Transportation Planning and Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gtpt20

Traffic volume prediction on low-volume roadways: a Cubist approach

Subasish Das

TRANSPORTATION PLANNING AND TECHNOLOGY
https://doi.org/10.1080/03081060.2020.1851452

Traffic volume prediction on low-volume roadways: a Cubist approach

Subasish Das

Texas A&M Transportation Institute, San Antonio, TX, USA

ABSTRACT

A significant aspect of the U.S. Department of Transportation's Highway Safety Improvement Program (HSIP) rulemaking is the prerequisite that states must gather and utilize Model Inventory of Roadway Elements (MIRE) for all public paved roads, including low-volume roadways (LVR). States are particularly not equipped with the ability to collect traffic volumes of LVRs due to issues such as budgetary constraints. One alternative is to estimate traffic volumes of LVRs using regression or machine learning (ML) models. The present study accomplishes this by developing a ML framework to estimate traffic volumes on LVRs. By using available traffic counts on low-volume roads in Minnesota, this study applies and validates three different ML models (random forest, support vector regression, and Cubist) to estimate traffic volumes. The models include various traffic and non-traffic (e.g. demographic and socio-economic) variables. Overall, the Cubist model shows better performance compared to support vector regression and random forests. Additionally, the Cubist approach provides rule-based equations for different subsets of data. The findings of this study can be beneficial for transportation communities associated with LVRs.

ARTICLE HISTORY
Received 8 October 2019
Accepted 24 September 2020

KEYWORDS
Traffic volume; low-volume roadways; machine learning; Cubist; case study

Introduction

Under the United States Department of Transportation's (USDOT) latest criteria, published in the Highway Safety Improvement Program (HSIP) Final Rule in March 2016, states will need to obtain annual average daily traffic (AADT) information along with other Model Inventory Roadway Elements (MIRE) – Fundamental Data Elements (FDEs) for all government-maintained highways, including non-Federal Aid System (NFAS) highways. The traffic safety sector encompasses a broad range of study fields, and crash data analysis is the most prominent among them. Crash data analysis is primarily used to evaluate the safety of a particular transit unit (e.g. arterial junctions); most model design methods focus on high-volume highways because appropriate information for these roadway types is more readily available.

The Highway Performance Monitoring System (HPMS) does not denote any specific method for sampling quantities of vehicles on NFAS highways. Many surveys emphasize

CONTACT Subasish Das s-das@tti.tamu.edu Texas A&M Transportation Institute, 3500 NW Loop 410, Room 315E, San Antonio, TX 78229, USA

the data-driven approach and thus concentrate on the event of the crash and its connection to a broad range of factors. The techniques included in the first version of the Highway Safety Manual (HSM) are commonly used to forecast crashes on a roadway segment or intersection (AASHTO 2010). The corresponding transport department selects a sampling technique and an AADT assessment method which typically uses the information on historical vehicle quantity or modeling methods for these highways. If the traffic count data is unavailable, then projections are made based on comparable roadway types. This method may lead to significant biases due to inaccurate estimates and the lack of current data.

There have been new advancements in traffic volume estimation within the last decade, specifically with statistical models and artificial intelligence (AI) or machine learning (ML) algorithms. The latest statistical and ML methods include spatiotemporal modeling, multivariate analysis, data mining techniques, and random parameters to empirical Bayesian and full-Bayesian hierarchical approaches. Without any preceding information of underlying processes, ML algorithms can identify non-linear and complex associations between an independent factor and a wide range of dependent factors. The key aim of ML models is to establish best fit models that provide high prediction accuracy. To accomplish the research goal in this study, four data sets were used to create a suitable database for the model development: (1) traffic volume counts of Minnesota low volume roadways (LVRs); (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. Using the merged dataset, this study utilized three robust ML models in estimating AADT on LVRs in Minnesota.

Literature review

Annual average daily traffic (AADT) estimation methods can be broadly divided into two major sections: methods with traffic volume counts and methods with no traffic volume counts.

Methods with traffic volume counts

Traffic volume count-based methods either rely on vehicle volume data acquired from ongoing counting locations, mobile vehicle recorders, or both. Current numbers along with socioeconomic data, network connection, and other information are used to estimate AADT values and to create regression designs for uncounted sections.

Traditional approach and sampling

To gather local road and street traffic count data, Barrett et al. (2001) developed a random sampling method with map dimensions of 0.1 miles for metropolitan regions and 0.2 miles for agricultural regions. The study experienced a complication when it tried to use map dimensions under 0.05 miles but failed due to software problems. In another study, Blume et al. (2005) developed a random sampling method using Florida census information to create a methodology to predict vehicle miles traveled (VMT). This research identified correlations between transport, demographic size, work size, and road size. Lloyd and French (2006) conducted a study to identify a sampling method to anticipate VMT projections on local highways organized by Pennsylvania municipalities. This study also used census information that was gathered at a county level and then associated to AADT concentrations. In another study, Jessberger et al. (2016) evaluated 14 years of data to develop a new method of estimating AADT that included any number of period increments.

Non-traditional approach and sampling

In Alberta, Canada, Sharma et al. (2001) established designs of neural networks to predict AADT on LVRs. Using the Classification and Regression Tree (CART), Dixon (2004) measured the annual growth level of AADT values. Local and low volume roadways indicated lower annual growth levels than roads with high traffic volumes. Gecchelea et al. (2011) studied TMG processes and clustering techniques to predict AADT more accurately. Gastaldi et al. (2012) indicated that the most precise projections would come from traffic data collected on weekdays. Conclusions were based on one-week annual traffic count estimate method in Italy. By creating an OLS model, Lowry (2014) estimated specific AADT scores in Idaho. The sub-sampling validation outcomes concluded that comparable levels of AADT precision could be obtained using about one-fifth of traffic count data.

Methods without traffic volume counts

Depending on non-traffic counts, models can generate AADT projections for individual sections or for a set of prevalent trait road sections. Models generating disaggregated assessments at segment level have a higher efficiency than models producing aggregated projections. Nevertheless, models for individual sections involve disaggregated information, which is often hard to acquire.

Disaggregated estimates

Zegeer et al. (1994) examined the link between road width and collision (AADT 2000 vehicles per day (vpd)) on rural LVRs. Roadways with comparatively wider shoulders reported reduced collision rates, whereas the shoulder type (surfaced or unpaved) was not statistically significant. Stamatiadis, Jones, and Aultman-Hall (1999) identified several influencing variables on LVRs, classified by an AADT of less than 1000 vpd. In another study using data from rural areas, Achwan and Rudjito (1999) examined the association of road characteristics and traffic volume. Liu and Dissanayake (2008) analyzed the collision variables that were the most related by creating logistical regression designs on gravel roads. Mohamad (1998) created a template of road forecast on Indiana district highways. Xia et al. (1999) estimated AADT for non-state highways in urbanized Florida regions and determined that traffic features, such as the number of routes, functional classification, and region sort were the most significant influencing predictors. McCord et al. (2003) used satellite imagery of elevated precision to estimate AADT. For a spatial forecast of AADT, Selby and Kockelman (2011) used standard kriging in uncounted Texas locations. Depending on Euclidean ranges, the research outcomes contrasted with those using network paths. Apronti, Herpner, and Ksaibati (2015) created a linear regression model and a logistic correlation method to anticipate AADT on LVRs in Wyoming. Findings concluded that both designs of regression are inexpensive, simple, and easy to execute. Das and Sun (2015) applied support vector regression technique to estimate traffic volumes on local roadways of eight parishes in Louisiana. In another study, Das and Ioannis (2020) used interpretable ML models to estimate traffic volumes on LVRs in Vermont.

Aggregated estimates

Shen, Zhao, and Ospina (1999) developed four multiple linear regression models to project AADT values for off-system roads in Florida. Each model produced aggregated estimates for different geographical sites. Seaver, Chatterjee, and Seaver (2000) determined the vehicle quantity on Georgia's non-state highways. Zhao, Li, and Chow (2004) conducted a regression analysis to detect feasible variables that influence monthly adaptation factors in selected agricultural regions in Florida. In Kentucky, Staats (2016) created six designs to evaluate aggregated local road vehicle quantities.

ML models

To measure the efficiency of estimating traffic volumes of LVRs, three ML regression techniques (Random Forest (RF), Support Vector Regression (SVR) and Cubist) were compared.

Cubist

Cubist, a rule-based ensemble regression model technique with separate linear regression equation subsequent for each terminal node, was developed by Quinlan (1996; 1992). The paths along the model tree are flattened into rules these rules are simplified and pruned. In comparison to ordinary regression, model trees have been shown to be more accurate. An additional technique to improve estimation uses similar training cases, or instances. Cubist ensembles are created using committees, which are similar to boosting. After the first model in the committee is created, the second model uses a modified version of the outcome data based on whether the previous model under- or over-predicted the outcome. For iteration m, the new outcome $y*$ is computed using the following equation:

$$y^*_{(m)} = y - (\hat{y}_{(m-1)} - y) \tag{1}$$

On the off chance that a test is under-predicted on the past cycle, the result is balanced so that another time it is more likely to be over-predicted to compensate. This alteration proceeds for each outfit emphasis.

Random forest (RF)

The importance of each variable was ordered using random forest (RF) algorithms. RF strategy is dependent on the bagging principle (Breiman 2001) and random subspace method (Ho 1998) that depends on building a compilation of decision trees with random predictors.

Out-of-bag error rate (OOB) and variable importance measures are the two vital byproducts of the RF method. OOB is the misclassification rate that decreases as the number of tree increment. The trees are grown to the maximum depth to reduce the bias and correlation. Gini impurity and classification accuracy are used as the measures of variable importance. This importance measure illustrates how much the mean squared error or the 'impurity' increases when the specified variable is randomly permuted.

Support vector regression (SVR)

In 1963, Vapnik and Lerner presented the Generalized Portrait algorithm, which has a fundamental algorithm for the development of Support Vector Machine (SVM). SVMs are the statistical learning theory algorithms implementing the structural risk minimization inductive principal to get excellent generalization on a restricted number of learning designs. Vapnik begun the field of statistical learning theory in 1974 (History of SVM 2020). On a premise of a distinguishable bipartition problem, Vapnik et al. presented the SVM framework in 1992 at the AT & T Bell Laboratories (2004). SVM aims to delineate the information x into a high-dimensional feature space F by using a nonlinear mapping in a way to execute linear regression in this space.

Whereas keeping up all the most highlights that characterize the maximal margin algorithm, the SV algorithm can also be applied to Support Vector Regression (SVR). The SVR approach accounts for the error approximation in data with the generalization of the model. With different forms of SVR, the classical model, ε-SVR, was discussed in Smola and Schölkopf (2004) and Cornejo-Bueno et al. (2016).

Database development

A wide range of transportation data were collected from LVRs in all Minnesota counties to estimate the traffic volume: (1) Minnesota LVR traffic count station data, (2) Minnesota road inventory database, (3) U.S. Census and American Community Survey (ACS) data, and (4) distance to major roadways (Interstate and U.S. highways) from the count stations.

Data sources

LVR traffic count data

Minnesota data contains traffic volume count data for 14,989 stations in 87 counties. LVRs consider the following three functional classes:

- Rural collector (6R): 4024 stations
- Rural local (7R): 6543 stations
- Urban local (7U): 4422 stations

Demographic and economic data

U.S. Census and American Community Survey (ACS) data. Demographic information on various spatial units are provided by the U.S. Census. Due to its higher relevance in modeling outcomes, this study used the Census block group level demographic data.

Conducted by the U.S. Census Bureau, the ACS is an ongoing national survey of U.S. households to gather a wide variety of information such as a primary travel mode from home to work. ACS is an imperative tool for tracking travel patterns. The ACS supplies estimates for various levels: (a) 1-year estimates, (b) 3-year estimates, and (c) 5-year estimates. Due to the large sample size, practitioners usually use 3- or 5-year ACS is more beneficial compared to 1-year estimates. The multi-year estimations have benefits of statistical consistency for small population subgroups and less populated areas (Shawn et al. 2017).

Longitudinal Employer-Household Dynamics (LEHD) data. Under the Local Employment Dynamics Partnership, the LEHD produces cost-effective, new, public-use data. Moreover, states correspond to share unemployment insurance earnings data and the Quarterly Census of Employment and Wages data with the U.S. Census Bureau. The LEHD information gives both work (known as Workplace Area Characteristic or WAC) and home (known as Residence Area Characteristic or RAC) Census block data. These files are released at the state level and totaled by home Census block and work Census block, respectively.

Distance to major highways: The network analyst extension of ArcGIS 10.4.1 is used to determine the distance from LVRs to the nearest interstate and major highways. The network analyst extension also has a tool called the origin-destination cost matrix. The network distance is used in order to identify the shortest route. This method uses a routed roadway layer that considers one-way directionality and elevation differences. The shortest route within the network is identified between AADT count stations on roads with functional classes 6R, 7R, and 7U and the closest intersection of interstate and US Route.

Data integration

Figure 1 shows the overall data merging steps. 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:

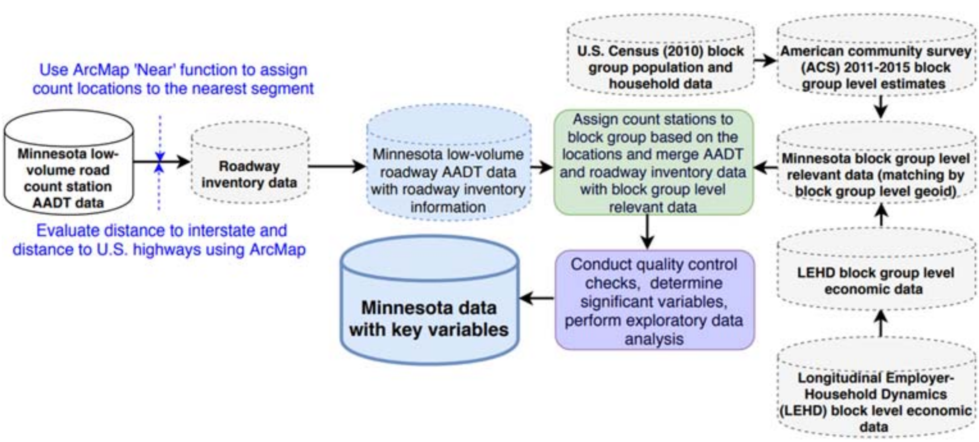

Figure 1. Data merging flowchart.

- Using ArcMap, select the count stations on LVRs. Assign the nearest road segment data to the count station using the 'near' function.
- From the ACS data, population, housing unit, and income data are needed to be selected. 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 6R, 7R, and 7U roadways and intersections of interstates and US Routes. This is accomplished using the origin-destination (OD) cost matrix tool within the ArcGIS network analyst extension.

Exploratory data analysis

The multi-collinearity was examined using the variance inflation factor (VIF), and eight variables are primarily selected for the model development. Since multi-collinearity increases the instability of coefficient estimates, the multicollinearity problem was remedied by expressing the model regarding key independent variables. Table 1 lists some key describe statistics of the key variables. The functional classifications described in Table 1 are 6R, 7R, and 7U; the mean, standard deviation, minimum, maximum, and interquartile range is given for the AADT, the population, the housing units in the block group, the number of occupants in the block group, work area characteristics, residential area characteristics, the distance to the interstate from the count station, and the distance to a U.S. highway from the count station for each of these classifications. Furthermore,

Table 1. Descriptive statistics of the key variables.

Functional classification	Attribute	Count	Mean	SD	Min	Max	IQR
6R	AADT	4024	425	487.8	5	4700	390
	Popu	4024	1133	554.2	437	6396	560
	HU	4024	591	297.7	210	2426	335
	Occu	4024	591	297.7	210	2426	335
	WAC	4024	285	403.9	2	6476	255
	RAC	4024	554	298.2	151	3532	320
	DistI	4024	45	42.4	0	214	54
	DistUS	4024	12	15.6	0	158	11
7R	AADT	6543	177	300.8	5	4700	150
	Popu	6543	1076	490.5	62	6396	497
	HU	6543	582	289.0	30	2426	318
	Occu	6543	582	289.0	30	2426	318
	WAC	6543	257	344.0	2	6001	255
	RAC	6543	516	270.3	25	3532	293
	DistI	6543	51	41.3	0	215	58
	DistUS	6543	11	16.1	0	158	11
7U	AADT	4422	1372	1150.7	5	5000	1460
	Popu	4422	1805	1130.8	0	9734	1195
	HU	4422	737	398.8	0	3220	458
	Occu	4422	737	398.8	0	3220	458
	WAC	4422	1402	2420.6	2	31,208	1280
	RAC	4422	928	617.3	33	4779	660
	DistI	4422	14	27.0	0	163	10
	DistUS	4422	4	4.4	0	26	5

Notes: Popu = Population in block group, HU = Housing units in block group, Occu = Number of occupants in block group, WAC = Work Area Characteristics (block group), RAC = Residential Area Characteristics (block group), DistI = Distance to Interstate from the count station, DistUS = Distance to U.S. Highway from the count station.

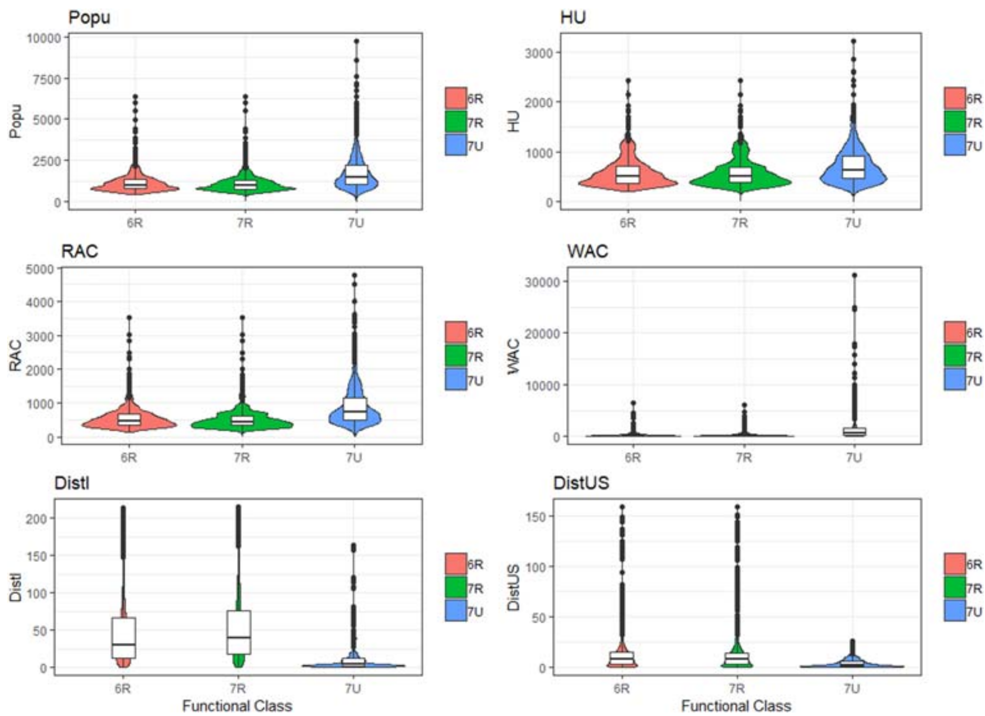

Figure 2. Box and violin plots of key variables.

box and violin plots for six variables are demonstrated in Figure 2. There are clear contrasts in AADT values and other variables by urban or rural locations and by functional class. From Figure 2, it is found that AADT, population density, and housing unit density in rural LVRs have higher mean and standard deviations compared to urban local and rural collector. In the same way, the mean and the standard deviation of RAC and WAC of urban local is considerably higher than the two other regions. But comparing the distance to interstate of these regions indicates the negligible difference between their means and their standard deviations.

Methodology

This study conducted a five-fold cross-validation procedure to execute validation on the dataset in an iterative fashion. To accomplish this, this study portioned the full dataset into five equal subsamples that were used successively to independently validate the trained based on the four remaining subsamples. This strategy guaranteed a decreased computation time for three different algorithmic techniques used in this study. The standard statistical measures used to evaluate model performance incorporate the coefficient of determination (R^2), Root Mean Square Error (RMSE), and mean absolute error (MAE). For example, RMSE is the standard deviation of the residuals, which is considered as a measure of the dispersion of the residual measures. It gauges the parameter values, the standard deviation of the error term with certain degrees of freedom or DOF (consider DOF as n). The formulation of RMSE can be expressed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \tag{2}$$

Smaller RMSEs are associated with smaller standard errors, which indicate better model fitness. This study used two open source R packages (cubist and caret) to apply the ML algorithms (Kuhn and Quinlan 2018; Kuhn 2018). The RMSE measures, as shown in Table 2, can provide evidence of model performances. The outcome measures show that the Cubist model yields the highest accuracy than other two ML models (RF and SVR). The three models were compared using their minimum, first quartile, median, mean, third quartile, and maximum values.

The Cubist model's performance was later used to create the rules-based SPFs. A committee is a boosting feature within the Cubist model where repetitive model trees are produced in succession. After the generation of the first tree, the following trees are formed based on adjusted versions to the training data result: if the model over-predicts a value, the following model is adjusted accordingly. In contrast to traditional boosting, the predictions from each model tree are not averaged based on stage weights within each committee; the final prediction is found by a simple mean of the predictions from each previous model tree. This study used the committee method to regulate the number of model trees. The Cubist model also uses nearest–neighbors to change the predictions from the rule-based model. First, a model tree (with or without consideration of any committee) is created. Once this model makes a sample prediction, Cubist finds its nearest neighbors and calculates the average of these training set points. Readers can consult Quinlan (1993) for more information about these adjustment criteria.

Cubist models can be actualized and utilized successfully with the determination of exceptionally few tunable model parameters. In most cases, because a number of rules will need to be optimized for the given regression problem, it makes this procedure exceedingly alluring as data driven tools for understanding complicated associations between dependent and independent variables. This study used the complete dataset for the final regression rules. Four different instances (AASHTO 2010; Blume et al. 2005; Jessberger et al. 2016; Dixon 2004) were chosen in the final stage of modeling performance. Based on the preparatory investigations, it was found that committees higher than 5 did not lead to extra changes in model estimation. The instances are limited to 7 to reduce computation time. The RMSE values generated from different tuning or committee-instance scenarios for 6R, 7R, and 7U are shown in Figure 3.

To see how proficient the estimate is in terms of the estimated variability or precision, one can quantify the coefficient of variation (i.e. the quotient of a standard deviation and a mean). Table 3 lists the model performance measures for the final models of the three

Table 2. RMSE Values for different algorithms.

Models	Minimum	First quartile	Median	Mean	Third quartile	Maximum
RMSE						
RF	17.72	18.01	18.98	18.86	21.02	24.51
SVR	17.68	18.23	19.16	19.23	22.41	24.68
Cubist	16.98	17.67	18.29	18.14	20.78	22.35

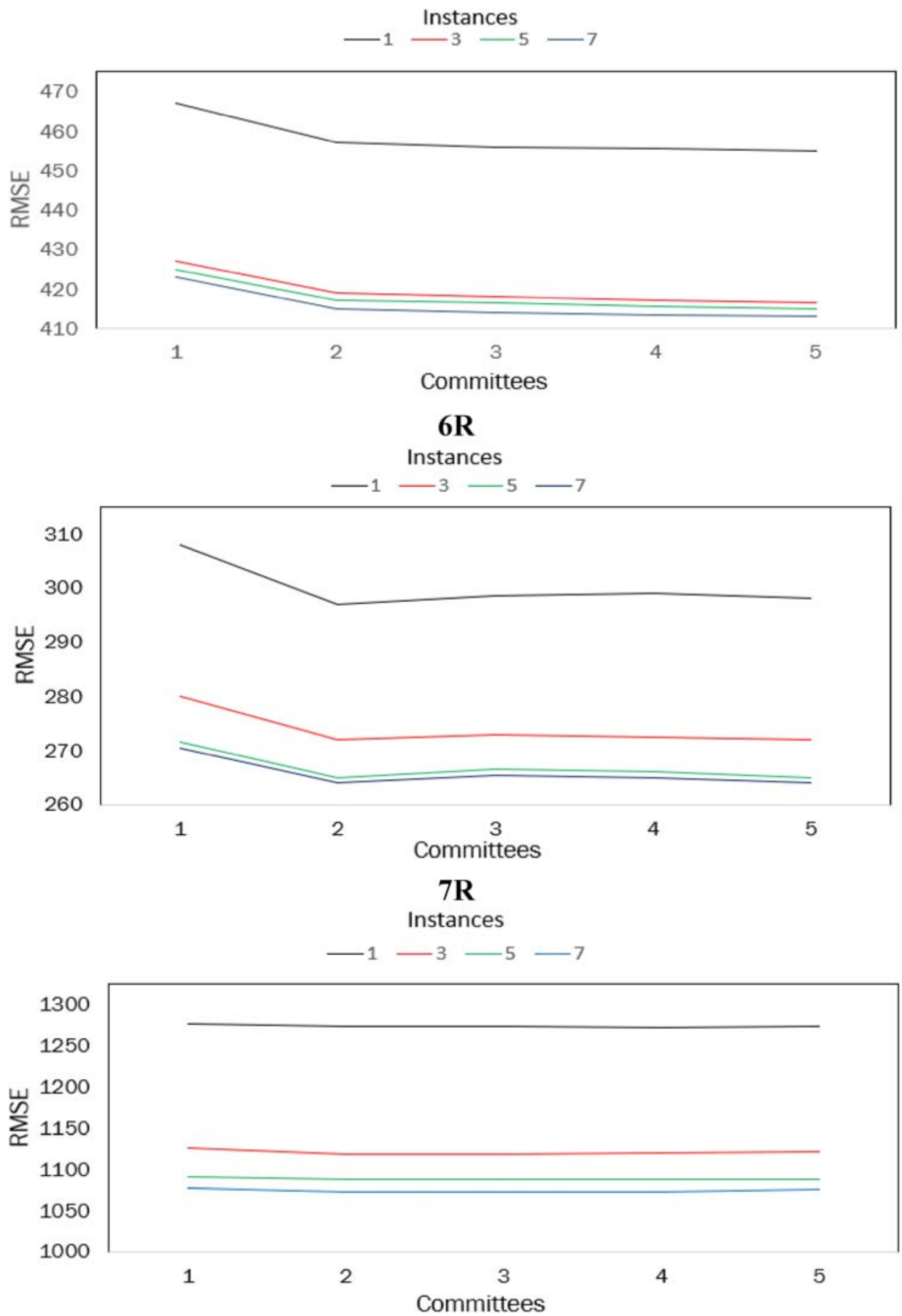

Figure 3. Tuning parameters and RMSE values.

Table 3. Model performances.

Functional class	Count stations	Committees	RMSE		R^2		MAE	
			Train	Test	Train	Test	Train	Test
6R	3216	5	413.100	464.835	0.3056	0.2757	257.164	279.491
7R	5234	5	262.934	304.118	0.2460	0.2075	132.358	144.212
7U	3536	5	1076.231	1270.798	0.1450	0.1025	818.509	916.211

functional classes compared in this table. The number of stations, committees, RSME, R^2, and MAE are listed for comparison purposes. The overall contention is that the Cubist model performs well in both training and test data.

Results and discussion

Where each tree divides, Cubist produces a linear model (after performing the feature selection) that allows terms for each variable used in the existing division or any division above it. Thus, the final prediction is a result of all the linear models ranging from the original node to the terminal node. The attribute usage percentages shown in Table 4 reflect all of the models used for the prediction. The data were evaluated in the table by calculating the mean error, relative error, and correlation coefficient. The presence of a relationship is increased with a higher correlation coefficient. The important predictor variables include population, WAC, and distance to Interstate, which is coherent with the training and test data. Distance to US and Housing units are the significant contributor in the urban local roadways. These two variables are not significant in rural models.

It is important to note that the Cubist algorithm relies on rule-based multivariate linear regression models rather than an uninterpretable ML model. This unique feature makes Cubist attractive to the researchers. A linear model developed from a rule can be used to predict traffic volume at a location as a function of exposure (i.e. population, RAC, WAC). Tables 5–7 list the generated rules and linear models developed by each rule for traffic volumes of different low volume roadways. It is worth noting that the sum of the number of cases is not needed to be the total number of cases. Each rule lists their number of cases, mean (range), estimation error, and model by rules. During interpreting each equation, it is important to recall that several rules can contemplate the identical segment feature if it is in the criteria of the generated rules.

Table 4. Attribute usage in the models.

	6R		7R		7U	
	Training data	Test data	Training data	Test data	Training data	Test data
Attributes						
DistI	86%	90%	69%	93%	89%	98%
DistUS	31%	3%	35%	25%	82%	60%
RAC	81%	39%	41%	47%	31%	38%
HU	14%	21%	40%	7%	91%	79%
WAC	77%	55%	51%	88%	67%	97%
Popu	69%	64%	68%	72%	80%	56%
Route_Syst	4%	–	27%	–	36%	58%
Data evaluation						
Average error	258.8	276.6	134.4	148.7	866.1	995.9
Relative error	0.80	0.90	0.78	0.87	0.94	1.07
Correlation coefficient	0.60	0.45	0.48	0.49	0.37	0.27

Table 5. Rules generated for rural minor collector (6R) roadways.

Rules	Cases	Average (range)	Est Err	Rule based models
Rule 1: WAC > 423, RAC ≤ 492, DistUS > 1.325892 and DistUS ≤ 9.029426	45	250.8 (25, 970)	151.9	$AADT_i = 416.6 - 5.4 Distl$
Rule 2: WAC > 308, WAC ≤ 423, DistUS > 1.325892 and DistUS ≤ 9.029426	99	335.5 (10, 2000)	232.4	$AADT_i = 2121.5 - 6.02\ WAC + 0.53\ RAC$
Rule 3: DistUS > 1.220559	2697	379.9 (5, 4550)	238.2	$AADT_i = 89.2 + 0.3\ RAC - 1\ Distl + 0.043\ Popu$
Rule 4: WAC > 423, RAC > 492, Distl > 1.990492, DistUS > 2.970464 and DistUS ≤ 9.029426	113	453.7 (25, 1650)	291.7	$AADT_i = -205.8 + 100.7\ DistUS - 0.05\ HU + 0.05\ RAC$
Rule 5: HU ≤ 730, WAC > 308, Distl > 1.990492 and DistUS > 9.029426	180	568.6 (10, 3000)	383.0	$AADT_i = 351.6 - 1.8\ Distl + 0.18\ WAC + 0.03\ RAC - 0.011\ Popu$
Rule 6: HU > 730, HU ≤ 1103, WAC > 308, Distl > 1.990492 and DistUS > 9.029426	100	653.8 (15, 2900)	436.8	$AADT_i = 6159.8 - 4.71\ HU - 1.11\ RAC - 9.4\ DistUS$
Rule 7: DistUS ≤ 1.220559	519	685.1 (5, 4700)	428.8	$AADT_i = 305.7 + 0.21\ WAC + 0.24\ RAC - 1.3\ Distl + 0.026\ Popu$
Rule 8: WAC > 308, Distl > 1.990492, DistUS > 0.09887236, and DistUS ≤ 0.491724	94	745.2 (5, 2450)	465.8	$AADT_i = 1179 - 444.4\ DistUS - 7.4\ Distl$
Rule 9: WAC > 308, WAC ≤ 432, and DistUS ≤ 0.491724	57	745.2 (5, 2450)	471.4	$AADT_i = 5328.7 - 2373.7\ DistUS - 12.35\ WAC$
Rule 10: WAC > 423, RAC > 492, DistUS > 1.325892, and DistUS ≤ 2.970464	26	850.4 (95, 2300)	833.7	$AADT_i = 2604.1 - 6.05\ HU + 4.69\ RAC - 0.635\ Popu$
Rule 11: HU > 1103, WAC > 308, Distl > 1.990492, and DistUS > 9.029426	65	859.7 (35, 3400)	649.3	$AADT_i = 2851.3 - 1.194\ Popu - 14.4\ Distl + 1.33\ RAC + 0.71\ WAC$
Rule 12: HU ≤ 406, WAC > 308, and DistUS ≤ 1.325892	58	940.6 (35, 3600)	705.6	$AADT_i = 6502.2 - 16.99\ HU - 231.4\ DistUS$
Rule 13: WAC > 308, Distl > 1.990492, DistUS > 0.491724, and DistUS ≤ 1.325892	74	1124.5 (90, 4000)	672.6	$AADT_i = 798 + 0.2\ WAC + 0.24\ RAC - 1.3\ Distl - 0.027\ Popu$
Rule 14: HU > 406, WAC > 432, and DistUS ≤ 0.09887236	25	1386.8 (520, 3750)	767.0	$AADT_i = 259.8 + 37040.6\ DistUS$
Rule 15: WAC > 308, and Distl ≤ 1.990492	29	1849.0 (115, 4550)	1074.4	$AADT_i = 1901.1$

Table 6. Rules generated for rural local (7R) roadways.

Rules	Cases	Average (range)	Est Err	Rule based models
Rule 1: WAC > 573, WAC ≤ 584, DistUS > 8.029039, and DistUS ≤ 22.27129	32	51.9 (5, 425)	38.7	$AADT_i = 304.2 - 15.02\ DistI + 16.8\ DistUS$
Rule 2: WAC > 175, WAC ≤ 584, DistI > 80.54391, and DistUS > 1.341689	335	78.5 (5, 910)	69.1	$AADT_i = 92.4 - 0.279\ WAC + 0.12\ HU - 1.7\ DistUS$
Rule 3: Popu ≤ 1294, WAC > 584, RAC > 493, and DistUS > 11.76125	29	99.5 (5, 580)	77.3	$AADT_i = -846.4 + 1.584\ WAC - 34ROUTE_SYST - 2.3DistUS - 0.059Popu - 0.26\ DistI - 0.02\ HU$
Rule 4: WAC ≤ 573, DistI > 16.03872, and DistUS > 1.341689	3016	108.5 (5, 2050)	82.1	$AADT_i = 51.7$
Rule 5: WAC ≤ 175	2866	119.3 (5, 2250)	84.6	$AADT_i = -23.5 + 0.11\ RAC + 0.058\ Pop - 0.7\ DistUS - 0.13\ DistI + 0.01\ HU$
Rule 6: WAC > 175, DistUS > 1.341689, and DistUS ≤ 8.029039	826	143 (5, 2600)	117.9	$AADT_i = 62.3 - 8.7\ DistUS + 0.28\ RAC - 0.22\ HU - 0.113\ WAC - 0.49 + 8\ ROUTE\_SYST$
Rule 7: HU ≤ 603, WAC > 175, and DistUS > 0.3158301	1129	204 (5, 4700)	190.1	$AADT_i = 996.3 - 662.6\ DistUS - 0.781\ Popu + 1.17\ HU$
Rule 8: WAC > 175, DistI > 2.920889, and DistI ≤ 14.28077	434	236.7 (5, 2050)	177.9	$AADT_i = 17.1 + 1.61\ DistI + 3.5\ DistUS - 0.05\ HU + 0.04\ RAC + 0.018\ Popu$
Rule 9: Popu ≤ 1488, WAC > 175, WAC ≤ 782, DistUS > 0.1038099, and DistUS ≤ 1.341689	309	298.5 (5, 4000)	284.2	$AADT_i = 115.6 - 0.521\ WAC - 6.6\ DistUS + 1.2\ DistI - 0.05\ HU + 0.008\ Popu$
Rule 10: WAC > 432, WAC ≤ 573, RAC ≤ 531, DistI > 16.03872, DistI ≤ 80.54391, and DistUS > 8.029039	23	337.6 (5, 1050)	347.4	$AADT_i = 1988.5 - 5.33\ RAC + 18.7\ DistI$
Rule 11: WAC > 584, and DistUS ≤ 11.76125	329	370.7 (5, 3200)	300.3	$AADT_i = 287.6 - 3.75\ DistI$
Rule 12: Popu ≤ 1488, HU ≤ 603, WAC > 175, WAC ≤ 782, DistUS ≤ 1.341689	276	378.2 (5 to 4000)	270.6	$AADT_i = 218.7 + 1.44\ HU - 0.817\ Popu + 0.512\ WAC$
Rule 13: WAC > 175, DistUS ≤ 1.058187	480	392.1 (5, 4000)	313.0	$AADT_i = 331.1$
Rule 14: WAC > 175, WAC ≤ 584, DistI > 2.920889, DistI ≤ 80.54391, and DistUS > 22.27129	45	425.9 (5, 2050)	406.7	$AADT_i = 904.5 + 4.7\ RAC - 2.431\ Popu$
Rule 15: DistI ≤ 2.920889	144	436.3 (5, 4700)	335.5	$AADT_i = 316.1$
Rule 16: WAC > 432, WAC ≤ 573, RAC > 531, DistI ≤ 80.5439, DistUS > 8.029039	46	447.2 (30, 1600)	324.5	$AADT_i = 1801.9 - 1.77\ RAC$
Rule 17: WAC > 782	279	449.6 (5, 3350)	401.1	$AADT_i = 392.6$
Rule 18: HU ≤ 603, WAC > 180, WAC ≤ 782, DistI > 65.43855, DistUS > 0.3158301, and DistUS ≤ 1.341689	36	457.9 (15, 1450)	468.3	$AADT_i = 4270.8 - 1080.8\ DistUS + 4.16\ HU - 3.35\ RAC - 18.95\ DistI - 1.544\ Popu$
Rule 19: WAC > 584, RAC ≤ 493, and DistUS > 11.76125	57	483.8 (10, 1550)	415.3	$AADT_i = -2781.5 + 8.45\ RAC + 10.34\ DistI - 6.5\ DistUS + 0.03\ WAC - 0.02\ HU$
	23	485.9 (10, 2050)	385.7	$AADT_i = -4741.8 + 22.043\ WAC$

(*Continued*)

Table 6. Continued.

Rules	Cases	Average (range)	Est Err	Rule based models
Rule 20: WAC > 180, WAC ≤ 290, and DistUS ≤ 0.1038099				
Rule 21: Popu > 1488, WAC > 175, WAC ≤ 782, and DistUS ≤ 1.341689	64	512.3 (20, 2700)	390.7	$AADT_i = 437.7$
Rule 22: WAC > 175, WAC ≤ 782, DistUS > 1.058187, and DistUS ≤ 1.341689	46	513.3 (5, 3850)	468.2	$AADT_i = 206.5 - 0.89\ DistI + 0.094\ WAC$
Rule 23: WAC > 175, WAC ≤ 573, DistI > 14.28077, DistI ≤ 16.03872, and DistUS > 8.029039	35	522.0 (5, 1750)	398.8	$AADT_i = -1072.2 + 72.2\ DistUS + 3.124\ WAC + 0.93\ DistI - 0.02\ RAC + 0.01\ Popu$
Rule 24: WAC > 290, and DistUS ≤ 0.1038099	67	621.7 (5, 2650)	492.3	$AADT_i = 624.7 + 0.09\ WAC + 0.04\ Popu - 0.44\ DistI$
Rule 25: Popu > 1294, WAC > 584, and DistUS > 11.76125	60	807.6 (10, 4700)	662.7	$AADT_i = 2195.8 - 35.9\ DistUS - 1.58\ RAC + 0.938\ WAC + 6.56\ DistI - 0.18\ HU$

Table 7. Rules generated for urban local (7U) roadways.

Rules	Cases	Average (range)	Est Err	Rule based models
Rule 1: DistI > 26.37144	509	737.0 (5, 3900)	446.1	$AADT_i = 794.1 + 0.13\,HU - 1.9\,DistI - 0.04\,Popu + 0.008\,WAC - 0.03\,RAC - 4\,DistUS + 7\,ROUTE\_SYST$
Rule 2: DistI ≤ 26.37144, and DistUS > 8.917374	384	893.5 (15, 4900)	598.3	$AADT_i i = 335.5 + 0.364\,WAC - 9.3\,DistI - 0.07\,Popu + 0.17\,HU - 7\,DistUS$
Rule 3: HU ≤ 2155, RAC > 1127, DistI ≤ 26.37144, and DistUS ≤ 8.917374	721	1338.6 (10, 4950)	874.5	$AADT_i = -752.5 - 1.01\,Popu + 1.5\,RAC + 1.84\,HU - 2.7\,DistI + 3\,DistUS$
Rule 4: HU > 2155	38	1461.2 (45, 5000)	1149.0	$AADT_i = -21475 - 244.1\,DistI + 9.05\,HU - 435\,DistUS + 389\,ROUTE\_SYST$
Rule 5: HU > 595, DistI ≤ 26.37144, and DistUS ≤ 4.564781	1119	1539.6 (5, 5000)	999.7	$AADT_i = 1224.4 + 169\,DistUS - 27.3\,DistI + 0.88\,HU - 0.53\,RAC + 36\,ROUTE\_SYST$
Rule 6: RAC ≤ 1127, DistI ≤ 26.37144, and DistUS ≤ 8.917374	1884	1646.9 (5, 5000)	1023.6	$AADT_i = 1140.3 + 83\,DistUS - 0.57\,HU - 3.3\,DistI - 0.02\,Popu$
Rule 7: HU ≤ 595, WAC > 457, DistI ≤ 26.37144, and DistUS ≤ 8.917374	462	1709.0 (60 to 4950)	966.8	$AADT_i = 3047 - 3.1\,HU - 60\,DistUS - 3\,DistI - 0.04\,RAC$
Rule 8: HU > 595, WAC > 457, RAC ≤ 1127, DistUS > 4.564781, and DistUS ≤ 8.917374	170	1841.3 (5, 5000)	1115.3	$AADT_i = -3224.6 - 124.3\,DistI + 595\,DistUS + 4.48\,HU - 2.87\,RAC + 90\,ROUTE\_SYST$
Rule 9: HU > 595, WAC > 457, RAC ≤ 1127, DistI ≤ 26.37144, and DistUS ≤ 0.5526485	66	2516.7 (125, 4950)	1407.4	$AADT_i = -1194.8 + 4.99\,RAC$

Conclusions

In transportation engineering, traffic volume is an important measurement used in safety and operational design. Low-volume roads (LVRs) are major components of a road network in any locality. However, traffic count-stations are often limited to roadways with high functional classifications, and such information is rarely available for LVRs. As LVR traffic continues to grow due to economic growth, estimating traffic volume becomes a necessity. A recent report indicated the importance of data-driven and innovation approaches to estimate traffic volume on LVRs. In previous studies, both statistical and ML models have been used in estimating traffic volumes. Regression models have been extensively used to estimate traffic volumes in many studies. However, regression models, in general, examine the average effects of the factors and ignore subgroup or cluster effect. As a result, interventions are often geared towards the mean effect, without consideration of any subgroup effect. On the other hand, ML models provide better prediction by considering the subgroup effect. Therefore, these models are not useful for practitioners due to their lack of interpretability.

This paper has demonstrated the value of using rule-based analysis methods to identify subgroups with heterogeneous profiles without imposing assumptions on the subgroups or method by using traffic count station data on the LVRs in, in this case, Minnesota. Generally, Cubist is characterized by a better performance in predicting traffic volume. Instead of an uninterpretable ML model, consideration of rule-based multi-variate linear regression models makes Cubist poised to deliver a model explanation. The regression models from each rule predict annual average daily traffic (AADT) for a particular LVR. This study has shown that ML algorithms such as Cubist are more robust compared to statistical models because no hidden assumptions are required. The rule-based estimation models are useful for traffic engineers for easy interpretation and decision making to improve traffic volume estimation on LVRs.

This study has, however, some limitations. First, the numbers of variables used in this analysis are limited. Second, results from the network-level conventional traffic volume estimation methods are not compared with the current model outcomes. Future replications with additional roadway geometry and demographic data are needed to gather better estimates of traffic volumes on LVRs.

Acknowledgements

The contents of this paper reflect the views of the author and not the official views or policies of the Minnesota Department of Transportation (MnDOT). The author likes to thank Dr. Ioannis Tsapakis and Dr. William Holik for their help.

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Subasish Das http://orcid.org/0000-0002-1671-2753

References

AASHTO. 2010. *Highway Safety Manual*. AASHTO, Washington D.C.

Achwan, N., and D. Rudjito. 1999. "Accident Characteristics on Low-Volume Roads Indonesia." *Transportation Research Record: Journal of Transportation Research Board* 1652: 103–110.

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

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

Blume, K., M. Lombard, S. Quayle, P. Worth, and J. Zegeer. 2005. "Cost-Effective Reporting of Travel on Local Roads." *Transportation Research Record: Journal of the Transportation Research Board* 1917: 1–10.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Cornejo-Bueno, L., J. N. Borge, E. Alexandre, K. Hessner, and S. Salcedo-Sanz. 2016. "Accurate Estimation of Significant Wave Height with Support Vector Regression Algorithms and Marine Radar Images." *Coastal Engineering* 114: 233–243.

Das, S., and I. Ioannis. 2020. "Interpretable ML Approach in Estimating Traffic Volume on LVRs." *International Journal of Transportation Science and Technology* 9 (1): 76–88.

Das, S., and X. Sun. 2015. *Developing a Method for Estimating AADT on All Louisiana Roads*. Report No. FHWA/LA.14/548. Baton Rouge: Louisiana Department of Transportation and Development.

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

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

Gecchelea, G., R. Rossia, M. Gastaldia, and A. Caprinia. 2011. "Data Mining Methods for Traffic Monitoring Data Analysis: A Case Study." *Procedia Social and Behavioral Sciences* 20: 455–464.

History of SVM. 2020. http://www.svms.org/history.html.

Ho, T. K. 1998. "The Random Subspace Method for Constructing Decision Forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8): 832–844.

Jessberger, S., R. Krile, J. Schroeder, F. Todt, and J. Feng. 2016. "Improved Annual Average Daily Traffic Estimation Processes." *Transportation Research Record: Journal of the Transportation Research Board* 2593: 103–109.

Kuhn, M. 2018. *Classification and Regression Training*. R Package Version 6.0-81.

Kuhn, M., and R. Quinlan. 2018. *Cubist: Rule- and Instance-based Regression Modeling*. R Package Version 0.2.2.

Liu, L., and S. Dissanayake. 2008. *Examination of Factors Affecting Crash Severity on Gravel Roads*. Manhattan: Kansas State University.

Lloyd, J. F., and M. S. French. 2006. *Stratification of Locally Owned Roads for Traffic Data Collection*. Final Report FHWA-PA-2006-009-050210. Harrisburg: Pennsylvania Department of Transportation.

Lowry, M. 2014. "Spatial Interpolation of Traffic Counts Based on Origin–Destination Centrality." *Journal of Transport Geography* 36 (0): 98–105.

McCord, M., Y. Yongliang, Z. Jiang, B. Coifman, and P. Goel. 2003. "Estimating Annual Average Daily Traffic from Satellite Imagery and Air Photos." *Transportation Research Record: Journal of the Transportation Research Board* 1855: 136–142.

Mohamad, D. 1998. "An Annual Average Daily Traffic Prediction Model for County Roads." *Transportation Research Record: Journal of the Transportation Research Board* 1617: 99–115.

Quinlan, J. 1992. "Learning with Continuous Classes." *Proceedings of the 5th International Conference on Artificial Intelligence '92*, Singapore, 343–348.

Quinlan, J. 1993. "Combining Instance-Based and Model-Based Learning." *Proceedings of the 10th International Conference on Machine Learning*, San Mateo, CA, 236–343.

Quinlan, J. 1996. "Improved Use of Continuous Attributes in C4." *Journal of Artificial Intelligence Research* 4: 77–90.
Seaver, W. L., A. Chatterjee, and M. L. Seaver. 2000. *Estimation of Traffic Volume on Rural Non-State Roads*. Research Report. Knoxville: University of Tennessee.
Selby, B., and K. Kockelman. 2011. *Spatial Prediction of AADT in Unmeasured Locations by Universal Kriging*. Research Report. Austin: The University of Texas.
Sharma, S., P. Lingras, F. Xu, and P. Kilburn. 2001. "Application of Neural Networks to Estimate AADT on Low-Volume Roads." *Journal of Transportation Engineering* 127 (5): 426–432.
Shawn, S., I. Sener, M. Martin, S. Das, E. Shipp, R. Hampshire, K. Fitzpatrick, et al. 2017. *Synthesis of Methods for Estimating Pedestrian and Bicyclist Exposure to Risk at Areawide Levels and on Specific Transportation Facilities*. Washington: Federal Highway Administration.
Shen, L. D., F. Zhao, and D. I. Ospina. 1999. *Estimation of Annual Average Daily Traffic for Off-System Roads in Florida*. Final Report. Tallahassee: Florida Department of Transportation.
Smola, A. J., and B. Schölkopf. 2004. "A Tutorial on Support Vector Regression." *Statistics and Computing* 14 (3): 199–222.
Staats, W. 2016. *Estimation of AADT on Local Roads in Kentucky*. (Unpublished M.S. Thesis). University of Kentucky.
Stamatiadis, N., S. Jones, and L. Aultman-Hall. 1999. "Causal Factors for Accidents on Southeastern Low-Volume Rural Roads." *Transportation Research Record: Journal of Transportation Research Board* 1652: 111–117.
Xia, Q., F. Zhao, Z. Chen, L. Shen, and D. Ospina. 1999. "Estimation of Annual Average Daily Traffic for Nonstate Roads in a Florida County." *Transportation Research Record: Journal of the Transportation Research Board* 1660: 32–40.
Zegeer, C. V., R. Stewart, F. Council, and T. R. Neuman. 1994. "Accident Relationship of Roadway Width on Low-Volume Roads." *Transportation Research Record* 1445: 160–168.
Zhao, F., M. T. Li, and L. F. Chow. 2004. *Alternatives for Estimating Seasonal Factors on Rural and Urban Roads in Florida*. Final Report BD015-03. Tallahassee: Florida Department of Transportation.