Leveraging Probe Data and Machine Learning to Derive and Interpret Macroscopic Fundamental Diagrams Across U.S. Cities

Ling Jin, Xiaodan Xu, Yuhan Wang
Energy Technology Area, Berkeley National Laboratory
Berkeley, CA 94720
ljin, xiaodanxu, yuhan_wang@lbl.gov

Kaveh Farokhi Sadabadi
Center for Advanced Transportation Technology University of Maryland
College Park, MD 20742
kfarokhi@umd.edu

Alina Lazar
Youngstown State University Department of Computer Science and Information
Youngstown, OH 44555
Email: alazar@lbl.gov

Duleep Rathgamage Don
Kennesaw State University School of Data Science and Analytics
Marietta, GA 30060
drathgam@kennesaw.edu

Zachary Needell, C Anna Spurlock
Energy Analysis and Environmental Impacts Division
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
ZANeedell, caspurlock@lbl.gov

Mahyar Amirgholy,
Kennesaw State University Civil and Environmental Engineering Department
Marietta, GA 30060
mahyar.amirgholy@kennesaw.edu

Mona Asudegi
Office of Transportation Policy Studies
Federal Highway Administration
Washington DC, D.C. 20590
mona.asudegi@dot.gov

2023 IEEE International Conference on Big Data (BigData) | 979-8-3503-2445-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/BigData59044.2023.10386591

***Abstract*— Macroscopic fundamental diagram (MFD) captures an orderly relationship among traffic flow, density, and speed at the network level. Understanding network-wide traffic through MFDs can optimally allocate demand to existing networks, improving performance by maximizing network production and avoiding congestion. However, due to historical data limitations, empirically derived MFD models are sparse in the literature, especially for the U.S. cities. Leveraging a large-scale and granular census-tract-level flow and density derived from vehicle probe data, this research is the first to develop a machine learning approach to both derive MFD models and interpret their underlying difference among urban networks across the entire United States. Among the four machine learning methods tested here XGBoost is found to deliver the best performance to predict the network traffic flow for given vehicular density and location attributes. Interaction Shapley Additive explanation (SHAP) values are used to interpret the factors, such as land use, transportation infrastructure, and network topology, that influence the flow-density relationships among locations. The analysis framework developed in this work can generate data-driven MFDs and a deeper understanding of their shape dependence on network, infrastructure, and land use characteristics, which can be used by transportation authorities to derive and optimize location-specific MFDs facilitating more informed management and planning decisions at the network level.**

Keywords—macroscopic fundamental diagram, United States, vehicle probe data, machine learning models, TreeExplainer, Interaction Shapley values.

I. Introduction

Traffic in an urban network becomes congested once a critical number of vehicles is reached. Macroscopic fundamental diagrams (MFD) describe an orderly and consistent relationship between average vehicle flow and average traffic density when both are measured across a certain urban network. Such relationships have been proven to exist with simulation and empirical data in field studies [1]–[3]. The MFD (see Fig. 1) usually exhibits an uncongested branch, when increasing the number of vehicles in the network (indicated by traffic density) increases the travel production (indicated by space mean flow), and a congested branch, when the opposite is true. The urban network system's capacity and critical density are reached at the boundary between the two phases (Fig. 1). The shape of MFDs depends on traffic signal settings, block lengths, free-flow speeds, and routing behaviors that are specific to a given network location [4], [5].

MFD model is one of the most famous examples of parsimonious traffic models for the aggregate behavior of large

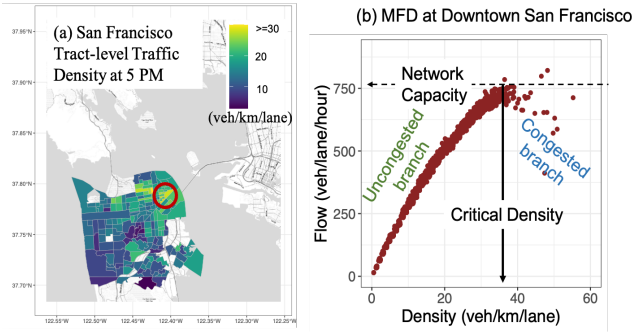

Fig. 1. (a) San Francisco tract-level traffic density at 5 PM; (b) Observed flow-density scatter at an example subregion of downtown San Francisco with illustration of network capacity and critical density and traffic regimes using data from this study.

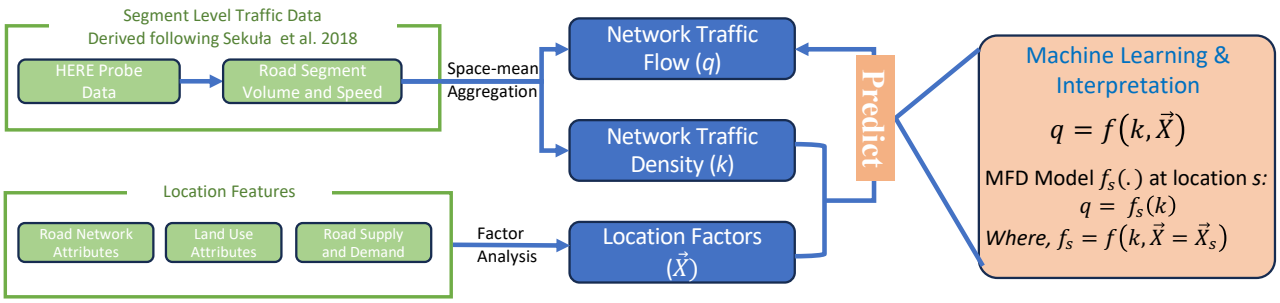

Fig. 2. Overview of data process and analysis of the paper.

systems with many agents [6]. Understanding network-wide traffic through MFDs can optimally allocate demand to existing networks, improving performance by maximizing network production and avoiding congestion. With reduced computational complexity and improved system-level representation and interpretability, MFDs are well suited to analyzing a large space of policy options and uncovering general insights into large-scale strategies. Example applications include perimeter flow control [7], [8], area-wide congestion pricing [9], [10], space allocation [11], street network configuration [12], [13], vehicle routing [14] and regional evacuation [15].

Despite wide application of MFDs, MFD models (i.e. the flow-density relationships) have only been empirically derived for a limited number of networks such as [4],[16], [17][18]. Literature has been particularly sparse in empirically derived MFDs in U.S. urban locations. For example, only one U.S. city (Los Angeles) was included in a recent study that estimated MFD functional forms in 41 cities (mostly in Europe) around the world using existing traffic monitoring systems located on main urban roads [18].

Furthermore, the empirical functional form, $f_s(.)$ that describes MFD models (i.e., flow as a function of density) at a given location s, were typically predetermined as multi-regime linear, polynomial, or exponential functions (see review in [19]) with parameters not readily transferrable from one location to another. However, machine learning models, on the other hand, when trained with large-scale flow-density data and location related input features across network types (see Equation (1)), hold the potential to derive MFD models flexibly for any given network based on location-specific characteristics (Equation (2)).

$$q = f(k, \vec{X}) \tag{1}$$

Where q is the network average flow, k is the network average density, $\vec{X}$ are the location features, and $f(.)$ represents the relationship learned by machine learning models. Thus the MFD model $f_s(.)$ at any given location s with location features $\vec{X} = \vec{X}_s$ can be derived as:

$$q = f_s(k) = f(k, \vec{X} = \vec{X}_s) \tag{2}$$

The lack of empirically-derived, machine-learning based MFD models and subsequent understanding of MFD differences across network locations in the U.S. is mainly due to limited availability of traffic volume data (traffic flow). Unlike speed data, which is readily available network-wide through probe vendors (e.g., INRIX, HERE, TomTom), reliable volume data generally exists only at sparsely-located continuous count stations. In recent work, Sekuła et al. [20] developed and applied a novel approach for estimating hourly volumes that combines a widely-used profiling method [21] and an artificial neural network (ANN) model trained with vehicle probe data for the state of Maryland. Sadabadi et al. (described in [22]) then expanded this method. They applied it to all 50 states in the U.S., leveraging a variety of data sources—most notably segment-level HERE probe data, including probe counts and speed—to estimate traffic volume and speed at the road segment level at 15-minute intervals.

To address the aforementioned research gaps related to empirically derived and location-flexible MFD models across U.S. urban areas, this paper develops the *first* application of machine learning methods to derive the empirical flow-density relationships MFD models by leveraging the newly available volume and speed estimates at the road segment level nation-wide from HERE probe data [22]. We seek to evaluate the ability of machine learning methods to predict location-dependent flow-density relationships and determine important location factors that underly the differences in the resulting shape of MFD curves. We particularly focus on the differences in critical density and network capacity (as illustrated in Fig. 1) that delineate the boundary of the network traffic between being in the uncongested and congested branches of the MFD curves.

First, we compare the performance of four machine learning methods. Then, TreeExplainer [23] is used to identify and interpret important factors influencing the flow-density relationships across different locations, including a wide range of transportation supply and demand characteristics such as road network topology, land use, transportation infrastructure, and demand characteristics. The overview of the data and analysis process is illustrated in Fig. 2.

The rest of the paper is organized into the following sections. Section II describes the data sources and preprocesses that generate network-level flow and density and location factors. Section III introduces the machine learning methods and how they are applied to our data; and the interpretation method used. Section IV presents the results, including the performance of the machine learning models and an interpretation of important factors. Section V concludes the paper.

II. Input Data Preprocesses and Description

A. Network Level Flow and Density Data Process

1) Road Segment-level Volume and Speed Estimation

Three months of HERE probe data (Sept - Nov) in 2019 have been licensed for the full U.S. network geometry, traffic counts, speeds, number of probes, and weather data are pre-processed and ultimately conflated to prepare the data for model calibration. A fully connected feedforward multi-layer Artificial Neural Network (ANN) model is applied to calibrate, test, and validate consistent models to estimate traffic counts (volume) for road segments belonging to different functional road classes (FRCs) at 15-minute granularity. Performance is evaluated in comparison with existing traffic count observations. A detailed description of the traffic volume estimation, model calibration, and validation strategy is presented in [20] [22] and thus *not* reported in this paper. The segment-level volume estimation and HERE probe vehicle reported average link speed are used in the next step for deriving network level flow and density.

2) Network-level Flow and Density Aggregation at Urban Census Tracts

The vehicle flow (q_i) at a given road segment (i) is computed by averaging the volume data over monitored lanes and over the observation period to get the number of vehicles per lane per second. Then the harmonic mean speed (v_i) is derived from the observed speed data collocated with the volume monitor. Traffic density (k_i) is derived using the macroscopic flow equation $q = kv$. Then network-level average flow ($\hat{q}$) and density ($\hat{k}$) are essentially the spatially weighted average of all the individual links for the given spatial unit [21] shown below in Equations (3) and (4), where l_i is the segment length and n_i is the number of lanes for segment i.

The aggregation is performed for each census tract, with a typical size of 0.6 to 1 km^2 in densely populated urban areas. The choice of spatial unit is to ensure the spatial alignment with the available location features to avoid interpolation errors as well as to limit network inhomogeneity that may arise from aggregation.

Freeway segments (FRCs 1 and 2), which generally account for less than 3% of the total lane miles in urban tracts as defined in the transportation typology [24], are excluded in the aggregation to avoid the influence of higher speed and volume from these non-typical road types in urban areas.

$$\hat{q} = \frac{\sum_i q_i n_i l_i}{\sum_i n_i l_i} \tag{3}$$

$$\hat{k} = \frac{\sum_i k_i n_i l_i}{\sum_i n_i l_i} \tag{4}$$

The upper bound of the flow-density scatter that represents the MFD relationships are used for training the ML models. The upper-bound flow and density values correspond to the top 20% of the flow values per density bin are used as the outcome flow, with each density bin corresponding to 1/50 of the density range observed.

Finally, to ensure the derived MFD models captures homogeneous traffic patterns at the census-tract level, the study only selects urban census tracts with land areas less than 10 km^2 [18].

B. Location Attributes Process

In addition to the network flow and density derived from HERE data, various transportation supply and demand characteristics are also collected from various data sources and aggregated at the census tract level to predict tract level flow at a given density, i.e., the MFD. These features help explain how land use, transportation infrastructure, network topology, and travel demand may affect flow-density relationships across different locations. A total of 38 location attributes are included:

- Land use attributes: e.g., development intensity and fraction of land use types;
- Network attributes: e.g., network circuity, dead-end fraction, intersection density, street length, percentage of road functional classes;
- Road supply and demand characteristics that may affect the network utilization and thus influence the MFD shape: e.g., lane-meter per capita, job and population density, job-housing balance.

The input features, including variable names and descriptions, and their data sources are provided in Table A1 in the Appendix.

Due to the large number of input features and high correlation among them, we applied factor analysis to reduce the dimensionality and derive interpretable location factors. An Exploratory Factor Analysis (EFA) is performed using the Python package 'FactorAnalyzer' with data from 19,361 census tracts after removing tracts larger than 10 km^2 in land area or with missing values.

III. Machine Learning and Interpretation Methods

A. Machine Learning Methods

We applied four machine learning methods (briefly described below) to predict network flow from given density and location factors. A total 16,808,176 network-level data points from 19,361 census tracts are used. The data are split into 80% for training and 20% for testing, with network density and location factors as input features and network flow as the outcome.

1) Random Forest.

This algorithm [25] builds an ensemble of decision trees, or tree predictors, which depend on randomly and independently sampled vectors over the same distribution. The strength, correlation, and monitor error are closely followed to track the growing features in response to the branches splitting.

In this study, the random forest regressor from the 'scikit-learn' package is used [26] to train the random forest model. The

hyperparameters of tree size are tuned to achieve the best model accuracy (or lowest squared error).

2) *XGBoost.*

This algorithm is based on the standard gradient boosting methods but employs a new regularization technique, instead of optimizing the loss function, to minimize overfitting [27]. This tactic allows XGBoost to be faster and more robust during tuning.

In this study, the 'XGBoost' package [27] is used to estimate the gradient boosting tree. 'XGBoost' allows parameter tuning for a variety of hyperparameters, and the notable hyperparameters, including learning rate, tree size and regularization terms, are tuned to minimize the squared error of the model.

3) Support vector machine (SVM)

This algorithm is another ML method that addresses nonlinearity in the data (Hastie et al., 2009). SVM regression works by projecting input factors into linear-separatable spaces and finding the best fit linear function in that space. The projection is performed using various linear or nonlinear kernel functions. SVMs are one of the most robust prediction methods, insensitive to outliers and less prone to overfitting when using the 'loss+penalty' function as the objective.

In this study, due to the low scalability of SVM regression on large a dataset, an ensemble approach is adopted to combine the predictions from a large number of SVM regressors, with each SVM trained on a smaller subsample (10,000 samples) from the training data. The Radial Basis Function (RBF) kernel is adopted for the nonlinear projection of input factors, as RBF can combine multiple polynomial kernels multiple times of different degrees efficiently, and outperforms other kernels.

4) Neural Network - Multilayer Perceptron (MLP)

This algorithm is one of the simplest multi-layered neural network architectures, consisting of a hierarchical structure of layers containing individual artificial neurons [28]. For the current application, we implement an MLP architecture with 3 hidden layers, with 100, 50, and 5 neurons, respectively. The Adaptive Movement Estimation algorithm (ADAM) [29] is an extension of the stochastic gradient descent that automatically updates the learning rate by taking into account the average of the second moments of the gradients. We employed ADAM with a starting learning rate of 0.01. The loss function for this regression task is the Mean Squared Error (MSE). We train the model for 10k epochs.

B. Interpretation of Location Factors

1) Importance Ranking using Interaction SHAP Values

In this study, the importance of the location factors lies in their interaction effects with density, that is, their ability to influence the prediction of outcome flow as an interacting factor with input density. However, traditionally, local explanations based on feature attribution assign a single number to each input feature. Such simplified representation comes at the cost of combining main and interaction effects. For some ML methods, especially tree-based methods, SHAP also provides measurements of local interaction effects under TreeExplainer based on a generalization of Shapley values [23], [30], which was leveraged in transportation research [31]. The interaction SHAP values allocate credit not just among each factor, but among all pairs of factors, to separate out main and interaction effects for individual model predictions and uncover important patterns of joint effects of factor combinations. For TreeExplainers, the SHAP interaction value is defined as:

$$\phi_{i,j}(f,x) = \sum_{S \subseteq \mathrm{M} \setminus \{i,j\}} \frac{|S|(M-|S|-2)!}{2(\mathrm{M}-1)!} \nabla_{ij}(f,x,S) \qquad (5)$$

And,

$$\nabla_{ij}(f,x,S) = f_x(S \cup \{i,j\}) - f_x(S \cup \{j\}) - f_x(S \cup \{i\}) + f_x(S) \qquad (6)$$

Where $\phi_{i,j}(f,x)$ is the interaction SHAP value between factor i and j, for the estimated model $f(.)$ and specific input x; S is the subset of factors; M is the set of all m input features; f_x is conditional expectation function of the output under input x and estimated model $f(.)$.

In this study, the average interaction SHAP values of the location factor and density pairs are used to rank the importance of and help interpret location factors influencing the flow-density relationship (i.e., the MFD shapes).

2) Location Dependence of MFD Shapes with a Focus on Network Capacity and Critical Density

The shapes of the MFD curves vary by network locations. In particular, the critical density and network capacity are two important traffic control parameters related to MFD shapes. These two parameters (as illustrated in Fig. 1) delineate the boundary between uncongested and congested branches of the MFD curves, representing the optimal performance of the network. In this study, the network capacity is derived from both observed flow and the predicted flow from the machine learning MFD models, as the 99th percentile of the flow values (observed or predicted). The critical density is the network density associated with the network capacity flow value. The location dependence of MFD shapes will be investigated by comparing the predicted and observed relationships between these MFD shape-related parameters (critical density and network capacity) and location factors.

IV. Results and Discussions

A. Data Preprocessing Results

1) Aggregated Network Flow and Density

The road segment level HERE data are aggregated to derive flow and density in 9,528 census tracts and used for final MFD estimation, with sufficient coverage for major U.S. cities and urban areas as indicated in Fig. 3.

The tract-level median density at 5:00PM is shown for urban tracts across the U.S. in Fig. 3 with higher densities appearing in major urban areas known for experiencing chronic congestion

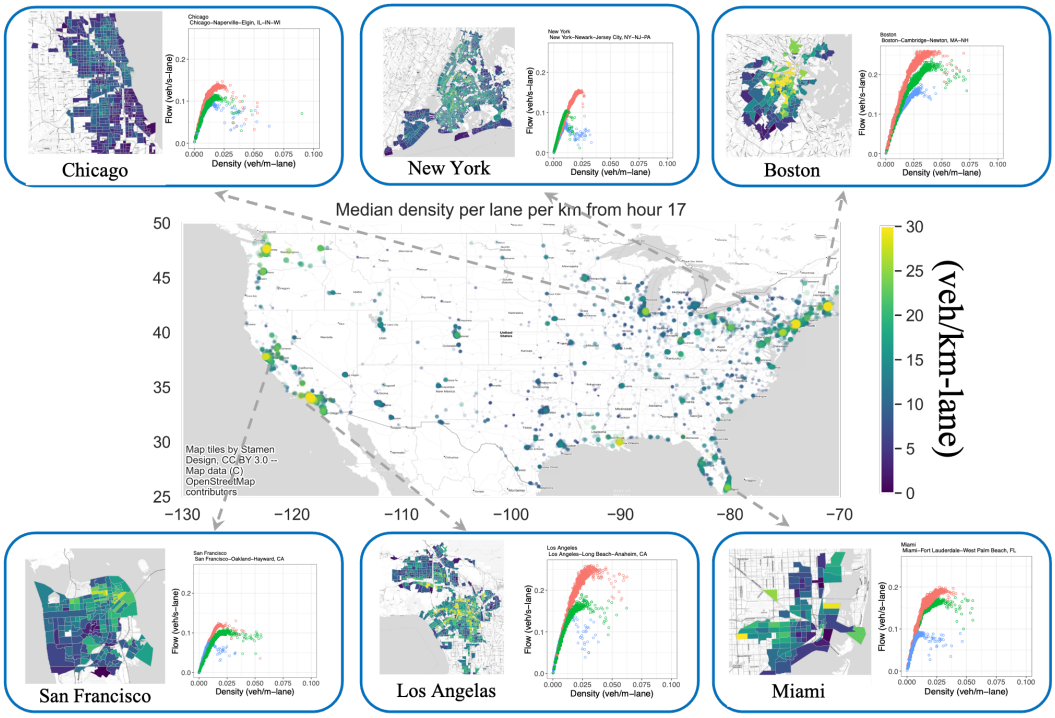

Fig. 3. Network traffic density at 5:00PM in urban tracts across the U.S. with inserts showing the zoomed-in view of the tract-level density in six cities and the flow-density scatters (i.e., observed MFDs) for three randomly selected urban tracts in each city.

according to the Texas Transportation Institute's Urban Mobility Report [32]. Zoomed-in views of six selected cities are provided in Fig. 3 together with the observed MFDs derived from HERE probe data at three randomly selected tracts in each city.

The observed MFDs shown in Fig. 3. give a good example of MFD shapes varying by location of the networks. Overall, the network capacities are below 0.14 #veh/sec-lane, with the lowest in Chicago (0.10 #veh/sec-lane). Urban road networks in Boston and Los Angeles are observed to have greater capacity with mean values of 0.18 and 0.19 #veh/sec-lane, respectively. In addition to the variation between cities, from the MFD plots we can see, the MFD curves exhibit within-city variation.

2) Location Factors Derived

Through parallel analysis, the top 13 factors are selected to achieve the balance between variance explained and interpretability of the factors. Fig. 4 depicts the 13 factors indicated by the columns, and how the raw features (y-axis) are loaded on them with complete explanation and description presented in the Appendix Table A2.

Among the resulting factors, "freeway", "development level" and "non-freeway arterial" indicate density of major roads and level of urbanization, combining various network and traffic attributes. Factors such as "network connectivity", "network complexity", "core-edge network" and "network circuity" represent the network topology mostly relying on network features from OpenStreetMap [33]. Factors including "mixed-use districts", "bike potential", "walk potential", "job hub", and "median travel", capture the demand characteristics

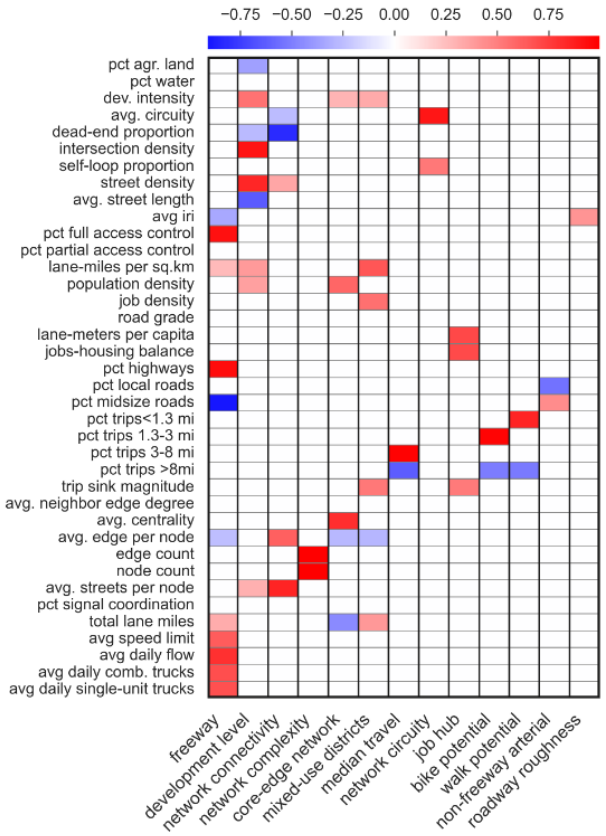

Fig. 4. Raw features (y axis) and their factor loadings on the 13 derived location factors (x-axis).

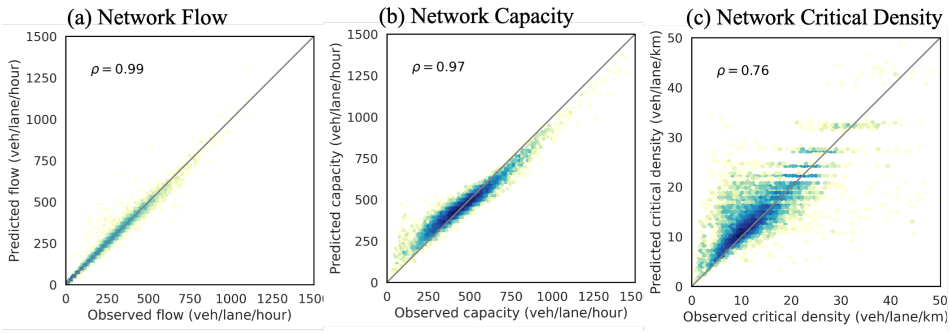

Fig. 5. Comparison between observed and XGBoost model predicted values for (a) network flow, and derived MFD shape-related parameters (b) network capacity, and (c) critical density. Color indicates the density of the points.

TABLE I. MODEL PERFORMANCE METRICS ON 20%* TESTING DATA

ML Models	Performance Measures			
	R^2	MAE (veh/lane-hr)	RMSE (veh/lane-hr)	MAPE+ (%)
XGBoost	0.984	12.12	22.67	6.1
Random Forest	0.982	12.28	24.2	6.2
Ensembled SVM	0.913	27.27	53.02	12.7
MLP	0.953	22.65	39.25	11.1

* Testing data size: 3,361,636
+ data size 2,143,156 after flow (100 veh/lane-hr) and density (0.006 veh/m-lane) cutoff thresholds applied. (the cut-off is to remove observations with near zero flow and thus potentially high percentage error even if absolute error is low)

of each tract. Finally, "roadway roughness" suggests the vertical alignment of the roads and easiness of driving on those roads. Those factors help capture the major location-specific infrastructure and traffic characteristics, and can affect the MFD trends due to their potential impacts on traffic flow and network utilization.

B. Performance of Machine Learning Models

The model performance, when predicting network flow from given density and location factors in the 20% testing data, is evaluated using four metrics: R^2, mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) in TABLE I. Across all four metrics, XGBoost consistently shows the best performance among the machine learning models evaluated here. This conclusion can also be visually confirmed by comparing the observed vs predicted network flows from the XGBoost model in Fig. 5 (a).

Taking the best performing XGBoost model, we further evaluate its ability to capture two of the MFD shape parameters: network capacity and critical density, derived from the observed vs. predicted MFD curves in Fig. 5 (b) and (c). The comparison indicates reasonable agreement between modeled and observed turning points of the MFD curves, with correlation of 0.97 and 0.76 for network capacity and critical density, respectively, across U.S. urban tracts.

C. Influence of Location Factors on MFD Shapes

1) Importance Ranking of Location Factors

Coupled with the best performing XGBoost model, TreeExplainer uncovers the influence of location factors on MFD shapes learned by the model using the interaction SHAP values. Fig. 6(a) presents the importance ranking of the location factors according to their interaction SHAP values with density.

The top-ranking factors are mostly related to network topology (such as network connectivity, network complexity, core edge network, and network circuity), transportation infrastructure characteristics—such as composition of the road functional classes (freeway and non-freeway arterial) and roadway condition (roadway roughness)—and land use factors (such as mixed-use districts and development level).

In contrast, the demand and trip-related factors (such as trip distance related factors: median travel, job hub, bike potential and walk potential) are ranked at the bottom.

This ranking of location factors aligns well with existing literature. The shape of MFDs has been considered in the literature to be mainly determined by the urban road structure and network topology, traffic control, and the level of inhomogeneity in the distribution of traffic. Although it is still under debate whether the MFD shape depends on demand characteristics such as trip origins and destinations and route choice, most of the MFD literature assumes it is more or less independent of demand when the trip length remains roughly constant [34].

Land use characteristics, such as the development level and mixed land use with development including both commercial and residential buildings, whose importance is revealed in this study, have rarely been investigated in the literature. These land-use related attributes, especially mixed-use level, may influence the road network utilization and homogeneity of the traffic. For example, networks primarily serving as job centers may induce traffic flows that are uni-directional during rush hours and therefore decrease the level of two-way lane utilization, resulting in inhomogeneous congestion. More importantly, the direction of the influence of these location factors (discussed in the next section) can help future design of transportation infrastructure and land use planning to mitigate traffic congestion and improve network performance.

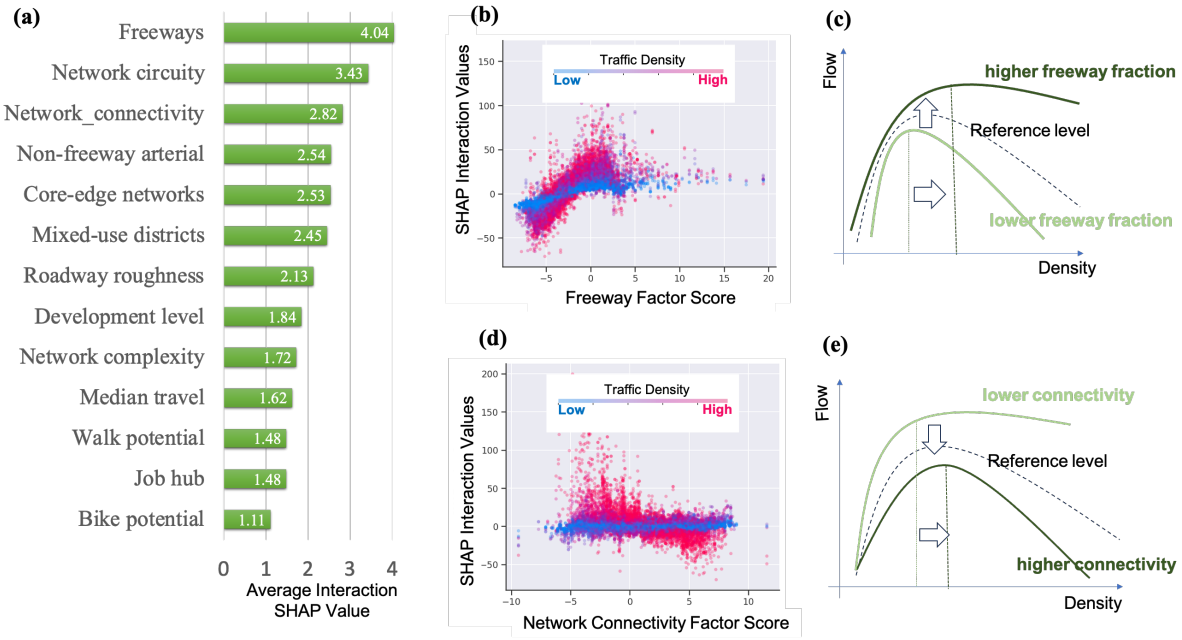

Fig. 6. (a) The importance ranking of location factors according to their interaction SHAP values with network density; local interaction SHAP values between density and two example location factors (b) freeway factor and (d) network connectivity factor. The corresponding graphical interpretation of MFD shape changes are illustrated in (c) and (e).

2) Interpretation of Location Dependence of MFD Shapes

The dependence of MFD shapes on location factors can be closely examined by the local SHAP interaction values. The implications on how the consequent network critical density and capacity from MFD shapes change from these factors can be subsequently derived. Two examples are shown in Fig. 6 (b) and (d) for "freeway" and "network connectivity" factors. Positive interaction SHAP values indicate a relative increase in the predicted flow, whereas negative ones indicate a relative decrease.

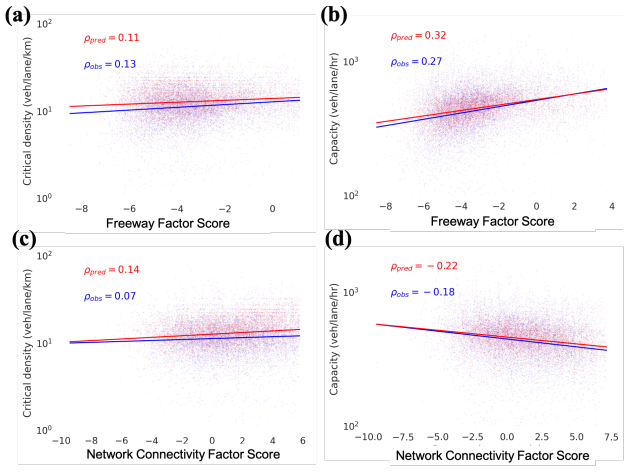

Fig. 7. Association between example location factors and MFD shape parameters derived from observed (blue) and predicted (red) MFD curves (ρ = correlation coefficient): (a) critical density and freeway factor; (b) network capacity and freeway factor; (c) critical density and network connectivity factor; (d) network capacity and network connectivity.

We can see in Fig. 6 (b) that the presence of a higher fraction of freeways in the network slightly increases flow under the low density (uncongested—indicated by the blue dots) conditions and greatly increases flow under the high density (more congested—indicated by red dots) conditions. This results in the deep green MFD curve in Fig. 6 (c) relative to a reference level (dashed black curve). In contrast, a lower fraction of freeways in the network slightly decreases the flow under the low density (blue dots) conditions and greatly decreases the flow under the high density (red dots) conditions. This results in the light green MFD curve in Fig. 6 (c) relative to a reference level. Taken together, MFD shapes under high vs low fractions of freeways shown by dark green vs. light green in Fig. 6 (c) indicate that both the capacity and critical density of the network increase with the fraction of freeways.

In contrast, when we apply the interpretation of interaction SHAP values in Fig. 6 (d) to the MFD shape changes illustrated in Fig. 6 (e), we can see increasing network connectivity has a *trade-off/opposite* effect on network capacity and critical density. Higher connectivity of the road network can accommodate more vehicles before congestion sets in, indicated by the increase in critical density, however, it also decreases the network capacity at the same time.

Fig. 7 confirms, both from observed and predicted MFD curves, the association of critical density and network capacity with the two location factors illustrated in Fig. 6. We can see indeed that critical density and network capacity both have a positive association with the freeway factor, while they have the opposite associations with the network connectivity factor. The complete list of associations between location factors and MFD shape parameters (critical density and network capacity) derived from observed and predicted MFD curves, is presented in Table II.

TABLE II. ASSOCIATION BETWEEN MFD SHAPE PARAMETERS (CRITICAL DENSITY AND NETWORK CAPACITY) AND LOCATION FACTORS

Location Factors	Network Capacity		Network Critical Density	
	predicted	observed	predicted	observed
Freeway fraction	+ + +	+ + +	+ +	+ +
development level	−	−	+ +	+ +
network connectivity	− − −	− − −	+ +	+ +
network complexity	+ +	+ +	− −	− −
core-edge network	+	+	+ + +	+ + +
mixed-use districts	+	+	+ + +	+ + +
median travel	+ +	+ +	+ +	+ +
network circuity	+ + +	+ + +	− − −	− −
non-freeway arterial fraction	+ + +	+ +	+ + +	+ + +
roadway roughness	− −	−	+ + +	+ +

Association according to correlation ρ: $(0,0.05]$ +; $(0.05, 0.15]$++, > 0.15 +++; $[-0.05, 0)$ -; $[-0.15, -0.05)$ --, ≤ -0.15 ---.

The directionality of the associations between these MFD shape parameters and the location factors is closely aligned between model prediction and observations (Table II). An important observation is that a particular location factor's association with critical density and network capacity is not always in the same direction, indicating a potential tradeoff in traffic controls. For example, greater road network connectivity and development level may accommodate more vehicles before the congestion sets in, as indicated by their positive association with critical density, however, they may also decrease the network capacity at the same time, as indicated by their negative association with network capacity. On the other hand, some location attributes pertinent to urban/transportation planning, such as mixed-use development, may contribute to enhanced network performance by increasing both network capacity and critical density. The underlying mechanism from the data-driven associations observed here warrants further research to confirm a causal relationship, before such insights can be applied in practice.

V. CONCLUSIONS

Macroscopic fundamental diagram is a parsimonious modeling tool used in urban traffic management for capturing the interrelationship between vehicular flow, density, and speed at a network-wide level. However, in practice, due to historical data limitations, empirically derived MFD models are sparse in the literature especially for U.S. cities. Leveraging large-scale and granular census-tract-level flow and density data derived from vehicle probes, this paper has presented the first application of machine learning methods to both deriving MFD models and interpreting the important location factors underlying different MFD shapes of urban networks across the entire United States.

Among the four machine learning methods tested, XGBoost is found to deliver the best performance to predict the network traffic flow for given vehicular density and location attributes. In particular, predictions from XGBoost effectively capture both local flow values of a given network density and the key traffic control parameters related to MFD shape, i.e., critical density and network capacity.

The interaction Shapley Additive explanation (SHAP) values were used to determine the importance of and understand location factors, such as land use, transportation infrastructure, and network topology, that influence the shape of MFD curves. We find top-ranking factors are mostly related to network topology, transportation infrastructure, and land use, whereas demand and trip related factors are ranked at the bottom. The ranking of these location factors is largely aligned with the literature.

The directionality of the associations between MFD shape parameters (network capacity and critical density) and location factors have good agreement between model predictions and observations, both confirming the model's ability to capture changes of MFD shapes across locations and revealing potential synergistic and tradeoff effects of land use and network design to be considered in transportation and land use planning.

The analysis framework developed in this work can generate data-driven MFDs and a deeper understanding of their shape dependence on network, infrastructure, and land use characteristics, which can be used by transportation authorities to derive and optimize location-specific MFDs facilitating more informed management and planning decisions at the network level.

ACKNOWLEDGMENT

This research was funded by the Federal Highway Administration (FHWA) Office of Transportation Policy Studies, under Interagency Agreement 693JJ318N300068 with the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory. Any statement or results in this paper reflects the authors' perspectives and not necessarily the point of view of FHWA.

REFERENCES

[1] N. Geroliminis and C. F. Daganzo, "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings," *Transportation Research Part B: Methodological*, vol. 42, no. 9, pp. 759–770, Nov. 2008, doi: 10.1016/j.trb.2008.02.002.

[2] C. F. Daganzo and N. Geroliminis, "An analytical approximation for the macroscopic fundamental diagram of urban traffic," *Transportation Research Part B: Methodological*, vol. 42, no. 9, pp. 771–781, Nov. 2008, doi: 10.1016/j.trb.2008.06.008.

[3] C. F. Daganzo, "Urban gridlock: Macroscopic modeling and mitigation approaches," *Transportation Research Part B: Methodological*, vol. 41, no. 1, pp. 49–62, Jan. 2007, doi: 10.1016/j.trb.2006.03.001.

[4] C. F. Daganzo, V. V. Gayah, and E. J. Gonzales, "Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability," *Transportation Research Part B: Methodological*, vol. 45, no. 1, pp. 278–288, Jan. 2011, doi: 10.1016/j.trb.2010.06.006.

[5] J.-T. Girault, V. V. Gayah, I. Guler, and M. Menendez, "Exploratory Analysis of Signal Coordination Impacts on Macroscopic Fundamental Diagram," *Transportation Research Record*, vol. 2560, no. 1, pp. 36–46, Jan. 2016, doi: 10.3141/2560-05.

[6] C. F. Daganzo, V. V. Gayah, and E. J. Gonzales, "The potential of parsimonious models for understanding large scale transportation

systems and answering big picture questions," *EURO J Transp Logist*, vol. 1, no. 1, pp. 47–65, Jun. 2012, doi: 10.1007/s13676-012-0003-z.

[7] N. Geroliminis, J. Haddad, and M. Ramezani, "Optimal Perimeter Control for Two Urban Regions With Macroscopic Fundamental Diagrams: A Model Predictive Approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 348–359, Mar. 2013, doi: 10.1109/TITS.2012.2216877.

[8] J. Haddad and N. Geroliminis, "On the stability of traffic perimeter control in two-region urban cities," *Transportation Research Part B: Methodological*, vol. 46, no. 9, pp. 1159–1176, Nov. 2012, doi: 10.1016/j.trb.2012.04.004.

[9] A. Loder, M. C. J. Bliemer, and K. W. Axhausen, "Optimal pricing and investment in a multi-modal city — Introducing a macroscopic network design problem based on the MFD," *Transportation Research Part A: Policy and Practice*, vol. 156, pp. 113–132, Feb. 2022, doi: 10.1016/j.tra.2021.11.026.

[10] N. Zheng, R. A. Waraich, K. W. Axhausen, and N. Geroliminis, "A dynamic cordon pricing scheme combining the Macroscopic Fundamental Diagram and an agent-based traffic model," *Transportation Research Part A: Policy and Practice*, vol. 46, no. 8, pp. 1291–1303, Oct. 2012, doi: 10.1016/j.tra.2012.05.006.

[11] N. Zheng and N. Geroliminis, "On the Distribution of Urban Road Space for Multimodal Congested Networks," *Procedia - Social and Behavioral Sciences*, vol. 80, pp. 119–138, Jun. 2013, doi: 10.1016/j.sbspro.2013.05.009.

[12] J. Ortigosa, M. Menendez, and V. V. Gayah, "Analysis of Network Exit Functions for Various Urban Grid Network Configurations," *Transportation Research Record*, vol. 2491, no. 1, pp. 12–21, Jan. 2015, doi: 10.3141/2491-02.

[13] J. Ortigosa, V. V. Gayah, and M. Menendez, "Analysis of one-way and two-way street configurations on urban grid networks," *Transportmetrica B: Transport Dynamics*, vol. 7, no. 1, pp. 61–81, Dec. 2019, doi: 10.1080/21680566.2017.1337528.

[14] M. Yildirimoglu and N. Geroliminis, "Approximating dynamic equilibrium conditions with macroscopic fundamental diagrams," *Transportation Research Part B: Methodological*, vol. 70, pp. 186–200, Dec. 2014, doi: 10.1016/j.trb.2014.09.002.

[15] Z. Zhang, S. A. Parr, H. Jiang, and B. Wolshon, "Optimization model for regional evacuation transportation system using macroscopic productivity function," *Transportation Research Part B: Methodological*, vol. 81, pp. 616–630, Nov. 2015, doi: 10.1016/j.trb.2015.07.012.

[16] C. F. Daganzo and N. Geroliminis, "An analytical approximation for the macroscopic fundamental diagram of urban traffic," *Transportation Research Part B: Methodological*, vol. 42, no. 9, Art. no. 9, Nov. 2008, doi: 10.1016/j.trb.2008.06.008.

[17] N. Geroliminis and C. F. Daganzo, Eds., *Macroscopic modeling of traffic in cities*. 2007.

[18] A. Loder, L. Ambühl, M. Menendez, and K. W. Axhausen, "Understanding traffic capacity of urban networks," *Scientific Reports*, vol. 9, no. 1, Art. no. 1, Nov. 2019, doi: 10.1038/s41598-019-51539-5.

[19] L. Ambühl, A. Loder, M. C. J. Bliemer, M. Menendez, and K. W. Axhausen, "A functional form with a physical meaning for the macroscopic fundamental diagram," *Transportation Research Part B: Methodological*, vol. 137, pp. 119–132, Jul. 2020, doi: 10.1016/j.trb.2018.10.013.

[20] P. Sekuła, N. Marković, Z. Vander Laan, and K. F. Sadabadi, "Estimating historical hourly traffic volumes via machine learning and vehicle probe data: A Maryland case study," *Transportation Research Part C: Emerging Technologies*, vol. 97, pp. 147–158, Dec. 2018, doi: 10.1016/j.trc.2018.10.012.

[21] D. Schrank, B. Eisele, T. Lomax, and J. Bak, "Appendix A: Methodology for the 2015 urban mobility scorecard," Texas Transportation Institute, Technical report, 2015.

[22] Kaveh Farokhi Sadabadi; Zachary Vander Laan; Przemyslaw Sekula; Ling Jin; Xiaodan Xu; Mahyar Amirgholy; etc., "Estimating the Macroscopic Fundamental Diagram Using Probe Data: A Machine Learning Approach." submitted to TRB 2023.

[23] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[24] N. Popovich *et al.*, "A methodology to develop a geospatial transportation typology," *Journal of Transport Geography*, vol. 93, p. 103061, May 2021, doi: 10.1016/j.jtrangeo.2021.103061.

[25] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[26] F. Pedregosa *et al.*, "Scikit-Learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. null, pp. 2825–2830, Nov. 2011.

[27] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[28] D. Ruppert, *The elements of statistical learning: data mining, inference, and prediction*. Taylor & Francis, 2004.

[29] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." arXiv, Jan. 29, 2017. doi: 10.48550/arXiv.1412.6980.

[30] K. Fujimoto, I. Kojadinovic, and J.-L. Marichal, "Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices," *Games and Economic Behavior*, vol. 55, no. 1, pp. 72–99, Apr. 2006, doi: 10.1016/j.geb.2005.03.002.

[31] L. Jin *et al.*, "What Makes You Hold on to That Old Car? Joint Insights From Machine Learning and Multinomial Logit on Vehicle-Level Transaction Decisions," *Frontiers in Future Transportation*, vol. 3, 2022, Accessed: Jul. 27, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/ffutr.2022.894654

[32] Texas Transportation Institute, "Urban Mobility Report." https://mobility.tamu.edu/umr/.

[33] G. Boeing, "A multi-scale analysis of 27,000 urban street networks: Every US city, town, urbanized area, and Zillow neighborhood," *Environment and Planning B: Urban Analytics and City Science*, vol. 47, no. 4, pp. 590–608, May 2020, doi: 10.1177/2399808318784595.

[34] J. A. Laval and F. Castrillón, "Stochastic approximations for the macroscopic fundamental diagram of urban networks," *Transportation Research Part B: Methodological*, vol. 81, pp. 904–916, Nov. 2015, doi: 10.1016/j.trb.2015.09.002.

APPENDIX

TABLE A1. SUMMARY OF TRANSPORTATION SUPPLY AND DEMAND FEATURES USED FOR MFD PREDICTION.

Variable	Description	Source
pct agr. land	Fraction of land area is agriculture land	(Dewitz, 2019)
dev. intensity	Fraction of developed land area	
pct water	Fraction of area is water surface	(U.S. Census Bureau, TIGER, 2019)
Network attributes		
avg. circuity	Total edge length/sum of great circle distances between the network nodes indecent to each edge.	(Boeing, 2017)
dead end proportion	Fraction of network nodes that are dead ends	
self-loop proportion	Fraction of edges with single incident node	
intersection density	The density of intersection nodes per area (nodes/km^2)	
street density	The ratio between street length to the area (1/km)	
avg. street length	Average road edge length in undirected network (m)	
avg. streets per node	Average number of streets that emanate from each node	
avg. edge per node	Average number of inbound and outbound edges incident to the nodes	
edge count	Count of network edges from OSM, normalized by lane miles (#/ln-mile)	
node count	Count of network nodes from OSM, normalized by lane miles (#/ln-mile)	
avg. centrality	Average of all degree centralities in the network, where centrality is defined as the fraction of nodes that each node is connected to	
avg. neighbor edge degree	Average degree of nodes in the neighborhood of each node	
pct full access control	Fraction of lane miles with full access control	(BTS NTAD, 2021)
pct partial access control	Fraction of lane miles with partial access control	
avg. iri	Average of International Roughness Index (IRI) all lane miles (inches/mile)	
avg speed limit	Average speed limit (weighted by lane mile) (mph)	
pct highways	Fraction of lane miles are highway (FHWA class 1 and 2)	
pct local roads	Fraction of lane miles are local roads (FHWA class 6 and 7)	
pct midsize roads	Fraction of lane miles are arterials (FHWA class 3 - 5)	
pct signal coordination	Fraction of road with coordinated signal (weighted by road count)	
lane-miles per sq.km	The ratio between total lane miles and land areas (lane mile/km^2)	
total lane miles	Total lane miles from all FHWA road classes (mile)	
road grade	Average road grade of all roadways (%)	
Aggregated traffic		
avg. daily flow	Average daily traffic volume for all traffic, normalized by lane miles (veh/lane/day)	(BTS NTAD, 2021)
avg. daily comb. trucks	Average daily traffic volume for combination trucks, normalized by lane miles (veh/lane/day)	
avg. daily single-unit trucks	Average daily traffic volume for single-unit trucks, normalized by lane miles	
Trip Generation/demand characteristics		
job density	Jobs per land area (jobs/km^2)	(Census Bureau, LODES, 2017)
population density	Population per land area (person/km^2)	(Census Bureau, ACS, 2018)
pct trips<1.3 mi	Fraction of commute trips within 1.3 miles	(Census Bureau, LODES, 2017)
pct trips 1.3-3 mi	Fraction of commute trips between 1.3 miles and 3 miles	
pct trips 3-8 mi	Fraction of commute trips between 3 miles and 8 miles	
pct trips >8mi	Fraction of commute trips above 8 miles	
trip sink magnitude	Trip sink magnitude = number of work trip destinations/number of work trip origins (homes)	
jobs-housing balance	Total jobs/total population by tracts	(Census Bureau, LODES, 2017) (Census Bureau, ACS, 2018)
lane-meters per capita	Total lane distance/population (m/person)	(BTS NTAD, 2021) (Census Bureau, ACS, 2018)

TABLE A2. DESCRIPTIONS OF LOCATION FACTORS DERIVED AND THEIR COMPONENTS.

Factor	Component
freeway	High freeway fraction and traffic volume
development level	More developed land with dense and short streets
network connectivity	More streets are connected
network complexity	More edges and nodes per lane-mile available in a network
core-edge network	Featuring a few nodes with high centrality
mixed-use districts	Mix of high job and residential density
median travel	Many trips between 3 and 8 miles
network circuity	The road network is more circular, less straight lines
job hub	Industry areas with high employment and low residential population
bike potential	Many trips between 1.3-3 miles
walk potential	Many trips under 1.3 miles
non-freeway arterial	More streets are non-freeway arterials
roadway roughness	High roughness and steep roads