En Route: Towards Vehicular Mobility Scenario Generation at Scale

Roozbeh Ketabi, Babak Alipour, Ahmed Helmy
Computer and Information Science and Engineering
University of Florida, Gainesville, FL
Email: {roozbeh, babak.ap, helmy}@ufl.edu

***Abstract*—Vehicular mobility scenarios are utilized to study vehicular networks and transportation systems. However, the generation of vehicular simulation scenarios at scale poses several research challenges. Large-scale vehicular datasets (in geographic coverage and time span) are not easily or publicly available, which hinders the generation of data-driven scenarios. In this paper, we introduce a systematic method, called *En Route*, to generate vehicular mobility scenarios from traffic datasets such as one derived from thousands of available traffic webcams covering major cities around the world. Our framework includes data-driven components for estimation of traffic density, flow, road occupancy, as well as origin-destination (O/D) matrix estimation, trip generation, and route/navigation calculations. By applying the framework, we explore the city of London using the dataset of ≈100 traffic cameras throughout the city. We utilize available taxicab trips and traffic measurement datasets as guidelines for reasonable estimation of flow values, trip generation and O/D matrix. Our initial study shows reproducible step-by-step procedures, detailing parameter choices and settings, and measuring effects of changing those settings on the scenario outcomes. The results show a clear relation between flow and occupancy, and that trip duration follows a Lognormal distribution. Also, using traffic-aware routing (vs. shortest path) results in less congestion and more completed trips for a given simulation time. Queue distributions are obtained showing that over 90% of the intersection queues are 15 meters long (have average of 4 cars), and 90% of the roads carry less than 20 vehicles/km with average speed of ≈22.3 mph. Future studies shall provide multi-city simulations with further analysis.**

I. Introduction

Mobility modeling and simulation has been an active area of research for the past decade. It benefits the design and performance evaluation of existing and emerging wireless networks, mobile social networks and transportation, to name a few. Although there have been successful efforts to build libraries of pedestrian wireless users (e.g. crawdad.org, cise.ufl.edu/~helmy/MobiLib.htm), to aid data-driven mobility modeling, there have hardly been any large-scale libraries for vehicular traces. Much of the data for research in the transportation field is not publicly available for various reasons (privacy, industry practice, etc.). Hence, there exists a research challenge to develop simulation scenarios and vehicular mobility models, especially with reproducible results. One of our main goals is to generate scenarios to help investigate microscopic vehicular mobility. This fine level of granularity can facilitate the study and understanding of complex spatio-temporal characteristics of vehicular mobility.

This paper presents our first study to address this problem in a systematic way through our *En Route* framework. The framework includes elaborate steps, starting from the dataset, to origin-destination (OD) matrix and vehicle flow generation and finally simulation scenario generation and evaluation. We utilize several datasets, including planet-scale imagery data from thousands of webcams around the world, two datasets from taxicabs and a set of traffic measurements on highways.

By applying the framework to processed imagery data of the city of London (Oct. 11 2010, 10-10:30AM), considering multiple routing schemes, we investigate the effects of parameter alteration on number of trips and some initial measurements on outputs of the simulation. We observe that average trip time does not affect the number of trips started while increase in number of shortest paths results in fewer number of trips. Trip duration from simulation outputs also conforms to Lognormal distribution. ≈90% of roads are less than 10% occupied and have almost 20 vehicles/km. Intersection queues have average of 15 meters (per lane) in length for 90% of all lanes and in all cases traffic is heavier near cameras.

Our contributions in this work include: introduction of the En Route framework integrating the various pieces enabling generation of suitable scenarios and measurement system, study of system parameters and outputs based on a large scale scenario (London) and providing the scenarios for future studies and finally evaluating the scenario by the means of trip and traffic characteristics. We plan to extend our work in the future to other cities and longer times and share out datasets and generated scenarios and tools. The rest of the paper is organized as follows: Section II discusses related works. Section III presents the dataset and framework. Section IV provides system analysis, results and comparison with the original data. Section V concludes the study.

II. Related Work

The related work lies in three main areas; datasets for mobile vehicles, OD matrix estimation, and vehicular mobility simulations. In this section we briefly summarize the state-of-the-art in these areas and how they relate to this paper. Various methods of urban measurement are used to collect traffic information throughout cities and highways such as induction loop detectors data (including vehicle flow, speed and occupancy), traffic cameras (image data), radar and sonar sensors [1] but unfortunately not many complete and large scale datasets are available/accessible. Thakur *et al.* [2] in addition to providing us with a large scale dataset, have analyzed traffic patterns which facilitates understanding of

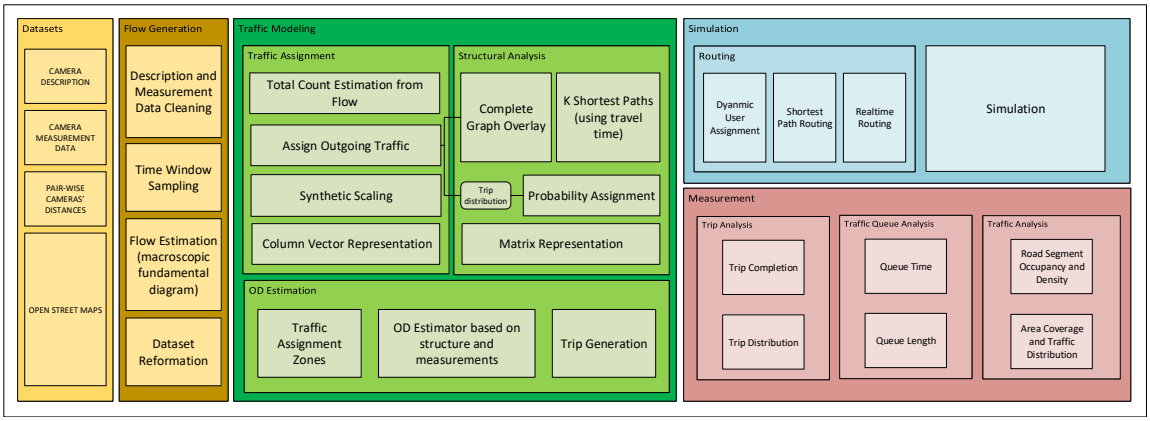


Fig. 1. Block diagram of *En Route* framework. The core component is the O/D matrix estimation works on the processed planet scale imagery data and provides the configuration for routing and simulation. Measurement system analyzes final simulation outputs and intermediate results.

vehicular mobility. Origin/Destination matrix has widely been used as a concept and a tool to represent and simulate vehicular traffic [3]. This information can be used in a routing algorithm (e.g. shortest path algorithms) to generate routes to be taken by vehicles traveling between each pair of origins and destinations. Generating (estimating) OD matrices is a well known problem in the transportation community [4][5][6]. Various studies have been conducted to model traffic demand especially using origin/destination matrices. Moghadam *et al.* [7] have designed a smart O/D estimation technique that is easily achievable using the state of the art programming languages and tools. Tools and ideas for generating mobility models are in demand in both transportation science and networking science communities. For instance, Karnadi *et al.* [8] have proposed MOVE, a mobility model generation architecture that can be used for VANET simulations. Noori has done multiple studies involving generation of scenarios (using SUMO [10] and VEINS framework) for various purposes considering VANETs and vehicle based communication including analysis of emergency vehicles response time and reducing travel time [9][11][12]. For this study, we rely on SUMO with 3 flavors of routing for running the simulations.

III. En Route

We propose *En Route*; a systematic framework (shown in fig. 1) capable of modeling traffic by means of an O/D matrix and generating simulation scenarios for mobility in urban environments. This framework has a modular design so changes to one module can be easily applied (as far as the interface between them is maintained). Main components of the framework are: dataset, traffic modeling, routing and simulation and finally measurement. We also incorporate datasets from Berkeley [13], San Francisco (SF) [14], and Beijing [15] to gain analytical insight for design and evaluation of the framework. In order to generate reasonable scenarios, we need valid routes which can be computed on the trips generated based on origin and destination (O/D) matrix as the tool to represent the traffic needs of the scenario. The framework, processes the data into the required format and then uses it to estimate the O/D matrix. Next we will describe each component in finer details.

A. Dataset

A dataset suitable for traffic modeling consists of description of observation points and their geo-coordinates, time to travel between each pair (potentially extracted using third party location/map software) and measurements of flow values (number of cars observed in some unit time). This information can help with modeling the structure of a city in addition to transforming and assigning the measurements in order to estimate the O/D matrix.

Planet scale imagery dataset provides traffic cameras and their geo-information, pairwise distances and travel times (estimated using Google Maps API) and timestamped values of density extracted using background subtraction algorithms [16]. These values if scaled to [0,1] for each camera, represent a measure for congestion in view point of the camera. Open Street Maps (OSM) augments the dataset by adding geographical information (i.e. nearest road segments) as well as providing the maps of cities under investigation for simulation. The original dataset consists of over 10 urban areas including several major cities around the globe (New York City, London, Sydney, etc.). For the purposes of this study we focus on the city of London, which contains $\approx$180 cameras.

Data requirements of the system demand values corresponding to total number of actual vehicles during the time window under consideration. We target the time window of 30 minutes as going to longer time windows results in less accuracy and going to shorter times results in too many small matrices (this level of detail is not required for a scenario at the scale of a city). Simulation of longer time periods is achievable by using multiple matrices for different (potentially contiguous) times.

Scaled pixel densities represent density of the traffic at corresponding camera's view. Occupancy refers to the amount of time an induction loop detector is active and represents a measure of traffic density. By assuming a linear relation between occupancy and pixel density, we can in turn relate the pixel densities to flow values. This is achievable based on

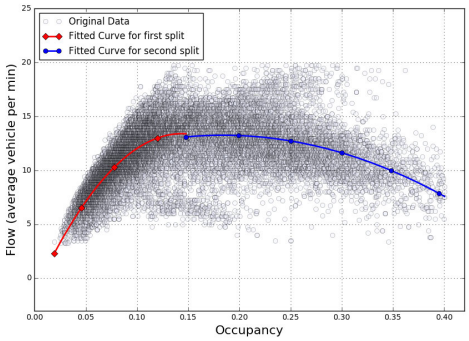


Fig. 2. Relationship between flow and occupancy as suggested by Macroscopic Fundamental Diagram

studies that have investigated the relationship between occupancy, speed and flow. Existence of Macroscopic Fundamental Diagram (MFD) [17] suggests how flow measurements change with changes in occupancy. Observations suggest existence of a maximal flow with increase in occupancy and then flow value tends to drop as the road becomes more occupied. We estimate the MFD parameters using two second-degree polynomials and associate density values to corresponding flow values. To observe and validate this conjecture, we studied the traffic measurements data of Berkeley [13] and put the estimated parameters as a basis to estimate flows of other cities. This parameter might not have an optimal value (depending on the application of the scenario and whether realism is favored) and requires further studies. To reduce the variance and noise in this data, we aggregated every 100 data points and replaced them with their mean. Then to determine the cutoff point, we used the occupancy at which the flow maximizes and fitted two second-degree polynomials for each split. Figure 2 presents the relationship between flow (average vehicle per minute) and occupancy as well as the two polynomials.

B. Traffic Modeling using an O/D Matrix

At the core of the system, lies the O/D matrix. A matrix cell $OD_{i,j}$ represents the number of cars that travel from origin i to destination j (for the time window that an OD matrix represents e.g. 30 minutes). Moghadam *et al.* [7] have proposed a method of estimating O/D matrices based on induction loop detector measurements that span a section of city of Los Angeles with origins and destinations chosen with user's knowledge of city. This ignited the initial idea for our work. In order to get a systematic estimate of the values corresponding to an O/D matrix, a matrix representation of the city structure (matrix A) and a vector containing the estimates of flow values are required (vector b).

1) Traffic Assignment: Flow of vehicles (number of vehicles in a time period, e.g. a minute) is a common metric to study traffic patterns. Summation of per minute flow values provides the total number of vehicles expected in the vicinity of cameras (for each camera). This total value, then will be assigned to outgoing edges on the overlay graph (explained in III-B2). Since the overlay graph is assumed to be complete, an edge exists from each point (camera) to all the others. The total value at a given point is assigned to outgoing edges

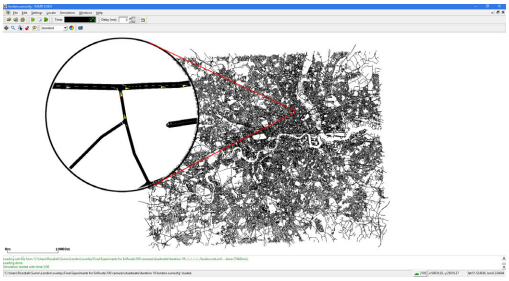
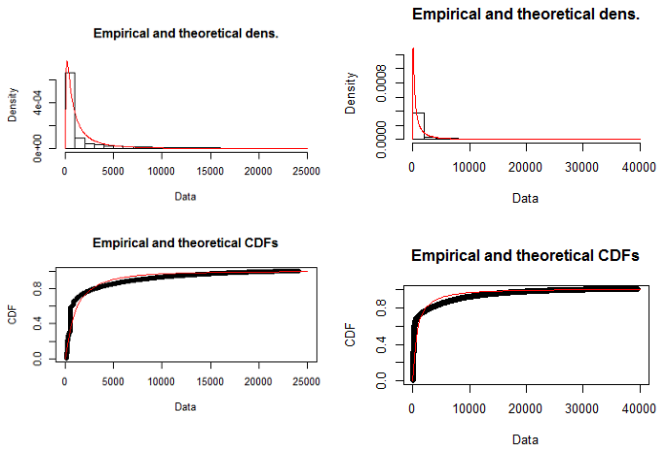


Fig. 3. Complete map of London imported in SUMO (more than 300K road segments) with a zoomed in sample of an intersection with cars.

proportional to the probability of a trip between the point and the other end of the edge. Probability is drawn from a Lognormal distribution for trip travel times (travel time between each pair of points). Using this information an $N \times N$ matrix representation of the graph with traffic count as weights (on the edges) can be formed and flattened out into a vector representation ($N^2 \times 1$) suitable for O/D estimation (vector b).

2) Structural Analysis: Using cameras' geo-coordinates, we can assume an overlay complete graph with cameras acting as vertices and edges weighted by the time it takes to drive between the nodes on each end of that edge. This graph helps us with estimating the value (number of trips) for each origin/destination pair as each edge in the graph also corresponds to an O/D pair. Lognormal distribution is the de facto distribution for modeling trip duration used in the literature. We verified that trip times fit a Lognormal distribution well by analyzing two publicly available datasets from Beijing [15] and San Francisco taxicabs [14]. To extract trips from GPS traces, we used a criteria of speed < 0.01 m/s, dwell time $>= 120$s and change in latitude and longitude $< .00005$ to detect arrival of a trip [18]. Figure 4 presents empirical and theoretical densities and CDFs based on Beijing and San Francisco taxicab trips against a Lognormal distribution, visualizing the goodness of fit. In order to establish a relation between O/D pairs and graph edges we compute the k-shortest paths between each pair of nodes in the overlay graph. We assign probabilities using Lognormal distribution by assuming for each path, the probability to take that path is proportional to the probability extracted using Lognormal distribution on travel time computed for that path. This information is represented in a matrix (called A) where cell $A_{i,j}$ is the probability to take ith edge on a trip between jth O/D pair. This ultimately helps identify which portion of the traffic passing through an edge is sourced and destined to vertices of that edge which in turn can translate to the value for the O/D pair.

3) OD Estimation: Estimating an OD matrix is achieved by solving a system of linear equation $Ax = b$ where A is a square matrix of size n (each matrix row represents an edge in the complete graph and each column represents an O/D pair) and the cell $A_{i,j}$ represents the probability of the ith edge being used for jth O/D pair, and b is the vector of flow values and x will be our O/D matrix (flatted out in a n^2 by 1 vector instead of n by n matrix). Although the problem is determined in this case, we need non-negative solutions (negative number of



(a) (b)

Fig. 4. Distribution of Beijing (a) and San Francisco (b) taxicab trips fit Lognormal

trips is undefined) and therefore the system is solved such that it minimizes the least square error by assigning non-negative values (non-negative least squares approach).

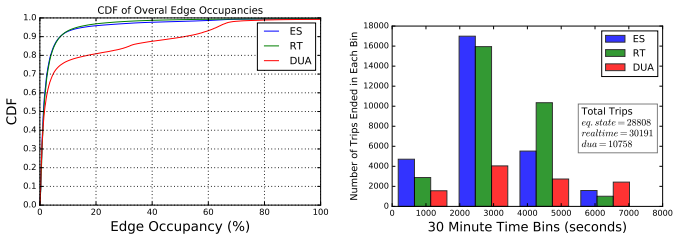
With OD matrix available, we are able to generate trips by randomly assigning departure times within the OD time window (e.g. 30 minutes) between each origin and each destination. Each origin or destination is called a Traffic Assignment Zone (TAZ) and will consist of three closest edges to the camera's coordinates (more than 3 edges can be used if more spread of traffic at assignment zones is preferred).

C. Simulation

To run a scenario, generated trips based on the O/D matrix should be routed on the map of the city under consideration. A minimal set of required configurations to run a scenario using SUMO v0.28 is a route and a network configuration (map of the city). Network configuration may be acquired using Open Street Maps and converted into SUMO compatible format.

D. Setup

1) Route Generation: A simple shortest path in time (i.e. longer road with higher speed limit may be faster) is usually the method people utilize to reach a destination. More recently, with advances of navigation tools and almost real-time availability of traffic information, people can also find the fastest route to their destination considering the traffic. We consider three different routing schemes: shortest path (1 iteration of dynamic user assignment-DUA), 10 iterations of DUA toward an equilibrium state (ES) and traffic aware real-time routing (referred to as RT). ES and RT provide better routes in terms of trip completion times and heuristically match how people navigate nowadays. According to figure 5a using ES and RT results in similar performance while shortest path assignment tends to put more traffic on fewer road segments (hence more over-occupancy). Meanwhile, an interesting observation from figure 5b is the difference between the number of trips completed within two hours of simulation when ES or RT is used and the number of completed trips using only simple shortest path routing (near 30K trips completed vs around 10K). Also



(a) (b)

Fig. 5. Measurement from city of London using 100 cameras. a) Cumulative distribution of occupancy values reported based on one minute measurements. b) Multi-histogram of trips completed within each 30 minutes of the simulation. While ES and RT seem to perform similarly, DUA produces more congested roads and as a consequence, less trips finished within 2 hours.

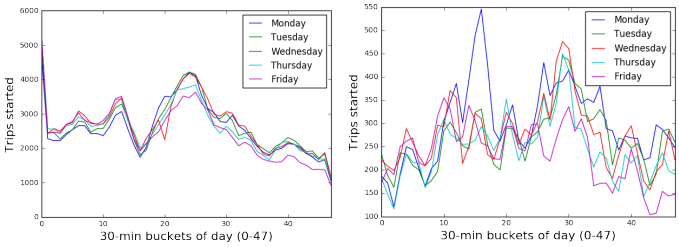
using 10 iterations of dynamic user assignment most of the trips (over 58%) are completed within the second 30 minutes of the simulation which conforms to the expected trip duration of 10 to 20 minutes (depending on city characteristics) and start times within the first 30 minutes. From computation point of view, using 100 traffic assignment zones, ES took slightly over 20 hours while using RT took even longer time of over 30 hours for 30K trips in 2 hour total simulation time.

IV. Evaluation

The system is proposed with flexibility to provide diverse scenarios in mind. We base our evaluation on comparing simulation results to the original dataset along with investigating the relationship of system parameters and outputs.

1) Parameters: Although not an exhaustive list of parameters, following have been identified as knobs that can be tweaked to achieve a variety of scenarios with reasonable (qualitatively/realistically feasible) outputs: *Lognormal mean and standard deviation* as trip duration distribution for assigning probabilities to edges of overlay graph, *Number and choice of shortest paths* used on overlay graph, *Number of the cameras* in the subset of total cameras used as origins and destinations, *synthetic scaling factor* to increase the number of trips to generate and *Flow estimation parameters* (min and max occupancy and their corresponding flow values in addition to maximum flow value and its corresponding occupancy).

2) Outputs: Trips act as the connecting agent between the previously mentioned system parameters and the following measurements. Number of trips generated can be seen as the output of traffic modeling and as input to the routing module of the simulator. In addition to scenario configuration file and computed routes (scenarios as the product), various supported outputs may be acquired from the simulator (as product of the scenarios). We focus on measurements of *edge* based information, edge/lane *queue* statistics and *trips* information. We investigate the relationship among inputs, tweak points and outputs in London. Using 20, 50, 100 and 150 cameras resulted in synthetic scaling factor of approximately 5.5, 3.5, 2.5 and 2.1 respectively. This drop in synthetic scaling factor demonstrates the convergences to more realistic scenarios by using more observation points.



(a) (b)

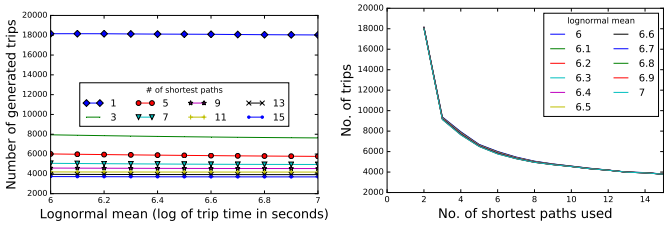
Fig. 6. Trips started in 30-minute windows of a day for Beijing (a) and San Francisco (b)

The milestone of 30K trips (*starting* in 30 minutes) is chosen as such to stress the system. In order to have a realistic guideline, we studied Beijing and San Francisco datasets. Figure 6 represents the time series for the number of trips started in 30-minutes windows of a day for several weekdays. The lower number of trips started in SF can be explained by the lower number of taxicabs participating in data collection. Nonetheless, There are visible hikes around early morning and afternoon hours, which could be explained by work commutes (and can be used to generate target number of trips for specific times of day). If realism is favored for a scenario, desired number of trips chosen by further study of the urban environment can be set as the milestone for the system.

Figure 7 visualizes the relationship between number of shortest paths used and mean for Lognormal distribution for probability assignment, and number of generated trips. Changing mean parameter from 6 to 7 has negligible effect whereas using fewer paths produces higher number of trips. A suggested number of paths to consider would be between 3 and 5 so that a large span of city is covered and at the same time the rate of decrease in number of trips also drops. Smaller number of paths will result in heavier traffic jams. To further distribute the traffic, instead of the first few shortest paths, those that have more uncommon edges or highly varying travel times may be used (choice of path).

To compare the outputs of the simulation versus the original dataset, various situations considering the choice of result and scaling is shown in fig. 8. Resulting occupancies reported as one minute measurements on all road segments used throughout the simulation, are generally small (90% of values are less than ≈0.1) while for original dataset values are considerably more (90% of values are less than ≈0.5). This is somewhat expected as the cameras are usually placed in places with heavier traffic. CDF of results near the cameras endorses this conjecture (as 90% of values are less than ≈0.25 vs 0.1 for all roads). Results, when scaled to range [0,1], show a closer behavior to original dataset. Occupancy of Berkeley highways is also plotted to give a basis to compare our urban scenario vs a highway one (highways tend to have more stable occupancy). Another observation is the similarity of results near cameras and overall when rescaled which corroborates that cameras are a good representative of the overall traffic.

3) Measurement Results: As a sample of the outputs of the generated scenarios, various plots on length and time of



(a) (b)

Fig. 7. (a) Almost no change in number of trips with varying lognormal mean parameter from 6 to 7. (b) Number of generated trips declines using more shortest paths. Change in Lognormal mean does not affect the behavior.

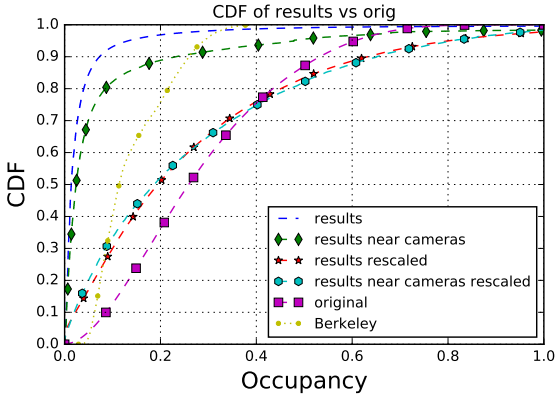


Fig. 8. Cumulative distribution of occupancy values. Berkeley corresponds to the highways near the city of Berkeley, CA and is drawn for comparison. Road segments near the cameras are more occupied but rescaled to [0,1] they closely represent the overall result and are more similar to the original dataset.

queues, edge based density (vehicles/km), traffic, speed and trip duration distribution are presented and discussed in 9:

- 90% of the queues at intersections have lengths of less than ≈15 meters which take ≈20 seconds to empty up.
- 90% of edges have 20 vehicles per kilometer or less overall. In vicinity of the cameras this value becomes 60 vehicles/km (cameras define the traffic assignment zones and therefore heavier traffic is expected close to their location). 90% of edges have ≈10 vehicles (≈18 for vicinity of cameras). Distribution of vehicles near the cameras seems to follow an exponential trend. This can be particularly interesting for vehicular mobility modeling and is due for in-depth analysis. Reported values for speed (average speed per minute for each edge) seems to be mostly distributed around a mean value of ≈10 m/s (36 km/h or 22.3 mph) which is heuristically reasonable for an urban scenario (considering city map constraints).
- Trip duration is observed to follow a Lognormal distribution. More than 50% of trips finish within the second 30 minutes (≈66% of trips are completed in the first hour), in line with the average trip time of ≈2000 seconds (> 30 min).

V. Conclusion and Future Plans

We presented a novel scenario generation framework that provides an O/D matrix compatible with common tools and simulators. These scenarios can be adjusted for purposes of different studies whether to better understand reality or to test an idea or model. Generated scenarios can be further

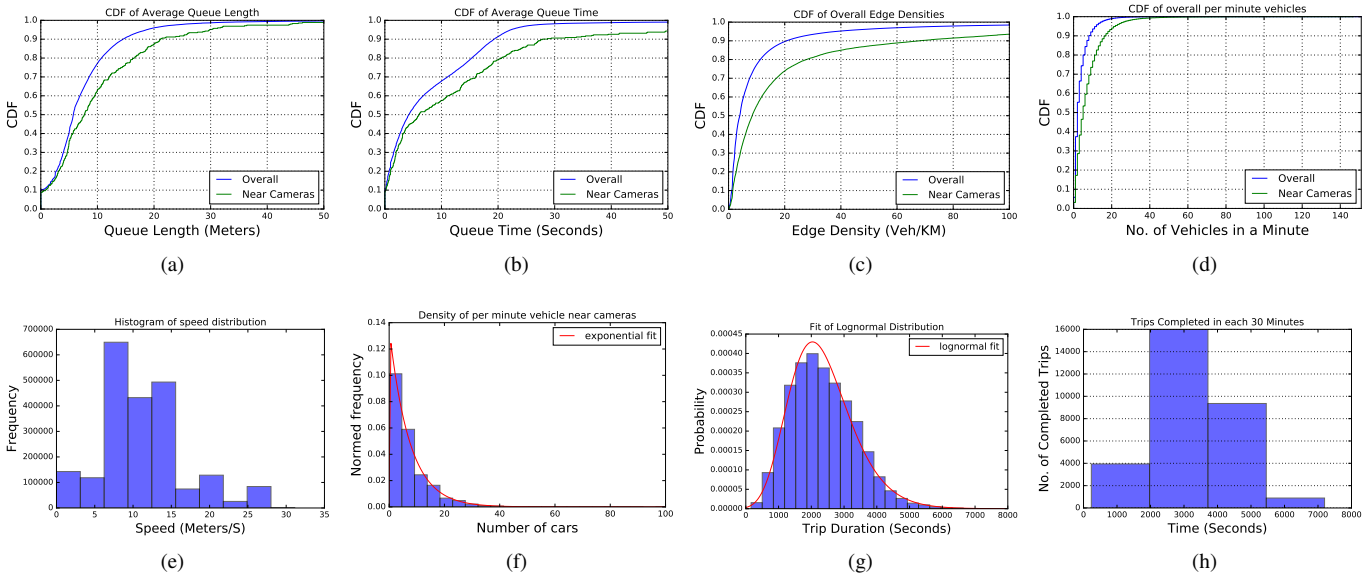


Fig. 9. Visualization of various measurements from simulation outputs. (a) and (b) explain the length and time of intersection queues. (c) and (d) focus on overall placement of vehicles over edges (roads). (e) shows the distribution of average per minute speeds and (f) demonstrate the distribution of vehicles near cameras. (g) presents the distribution plot of trip duration with a Lognormal fit and (h) shows the number of trips completed in 30 min. bins.

studied for self-similarity, causality, time series analysis or more complex statistics in order to understand not only the quality of scenarios but also the dynamics of real world urban mobility. Other cities available in the imagery dataset, or any other compatible dataset, can be studied to generate scenarios covering vehicular mobility of those cities. More in-depth parameter analysis bears further research. In order to model the behavior of vehicular mobility considering individual, mutual and social aspects, a dataset that contains detailed movement of individual vehicles is required. Vehicular mobility seems more challenging compared with pedestrians as not only the locations that a car visits is important but also how one travels between them matters (routes and duration). We plan to continue this work using generated scenarios towards proposing a framework to model mobility of vehicles that can help with benchmarking a vehicular/mobile networking system.

Acknowledgment

Authors would like to thank Gautam Thakur, Hamed Noori, Kayvan R Moghadam and Bhaskar Krishnamachari for their work and guidance that gave us a valuable insight throughout this study. This project was funded through NSF project 1320694.

References

[1] L. A. Klein *et al.*, "Traffic detector handbook," Tech. Rep., 2006.

[2] G. S. Thakur, P. Hui, H. Ketabdar, and A. Helmy, "Spatial and temporal analysis of planet scale vehicular imagery data," in *ICDM 2011*, IEEE.

[3] J. d. D. Ortúzar and L. Willumsen, *Modelling transport*. Wiley, 1995.

[4] K. Ashok *et al.*, "Dynamic origin-destination matrix estimation and prediction for real-time traffic management systems," in *International Symposium on the Theory of Traffic Flow and Transportation*, 1993.

[5] H. Yang *et al.*, "Estimation of origin-destination matrices from link traffic counts on congested networks," *Transportation Research Part B: Methodological*, vol. 26, 1992.

[6] M. G. Bell, "The estimation of an origin-destination matrix from traffic counts," *Transportation Science*, vol. 17, 1983.

[7] K. R. Moghadam *et al.*, "Traffic matrix estimation from road sensor data: A case study," in *Proceedings of SIGSPATIAL 2015*, ACM.

[8] F. K. Karnadi *et al.*, "Rapid generation of realistic mobility models for vanet," in *WCNC 2007*, IEEE.

[9] H. Noori, "Realistic urban traffic simulation as vehicular ad-hoc network (vanet) via veins framework," in *FRUCT 2012*.

[10] M. Behrisch *et al.*, "Sumo–simulation of urban mobility: An overview," in *Proceedings of SIMUL 2011*, ThinkMind.

[11] H. Noori and M. Valkama, "Impact of vanet-based v2x communication using ieee 802.11 p on reducing vehicles traveling time in realistic large scale urban area," in *ICCVE 2013*, IEEE.

[12] H. Noori, "Modeling the impact of vanet-enabled traffic lights control on the response time of emergency vehicles in realistic large-scale urban area," in *ICC 2013*.

[13] J. C. Herrera *et al.*, "Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment," *Transportation Research Part C: Emerging Technologies*, vol. 18, 2010.

[14] M. Piorkowski *et al.*, "A Parsimonious Model of Mobile Partitioned Networks with Clustering," in *COMSNETS 2009*.

[15] W. Zhang *et al.*, "Mobile sensing in metropolitan area: Case study in beijing," in *Ubicomp 2011*.

[16] G. S. Thakur, P. Hui, and A. Helmy, "Modeling and characterization of urban vehicular mobility using web cameras," in *Infocom Computer Communications Workshop 2012*, IEEE.

[17] N. Geroliminis and C. F. Daganzo, "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings," *Transportation Research Part B: Methodological*, vol. 42, 2008.

[18] L. Gong *et al.*, "Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies," *Procedia - Social and Behavioral Sciences*, vol. 138, 2014.