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.

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, Routing Validation

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 the 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. 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.

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. 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.

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 route for each of the possible origin-destination pairs.

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 [13] and urban traffic planning [2, 10].

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. [10] use OD and underlying traffic road network as feeds to the SUMO [1] traffic generator to study traffic pattern for city of Köln in Germany. Another example is VanetMobiSim [4] which is a simulator concerned with both the traveling path choice of commuters (Macro-mobility) as well as the individual car interactions (Micro-mobility).

Traditionally, urban traffic OD matrix is estimated through individual household surveys on their commuting habits and residential commuting needs [10], 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 provides a prior belief about the general shape of the OD matrix. Munuzuri et al. [7] do OD estimation for city of Seville. To avoid the cost of data gathering through surveys, the authors construct their model on entropy maximization and algorithmic solutions such as Frank-Wolfe’s linear approximations. Authors in [14] use GPS data from taxis in Shanghai to infer the OD matrix. Other major branch of works in this context have used traffic counts [8, 2, 7] to get an estimate on OD matrix of the targeted region. Van Zuylen and Willumsen [11] 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 other 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 [12] 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 [9]. 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 [5]. However, as we explain later on, the basic model assumptions of such approaches do not necessarily hold for arterial traffic patterns.

3. PROBLEM MODEL AND SOLUTION

Given a certain geographical region, the road network lying in the area can be modelled as a graph \( G(V, E) \) where \( V \) is the set of nodes and \( E \) is the set of edges of the graph. The road network intersections will map onto 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 \([|V| \times |V|]\) with zero diagonal elements. Each element of the matrix, \( o_{ij} \), 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. 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 \( i \) that uses road segment \( j \).

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.

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 \( x \in \mathbb{R}^{n \times 1} \).

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 the average traffic speed \( s \) as part of our collected data and the road network details we can compute \( d \) which is the vector of link traverse time delays.

Next, we can build the routing matrix \( A \in \mathbb{R}^{n \times |E|} \) as described in details in[6]. An element \( a_{ij} \) represent the portion of flow of the OD indexed \( j^{th} \) in the set of all possible flows, \( E \), that is passing through road segment \( i \in E \) and will be equal to \( J_i \). There is a linear equation between the actual size of OD pairs and the flow of traffic on each link in terms of \( A \):

\[
Ax = b
\]  

In which, \( b \) is the data that sensors are reporting on traffic count on each link. If we consider to solve the equation (1) for the unknown possible OD pairs, it is an under determined system of linear equations i.e. \( A \) is a low rank matrix.

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. Here, 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 nominee for a 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. By narrowing down our search space, the new set of candidate ODs will be a subset of \( E \) and will be shown by \( \tilde{E} \). We also have \( \tilde{E} \) which is the union of the links that are used by \( \tilde{E} \). And hence equation 1 would change to \( A\tilde{x} = \tilde{b} \).

Where \( \tilde{A} \in \mathbb{R}_{\tilde{E}}^{n \times |\tilde{E}|} \) is a sub-matrix of \( A \) with those elements of \( A \) that have corresponding elements in \( \tilde{E} \times \tilde{E} \). Also \( \tilde{x} \in \mathbb{R}_{\tilde{E}}^{n \times 1} \) and \( \tilde{b} \in \mathbb{R}_{\tilde{E}}^{n \times 1} \) are sub-vectors of \( x \) and \( b \) respectively with the same logic.

The new equation 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.

$$\arg\min_x ||Ax - b||_2^2$$
subject to: $x \geq 0$ (2)

We should note that $A$ depends on drivers routing preferences. In fact we may end up with different results for different assumption of drivers routing preferences. To address these issues, we choose a cross-validation technique in which we randomly select 80% of road segments from $\hat{E}$, call the new set $\hat{E}_1$ and the remaining of the set $\hat{E}_2$. We reconstruct optimization 2 into another optimization problem such that it only comprises rows corresponding to $\hat{E}_1$. The solution to the new optimization formulation, $\hat{x}_1$ can be now evaluated with respect to the set of remaining road segment, $\hat{E}_2$.

The final optimization problem 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.

4. 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. This area includes 5498 intersections and 7584 distinct arterial road segments. The detailed information of the map including the road segment ID, name, length, bearing, start and end location and latitude and longitude are all extracted from Open Street Map (OSM) [3]. The LA Metro data are gathered through induction loop sensor in arterial road segments. The information in each sensor has been stored each minute.

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. The set of candidate nodes with their location with respect to the map is shown in figure 3. $\mathcal{L}$, set of all possible OD pairs, 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 $A$ 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 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 route. The split of traffic happens in a way that in general the paths with the shorter delay take 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 1 illustrates the results confirming this behaviour for all three different time slots. Figure 1 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. More information can be found in [6].

The full list of estimated traffic flows is accessible in the extended version of this paper [6]. The sizes are the estimated number of vehicles that are traveling from the corresponding source to destination per one minute. By looking into the resulted estimates different traffic demands can be revealed. To further assist interpreting the results we have visualized them on the map for different flow sizes and time window intervals.

The resulted medium size flows (between 15 to 20 vehicles per minute) have been shown with respect to the map in figure 2. Each arrow in the map corresponds to one flow, the arrow head indicates the destination and the tails indicates the origin of the flow. 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. More results and discussion on this matter can be found in [6].

5. 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. There are dif-
ferent directions in continuation of this work with promising outlook. An important direction is the zoning problem in which the set of candidate OD pairs can be inferred from the data itself rather than using meta-data that may not be always reliable. Also, 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. It may be worthwhile as well to explore more sophisticated approaches that allow for more numerous OD pairs to provide a more fine-grained traffic matrix estimation.

6. REFERENCES


