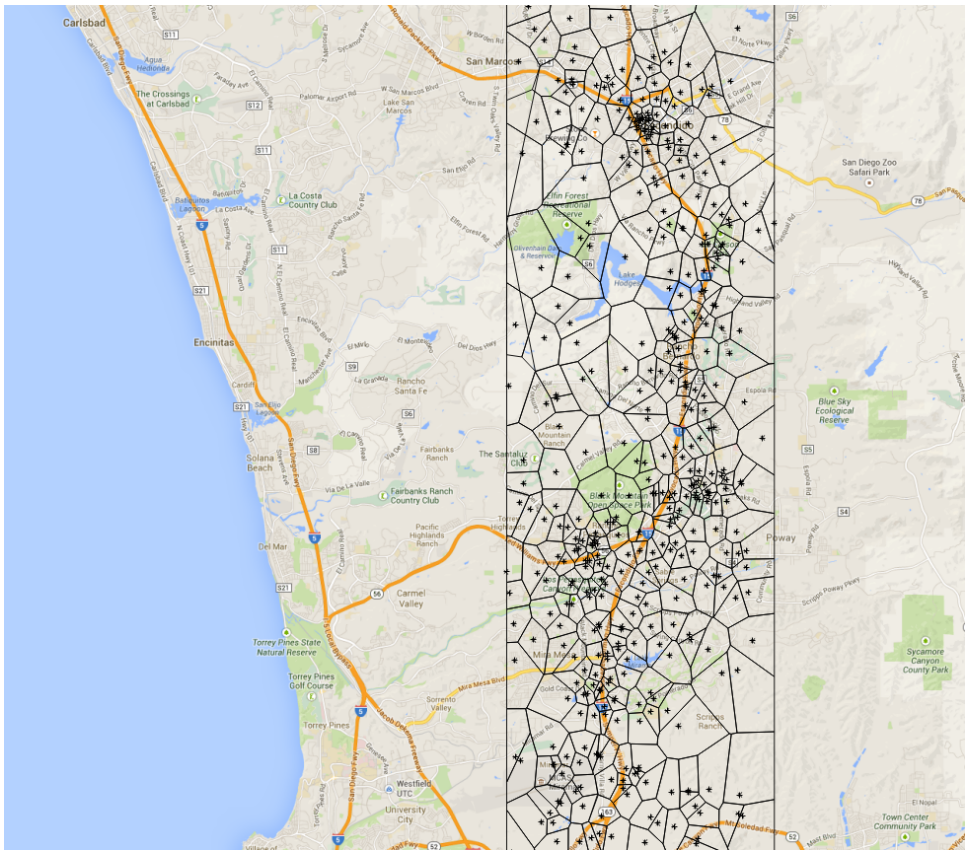


FIGURE 2 : Voronoi partition of the cell tower space for the I-15 corridor according to cell tower locations modeled based on population density in the corridor.

density.

PROBLEM FORMULATION

The rest of the article is organized as follows: Firstly, we present the model of the transportation network used in this algorithm and the assumptions about cell phone networks and the data acquired from the cell tower infrastructure. Secondly, a convex optimization framework is presented based on the combination of these models that enables efficient inference of vehicle density on links in a road network. Finally, experimental results are presented to validate the efficiency and accuracy of our model in which the ground truth is obtained from a microsimulation of a weekday morning rush hour on the I-15 corridor in San Diego County.

Modeling

We assume that a road network can be described as a directed graph $G = (V, E, w)$, where V is the set of nodes on the network, and E is the set of links, (Eppstein and Goodrich *(17)*). The variable w is a tuple for each link that encodes a set of properties of the associated link. For the purposes of this article, each tuple contains the link's length, number of lanes, and its "importance" to the network – an empirically determined parameter which captures information such as which roads are major thoroughfares and highways, and which are neighborhood streets that are less traveled. The number of vehicles on a given link $e \in E$ at a specific time t is denoted by n_e^t.

The road network is associated with a set of cell phone towers, T, each described by a geographical location and a history of the number of connections to that tower. We assume that there is a static coverage area of each tower $k \in T$. We assume that the coverage model can be represented by a Voronoi partition of the geographic region, Figure 2, according to distance from the nearest cell tower, as modeled by Yuan and Raubal *(10)*, Gonzalez et al. *(18)*, Steenbruggen et al. *(11)*, and many others. We call E_k the set of links that could possibly be connected to tower k. With hard boundaries, such as the Voronoi partition described here, this set contains only links that intersect the tower's Voronoi cell. However, if one were to develop a stochastic model, this set would be the set of all links that have a nonzero probability of having a cell phone user on that link and connected to the tower u.

For a given time interval $[t,\ t+\Delta t]$, each tower has a count of the calls placed in that cell during the time interval, which we call y_k^t. To simplify notation, we will assume the rest of the analysis applies only to a specific time interval, and so we will drop the indication of the time interval, and for notational convenience refer to these quantities by, for example, y_k. To use cell phone data to estimate vehicle density, we must make an assumption about the relationship between cell tower usage and road network utilization. In this model, we assume that the two are proportional. We assume that we can get the empirical coefficient of proportionality, such that the number of cars in the region $c_k = m_k y_k$, where $m_k > 0$ is a list of known constants, as proposed by Wang et al. *(9)*. For instance, if we assume a given carrier provides cell service to 50% of the population in all regions of the network, and that y_k accounts for only connections made by vehicles in motion, and that there are an average of 1.3 persons per vehicle, then $m_k = \frac{1.3}{0.5} = 2.6\ \forall k \in T$. In matrix form, we write $M = \mathrm{diag}(m_k)_{k\in T}$, and so $c = My$.

Optimization Framework

In our model, the prediction of the number of vehicles $\hat{c}_k$ in the coverage area of a cell tower k is a sum of the number of cars on each road segment that is entirely in the cell tower coverage region, plus the sum over of the number of cars on links that cross into that region, times the fraction of the link that is in the given cell. So we can define a prediction of the number of cars in the kth cell as

$$\hat{c}_k = \sum_{e\in E_k} n_e P(E_k|e) \tag{1}$$

$$= \sum_{e\in E_k} q_{k,e}\hat{n}_e \tag{2}$$

$$q_{k,e} = P(E_k|e) \tag{3}$$

where $P(E_k|e)$ is the probability of one being in the coverage region E_k given that he or she is on link e, and $\hat{n}_e$ is the predicted number of cars on link e.

In matrix form, the number of cars $\hat{c} := (\hat{c}_k)_{k\in T}^{\mathsf{T}}$ if a aggregation of the number of cars on each link $\hat{n} := (\hat{n}_e)_{e\in E}^{\mathsf{T}}$ such that $\hat{c} = Q\hat{n}$, where $Q = (q_{k,e})_{k\in T, e\in E} = (P(E_k|e))_{k\in T, e\in E}$. Hence, we expect the best estimate of the link densities $\hat{n}$ to be one that minimizes the error in estimating the number of phones connected to a cell tower, $y = (y_k)_{k\in T}^{\mathsf{T}}$. This can be formulated as the following optimization program.

$$\begin{aligned} \underset{\hat{c},\hat{n}}{\text{minimize}} \quad & \|y - M^{-1}\hat{c}\|_2 \\ \text{s.t.} \quad & \hat{c} = Q\hat{n} \end{aligned} \tag{4}$$

As the objective function is a sum of squares and all constraints are linear in the problem variables, $\hat{c}$ and $\hat{n}$, the problem is a quadratic program. We can see, though, that if any of our cells have more than one link in them, then this problem would be have many solutions that minimized the loss, as vehicles could always be removed from one link and placed on the other without affecting the total number of vehicles in the cell. However, it is intuitively the case that some distributions of vehicles onto the links within a tower's coverage region are more likely than others– it is unlikely that everyone is on neighborhood roads and no one is on the highway, especially if there are cars on that highway in the adjoining region.

To resolve this issue, we propose a set of constraints on the set of links that reduces the degrees of freedom. In this model, we define a similarity function on links, $s(e,e')$ which measures the similarity of e' to e. This similarity measure should express how density on one road segment will affect the density on another road segment. We then define the expected similarity between a link and a cell phone region as

$$S(e,E_u) = \sum_{e' \in E_u} s(e,e')P(e'|E_u) = \sum_{e' \in E_u} s(e,e')p_{e',u} \text{ with } p_{e',u} = P(e'|E_u) \tag{5}$$

where $S(e,E_u)$ is the expected similarity of e to the u-th tower, and $P(e'|E_u)$ is the probability of being on link e' given that one are connected to cell tower u. $S(e,E_u)$ encodes the rate at which density varies on link e with densities on other links through their relationship to the spatial region covered by tower u. The similarity need not be symmetric, but in this case, care should be taken such that the result measures the similarity of e' to e. Later, we shall propose a set of similarity metrics that are simple and can be extracted from the graph representation of the road network.

Because the expected similarity captures the effect of a specific cell region on a link's vehicle density, the estimate, $\hat{n}_e$, of the total vehicle density on each link is the sum of the effect of each tower on the corresponding link, weighted by a single quantity for each region, α_u, $u \in T$, which captures how much that region contributes to traffic on the network.

$$\hat{n}_e = \sum_{u \in T} S(e,E_u)\alpha_u \tag{6}$$

And so, by eliminating variables and simplifying the form, we can write the objective as a least squares in the $\alpha = (\alpha_u)^\intercal$ variables.

$$\underset{\alpha}{\text{minimize}} \quad \sum_{k \in T} \left(y_k - M^{-1} \sum_{e \in E_k} q_{k,e} \sum_{u \in T} \sum_{e' \in E_u} s(e,e')p_{e',u}\alpha_u \right) \tag{7}$$

which we can write in matrix form as

$$\underset{\alpha}{\text{minimize}} \quad \left\| y - M^{-1}QSP\alpha \right\|_2^2 \tag{8}$$

This program can be interpreted as a way of estimating the effect that each region has on vehicle density in the network from indirect measurements of call volume.

In this current formulation, it is possible that noise or modeling error could allow the choice of α that minimizes this error to lead to a negative count on a link. This would be an infeasible in reality, as would the case where more vehicles are on the link than the link's capacity could carry. To prevent infeasible solutions, we add an inequality constraint that for the estimated count on each link e, $0 < (SP\alpha)_e < (n_{\max})_e$.

The current objective is not guaranteed to be strongly convex unless the matrix QSP is full rank; to guarantee strong convexity and help reduce over-fitting to noise, we propose adding a regularization term to the objective. The most commonly used regularization is ℓ_2 regularization, which results in a least-squares least-norm problem, (Boyd and Vandenberghe *(19)*). However, ℓ_1 regularization still results in a QP, but encourages more weight near zero and in the tails of the distribution by placing a Laplacian prior on the link densities, (Tibshirani *(20)*). Therefore, ℓ_1 regularization is a more accurate form of regularization for this problem. We provide numerical results to verify this claim in the Experiment section.

Thus, our final formulation can be posed as the following quadratic program,

$$\begin{aligned} \underset{\alpha}{\text{minimize}} \quad & \left\| y - M^{-1}QSP\alpha \right\|_2^2 + \lambda \mathbf{1}^\intercal SP\alpha \\ \text{s.t.} \quad & 0 \leq SP\alpha \leq n_{\max} \end{aligned} \tag{9}$$

Model selection

We have a number of road and cell network modeling decisions that need to be made: firstly, we need a model for the probability $P(e|E_k)$, and secondly, we need a model for the similarity function, $s(e,e')$. Figure 3 illustrates how these models are incorporated along with the optimization program into the framework. We present a simple model and in our experiment, describe the effects of choosing this model on the performance of this method. In this model, we assume that if a call is made through a cell, it could come from any link in that cell. The total area covered by a link is the length of the link, ℓ_e times the number of lanes N_e. The length of the link must be further divided into the fraction of it that lies within the Voronoi partition of tower t, f_{et}. These properties are assumed to be known in the road network, from a map of the region.

This area can then be weighted by a factor according to the road type, as described in the model assumptions, that captures how likely this road is to be traveled by commuters, local traffic, and how often these groups place calls. Pozdnoukhov and Kaiser *(16)* proposes using a weight of 1 for busy streets, and weight of 0.5 for local roads. In this model, we use a weight of 1 for highways or freeways, and a weight of 0.5 for all other roads. This modeling choice could be adjusted to improve estimates to better match what is observed in a real road network of interest. We model the probability of being on a link e when placing a call in the cell k which covers E_k as

$$P(e|E_k) = \frac{N_e \ell_e f_{ek} w_e}{\sum_{e' \in E_k} N_{e'} \ell_{e'} f_{e'k} w_e} \tag{10}$$

Also, we can estimate the probability that one is in a cell given that he or she are on a link is just the fraction of that link which is in each cell. In our model, for most links, this is either 1 or 0. However, for links that cross between two towers, it is:

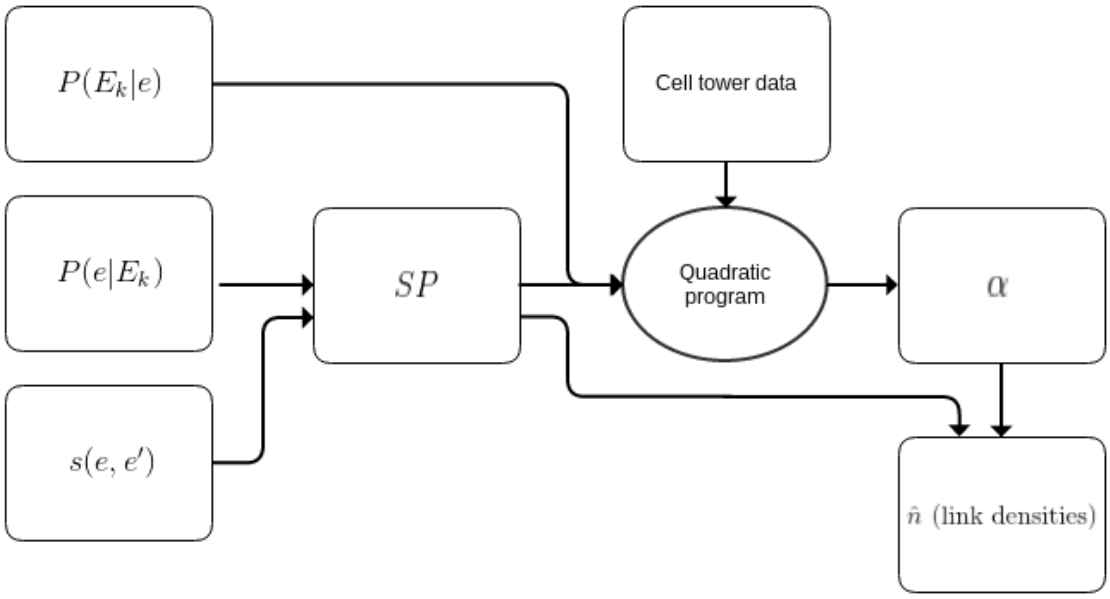


FIGURE 3 : Model inputs to the framework.

$$P(E_k|e) = f_{ek} \tag{11}$$

Finally, the choice of similarity function must be made. We propose using a simple function, inspired by the literature on kernel learning (Borgwardt and Kriegel *(21)*), based on shortest path distance. This captures the fact that the density on links are likely to correlated to links that are nearby. Given the shortest path distance, $d(e,e')$, we define the similarity function,

$$s\left(e,e'\right) = \exp\left(-\beta d\left(e,e'\right)^2\right) \tag{12}$$

This leaves us with one hyper-parameter for this similarity function. Along with the regularization parameter, λ, we are left with only two hyper-parameters, a reasonable number to be able to estimate in a cross validation scheme. Performance can be measured in terms of squared error of counts on a set of links with known counts as well as squared error of number of calls in a given cell.

In this sense, the α parameters describe how density is distributed among regions of the road network, and the P matrix describes how density associated with a region is divided up among the links in that region. Furthermore, Q describes how links are aggregated together to get the number of cars contributing to connections in a cell phone region.

NUMERICAL EXPERIMENTS

We assessed the accuracy of this formulation by collecting data from a microsimulation and then comparing the estimates generated by the model with the true counts measured in the microsimulation. We present a method for assessing the accuracy of a model for vehicle density through this framework, and discuss how the choice of model used can affect the accuracy of the results.

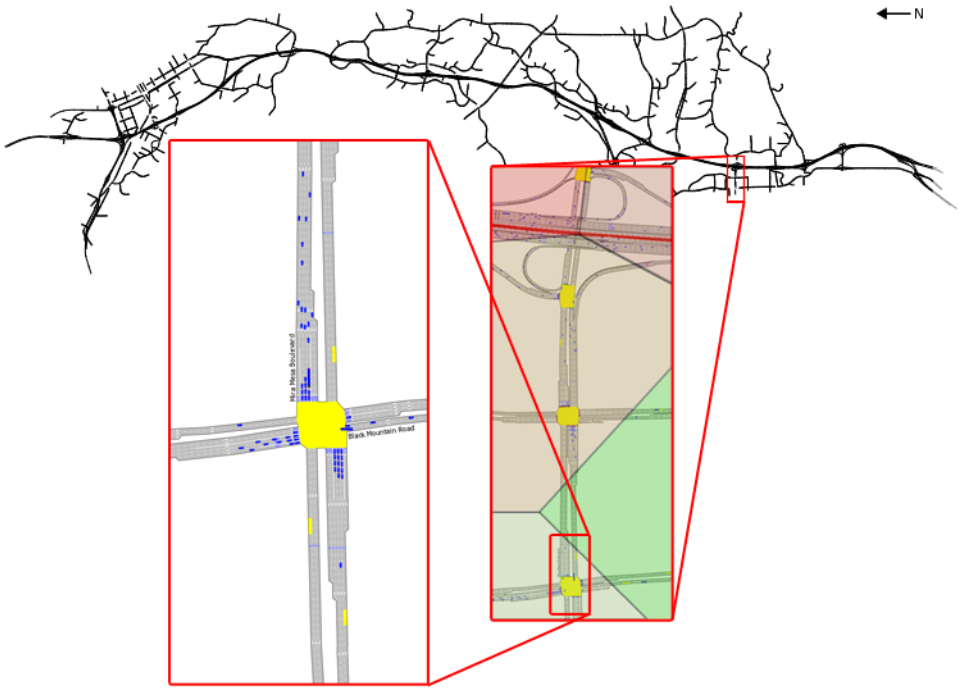


FIGURE 4 : Screen captures from Aimsun microsimulation shows full view of network used in microsimulation, a couple blocks with cell tower coverage map overlay, and a close up with detail of vehicle trajectories.

Methods

A calibrated *Aimsun 8* microsimulation was used to simulate the morning rush hour in the I-15 corridor, from I-15 mile marker 12 to 33. The map included arterial roads in the area as well, and in total, there were 3273 links in the network. The microsimulation was fed a calibrated OD demand model developed by SANDAG, (Miller and Skabardonis *(22)*). The simulation was run with a 0.85 second step interval with 370,000 unique vehicles simulated, and all vehicle trajectories were logged every time step and stored on the cloud in a PostGIS spatial database for analysis. Cell phone tower locations were simulated for the same region, using a mixture model of a uniform distribution over the region, and a kernel density estimated distribution along roads in the network, with a Gaussian kernel. In total, 300 cell towers were distributed over the network covering roughly 300 sq. miles. The cell tower coverage regions were calculated by performing a Voronoi tessellation on the cell tower locations. Figure 2 shows the resulting coverage map. These tessellations were saved to a table in the PostGIS spatial database with the microsimulation trajectories for analyzing coverage. Figure 4 shows screenshots of the Aimsun model on the network.

In total, the entire table of vehicle trajectories is approximately 30 GB. Spatial and temporal indices were constructed for trajectory data in the database, though care is still needed to be taken to efficiently query the data, as each time slice returns a large number of rows. The trajectory data was aggregated over a time interval by counting the number of unique vehicle identifiers on each link in the network to compute n^t_{true}, for a time slice t. Similarly, the number of connections for each cell tower was computed by counting all vehicles whose location during a specific time slice

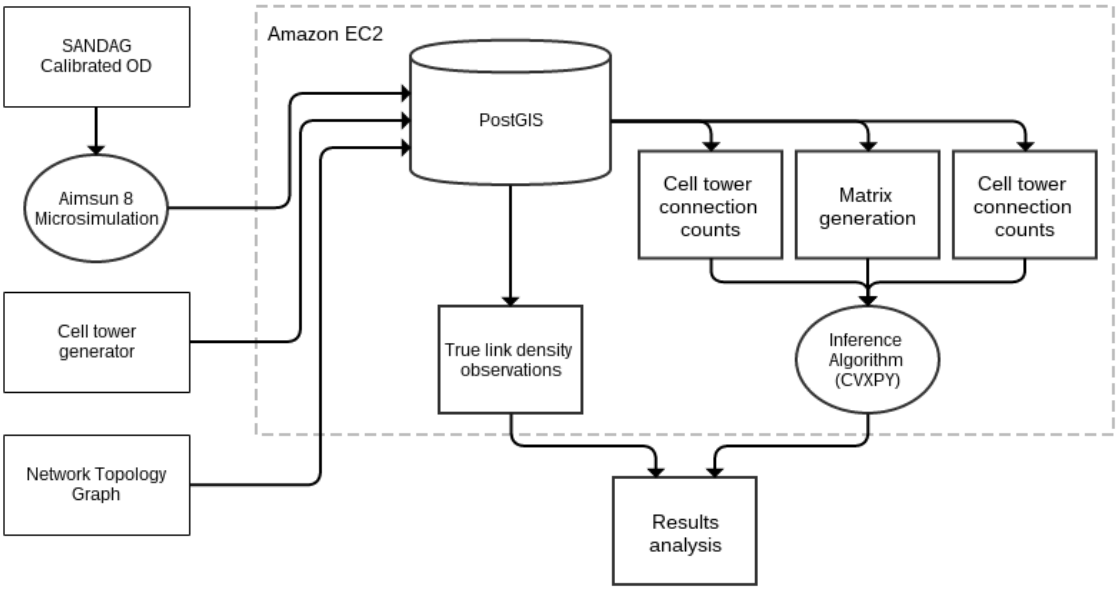


FIGURE 5 : System diagram for simulation and data analysis method.

was in that tower's Voronoi cell. This aggregate was scaled by the penetration rate to simulate the number of cell tower connections observed.

We sample 10 minute intervals from the morning rush hour, 5:00 am to 10:00 am. In each interval, we counted the average number of cars on each link and the average number of cars in each cell. We assume a 25% penetration rate among drivers and assume that all cars are connected to the nearest cell tower in modeling cell phone connections in the microsimulation. We use these same assumptions in choosing M and P in our model for density estimation.

We solve the optimization problem in terms of $\hat{n}$ and α using the python optimization package `cvxpy` with the `SCS` second order cone solver, (Diamond et al. *(23)*). We adjust the results for the penetration rate and compared our estimates of $\hat{n}$ to the true counts on each link, as extracted from the microsimulation results. We tested two different probability distributions for $P(e|E_k)$: the first used the true distribution derived from the simulation, and the second used the estimate we proposed above. An overview of the system for simulating this method is shown in Figure 5.

To assess the accuracy of this model, we suggest using the R^2 correlation coefficient of the link counts to judge the accuracy of the prediction. Because our model is a constrained linear one, R^2 corresponds to the fraction of variation captured by the estimates.

Distribution of Vehicle Density

We present the histogram of the distribution of n_{true}, to justify our use of the ℓ_1-norm for regularization. In Figure 6, the empirical distribution of the number of cars on a given link is shown. This is calculated from the true distribution found from mining the microsimulation. $p(n)$ is the probability that a given link has n cars on it. The histogram shows that the true distribution has more mass in the tails and near 0 than an exponential distribution, but it is still much better than a normal distribution. Tibshirani *(20)* shows that for an appropriate choice of λ, the $\lambda\|SP\alpha\|_1$ term in the objective represents a Laplacian (two-sided exponential) prior over $\hat{n}$. The ℓ_1-norm is the smallest p-norm which is still convex, so despite the fact that we would prefer to add a regularization that

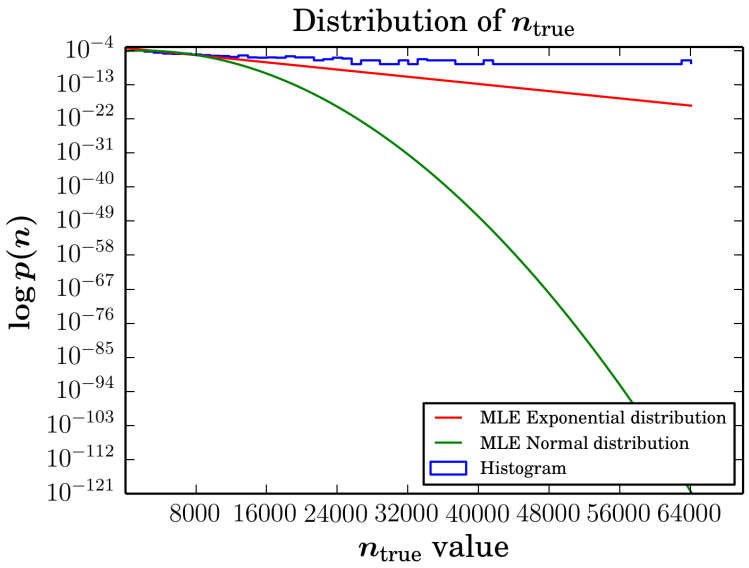


FIGURE 6 : Distribution of number of cars on links in simulation.

would put more weight at zero and the tails, using ℓ_1 regularization is a decent convex relaxation. We make this into an exponential distribution by imposing the non-negativity constraint on $\hat{n}$.

Prediction accuracy

Using the true distribution of $P(e|E_k)$, we observe accurate estimates of link densities, consistent with the microsimulation. The estimates produced by the model have an R^2 correlation coefficient as high as 0.998, as shown in Figure 7a. What we observe in Figure 8a is that with this model for P, larger β does better. We can see that

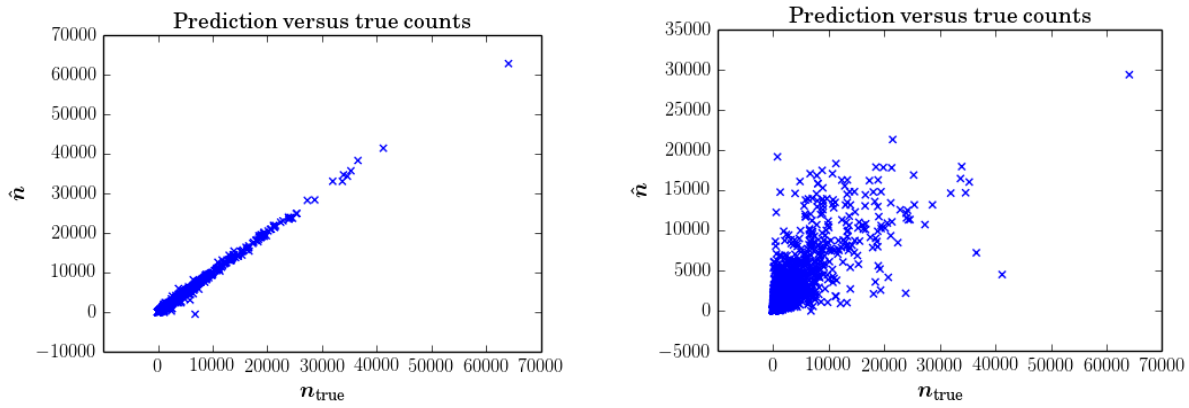
$$\lim_{\beta \to \infty} s(e,e') = \mathbb{I}_{\{e=e'\}} \tag{13}$$

where $\mathbb{I}$ which is 1 when its argument is true, and 0 otherwise. So, a larger β corresponds to the correlation matrix being "simplified" out of our model. This is not too surprising, as having the exact probability distribution means that the estimates don't need to be smoothed onto similar links, so mapping each onto only itself is the best one can do.

However, what is surprising is that this tends to be true for our simple model of the distribution as well, as shown in Figure 8b, despite the fact that this model does not completely characterize the true distribution of vehicles on links. We believe this is an indication that the choice of $s(e,e') = \exp(-\beta d(e,e'))$ is not the best choice for describing the relationship between links in a road network. More investigation into this is warranted, to look for a better choice of S that does a better job of infusing the behavior of traffic into the model. The prediction versus true distribution scatter plot is shown for the best choice of hyper-parameters ($R^2 = 0.767$) in Figure 7b.

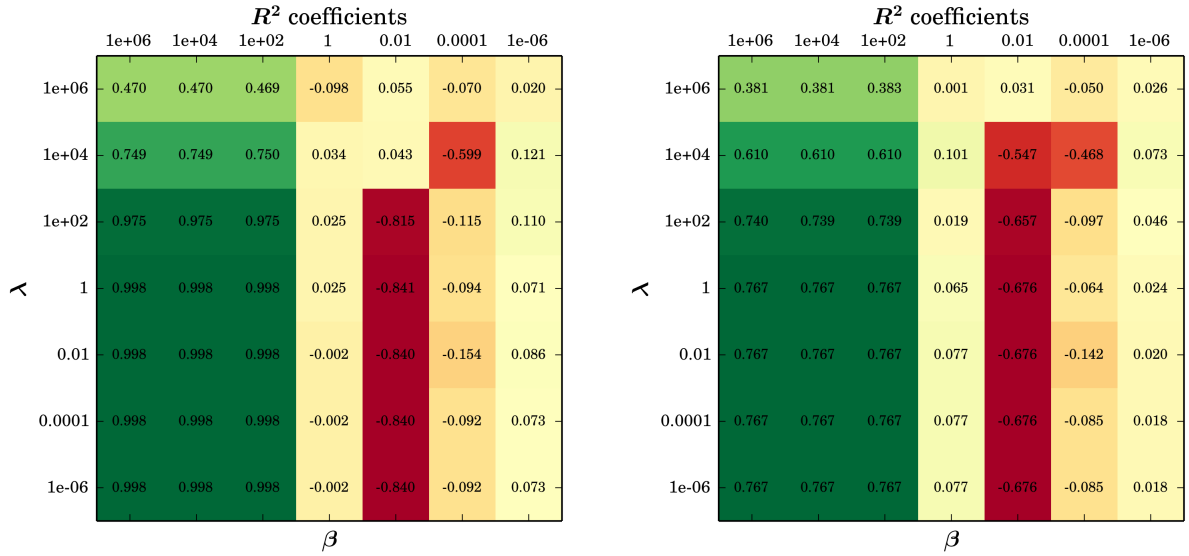
Adaptability of Model

Having the true distribution for P makes a significant difference, so finding a better model for this is crucial to further improving the results. For example, Traag et al. *(24)* proposes a more sophis-



(a) Exact P matrix. (b) Estimated P matrix.

FIGURE 7 : Car counts scatter plot.

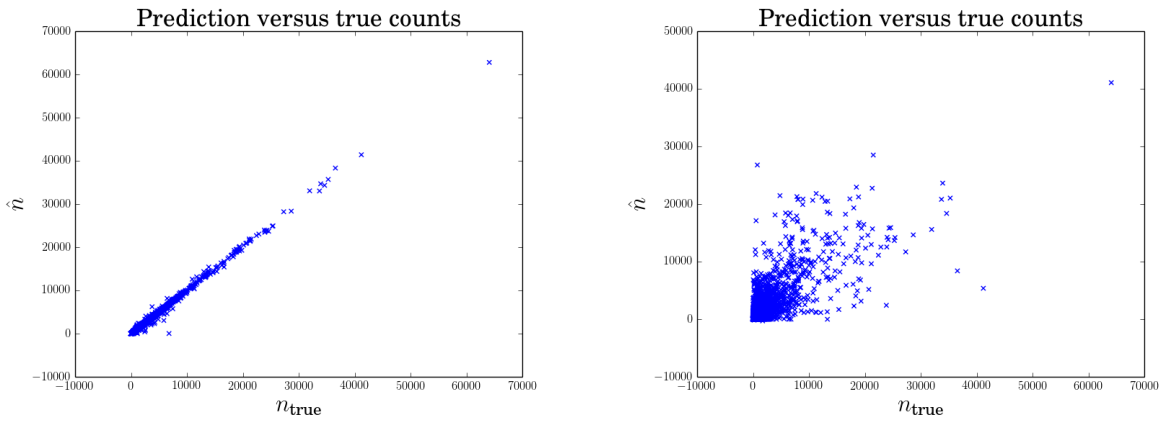


(a) Exact P matrix. (b) Estimated P matrix.

FIGURE 8 : Effect of hyper-parameters on prediction accuracy.

ticated model for fuzzy Voronoi tessellation that captures the uncertainty of connections near cell tower boundaries. However, the model we have presented for P is simple, and its implementation in this framework is clear. To demonstrate that the framework can easily be adapted to implement a different model for $P(e|E_t)$ and $P(E_t|e)$, We choose another model, implement it in our microsimulation, and show that our framework produces estimates with similar error as presented above in the case of the simpler model.

In this model, we assume that there is a 70% chance that a phone is connected to the nearest tower, and a 30% chance that it is connected to any of the surrounding towers. We use this model when assigning counts to towers in the simulation. Figures 9a and 9b show that when we model this in P and Q in the framework, the model works roughly as well, with R^2 accuracy of 0.998 when the true distribution is known and R^2 of 0.74 when we use the following estimate:



(a) Exact P matrix.

(b) Estimated P matrix.

FIGURE 9 : Car counts with overlapping cell coverage.

$$P(e|E_k) = \frac{\gamma N_e \ell_e f_{ek} w_e}{\gamma \sum_{e' \in E_k} N_{e'} \ell_{e'} f_{e'k} w_e + \frac{1-\gamma}{|\operatorname{adj} k|} \sum_{u \in \operatorname{adj} k} \sum_{e' \in E_u} N_{e'} \ell_{e'} f_{e'k} w_e} \tag{14}$$

$$P(E_k|e) = \gamma \cdot f_{ek} + \frac{1-\gamma}{|\operatorname{adj} k|} \sum_{u \in \operatorname{adj} k} f_{eu} \tag{15}$$

where $\operatorname{adj} k$ is the set of cells adjacent to tower k, and $\gamma = 0.7$.

MODEL ANALYSIS AND DISCUSSION

These results provide evidence that the framework is effective in reproducing the vehicle densities on links in a road network, given that S, P, and Q are well modeled. However, we see that it is sensitive to errors on P and S. This shows that more effective modeling of these probabilities, perhaps based on the traffic patterns or driving behaviors, could be used to improve these results. While consistency with a simulation is nice, the reality of complexities in practical road networks and noise in cell tower data present challenges for building a practical implementation of the model employed by this network.

The key models that we explored in this experiment were $P(e|E_k)$ and $s(e,e')$, as modeling of $P(E_k|e)$ was simple to choose based on the model used in our simulation. We demonstrated that a more sophisticated model of $P(e|E_k)$ was key to our results, and suggested that a better choice of $s(e,e')$ could make it more useful to the framework as well. However, by taking a more careful look at where these appear in the framework, we can see that the two are closely related.

We rewrite (9) in a way that highlights this point. The implicit variable $\hat{n}$ can be written in to the program as variables constrained by an equality relation to the α:

$$\begin{aligned} \underset{\alpha,\hat{n}}{\text{minimize}} \quad & \left\| y - M^{-1} Q \hat{n} \right\|_2^2 + \lambda \mathbf{1}^\intercal \hat{n} \\ \text{s.t.} \quad & 0 \le \hat{n} \le n_{\max} \\ & \hat{n} = SP\alpha \end{aligned} \tag{16}$$

But in this formulation, α only appears in the constraints of $\hat{n}$. Otherwise, it is a free variable with no other constraints or effect on the objective. So, we can rewrite that constraint as:

$$\hat{n} \in \text{range}(SP) \tag{17}$$

and remove α entirely from the problem.

This constraint restricts the link densities to a $|T|$-dimensional subspace of $\mathbb{R}^{|E|}$ that should represent reasonable relationships of the vehicle densities between links that are related to one another on the roadway. Separating it into S and P, however, does provide value: it defines a useful interpretation of how this subspace can be constructed. However, some may find this second interpretation instructive, as it removes the implicit assumption that P is a stochastic matrix and may allow more sophisticated models for describing how vehicle densities are related to one another on the road network.

An important consideration of these experiments is that they are simulations, which do not fully verify the practicality of the modeling assumptions in the present work for real traffic inference problems. We recognize the challenges of validating a model such as this, and suggest experiments be performed on a real network and validated against existing sensors wherever available to investigate the modeling assumptions and accept or refine them as necessary. Choosing an experiment location would be key: an area must be selected with many sensors against which to compare estimates and a realistic cell phone network in the area.

CONCLUSION

This article presents a convex optimization framework for estimating vehicle densities on the links in a road network from call volumes collected from cell towers. This provides higher penetration of vehicle density estimates in the road network, as cellular infrastructure has much better distribution than loop detectors in the road. It can be used to provide estimates on a 10-to-15 minute time interval, which would be useful to traffic managers in choosing adaptive signal timing plans.

The framework is posed as a quadratic program, and so it can be solved efficiently in $O(n^3)$ time, where n is the number of towers in the road network. Typically, this is less than 1000. For an area the size of the I-15 corridor, there are about 300 cell towers. In our results, we demonstrate that it can feasibly be implemented for a road network of the scale found in integrated corridor management projects, such as the I-15 corridor in San Diego.

The framework is fed with a model for the probability of being on a specific link given that you are connected to a specific cell tower, and a model for the relationship between vehicle densities on every pair of links. Careful choice of these models has a significant impact on the performance of the estimates computed according to this framework. We presented an alternative perspective of the modeling task here, based on constraining the subspace of feasible link density observations.

In the experiment, we demonstrated that both simple and sophisticated models of cell tower connections can be handled robustly by the framework, demonstrating that more sophisticated modeling which may allow improved results would not break any implicit assumptions made in the convex optimization formulation.

Further refinement of this method involves developing more realistic models to be used in this framework, that would allow this to be used in practice for corridor management. This includes more realistic modeling of link density similarity, and cell tower connection dynamics. This could

also be extended to include traffic dynamics by using the estimates of densities from the last time step to construct a better model for the next time step.

Acknowledgements
The authors would like to thank Dimitrios Triantafyllos for his assistance with the Aimsun micro-simulation used for this study. They would also like to thank Ahmed El Alaoui for his help in the early stages of modeling the problem formulation.

REFERENCES

[1] Fontaine, M. D. and B. L. Smith, Part 1: Freeway operations: Probe-based traffic monitoring systems with wireless location technology: An investigation of the relationship between system design and effectiveness. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1925, No. 1, 2005, pp. 2–11.

[2] Lee, J. and B. M. Williams, Development and evaluation of a constrained optimization model for traffic signal plan transition. *Procedia-Social and Behavioral Sciences*, Vol. 17, 2011, pp. 490–508.

[3] Cheng, P., Z. Qiu, and B. Ran, Particle filter based traffic state estimation using cell phone network data. In *Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE*, IEEE, 2006, pp. 1047–1052.

[4] Herrera, J. C., D. B. Work, R. Herring, X. J. Ban, Q. Jacobson, and A. M. Bayen, Evaluation of traffic data obtained via GPS-enabled mobile phones: The *Mobile Century* field experiment. *Transportation Research Part C: Emerging Technologies*, Vol. 18, No. 4, 2010, pp. 568–583.

[5] Work, D. B., S. Blandin, O.-P. Tossavainen, B. Piccoli, and A. M. Bayen, A traffic model for velocity data assimilation. *Applied Mathematics Research eXpress*, Vol. 2010, No. 1, 2010, pp. 1–35.

[6] Calabrese, F., M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. *Intelligent Transportation Systems, IEEE Transactions on*, Vol. 12, No. 1, 2011, pp. 141–151.

[7] Toole, J. L., S. Colak, F. Alhasoun, A. Evsukoff, and M. C. Gonzalez, The path most travelled: Mining road usage patterns from massive call data. *arXiv preprint arXiv:1403.0636*, 2014.

[8] Ratti, C., S. Williams, D. Frenchman, and R. Pulselli, Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning b Planning and Design*, Vol. 33, No. 5, 2006, p. 727.

[9] Wang, P., T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González, Understanding road usage patterns in urban areas. *Scientific reports*, Vol. 2, 2012.

[10] Yuan, Y. and M. Raubal, Extracting dynamic urban mobility patterns from mobile phone data. In *Geographic Information Science*, Springer, 2012, pp. 354–367.

[11] Steenbruggen, J., E. Tranos, and P. Nijkamp, Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy*, 2014.

[12] Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira Jr, and C. Ratti, Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, Vol. 26, 2013, pp. 301–313.

[13] Rose, G., Mobile phones as traffic probes: practices, prospects and issues. *Transport Reviews*, Vol. 26, No. 3, 2006, pp. 275–291.

[14] Iqbal, M. S., C. F. Choudhury, P. Wang, and M. C. González, Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, Vol. 40, 2014, pp. 63–74.

[15] Jiang, S., G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M. C. González, 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.

[16] Pozdnoukhov, A. and C. Kaiser, Area-to-point kernel regression on streaming data. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on GeoStreaming*, ACM, 2011, pp. 33–36.

[17] Eppstein, D. and M. T. Goodrich, Studying (Non-Planar) Road Networks Through an Algorithmic Lens. *CoRR*, Vol. abs/0808.3694, 2008.

[18] Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabasi, Understanding individual human mobility patterns. *Nature*, Vol. 453, No. 7196, 2008, pp. 779–782.

[19] Boyd, S. and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2009.

[20] Tibshirani, R., Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, pp. 267–288.

[21] Borgwardt, K. M. and H.-P. Kriegel, Shortest-path kernels on graphs. In *Data Mining, Fifth IEEE International Conference on*, IEEE, 2005, pp. 8–pp.

[22] Miller, M. A. and A. Skabardonis, *San Diego I-15 Integrated Corridor Management (ICM) System: Stage II (analysis, Modeling, and Simulation)*. California PATH Program, Institute of Transportation Studies, University of California at Berkeley, 2010.

[23] Diamond, S., E. Chu, and S. Boyd, *CVXPY: A Python-Embedded Modeling Language for Convex Optimization, version 0.2*. `http://cvxpy.org/`, 2014.

[24] Traag, V. A., A. Browet, F. Calabrese, and F. Morlot, Social event detection in massive mobile phone data using probabilistic location inference. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, IEEE, 2011, pp. 625–628.