A Data-Driven Approach to Characterize the Impact of Connected and Autonomous Vehicles on Traffic Flow

Amir Bahador Parsa [a,1], Ramin Shabanpour [a], Abolfazl (Kouros) Mohammadian [a], Joshua Auld [b], Thomas Stephens [b]

[a] *University of Illinois at Chicago, 842 W Taylor Street, Chicago, IL 60607, USA*
[b] *Argonne National Laboratory, 9700 S Cass Ave Lemont, IL 60439, USA*

Abstract

The current study aims to present a model to characterize changes in network traffic flows as a result of implementing connected and autonomous vehicle (CAV) technology based on traffic network and built-environment characteristics. To develop such a model, first, POLARIS agent-based modeling platform is used to predict changes in average daily traffic (ADT) under CAVs scenario in the road network of Chicago metropolitan area as the dependent variable of the model. Second, a comprehensive set of variables and indicators representing network characteristics and urban structure patterns are generated. Three machine learning models namely K-Nearest neighbors, Random Forest, and eXtreme Gradient Boosting are developed and validated to establish the relationship between network characteristics and changes in ADT under CAVs scenario. The estimated models are found to yield acceptable performance. In addition, SHapley Additive exPlanations (SHAP) analysis tool is employed to investigate the impact of important features on changes in ADT, which discloses the most important link properties, network features, and demographic information in predicting change in ADT under the analyzed CAVs scenario.

Keywords: Connected and Autonomous Vehicles, POLARIS, Traffic Flow, Machine Learning

---

1. Introduction

Emergence of the connected and autonomous vehicles (CAVs) is a controversial topic in transportation community since they are expected to revolutionize both human mobility and goods transport in near future. In the United States, it is predicted that penetration of CAVs in light-duty-vehicle fleet will be up to 24.8% by the year 2045, if the technology price annually drops by 5%, and Americans' willingness to pay (WTP) increases by 5% in each year (*1*). Accordingly, with 10% annual price reductions and WTP increases, penetration of CAVs can reach up to 87.2% (*1*). It is also reported in another study that, if CAVs prices decrease at rates of 15% or 20% per year, it is expected that their market share will be homogeneous and near 100% by the year 2050 (*2*).

---

[1] Corresponding author. Tel: +1 (312) 996-9840.
Email addresses: aparsa2@uic.edu (A. B. Parsa), rshaba4@uic.edu (R. Shabanpour), kouros@uic.edu (A. K. Mohammadian), jauld@anl.gov (J. Auld), tstephens@anl.gov (T. Stephens).

Aligned with the rapid technological advancements in this area, most of major car companies have announced that they will release their fully autonomous vehicles in the next decade (*3*, *4*).

Given that CAVs have the potential to substitute the current vehicle fleet (*5*), a growing number of studies have focused on evaluating the impact of this new technology on different aspects of transportation systems. One aspect that is expected to be substantially affected by emergence this technology is travel demand (*6*). It is predicted that an AV fleet size of only one-third the number of private vehicles would be enough to meet the demand generated today (*7*). Potential change in peoples' preferences towards their vehicle ownership, mode of travel, and timing and sequence of travels are some examples of impacts of CAVs on travel demand. Travel safety is another dimension which is expected to be greatly affected as vehicle to vehicle communication systems can reduce the chance of collision of vehicles (*8*). There are several studies evaluating the effects of CAVs on travel safety (e.g., *6*–*8*). Most of these studies have reported a considerable enhancement of safety as a result of CAVs deployment (*11*–*13*). A report published by KPMG indicates that about 90% of all types of vehicle accidents can be eliminated when CAVs substitute current vehicle fleet (*14*). Quite a few studies have also investigated the impact of connected vehicles or CAVs on energy consumption and emission and found inconclusive results with respect to environmental impacts of the technology (*15*–*18*).

Network traffic condition is another major dimension of transportation system which is anticipated to be affected by CAVs technology (*19*–*22*). Analysis of flow-density diagram has shown that increasing partial penetration of CAVs can result in more stable traffic stream (*15*). Stability of traffic is found to be higher under CAVs scenario since automation and connection between vehicles can prevent shockwave formation (*23*). Regarding congestion, researchers found out that CAVs can benefit travel time through smoothing the traffic (*15*, *24*, *25*). Analysis of capacity under CAVs scenarios indicates that CAVs penetration rate of 75% increases the capacity by 25-35% (*14*). In another study, impact of CAVs on heterogenous traffic flow is simulated under different penetration rates. It is reported that by increasing CAVs penetration rate up to 30%, capacity increases at a slow pace, and passing that penetration rate will result in faster capacity enhancements (*26*). A 50% penetration rate of CAVs can increase vehicle miles traveled (VMT) by 20%. Increasing penetration rate to 95% can result in 35% increase in VMT (*14*).

Studies focusing on the impact of CAVs on transportation network are suffering from dearth of CAVs historical data, especially at the large scale, which can certainly affect reliability and accuracy of their results. Recently, a number of transportation simulation platforms have started to incorporate vehicle automation and connectivity features into their simulation process. For instance, several researchers have used VISSIM to simulate the impact of AVs or CAVs on highway capacity (*14*), car following behavior (*27*), emergency evacuation (*28*), etc. Zhang and Cassandras combined MATLAB and VISSIM to simulate the impact of CAVs on performance of a single urban intersection (*16*). To cope with limitation of microscopic simulation models, Talebpour and Mahmassani, proposed a novel acceleration framework as an alternative (*23*). Amoozadeh et al. employed VENTOS simulation framework to analyze impact of CAVs on different aspects of transportation system (*29*). In addition, other researchers proposed new simulation frameworks such as microscopic simulation framework of Rios-Torres and Malikopoulos to understand interaction of CAVs and human driven vehicles (HVs) at on-ramp merging area (*15*), and a java-based algorithm by Yang et al., to predict total flow and demand ratio of CAVs at intersections (*21*).

POLARIS, as an advanced transportation simulation framework, is another simulation platform which is recently equipped with new modules to incorporate the simulation of CAVs

(*30*). The framework is developed by Argonne National Laboratory for Chicago and Detroit regions. The POLARIS modeling suite is an open-source agent-based modeling platform specifically designed for simulating large-scale transportation systems. The platform has been used to successfully simulate ITS interactions with regional demand, statewide long-distance passenger travel, and evacuations (*31*). Polaris is designed as a continuously integrated activity-based model and network supply model, where individuals plan and schedule their activities dynamically, engage in simulated travel, and re-plan activities on the fly due to changing traffic conditions, new information or external control. The dynamic, integrated nature of POLARIS means that it is well suited for simulating vehicle connectivity and automation and the impact on individual travel behavior.

Using CAVs simulation results in POLARIS (*32*), the current study aims to present a data-driven model to relate changes in network traffic flows as a result of implementing connected and autonomous vehicle (CAV) technology to traffic network and built-environment. It is worthwhile to note that the objective of this study is not assessing the impacts of CAVs, rather it aims to provide a data-driven model to model traffic flow changes as a function of network characteristics and built-environment factors. The proposed model can be applied in other geographical contexts where a CAV-based network simulator is not available. To develop such a model, we used results of simulations of CAVs in the Chicago metropolitan area, which were generated by POLARIS agent-based platform (*32*), specifically taking the changes in average daily traffic (ADT) as the dependent variable of the model. We have also integrated several other data sources along with feature engineering through link-based analysis to train three powerful machine learning models, K-Nearest Neighbors (KNN), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost) to find out and analyze significant features and their level of importance in predicting changes in ADT of links under fully CAVs scenario. It is worth noting that although frequently-used machine learning techniques have performed very well in transportation studies (*33*), more advanced techniques such as deep learning and XGBoost along with a powerful analysis tool, SHAP, are recently employed and resulted in more robust and great performance (*34*–*36*).

The remainder of this paper is organized as follows. First, different sources of data and feature generation are described in detail. Second, machine learning techniques employed in this study are explained in the methodology section. Then, in the results section, final models are presented and performance of them are compared. Finally, conclusion and limitations of this study are discussed.

2. Data

2.1. POLARIS Simulation Output

As previously pointed out, one of the major challenges in conducting research on CAVs implications is the lack of historical data. Result of traffic simulation platform under CAVs scenarios can be an acceptable alternative to the historical data. In this study, we use results of simulations (29) of CAVs that were generated using the POLARIS platform that is developed by Argonne National Laboratory for Chicago and Detroit regions is employed (*30*).

In (29) CAVs were represented in POLARIS by modifying several aspects of the simulation to account for the expected impacts. For example, travelers who use an AV are assumed to have a lower value of travel time, due to the reduced burden of driving, so the travel time utility parameters for all choices (mode, timing, destination, etc.) within the demand models were reduced for AV drivers based on literature review, with the value of time ranging from 100% down to 50% of the current value. Traffic flow impacts were represented using empirically estimated link capacity changes with cooperative adaptive cruise control (CACC) penetration from (*37*). This function relates increases in link capacity to the penetration of CACC-equipped vehicles in a vehicle stream. For more information about the implementation of CAV scenario in the POLARIS framework, the reader is referred to (*32*).

In this study, we used POLARIS simulation results for two extreme scenarios: 0% penetration of CAVs (base scenario) and 100% penetration of CAVs (CAVs scenario) and calculated the difference of traffic flow for links of the Chicago network between these two scenarios. Traffic flow of links is simulated for a duration of 24 hours under the two scenarios. The daily traffic flows are referred to as ADT in this study, assuming that the POLARIS simulation results for a whole day period is a representative of the traffic condition during the year. Accordingly, the target variable is calculated through **Equation 1**.

$$\Delta\,\mathrm{ADT} = \mathrm{ADT}_{\mathrm{CAVs\ scenario}} - \mathrm{ADT}_{\mathrm{base\ scenario}} \tag{1}$$

Total number of 22,465 links from Chicago traffic network are considered in the POLARIS platform to generate traffic flow. **Figure 1** shows the average value of change in ADT across different five road types: Freeway (8.5% of roads), Expressway (1.1% of roads), Ramp (9.3% of roads), Major & Minor (71.9% of roads), Collector & Local (9.2% of roads). Based on this figure, average change in ADT is significantly higher for freeway and expressway. Furthermore, classifying links into central business district (CBD) and non-CBD groups displays a tangible difference between the average change in ADT in links located inside and outside the CBD.

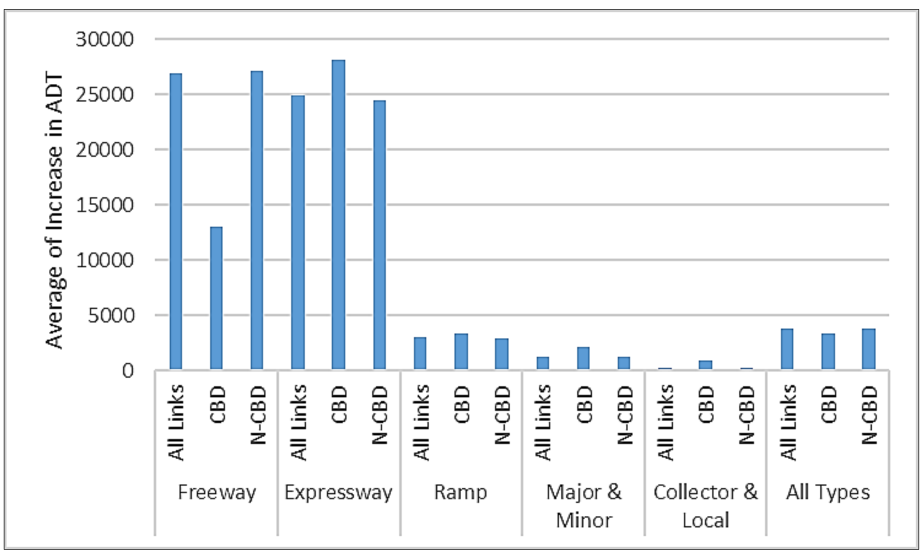

Figure 1. Average of increase in ADT under CAVs scenario

Figure 2 displays the change in ADT across the study area. According to this figure, freeways and expressways connecting downtown to suburban areas are impacted more than the other road types by CAVs scenario. That means, by implementing CAVs technology, changes in ADT of these expressways and freeways can exceed 32000 vehicles per day.

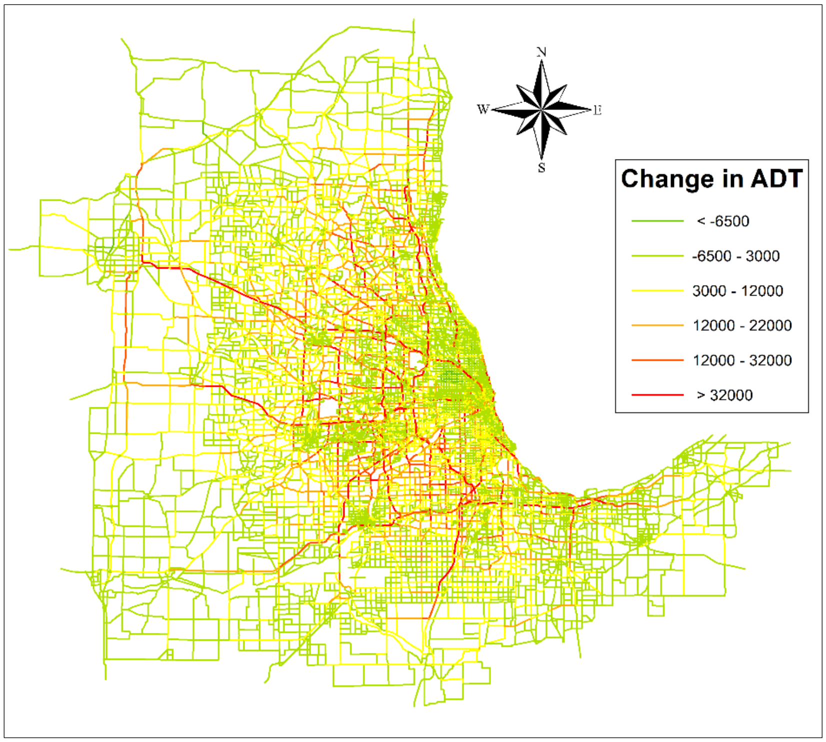

Figure 2. Change in ADT across the study area

2.2. Link Properties Data

Since our target variable, change in ADT, is a link-based parameter, all the features should be generated at the link level as well. Link properties include type of road, slope, length, and number of lanes. Intuitively, road type and other features such as connectivity are expected to have meaningful impacts on traffic flow. Finally, number of lanes is another important attribute of links which is used in training the models. **Figure 3** shows the proportion of links with different number of lanes and the average of target variable for links with different number of lanes. Based on this figure, most of links in the data have 2 lanes. Accordingly, increasing number of lanes can increase the change in ADT.

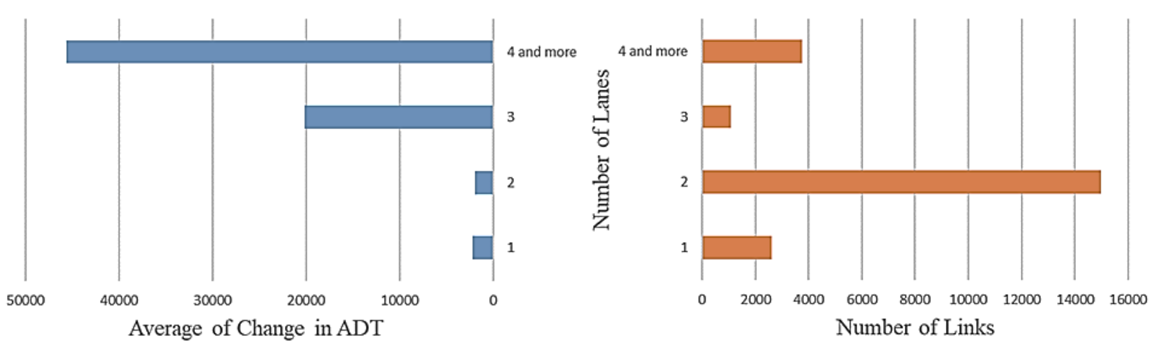

Figure 3. Number of lanes and average change in ADT

2.3. Network Data

Features generated from traffic network of Chicago include connectivity index, distance to CBD, road density, and intersection density. Connectivity index is a feature which is generated for each link to represent the role of the link in creating linkage in the traffic network. It is expected that increasing number of links connected to the start node and end node of a link increases connectivity of that link. Therefore, connectivity index is defined through **Equation 2** for each link.

$$Connectivity\ index = {}^{(\#\ of\ links\ connected\ to\ the\ start\ node\ +\ \#\ of links\ connected\ to\ the\ end\ node)}/_{Link\ length} \quad (2)$$

Distance to CBD is another traffic network feature which is generated based on the distance between each link and the CBD of Chicago. To do so, first centroid of Chicago CBD and all the links are specified in the ArcMap and then the distance between the centroid of each link and centroid of CBD is calculated and assigned to the links.

Finally, road density and intersection density are two traffic network features that are generated using the Environmental Protection Agency's (EPA) Smart Location Database (*38*). EPA is a comprehensive source of data which includes demographic, employment, and built environment information for every census block group in the US. As mentioned earlier, all variables in this study should be prepared at the link level. Therefore, attributes of each census block group are assigned to the links which are passing through that block group. However, since there are many links which pass through multiple block groups, weighted average of attributes of those block groups is assigned to the link passes through them.

2.4. Demographic Data

Demographic data is another category that includes some features such as population, vehicle ownership, job per household, job density within 45-minute drive, etc. Similar to road and intersection density features, these features are generated from the EPA Smart Location Database. Hence, they are assigned from block groups to the links, as was done with the density of roads and intersections.

2.5. Transportation Data

Another type of data used in this study is transportation related data. Although there are several transportation related features in the EPA Smart Location Database, trip equilibrium index is selected to be used in model training. Trip equilibrium index is generated by calculating trip productions and trip attractions of block groups in such a way that values closer to one indicate that trip making at block group level is more balanced. Although we tested several transportation-related variables in the EPA Smart Location Database, only trip equilibrium index is found to be significant in the models.

2.6. Land-Use Data

Another source of data used in this study is land use data provided by Chicago Metropolitan Agency for Planning (CMAP) which includes very detailed land-use information for the Chicago metropolitan area. CMAP land-use types can be divided into eight groups of 1) Residential, 2) Commercial, 3) Institutional, 4) Industrial, 5) Transportation, communication, utilities, and waste land uses, 6) Agriculture, 7) Open space, and 8) vacant/under construction. In order to assign land use variables to the links, a comprehensive GIS-based analysis has been conducted and different sizes of buffer area are created and tested around links. Having analyzed different sizes of buffer area, an area which covers 150 meters around a link is selected as the preferred size of buffer area for this study. Accordingly, for each link's buffer area, percentage of area which is covered by each land use type is calculated and assigned to that link. Further, **Table 1** shows the final set of explanatory variables used in the next step to train the models.

Table 1. Description of explanatory variables

Variable	Description	Mean
Link Properties		
Freeway	1: if link type is freeway; 0: otherwise	0.08
Expressway	1: if link type is Expressway; 0: otherwise	0.01
lanes	Number of lanes	2.32
Network		
Dist_CBD	Distance from centroid of link to the centroid of CBD	32673.4
Connectivity	Role of link in making connection between links of network	0.027
Road_den	Total road network density	18.24
Intersect_den	Street intersection density	84.91
Demographic		
Job	Jobs within 45 minutes auto travel time	264676
HH_1veh	Number of households in block group that own 1 auto	260.87
HH_2veh_	Number of households in block group that own 2 or more auto	355.44
Pop_work_aged	Percent of population that is working aged	0.77
G_pop_den	Gross population density (people/acre) on unprotected land	14.53
Jobs_HH	Jobs per household	63.52
Entertain_job	Entertainment jobs within a 5-tier employment classification scheme	233.70
Pop	Population of block group	1673.87
Transportation		
Trip_equ_ind	Trip equilibrium index	0.41
Land-use		
Residential	Area of buffer zone around link covered by residential land use	0.024
Commercial	Area of buffer zone around link covered by commercial land use	0.017
Transport	Area of buffer zone around link covered by transportation related land use	0.011

3. Methodology

Three Machine Learning techniques, namely, K-Nearest Neighbors (KNN), Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) are employed in this study due to their high estimation accuracy compared to the other ML and statistical models. A brief introduction to these models is provided in the following sub-sections.

3.1. K-Nearest Neighbors

One of the most popular supervised machine learning techniques, which is widely used for classification and regression, is the K-Nearest Neighbors technique. In this study, the KNN regression algorithm is used in which the output is a continuous value (change in ADT).
Based on the training data points, which are described by multiple attributes, a feature space is formed, and each record is positioned in this space. Then, each unknown record (i.e., a data point from test data) is located in the feature space based on the value of its attributes and the KNN

technique looks for the k nearest neighbors for this record in the training data points. Thus, the value of target variable for this record is predicted based on the arithmetic average of the value of target variable of those data points which are selected as the k nearest neighbors.

To measure the distance in order to find the closest (i.e., most similar) data points, different distance metrics could be used such as Euclidean distance which is one of the famous ones. **Equation 3** represents the Euclidean distance between two points of X_1 = $(x_{11}, x_{12}, \ldots, x_{1n})$ and X_2 = $(x_{21}, x_{22}, \ldots, x_{2n})$ with n attributes.

$$\text{Euclidean distance } (X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \quad (3)$$

Finally, performance of the model is evaluated by comparing true value of test data points to the values which are predicted for these test data points by the model.

3.2. Random Forest

Random Forest (RF) is a Machine Learning technique which utilizes combination of several random Decision Trees (DTs). In DT technique, during the training process a feature selection method is used in order to choose the best attribute to be used at each node of the tree; this heuristic procedure also determines how to best split the node to two or more branches. Among different functions such as Mean Squared Error (MSE), Friedman_MSE, and Mean Absolute Error (MAE) to measure the quality of a split, the MSE technique is used in DT regressor model of this study which is equal to variance reduction as the feature selection criterion. **Equation 4** presents the MSE function:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \quad (4)$$

In this equation, μ is the average of x_i when i goes from 1 to n. Splitting on nodes is accomplished through reduction of variance in such a way that the weighted variance of lower level nodes should be less that the variance of upper level node.

RF is capable of working with categorical and numerical data. One disadvantages of DT is that they are sensitive to the data on which they are trained. Hence, changing the training data can significantly impact the resulting DT. To this end, aggregating several trees can result in higher accuracy and decrease the probability of overfitting which might happen in an individual tree.

In the RF, a technique called Bootstrap Aggregation which is also known as bagging is used to combine DTs. Bagging is a powerful method which is used to combine machine learning techniques in order to achieve higher accuracy than the individual machine learning technique. That is, different DTs are trained in parallel on different samples, selected randomly with replacement from the data, and the aggregation of these trees would be the output prediction of the RF model.

3.3. eXtreme Gradient Boosting (XGBoost)

Although RF usually performs well by combining a large number of DTs and taking average of their outputs, DTs are generated independently in this technique. On the other hand, a more advanced model called XGBoost which is created from gradient boosted decision trees can improve the model performance through combining DTs in such a way that each new tree is impacted by previously trained trees, and this can help to reduce errors. In this ensemble learning technique, there are more parameters which need to be tuned to maximize model performance. Proper parameter tuning is essential for XGBoost to avoid overfitting or being too complex. It is also worth noting that RFs combine the results at the end of modeling procedure while XGBoost does it along the process.

The parameters which should be tuned for XGBoost are as follows. First, number of iterations which is the number of trees fitted in the model. Second, maximum depth of the tree which is maximum number of splits and increasing this parameter can cause overfitting. Third, subsample which is the fraction of observations randomly selected for the training instances and can prevents overfitting. Forth is the learning rate used to shrink the weights and change the impact of each individual tree at each step which results in a more robust model. Next parameter is colsample_bytree which is subsampling the columns and can help prevent overfitting. The last two parameters are lambda and alpha that are L2 and L1 regularization terms on weights, respectively, and increasing their value makes the model more conservative. In this study parameters are tuned, and their values are as follows. The optimal XGBoost hyper-parameters values after cross-validation process are: Number of iterations: 700, Max Depth: 7, Subsample: 0.8, Colsample bytree: 0.4, Lambda: 1.5, Alpha: 0.2, Learning Rate: 0.02.

3.4. SHapley Additive exPlanations (SHAP)

Interpreting output of machine learning techniques is often challenging. However, SHapley Additive exPlanations (SHAP) is a powerful tool for this which was proposed by Lundberg and Lee (*39*). SHAP is based ongame theory rules (*40*) and local explanations (*41*), and it can provide a means for estimating the contribution of each feature to the output of the model. Given an XGBoost model with a set of N features is used to predict an output $v(N)$, SHAP values are determined using several axioms to allocate the contribution of each feature through **Equation 5.**

$$\phi_i = \sum_{S\subseteq N\{i\}} \frac{|S|!(n-|S|-1)!}{n!}[v(S\cup\{i\}) - v(S)] \tag{5}$$

Where ϕ_i is contribution of feature i in the model output, and it is allocated based on their marginal contribution (*42*). A linear function of binary features g is defined based on an additive feature attribution method shown in **Equation 6** where M is the number of input features and $z' \in \{0,1\}^M$, equals to 1 when a feature is observed, otherwise it equals to 0 (*39*).

$$g(z') = \phi_i + \sum_{i=1}^{M} \phi_i z_i' \tag{6}$$

4. Results

To train the models, 70% of the data is randomly selected for training and the remaining 30% is used to validate the models. In addition, a 5-fold cross-validation procedure is applied on the training data. Therefore, at first the training data is divided to five subsamples randomly, and then four subsamples are used to train the models while the remaining subsample is used as the validation data. We repeated this procedure 5 times so that each subsample is used exactly once as the validation data. This procedure helps us to measure whether a model is performing well consistently.

Validation of three models shows that the KNN model, for which the optimal number of neighbors is found to be six, results in the accuracy of 83.5%, the RF model achieves accuracy of 87.1%, and XGBoost yields the accuracy of 89.7%. Thus, XGBoost outperforms the other two models in terms of accuracy. In the **Figure 4**, true values and predicted values of test dataset are plotted for KNN, RF, and XGBoost techniques.

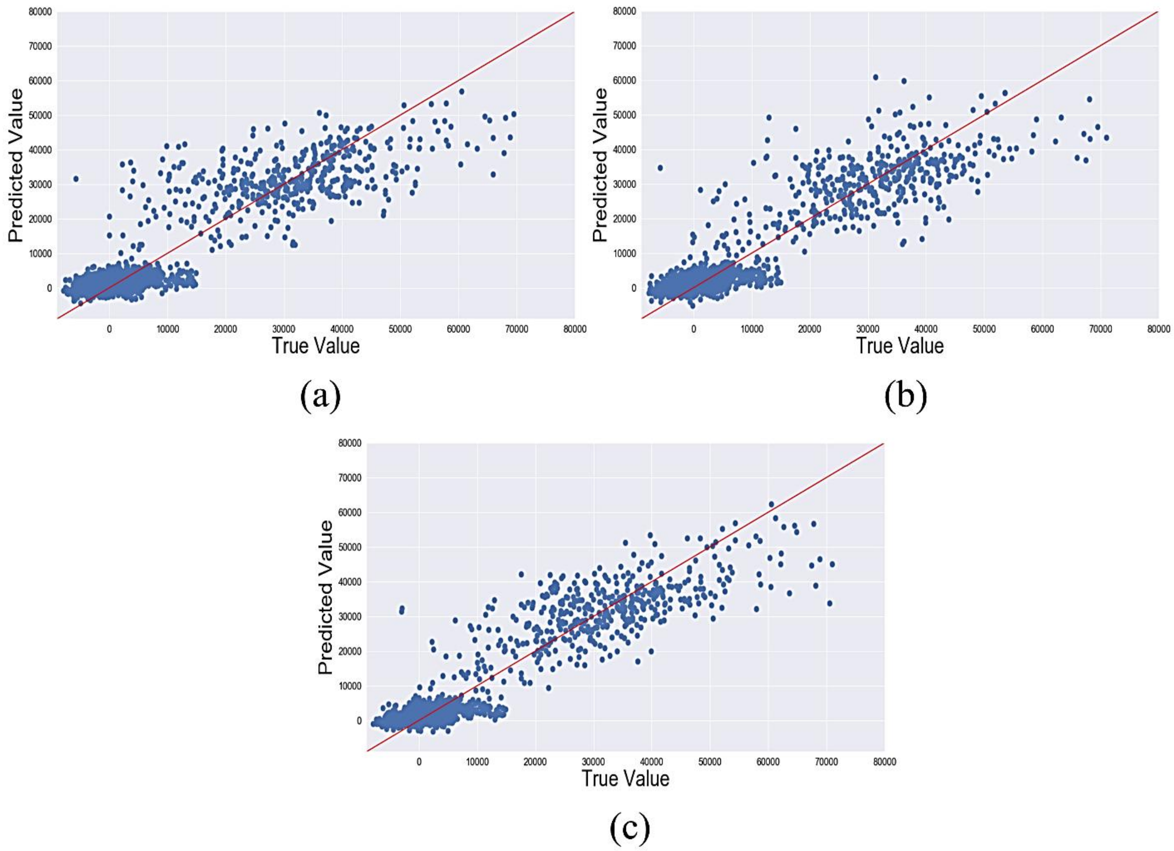

Figure 4. Predicted values against true values: (a) KNN, (b) RF, (c) XGBoost

After training the models, SHAP values of every feature are plotted in the **Figure 5** to show which features are most important for the model as well as how these features can impact the XGBoost model. In this figure, first 11 important features are sorted by the sum of SHAP value magnitudes, then distribution of the impacts each feature has on the model output are displayed using SHAP values. The color spectrum from blue to red represent the magnitude of feature values from low to high, respectively.

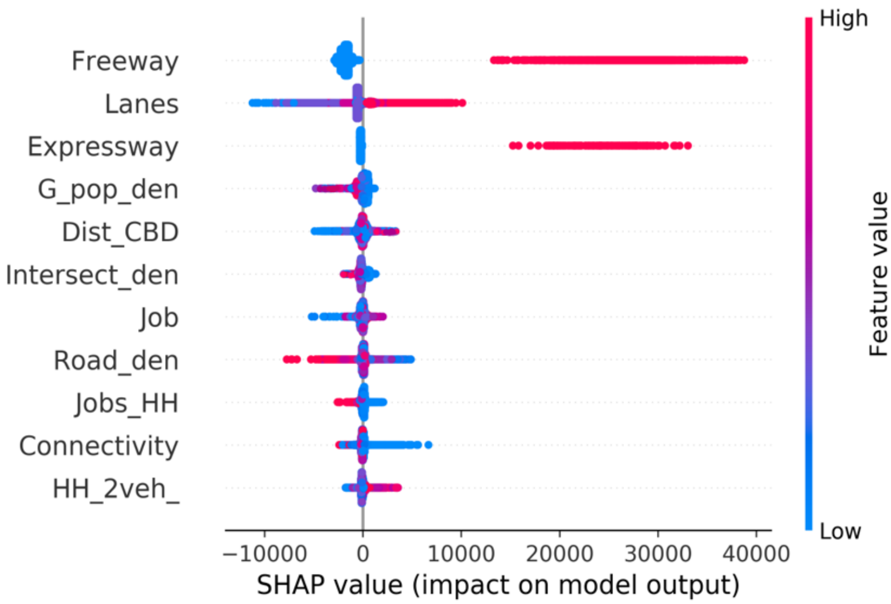

Figure 5. SHAP summary plot

Based on **Figure 5**, link properties including type of the roadway and number of lanes have the highest impact on the target variable (i.e., change in ADT) in such a way that increasing number of lanes and changing road type to Freeway and Expressway can increase likelihood of higher ADT in the CAVs scenario. Interestingly, next most important feature is gross population density so that for the roads passing through zones with denser gross population the change in ADT between base and CAVs scenario can decrease. The next important feature is distance to CBD, and it has a direct impact on the target variable meaning for the roads close to the CBD the change in ADT is less than that of roads far from the CBD. It could stem from that traffic of roadways which are close to the CBD are already higher than other roadways so that the impact of CAVs in increasing the ADT is less for these roads. **Figure 1** can also show the increase in ADT under CAVs scenario is slightly less for CBD roadways.

Intersection density of the zones through which roadways are passing is the next important features. However, based on **Figure 5,** when intersection density is lower, ADT would increase slightly more. Next important feature is number of jobs near the roadway and this feature has a direct impact on the target variable meaning that when there are more job opportunities around a road, the change in ADT would be higher. Road density is the next important feature which has a similar impact to intersection density. The next two features are jobs per household and connectivity, respectively, which have indirect impact on the target variable. That is, for lower values of these features impact of CAVs on change in ADT increases. Finally, according to **Figure 5**, when number of households with two or more vehicles increases in a block group, ADT of roadways passing through this block group is expected to increase more under CAVs scenario.

It is worth noting that the reasons provided in this section are not definite and we tried to analyze features based on the observed data and our understanding about it. In addition, although some features might seem correlated, it doesn't impact the performance of the models, especially in tree-based models.

5. Conclusion

This study presented a data-driven model to relate changes in network traffic flows as a result of implementing CAV technology to characteristics of the traffic network and built environment. To develop such a model, we used changes in ADT under CAVs scenario in traffic network of Chicago metropolitan area, which is generated by POLARIS agent-based platform. Using other sources of data and feature engineering techniques, three machine learning models, KNN, RF and XGBoost, are trained to predict impact of CAVs on traffic flow based on link-based features. Changes in daily traffic flows of traffic network links is an indicator considered in this study and using data-driven methods, it was modeled at the regional level and cross-validated in the same context. This study demonstrates approaches that are useful for identifying the most important factors that influence the changes in traffic flow attributable to widespread adoption of CAVs and for quantifying the importance of each of these factors. We demonstrated these methods using results of previous simulations of a CAVs scenario in the POLARIS (from (*32*)), and we took advantage of different sources of data and powerful machine learning techniques to model the impacts of CAVs on ADT.

It is found that traffic flows will most likely increase in most of the road types in case of fully CAVs scenario. SHAP feature analysis also shows that properties of links have the highest impact on target variable. Gross population density is the next important feature which has an indirect impact on ADT. Next, distance of links from the CBD as well as other network features are the second most important, and finally, attributes of block groups around the links such as demographic, transportation and land uses are, respectively, less important, but still significant features in predicting traffic flow in the CAV scenario analyzed.

Results of this study offer powerful methods that we validated for the Chicago metropolitan area. Future work should test and hopefully validate these methods in analyzing simulations under other conditions, such as different levels of CAVs penetration, or in other geographical contexts, or perhaps for transferring simulation results from one geographical area to others. This is important, since agent-based transportation demand models that model appropriate behaviors and choices for metropolitan areas are difficult to develop and validate, and methods to analyze and generalize results from existing models would be very valuable.

Acknowledgments

The authors gratefully acknowledge the sponsorship of the Systems and Modeling for Accelerated Research in Transportation (SMART) Mobility Laboratory Consortium, an initiative of the Energy Efficient Mobility Systems (EEMS) Program, managed by David Anderson of the Vehicle Technologies Office of the U.S. Department of Energy. This study was conducted by the University of Illinois at Chicago and Argonne National Laboratory, a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC under Contract No. DE-AC02-06CH11357. The authors are solely responsible for the findings of this research which do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

1. Bansal, P., and K. M. Kockelman. Forecasting Americans ' Long-Term Adoption of Connected and Autonomous Vehicle Technologies. *Transportation Research Part A*, Vol. 95, 2017, pp. 49–63. https://doi.org/10.1016/j.tra.2016.10.013.
2. Talebian, A., and S. Mishra. Predicting the Adoption of Connected Autonomous Vehicles : A New Approach Based on the Theory of Di Ff Usion of Innovations. *Transportation Research Part C*, Vol. 95, No. June, 2018, pp. 363–380. https://doi.org/10.1016/j.trc.2018.06.005.
3. Bimbraw, K. Autonomous Cars: Past, Present and Future a Review of the Developments in the Last Century, the Present Scenario and the Expected Future of Autonomous Vehicle Technology. *12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, Vol. 01, 2015, pp. 191–198.
4. Taiebat, M., A. L. Brown, H. R. Safford, S. Qu, and X. Ming. A Review on Energy, Environmental, and Sustainability Implications of Connected and Automated Vehicles. *Environmental science & technology*, 2018, pp. 11449–11465.
5. Sheng, S., E. Pakdamanian, K. Han, B. Kim, P. Tiwari, I. Kim, and L. Feng. A Case Study of Trust on Autonomous Driving. *arXiv: 1904.11007 [cs. HC]*, 2019.
6. Lee, Y.-J., and A. Nickkar. Optimal Automated Demand Responsive Feeder Transit Operation and Its Impact. 2018.
7. Spieser, K., K. Treleaven, R. Zhang, E. Frazzoli, D. Morton, and M. Pavone. Toward a Systematic Approach to the Design and Evaluation of Automated Mobility-on-Demand Systems: A Case Study in Singapore. *Road vehicle automation*, 2014, pp. 229–245.
8. Mudalige, U. P. Fast Collision Detection Technique for Connected Autonomous and Manual Vehicles. *U.S. Patent*, Vol. 2, No. 12, 2013.
9. Arvin, R., M. Kamrani, A. J. Khattak, and J. Rios-Torres. Safety Impacts of Automated Vehicles in Mixed Traffic. *97st Annual Meeting of Transportation Research Board*, 2018.
10. Arvin, R., M. Kamrani, and A. J. Khattak. How Instantaneous Driving Behavior Contributes to Crashes at Intersections: Extracting Useful Information from Connected Vehicle Message Data. *Accident Analysis & Prevention*, Vol. 127, 2019, pp. 118–133.
11. Li, T., and K. M. Kockelman. Valuing the Safety Benefits of Connected and Automated Vehicle Technologies. *Transportation Research Board 95th Annual Meeting*, 2016, pp. 1–22.
12. Li, S. ., Y. Zheng, K. Li, Y. Wu, J. . Hedrick, F. Gao, and H. Zhang. Dynamical Modeling and Distributed Control of Connected and Automated Vehicles: Challenges and Opportunities. *IEEE Intelligent Transportation Systems Magazine*, Vol. 9, 2017, pp. 46–58. https://doi.org/10.1109/MITS.2017.2709781.
13. Talebpour, A., H. S. Mahmassani, and F. E. Bustamante. Modeling Driver Behavior in a Connected Environment and Mobile Wireless Telecommunication Systems. *Transportation Research Record*, 2016. https://doi.org/10.3141/2560-09.
14. Bierstedt, J., A. Gooze, C. Gray, J. Peterman, L. Raykin, and J. Walters. Effects of Next-Generation Vehicles on Travel Demand and Highway Capacity. *FP Think Working Group*, 2014.
15. Rios-torres, J., A. A. Malikopoulos, and S. Member. Impact of Partial Penetrations of Connected and Automated Vehicles on Fuel Consumption and Traffic Flow. *IEEE TRANSACTIONS ON INTELLIGENT VEHICLES*, Vol. 3, No. 4, 2018, pp. 453–462.
16. Zhang, Y., and C. G. Cassandras. The Penetration Effect of Connected Automated Vehicles in Urban Traffic : An Energy Impact Study. *IEEE Conference on Control Technology and Applications*, 2018, pp. 1–6.
17. Greenblatt, J. B., and S. Saxena. Autonomous Taxis Could Greatly Reduce Greenhouse-Gas Emissions of US Light-Duty Vehicles. *Nature Climate Change*, Vol. 5, No. September, 2015. https://doi.org/10.1038/NCLIMATE2685.
18. Ahangari, S., Z. Rashidi Moghaddam, M. Jeihani, C. Chavis, H. Chen, H. Rakha, and K. Kang. Investigating the Effectiveness of an Eco-Speed Control System in the Vicinity of Signalized

Intersections Using a Driving Simulator. 2019.

19. Kidando, E., R. Moses, M. Ghorbanzadeh, and E. E. Ozguven. Traffic Operation and Safety Analysis on an Arterial Highway : Implications for Connected Vehicle Applications. *21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2753–2758.
20. Mahmassani, H. S. M. 50th Anniversary Invited Article — Autonomous Vehicles and Connected Vehicle Systems : Flow and Operations Considerations. *Transportation Science*, No. July 2019, 2016.
21. Yang, K., S. I. Guler, and M. Menendez. Isolated Intersection Control for Various Levels of Vehicle Technology : Conventional , Connected , and Automated Vehicles. *Transportation Research Part C*, Vol. 72, 2016, pp. 109–129. https://doi.org/10.1016/j.trc.2016.08.009.
22. Azizi, L., M. S. Iqbal, and M. Hadi. Estimation of Freeway Platooning Measures Using Surrogate Measures Based on Connected Vehicle Data. *97st Annual Meeting of Transportation Research Board*, No. March, 2018.
23. Talebpour, A., and H. S. Mahmassani. Influence of Connected and Autonomous Vehicles on Traffic Flow Stability and Throughput. *Transportation Research Part C*, Vol. 71, 2016, pp. 143–163. https://doi.org/10.1016/j.trc.2016.07.007.
24. Approach, I. D. M., M. Zhou, X. Qu, S. Jin, A. M. Human, and D. Behavior. On the Impact of Cooperative Autonomous Vehicles in Improving Freeway Merging : A Modified. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 18, No. 6, 2017, pp. 1422–1428. https://doi.org/10.1109/TITS.2016.2606492.
25. Nezafat, R. V., E. Beheshtitabar, M. Cetin, E. Williams, and G. F. List. Modeling and Evaluating Traffic Flow at Sag Curves When Imposing Variable Speed Limits on Connected Vehicles. *Transportation Research Record*, Vol. 2672, 2018, pp. 193–202.
26. Ye, L., and T. Yamamoto. Modeling Connected and Autonomous Vehicles in Heterogeneous Traffic Flow. *Physica A*, Vol. 490, 2018, pp. 269–277. https://doi.org/10.1016/j.physa.2017.08.015.
27. Kockelman, K., P. Avery, P. Bansal, D. Stephen, P. Bujanovic, T. Choudhary, L. Clements, G. Domnenko, D. Fagnant, J. Helsel, M. Levin, J. Li, T. Li, L. Loftus-, A. Nichols, M. Simoni, and R. Hutchinson. Implications of Connected and Automated Vehicles on the Safety and Operations of Roadway Networks : A Final Report. Vol. 7, 2016.
28. Al-Ahad, E., and R. Md Sharikur. Effects of Connected and Autonomous Vehicles on Contraflow Operations for Emergency Evacuation: A Microsimulation Study. *Transportation Research Board 97th Annual Meeting*, 2018.
29. Amoozadeh, M., A. Raghuramu, C. Chuah, D. Ghosal, H. M. Zhang, and J. Rowe. Security Vulnerabilities of Connected Vehicle Streams and Their Impact on Cooperative Driving. *IEEE Communications Magazine*, Vol. 53, No. June, 2015, pp. 126–132. https://doi.org/10.1109/MCOM.2015.7120028.
30. Auld, J., M. Hope, H. Ley, V. Sokolov, B. Xu, and K. Zhang. POLARIS: Agent-Based Modeling Framework Development and Implementation for Integrated Travel Demand and Network and Operations Simulations. *Transportation Research Part C: Emerging Technologies*, Vol. 64, 2016, pp. 101–116. https://doi.org/10.1016/j.trc.2015.07.017.
31. Argonne National Laboratory. POLARIS Transportation System Simulation Tool.
32. Auld, J., O. Verbas, M. Javanmardi, and A. Rousseau. Impact of Privately-Owned Level 4 CAV Technologies on Travel Demand and Energy. *Procedia Computer Science*, Vol. 130, 2018, pp. 914–919. https://doi.org/10.1016/J.PROCS.2018.04.089.
33. Parsa, A. B., H. Taghipour, S. Derrible, and A. (Kouros) Mohammadian. Real-Time Accident Detection : Coping with Imbalanced Data. *Accident Analysis and Prevention*, Vol. 129, No. January, 2019, pp. 202–210. https://doi.org/10.1016/j.aap.2019.05.014.
34. Parsa, A. B., A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian. Toward Safer Highways, Application of XGBoost and SHAP for Real-Time Accident Detection and Feature Analysis. *Accident Analysis & Prevention*, 2020.

35. Parsa, A. B., R. S. Chauhan, H. Taghipour, S. Derrible, and A. Mohammadian. Applying Deep Learning to Detect Traffic Accidents in Real Time Using Spatiotemporal Sequential Data. *arXiv preprint*, Vol. arXiv:1912, 2019.
36. Movahedi, A., and S. Derrible. Interrelated Patterns of Electricity, Gas, and Water Consumption in Large-Scale Buildings. (under Review). *engrXiv*, 2020. https://doi.org/https://doi.org/10.31224/osf.io/ahn3e.
37. Shladover, S. ., C. Nowakowski, X. Lu, and R. Hoogendoorn. Using Cooperative Adaptive Cruise Control (CACC) to Form High-Performance Vehicle Streams. Microscopic Traffic Modeling. *eScholarship. Available at: http://escholarship.org/uc/item/3m89p611* .
38. Ramsey, K., and A. Bell. The Smart Location Database: A Nationwide Data Resource Characterizing the Built Environment and Destination Accessibility at the Neighborhood Scale. *Cityscape*, 2014, pp. 145–162.
39. Lundberg, S. M., and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.
40. Štrumbelj, E., and I. Kononenko. Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowledge and information systems*, 2014, pp. 647–665.
41. Ribeiro, M. T., S. Singh, and C. Guestrin. Why Should i Trust You?: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.
42. Shapley, L. S. A Value for N-Person Games. *Contributions to the Theory of Games*, 1953, pp. 307–317.