Urban Traffic Simulation: Network and Demand Representation Impacts on Congestion Metrics

Aaron Faltesek
Southern Methodist University, aaron.faltesek@gmail.com

Balasubramaniam Dakshinamoorthi
Southern Methodist University, bdakshinamoorthi@smu.edu

Sreeni Prabhala
Southern Methodist University, sprabhala@smu.edu

Akbar Thobani
Southern Methodist University, athobani@smu.edu

Anu Kuncheria
University of California, Berkeley LBNL, anu_kuncheria@berkeley.edu

See next page for additional authors

Follow this and additional works at: https://scholar.smu.edu/datasciencereview

Part of the Data Science Commons, and the Urban Studies and Planning Commons

Recommended Citation
Faltesek, Aaron; Dakshinamoorthi, Balasubramaniam; Prabhala, Sreeni; Thobani, Akbar; Kuncheria, Anu; and Macfarlane, Jane (2021) "Urban Traffic Simulation: Network and Demand Representation Impacts on Congestion Metrics," SMU Data Science Review: Vol. 5: No. 2, Article 8.
Available at: https://scholar.smu.edu/datasciencereview/vol5/iss2/8

This Article is brought to you for free and open access by SMU Scholar. It has been accepted for inclusion in SMU Data Science Review by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
Urban Traffic Simulation: Network and Demand Representation Impacts on Congestion Metrics

Authors
Aaron Faltesek, Balasubramaniam Dakshinamoorthi, Sreeni Prabhala, Akbar Thobani, Anu Kuncheria, and Jane Macfarlane

This article is available in SMU Data Science Review: https://scholar.smu.edu/datasciencereview/vol5/iss2/8
Urban Traffic Simulation: Network and Demand Representation
Impacts on Congestion Metrics

Balasubramaniam Dakshinamoorthi 1, Sreeni Prabhala 1, Akbar Thobani 1, Aaron Faltesek 1,
Anu Kuncheria 2, and Jane Macfarlane 2

1 Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
2 UC Berkeley, Institute of Transportation Studies,
Lawrence Berkeley National Laboratory
107C McLaughlin Hall, Berkeley CA 94720-1720
{bdakshinamoorthi, sprabhala, athobani, afaltesek}@smu.edu
{anu_kuncheria, janemacfarlane}@berkeley.edu

Abstract.—Traffic simulations are often used by city planners as a basis for predicting the impact of policies, plans, and operations. The complexities underpinning traffic simulations are often not described in detail yet can significantly impact the simulation outcome. Conflating underlying data for simulations is complex and hinders the interest in this type of exploration. This paper aims to elucidate critical features of traffic simulations that drive the generated metrics of the modeled urban environment. Specifically, this paper examines differences in two road graph networks for the metropolitan region of Houston, TX: a reduced network composed of 45,675 road links and an expanded network consisting of 729,753 road links. This paper will also cover collecting, refining, the feature extracting, and mapping matching real-world data to the simulated data. The traffic dynamics are generated by a simulator called Mobiliti. Two scenarios are explored: a baseline shortest travel time with 50% of the vehicles enabled to dynamically route to reduce travel time (B50), and a User Equilibrium Travel time (UET) scenario that results from a quasi-dynamic traffic assignment optimization. The resultant dynamics of these routing algorithms generate speeds and flows on the road graph links. The demand model trips are characterized by key features like travel times, delay times, and vehicle miles traveled. Validation with real-world data is presented using open-source Texas Department of Transportation data. The validation results of the various simulations provided evidence that the expanded network resulted in a more accurate simulation.

1 Introduction

Advanced traffic management is a key area of focus for city planners to bring about traffic flow efficiencies into their planning and operations. They often use traffic simulations to predict the results of their policies, plans, and procedures. As urbanization increases, there has been added pressure on existing transportation infrastructure. This increase in urban population has led to congestion, leading to increased pollution, wasted resources, and a reduction in aggregate productivity. New modes of traffic, such as autonomous vehicles, ride-share services, drones, and hyperloop tunnels, add additional complexity to city planners’ current traffic simulation.
and forecast processes. Frequently these traffic simulations are based on static, non-real-time data, which results in inaccurate simulations. This has resulted in a drive towards real-time data collection with dynamic traffic rerouting models, which help city planners to predict traffic demand and congestion accurately. These models improve traditional traffic planning systems that are too slow to adapt to new changes. In addition, dynamic routing models help develop cost-effective routes that reduce congestion and traffic delays [15][16].

The efficiency of traffic routing models is measured using various metrics like Vehicle Miles Traveled (VMT), Average Daily Traffic (ADT), congestion delay, fuel consumption, productivity consumption, and Volume-to-Capacity (V/C) ratio. VMT is considered one of the key metrics to estimate forecasts - it indicates travel demands and traffic behavior for city planners [12]. Transportation agencies use it to perform various analyses, including assigning transportation resources, computing fuel consumption, and assessing traffic patterns and impacts.

The traffic routing models require collecting large amounts of real-time data of various kinds, including data from mobile platforms and applications. Most of the time, such data is owned by private companies, resulting in challenges with the data collection process. Data privacy rules and regulations also hinder mobility data collection. Sometimes real-world data for specific scenarios may not readily be available to the planners. E.g., if a city is attempting to plan for a natural disaster in the future, there will not be real-world data to leverage for this analysis. To circumvent these issues, city planners have created data using simulations that produce synthetic data that can be compared against actual data using different algorithms. These simulations can run through various scenarios and produce data for them that in turn informs city planners. One issue with such simulations is that they often rely on static datasets often collected from census data, land usage, surveys, etc. This data collection method may result in slowness to pick up on new trends like the decrease in public transportation usage and lack of infrastructure funding. By introducing dynamic datasets, a simulation can expand on the number of scenarios provided to city planners and respond more quickly to a changing environment. The traffic simulation solutions often use algorithms to optimize the flow through the traffic network. The Frank–Wolf algorithm is one such optimization algorithm used very widely to address traffic optimization.

The traffic simulations are measured by performance time and accuracy when compared to real-world scenarios. The performance of the simulation is related to how quickly a simulator can produce predicted traffic for a given geographic area and timeline. A simulator that can produce results in the shortest time possible enables more extensive and more varied simulations. A second method to measure simulations is by validating the simulation results against real-world data using various traffic metrics discussed above (VMT, ADT, etc.).

Running traffic simulations for various real-world scenarios requires simulation software programs that are faster, powerful, and scalable, as well as high-end computing resources that can support such programs. In addition, city planners will need to have an accurate representation of the traffic network and demand models to come up with effective simulations. Due to various challenges associated with getting access to systems and resources and collecting the data, city planners often utilize reduced networks to run the simulations quickly and analyze the results to help them plan their policies and operations.
This paper will analyze and examine the differences between traffic flow simulations run using reduced and expanded road network topologies for the Houston traffic district. For these two network configurations, the analysis will focus on comparing two different dynamic routing simulations - baseline with 50% dynamic routing and User Equilibrium Travel time (UET) optimization, to assess which one results in better traffic dynamics for the region. These scenarios are analyzed for efficiency based on traffic metrics, as well as the feasibility of implementation. This paper will also discuss which of these simulations are more realistic through an assessment against real-world traffic metrics for the region.

2 Literature Research

Traffic congestion harms the productivity, employment, and economic growth of urban areas [12]. City planners continuously strive to understand traffic patterns to implement strategies to optimize infrastructure and controls to manage traffic congestion better. Traffic simulations are a valuable tool that supports this analysis. Urban traffic congestion can be studied at different levels of scale. It can be reviewed at the microscopic scale, which focuses on individual vehicles' mobility [14]. It can also be studied at the mesoscale, which analyzes a group of vehicles over a set distance. A third way is to research congestion at the macroscopic level, which is focused on the complete road flow and general traffic density [13]. This paper will perform analysis at the mesoscopic and macroscopic levels. There are additional levels of analysis on traffic congestion, such as Stay Point Analysis. This type of analysis involves leveraging fixed-point sensors, such as cell phone towers, to model traffic congestion [6]. There is also activity-based pattern analysis [3], which helps identify the drivers' behavioral patterns and provides key insights to the city planners and law enforcement. It is essential to review these different types of analysis techniques, even though this paper will use all of them.

Some studies examine traffic dynamics in various simulations by changing the volume of the traffic. By changing this demand model but holding road capacity static, these studies can examine the vulnerability of the existing network [8]. This paper will do the opposite and have the demand model static but alter the underlying road network. Instead of examining the road network capacity at the road level, this paper will discuss the road network capacity by comparing a reduced vs. expanded network. Additionally, "[recent] studies learn about human behavior in cities by using data collected from location-aware technologies - they infer preferences in travel decisions that are needed to calibrate existing choice modeling frameworks" [9]. The advent of location positioning technologies has created tremendous amounts of data and digital footprints. Several studies have leveraged these data to represent commuter patterns correctly. It has also been inferred that U.S. Census data and data detected by GPS are highly correlated, which could also be used to develop better trip models [4]. These are a few ways to model demand, but alternative demand models and user behavior/decision-making preferences will be discussed in separate sections for this paper.
The constant increase in commuters and decrease in public transit travel due to a lack of infrastructure and funding have exacerbated the traffic congestion issues worldwide. The yearly costs of congestion have doubled between 2007 and 2013. For example, cars are traveling 15 percent slower in Manhattan than five years ago [2]. As per the World Bank Group, "Cities are home to more than half of the world population, and they are expected to add another 2.5 billion new residents by 2050." [22]. This combination of growing cities and increased congestion underlines the importance of accurate traffic modeling by city planners.

It is important to understand the scope and impact of traffic in the U.S. as this study will focus on the U.S. city of Houston. The U.S. Department of Energy outlines that each year "the transportation sector accounts for ~30% of total U.S. energy needs and 70% of U.S. petroleum consumption" [7]. In this paper, it will be important to analyze traffic metrics by fuel consumption and travel time and miles traveled. The fuel consumption will also touch on ethical issues when it comes to city planning and its impacts on traffic dynamics.

2.1 Urban Traffic Simulation

This work leverages a simulation capability provided by Lawrence Berkeley National Laboratory (LBNL) called Mobiliti [5]. LBNL is a Department of Energy national laboratory that researches urban traffic modeling. Mobiliti is a parallel discrete event simulation for large-scale urban traffic networks that run on distributed memory parallel computing platforms at the supercomputer center at LBNL. The simulation models the effects of link congestion, timing constraints, and road network storage capacity constraints. Because of the parallel computing implementation, Mobiliti can simulate one model day of the San Francisco Bay Area with 19 million vehicle trips with a road network composed of 0.5 million nodes and 1 million links in a matter of minutes.

"The Mobiliti Platform provides a mechanism to generate a variety of scenarios from three underlying datasets: a description of the road network in terms of the network graph (nodes and links) and link attributes; a demand model that describes the trips that load the network; and fuel maps that characterize fuel use for a vehicle that is associated with each trip." [5]. Figure 1 details the flow for generating scenario descriptions. Mobiliti causes the network loading from the demand as defined by a specified approach to vehicle routing and captures the consequent metrics in terms of speeds and flows for every 15 minutes in the simulated day. Demand is represented by trip legs and characterized by trip times, delays, and energy consumption. Both reflect the network loading that is a function of the routing strategy.
Similar systems to Mobiliti include BEAM and MANTA. BEAM extends MATSim, a Multi-Agent Transportation Simulation framework. This paper will use a network in the MATSim format for compatibility with the Mobiliti simulation platform. MANTA, Microsimulation Analysis for Network Traffic Assignment, leverages GPUs (graphics processing units) to run simulations quickly [24]. Both the systems use reduced travel demand models and reduced road networks to reduce the computational complexity.

3 Methods

3.1 Data

Simulated Traffic Data
This paper will leverage the Mobiliti simulation made available by Lawrence Berkeley National Laboratory (LBNL). Mobiliti can use a variety of methods for routing the vehicle in the simulation. This paper chose two types of simulations for comparison:
- User Equilibrium Travel Time (UET): UET is a quasi-dynamic traffic assignment method [5] for assigning routes to the simulated vehicles. It is founded on the Frank-Wolfe algorithm, which optimizes each vehicle's travel time over a specified time. Mobiliti derives the optimized routes and then assigns the routes to each vehicle during the initialization of the simulation [25].
- Baseline-50: In this case, Mobiliti takes a random sample of 50 percent of the vehicles and allows them to reroute from an initial shortest path route. These vehicles will adjust their routes to find shorter travel times if they encounter significant congestion. This behavior is similar to the effects of a driver using a navigation application.

Houston Road Networks
As mentioned in the literature review, many simulation efforts will reduce the network and/or the travel demand to reduce the computational complexity and get faster turnarounds for running the simulation. This analysis aims to understand the consequences of reducing the network and demand to determine if the resulting metrics
align with real-world data. Due to time constraints, only network reduction was evaluated. Future work will consider reductions in the demand profile.

An external organization in Texas provided a reduced road network for the Houston area. The boundaries defined the area: the City of Conroe in the North, Mont Belvieu in the East, Freeport in the South, and Sealy in the West. The reduced network encompassed this geographic area but lacked many reduced-speed roads, such as the residential roads and tertiary roads. The network is comprised of 45,697 links and 22,943 nodes.

A complete road network was produced using Open Street Map (OSM). The process for generating a Mobiliti compliant network from OSM is described in Section 3.2. Data retrieved from the OSM file for the Houston traffic district contains 754,376 ways/links. The challenge with using open-data sources is the quality of the data can vary greatly. In this case, the OSM network provides lane information for only 131,325 of the ways/links. However, the simulation requires the number of lanes on each link to estimate the capacity and link dynamics. A random forest model was created to resolve this issue that used the links with lane specifications as a training set for estimating the number of lanes for lanes that didn't include the number of lanes.

Both the networks were used as input to the Mobiliti simulation. The network dynamics can be compared using various traffic metrics and validation against the field data.

Regional Travel Demand Model
Demand models define a set of trip legs by defining an origin node and a destination node along with the starting time for the trip. The external organization also provided the travel demand model used for this analysis that provided the reduced network. It included over 19M trip legs, specified as origin node, destination node, and time of day that the trip leg started.

Real-World Traffic Data
Finally, real-world traffic data was collected for the greater Houston area using open-source TXDOT information. This data was visually aligned with the reduced and expanded networks and was used to validate the simulations' results.

3.2 Creating the Expanded Road Network for Houston

The initial challenge was to build a complete network for the Houston region for comparison against the reduced network. This network needed to be extensive and available for use in this research paper. OpenStreetMap (OSM) is an open-source map that is freely available for public use. "OpenStreetMap is a map of the world, created by people like you and free to use under an open license." [19].

Extracting Houston data from the larger OSM map required the use of the BBBike extract tool [26]. Using this tool, the road network details for an expanded area around Houston were extracted. The OSM map is extracted as an OSM (.osm) file, an XML representation of the map. This XML representation contains three types of entities - nodes, ways, and relations. These nodes, ways, and relations detail many geographic objects, many of which are not relevant for traffic simulation.
In the face of memory and compute constraints, this initial OSM file for Houston must be reduced to nodes, ways, and relations that define just the road network accessible by vehicles. Additional data such as golf paths or buildings needed to be stripped out of the map. To filter down to the relevant streets, the OSM highway tag was used to pull only data with the following values: motorway, trunk, primary, secondary, tertiary, residential, motorway link, trunk link, primary link, secondary link, and tertiary link. This was accomplished using the OSMfilter tool [20].

The OSM network uses bi-directional links to represent two-way roads. However, the Mobiliti simulation models link dynamics for each direction, and as such, a unidirectional link representation had to be generated. To accomplish this, a tool that creates road networks for MATSim was used. MATSim is an open-source framework for running Meso-level traffic simulations that also requires a unidirectional road network. To convert the OSM file to a MATSim network, a tool called JSOM was used along with a MATSim plugin [21]. A portion of the reduced OSM map converted to the MATSim network is shown in Figure 3.
The MATSim network was exported as shapefiles that represent the nodes and links of the network. These shapefiles were then converted to the data representation that Mobiliti required. This process required transforming the reference coordinate system from Projected CRS: EPSG:3857 to Geographic 2D CRS: EPSG:4326. Once this was accomplished, the shapefiles were converted to CSV files that Mobiliti could ingest. The map visualizations in Figure 4 were generated to display the differences between the two networks to verify the integrity of the newly developed full network. These show the dramatic difference in link density between the two networks.

**Figure 4:** The left image represents the OSM expanded network; the right image represents the reduced network.

**Random Forest Model**

At this stage, generating the complete network was nearly finished and ready for use in the Mobiliti simulation. However, one attribute, the number of lanes, was missing in a significant amount of the links. Because OSM is open-sourced, contributors are asked to select mandatory attributes, while some are left optional. As such, some users did not populate the number of lanes. To resolve this issue, a random forest model was built.

Of the 754,376 links in the expanded network, only 131,325 links contained a value for lanes. This meant that 83% of the links needed to have their number of lanes estimated. The 131,325 links were split into test and train datasets with a 25:75 split. Then a Random Forest model was used to generate the number of lanes for the rest of the links. The three predictor variables key to estimate the number of lanes is length, capacity, and free speed.

<table>
<thead>
<tr>
<th>Length</th>
<th>Capacity</th>
<th>Free Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.19724007</td>
<td>0.73000833</td>
<td>0.07275159</td>
</tr>
</tbody>
</table>

**Table 1:** The feature importance for the random forest model.

The feature importance details of this model are as expected, as the link capacity is most important for predicting the number of lanes, along with the length of
Lastly, the free speed carries some weight, although smaller than the other two features.

After the model is trained and feature importance is evaluated, the model is tested against the test dataset of 32,832 links. Against this test dataset, the model resulted in a mean absolute error of 0.15 lanes. This is extremely good, particularly when considering the random forest model is outputting continuous values and not whole numbers. When the output is rounded to the nearest whole number, the model resulted in an accuracy of 93.29%. This model is fairly accurate and devoid of any extreme outliers in its predictions. With the model trained and tested against links with existing values for the number of lanes attribute, the next step is to predict the number of lanes on the links without a value. The results showed that there were no extreme outliers.

**Updating the Travel Demand Model for the Expanded Network**

To complete the datasets that Mobiliti requires, a travel demand model had to be created for the newly created expanded network. For this purpose, the provided reduced network demand model was transformed to fit the expanded network by map matching the reduced demand model origin and end nodes to the nearest node in the extended network. A visual heat map of where the demand is distributed on the network is shown in Figure 5. Concerns about this transformation and the new demand model's ability to utilize the expanded network are addressed in the conclusion section. The heat map shows that the demand distribution between the two networks heavily overlaps, and the demand profile is maintained across the two networks.

**3.3 Simulation Scenarios**

The two simulation scenarios, Baseline-50 and UET, were investigated for both the reduced and the complete regional network model, resulting in four simulation datasets.
A simulation dataset is composed of five data sets (TSV and CSV files) described in the following table.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Description</th>
<th># Records / Granularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legs</td>
<td>Simulation details for most of the trips from the demand model, including start and end nodes, times, delay, fuel consumed, and distance traveled</td>
<td>18M+ trips</td>
</tr>
<tr>
<td>Links - Average Flow Rates</td>
<td>Vehicle flow rates in vehicles per second for each link for 15 min. intervals for a daily simulation (i.e., 96 values for each link, 750K links)</td>
<td>Per link</td>
</tr>
<tr>
<td>Links - Average Speeds</td>
<td>Average speeds in meters per second from the simulation for each link for 15 min. intervals for a daily simulation (i.e., 96 values for each link, 750K links)</td>
<td>Per link</td>
</tr>
<tr>
<td>Links – Fuel Consumption</td>
<td>Fuel consumed for each link for 15 min. intervals for a daily simulation (i.e., 96 values for each link, 750K links)</td>
<td>Per link</td>
</tr>
<tr>
<td>Links by Count</td>
<td>Summary values for 750K links from the simulation, including capacity and utilization</td>
<td>Per link</td>
</tr>
</tbody>
</table>

Table 2: Summary of input datasets

### 4 Analytics of the Simulation Results

This section details the analytics of the datasets. First, a detailed comparison of the characteristics of the reduced and expanded network is presented. Second, the real world does not follow optimized routes, and the Baseline-50 scenario was chosen for presentation in this paper. The results for this scenario for both networks are presented. Finally, real-world data from TXDOT provided a basis for comparison.

#### 4.1 Differences in Network Length

The Reduced Network contains 45,675 links for a total cumulative distance of 27,842,105 meters. In contrast, the expanded network contains 729,753 links, for a total cumulative distance of 122,311,419 meters. The ratio of reduced links counts to
expanded links count are approximately 1:16. The ratio of total length in meters between the two networks is approximately 1:4. This difference is primarily related to the expanded network's inclusion of lower-level arterials and residential links, shorter in length than the average link. Further analysis on these differences will examine capacity, free speed, and length distributions and visualize these two networks overlaid on top of each other.

Figure 6: This left image is a map of the greater Houston area that overlays the reduced network over the expanded network. The orange lines represent links in the expanded network. The blue lines represent links in the reduced network. The right image is a zoomed-in view of the Houston area, highlighting the locations where the orange/expanded network links are not present in the reduced/blue network.

These images provide a striking visual contrast for just how extensive the complete network's coverage is compared to the reduced network. This is particularly noticeable around the residential roads. This difference expands beyond the residential roads, with large arterial roads visible in the expanded network but lacking in the reduced network.

Figure 7: The above figure displays the network when filtering out links with 0 vehicle counts.
Because the travel demand was so similar in the two network cases, the simulations run for both the expanded and reduced network resulted in an under-utilization of the residential links in the expanded network. Many of these smaller links received no demand and had a vehicle count of 0. Even so, after filtering out these links, it is clear there are still some links in the expanded network that were not present in the reduced network that did receive traffic from the simulation. These larger arterial roads likely contributed to the VMT differences, which will be discussed later in the paper.

4.2 Comparison of Reduced and Expanded Network Attributes

Figure 8: This diagram is a comparison of the two networks. The x-axis groups the data by the functional link class. The y-axis represents the cumulative length of the links in meters.

The plot in Figure 8 highlights the differences between the reduced and the expanded network. The expanded network includes over 80,000,000 million meters of residential links, while the reduced network includes none. Unexpectedly, the total length of the "Highway" and "Freeway" links was larger in the reduced network. As the expanded network was intended to encompass all the links within the reduced network, this difference is surprising. It would be expected for these two functional classes to result in similar total length, if not more for the expanded. This discrepancy can be attributed to a difference in classifying the links into a unified functional class system. In OSM, several different functional class types do not match one-to-one with the reduced network classification.

https://scholar.smu.edu/datasciencereview/vol5/iss2/8
**Figure 9:** This diagram is a comparison of the two networks. The x-axis groups the data by bins that define the link’s free speed in meters per second. The y-axis represents the cumulative length of the links in meters, grouped by their bins.

The plot in Figure 9 examines the free speed. The fact that there is a greater distance of low free speed links in the expanded network is expected. What is unexpected is the fact that the reduced network has a larger distance of high free speed links. It would be expected that these high free speed links would match. This could result from an issue in the reduced or expanded network with how certain links are classified and what speeds these links are assigned.

**Figure 10:** This diagram is a comparison of the two networks. The x-axis groups the data by bins that define the link’s capacity. The y-axis represents the cumulative length of the links in meters, grouped by their bins.
The total length of high-capacity links mirrors the results from the free speed and length comparison. There is vastly more coverage of low-capacity links in the expanded network vs. the reduced network. The reduced network has greater coverage when measured by cumulative distance for the higher capacity links.

![Link Length Comparison](chart.png)

**Figure 11**: This diagram is a comparison of the two networks. The x-axis groups the data by bins that define the link's length in meters. The y-axis represents the cumulative length of the links in meters, grouped by their bins.

This diagram shows the cumulative length of links when binned by the length of their link. This indicates that the reduced network contains a greater cumulative distance for longer links in the three top bins. For all other bins, the expanded network includes a greater cumulative distance. These differences can be attributed to the residential roads in the expanded network that are not present in the reduced network. This assumes that the residential streets will be segmented into smaller lengths.

### 4.3 Network Conflation

A subset of links was conflated for comparison to understand the attribute differences between the two network representations. The geometric representation of links (shape and length) in the two networks is dissimilar, adding challenges in conflation. ArcGIS’s aligned and spatial tools were used to conflate a subset of the network. An area in downtown Houston consisting of 316 links (in the expanded network) was matched with the corresponding links in the reduced network. Figure 12 shows the capacity differences between the two networks. Of the matched subset, 65% of links in the expanded network had capacities higher than the reduced network by at least 500 vehicles/hr. For 45% links, the expanded network's capacity is at least two times that of the reduced network. However, when comparing the speeds, 80% of the expanded
network links have lower speeds than the links in the reduced network. The number of lanes remained consistent for most connections between the networks.

Figure 12: The capacity difference between expanded and reduced networks

The violet-colored links in Figure 12 represent the links with the capacity difference greater than 4,000 vehicles/hr. Fannin and San Jacinto are the two main streets in that category. For Fannin, the capacity is 6,000 vehicles/hr. and 1,219 vehicles/hr. for expanded and reduced networks, respectively. The speeds are 45 mph and 50 mph, and the numbers of lanes are 5 and 4 respectively in the two networks. The expanded network capacity number looks valid based on the number of lanes and free speed of the link. The link is also classified as a highway in the reduced network.
4.4 Traffic Metrics Comparison

Measuring the efficiency of traffic dynamics is often accomplished by comparing the aggregate results. Examples of aggregated metrics include VMT (Vehicle Miles Travelled), VHD (Vehicle Hours of Delay), estimated productivity loss from driver delay, and system-level fuel consumption. As part of this analysis, both the real-world sensor data and simulated data are visualized to aid in the presentation and validation of the simulations. The VMT, Fuel Usage, Delay, and Congestion Link Length (link volume/link capacity > 1.0) metrics for the four scenarios are summarized in the following table.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>VMT - million miles</th>
<th>Fuel Consumed - million gallons</th>
<th>Delay - million minutes</th>
<th>Productivity Loss - million dollars</th>
<th>Congested Links Length - miles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced Network - Baseline 50</td>
<td>177.41</td>
<td>31.10</td>
<td>18.96</td>
<td>47.28</td>
<td>1,274</td>
</tr>
<tr>
<td>Expanded Network - Baseline 50</td>
<td>181.55</td>
<td>34.27</td>
<td>35.19</td>
<td>50.74</td>
<td>5,726</td>
</tr>
<tr>
<td>Reduced Network - UET</td>
<td>176.80</td>
<td>31.23</td>
<td>7.22</td>
<td>45.04</td>
<td>342</td>
</tr>
</tbody>
</table>

Figure 13: Speed difference and Lane differences between the two network representations.
Simulation | VMT - million miles | Fuel Consumed - million gallons | Delay - million minutes | Productivity Loss - million dollars | Congested Links Length - miles
--- | --- | --- | --- | --- | ---
Expanded Network - UET | 180.46 | 34.41 | 23.91 | 48.96 | 2,185

Table 3: Comparison of traffic metrics across simulations

The above summary indicates that the total miles traveled and the fuel consumption went up with the simulations using the expanded network. The reason for this is the increase in the delay with the expanded network. The Mobiliti simulation has a storage model as part of the modeled link dynamics. The storage model will not allow the physical capacity to be exceeded and thus queues vehicles on upstream links until there is room for the vehicles to enter the downstream link. As such, the naive distribution of the travel demand onto the expanded network likely violated the assigned link capacities. This would result in vehicles having an unrealistic wait time to get started as they wait to be released into the network. Over time further investigation of the high demand links - their demand profile v.s capacity will reveal if this is the case. Due to the time constraints of the current exercise, this will be the focus of future work. Next, the analysis will look at the miles traveled by the road functional class across the four simulations.

![Graph](image)

Figure 14: Plot depicting areas where the reduced network’s travel time is significantly higher than the expanded network.

While the expanded network has a significant number of residential miles, most of them did not demand during the simulation. A potential reason for this is that the demand model for a reduced network is used to run the simulation for the expanded network, causing the demand to be skewed towards the links covered in the reduced network. Next will compare the travel times for individual trips for the Baseline 50 simulations between reduced and expanded networks.
Figure 15: Plot depicting areas where the travel time with the reduced network is significantly higher (i.e., minimum travel time 25 min. and difference greater than 25%) compared to expanded network; Point Count represents the number of trips.

Figure 16: Plot depicting areas where the travel time with the expanded network is significantly higher (i.e., minimum travel time 25 min. and difference greater than 25%) compared to reduced network; Point Count represents the number of trips.

Below is the histogram of significant differences in travel time between the networks to assess the magnitude of the difference.
Both the above plots infer that many trips have resulted in higher travel time when simulated using the expanded network. Many of these are concentrated around the Sam Houston Tollway segment to the southwest of the downtown area. This calls for further analysis in the future, as with access to a more extensive and more complete network, it would be expected to see improvement in travel times.

Figure 17: Histogram of significant travel time differences (i.e., minimum travel time 25 min. and difference greater than 25%) between reduced and expanded networks.

Figure 18: This figure shows the two networks with the stroke of the links weighted based on the vehicle count.
**Figure 19:** This diagram shows these two networks overlaid on top of each other.

**Figure 20:** The above shows an arterial link found in the expanded network but not the reduced one. It received minimal traffic, 2611 vehicles, relative to the other links in the networks. These arterial links contributed to the difference between the simulations, but these differences were not drastic.
4.5 Demand Distribution Heat Map

Figure 21: The above heat maps represent each node and are weighted by the number of trips that end at the node. The left image represents the demand model for the reduced network. The right image represents the demand for the expanded network.

The demand model originated with the reduced network. The demand model can be understood by examining the start and end nodes within the legs data generated for each simulation. Each leg or trip will have a start node and an end node. By aggregating the legs/trips by grouping on the start and end nodes, a visual heat map of where the demand is distributed on the network can be created. As previously discussed, the reduced demand model was transformed to fit the expanded network by map-matching the reduced demand model start and end nodes to the nearest expanded network nodes. Concerns about this transformation and the new demand model's ability to utilize the expanded network are addressed in the conclusion section. The heat map shows that the distribution of the demand between the two networks heavily overlaps.

<table>
<thead>
<tr>
<th>Network</th>
<th>Start Nodes Count</th>
<th>End Nodes Count</th>
<th>Total Start/End Nodes</th>
<th>Total Candidate Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced Network</td>
<td>14,987</td>
<td>15,386</td>
<td>19,148</td>
<td>22,943</td>
</tr>
<tr>
<td>Expanded Network</td>
<td>14,323</td>
<td>14,950</td>
<td>18,249</td>
<td>318,814</td>
</tr>
</tbody>
</table>

Table 4: The differences in the demand model counts by the network.

The two resulting demand models have now been aggregated and visualized as the heatmap in Figure 21. The reduced demand model utilized 14,987 unique start nodes and 15,386 unique end nodes. A total of 19,148 unique nodes out of 22,943 total
possible nodes were used as either start or end nodes in the reduced networks demand model. The expanded demand model utilized 14,323 unique start nodes and 14,950 unique end nodes. A total of 18,249 unique nodes out of 318,814 total possible nodes were used as either start or end nodes in the reduced networks demand model. Based on these counts, the expanded demand model does utilize the expanded network's capacity fully. With approximately 19 million trips, these trips could be better distributed across the expanded network to use the candidate nodes better. This would introduce some complexity as this redistribution of the demand model must be considered when analyzing differences in the simulations.

4.6 Simulation Data Validation and Evaluation Methods

Real traffic data (ground truth) can be acquired in many ways. Fixed cameras mounted on highways or buildings and traffic monitoring cameras are heavily used to capture the traffic details. Modern devices and technologies such as GPS and cell phones also provide vehicle positioning, speed, traffic flow, and other attributes. In addition, the U.S. Department of Transportation has also started a program called Next Generation Simulation (NGSIM) [17], capturing traffic data on various highways ten frames per second. The simulation data validation can be conducted in two main ways:

1) By using ground truth recorded data collected by the methods described above
2) Using synthetic data and statistical methods such as Intelligent Driver Model (IDM) [18]

The model validation in this analysis will primarily rely on the ground truth public data provided by Houston District TXDOT, Houston TranStar, and Texas A&M University organizations. Data acquired from these sources is merged and compared with the simulation data. This will help with understanding the accuracy of the simulation model and any variances therein.

The validation procedure involves checking daily traffic counts (ADT) field data collected from TXDOT across various simulations. The validation results showed that the expanded network baseline 50% dynamic routing simulation is the nearest representation of the real work traffic. The expanded network resulted in 18% relative absolute error vs. the reduced network yield of 24% relative absolute error.
Table 5: The above are sample validation results comparing field data against the expanded and reduced simulations for Average Daily Traffic (ADT). Figure 22: The above shows the correlation between the field data and the expanded network. The $R^2$-value = 0.9211.
Figure 23: The above shows the correlation between the field data and the reduced network. The $R^2$-value = 0.9428.

Below is a further evaluation of one of the outlier nodes from the above comparison, i.e., the node at the intersection of McKinney & Louisiana, with a relative error of 83%. Both validation and real-count data are accurate, as shown in the below images.

Figure 24: Location of the node with the highest variation to the actual ADT value

**Speed Validation**

Initial visual comparisons of the speed data against field data showed that the expanded network yielded closer results than the field data. This was particularly pronounced at the rush hour times of 7:30 AM-9:30 AM and 3:30 PM-6:30 PM, as depicted in Figure 25.

Figure 25: Validation of speed profile for one of the intersections
Vehicle Miles Travel (VMT) Validation

For VMT validation, this paper compared the VMT values from the simulation against the field data for the five counties making up the Houston Traffic District. The table below showcases that both expanded and reduced networks performed relatively well compared to the field VMT data.

<table>
<thead>
<tr>
<th>County Name</th>
<th>Field VMT</th>
<th>Expanded Network VMT</th>
<th>Reduced Network VMT</th>
<th>% Relative Error - Expanded Network</th>
<th>% Relative Error - Reduced Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazoria</td>
<td>8.6</td>
<td>7.6</td>
<td>7.5</td>
<td>11%</td>
<td>13%</td>
</tr>
<tr>
<td>Fort Bend</td>
<td>11.3</td>
<td>14.5</td>
<td>13.9</td>
<td>-9%</td>
<td>-4%</td>
</tr>
<tr>
<td>Galveston</td>
<td>7.0</td>
<td>6.4</td>
<td>6.3</td>
<td>8%</td>
<td>9%</td>
</tr>
<tr>
<td>Harris</td>
<td>117.7</td>
<td>128.3</td>
<td>125.3</td>
<td>-9%</td>
<td>-6%</td>
</tr>
<tr>
<td>Montgomery</td>
<td>15.0</td>
<td>16.1</td>
<td>15.9</td>
<td>-8%</td>
<td>-6%</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td>9%</td>
<td>8%</td>
</tr>
</tbody>
</table>

Table 6: Validation of VMT by County

5 Discussion

This paper started with whether existing reduced networks would yield as accurate traffic simulation results as expanded networks. This is an important question to answer when positioned within the context of aiding city planners. City planners often rely on simulations to help with forecasting the impact of changes on the city. With rapid urbanization occurring, cities need to run various simulations on time and against multiple possible networks. This paper was able to leverage the Mobiliti simulation platform to provide quick simulation results.

This paper started with a pre-generated reduced network for the greater Houston area and required that the expanded network be generated. This process was thoroughly covered in section 3.7. Creating this expanded network proved to be a more challenging than expected endeavor. As city planners and traffic simulations continue to develop, the question of how to create expanded networks for cities will be aided by the lessons learned in this paper. One important lesson learned from this paper is that open-source data is not always complete. The number of lanes for each link was missing for 83% of the data. In the form of a random forest model, this required machine learning to support populating this data where it was absent. Depending on the attributes needed for a simulation, it is important to understand that creating an expanded network comes at the cost of increased time and complexity in generating the network. This paper should aid future research and experimentation with creating expanded road networks. Ultimately an expanded network was created, although it differed from the reduced network in some unexpected ways.

Once the expanded network was created, the next step was to run simulations against the reduced and expanded networks to analyze their differences. This was accomplished by leveraging the Mobiliti platform, which allowed for quick results. Initial expectations of this comparison were that the expanded network would improve traffic dynamics/metrics. The reasoning for this is that an expanded network allows for additional options when routing vehicles according to the demand model. This expectation was upended when analysis of the differences between the two simulations
showed that the simulation run against the expanded network resulted in greater VMT, fuel usage, and vehicle delay. The differences between the simulation results for expanded and reduced networks can be seen in section 4.4. The aggregate vehicle miles traveled, VMT, was 2.3% higher in the expanded network. The fuel consumed in the expanded network was 10% higher. The aggregate vehicle delay was nearly double in the expanded network when compared against the reduced network. This was an unexpected outcome, but after analysis of the differences between the two networks, this could be explained by differences in the two networks. As outlined in Section 4.2, the expanded network contained fewer high-speed links and fewer high-capacity links when measured by distance (meters). Figure 18 in section 4.4 shows the vehicle counts by links between the reduced and expanded network. This helps to visualize where the demand was routed and aids in answering the question of why the expanded network resulted in higher traffic dynamics. Figures 18 and 19 show heavy overlap between the two networks. Some of this is to be expected, but there was also the expectation that the simulation would utilize the additional expanded network links. This was not the case, as shown in Figure 20. This is an example arterial link that exists in the expanded network but not in the reduced network. This additional arterial network received a small amount of vehicle traffic. Additionally, many of the expanded network links received no demand at all. This leads into section 4.5 - Heat maps of the demand model.

Section 4.5 helps to explain why the expanded network simulations utilized additional links so little. The demand model for the expanded network was the same as the reduced network. This reduced network was confined to a total of ~22,000 nodes to assign vehicle trips to for their source and destination. Alternatively, the expanded network contained ~320,000 possible nodes for assignment of demand/trips. Since the demand model was retained between the two networks, they both used ~19,000 nodes as the source and destination points. This leaves many expanded network nodes untouched as either the source or the destination points. The heatmap in Figure 21 helps to visualize the similarity between the two demand models. Their concentrations are nearly the same.

After summarizing the aggregate traffic dynamics differences between the simulations, the next step was to compare the simulations against field data. This was done in section 4.6. The results of these comparisons against validation data show that the expanded network resulted in improved simulation accuracy when compared against the field data. The vehicle counts for sample links in the expanded network resulted in a relative error of 18%. The vehicle counts for the same sample links in the reduced network resulted in a relative error of 24%. This can be found in Table 4. Additionally, initial samples of the speed of vehicles by links yielded better results in the expanded network than the reduced network. This was particularly pronounced around rush hour times.

5.1 Future Work

The demand model was minimally transformed between the reduced and expanded network simulations. This demand model was initially generated on the reduced network. This limited the origin and destination nodes available to the demand model.
when transferred over to the expanded network. The expanded network contained nearly 16 times more links that could be assigned the demand. This contributed to the underutilization of the expanded network's additional links because the demand source and destination nodes were still centered around the original reduced network. Future work would benefit by exploring an updated demand model that redistributes the demand according to the expanded network. This redistribution could be accomplished by breaking up the expanded network into smaller geographic areas and redistributing all the demand within those areas based on demographic data.

The expanded network was created using OSM, which is open-source data. Such data is dependent on the users who are contributing to the dataset. These users vary by region, with some high-quality power users often contributing greater to differences between regions of the map. Further work to refine the real network to account for variation in the quality would be important. The random forest model used to predict the number of lanes by the links could be further improved by additional data sources. It outlines how simulation results can be analyzed and verified against real-world data. Hopefully, this paper provides a guide for future traffic simulation work about creating expanded networks from open-source data like OSM.

5.2 Ethics

Analysis of traffic is intended to be used by city planners and other organizations interested in traffic management. Often this type of analysis is reduced to traffic mileage reduction. This does not consider other factors such as the environmental impact, noise pollution, and risk to pedestrians. There are other demographic considerations to consider as well. A simulation may minimize fuel consumption and miles traveled, but the increased risk to children is not supposed to increase traffic past schools. Any insight taken from simulations and this paper needs to be analyzed through the larger lens of the impact on society. A narrow-minded approach to traffic analysis is sometimes beneficial for scientific experimentation. Still, before real-world action can be taken, it must be reviewed to consider the various and complex factors that make up the impact on society.

Additionally, this analysis has focused on the Houston area. Globally, some factors have a more significant effect than in the Houston area, such as road quality/safety and extreme crime areas. It would be important to consider these factors when expanding the analysis to other parts of the world.

The validation data was gathered through online sources. However, the team ensured it is covered under the open-source licensing terms and was available via verified government and private sites. The team also confirmed the data sources do not violate any privacy laws, mention any individual, cause any emotional or personal damage. Team also made sure all the sources of data, research, and papers were adequately documented, authors and sources were given the due credit. As a part of ethics guidance, this paper also represents all the data and research in full transparency. The paper honestly reflects all the findings and accurately represents the data and opinions.
6 Conclusion

Analyzing traffic dynamics using various networks and simulations is a complex and challenging task - major variables to calibrate include the networks themselves, the simulations, and the demand models used. Then determining which combination of variables provided the most accurate outcome introduces an additional task. The result is a valuable insight for city planners, corporations, and residents of the studied area. As an evolving field of study, this paper will support future work in creating networks for simulations and subsequent city planning.

The initial scope of this paper expanded as the challenge of generating an expanded network became apparent. By detailing how this was done and which issues were identified and overcome, others will build off these processes. Additionally, the expanded network leveraged open-source data and open-source tools. This enables other researchers to easily replicate and create expanded networks for their specific geographic areas of interest.

Lastly, the expanded network increased traffic metrics, yet mixed results regarding the matched field data were an unexpected but interesting lesson. Although the expanded network seemed to perform better in some metrics than the reduced network against the field data, it still has room for improvement. In the future work section, this paper discussed how this expanded network could be improved and how the demand model could be fit to utilize the expanded network better.

Acknowledgments:
Jacquelyn Cheun, Ph.D. – Capstone Professor – SMU
Jane Macfarlane, Ph.D. – Capstone Professor – UC Berkeley LBNL
Anu Kuncheria, Ph.D. Candidate – Capstone Advisor – UC Berkeley LBNL

References

3. S. Jiang, J. Ferreira and M. C. Gonzalez, "Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore," in IEEE Transactions
4. Brita Sadeghinasr1, Armin Akhavan1, Qi Wang, Ph.D. (n.d.). Estimating Commuting Patterns from High Resolution Phone GPS Data. Department of Civil and Environmental Engineering, Northeastern University, 360 Huntington Ave, Boston, MA 02115.
6. Yihan Cai1, Menghan Tian1, Weidong Yang1, Yi Zhang2. (n.d.). Stay Point Analysis in Automatic Identification System, Trajectory Data School of Computer Science, Fudan University, Shanghai, China.


22. JONATHAN WOETZEL, DR. SHANNON BOUTON - Activating more private capital to make cities more sustainable; https://blogs.worldbank.org/sustainablecities/activating-more-private-capital-make-cities-more-sustainable


Appendix

A.1. Quartile Comparisons of Reduced and Expanded Network Attributes

![Link Capacity Comparison](image)

**Figure A.1.1:** The capacity in the expanded network has a large outlier, due to a road segment with a high number of lanes. The 25th, 50th, and 75th percentiles result in a lower capacity in the expanded network relative to the reduced network. This is expected as the expanded network includes smaller links such as residential roads.
Freespeed is a measurement of the speed at which a vehicle can traverse a link, measured in meters per second. This freespeed differs between the two networks in an expected way. The expanded network has smaller roads and therefore the bottom 25th, 50th, and 75th percentiles fall well below those for the reduced network. The maximum speeds effectively match between the two networks as they both encompass the highways.

**Figure A.1.2:** Freespeed comparison.

**Figure 18:** The link length comparison mirrors the previous analysis. The expanded network at lower percentiles measures lower in link length relative to the reduced network. This is expected due to smaller roads in the expanded network.