Traffic Matrix Estimation from Road Sensor Data: A Case Study*

Keyvan R. Moghadam
Viterbi School of Engineering,
RTH 418
3710 McClintock Ave.
Los Angeles, CA 90089
rezaeimo@usc.edu

Quynh Nguyen
Viterbi School of Engineering,
RTH 418
3710 McClintock Ave.
Los Angeles, CA 90089
quynhngu@usc.edu

Bhaskar Krishnamachari
Viterbi School of Engineering,
RTH 410
3710 McClintock Ave.
Los Angeles, CA 90089
bkrishna@usc.edu

Ugur Demiryurek
Viterbi School of Engineering,
PHE 306
3737 Watt Way.
Los Angeles, CA 90089
demiryur@usc.edu

ABSTRACT

We present a study which aims to infer the vehicular traffic origin-destination matrix for the Los Angeles Downtown Area, from a unique real-world LA Metro data source which comprises sensor information of traffic counts and speeds obtained in real-time from LA arterial road intersections. We review the possible solution approaches and discuss the one is used here in details. The final results are presented for three different time intervals with different traffic regimes of the same day. The validity of the approach and some major applications of the inferred origin-destination matrix is discussed at the end.

Categories and Subject Descriptors

I.6.4 [**Simulation and Modeling**]: Model Validation and Analysis; H.4.m [**Information Systems Applications**]: Miscellaneous

Keywords

Origin Destination Estimation, OD Matrix, Routing Validation, Drivers preference inference

1. INTRODUCTION

Los Angeles is the second largest city of the United States, with a population of about 3.8 million. Over the years, it has been developed with an emphasis on private transportation. It is spread over a large area (more than 500 square miles) that allows it to accommodate large numbers of single passenger commuters. However, the rate of population growth has made the city face an increasingly challenging problem to plan and manage its traffic in an efficient way.

For the first time, this work presents and utilizes a unique set of data that has been gathered by LA Metro, the major bus and rail service operator in this area. The data comprises the traffic count and traffic speed in most of the arterial intersections of Los Angeles city. We are working to analyze this dataset to get a better understanding of the underlying behaviour of drivers in the LA area, which we will use later on to plan for traffic control.

A fundamental quantity we seek that could give us a useful representation of the driving habits of people in certain geographical region is the Origin-Destination matrix (OD matrix) for the vehicular traffic. Each element in OD matrix corresponds to a particular pair of origin and destination points in the city, and it indicates the rate of travelers that will move from the origin to the given destination (for a particular period of time). Once we have the OD matrix we can use it to develop various applications which are concerned with urban traffic conditions and traffic control. We are also interested in understanding the routing preference of drivers for any given pair of source and destination. For each OD pair, there are multiple different possible paths that commuters associated with that OD may choose. The preference of certain paths may depends on drivers habit, the estimated distance and time of travel through the path.

We present an inference method to estimate the OD traffic matrix for downtown LA based on the measured traffic counts on sampled intersections. This is an under-determined problem in general and there could be multiple solutions that fits the sensor readings of the traffic counts. To pin down the most accurate inference we will incorporate other sources of urban information (Metadata) along-side with traffic counts. Namely, we use a smaller set of nodes, called here as hotspots, which are candidates for origin-destination pairs of considerable size. We also incorporate the travel distance distribution as an additional information input to the OD matrix estimation to eliminate the OD pairs with travel distances that are unlikely to happen.

*This work has been made possible by the generous support of Integrated Media System Center (IMSC) at USC.

WOODSTOCK '97 El Paso, Texas USA

Knowing that the problem solution also depends on our model for the drivers routing preferences, for the first time, we use a cross-validation method to evaluate our solution as well as to infer the proportion of the traffic using each possible routes for each of the possible origin-destination pairs.

The data set is new and to our knowledge our work represents the first attempt to get an estimation of the true vehicular traffic matrix in the LA region. We will also discuss different ways that this estimation may be used to develop relevant applications for urban traffic control, etc.

The rest of the paper is organized as follows. Section 2 reviews similar work that has been done on other data sets of the same kind. It will also review different methods that have been proposed and used in prior work for OD estimation. Section 3 states the problem in a formal way and describes the data set and details of the cleaning steps required. We talk about the solution approach of inferring the OD matrix based on our data in Section 4. The estimated OD matrix is presented in Section 5 and different traffic demand patterns that can be emerged from it are discussed in details. Finally we address the the validity of the approach and touch upon the use of estimated OD matrix in Section 6. Section 7 concludes the paper with possible future directions of the work.

2. RELATED WORK

OD matrices have been used in different contexts in order to provide the underlying truth about the demands that drive system behaviour in various networks. Examples of this are network tomography [16] and urban traffic planning [2, 13].

In the literature of urban planning, there are many works relying on OD matrix of a certain region to synthesize realistic vehicle mobility traces. For instance, Uppoor *et al.* [13] use OD and underlying traffic road network as feeds to the SUMO [1] traffic generator to study traffic pattern for city of Koln in Germany. Another example is VanetMobiSim [5] which is a simulator concerned with both the traveling path choice of commuters (Macro-mobility) as well as the individual car interactions (Micro-mobility). It is built upon OD input to emulate the urban traffic traces for use in telecommunication VANETs.

In the context of network tomography the OD matrix represents the data demand from one point to another point in a communication network infrastructure. It is easy to get a precise estimate of the OD matrix if we could sniff packets at routers and read their headers. Even on occasions where reading the packet headers is not possible, knowing the network routing policies and assuming Poisson arrivals for the traffic [6] helps to get a precise OD estimate solely based on packet counts in the routers.

On the other hand, it is more complicated to determine the OD-matrix for a real-world urban traffic network. The equivalent of packet header readings in urban traffic setting is tracking the GPS traces of individual vehicles [20] which is not always possible due to privacy and accessibility constrains. Moreover, Poisson arrivals assumption does not hold for urban traffic networks and we can not have access to drivers routing preferences beforehand.

Traditionally, urban traffic OD matrix is estimated through individual household surveys on their commuting habits and residential commuting needs [13], roadside interviews and plate methods [2]. However, while it can be very costly to do that, it is not giving a fine grained precise estimation and rather provide a prior belief about the general shape of the OD matrix. Munuzuri *et al.* [9] do OD estimation for city of Seville. To avoid the cost of data gathering through surveys, its authors construct their model on entropy maximization and algorithmic solutions such as Frank-WolfeâĂŹs linear approximations. Authors in [20] use GPS data from taxis in Shanghai to infer the OD matrix. Other major branch of works in this context have used traffic counts [10, 2, 9] to get an estimate on OD matrix of the targeted region. Van Zuylen and Willumsen [14] use traffic counts gathered through induction loop in highways in Amsterdam to model trip generation.

The major drawback with the OD estimation technique using the traffic counts is the visibility issue i.e. what we sense (traffic counts on each road segment) does not provide full observability to the system OD matrix elements, or in another words the OD matrix cannot be inferred unambiguously in general from the traffic data. Different ways have been considered by researchers to pick the best OD from the set of possible ODs that fit the observation. Some have considered the maximum entropy approach [15] in which among all OD matrices in the feasible set, the one with the maximum entropy will be picked. Some have used a prior belief that maybe resulted from surveys combined with Bayesian inference to find the best OD [12]. Statistical approaches that leads to parameter estimation by means such as maximum likelihood maximization have been also used to choose the OD which is most likely to happen [6]. However, as we explain later on, the basic model assumptions of such approaches do not necessarily hold for arterial traffic patterns.

As we discuss, we have undertaken in this work a different approach, which combines on the one-hand a reduction of the order of the model (going with an aggregated smaller set of OD points) and on the other hand the incorporation of more information inputs such as the trip distance distribution, in order to obtain a unique OD matrix in a tractable manner. We also use a validation technique to infer the routing matrix (Driver's rout preference) as well as to evaluate our solution.

3. PROBLEM STATEMENT

Origin-Destination matrix reflects the underlying behavioural structure of commuting needs of people residing in a certain geographical region. Once one have an estimation of this matrix they will be able to understand the traffic needs and accommodate the need considering different restrictions that may be imposed.

Given a certain geographical region, the road network lying in the area can be modelled as a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges of the graph. The road network intersections will map onto the the graph nodes and for each road segment connecting two intersection directly, there is a corresponding edge in the graph.

The OD matrix is a matrix of the size $|\mathcal{V}| \times |\mathcal{V}|$ with zero diagonal elements i.e. there is no trip starts and ends at the same spot. Each element of the matrix, od_{ij}, represents the number of cars which start their journey at the corresponding intersection, i, destined for the other intersection, j, in the mapped graph of road city network within unit of time.

The goal is having the available valid counts of vehicles in the unit of time and their speeds on a set of road segments,

to infer an OD matrix that fits the best to our knowledge of the city as well as our data.

To achieve this, we need to know the amount of traffic which is contributed by each OD pair to the traffic count of each road segment. This is, in fact, affected by the routing choice of the commuters driving between each origin and destination pair. The choice of the drivers might depends on many factors, including the shortest path, fastest rout and their particular driving habits.

For a given commuters routing preference we can construct the routing matrix. Which is a matrix with columns corresponding to possible OD pairs and row corresponding to the arterial road segment. An element of the matrix indexed by i, j is then represents the proportion of the traffic between OD pair j that uses road segment i.

While works has been done to estimate drivers behaviour in choosing different paths based on survey or GPS traces [18], here we are interested in inferring the routing preference of drivers based on traffic congestion and delay estimates of different routes. Where both traffic congestion and delay estimates can be extracted from the traffic counts and average speed in the arterial road segments.

4. SOLUTION APPROACH

We start with some notations. We use $\mathbb{R}$ to indicate real numbers and $\mathbb{R}_+$ for non-negative real numbers. If $\mathcal{A}$ and $\mathcal{B}$ are given sets then ${\mathbb{R}_+}^{\mathcal{A}}$ is the set of all vectors of size $|\mathcal{A}|$ with values that are non-negative and are indexed by elements of $\mathcal{A}$. With the same logic ${\mathbb{R}_+}^{\mathcal{A}\times\mathcal{B}}$ is the set of matrices with non negative real value that are indexed by elements of $\mathcal{A} \times \mathcal{B}$. In the rest of the paper we use the following notation to formulate our model:

- $\mathcal{E}$: set of directed edges/road segments.
- $\mathcal{V}$: set of nodes/intersections.
- $\mathcal{G}(\mathcal{V}, \mathcal{E})$: The directed graph that represents intersections and road segments.
- $\forall e \in \mathcal{E}, w_e, \eta_e \in \mathcal{V}$: w_e is the node connected to the tails of edge e and η_e is the node connected to the head of that edge.
- $\mathfrak{L} = \{\mathcal{V} \times \mathcal{V}\}$: Ordered set of pairs of nodes. This corresponds to all possible OD pairs.
- $\phi_j, \delta_j \in \mathcal{V}$: Respectively the source and destination of the j^{th} element in $\mathfrak{L}$.
- $f \in {\mathbb{R}_+}^{\mathcal{E}}$: Vector of traffic counts. Each element represents average count of cars leaving each road segment per unit of time.
- $x \in {\mathbb{R}_+}^{\mathfrak{L}}$: Vector of flows. Each element of the vector represent average count of cars that travel from the origin to the destination of the corresponding OD pair in unit of time.
- $s \in {\mathbb{R}_+}^{\mathcal{E}}$: The average speed of traffic in every road segment in the given period of interest.
- $d \in {\mathbb{R}_+}^{\mathcal{E}}$: The average time it takes a car to traverse each edge in the current traffic load.
- $\mathcal{M}^j$: Set of all possible routs corresponding to OD pair, indexed j^{th} in $\mathfrak{L}$.
- $D^j \in {\mathbb{R}_+}^{\mathcal{M}^j}$: Vectors of the delay associated with each path in $\mathcal{M}^j$.
- $\mathcal{P}^j$: Set of all usable routs corresponding to OD pair, indexed j^{th} in $\mathfrak{L}$.
- $J^e \in \mathbb{N}^{\mathfrak{L}}$: Vector of link $e \in \mathcal{E}$ utilization by different ODs.
- $A \in {\mathbb{R}_+}^{\mathcal{E}\times\mathfrak{L}}$: Routing matrix.

To solve the problem described in section 3 , we consider the traffic network to be in its equilibrium and model the traffic system dynamics with steady flows of cars. Each flow corresponds to one origin destination pair in the OD matrix. The flow sizes are shown by the elements of the vector $x^{|\mathcal{E}|\times 1}$. It is expected that each driver in the flow would take the path to the destination that is the shortest either with respect to the time or the distance. When there are multiple paths with same range end to end delay, drivers may choose each of available paths based on a probability distribution that may change depending on drivers priority and estimation accuracy. It has been assumed that most of drivers do not have a delay estimation drastically different from the reality. This is given the fact that they usually construct their estimation based on online traffic maps and past experiences.

The traversing time of each link can vary in time as traffic congestion varies from time to time. Knowing that the traffic congestion have slow transients, we can get a realistic view of traffic by averaging over the speed of passing vehicles in each intersection. This will also help to account for the effect of traffic lights. Given s as part of our collected data and the road network details we can compute d which is the vector of link traverse time delays.

For an OD indexed j in $\mathfrak{L}$ a possible loop free end to end travel path can be defined as follows:

$$\mathcal{M}^j = \{m^j | m^j = m'^j \cap m''^j\} \tag{1}$$

Where:

$$\begin{aligned} m'^j &= \{e | e \in \mathcal{E}, (\exists! l \in m^j : \eta_e = \omega_l) \vee (\eta_e = \delta_j)\} \\ m''^j &= \{e | e \in \mathcal{E}, \exists! l \in m^j : \eta_l = \delta_j, \exists! l \in m^j : \omega_l = \phi_j\} \end{aligned} \tag{2}$$

In particular M^j is the ordered set of all distinct sets m^j. Where m^j is the collection of all edges that form a path from ϕ_j to δ_j. having the average traverse time of each link we can calculate the D^j as follows:

$$D_i^j = \sum_{e \in \mathcal{M}_i^j} d_e \tag{3}$$

In which D_i^j is the i^{th} row of D^j and $\mathcal{M}_i^j$ is the i^{th} element of $\mathcal{M}^j$.

Once a traveler wants to choose its travel path to its destination, there might be multiple paths with travel time in the same range of the shortest path delay. We name the set of paths that might be used by a traveller in the current network traffic state $\mathcal{P}^j$ and define it as follows:

$$\mathcal{P}^j = \{m^j | m^j \in \mathcal{M}^j, D^j_{m^j} \leq \alpha \min_i D_i^j\} \tag{4}$$

Where α is a constant greater and close to 1 and is added to capture the drivers preference and traffic estimation quality. The lower the α correspond to willingness of the drivers

to take routes with longer delays in order to get to their destinations. This willingness may come from bad estimations or other priorities such as selecting main arterial routes or ones with shorter travel distances.

We can consider a probability distribution vector h^j over the set of possible paths ($\mathcal{P}^j$) of each OD pair j to represent drivers priorities in rout selection. In this case, $x^j * h_i$ for $i \in \{1, \ldots, |\mathcal{P}^j|\}$ represents the actual portion of the flow taking path indexed by i from the set of ordered paths $\mathcal{P}^j$.

Knowing what paths each flow may take for each possible origin-destination we define J^e as follows:

$$J^e_j = \sum_{m \in \mathcal{P}^j} h^j_m \sum_{e' \in m} \mathbb{1}_{e'=e} \tag{5}$$

J^e_j is the j^{th} element of J^e and indicates the proportion of x^j that are using link e.

We can now build the routing matrix $A^{|\mathcal{E}|\times|\mathfrak{L}|}$. The elements of the routing matrix are in range of $[0, 1]$. An element $a_{i,j}$ represent the portion of flow of the OD indexed j^{th} in $\mathfrak{L}$ that is passing through road segment $i \in \mathcal{E}$ and will be equal to J^i_j.

Hence given the road network within a certain traffic state we can compute the corresponding routing matrix A. There is a linear equation between the actual size of OD pairs and the flow of traffic on each link in terms of A:

$$Ax = b \tag{6}$$

In which, b is the data that sensors are reporting on traffic count on each link and A is the routing matrix at the current traffic setup.

Equation (6) gives a straight forward way to compute traffic counts on each link knowing the OD pairs sizes. However, the reverse problem has more into it. If we consider to solve the equation for the unknown possible OD pairs, (6) is an under determined system of linear equations i.e. A is a low rank matrix. In other words, having the counts traffic on each link does not reveal all the underlying truth about the origin and destination of the traffic flows.

In order to narrow down our search for a plausible (close to reality) OD matrix, we have to incorporate other sources of information than only the traffic counts. As mentioned in previous sections some have used the Poisson distribution for OD sizes within a probabilistic framework to pin down one solution by solving a maximum likelihood optimization. Although i.i.d Poisson distributed OD size model might be acceptable for information packets in worldwide web or partially in highway traffic networks, they are not suitable for arterial traffic networks. One can get an intuition on this by looking into the first example in [16]. The Poisson assumption in the example has led to absolute preference of shorter OD pairs to the longer ones, which is not a justifiable preference in the case of arterial traffic networks.

Here, we incorporate two other class of information to narrow down our search space and get a more realistic estimation of ODs. We first identify the candidate spots that are more likely to be the source or destination of major traffic travel flows. These candidate nodes can be found by identifying the major hot spots of the city, clusters of residential areas and major highway entrances.

Having a set of candidate nodes any pair of them can be a candidate major OD pair. At this point we use our knowledge about the urban travel distance distribution to further limit the candidate sets by eliminating those pairs which are less likely to be valid. In other words, we choose the second source of our information to be the meta-data on urban commute distance distribution.

Trip distance distribution indicate the portion of the total trips that are made by the cars such that the length of the trip falls within a given range. It has been shown that the log normal distribution is a good fit for urban trip distance distribution [3]. An example of a log-normal distribution can be seen in Figure 1. As you can see, There is a distance threshold that having a trip distance below which is very unlikely. Intuitively it means that people are not likely to use vehicles to trip distances that are not long enough. We use this fact to further remove OD pair candidates that their corresponding trip distance would fall bellow the given threshold of trip distance distribution. Here we consider this threshold to be equal to 500 meters.

By narrowing down our search space, the new set of candidate ODs will be a subset of $\mathfrak{L}$ and will be shown by $\hat{\mathfrak{L}}$. We also have $\hat{\mathcal{E}}$ which is the union of the links that are used by $\hat{\mathfrak{L}}$.

$$\hat{\mathcal{E}} = \cup_{j \in \hat{\mathfrak{L}}} \cup_{m \in \mathcal{P}^j} m \tag{7}$$

And hence equation 6 would change to the following one:

$$\hat{A}\hat{x} = \hat{b} \tag{8}$$

Where $\hat{A} \in {\mathbb{R}_+}^{\hat{\mathcal{E}}\times\hat{\mathfrak{L}}}$ is a sub-matrix of A with those elements of A that have corresponding elements in $\hat{\mathcal{E}} \times \hat{\mathfrak{L}}$. Also $\hat{x} \in \mathbb{R}^{\hat{\mathfrak{L}}}_+$ and $\hat{b} \in \mathbb{R}^{\hat{\mathcal{E}}}_+$ are sub-vectors of x and b respectively with the same logic.

Equation (8) will not have more than one answer if the following is true:

$$\forall \mathfrak{L}_1 \neq \mathfrak{L}_2 \subset \hat{\mathfrak{L}} : \cup_{j \in \mathfrak{L}_1} \cup_{m \in \mathcal{P}^j} m \neq \cup_{j \in \mathfrak{L}_2} \cup_{m \in \mathcal{P}^j} m \tag{9}$$

In other words no two distinct subset of candidate ODs should have the same set of contributing links to their usable paths. If this statement does not hold for any two subsets of ODs, those two subsets are not distinguishable from each other based on the traffic counts.

On the other hand, even if (9) holds, equation (8) might be overdetermined. This is partly due to the number of small sized ODs that we have omitted. In this case the best possible estimate is the one with the smallest error. This leads in turn to solve the following optimization problem.

$$\begin{aligned} &\text{argmin}_{\hat{x}} ||\hat{A}\hat{x} - \hat{b}||_2^2 \\ &\text{subject to:} \quad \hat{x} \geq 0 \end{aligned} \tag{10}$$

However, we still need to assess the solution given by 10 for our problem. We should also note that $\hat{A}$ depends on the way we construct h^j and for different OD pairs. In fact, we may end up with different results for different ways of constructing h^j. To address these issues, we choose a cross-validation technique in which we randomly select %80 of road segments from $\hat{\mathcal{E}}$, call the new set $\hat{\mathcal{E}}_1$ and the remaining of the set $\hat{\mathcal{E}}_2$. We reconstruct optimization 10 into another optimization problem such that it only comprises rows corresponding to $\hat{\mathcal{E}}_1$. The solution to the new optimization formulation, $\hat{x}_1$ can be now evaluated with respect to the set of remaining road segment: $\hat{\mathcal{E}}_2$. Based on this, we define the average error as follows:

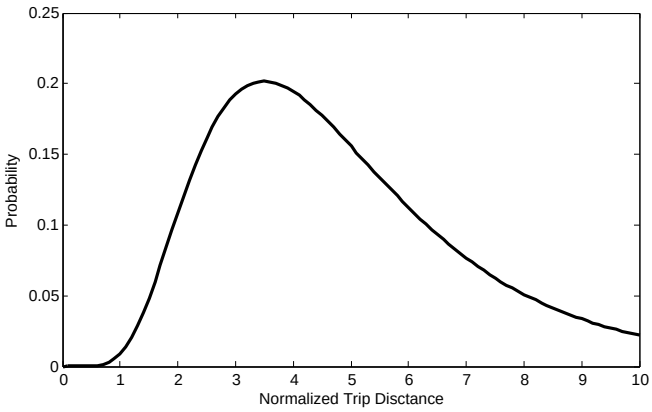


Figure 1: A typical log-normal distribution with $\sigma = 0.5, \mu = 1.5$

$$z = \frac{||\hat{A}_2\hat{x}_1 - \hat{b}_2||_2}{||\hat{\mathcal{E}}_2||} \tag{11}$$

Where $\hat{x}_1$ is the solution of the optimization formulation 12 and $\hat{A}_i$ is the sub-matrix achieved by selecting the rows of $\hat{A}$ that correspond to elements of $\hat{\mathcal{E}}_i$.

$$\begin{aligned} &\text{argmin}_{\hat{x}}||\hat{A}_1\hat{x} - \hat{b}_1||_2^2 \\ &\text{subject to:} \quad \hat{x} \geq 0 \end{aligned} \tag{12}$$

Optimization (12) is a linear least square optimization problem that can be solve by many different convex optimization tool. Here we use the *scipy.optimize* package of Python to solve it.

5. RESULTS

In this section we present the OD estimated results for the rectangular region including USC and downtown Los Angeles. The area of focus spread from (34.0730, -118.3060) to (34.0170, -118.1950) and has dimensions 6227×10922 meters. The map of the area is shown in Figure 2. This area includes 5498 intersections and 7584 distinct arterial road segments. The detailed information of the map including the the road segment ID, name, length, bearing, start and end latitude and longitude are all extracted from Open Street Map (OSM).

The LA Metro data are gathered through induction loop sensor in arterial road segments. The information in each sensor has been stored each minute and comprises of the followings:

- ID: Unique ID for the sensor.
- Link ID: Unique ID for the road segment.
- Link Type: Highway/Arterial.
- On Street: Sensor located on which street.
- From Street: Sensor is located on the lane that leave off this street.
- To Street: Sensor is located on the lane that goes toward this street.
- Date: Day of the reading.
- Time: Exact time of the reading.
- Latitude and Longitude: GPS Specific of the sensor.
- Occupancy: Percentage of the time Where the loop was occupied in the last reported minute.
- Speed: The average speed of the traffic flow in the reported minute.
- Volume: Count of cars passing over the loop sensor in the past minute.
- HovSpeed: Specifies if there is a carpool lane.
- Status: "OK" if the sensor reading is valid.

Since the sensors might fail to record the data from time to time or they may not working for a periods of the time, a cleaning phase is performed to prune the unreliable data. Also a processing phase is performed to let the valid data be available for each sensor in each 5 minute interval.

The LA Metro data is used alongside information extracted from Openstreetmap [4]. Openstreetmap is an open source user based database that provides most up to date maps of selected regions. The data set from Openstreetmap comprises the followings:

- Node ID: Unique ID for each intersection.
- Neighbour ID: For each node set of neighbouring node IDs are listed.
- Longitude, Latitude: Each node ID has its own Longitude and Latitude.
- Segment ID: Unique ID for each road segment.
- Segment end: the two end vertices's IDs of each road segment.

Another level of processing is done at this point to map the segments ID from the Openstreetmap database to induction loop sensor ID in the LA Metro data set. It is done here by comparing the Longitude and Latitude of the each sensor with of the same of each intersection in the Openstreetmap data base. From the set of the candidate segments connected to the selected intersection, we choose the segment that fits to the other specification of the sensor in the LA Metro data set.

Once we integrate our two databases, we could define our problem. The goal is having the available valid counts of vehicles and their speeds on a set of road segments, to infer an OD matrix that fits the best to our knowledge of the city as well as our data.

The LA Metro data set contains reading from 1086 distinct sensors in the area of focus. Each sensor reports the average speed and aggregate number of vehicles passing over them each one minute. We have merged the sensor location data with the OSM map data and find the exact location of each sensor with respect to the road network. Having 1086 number of sensors means that we have reading from 1086 number of edges of our modeling graph.

In the next step we have identified a set of candidate nodes that can participate in potential origins or destinations of major flows. This set has been chosen based on

meta data indicating the residential block, major highway, business block and potential hot spots of the city. There are 23 distinct nodes in this set, from which 7 nodes represent potential origins or destinations that fall outside of our map boundary. The set of candidate nodes with their location with respect to the map is shown in Figure 3.

$\hat{\mathfrak{L}}$ is constructed afterward from the cross product set of chosen nodes and using a typical trip distance distribution with a cut off value of 1640 feet (500 meters).

We focus on three different state of the traffic network through a day, morning and evening rush hours, alongside the afternoon traffic. Three different time windows of size 30 minutes are selected corresponding to each state which are respectively start at 7:00 am, 7:00 pm, and 1:00 pm. The collected data in each time window is averaged for each sensor to reduce the noise and anomalies effect. Next, the routing matrix $\hat{A}$ is constructed for each time window using the extracted information from OSM, sensory readings and different scenarios of the routing preferences. At this stage we evaluate the average error of the solution based on the cross-validation approach mentioned in the previous section. This way we can evaluate the performance of our solution for a given routing preferences. By finding the setting where the error is the smallest, we get an understanding on how on average the commuters behave when it comes to route selection.

Our results show that the small average error is achieved when the traffic is assigned to routes that have an end to end delay in the range of the shortest path rout. The split of traffic happens in a way that in general the paths with the shorter delay takes the higher portion of the traffic for a given OD pair. However an interesting observation is that the set of chosen routes by drivers not only have delays in the range of the shortest path but also have low *correlations* with each others. Here, we use the term correlation for two different paths between the same origin and destination to reference the average percentage of the length that they have in common. Figure 5 illustrate the results confirming this behaviour for all three different time slots.

In general when sorting the available paths based on their end to end delay, those on the top of the list have high correlations. We create a delay sorted list of low correlated paths for each OD pairs by adding the constrains that no path can be added unless it has less that 20% correlation with the ones above it in the list. Figure 5 also shows that on average the traffic split between the best 3 uncorrelated routes in the rush ours while this number is 2 in the light afternoon traffic.

Some of the major size resulting estimated traffic flows are listed in table 1. The full list of results is accessible in the extended version of this paper [8]. The sizes are the estimated number of vehicles that are traveling from the corresponding source to destination per one minute. By looking into the resulting estimates different traffic demands can be revealed. To further assist in interpreting the results we have visualized them on the map for different flow sizes and time window intervals.

Figure 4 shows the flow size distribution for the three different time intervals. As it can be seen in the afternoon time slot, the aggregate traffic flow is less compared to the morning and evening rush hours and there are more flows with small sizes, indicating more erratic behavior of travelers. In the morning and evening time interval there are a few flows

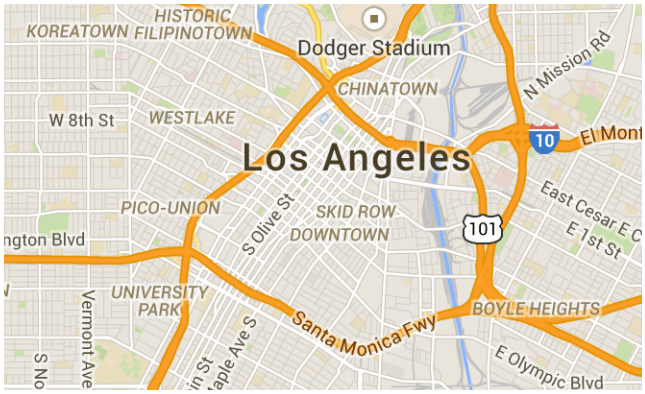


Figure 2: The area of focus. It is rectangle of size 6227 × 10922 meters including the Los Angeles Downtown area and USC Park Campus. Spread from (34.0730, -118.3060) to (34.0170, -118.1950).

with large sizes suggesting multiple flocks of cars traveling the same path. These flows have been shown with respect to the map in figure 6. Each arrow in the map corresponds to one flow, the arrow head indicates the destination and the tails indicates the origin of the flow. In both morning and evening time intervals one of the major flows is the one with ID 436 representing the vehicles that are passing downtown going to the north-east. It is also interesting to note that none of the flows with size greater than 25 per minute originate from or destined in downtown financial area, suggesting the major OD flow in the map are those which are passing the map center. This is while when looking in the afternoon time interval, the biggest size OD flows only contains short travel distances and happening only in the residential blocks, mostly in the west-lake area.

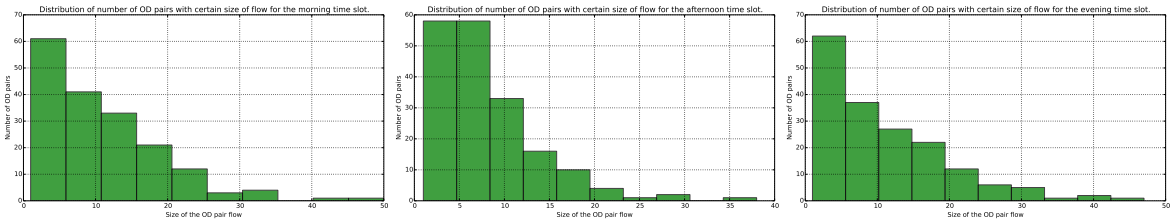
When looking into the flows with sizes between 21 and 25 vehicles per minute, there are many flows in the morning destined for downtown financial area and a few others originated from residential areas going toward destinations resides outside of them map. This is while there is not such pattern in the evening time interval.

The flows with medium sizes (between 15 to 20 vehicles per minute) are shown in figure 7. Interesting patterns can be observed as well for these flow sizes. As an example in the morning time slot most of the flows are originated from or destined for highway entrances. Remembering that each flow is a demand for the road network, we can see that highway 10 has the most demand in the network, both in the morning and in the evening.

6. DISCUSSION

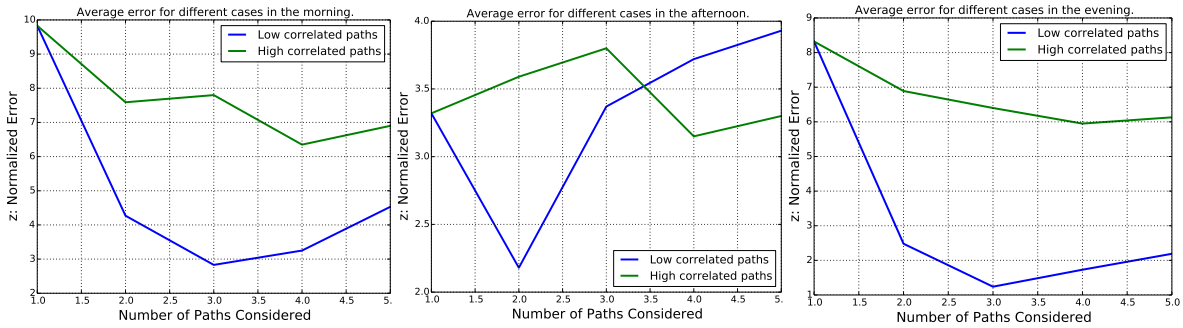
This section touches upon two different aspects that are very much needed to justify this work. First, is the concern of accuracy. It is important to know how reliable are the estimated results and more generally under what conditions the results can be considered valid. Second, is how helpful the estimated OD matrices can be. We address this one by going through some of the most important technologies and applications that could rely heavily on OD matrix estimation.

6.1 Accuracy



(a) Morning, 7:00am (b) Afternoon, 1:00pm (c) Evening, 7:00pm

Figure 4: distributions of the size of the flows in different time periods of the day.



(a) Morning, 7:00am (b) Afternoon, 1:00pm (c) Evening, 7:00pm

Figure 5: The average error for different scenarios of routing preferences.

Figure 3: Selected potential sources and destinations based on the travel distance expected behaviour with denser sampling in the area of focus with respect to the map details

The accuracy of the estimated OD sizes depends on many factors such as the accuracy of induction loop readings, drivers behaviour modeling, existence of traffic equilibrium and etc. But most importantly it depends heavily on the choice of the candidate ODs. To be more precise, for the estimated flow sizes to be close to reality the set of candidate OD's ($\hat{\mathfrak{L}}$) needs to have satisfy the following conditions.

1. $\hat{\mathfrak{L}}$ should satisfy criteria 9.
2. The major ODs in $\mathfrak{L}$ should be clusterable and $\hat{\mathfrak{L}}$ should be a close representation of the clusters.

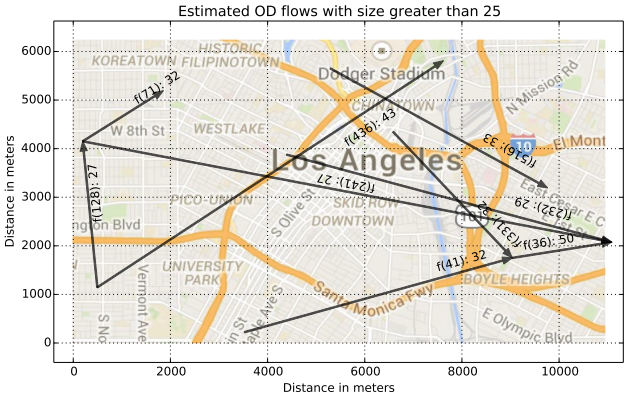
The first condition is what we discussed earlier: for a given set of OD candidate pairs the estimation result is unique if criteria in 9 holds. Otherwise there would be more than one answer to the system of linear equations and there is no way to tell them apart.

The second condition is observant of the effect that eliminating the rest of OD pairs can have on the resulting estimate. Let's assume that we have all the non-negligible OD pairs with their actual sizes. We can put different OD pairs into one cluster if their sources are spatially close together as well as their destinations. For example all the flows that start from a residential region of few blocks size and ending up in downtown area can be seen as one cluster. The underlying set of non-negligible OD pairs of a given urban area, $\mathfrak{L}$, is clusterable if we can identify clusters such that most of the large size OD pairs fall under one and only one cluster. Under such condition, each cluster can be represented by one OD pair. As a result, the resulting estimate of each candidate OD size will be the estimate of aggregate size of the ODs within the corresponding cluster. In fact, when we choose candidate nodes we are looking to form the most likely clusters that may exist for the underlying OD structure of the city.

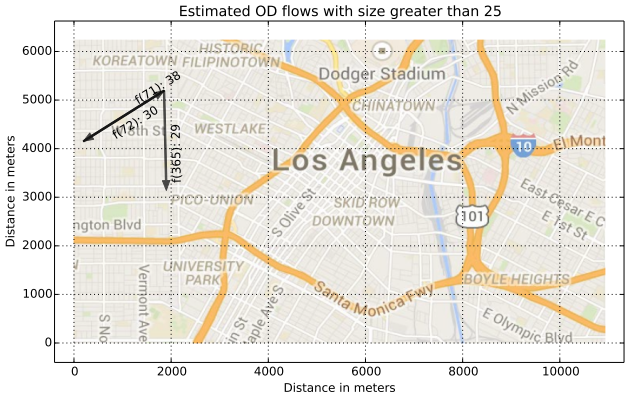
6.2 Applications

Estimating the underlying OD structure of a given urban area is the first building block for many different purposes. Here we identify some of the important applications that can make use of the OD matrix.

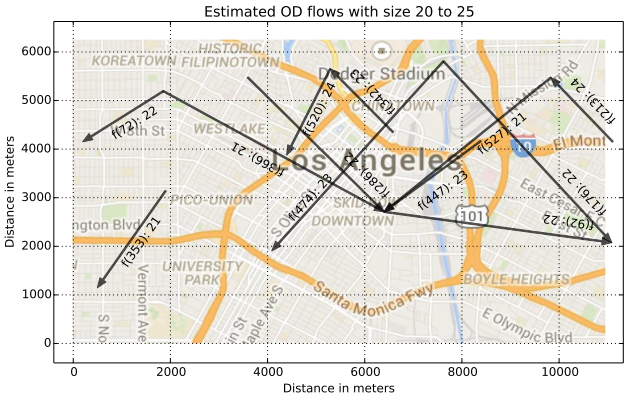
1. **Urban Planning:** Urban planning is a complicated topic with lots of different aspects[17]. The most important aspect of this broad topic is the optimum design and regulation of transportation network in a way that could handle the demands of urban commuters as well as meets the environmental standards[7]. We can not design an appropriate policy for a transit network



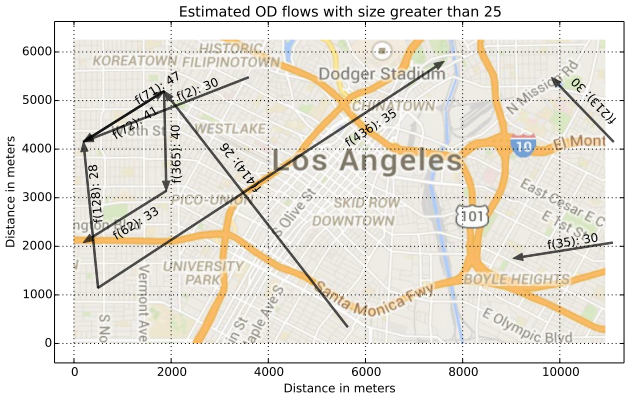
(a) 7:00 am, flow size > 25



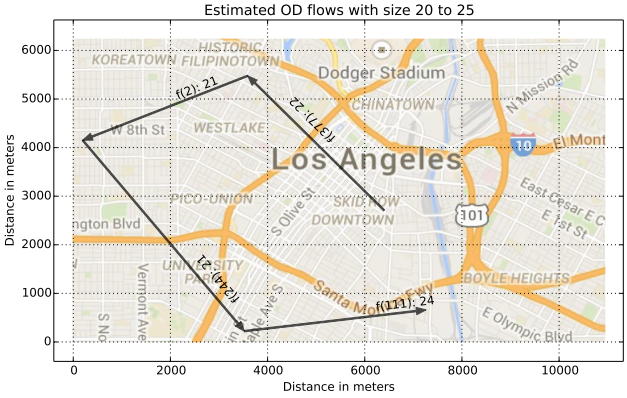
(b) 1:00 pm, flow size > 25



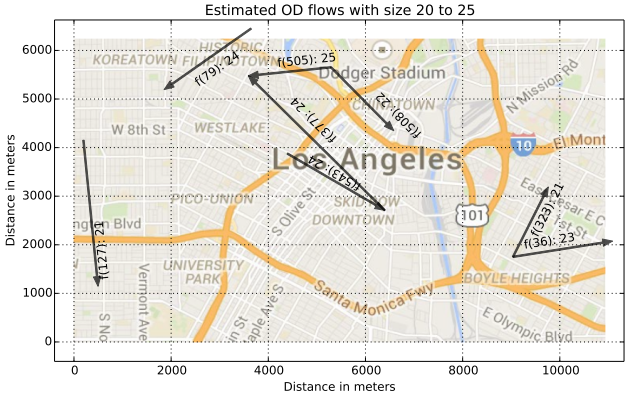
(c) 7:00 am, flow size: (20,25]



(d) 7:00 pm, flow size > 25

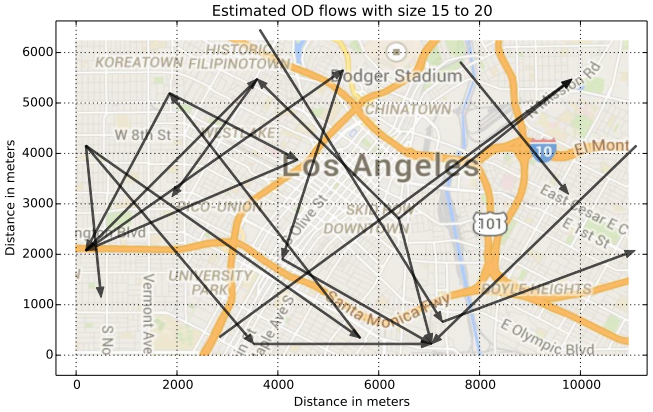


(e) 1:00 pm, flow size: (20,25]

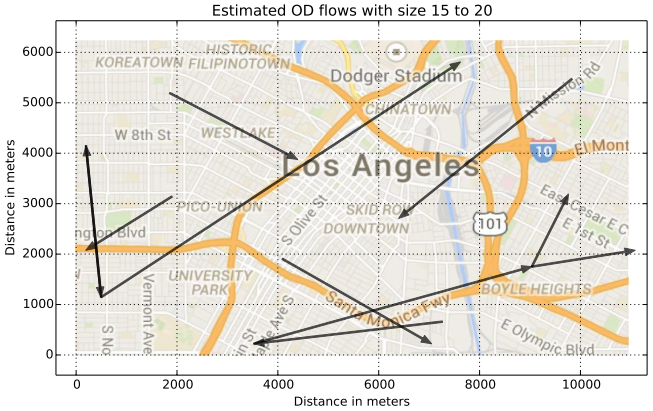


(f) 7:00 pm, flow size: (20,25]

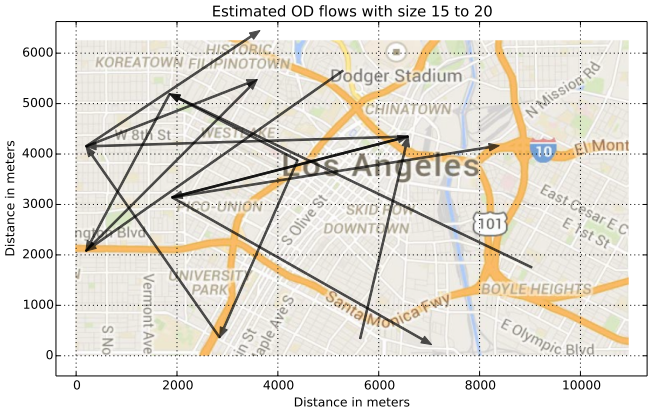
Figure 6: Estimated OD flows in different time periods of the same day.



(a) Morning, 7:00am



(b) Afternoon, 1:00pm



(c) Evening, 7:00pm

Figure 7: Estimated OD flows with sizes 15 to 20 in different time periods of the same day.

OD ID	Source	Destination	X_m	X_a	X_e
1	5499	2018	0	12	16
2	2018	5499	15	21	30
9	5504	2018	15	8	14
14	2018	5506	13	3	20
35	5502	1816	15	11	30
36	1816	5502	50	16	23
42	1816	5506	14	15	13
44	165	5499	10	7	16
62	317	5501	0	18	33
68	317	5505	11	7	16
72	4006	5499	22	30	41
75	5501	4006	10	6	9
85	5499	4519	0	7	12
95	5505	4519	12	9	8
97	5506	4519	11	13	10
102	1703	5500	9	11	9
106	1703	5502	17	13	12
112	1703	5506	0	19	10
119	5502	3858	14	8	11
128	4667	5499	27	17	28
137	5505	4667	7	15	15
195	5506	2747	0	0	13
196	2747	5506	13	0	0
202	635	5501	11	15	16
213	5500	3790	24	12	30
221	5505	3790	8	6	12
230	2498	5501	16	8	15
232	2498	5502	29	8	14
241	5499	5502	27	0	12
286	2018	4519	22	11	14
365	4006	317	12	29	40
543	2498	4519	15	13	24

Table 1: Estimated size of candidate OD pairs for three different times in a day of 05/28/2014.

if we don't have a good estimates of what this network need to handle. On the other hand, once we have the OD flows for a given urban area, we are aware of the vehicular traffic demands. We can plan to satisfy the demands with respect to different criteria that might be imposed because of some other objectives. It enables us to add or remove road segments into the road network. More dynamically, we will be able to manipulate speed limits, traffic signalling and stop signs to force people to take more desirable routs, so that the individual selfish behaviour results in communal benefits as well.

2. **Vehicular Network Data Delivery:** Vehicular Networks can be seen as an alternate medium for data delivery. Many have even considered it as a potential candidate for 5G wireless networks [11]. Having the OD flow sizes alongside with the road network map is essential to estimate the vehicle pairwise contact rate based on which we can plan and choose best strategy to achieve acceptable throughput of the system.

3. **Distributed Sensing:** There are many different environmental variables, such as noise, pollution, temperature and etc, that can be seen as a random field spread over large areas. The traditional way of measur-

ing these variables is through fixed stationary stations. However, because of the cost and implementation constrains of such stations, they can not provide a fine grain description of the objective field. This in turn, makes the vehicular networks a much better and more effective mediums for the purpose of urban monitoring[19]. However, the challenge is how best to make use of such complicated networks. In an ideal scenario we should be able to predict the future path of each vehicle so to decide how to combine and compress collected spatial data through the field without lots of overhead communication. And this is where OD matrix of the given area will become useful.

7. CONCLUSION

This work has presented a unique data set of traffic counts in LA for the first time. Based on this data set, we have shown how to extract useful information about traffic pattern in downtown LA, in the form of origin-destination traffic matrices for different time intervals. We have also used a validation technique that reveals interesting behaviour of urban commuters when it comes into route selection. Besides the algorithm we have used for the OD estimation, result evaluation, the details of the data set, cleaning and processing steps have been explained and we have also provided a discussion of the applications that can be supported by this set of results. In this work, we considered a fixed value for each time window (30 minutes). As a next step, we plan to evolve the modeling to a dynamic setting and try to capture how the OD pair flows are changing smoothly over time given the data. Also, it may be worthwhile to explore more sophisticated approaches that allow for more numerous OD pairs to provide a more fine-grained traffic matrix estimation.

8. REFERENCES

[1] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz. Sumo-simulation of urban mobility-an overview. In *SIMUL 2011, The Third International Conference on Advances in System Simulation*, pages 55–60, 2011.

[2] S. Bera and K. Rao. Estimation of origin-destination matrix from traffic counts: the state of the art. *European Transport Trasporti Europei*, 2011.

[3] G. C. Dandy and E. A. McBean. Variability of individual travel time components. *Journal of Transportation Engineering*, 110(3):340–356, 1984.

[4] M. Haklay and P. Weber. Openstreetmap: User-generated street maps. *Pervasive Computing, IEEE*, 7(4):12–18, 2008.

[5] J. Härri, F. Filali, C. Bonnet, and M. Fiore. Vanetmobisim: generating realistic mobility patterns for vanets. In *Proceedings of the 3rd international workshop on Vehicular ad hoc networks*, pages 96–97. ACM, 2006.

[6] M. L. Hazelton. Estimation of origin–destination matrices from link flows on uncongested networks. *Transportation Research Part B: Methodological*, 34(7):549–566, 2000.

[7] L. J. LeBlanc, E. K. Morlok, and W. P. Pierskalla. An efficient approach to solving the road network equilibrium traffic assignment problem. *Transportation Research*, 9(5):309–318, 1975.

[8] K. Moghadam, Q. Nguyen, B. Krishnamachari, and U. Demiryurek. Traffic matrix estimation from road sensor data: A case study (extended), *URL:http://anrg.usc.edu/www/papers/odest.pdf*, 2015.

[9] J. Munuzuri, J. Larraneta, L. Onieva, and P. Cortes. Estimation of an origin-destination matrix for urban freight transport. application to the city of seville. In *The 3rd International Conference on City Logistics*, 2004.

[10] S. Nguyen. Estimating origin destination matrices from observed flows. *Publication of: Elsevier Science Publishers BV*, 1984.

[11] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, et al. Scenarios for 5g mobile and wireless communications: the vision of the metis project. *Communications Magazine, IEEE*, 52(5):26–35, 2014.

[12] O. Z. Tamin, H. Hidayat, and A. K. Indriastuti. The development of maximum-entropy (me) and bayesian-inference (bi) estimation methods for calibrating transport demand models based on link volume information. In *Proceedings of the Eastern Asia Society for Transportation Studies*, volume 4, pages 630–647, 2003.

[13] S. Uppoor, O. Trullols-Cruces, M. Fiore, and J. Barcelo-Ordinas. Generation and analysis of a large-scale urban vehicular mobility dataset. *Mobile Computing, IEEE Transactions on*, 2013.

[14] N. J. van der Zipp and R. Hamerslag. Improved kalman filtering approach for estimating origin-destination matrices for freeway corridors. *Transportation Research Record*, (1443), 1994.

[15] H. J. Van Zuylen and L. G. Willumsen. The most likely trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological*, 14(3):281–293, 1980.

[16] Y. Vardi. Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, 91(433):365–377, 1996.

[17] V. R. Vuchic. *Urban transit: operations, planning, and economics*. 2005.

[18] Y. Wang, Y. Zheng, and Y. Xue. Travel time estimation of a path using sparse trajectories. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 25–34. ACM, 2014.

[19] X. Yu, H. Zhao, L. Zhang, S. Wu, B. Krishnamachari, and V. O. Li. Cooperative sensing and compression in vehicular sensor networks for urban monitoring. In *Communications (ICC), 2010 IEEE International Conference on*, pages 1–5. IEEE, 2010.

[20] W. Zhang, S. Li, and G. Pan. Mining the semantics of origin-destination flows using taxi traces. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 943–949. ACM, 2012.