International Journal of Transportation Science and Technology 9 (2020) 76–88

Contents lists available at ScienceDirect

International Journal of Transportation Science and Technology

journal homepage: www.elsevier.com/locate/ijtst

Interpretable machine learning approach in estimating traffic volume on low-volume roadways ☆

Subasish Das [a],*, Ioannis Tsapakis [b]

[a] Texas A&M Transportation Institute, College Station, TX 77843, United States
[b] Texas A&M Transportation Institute, San Antonio, TX 78229, United States

ARTICLE INFO

Article history:
Received 26 January 2019
Received in revised form 23 July 2019
Accepted 29 September 2019
Available online 15 October 2019

Keywords:
Traffic volume
Low-volume roads
AADT
Machine learning
Interpretable machine learning

ABSTRACT

Many state and local agencies are currently facing challenges concerning the collection and estimation of traffic volumes, particularly regarding the collection of annual average daily traffic (AADT) on low-volume roads. To overcome these challenges, there is a need to develop new affordable methods to collect data and estimate traffic volume on low-volume roadways. In this study, the research team developed an innovative interpretable machine learning framework and applied it to low-volume roads in Vermont to estimate traffic volumes. This study used several databases (e.g., U.S. Census, the American community survey) to prepare the final dataset for the model development. The findings show that population density and work area characteristic (WAC) density are the best predictors in estimating AADT. The model outcomes show that the machine learning models yield better estimates than the conventional parametric statistical methods. By improving the accuracy of AADT estimations, this study contributed to traffic monitoring and safety improvement, and it can help reduce costs of data collection. This study developed the top five decision rules for three types of low-volume roadways. Stakeholders can use the findings of this study to meet the new requirements pertaining to availability of AADT estimates for low-volume roads. Additionally, the best fit estimates and the developed rules from the current study could enhance the predictive power of the SPF development for the low-volume roadways in Vermont and therefore improve the decision-making process.

1. Introduction

The Moving Ahead for Progress in the 21st Century Act (MAP-21) mandates data collection of traffic volumes to support the development of states' Highway Safety Improvement Programs (HSIPs). The HSIP requires states to use a subset of Model Inventory of Roadway Elements (MIRE) for data collection on all public roadways, including local low-volume roads. However, state and local agencies are currently facing challenges concerning the collection and estimation of traffic volumes on low-volume roads, and particularly the collection of annual average daily traffic (AADT), the average 24-hour traffic volume at a roadway segment over an entire year. The State Departments of Transportation (DOT) use AADT as an important measure of roadway safety, roadway planning, roadway design, traffic operation, pavement maintenance, and air quality

Peer review under responsibility of Tongji University and Tongji University Press.

* Corresponding author.

E-mail address: s-das@tti.tamu.edu (S. Das).

https://doi.org/10.1016/j.ijtst.2019.09.004
2046-0430/

assessment. The DOTs and local transportation agencies use AADT count programs to collect traffic volume data. However, these traffic count programs are primarily focused on higher functional class roadways such as interstate or principal arterials, and traffic counting for low-volume roadways is highly selective and irregular. Additionally, low volume roadways account for a significant portion of the roadways in the U.S. Therefore, improving the estimation of traffic volume on these roadways will help develop low-volume roadway-based safety performance functions, and help improve the overall safety of these roadways.

Several studies have explored different statistical and machine learning models for estimating AADT on low-volume roads. However, an interpretable machine learning framework has not been developed in any of these prior studies. This study aims to mitigate this research gap by applying three robust machine learning models to estimate AADT on low-volume roadways of Vermont. To accomplish the research goal, several data sets were used to prepare a suitable database for the model development: (1) traffic volume counts of Vermont low-volume roadways, (2) geometric features from the road inventory database, (3) block group level demographic and economic variables from U.S. Census and American Community Survey (ACS) data, and (4) distance to major roadways (Interstate and U.S. highways) from the count stations. This study demonstrates the applicability of an interpretable machine learning framework for estimating AADT on low-volume roadways.

This paper is organized as follows. First, there is a review of the literature concerning AADT estimation methods. Next, the concepts of interpretable machine learning are reviewed and the data set is described. The paper concludes with a discussion of results, conclusions, and limitations.

2. Literature review

Researchers have employed many statistical techniques to estimate traffic volumes on various types of roadways. Table 1 provides critical information from previous studies on AADT estimation. Although the discussion of the studies in this literature review is brief, this table provides supplementary information about these studies. Among the plethora of methods and tools, regression analysis is the most frequently used method due to its ability to assess robust structural relationships while maintaining simplicity and interpretability. States that have developed regression models for AADT estimation include Indiana (Mohamad, 1998), Florida (Xia et al., 1999; Shen et al., 1999; Barrett et al., 2001; Pan, 2008), Georgia (Seaver et al., 2000), Kentucky (Zhao et al., 2004; Staats, 2016), and Wyoming (Apronti et al., 2015). In Florida, several studies were conducted to improve modeling performances over the years (Xia et al., 1999; Shen et al., 1999; Barrett et al., 2001; Pan, 2008). Morley et al. used a routing algorithm to refine the AADT estimates (Morley and Gulliver, 2016).

Several research studies applied different machine learning algorithms to estimate AADT. For example, Sharma et al. (2001) applied a neural network (NN) to improve estimation accuracies. To develop the NN model, this study used fifty-five automatic traffic recorder sites located on low-volume rural roads in Alberta, Canada. Dixon (2004) applied a classification and regression tree (CART) to estimate AADT for highways in rural Idaho. This method provided acceptable mean absolute percent errors, which were less than 10 percent for a 10-year forecast. Eom et al. (2006) performed a spatial linear regression to estimate AADT in North Carolina; the predictive capability of this model was significantly better than that of the ordinary regression model. Selby and Kockelman (2011) used the universal kriging method to predict AADT across the Texas roadway network. The results of this method indicated reduced errors over non-spatial regression models. Sun and Das (2015) developed support vector regression (SVR) models for eight different parishes in Louisiana; the SVR models performed better than regression models. This study also conducted a sensitivity analysis on the developed models. By using satellite images and aerial photographs, McCord et al. (2003) performed an image processing application for AADT estimation.

As the current study primarily focuses on estimation accuracy and interpretability, studies associated with the weighting factor and travel demand modeling are not included in the literature review. Several other studies used clustering and other modeling techniques (Gecchelea et al., 2011; Gastaldi et al., 2012; Lowry, 2014). Table 1 provides a brief overview on the state-of-the-art AADT studies.

Although there are many methods of AADT estimation, there is still a need to develop a robust framework for AADT estimation. The current study aims to contribute to the existing AADT estimation literature by developing an interpretable machine learning framework for low-volume roadways in Vermont. The model performance and interpretability provide strong support for the notion of using this method for AADT estimation.

3. Concepts of interpretable machine learning (IML)

Machine learning is a method using training machines to learn the patterns and connections in the data and to utilize this knowledge to make predictions. There are two major categories of machine learning: supervised learning and unsupervised learning. Supervised learning has an identified input and a desired output. The goal of supervised learning is to map the inputs to outputs. Diversely, unsupervised learning is more similar to clustering in that it has no pre-determined targets. The purpose of unsupervised learning is to let the algorithms themselves discover the non-trivial hidden patterns in the data through knowledge extraction. To achieve the goals of both supervised and unsupervised learning, these learning tools utilize machine learning algorithms that represent a set of rules to be followed by machines or computers.

Table 1
AADT estimation studies.

Ref	State	Modeling technique	Variables	Model performance
1	Indiana	Multiple Regression Analysis	Location type, easy access to highways, county population, and total arterial mileage of a county	– R^2 = 0.77 at a 95 percent confidence interval – Mean Squared Prediction Error (MSPR) = 0.051
2	Florida	Multiple Regression Analysis	Number of lanes, functional class, area type, population, housing units, vehicle ownership, employment status, school enrollment in the surrounding area, hotel occupancy, and accessibility to state/nonstate roads.	– The final model explained 63 percent of the ADT variability
3	Florida	Multiple linear regression	Land use, urban population sizes, labor force, per capita income, taxable sales, total lane miles of state roads, number of registered automobiles, and number of lanes	– Statewide model: R^2 = 0.2538 – Rural area model: R^2 = 0.3486 – Small-medium, urban area model: R^2 = 0.6937 – Large metropolitan area (Broward County) model: R^2 = 0.6069
4	Georgia	Multiple linear regression	45 variables in eight major categories: population demographics, transportation, housing, income, urbanization, employment, farming, and education	– R^2 = 0.89 for non-state roads
5	Alberta, Canada	Neural Network	No variable. This is an application of estimating AADT using Neural Network approach from direct sample count of vehicles.	– Two 48-hour counts within a year can reduce the 95th percentile errors to about 25 percent in most cases
6	Florida	Multiple linear regression	Number of lanes, truck factor, functional class, direction of travel, population density, households, hotel/motel rooms, employment, avg. income, distance from a Continuous count site (CCS) to coastlines, metropolitan areas, state borders, highway interchanges, accessibility to interstates, network density, and geographical locations of CCSs.	– R^2 = 0.0941 for January – R^2 = 0.5581 for December (both for rural areas)
7	Ohio	Empirical Analysis	Number of cars, number of trucks, and weighted avg. of car and truck speed	– Compared satellite images and aerial photos to the corresponding ground-based estimates and found smaller differences for images leading to longer equivalent traffic count duration
8	Idaho	Classification and regression tree (CART)	Functional class, county population, annual county population growth rate, and AADT	– CART method provided acceptable mean absolute percent errors, which were less than 10 percent for a 10-year forecast
9	Kentucky	Linear Regression	Population, average per capita income, employment, county-wide total earnings, and licensed drivers	– Urban model: R^2 = 0.0002 – Rural model: R^2 = 0.263–0.41
10	North Carolina	Spatial regression model	Speed, median income, number of lanes, land use, and functional class	– For urban arterials and local roads, weighted least squares and restricted maximum likelihood provided more accurate predictions than ordinary least squares regression
11	Florida	Linear Regression	Socioeconomic, roadway, and land use	– R^2 = 0.378 for rural highway model – R^2 = 0.418 for rural local street model
12	Texas	Universal Kriging for Spatial Prediction	Speed limit, number of lanes, and functional class	– Errors tended to be lower at locations with higher counts and more nearby count locations
13	Venice, Italy	Clustering methods	Volumes, equal shape, and orientation of the coordinate axes	– Estimation errors for passenger cars (9.99–11.07) percent are acceptable, but those for trucks are higher (16.28–18.09) percent
14	Venice, Italy	Fuzzy set theory, Neural networks	Number of lane, traffic volume on a single lane, temporal traffic pattern, and vehicle type	– Model finds traffic volumes collected on weekdays resulted in the most accurate estimates – Recreational roads have larger errors due to higher unpredictability of traffic volumes
15	Idaho	Ordinary least squares (OLS)	Number of lanes, speed limit, and adjacent land use, and centrality	– R^2 = 0.95 – The median absolute percent error of the calibration and validation datasets were 34 percent and 22 percent, respectively

Table 1 (*continued*)

Ref	State	Modeling technique	Variables	Model performance
16	Wyoming	Linear and Logistic Regression	Pavement type, access, land use, population, number of households/units, number of employed civilians, income, employment density, housing unit density, income density, and population density	– R^2 of the model was 0.64, and its percentage root mean square error was 73.4 percent – Models can be used for quick estimation but should not use when high level of prediction accuracy is required
17	Louisiana	Support vector regression (SVR)	Population, demographic characteristics, distance to permanent counts, and number of jobs	– SVR models tend to underestimate high AADT values and somewhat overestimate the low AADT values
18	Kentucky	Ordinary linear regression and linear model with a Poisson distribution and a log link function	Direct access to an expressway, employment and population buffer, distance to population center, accessibility to regional employment centers, land use variables, residential vehicle registrations, and curvature data	– The average AADT estimation errors ranged between 134 percent (overestimated AADTs) to 38 percent (underestimated AADTs)
19	UK	Regression Model	Route importance, road type, AADT on the nearest major road, area type (urban or rural)	– AADT on minor roads and major road show good agreement to the traffic count and fall along the identity line

Conventional statistical modeling is advantageous in terms of its interpretation ability. However, a significant drawback of this method is its pre-determined assumptions that must be preemptively defined, and deviations from these assumptions could result in biased solutions. The machine learning method can effectively handle big data; and in some cases, the entire process is computationally intensive. This method also has excellent flexibility in finding inherent associations in the data without predetermining the assumptions. The basis of the machine learning method is algorithms. Numerous algorithms have been developed to solve problems within many different branches of science and engineering. All algorithms have different benefits and disadvantages as they solve problems throughout different domains.

As the machine learning tools are based on algorithms, these methods are known as black-box method. The limitation of interpretation makes machine learning model difficult to present to the practitioners. There are a significant number of algorithms that can make precise predictions with almost zero interpretation of power. Conversely, the performance of the predictive model is sufficient enough for many practical problems as there would be no severe consequence if an unexpected error occurred. However, for many context sensitive problems, interpretability is crucial to solving the problem. Doshi-Velez and Kim emphasized the need for interpretability due to the incompleteness of the formalization of the problem; this means that the problem not only needs to be correctly predicted, but also explained (Doshi-Velez and Kim, 2017). It is important to note that algorithms are susceptible to pick up any pattern, including bias, in the training data. There is a high likelihood that a model may discriminate against a characteristic or trait in order to maximize the prediction accuracy (Christoph, 2018). The Research Team used open source software R packages (DALEX, and pdp) to perform modeling and data visualizations (Biecek, 2018; Greenwell, 2018) in this study.

3.1. Interpretability using partial dependence (PDP)

In recent years, several interpretable machine learning techniques were introduced by the researchers. The techniques include partial dependence plot (PDP), feature interaction, feature importance, Individual conditional expectation (ICE), Local interpretable model-agnostic explanations (LIME), global surrogate models, and Shapley value explanations (Friedman, 2001; Friedman and Popescu, 2018; Fisher et al., 2018; Ribeiro et al., 2016; Lundberg and Lee, 2016). These methods have both advantages and disadvantages. Selection of the interpretation method is based on the research question. For example, ICE has advantages in discovering heterogeneous relationships. The drawback is that ICE can display only one variable. LIME can be used for different kind of data types including text and images. The key drawback of this method is the instability of explanations. Global surrogate models can draw conclusions about the model but not about the data. Feature importance, feature extraction, and Shapley value explanations models are usually computationally expensive. PDP is a classical method that visually presents the average partial relationship between one or more features and the predicted outcome. PDS can illustrate how the feature or group of features influence the prediction. However, PDP also has some disadvantages. Heterogeneous effects can remain hidden because PDPs only show the average marginal effects (Biecek, 2018). However, based on the current research problem, the Research Team used PDP as the appropriate tool to explain the estimates of AADT.

The average partial relationship in the PDP is also called marginal effects, marginal means, and predictive margins in various studies. This model-agnostic method can be applied to most machine learning methods (linear, monotonic, and more complex ones) with great flexibility. Following is the formula of the partial dependence function:

$$\widehat{g}_{x_k}(x_K) = E_{x_i}\left[\widehat{g}(x_K, x_I)\right] = \int \widehat{g}(x_K, x_I)dP(x_I) \tag{1}$$

where

$\widehat{g}$: machine learning model
x_K: one or two features of interest
x_I: other features were used in the model
x_K and x_I consist of a whole set of features were used in the machine learning model. Integrating out features in set x_I in the model $\widehat{g}$ gives the marginal effect of the features in set x_K, which only shows the relationship between features in x_K and model $\widehat{g}$.

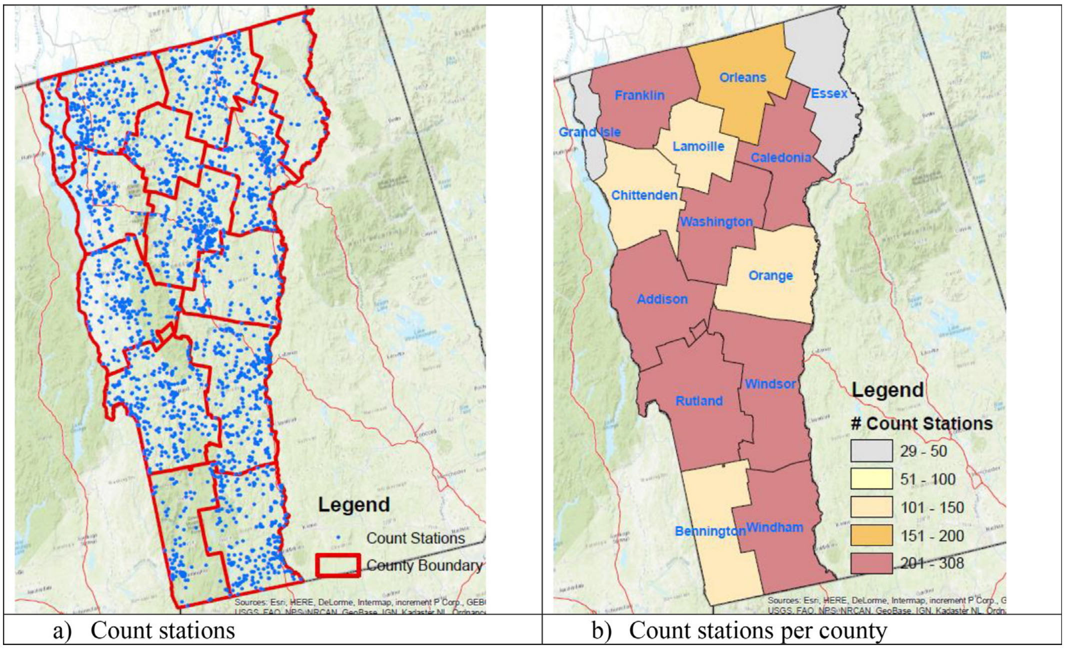
PDPs intuitively show an average estimation with all data points representing a particular value, as well as how much this average will change when all data points have a different value. This simple procedure provides insight to what happened in the black box while training the model. However, the interpretation only works well when the independence assumption holds. Based on the model above, the effect of interaction between those features remains in the model after the integration. Thus, partial dependence plots become questionable because the plots do not purely show the relationship between the feature of interest and the predicted outcome. Another concern is that plots can only handle one feature or, at a maximum, two features. Moreover, PDP shows the average estimation, which could be a concern when PDPs show a straight line. This straight line could represent the testing feature having no impact on the outcome, or that the impacts of the feature are canceled out because half of the data show the positive relationship of the feature and outcome while the other half shows the negative relationship (Biecek, 2018; Doshi Velez and Kim, 2017).

4. Data description

A wide range of transportation data was collected from across all Vermont counties in order to estimate the traffic volume on Vermont's low-volume roadways. The data included: (1) Vermont low-volume roadway traffic count station data, (2) demographic and economic data, and (3) distance to major roadways (Interstate and U.S. highways) from the count stations.

4.1. Low-volume roadway traffic count data

Vermont data contains traffic volume count data for 2369 low-volume roadway short-term count stations in 14 counties. Fig. 1 illustrates the locations of the stations in different counties. Fig. 1a shows the location and distribution of the count stations across the specified regions with each blue dot representing one station. It also shows which counties have the highest concentration of stations and where, approximately, those stations are located. Fig. 1b shows the heat map of count


Legend
Count Stations
County Boundary
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance
Franklin
Orleans
Essex
Grand Isle
Lamoille
Caledonia
Chittenden
Washington
Orange
Addison
Windsor
Rutland
Bennington
Windham
Legend
# Count Stations
29 - 50
51 - 100
101 - 150
151 - 200
201 - 308
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance

a) Count stations

b) Count stations per county

Fig. 1. Count stations on low-volume roadways in Vermont.

stations in each county. The counties are classified into categories based on the number of count stations. Essex and Grand Isle county have 29–50 stations; Bennington, Chittenden, Lamoille, and Orange county have 101–150 stations; Orleans county has 151–200 stations; and Addison, Caledonia, Franklin, Rutland, Washington, Windham, and Windsor county have 201–308 stations. Grand Isle county has the fewest number of count stations with 29, and Windsor county has the greatest number with 308. Low-volume roadways consider the following three functional classes:

- Rural collector (6R): 451 stations
- Rural local (7R): 1655 stations
- Urban local (7U): 263 stations

4.2. Demographic and economic data

4.2.1. U.S. Census and American Community Survey (ACS) data

The U.S. Census provides demographic information on various spatial units. The smallest unit is known as 'Census block,' and Census block group consists of several Census blocks. This study used the Census block group level demographic data due to its higher relevance to the modeling outcomes. The ACS is an ongoing national survey of U.S. households that is conducted by the U.S. Census Bureau to gather a wide variety of information such as a primary travel mode from home to work. The ACS is an essential tool for tracking travel patterns; it provides estimates for different levels: (a) 1-year estimates, (b) 3-year estimates, and (c) 5-year estimates. Many studies have used ACS data (Turner et al., 2017, 2018, 2019; Das et al. 2019). It is important to note that using 3-year or 5-year ACS is the most beneficial due to the large sample size relative to 1-year estimates. The multi-year estimates have advantages of statistical reliability for less populated areas and small population subgroups (Shawn et al., 2017).

4.2.2. Longitudinal Employer-Household Dynamics (LEHD) data

The LEHD program is a part of the Center for Economic Studies at the U.S. Census Bureau. It produces new, cost-effective, public-use information combining federal, state, and Census Bureau data on employers and employees under the Local Employment Dynamics Partnership. Furthermore, states agree to share unemployment insurance earnings data and the Quarterly Census of Employment and Wages data with the Census Bureau. Therefore, the LEHD program combines these administrative data, additional administrative data, and data from censuses and surveys. The LEHD data provides both home and work Census block data. Home level data is known as Residence Area Characteristic (RAC) data files, and work level data is known as Workplace Area Characteristic (WAC) data files. These files are released at the state level and are totaled by home Census block and work Census block, respectively.

4.3. Distance to major highways

Low volume roadways serve as collectors and minor arterials carrying traffic from local connector roads to major arterials. Thus, AADT on rural low-volume roadways is associated with the roadway's distance from interstates and/or major highways. To derive the shortest path from a low-volume roadway to the closest major highway or interstate, several processes within ArcGIS were used. First, every intersection of the Vermont centerline roadway network was located and added to a new feature class. The centerline roadway network is a two-dimensional network; in essence, it does not account for the elevation differences between overpasses and underpasses. As a result, the number of intersections found is greater than the actual number of traversable intersections. The naming convention within the centerline roadway network file includes a designation for roadway type (i.e., Interstate, US Route, etc.) and indicates whether the roadway is a ramp. The entire intersection feature class is parsed into two separate feature classes, one for intersections with interstates and one for intersections with U.S. Routes. In the case of interstates, only intersections with ramps are considered since ramps are used to access interstates. For U.S. Routes, all intersections with U.S. Routes are considered.

To find the distance between low-volume roads and the nearest interstate and major routes, the network analyst extension of ArcGIS 10.4.1 is used. There is a tool within the network analyst extension called origin–destination cost matrix. In order to find the shortest route, the network distance is used. This requires the use of a routed roadway layer that accounts for one-way directionality and elevation differences. The shortest route along the network is found between AADT count stations on roads with functional classes such as rural collector (6R), rural local (7R), and urban local (7U) roadways and the nearest intersection of an interstate and U.S. Route.

4.4. Data integration steps

The data preparation process is illustrated in the flowchart of Fig. 2. The data preparation works involve two software tools: ArcGIS 10.4.1 from Esri and open source tool R. The following steps were taken to develop the database:

- Using ArcMap, select the count stations on low-volume roadways. Assign the nearest road segment data to the count station using the 'near' function.

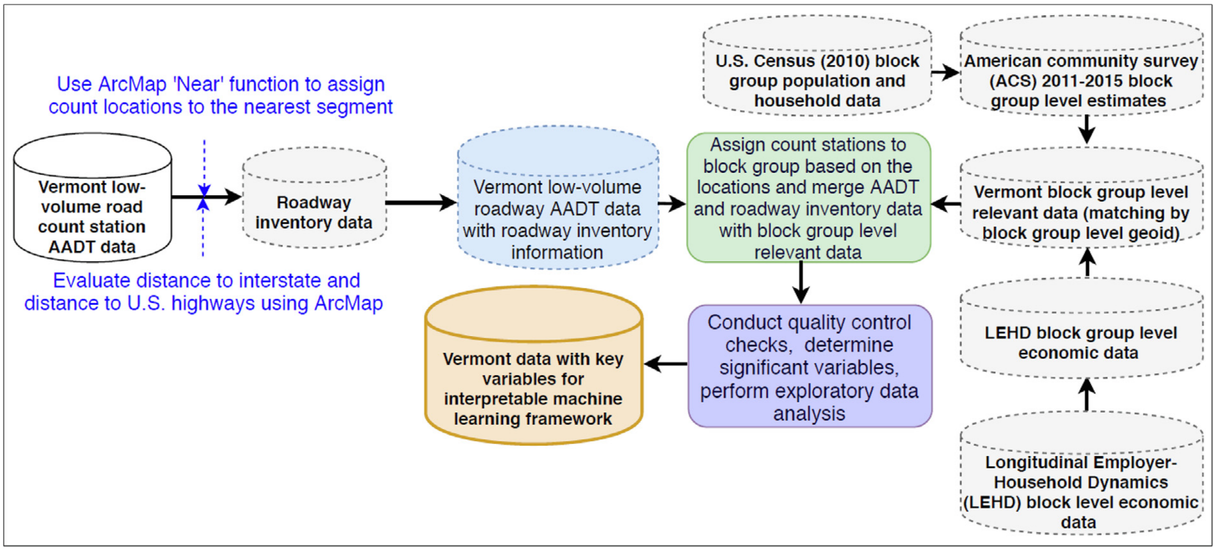

Use ArcMap 'Near' function to assign count locations to the nearest segment
Vermont low-volume road count station AADT data
Roadway inventory data
Evaluate distance to interstate and distance to U.S. highways using ArcMap
Vermont low-volume roadway AADT data with roadway inventory information
Assign count stations to block group based on the locations and merge AADT and roadway inventory data with block group level relevant data
U.S. Census (2010) block group population and household data
American community survey (ACS) 2011-2015 block group level estimates
Vermont block group level relevant data (matching by block group level geoid)
LEHD block group level economic data
Longitudinal Employer-Household Dynamics (LEHD) block level economic data
Conduct quality control checks, determine significant variables, perform exploratory data analysis
Vermont data with key variables for interpretable machine learning framework

Fig. 2. Flowchart of the data preparation.

- From the ACS block group geodatabase, select tables with population, housing unit, and income data. Assign count stations to the intersected block group level information.
- From the block level LEHD data, calculate block group level RAC and WAC values. Assign these data to the merged data.
- Determine the shortest network distance between AADT count stations on functional class rural collector (6R), rural local (7R), and urban local (7U) roadways and intersections of interstates and US Routes. This is accomplished using the origin–destination cost matrix tool within the ArcGIS network analyst extension.

4.5. Descriptive statistics

The multi-collinearity was examined using the variance inflation factor (VIF). Eight variables are primarily selected for the model development (density of several variables was calculated by dividing the count with the land area of the subsequent block group). Since multi-collinearity increases the instability of coefficient estimates, the multicollinearity problem was remedied by expressing the model regarding key independent variables. The final variables used for development of the model are AADT, population density, household (HH) density, WAC density, RAC density, distance to Interstate, and distance to U.S. Highway. Table 2 lists the mean and standard deviation of these variables. It shows that AADT, population density, and housing unit density vary from one functional class to another functional class. For example, traffic volume of rural local (7U) roadways have lower mean and standard deviation than the other two roadways. On the other hand, rural collector (6R) have lower mean and standard deviation in the demographic variables such as population density and housing unit density when compared with other two roadways. These statistics are intuitive due to the nature of land use and demography of the rural local roadways.

Furthermore, box and violin plots for three facility types are shown in Fig. 3. The violin plots in Fig. 3 show that the population density for rural collector (6R) and rural local (7R) roadways have mean values that are very small, and the kernel density is highly concentrated towards zero.

Table 2
Mean and standard deviation of the key variables.

Variable	Rural Collector (6R)		Rural Local (7R)		Urban Local (7U)	
	Mean	Std.	Mean	Std.	Mean	Std.
AADT (vehicles per days)	633	575	359	467	848	1153
Population density (population per sq. mile)	72.92	182.09	101.56	278.50	1432.21	1983.32
Housing unit density (household per square mile)	37.41	83.64	52.91	141.21	693.34	990.49
RAC density (RAC per square mile)	33.69	79.59	47.93	125.02	701.34	995.31
WAC density (WAC per square mile)	24.53	67.81	54.80	235.83	1285.51	3485.55
Distance to interstate (mi)	14.61	12.97	16.81	13.72	13.59	16.18
Distance to U.S. highway (mi)	6.97	5.68	7.56	6.69	1.64	1.43

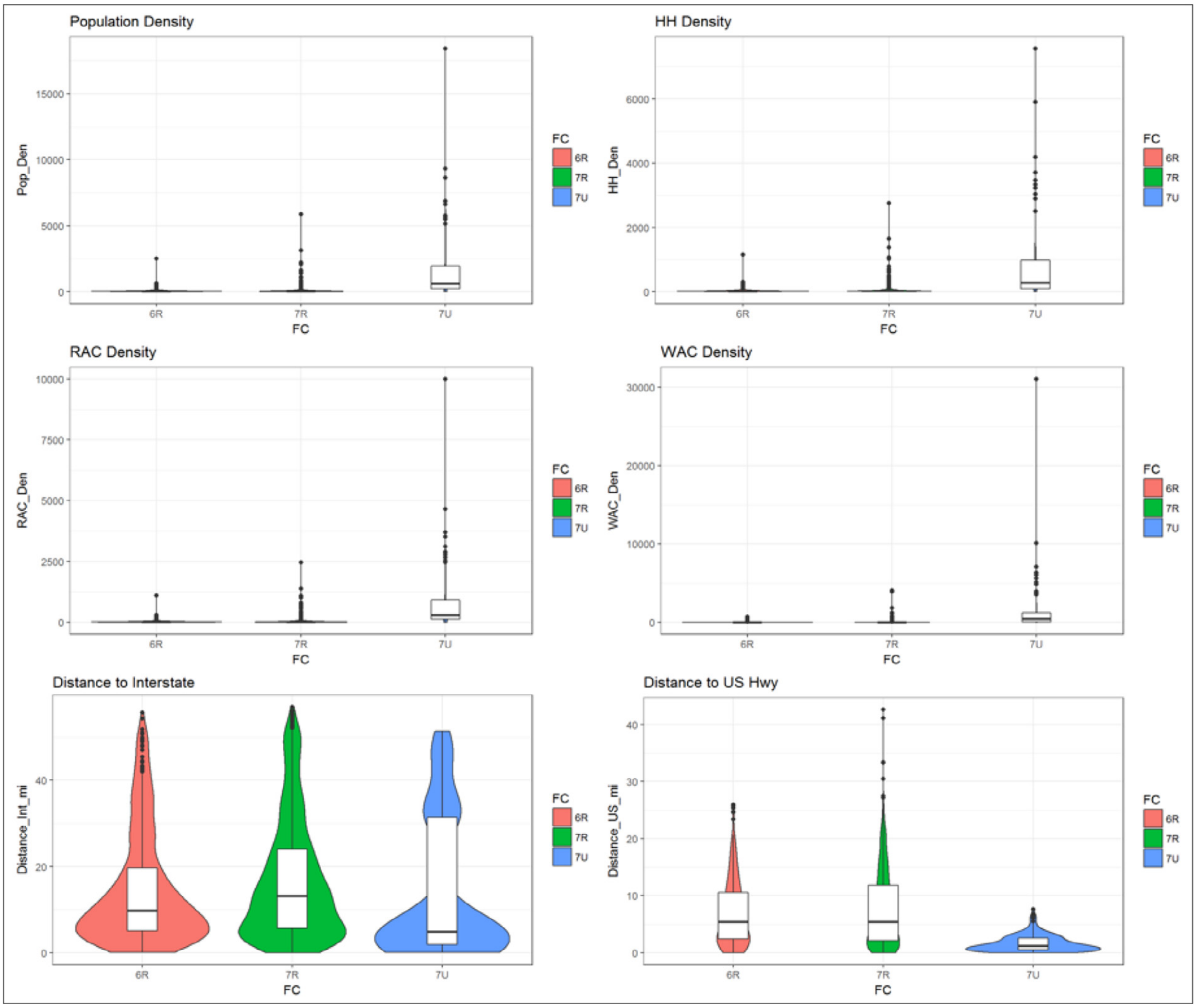


Fig. 3. Box and violin plots of the key variables.

Although the distribution is more widely spread for rural local (7R) than rural collector (6R) roadway; the mean value for urban local (7U) roadway is also relatively small; however, the distribution is much more widely spread. This same pattern of low mean values and the widest kernel distribution for urban local (7U) roadway is shown in the violin plots for HH density, RAC density, and WAC density as well. Fig. 3 also shows that the mean value for distance to interstate is lowest for urban local (7U) roadways with the highest mean value represented by rural local (7R) roadways. All three road types show a kernel distribution that is the most concentrated at or below the mean value. For distance to U.S. highway, urban local (7U) roadway has the lowest mean value while rural collector (6R) roadway has the highest; the kernel distribution is widely spread for rural local (7R) roadway while the distribution is heavily concentrated around the mean value. There are significant variations in AADT values and other variables by urban or rural locations and by functional class, but comparing the distance to interstate of these regions indicates the negligible difference between their means and standard deviations.

5. Methodology

5.1. Statistical and machine leaning models

After performing the feature-space analysis for the full dataset to identify any underlying correlations, the Research Team identified the optimal nature of the ultimate machine learning modeling technique. Two statistical models and three machine learning models were considered for this analysis:

- *Linear modeling (lm):* In this statistical model, the response variable is a function of one or more predictor variables, and therefore, it can be used to predict the response variable when only the predictor variable is known.
- *Generalized linear modeling (glm):* This statistical modeling is a generalization of linear modeling; it allows response variables to have error distribution models that are non-normal.
- *Random forest (rf):* The rf model, a machine learning model, operates by building multiple decision trees and merging them together in order to produce a more accurate and stable prediction. As each decision tree is created, the rf model introduces more randomness by searching for the best feature among a random subset. This results in a better, more diverse model.
- *Support vector machine (svm):* Machine learning model svm is a supervised learning models that plots data points, or support vectors, that lie closest to the decision surface or hyperplane. A svm estimates a situation by plotting these points in a finite-dimensional vector space, dividing them into distinct categories with gaps in between so that each dimension represents a "feature" of a particular object.
- *K nearest neighbor (knn):* A k-nearest neighbor (or k-NN for short) algorithm, a machine learning model, is a non-parametric model that categorizes an input by using its k nearest neighbors. A variety of distance functions (for example, Euclidean distance) can be used in this method.

5.2. Framework for AADT estimation on low-volume roadways

The current study developed a framework, as shown in Fig. 4, for the application of interpretable machine learning in estimating AADT on low-volume roadways. The steps are the following:

- Step 1: Acquire Traffic Volume Data. The first step is to collect AADT data from the available count stations on the low-volume roadway networks. AADT data is typically collected every year on these networks. Use growth factor or associated weighting parameter to evaluate AADT for the most recent year.
- Step 2: Collect Additional Demographic Data. In this step, the demographic and associated information are collected from the U.S. Census. If there is a need to calculate the distance between a station and major roadway networks, the users must evaluate those distance using GIS tools.

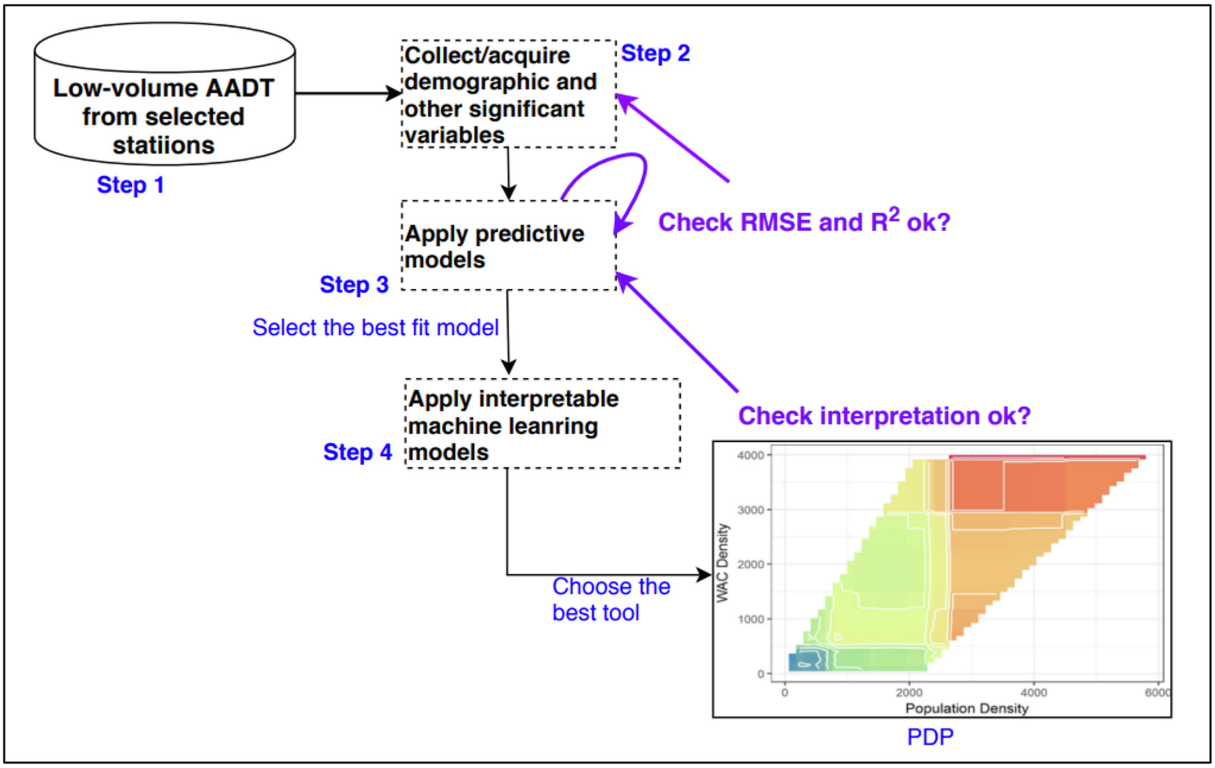


Fig. 4. IML framework for AADT estimation on low-volume roadways.

– Step 3: Apply Predictive Modeling. In recent years, many studies have developed an innovative machine learning tool for solving the regression problem. Several machine learning models can be examined by selecting the best-fit model with high R^2. This study applied two statistical models and three machine learning models to determine the most suitable model.
– Step 4: Apply Interpretable Machine Learning Model. This study developed PDPs for the combination of variable groups to determine the rules for AADT estimates.

6. Results and findings

Root mean square error (RMSE) and the coefficient of determination (R^2) are used to evaluate the model performance (see Table 3). RMSE represents the square root of the second sample moment of differences between observed and predicted values. A coefficient of variation is the proportion of the variance in the response variable that is predicted from the indicator variables. For RMSE, a value of 0 indicates perfect fit; the lower the RMSE value, the higher the model accuracy. For R^2, a value of 1 indicates the best fit model. For the random forest model, both RMSE and R^2 values indicate the better performance of ‘rf’ compared to the other models.

Fig. 5 shows a reversed empirical cumulative distribution function for absolute values from residuals. The illustration shows that a majority of residuals for the random forest method is smaller than residuals for the other models for all functional classes. As all of the models were performed using k (Eom et al., 2006) fold cross-validation, the model performances show the distribution of the RMSE and R^2 values as the measures of the model performances.

6.1. Rules generation from the PDPs

The variable importance measures generated from ‘rf’ algorithms show that population density and WAC density are the best predictors. Fig. 6 illustrates the PDPs based on these predictor variables. It shows different clusters of AADT values can

Table 3
Model performances for rural local models.

Model	Min	1st Qu	Median	Mean	3rd Qu	Max
RMSE values (7R)						
lm	322.702	385.735	443.571	448.373	501.106	621.211
glm	322.702	385.735	443.571	448.373	501.106	621.211
rf	309.173	329.208	360.162	396.309	458.895	566.300
svm	308.832	347.459	415.313	423.880	490.391	614.326
knn	294.652	342.255	380.132	410.567	472.337	561.772
R^2 values (7R)						
lm	0.008	0.043	0.070	0.105	0.140	0.346
glm	0.008	0.043	0.070	0.105	0.140	0.346
rf	0.108	0.221	0.287	0.289	0.367	0.455
svm	0.096	0.149	0.204	0.211	0.277	0.343
knn	0.092	0.168	0.227	0.225	0.279	0.427
RMSE values (7U)						
lm	474.509	744.646	914.491	1064.911	1155.051	2292.732
glm	474.509	744.646	914.491	1064.911	1155.051	2292.732
rf	425.504	705.392	838.260	992.118	1157.564	2129.099
svm	460.005	767.839	896.937	1066.632	1211.674	2288.859
knn	444.742	704.300	933.667	1066.792	1278.764	2189.031
R^2 values (7U)						
lm	0.000	0.012	0.037	0.081	0.149	0.357
glm	0.000	0.012	0.037	0.081	0.149	0.357
rf	0.000	0.055	0.138	0.197	0.253	0.672
svm	0.000	0.009	0.037	0.091	0.093	0.560
knn	0.000	0.011	0.051	0.091	0.150	0.470
RMSE values (6R)						
lm	340.325	362.438	426.554	478.494	515.318	923.433
glm	340.325	362.438	426.554	478.494	515.318	923.433
rf	293.680	346.096	392.908	456.797	524.332	893.574
svm	284.080	349.236	449.948	487.497	574.572	1002.404
knn	286.598	349.101	408.205	471.550	506.506	1006.798
R^2 values (6R)						
lm	0.032	0.084	0.267	0.275	0.376	0.677
glm	0.032	0.084	0.267	0.275	0.376	0.677
rf	0.068	0.212	0.326	0.360	0.458	0.773
svm	0.053	0.140	0.256	0.243	0.354	0.484
knn	0.044	0.140	0.243	0.264	0.320	0.614

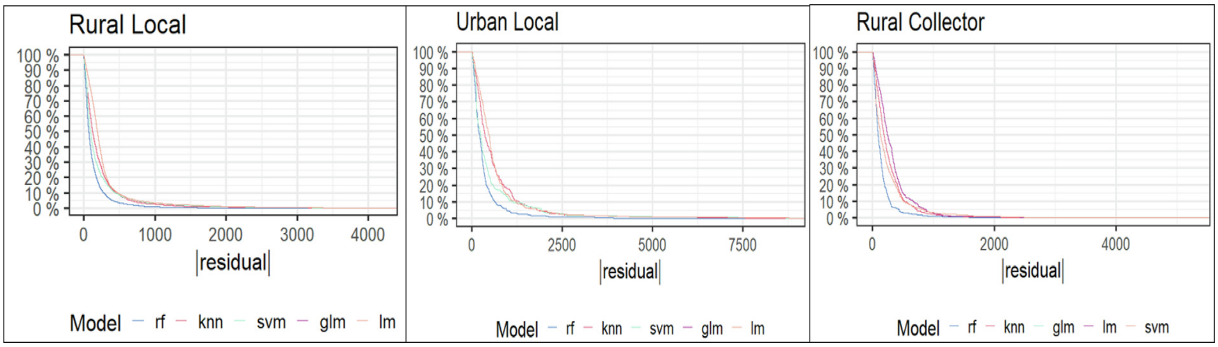


Fig. 5. Plots developed for Reversed Empirical Cumulative Distribution Functions.

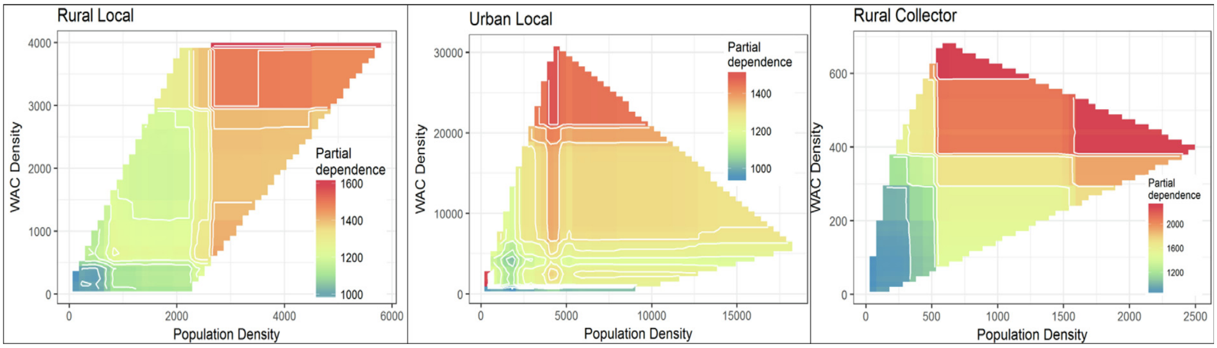


Fig. 6. PDPs for two variables.

be determined based on these two predictors. This plot can be translated into 'rules-based approximations', which can be used by different stakeholders. For example, *{Population density < 400 and WAC density < 400 → AADT < 400 vpd}* is a rule for rural local roadways (based on Fig. 6). Similarly, other rules can be developed from the PDPs. For this study, 10-fold cross validation approaches were applied. The sample size of rural collector (6R) and urban local (7U) are smaller than that of rural local (7R) roadways. Table 4 provides the top 5 rules for each functional class that are generated from the PDPs. Practitioners can use Table 4 to estimate traffic volume of low-volume roadways in Vermont. As the decision rules are based on only two major predictor variables, the implementation of this procedure is fairly simple.

Table 4
Rules generated from the PDPs.

Roadway	Rule No.	Rule	AADT range
Rural Local	Rule 1	Population density < 400 and WAC density < 400	<400 vpd
Rural Local	Rule 2	400 < Population density < 2100 and WAC density < 400	400–1000 vpd
Rural Local	Rule 3	400 < Population density < 2100 and 400 < WAC density < 1800	100–1300 vpd
Rural Local	Rule 4	2100 < Population density < 4200 and 700 < WAC density < 3000	1300–1500 vpd
Rural Local	Rule 5	Population density > 2100 and WAC density > 3000	1500–1600 vpd
Urban Local	Rule 1	Population density < 2100 and WAC density < 10,000	<1100 vpd
Urban Local	Rule 2	2000 < Population density < 12500 and WAC density < 10,000	1100–1200 vpd
Urban Local	Rule 3	4000 < Population density < 16,000 and 8000 < WAC density < 18,000	1200–1300 vpd
Urban Local	Rule 4	4000 < Population density < 12,000 and WAC density > 18,000	1400–1500 vpd
Urban Local	Rule 5	4000 < Population density < 5000 and WAC density > 20,000	1400–1500 vpd
Rural Collector	Rule 1	Population density < 400 and WAC density < 300	<1100 vpd
Rural Collector	Rule 2	400 < Population density < 500 and 400 < WAC density < 300	1100–1200 vpd
Rural Collector	Rule 3	400 < Population density < 1500 and 100 < WAC density < 400	1200–1500 vpd
Rural Collector	Rule 4	400 < Population density < 1500 and 400 < WAC density < 600	1500–1800 vpd
Rural Collector	Rule 5	500 < Population density < 2500 and 400 < WAC density < 700	1800–2000 vpd

7. Conclusion

Traffic volume data analysis is important for many transportation research areas, including roadway safety improvement and design, countermeasure determination, travel model calibration and validation, pavement design, and air quality compliance. However, AADT data are more easily available for higher functional classes, and only a small percentage of low-volume roads have accurate AADT data. Because low-volume roadways constitute a large portion of the U.S. roadway network, more studies are needed in traffic volume prediction on these roadways.

The main contribution of this study was the development of a robust interpretable machine learning framework that can be used to estimate AADT; this framework can be adopted by other researchers and practitioners. The findings of this study show that machine learning models can perform better than conventional linear regression models. Additionally, the research team found that population and work employment density are the best predictors for all three low-volume roadway classes in Vermont. In this study, the best fit machine learning model (random forest) has higher R^2 values in comparison to the statistical models. Compared with results from regression models, the best fit random forest model improved the accuracy of AADT for low-volume roadways significantly, from 0.45 to 0.77. The partial dependent plots developed for combination of variables show different clusters with the estimated AADT values. This study developed the top five decision rules for three functional classes of roadways. The best fit estimates and the developed rules from the current study could enhance the predictive power of the SPF development for the low-volume roadways in Vermont and therefore improve the decision-making process.

The current study does have a few limitations. The main limitation is the small sample size for two of the low-volume roadways (rural collector or 6R and urban local or 7U). With access to a larger sample size, the model performance can be improved. Additionally, the current framework is machine learning based, so structural equations are not available for interpretation of the results. However, PDPs are alternative tools which can transform the black box algorithm into interpretable estimations.

Declaration of Competing Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgements

The authors would like to thank Joshua Schultz, Maureen Carr, and Mandy White from the Vermont Agency of Transportation (VTrans) for providing us with data. The contents of this paper reflect the views of the authors and not the official views or policies of VTrans. The authors also like to thank Dr. William Holik and Songjukta Datta for their help during the data integration work. The authors are grateful for their anonymous reviews containing feedback and suggestions.

References

Apronti, D.T., Herpner, J.J., Ksaibati, K., 2015. Wyoming Low-Volume Roads Traffic Volume Estimation Final Report FHWA-WY-16/04F. Wyoming Department of Transportation.

Barrett, M., Graves, R., Allen, D., Pigman, J., Abu-lebdeh, G., Aultman-Hall, L., Bowling, S., 2001. Analysis of Traffic Growth Rates Final Research Report KTC-01-15/SPR213-00-1F. Kentucky Transportation Center, Lexington, KY.

Biecek, P., 2018. DALEX: explainers for complex predictive models. https://arxiv.org/abs/1806.08915. [Accessed on July 27, 2018]

Christoph, M., 2018. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable Accessed on July 27, 2018 https://christophm.github.io/interpretable-ml-book/.

Das, S., Tsapakis, I., Datta, S., 2019. Safety performance functions of low-volume roadways. Transp. Res. Rec.

Dixon, M., 2004. The Effects of Errors in Annual Average Daily Traffic Forecasting: Study of Highways in Rural Idaho Research Report. University of Idaho, Moscow.

Doshi-Velez, F., Kim, B., 2017. Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

Eom, J.K., Park, M.S., Heo, T.Y., Huntsinger, L.F., 2006. Improving the prediction of annual average daily traffic for nonfreeway facilities by applying a spatial statistical method. Transp. Res. Rec. 1968, 20–29.

Fisher, A, Rudin, C., Dominici, F., 2018. Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the 'Rashomon' Perspective. http://arxiv.org/abs/1801.01489. [Accessed July 27, 2018].

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. JSTOR.1189-1232.

Friedman, J.H., Popescu, B., 2018. Predictive learning via rule ensembles. Ann. Appl. Stat. JSTOR 916–954.

Gastaldi, M., Rossi, R., Gecchele, G., Lucia, L., 2012. Annual Average Daily Traffic Estimation from Seasonal Traffic Counts Research Report. University of Padova, Padova, Italy.

Gecchelea, G., Rossia, R., Gastaldia, M., Caprinia, A., 2011. Data mining methods for traffic monitoring data analysis: a case study. Proc. Social Behav. Sci. 20, 455–464.

Greenwell, B., 2018. pdp: A General Framework for Constructing Partial Dependence. https://github.com/bgreenwell/pdp. [Accessed on July 27, 2018].

Lowry, M., 2014. Spatial interpolation of traffic counts based on origin-destination centrality. J. Transp. Geogr. 36, 98–105.

Lundberg, S., Lee, S., 2016. An unexpected unity among methods for interpreting model predictions. http://arxiv.org/abs/1611.07478. [Accessed July 27, 2018].

McCord, M., Yongliang, Y., Jiang, Z., Coifman, B., Goel, P., 2003. Estimating annual average daily traffic from satellite imagery and air photos. Transp. Res. Rec. 1855, 136–142.

Mohamad, D., 1998. An annual average daily traffic prediction model for county roads. Transp. Res. Rec. 1617, 99–115.

Morley, D., Gulliver, J., 2016. Methods to improve traffic flow and noise exposure estimation on minor roads. Environ. Pollut., 1–9

Pan, T., 2008. Assignment of Estimated Average Annual Daily Traffic on All Roads in Florida Master Thesis. Civil Engineering Department, University of South Florida.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should I trust you? Explaining the predictions of any classifier. The Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM.

Seaver, W.L., Chatterjee, A., Seaver, M.L., 2000. Estimation of Traffic Volume on Rural Non-State Roads Research Report. University of Tennessee, Knoxville and University of Georgia, Athens.

Selby, B., Kockelman, K., 2011. Spatial Prediction of AADT in Unmeasured Locations by Universal Kriging Research Report. The University of Texas, Austin.

Sharma, S., Lingras, P., Xu, F., Kilburn, P., 2001. Application of neural networks to estimate AADT on low-volume roads. J. Transp. Eng. 127, 426–432.

Shawn, S., Sener, I., Martin, M., Das, S., Shipp, E., Hampshire, R., Fitzpatrick, Molnar, K.L., Wijesundera, R., Colety, M., Robinson, S., 2017. Synthesis of Methods for Estimating Pedestrian and Bicyclist Exposure to Risk at Areawide levels and on specific Transportation Facilities. Federal Highway Administration, 1200 New Jersey Avenue, Washington DC.

Shen, L.D., Zhao, F., Ospina, D.I., 1999. Estimation of Annual Average Daily Traffic for Off-System Roads in Florida Final Report. Florida Department of Transportation.

Sun, X., Das, S., 2015. Developing a Method for Estimating AADT on All Louisiana Roads Final Report FHWA/LA.14/158. Louisiana Department of Transportation, Baton Rouge, Louisiana.

Staats, W.N., 2016. Estimation of Annual Average Daily Traffic on Local Roads in Kentucky Master Thesis. Civil Engineering, University of Kentucky.

Turner, S., Benz, R., Hudson, J., Griffin, G., Lasley, P., Dadashova, B., Das, S., Texas A&M Transportation Institute, Texas Department of Transportation, Federal Highway Administration. Improving the Amount and Availability of Pedestrian and Bicyclist Count Data in Texas. 2019, p. 100.

Turner, S., Sener, I., Martin, M., White, L.D., Das, S., Hampshire, R., Colety, M., Fitzpatrick, K., Wijesundera, R. Guide for Scalable Risk Assessment Methods for Pedestrians and Bicyclists. 2018, p. 122.

Turner, S., Sener, I., Martin, M., Das, S., Shipp, E., Hampshire, R., Fitzpatrick, K., Molnar, L., Wijesundera, R., Colety, M., Robinson, S., Synthesis of Methods for Estimating Pedestrian and Bicyclist Exposure to Risk at Areawide Levels and on Specific Transportation Facilities, 2017, p. 93.

Xia, Q., Zhao, F., Chen, Z., Shen, L., Ospina, D., 1999. Estimation of annual average daily traffic for nonstate roads in a Florida County. Transp. Res. Rec. 1660, 32–40.

Zhao, F., Li, M.T., Chow, L.F., 2004. Alternatives for Estimating Seasonal Factors on Rural and Urban Roads in Florida Final Report BD015-03. Florida Department of Transportation, Tallahassee, Florida.