ABSTRACT

Title of Thesis:	NATIONWIDE ANNUAL AVERAGE DAILY TRAFFIC (AADT) ESTIMATION ON NON-FEDERAL AID SYSTEM (NFAS) ROADS BY MACHINE LEARNING WITH DATA MINING OF BUILT-IN ENVIRONMENT
	Qianqian Sun, Master of Science, 2020
Thesis Directed By:	Professor Lei Zhang, Department of Civil and Environmental Engineering

This study aims to address the nationwide gap in AADT data on NFAS roads in U.S. With a Spatial Autoregressive Model as a benchmark, two machine-learning approaches, i.e. Artificial Neural Network and Random Forest, show notable improvement in the accuracy of estimating AADT according to five measures, i.e. MSE, RSQ, RMSE, MAE, and MAPE. A data-mining of the built-in environment from three perspectives, i.e. on-road and off-road features, network centralities, and neighboring influences, paves the way for AADT estimation, which covers 87 variables in centrality, neighboring traffic, demographics, employment, land-use diversity, road network density, urban design, destination accessibility, etc. Data integration using different buffering sizes and statistical analysis of linearity and monotonicity promote the variable selection for estimation. When implementing the machine-learning approaches, not only the estimation performance is analyzed, but also the relationship between each variable and AADT, the interplays among variables, variable importance measures are thoroughly discussed.

NATIONWIDE ANNUAL AVERAGE DAILY TRAFFIC (AADT) ESTIMATION ON NON-FEDERAL AID SYSTEM (NFAS) ROADS BY MACHINE LEARNING WITH DATA MINING OF BUILT-IN ENVIRONMENT

by

Qianqian Sun

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2020

Advisory Committee:
Professor Lei Zhang, Chair
Professor Paul M. Schonfeld
Assistant Professor Vanessa Frias-Martinez

Table of Contents

Chapter 1: Introduction

Annual average daily traffic (AADT) is an important traffic parameter for federal, state, and local transportation agencies in making transportation planning and policy decisions. As an indispensable input of highway statistics, AADT is widely used for many transportation tasks, such as maintaining and evaluating highway projects, making decisions on transportation plans and policies, and conducting various transportation research and studies.

According to 2016 Highway Safety Improvement Program (HSIP) Final Rule, State agencies are required to have access to AADT on all paved roads open to public travel including Non-Federal Aid–System (NFAS) roads by 2026. In traffic monitoring guide, it is stated that AADT should be reported for Highway Performance Monitoring System (HPMS). In practice, traffic data mainly come from two data programs. First is permanent monitoring of continuous traffic flow 24 hours a day and 7 days a week throughout the entire year. Second is temporary monitoring that collects short-duration traffic data several times a year or once at several years, usually based on 24-hour, 48-hour, or 72-hour intervals. Traffic Monitoring Stations (TMS) that collect either continuous or short-duration traffic data are therefore distributed throughout the country to provide U.S. Traffic Monitoring Location Data. A couple of items are gathered, including traffic counts along with speed, vehicle class and vehicle weight.

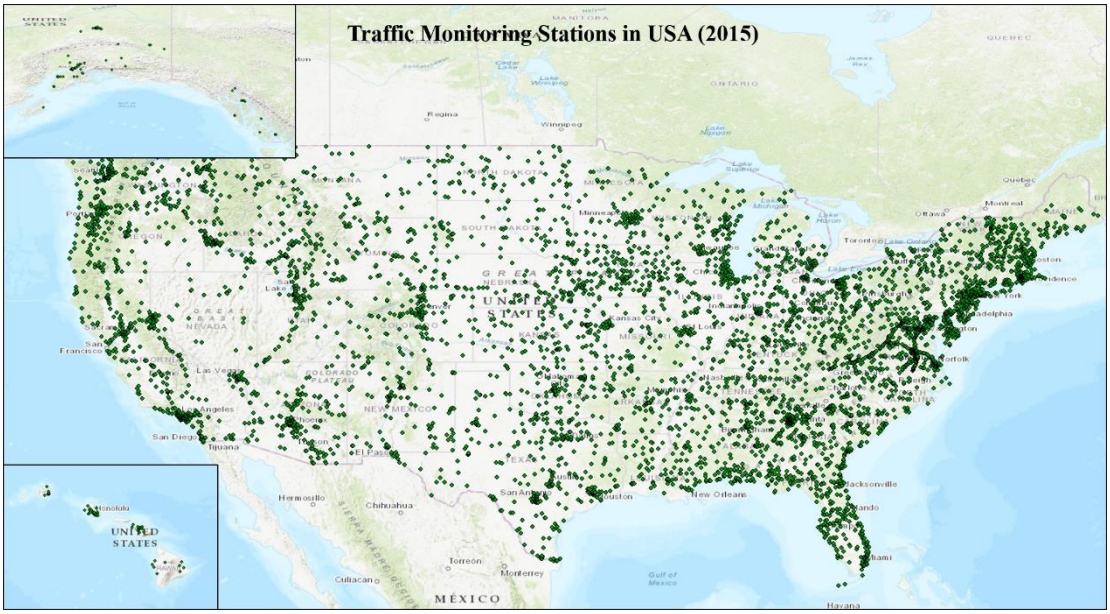

Figure 1-1 Distribution of traffic monitoring stations in USA

As shown in figure 1-1, 7430 Traffic Monitoring Stations that collect continuous or short-duration traffic data are working throughout the country as of 2015. Yet, there are around four million miles of highways according to National Transportation Statistics. A huge supply-demand gap obviously exists between data needs and available stations. Indeed, these TMSs provide consistent and well-structured traffic data at national level. However, implementing this program everywhere in the highway network is too prohibitive to be realistic. Roads of higher functional classes such as arterials are prioritized as a result. Data gap still exists on a large proportion of the highways, which is particularly serious on Non-federal Aid System (NFAS) roadways, that is rural minor collectors and locals in both rural and urban areas defined as FSystem = 6 and 7 respectively according to the Chapter 3 of 2016 HPMS Field Manual. Highway Statistics 2016 reports the annual Vehicle Miles Traveled (VMT) by functional classes, which shows that NFAS roads consist of a large portion of VMT. As table 1-1 shows, NFAS roads account for 15.1% of the annual

VMT, which is only second to urban interstate (17.66%) and urban other freeways and expressways (15.14%). As for lane miles, functional class local has a percentage of 45.7% in rural areas and 19.85% in urban areas. NFAS roads in total have a proportion as high as 71.4%. It is entirely conceivable that NFAS roads play an important role in carrying on personal and freight flows in our daily life.

Table 1-1 Vehicle miles traveled and lane miles on NFAS roads and others

Area	Functional class	VMT (millions)	Lane miles
Rural	Interstate	247152 (7.75%)	119158 (1.36%)
	Other freeways and expressways	34434 (1.08%)	24542 (0.28%)
	Other principal arterial	190090 (5.96%)	231532 (2.64%)
	Minor arterial	143525 (4.50%)	276684 (3.15%)
	Major collector	160066 (5.02%)	818994 (9.33%)
	Minor collector (NFAS)	47674 (1.49%)	517439 (5.90%)
	Local (NFAS)	127634 (4.00%)	4010341 (45.70%)
Urban	Interstate	563112 (17.66%)	105457 (1.20%)
	Other freeways and expressways	250325 (7.85%)	58943 (0.67%)
	Other principal arterial	482923 (15.14%)	237381 (2.71%)
	Minor arterial	411774 (12.91%)	296203 (3.38%)
	Major collector	207354 (6.50%)	278414 (3.17%)
	Minor collector	16487 (0.52%)	58584 (0.67%)
	Local (NFAS)	306424 (9.61%)	1741865 (19.85%)

Notes: 1. NFAS roads account for 15.1% of total VMT,
2. NFAS roads account for 71.4% of total lane miles,
3. Data source is Highway Statistics 2016.

Filling the data gaps has been challenging transportation agencies, practitioners and researchers for a long time. There is still no uniform methodology for AADT estimation on lower-level roads. Traditional methods utilize the short-duration traffic counts data to estimate AADT through developing adjustment factors, which is usually called factoring method. It is popularized for its being simple and easy to be applied. Generally, all roads first need to be classified into homogeneous groups based on a certain criteria such as functional class and the geographical units (e.g. counties). Then

within each group, the continuous TMSs serve as the source of adjustment factors to convert the short-duration traffic data into AADT. There is no doubt that the accuracy of this method heavily depends on the grouping process. This method is more effective on high-volume roads than low-volume roads since continuous TMSs are mainly located at roads of higher functional classes. Still, AADT estimation on NFAS roads with a desirable accuracy level is a difficult research problem. Researchers have been putting efforts on this topic by leveraging various methodologies. Statistical regression modeling is a major branch, which models the data generation process of AADT by capturing its distribution patterns such as Gaussian distribution and binomial distribution. Strong evidence from inferential tests enables it to yield good estimation results but in most of the time this is not the case in practice. This limits the performance of parametric models, where nonparametric modeling such as Artificial Neural Network (ANN) and Random Forest (RF) from machine learning family comes for AADT estimation. The reasons why machine learning outperforms others regarding this research topic are summarized as below.

A. Machine learning algorithms first and foremost perform well in terms of accuracy,
B. The relationship in real world among the features are usually non-linear. Sometimes it is too complicated to be modeled by mathematical models, where machine learning algorithms can handle high-order relationships,
C. Unlike parametric approaches, machine learning methods do not necessitate any statistical assumptions making it fairly flexible to be applied to data with or without a certain distribution pattern,

D. Machine learning algorithms are good at handling large-volume data just like NFAS roads in a time-efficient way. Not only statewide estimation but also nationwide estimation becomes feasible,
E. Machine learning algorithms are tolerant of data noises, while statistical models are sensitive to the disturbances from noisy data,
F. The application of machine learning algorithms is simple and easy benefiting from many well-developed and straightforward packages,
G. With more and more techniques that demystify the inner structures of trained model, machine learning gains more and more interpretability instead of being a complete black box.

Encountered with the limited data availability of traffic data, external databases are fully utilized to help estimate AADT. To some extent, traffic volume represents the strength of daily activities, which is closely related to the social demographic features of personals and the environment characteristics such as land use pattern, urban design, accessibility, road design, etc. Accordingly, as much as possible measures of the built-in environment are thoroughly analyzed to provide valuable inputs for AADT estimation. In this study, a total of 79 features from Smart Location Database (SLD), a well-structured nationwide database on built-in environment, are extracted for analysis. Additionally, the centrality of roads in the whole transportation network directly influences its capacity of delivering traffic. For example, if the shortest paths in the network are frequently passing one road, the traffic volume of this specific road should be higher than normal roads. Being enlightened from social network theory, this study makes a bold trial of employing centrality measures of road segment to help improve

AADT estimation. Furthermore, the interaction among the roads regarding traffic volume is not negligible and hence spatial dependence analysis among roads initiates the involvement of neighboring traffic features. In summary, all kinds of features from built-in environment are fully discussed to extract strong predictive factors as much as possible. Different ways of integrating spatial data are also compared to enhance the predictive power of variables. Finally, twelve predictors distinguish themselves to act as the potentially good predictors for AADT estimation. This feature exploration process provides an informational guide on future similar studies in this field.

Results show that machine learning algorithms, i.e. ensembling ANNs and RF, in this study, achieve high accuracy level measured from multiple ways including Mean Squared Error (MSE), R Squared Value (RSQ), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). A spatial autoregressive model, which is believed to be the best statistical model from literature review, is built as a benchmark model to be compared with the machine learning algorithms. Even though it works well, it is not as good as the ensembling artificial neural networks and random forest.

In summary, this study has the following objectives.

1. To develop a simple and easy, time-efficient and cost-effective procedure for conducting nationwide AADT estimation on NFAS roads and ensure a desirable accuracy meanwhile,

2. To consolidate public domain databases including HPMS data and SLD, both of which are ready-to-use, well-structured, comprehensive, and nationwide for a practical application purpose,
3. To conduct a deep data mining of built-in environment features from all kinds of aspects to provide potentially valuable predictors for AADT estimation as much as possible,
4. To analyze the relationship, either non-linear or linear, between features and AADT during feature selection process and to compare several spatial data integration methods based on the relationship analysis results,
5. To not only apply machine learning algorithms, i.e. artificial neural network and random forest, but also demystify the interactions among the features,
6. To compare the performance of artificial neural network and random forest with a benchmark model – a spatial autoregressive model.

Chapter 2: Literature review

2.1 Traditional factoring methods

The procedure for AADT estimation on lower-level roads by traditional factoring methods involves three steps. First, homogeneous permanent traffic count stations are classified into multiple groups. Then short-term traffic count stations are assigned to these groups. The continuous traffic monitoring data collected by permanent traffic count stations provide all types of factors, including hourly, daily, weekly, and monthly expansion factors. These factors are used for converting the short-duration traffic counts into AADT. This method is widely used across the country for its simplicity, effectiveness, and relatively low cost. However, the traditional factoring method has many deficiencies. Rossi et al. (1) summarizes error sources of factoring approach: determining the number of groups, identifying groups of road sections, and applying wrong expansion factors. Moreover, accuracy of AADT estimation based on factoring approach is very sensitive to the assignment of STTCs to PTC groups (2). Inappropriate assignments could lead to high estimation error. Additionally, assigning STTCs has always been a difficulty of this approach. Even though assignment methods, such as agglomerative hierarchical clustering method, k-means clustering method, etc. have been proposed to improve accuracy, they all have various deficiencies (1).

2.2 Statistical regression models

Many researchers propose regression models for AADT estimation and find that factors such as population, area type (rural or urban), per capita income, and roadway characteristics, etc. are significantly correlated with AADT (3, 4, 5). Zhao and

Chung (3) build linear regression models to estimate the AADT in Broward County, Florida using land-use and accessibility measures. They find that functional class (transformed nominal variables with numeric values) and number of lanes are significant factors in the estimation of AADT. Xia et al. (4) also suggest that roadway characteristics such as number of lanes and area type are significant factors in AADT estimation.

Zhao and Park (6) apply a geographically weighted regression (GWR) model to estimate the AADT of Broward County in Florida. Compared with an ordinary linear regression model, GWR allows the parameters of regression model to be locally-unique instead of globally-uniform. Spatial nonstationarity, meaning the relationship between independents and the dependent varies across the study area, is considered in the GWR model. Eom et al. (7) improve the general regression model by incorporating spatial statistical process. Three semivariogram models (i.e. Guassian, exponential, and spherical semivariogram) are compared for analyzing the spatial autocorrelation of data points; two interpolation methods (i.e.(ordinary Kriging and universal Kriging) are compared for estimating unknown data points. Shamo et al. (8) apply a linear spatial interpolation to interpolating the AADT in Washington State, in which different combinations of kriging techniques and variogram models are compared. This spatial modeling stands out for its capturing the spatial relationship among data points by geostatistical procedures. However, the feasibility of the interpolation method depends on not-sparsely-distributed spatial data points. Therefore, its applicability to estimating AADT on local roads at the link level lacks feasibility. Geol et al. (9) demonstrat that a correlation-based method can yield better estimates than traditional methods if traffic

volumes of road sections are significantly correlated with that of nearby roads. They propose a generalized-least-squares (GLS)-estimation-based method and apply it to Ohio intercity network, which is generated by Monte Carlo simulation. Even though the performance of this method in real network needs further investigation, it provides insights on the correlation issue among AADTs. Accordingly, some generalized regression models (10-12) and spatial statistical models (6-8, 13) were explored for estimating AADT.

2.3 Spatial statistical models

Spatial autoregressive models (SAM) (e.g. spatial lag model (SLM) and spatial error model (SEM)) give insights into the spatial autocorrelations in an OLS model of AADT estimation. They are more powerful techniques than GWR and they do not have the drawback like the Kriging-based method. Besides, SAM can be applied in various settings as long as spatial autocorrelations cannot be ignored in a normal regression model for AADT estimation. In this paper, spatial autoregressive models are set as the benchmark model for comparison purpose with machine learning algorithms. There are various spatial-statistical techniques utilized as revealed in the literature to estimate the AADT on different roadway functional classifications. For example, clusters of roads with similar volume level and functional classification can be created and used to apply spatial interpolation such as kriging, inverse distance weighting (IDW), neural neighbor (NN), and trend technique (29). Also, geographically-weighted regression models (GWR) have been proposed and applied for AADT estimation in a few recent studies. GWR model assumes that land-use and demographical variables are non-stationary across space, which means the statistical properties (mean, variance, etc.) of

variables are different among various locations. Therefore, the relationship between predictors and the response varies; this means that the parameters of dependent variables should not be fixed. At different locations, the influence of predictors on the response varies. For example, car ownership may have a greater influence on AADT at location A than location B. Then, its estimated parameter at location A could be larger than that at location B. The GWR model allows the parameters to be locally- rather than globally-estimated to reflect the mentioned non-stationarity. Local estimations mean that the parameters are more often determined by nearby observations than farther ones. In this way, local variations can be taken into consideration when exploring the relationship between independents and the dependent. Additionally, an important assumption for the GWR model is that the error terms are independent and identically distributed with zero means and constant variance. In the GWR model, a weighted window is moved over the data to estimate a set of parameters for each data point. Bi-square function and Gaussian function are two commonly-used weighting functions for the GWR model with one critical parameter – bandwidth.

The spatial-statistical method has several advantages. First, it considers the spatial non-stationarity of land-use and demographical variables, which is more reasonable than ordinary regression models when estimating the AADT. In addition, this methodology is economical in its data requirements because it uses existing traffic counts and does not require collection of additional count data. Moreover, spatial-statistical models can be implemented easily in standard GIS software packages that are readily available to local-road agencies. The methodology can also be updated easily in the future after the agencies receive new traffic-count data. Finally, the

methodology is straightforward and does not require complex procedures. It can be transferred and adapted rather easily to jurisdictions in other states to estimate their local-road AADTs.

2.4 Machine learning algorithms

In machine learning field, ANN and RF are two commonly used algorithms for prediction tasks. Karlaftis and Vlahogianni (14) thoroughly analyze the application of ANN in transportation research: ANN has been successfully applied as a data analytic method for solving transportation problems because of "their modeling flexibility, their learning and generalization ability, their adaptability, and their-generally-good predictive ability" (14). As summarized by Karlaftis and Vlahogianni (14), parameters of ANN are very adaptable. It is also good at addressing outliers and missing values and absorbing noises. ANN method is very practical in reality since no assumption is required and their nonlinear structure can capture complicated data patterns and model complex relationships (15). Duddu and Pulugurtha (11) implement a neural network model using back-propagation (BP) learning algorithm to estimate AADT and found that prediction results are better than that of the negative binomial count statistical model. Sharma et al. (16) also used a BP neural network to estimate AADT by setting hourly traffic volume factors as inputs. Zarei etc. use Random Forest as the prediction model for short-term traffic flow prediction (21). Hamner uses RF to predict travel flow in six and thirty minutes (22).

Chapter 3: Data processing

Data processing includes feature selection and feature engineering. The former involves three sections: on-road and off-road features, network centrality analysis, and spatial dependence analysis. The latter involves outlier detection and normalization. The local roads being analyzed in this research are the locals in rural and urban areas and minor collectors in rural only areas as defined as FSystem = 6 and FSystem = 7 respectively by FHWA according to HPMS field manual. All local roads for the entire United States are used for analysis. After data cleaning, there are 10490 road segments with reported AADT values for modeling specification and validation.

Two public domain databases are employed for AADT estimation in this research: 2012 HPMS data and 2012 SLD. The reason for choosing 2012 HPMS data is for the consistency with 2012 SLD to extract the most representative predictors. HPMS data provides nationwide AADT on each road segment as well as roadway attributes. The SLD is developed by the Environmental Protection Agency (EPA) for the entire U.S. and provides data on built-in environment characteristics such as demographics, employment, land use entropy, urban design, density, and destination accessibility at Census Block Group (CBG) level. For this part of feature selection, the way of merging CBG-based SLD features with the link-based AADT and road attributes affects estimation results as well. A state-of-the-practice way is building buffers, but there is no study examining the most appropriate size of the buffering. Thus, different buffer sizes are investigated and compared. These two databases provides 79 on-road and off-road features for analysis. Furthermore, network centrality

analysis based on social network theory is conducted for exploring useful predictors. The shapefile from HPMS is used for building transportation network. The HPMS data is also used for spatial dependence analysis.

3.1 Feature selection

3.1.1 On-road and off-road features

In order to improve local AADT estimation as much as possible, external databases, including SLD and HPMS databases, are fully utilized to provide influential predictors. Even though previous studies have already applied some built-in environment factors for AADT estimation, no study so far has fully investigated all potential predictors and especially their rationality. From the perspective of providing theoretical basics and practical guides, a long list of variables from SLD and HPMS datasets (i.e. 79 variables in total) is extracted and analyzed in terms of their potential relationship with AADT.

Simply speaking, either linear relationship or non-linear relationship exists between the predictors and AADT. Considering that common statistical methods are parametric-based while machine-learning methods does not necessitate a specific distribution of the variables, both Pearson Correlation Coefficient (r) and Spearman Rank-order Correlation Coefficient (rs) are employed to present the performance of all predictors. Pearson correlation measures the strength of linearity between two variables. Its coefficient falls into the range of minus 1 to 1. The closer the coefficient is to zero, the less linearity the test indicates. Spearman correlation measures the monotonicity between two variables in a non-linear way and it does not necessitate a

Gaussian distribution for both variables. For this research, detecting a statistically significant monotonicity by Spearman test contributes a lot to the predictor selection whether such correlation is linear or not since machine learning algorithms applied in this study are able to give full play to theirs advantage of handling non-linear relationships. It seems like Spearman correlation test is enough for selecting variables but Pearson correlation is also necessary since Spearman test might underestimate the linear correlation. In other words, the variables that show significant linearity by Pearson test might show insignificant association by Spearman test. Thus, both tools are utilized to select predictors.

Table 3-1 summarizes the linearity analysis results and table 3-2 summarizes the monotonicity analysis results. In general, built-in environment factors do not show an obvious linear correlation with AADT, by obvious it means the coefficient value is below 0.5 or above -0.5. This conclusion is expected since it is quite probable that a complex rather than a linear relationship exists between a predictor and the AADT. As shown from the results of Spearman Correlation test, non-linear relationships indeed widely exist between the built-in environment factors and AADT with a Spearman coefficient higher than 0.5 or smaller than -0.5.

Among the 79 factors from SLD and HPMS, 77 are continuous variables and 2 are dummy variables. The two dummy variables are UrbanCode (1 for rural sections, 2 for small urban sections, and 3 for urban sections), and FSystem (6 for minor collectors in rural area and 7 for locals in both rural and urban areas). Only the variables

with relative notable correlations are presented and discussed in this study. In the summary tables, the scheme of grouping is consistent with that of SLD.

Table 3-1 Linearity analysis by different buffers

Type	Groups	Feature	r	Groups	Feature	r	Mean
1-mile buffer	demographics*	autoown2p	0.412	road traits**	throughlane	0.409	0.329
		counthu	0.319		urbancode	0.367
		workers	0.313		fsystem	-0.321
		rhiwagewk	0.310	density*	d1c8ret10	0.368
		rmedwagew	0.305	urban design*	d3a	0.340
		autoown1	0.304		d3aao	0.321
		totpop	0.304		d3amm	0.308
		hh	0.301		d3bpo3	0.303
	employment*	e5ent10	0.322		d3apo	0.302
		e8ent10	0.322			
2-mile buffer	demographics*	autoown2p	0.419	density*	d1c8ret10	0.411	0.357
	urban design*	d3a	0.361		d1c5ent10	0.317
		d3bpo3	0.356	road traits**	throughlane	0.409
		d3apo	0.355		urbancode	0.367
		d3amm	0.334		fsystem	-0.321
		d3b	0.332	employment*	e5ent10	0.349
		d3bmm3	0.320		e8ent10	0.349
3-mile buffer	road traits**	throughlane	0.409	urban design*	d3apo	0.391	0.296
		urbancode	0.367		d3bpo3	0.391
		fsystem	-0.321		d3b	0.370
	density*	d1c8ret10	0.366		d3amm	0.356
					d3bmm3	0.333

Notes: *: the data source is smart location database,
**: the data source is highway performance monitoring system,
r: Pearson correlation coefficient,
Sample size is 10772.

Table 3-2 Monotonicity analysis by different buffers

Type	Groups	Feature	rs	Groups	Feature	rs	Mean
1-mile buffer	density*	d1c8ret10	0.572	demographics*	autoown2p	0.472	0.450
		d1a	0.430		hh	0.436	
		d1b	0.427		rhiwagewk	0.432	
	road traits**	fsystem	-0.534		counthu	0.431	
		urban_code	0.461		totpop	0.426	
	accessibility*	d5ae	0.485		autoown1	0.402	
		d5ar	0.470	urban design*	d3a	0.407	
	diversity*	d2aephhm	-0.409		d3bpo3	0.401	
	density*	d1c8ret10	0.598		fsystem	-0.534	0.467

2-mile buffer	accessibility*	d5ae	0.436	road traits**	urbancode	0.461	
		d5ar	0.421	urban design*	d3a	0.402	
	demographics*	autoown2p	0.421				
3-mile buffer	density*	d1c8ret10	0.624	urban design*	d3a	0.429	0.454
	road traits**	fsystem	-0.534		d3bao	0.428	
		urbancode	0.461		d3apo	0.428	
	diversity*	d2awrkemp	0.409		d3aao	0.412	
	accessibility*	d5ae	0.409		d3bpo3	0.405	

Notes: *: the data source is smart location database,
**: the data source is highway performance monitoring system,
1, 2, 3 … : the rank of correlation,
rs: Spearman coefficient,
Sample size is 10,772.

The way of integrating spatial data doubtlessly influences the inner relationship among features. Integrated features based on single buffers of different sizes are discussed.

3.1.1.1 Feature analysis based on 1-mile buffers

Regarding the Pearson test results of 1-mile based buffering, the most significant one is AutoOwn2P, which is the number of households in CBG that own two or more automobiles. This variable makes a lot sense since the more automobiles a household has the more traffic it generates. ThroughLane (number of through lanes), a major attribute of road geometry, directly influences traffic volume. D1C8Ret10 measures the gross retail (8-tier) employment density (jobs/acre) on unprotected land. It ranks third in terms of the strength of linearity with AADT. It is interesting that there are eight types of employment density including retail, office, industrial, service, entertainment, education, health care, and public sector but only retail employment density has a noticeable linear correlation with AADT. UrbanCode measures the

urbanization degree of the area. The more urbanized the area is, the more traffic volume there is as shown from results in the table 3-1 and table 3-2. From the group of urban design, D3a (total road network density) shows a relatively clear linear correlation with AADT, which is not difficult to understand in view of the definition of D3a.

As for machine learning algorithms, linearity between the predictors and response variables is not a prerequisite. As long as a significant correlation is presented from a statistical test such as spearman correlation test, this predictor cannot be ignored. As shown in table 3-2, some variables are significantly correlated with AADT with a Spearman coefficient higher than 0.45 or smaller than -0.5. D1C8Ret10 is the gross retail (8-tier) employment density (jobs/acre) on unprotected land, which ranks third regarding the linearity with AADT with a p-coefficient of 0.367549. Its non-linear correlation with AADT seems to be more obvious because of a quite high Spearman coefficient of 0.571981. Similarly, FSystem, AutoOwn2P, and UrbanCode also show high non-linear correlation with a Spearman coefficient of -0.534167, 0.471896 and 0.460472 respectively. Although they also outperform others in Pearson test, their non-linear correlation with AADT is more statistically significant than the linear correlation. Besides, two new groups of factors present non-linear correlation with AADT including destination accessibility and land use diversity. Destination accessibility in SLD measures number of jobs or working-age population within a 45-minute commute through automobile or transit (refer to SLD guide). Among the 12 variables on destination accessibility, D5ae (working age population within 45 minutes auto travel time, network travel time-decay weighted) and D5ar (jobs within 45 minutes auto travel time, network travel time-decay weighted) present obvious association with

AADT for a Spearman coefficient of 0.484990 and 0.469602. Land use diversity measures the entropy of mixed land use. There are various measures included in SLD, among which D2AEPHHM shows the most noteworthy association with AADT. It measures the diversity of employment and household. Detailed information can refer to the SLD user guide. It has a Spearman coefficient of -0.408495, which means the more diverse (i.e. equally mixed) the area is the less traffic there is. According to the calculation of D2AEPHHM in SLD, it is probable that households or some types of employment have a major influence on traffic volume. When the entropy gets larger, this influence gets weaker.

3.1.1.2 Feature analysis based on 2-mile buffers

It is noteworthy that data merging based on 2-mile single buffers yields more notable strength of linearity. Except that UrbanCode and FSystem remain the same Pearson coefficient value, all the other variables present a higher positive Pearson coefficient compared with the results from 1-mile based buffering. The average strength of linearity is calculated by averaging the absolute values of all Pearson coefficients. This yields 0.3571, which is higher than that in the case of 1-mile buffers, i.e. 0.3289. In consideration of linear relationship, 2-mile buffers are recommended. Among the 14 presented variables, 11 of which are overlapped with the variables showing significant linearity using the data merging method of 1-mile buffers, which shows the stability of these predictors.

There are fewer variables that have a Spearman coefficient higher than 0.4 or smaller than -0.4 than the other two cases (i.e. 1-mile based and 3-mile based

buffering). However, the average strength of association with AADT under 2-mile buffering is the largest, i.e. 0.4674 compared with 0.4495 (1-mile buffering) and 0.4538 (3-mile buffering). For this reason, 2-mile buffering outperforms the other two data merging ways. Among the variables with 2-mile buffering, all of the eight variables, including D1C8Ret10, D5ae, D5ar, AutoOwn2P, FSystem, UrbanCode, and D3a, present notable correlations with AADT in cases of both 1-mile and 2-mile buffering. These variables worth considering for their consistently showing the correlation with AADT. For reducing redundancy purpose, D5ar can be ignored for its similarity with D5ae. D5ar measures the jobs within 45 minutes auto travel time, network travel time-decay weighted and D5ae measures the working age population within 45 minutes auto travel time, network travel time-decay weighted.

3.1.1.3 Feature analysis based on 3-mile buffers

Once again, ThroughLane, UrbanCode, D1C8Ret10, and D3apo prove to be the ones with significant linear correlation with AADT just like the case in 1-mile buffering and 2-mile buffering. However, AutoOwn2P, D3a, and E5ENT10, all of which show strong linearity in cases of both the 1-mile and 2-mile buffering, do not show a notable linearity at all in the case of 3-mile buffers. What's more, the average strength of linearity based on 3-mile buffering is only 0.2958, which is smaller than that under the other two cases.

Familiar variables, D1C8Ret10, FSystem, UrbanCode, D2AEPHHM, D5ae, and D3a, appear again. The stability of these predictors distinguishes themselves as good candidate variables for AADT estimation. Similarly, D1C8Ret10 ranks the first

in terms of the strength of correlation with AADT just like the other two cases. In Pearson tests, this factor shows relatively high linear correlation than other factors as well. FSystem and UrbanCode' Spearman coefficients do not change with the data merging methods since they are the attributes along with the road segment from HPMS dataset. Data merging only applies to the polygon-based attributes from SLD database. D2AWRKEMP is a new variable, which only occurs in 3-mile based buffering tests. It measures the land use diversity, which is the household workers per job by CBG.

With the criteria of reducing redundancy, variables with similar definitions and probably high correlations with each other should not be completely included. Additional attention needs to be paid to testing the multicollinearity among predictors when the model to be applied has dependency prerequisite of predictors, such as regression models. Take demographics as an example, the variable counthu measures housing units and HH measures occupied housing units. Workers is the number of workers in CBG (home location). Totpop is the population. These four variables provide similar information to an extent. Besides, rhiwagewk is the number of workers earning $3333/month or more (home location) and rmedwagew is the number of workers earning more than $1250/month but less than $3333/month (home location). Only one of the two factors should be sufficient for analysis. Since AutoOwn2P shows a significant linear association with AADT, autoown1 (number of households in CBG that own one automobile) becomes a negligible predictor. The group of urban design measures the density of street network (D3a, D3aao, D4amm, D3apo) and street intersection (D3bpo3, D3b, D3bmm3). D3a measures the total road network density); D3aao, D3amm, and D3apo measure the network density by facility orientations (auto-

oriented, multi-modal, pedestrian-oriented, respectively). It is unnecessary to implement all these variables. Under different ways of buffering, D3a and D3apo should be sufficient for modeling. D3b measures the total intersection density; D3bpo3, D3bmm3 measure the intersection density by different intersection types (pedestrian-oriented with 3 legs and multi-modal with 3 legs). In comparison with D3a-related variables, all D3b related variables show less linearity for all cases of buffering.

In view of variable stability, some variables are also recommended for application. For example, FSystem (6 for minor collectors in rural area and 7 for locals in both rural and urban areas) shows evident linear correlation in all three cases. It makes a common sense because minor collectors usually have more traffic volume than locals just like their definitions. E5ENT10 (entertainment jobs within a 5-tier employment classification scheme) also show notable linear correlation with AADT in all cases. It has exactly the same Pearson coefficient value as E8ENT10 (Entertainment jobs within an 8-tier employment classification scheme). Only E5ENT10 is kept for analysis to reduce variable redundancy. Among all types of jobs, only the number of entertainment jobs shows evident association with traffic volume. It presents the strength of entertainment activities to some extent.

3.1.2 Network centrality analysis

In addition to on-road and off-road characteristics, the road network plays a role in influencing traffic flows. Suppose that the importance level of a road section or an intersection in the road network can reflect traffic volume to some extent. Specifically, if a road segment is frequently passed through by the shortest paths of node pairs in the

network, its traffic volume is expected to be higher than other links. If an intersection is connected with multiple legs, its importance is apparent. So how about the segments that are connected with multiple segments? In social network theory, there have been well-established methods for evaluating the centrality of a node or an edge in a network. Given an edge, the edge betweenness can be measured by the fraction of total shortest paths that go through this edge (Brandes, 2001). This theory is adopted to assess the centrality of road segments in transportation network. Degree centrality measures the importance of a node by the number of edges that it connects (Shaw, 1954). Similarly, the degree centrality of a road segment in this study is defined as the number of roads that the road of interest connects to. Due to a great number of edges (6,140,687 in total) and nodes in the national transportation network, the whole network is divided into subnetworks by States to save computation time.

3.1.2.1 Edge betweenness and AADT

Results in table 3-3 show that for some states, the association between edge betweenness and AADT is prominent for nine states with a Spearman coefficient of around 0.5 or more, including Maine, New York, New Jersey, Oklahoma, West Virginia, North Dakota, Mississippi, Arizona, and Delaware. Figure 3-1 is the comparison between edge betweenness and AADT for the State of Marine, which shows a quite similar distribution. For the States listed in table 3-3, the correlation between edge betweennness and AADT is obvious while for other states such relationship is not. There are some reasons. Firstly, the centrality measurement is very sensitive to the completeness of network. An incomplete network hurts the analysis. The transportation network used in this study is the shape file from HPMS, which is

not complete especially for local roads. Further improvement needs a more complete network. Secondly, the traffic volume is a reflection of the real world, which involves directional flows. The network used for analysis is the central line without directions. Further improvement can be adding the directions. Thirdly, for the purpose of saving computation time, the whole network is divided into 52 subnetworks for each state, which causes loss of accuracy since interstate connections are not included in network analysis. Moreover, future improvement could compute the edge betweenness based on OD matrix instead of all node-pairs in the network.

Table 0-3 Association strength between edge betweenness and AADT by States

State	rs	r	State	rs	r	State	rs	r
Maine	0.573	0.247	Mississippi	0.493	0.107	Connecticut	0.366	0.184
New York	0.543	0.026	Arizona	0.490	0.211	Maryland	0.336	0.010
New Jersey	0.521	0.220	Delaware	0.469	0.148	Illinois	0.314	0.043
Oklahoma	0.512	0.062	South Carolina	0.432	0.375	Pennsylvania	0.314	0.240
West Virginia	0.508	0.234	Wyoming	0.390	0.122	Nevada	-0.368	-0.176
North Dakota	0.499	0.269	District of Columbia	0.385	0.161	Others	< 0.3	< 0.3

Note: the analysis is based on the whole road network for each State

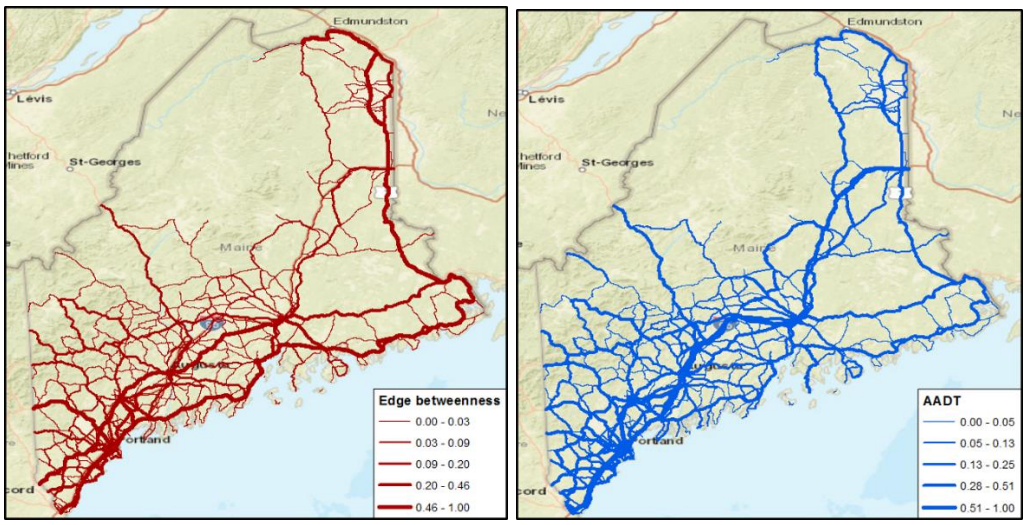

Figure 0-1 Edge betweenness and AADT of Marine

3.1.2.2 Degree centrality and AADT

As presented in table 3-4, the degree centrality of thirty-six states shows a signficant correlation with AADT with a Spearman coefficient (rs) greater than 0.5. And ten States' Spearman coefficients are greater than 0.4. The degree centrality is worth considering for AADT estimation based on the results in table 3-4. Taking Missouri as an example, the degree centrality and AADT of it are plotted in figure 3-2. A similar pattern can be clearly detected. Among these two centrality measures, only degree centrality is selected for the modeling section.

Table 0-4 Association strength between the degree centrality and AADT by States

State	rs	r	State	rs	r	State	rs	r
Missouri	0.722	0.473	West Virginia	0.615	0.557	California	0.515	0.347
Tennessee	0.705	0.457	Minnesota	0.603	0.310	Massacusetts	0.507	0.290
Kentucky	0.691	0.492	RhodeIsland	0.601	0.390	New Hampshire	0.499	0.457
Mississippi	0.687	0.513	Iowa	0.598	0.368	Maryland	0.489	0.298
Arkanas	0.682	0.488	South Carolina	0.590	0.447	Arizona	0.489	0.349
Virginia	0.665	0.456	North Dakota	0.577	0.460	Louisiana	0.488	0.348
Indiana	0.663	0.383	Washington	0.570	0.301	Colorado	0.465	0.266
New Mexico	0.657	0.364	Maine	0.566	0.444	Wyoming	0.460	0.385

Texas	0.657	0.402	Utah	0.558	0.317	Vermont	0.448	0.346
Hawaii	0.643	0.440	Connecticut	0.556	0.363	Florida	0.426	0.299
Deleware	0.643	0.481	New Jersey	0.547	0.329	Oregon	0.417	0.236
Pennsylvania	0.641	0.506	Idaho	0.543	0.330	Illinois	0.416	0.200
Wisconsin	0.637	0.381	South Dakota	0.538	0.359	Michigan	0.399	0.228
District	0.637	0.301	Alabama	0.538	0.356	Georgia	0.360	0.175
North Carolina	0.636	0.459	Nebraska	0.531	0.316	Montana	0.302	0.187
New York	0.619	0.397	Kansas	0.527	0.373	PA_NHS	NA	NA
Oklahoma	0.617	0.474	Ohio	0.527	0.325	Others	<0.2	<0.2

Note: 1. the analysis is based on the whole road network for each State,
2. There's no result for PA_NHS because of data problem.

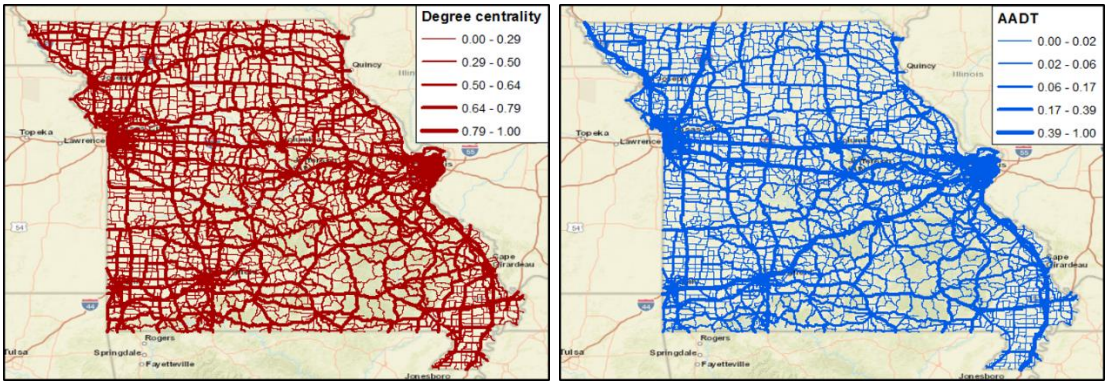

Figure 0-2 Degree centrality and AADT of Missouri

3.1.3 Spatial autocorrelation analysis

The traffic volume on one road segment is probably correlated with that on the neighboring segments. A Moran's I test on the dependent variable AADT confirms that spatial autocorrelation statistically does exist in the road network. Even though such spatial dependence issue cannot undermine the performance of machine learning algorithms because of the non-parametric nature of machine learning, taking care of this issue could help improve estimation accuracy. Therefore, several features regarding surrounding traffic features are investigated. It turns out that Naadt67 and Naadt5 are two neighboring factors that significantly correlate with AADT.

3.1.3.1 Global Moran's I test

The Moran's Index is developed by Moran in 1950. It is a generally used tool for testing the spatial autocorrelation. Given the AADT values of roadways and their geographical locations, the Moran's I test statistic is applied. Moran scatter plot (Figure 3-3) below shows the linear correlation between the AADT on one road (x-axis) and the AADTs on the surrounding roads (y-axis). The lagged-AADT (y-axis) here are calculated using K-Nearest Neighbors method with K = 10 and the neighbors' AADT are weighted by the inversed distance with power = 1. In other words, it is assumed that the AADT on one road is influenced by AADTs on at most 10 nearest neighboring roads and the influence of a neighbor decreases as the distance increases. The Moran's Index is 0.623 (the slope of the regression line), which is the strength of spatial autocorrelation. The points are concentrated on the upper-right quadrant indicating a positive spatial autocorrelation - a clustering of similar values. The Moran's I test statistic is quite significant because of a very small p-value of 2.2e-16. It is fairly convincing to reject the null hypothesis that the spatial distribution of AADTs is purely random. Under null hypothesis, i.e. without any spatial autocorrelation, the expected value of Moran's I is -7.5e-05 while the actual value is 0.623 and the variance is 1.4e-05. In summary, the positive spatial autocorrelation (clustering of similar values) among AADT values is not only strong (strength is 0.623) but also highly significant (99% confidence level).

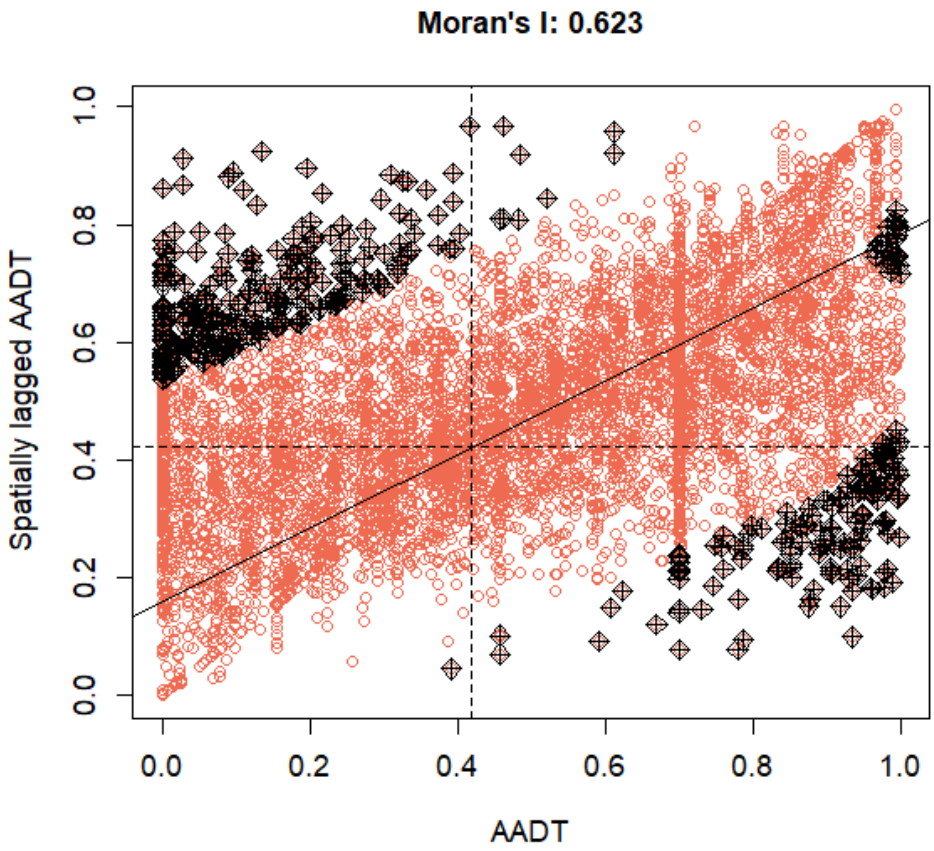

Figure 0-3 Moran scatter plot

3.1.3.2 Exploration of surrounding traffic features

This section investigates various surrounding features to consider spatial influence. Results in table 3-5 shows that the closest observed AADT of local roads has a stronger influence than that of higher level roads. The distance to the closest major collector also positively influences AADT. These two features are selected as the model inputs for AADT estimation.

Table 0-5 Correlation strength between surrounding traffic features with AADT

Feature	Definition	r	rs
Naadt67	The AADT value of the nearest local road in all area or minor collector in urban area only (FSystem = 6, 7)	0.840	0.905
Naadt	The AADT of the nearest road with (FSystem < 6)	0.153	0.202
Ndist5	The Euclidean distance to the nearest major collector (FSystem = 5)	0.094	0.315
Num0fRd	Number of road sections within a distance of 2 mile	-0.227	-0.236
Maxaadt	The max AADT value of nearby roads within a distance of 2 mile	0.279	0.042
Avgaadt	The average AADT value of nearby roads within a distance of 2 mile	0.039	-0.127

Note: sample size is 13335 for Naadt67, Ndist5, and Num0fRd2. For others, sample size is 3888 or 3397.

3.2 Feature engineering

Based on previous sections regarding on-road and off-road features, network centrality, and spatial dependence analysis, 12 candidate variables are selected from 87 features, which are summarized in table 3-6 along with Pearson correlation coefficient and Spearman coefficient and their definitions.

Table 0-6 Candidate features for modeling inputs

Feature	Definition	rs	r
Naadt67	The AADT value of the nearest local road in all area or minor collector in urban area only (FSystem = 6, 7)	0.91	0.84
D1C8Ret10	The gross retail employment density, jobs/acre, under 8-tier classification on unprotected land	0.60	0.41
UrbanCode	1 for rural sections, 2 for small urban sections, and 3 for urban sections	0.46	0.37
FSystem	6 for minor collectors in rural area and 7 for locals in both rural and urban areas	-0.53	-0.32
Degree	Number of road segments that are connected with	-0.44	-0.28
D5ae	Working age population within 45 minutes auto travel time, network travel time-decay weighted	0.44	0.14
AutoOwn2P	Number of households in CBG that own two or more automobiles	0.42	0.42
D5ar	Number of jobs within 45 minutes auto travel time, network travel time-decay weighted	0.42	0.15
ThroughLane	Number of through lanes	0.35	0.41
D3a	Total road network density	0.40	0.36
E5ENT10	Entertainment jobs within a 5-tier employment classification scheme	0.18	0.35
Ndist5	The Euclidean distance to the nearest major collector (FSystem = 5)	0.32	0.09

Note: rs: Spearman coefficient,
r: Pearson coefficient

Feature engineering paves the way for implementing machine-learning algorithms. It covers a series of techniques such as addressing missing and abnormal values, transforming data by logarithm or powers, normalizing or standardizing, one-hot encoding of nominal variable, etc. Following the feature selection process, all selected features from either HPMS data or SLD are integrated together, which yields a cross-sectional dataset consisting of 10490 observations without missing values. Then the Interquartile Range Rule, i.e. (Q1-1.5*IQR, Q3+1.5*IQR), is used to remove the outliers of AADT values and 464 observations are removed. Data transformation is not included to retain the interpretability of the results. Although scaling is not necessary for random forest modeling, it is necessary for the artificial neural network.

For comparison purpose, all features are scaled using min-max normalization and then increased by 0.1 to avoid zero values.

Chapter 4: Spatial autoregressive model: a benchmark

4.1 Ordinary least squares (OLS) model

4.1.1 Model specification

Beginning with an OLS model, the relationship between AADT and the twelve selected predictors is investigated. The two dummy variables, i.e. FSystem and UrbanCode, are treated as numerous variables in this linear model to keep the ordering information. Modeling results are summarized in table 4-1. For a linear regression model, all selected predictors are statistically significant except for AutoOwn2P (number of households in CBG that own two or more automobiles). D1C8Ret10, D3a, E5ENT10, ThroughLane, FSystem, UrbanCode, Naadty67 have quite significant coefficients at 99% confidence level. The other four predictors have significant coefficients at 95% confidence level.

Table 4-1 Summary of linear regression results

Predictors	**Estimates**	**CI**	**p**
(Intercept)	4469.48	3623.04 ~ 5315.92	< **0.001**
D1C8Ret10	919.26	672.92 ~ 1165.59	< **0.001**
D3a	27.76	13.93 ~ 41.59	< **0.001**
AutoOwn2P	0.00	-0.01 ~ 0.02	0.766
E5ENT10	-0.06	-0.06 ~ -0.05	< **0.001**
D5ar	0.00	0.00 ~ 0.00	**0.019**
D5ae	-0.00	-0.00 ~ -0.00	**0.041**
ThroughLane	200.28	143.63 ~ 256.92	< **0.001**
FSystem	-746.91	-890.44 ~ -603.38	< **0.001**
UrbanCode	165.28	94.23 ~ 236.32	< **0.001**
Naadt67	0.60	0.58 ~ 0.61	< **0.001**
Ndist5	-645.93	-1168.59 ~ -123.28	**0.015**
Degree	35.79	6.67 ~ 64.91	**0.016**
R2/R2 adjusted	0.669/0.667		

4.1.2 Multicollinearity and variance inflation factor (VIF)

To reduce the variable redundancy and improve the performance of OLS and spatial autoregressive model, the multicollinearity among the twelve selected variables is tested by VIF. Table 4-2 summarizes the results. D5ar and D5ae are two concerning predictors with very high VIF values. For the improved OLS and spatial autoregressive model, these two variables are excluded while they are kept when applying machine-learning algorithms.

Table 4-2 Multicollinearity test results with VIF

Predictors	VIF	Predictors	VIF
D1C8Ret10	3.09	ThroughLane	1.16
D3a	6.12	FSystem	4.12
AutoOwn2P	4.78	UrbanCode	2.11
E5ENT10	1.53	Naadt67	1.56
D5ar	49.03	Ndist5	1.34
D5ae	45.66	Degree	3.66

4.1.3 Improved OLS model

After removing the two predictors, i.e. D5ar and D5ae, and an insignificant predictor, i.e. AutoOwn2P, a new linear regression model is built, results of which are shown in table 4-3. With the improved regression model, R2 remains around 0.67.

Table 4-3 Summary of the improved linear regression results

Predictors	Estimates	CI	p
(Intercept)	4475.06	3630.45 ~ 5319.67	**< 0.001**
D1C8Ret10	1085.92	871.29 ~ 1300.55	**< 0.001**
D3a	28.96	19.43 ~ 38.49	**< 0.001**
E5ENT10	-0.06	-0.06 ~ -0.05	**< 0.001**
ThroughLane	194.59	138.27 ~ 250.91	**< 0.001**
FSystem	-745.39	-888.31 ~ -602.47	**< 0.001**
UrbanCode	160.61	89.75 ~ 231.46	**< 0.001**
Naadt67	0.60	0.58 ~ 0.61	**< 0.001**
Ndist5	-730.00	-1245.06 ~ -214.94	**0.005**

Degree	35.66	6.66 ~ 64.66	**0.016**
R2/R2 adjusted	0.667/0.667		

4.2 Spatial autoregressive model

It is intuitive that ADDT values are spatially autocorrelated. A Moran's I test on the dependent variable AADT confirms spatial autocorrelation and the need to proceed with a spatial autoregressive model. Therefore, spatial autoregressive models are developed to deal with the existing spatial dependence issue. There are two common types of spatial dependence, i.e. spatial lag and spatial error.

4.2.1 Spatial error model

When the error terms of the population regression equation are auto-correlated, the independence assumption of the error terms is violated and thus the spatial error model could be used to deal with the unknown random factors (nuisance spatial dependence). The spatial error model is defined with the equation:

$$y = \lambda W\varepsilon + X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

where: λ= nuisance parameter, W= spatial weights matrix, ε = error term, X = matrix of predictor variables, β = the coefficient matrix for the predictor variables.

4.2.2 Spatial lag model

Another approach is the spatial lag model (SLM) that can be used when the dependent variable in place i is influenced by independent variables in both place i and other places. This kind of spatial dependence not only violates the independence assumption of error terms, but also undermines the assumption of independent observations. Three types of spatial interactions must be addressed in a spatial lag

model: interaction effects among individual road segments (endogenous effects), exogenous group characteristics (contextual effects), and observed or unobserved characteristics that road segments have in common (correlated effects). SLM aims to handle spatially-lagged dependent variable by weighting neighboring values. SLM is specified by the equation:

$$y = \rho Wy + X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

where: ρ = spatial autoregressive coefficient, W = spatial weights matrix, X = the matrix of predictor variables, β = the coefficient matrix for the predictor variables, ε = the random error which follows a normal distribution.

4.2.3 Spatial autocorrelation diagnostics

Spatial weights matrix is generated by k-nearest neighbors method. Figure 4-1 shows the procedure for selecting the type of spatial autoregressive model. As it indicates, in the first step, an ordinary least square model is developed, followed by a spatial dependence diagnostic of OLS's error terms. These diagnostics calculate the Lagrange Multiplier of Error (LM-Error), Lagrange Multiplier of Lag (LM-Lag), Robust Lagrange Multiplier of Error (Robust LM-Error), and Robust Lagrange Multiplier of Lag (Robust LM-Lag). Based on the values of these diagnostics, the most appropriate model is selected.

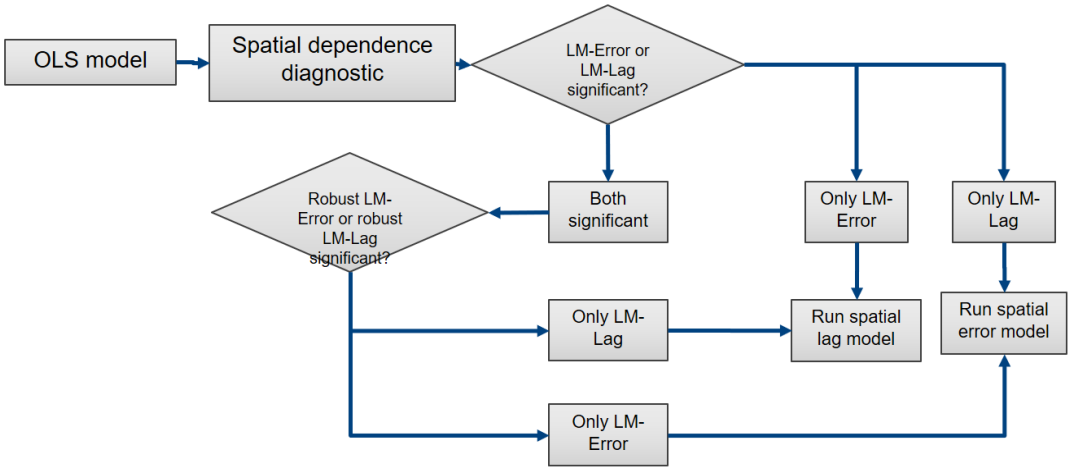

Figure 4-1 Procedure for selecting a spatial autoregressive model

Five tests are performed to assess the spatial dependence and detect the type of spatial autocorrelation (table 4-4). First, Moran's I test shows a significant positive spatial autocorrelation, i.e. roads with similar AADT values tend to form a cluster. Results of LM tests for a missing spatially-lagged dependent variable and dependent errors show significant LM-Lag and LM-Error values, indicating the presence of both spatial error and spatial lag. Robust LM-Lag indicates spatial lag dependence conditioned on missing spatial errors. Robust LM-Error indicates spatial error conditioned on the presence of spatial lagged dependent variable. Both LM and Robust LM tests are significant. Since LM-Lag indicator (p-value < 0.001) is more significant than LM-Error indicator (p-value = 0.03), spatial lag model is selected.

Table 4-4 Results of spatial autocorrelation diagnostics

Tests	Value	p
Moran's I	0.625	< **0.001**
Lagrange Multiplier (lag)	1064.15	< **0.001**
Robust LM (lag)	1461.80	< **0.001**
Lagrange Multiplier (error)	4.66	**0.03083**
Robust LM (error)	402.32	< **0.001**

4.2.4 Maximum Likelihood Estimation of the Spatial Lag Model

As shown in table 4-5, the ρ value (spatial autoregressive coefficient of spatial lag model) is quite significant with a p-value of an asymptotic t-test smaller than 2.22e-16. The likelihood ratio test on ρ is also quite significant with a p-value smaller than 2.22e-16. Compared with the OLS model, the AIC of spatial lag model is 138470, which is better than that of the OLS, i.e. 140650. The spatial lag model performs better than the OLS model.

Table 4-5 Summary of spatial lag model

Predictors	**Estimate**	**Std. Error**	**p**
(Intercept)	1537	374.33	< 0.001
D1C8Ret10	319	97.82	0.001
D3a	10.46	4.21	0.013
E5ENT10	-0.02	-0.003	< 0.001
ThroughLane	170.16	23.96	< 0.001
FSystem	-321.81	63.10	< 0.001
UrbanCode	20.73	30.89	0.5
Naadt67	0.33	0.0072	< 0.001
Ndist5	-609.35	226.32	0.007
Degree	49.43	12.61	< 0.001
Rho: 0.55463 LR test value: 2187.8, p-value: < 2.22e-16 Asymptotic standard error: 0.010312, z-value: 53.787, p-value: < 2.22e-16 AIC: 138470, (AIC for lm: 140650)			

Chapter5: Implementation of Machine learning algorithms

5.1 Ensembling artificial neural networks

5.1.1 Architecture design

The ANN model imitates a brain's biological neural network. It can learn from training process and no specific rules are needed for a learning task. A neural network consists of neurons (nodes) and edges (links) (Figure 5-1). Each neuron has a value and each link is assigned by a weight; computational process happens on links from input layer to hidden layer and then to output layer by weighting the values of the previous layer. The output is an aggregated sum of values through numerous non-linear transferring processes from layer to layer. Then there are two types of neural networks: feedforward and feedback. Feedforward neural network is nonrecurrent without any cycling while feedback neural network adjusts the weights based on the output's bias from target. For this study, a three-layer-based neural network work is applied including an input layer, a hidden layer, and an output layer. The Levenberg-Marquardt (LM) algorithm, also known as the damped least-squares method, is used to tune the weights. It works specifically with loss functions in the form of a sum of squared errors. The learning rate of LM algorithm is 0.1. Studies show that adding more layers seldom significantly improve the performance; one hidden layer is sufficient in most circumstances. Number of neurons in the hidden layer is nine based on a rule-of-thumb.

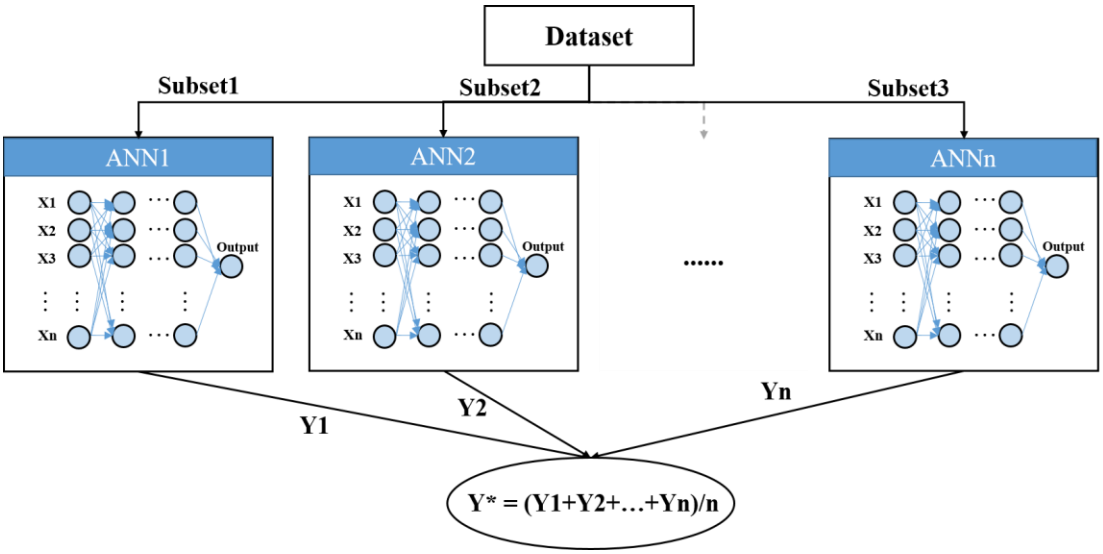

Figure 5-1 Structure of ensembling artificial neural networks

In order to improve the model robustness, an ensemble of ten neural networks are simultaneously trained. On the one hand, the overfitting problem can be detected through performance variation among the ten neural networks. On the other hand, averaging the estimations from multiple ANNs reduces the risk of overfitting and random disturbances. Each neural network is trained by a random sample of size 5614 from the original training dataset of size 8020, i.e. 70% sampling rate. These ten ANN models are applied to the same validation set to calculate the accuracy. The average of all estimations from the ten ANNs is the final estimation.

5.1.2 Training results and variable importance measure

The ensembling artificial neural networks consist of ten independent neural networks, each of which is trained by a random sample from the training dataset. The black links denote positive coefficients and the gray links represent negative coefficients. The strength of the linkage is presented by the width of the links. After training, all ten neural networks are shown in figure 5-2. The importance of each

variables is shown through the color depth. The greener the node is, the more important the input feature is.

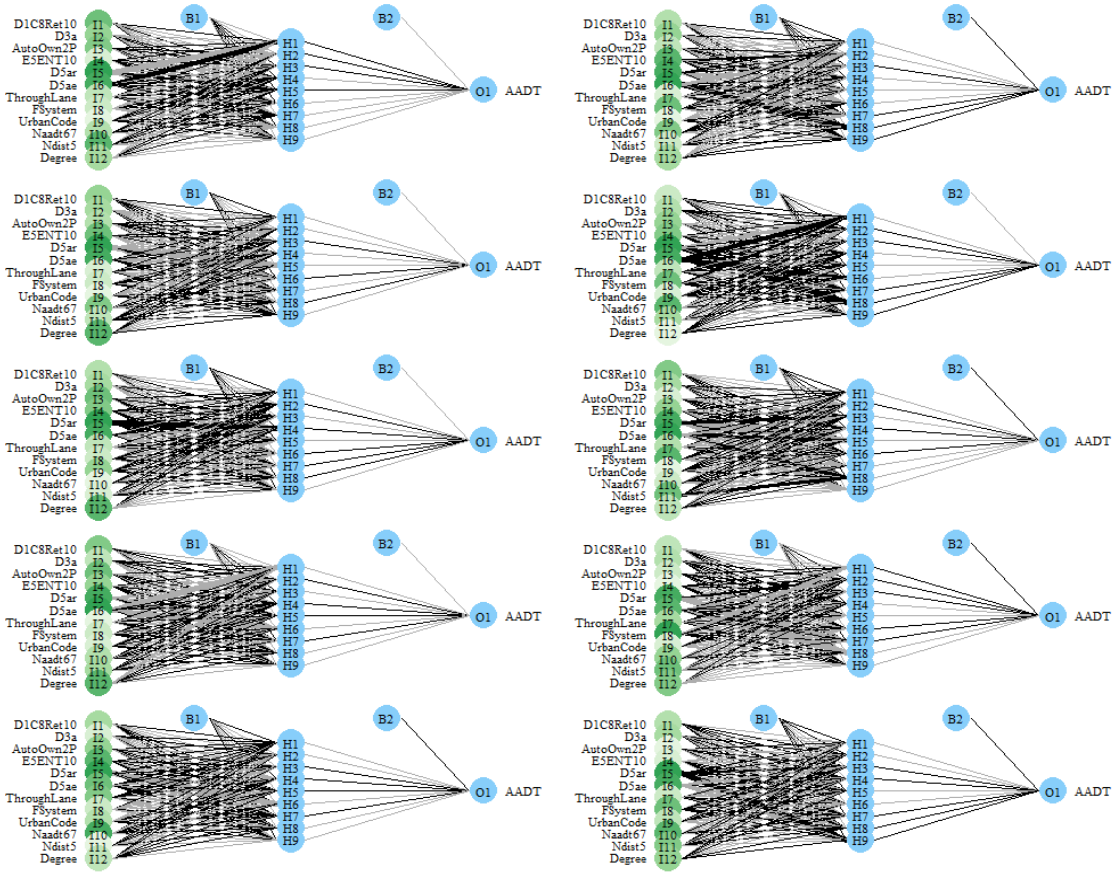

Figure 5-2 Architecture of artificial neural networks after training

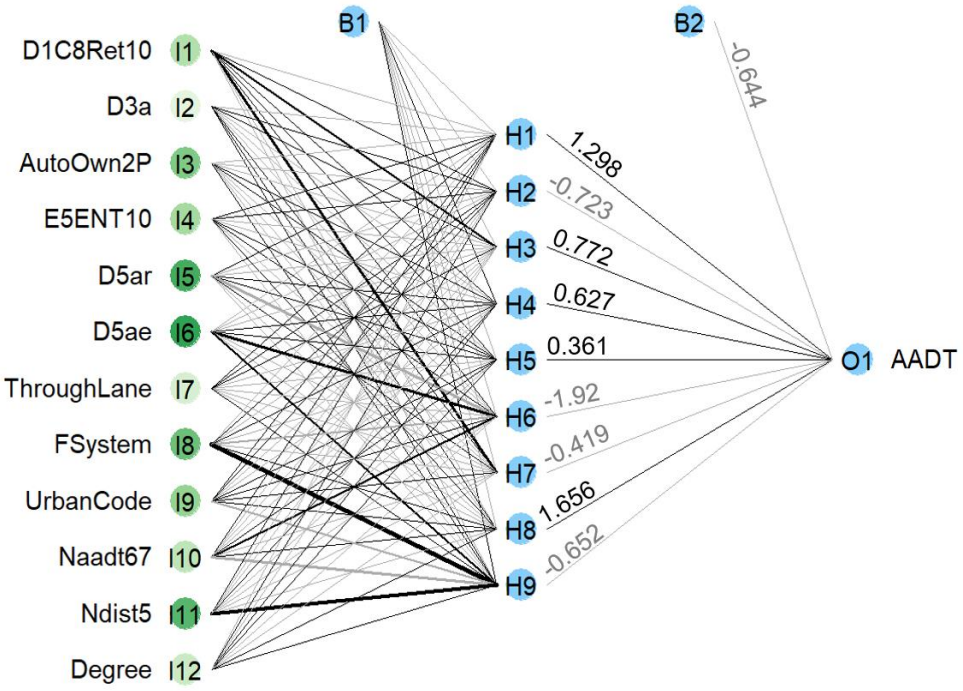

Figure 5-3 An example of trained neural network and importance rank of inputs

Figure 5-4 depicts the importance rank of input features of each trained ANN in the order of average importance rank. The importance of an input features is measured by the sum of all weights connecting the given input feature and the output AADT (Garson, 1991; Goh, 1995). The five most important input features for ANNs are D5ar (number of jobs within 45 minutes auto travel time), D5ae (working age population within 45 minutes auto travel time), FSystem (functional class), Naadt67 (the AADT of nearest NFAS road), and Ndist5 (the distance to the nearest major collector).

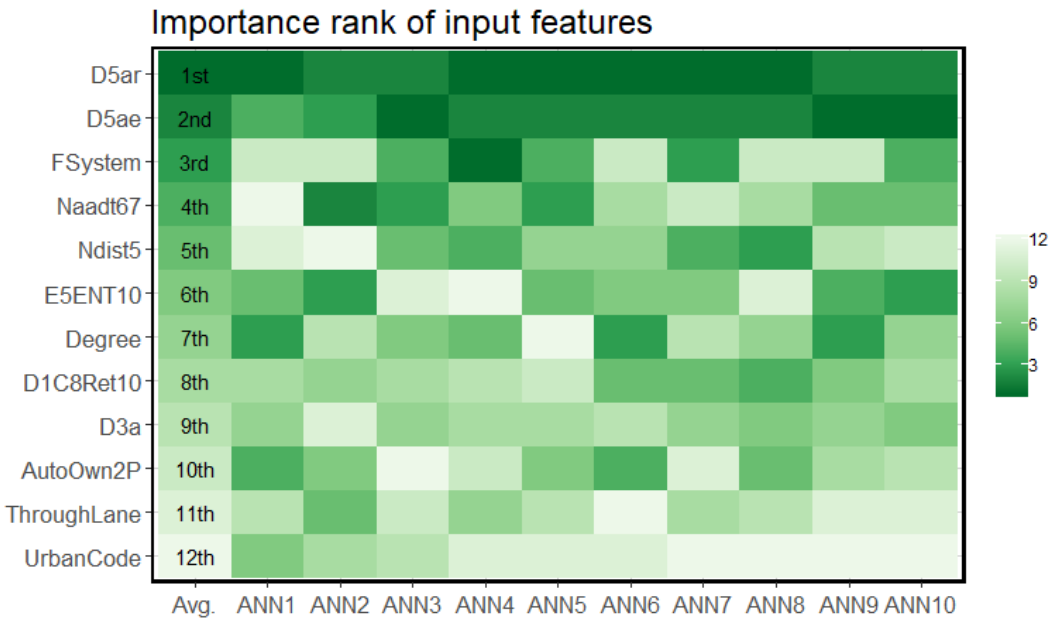

Figure 5-4 Importance rank of input features of ANN

5.1.3 Accuracy measures

Five measures are employed to evaluate the accuracy level: Mean Squared Error (MSE), R Squared Value (RSQ), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Both the accuracy of individual ANNs and the combined ANNs is plotted in figure 5-5. The difference among the ten individual neural networks is minimal. Even though some particular neural networks behave better than the combined network, the main benefit of assembling the ANNs is to improve the model's stability and robustness.

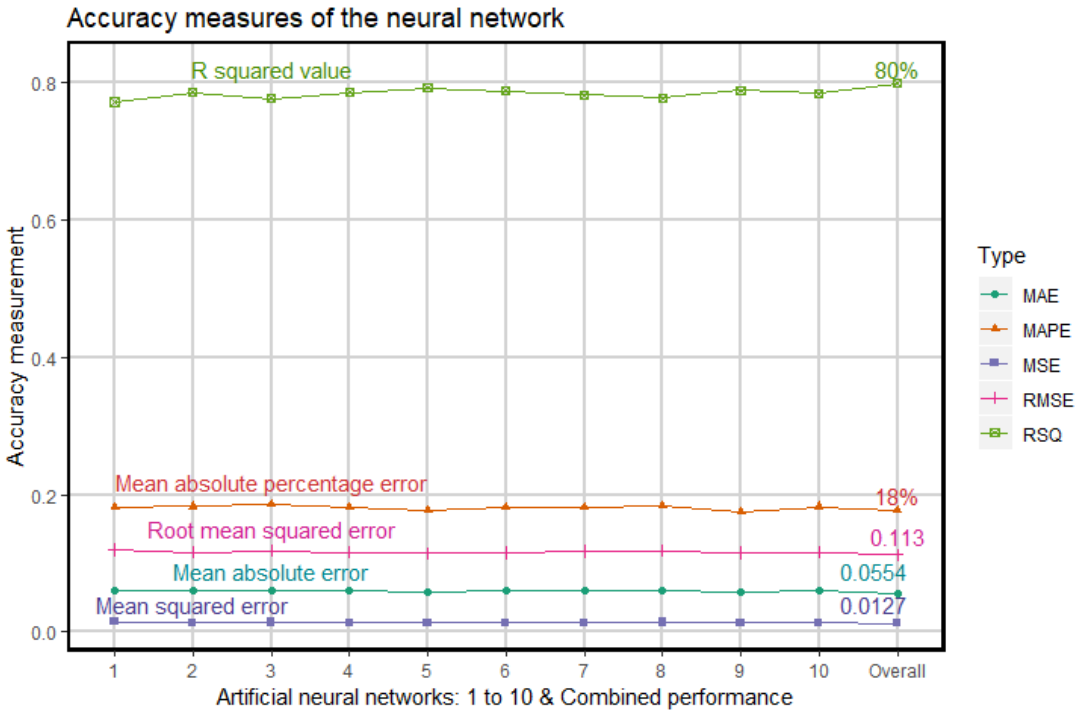

Figure 5-5 Performance of ensembling ANNs

5.2 Random forest

Random forest regression gains increasing popularity in prediction or estimation studies by virtue of many merits. Not only it usually achieves a high accuracy level but also it has good interpretability. Moreover, overfitting issue does not bother it because of its robust architecture derived from ensemble learning theory. Recently, various methods have been developed to demystify Random Forest such as variable importance measures and partial dependence plots (PDPs).

5.2.1 Architecture design

As depicted in figure 5-6, a random forest consists of a predefined number of decision trees – ntree, the magnitude of which is usually in hundreds. Each tree randomly extracts a portion of the original dataset in a way of bootstrap resampling

(sampling with replacement). Then each tree independently makes its own estimation using randomly selected features. The number of features that each node can select is controlled by the second model parameter – mtry. Finally, a voting process takes the estimations from all trees into account and makes the final decision usually by unweighted averaging, which is a bagging process that ensembles hundreds of tree models. Random forest benefits from this bagging feature to provide a more stable and accurate estimation. Adjusting the two parameters, ntree and mtry, contributes to improving the predictive performance. Since the parameter mtry influences the accuracy of each individual tree and simultaneously determines the correlation among the trees in an opposite direction, the model is more sensitive to the mtry value.

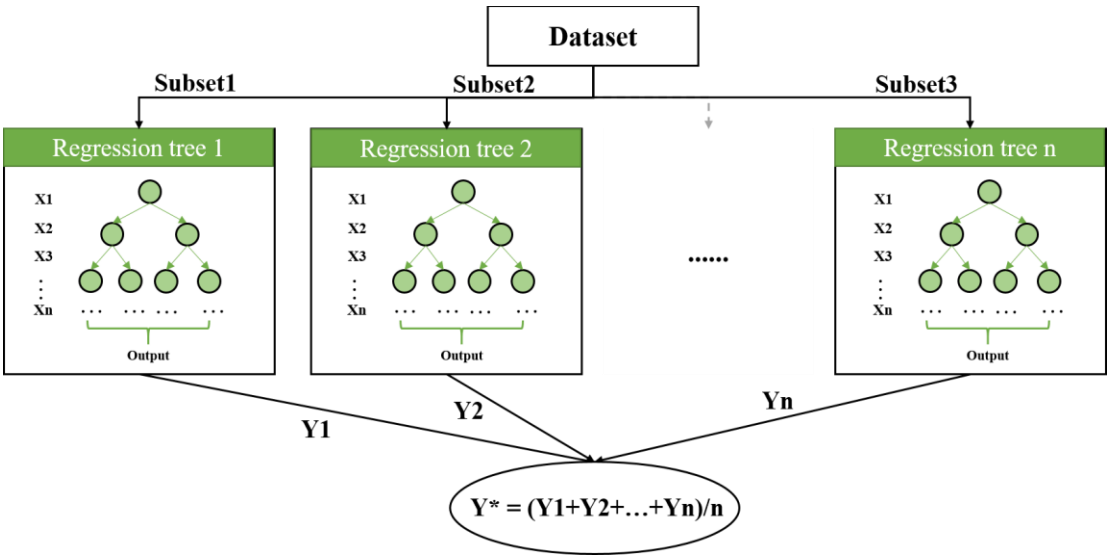

Figure 5-6 Architecture of random forest

5.2.2 Training results

The whole dataset is divided into training and testing part with a ratio of 80% to 20%. Multiple combinations of ntree and mtry are tested and theirs prediction

performances are compared as shown in figure 5-7. It is obvious that setting mtry as three generates the best result no matter what ntree is. Besides, Mean Squared Residuals (MSE) is the minimal and R Squared value (RSQ) is the maximal when ntree is 500 and mtry equals 3. Thus, this specification of parameters is used for modeling.

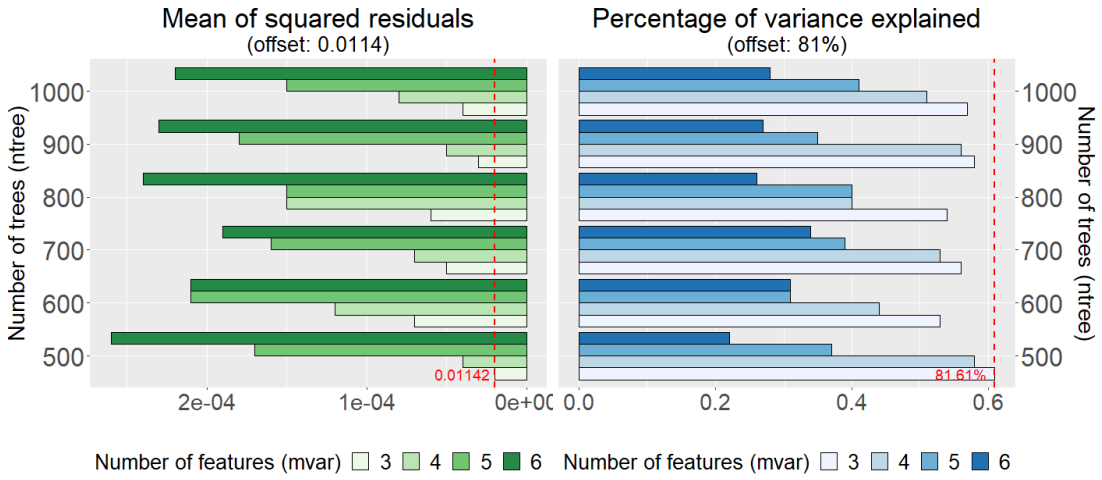

Figure 5-7 Training results of random forest

With three features randomly selected for each node, the random forest is built up on 500 trees through the training dataset. The learning curves in figure 5-8 show how MSE and RSQ change with more trees joining in. When there were 100 trees, the learning curves gradually stabilize to a constant level. After 500 trees are built, MSE decreases to 0.01142 and RSQ gets as high as 0.8161 meaning that 81.61% total variance can be explained. The model achieves a high goodness of fit.

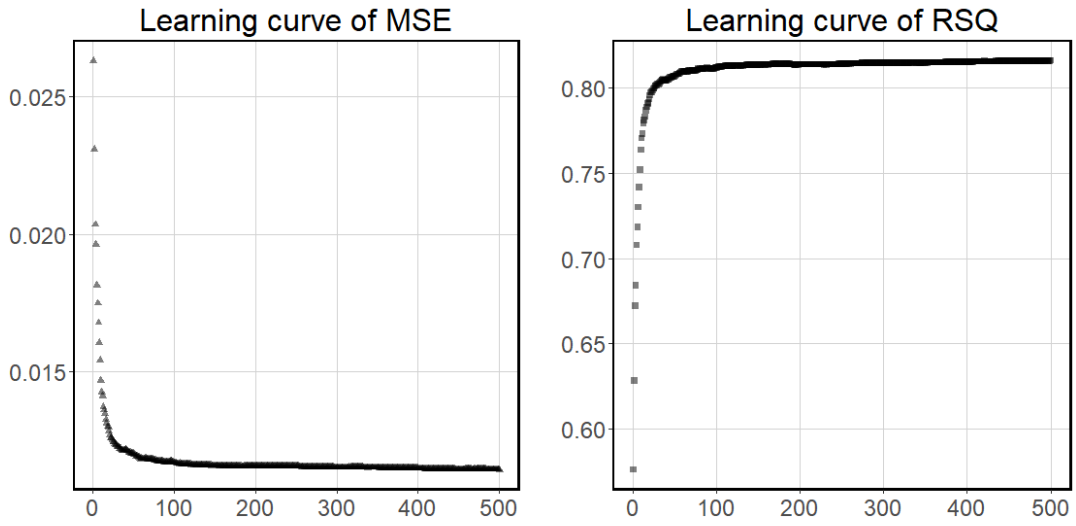

Figure 5-8 Learning curves on MSE and RSQ of the trained random forest

5.2.3 Interactions between predictors and AADT

How each predictor interacts with the response variable AADT needs further investigation. In 2001, Friedman proposed Partial Dependence Plots, which can visualize the marginal effect of a single predictor on the output of a machine-learning model, such as Random forest and Support Vector Machine, while averaging the effects of all other predictors. Along with the changes of a predictor, how is the response variable changing is plotted through PDP. The larger the range that PDP varies over along y-axis, the more influential the predictor is. Besides, various interactions including not just linear correlation are shown from figure 5-9. Among all 12 predictors, only FSystem shows a complete negative effect on AADT, which makes a lot sense because FSystem=6 represents minor collectors in rural area and FSystem=7 is locals with less traffic. For D1C8Ret10, D5ar, and D5ae, they show a strong sensitivity at the very beginning and then they become stabilized. The PDP of UrbanCode shows that the urban code increase from 2 (small urban sections) to 3

(urban sections) brings about a larger increase of AADT compared with the change from 1 (rural sections) to 2 (small urban sections). Other predictors overall show a positive influence on AADT even though some fluctuations occur.

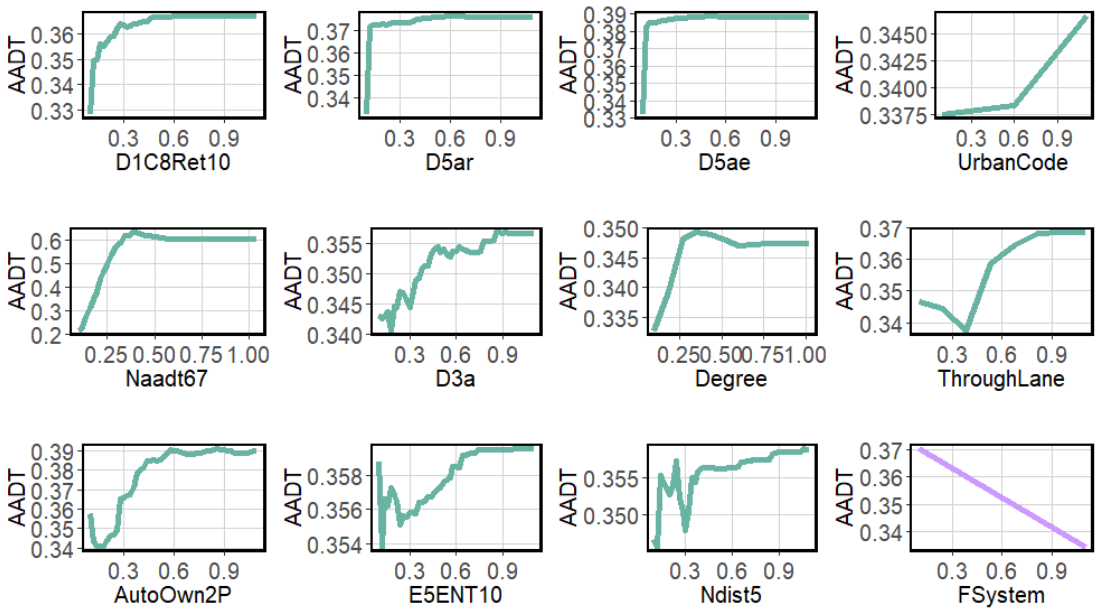

Figure 5-9 Partial dependence plots

5.2.4 Variable importance measures

Multiple methods are utilized to assess the importance of predicting features. First, a widely used measurement, that is percentage increase in mean squared error when permuting a single predictor, is applied. Second, according to the inherent structure of random forest, several methods from various aspects quantify the importance of predictors. Two representative methods, including times of splitting the root node and mean of minimal depth, are used and discussed. Finally, the importance ranks from these three different methods are summed up, through which a list of predicting features in order of priority is given.

5.2.4.1 Permutation-based measure

In a random forest, the contribution rate of each predictor can be measured from each tree. Then all these contributions are averaged and sometimes further normalized with the standard deviation. This yields the final importance score for a predictor. As for random forest regression, one widely used measure of importance is the percentage increase in mean squared error after permuting the predictor of interest. Using this measurement, the importance of all twelve predictors is plotted in figure 5-10. The most important predictor is the AADT value of the nearest local road segment, which is intuitive because of significant spatial dependence as discussed before. Then D1C8Ret10 (the gross retail employment density in number of jobs per acre under 8-tier classification on unprotected land), D3a (total road network density), D5ar (jobs within 45 minutes auto travel time, network travel time-decay weighted) ranks second, third, and forth with an importance measure around 50%.

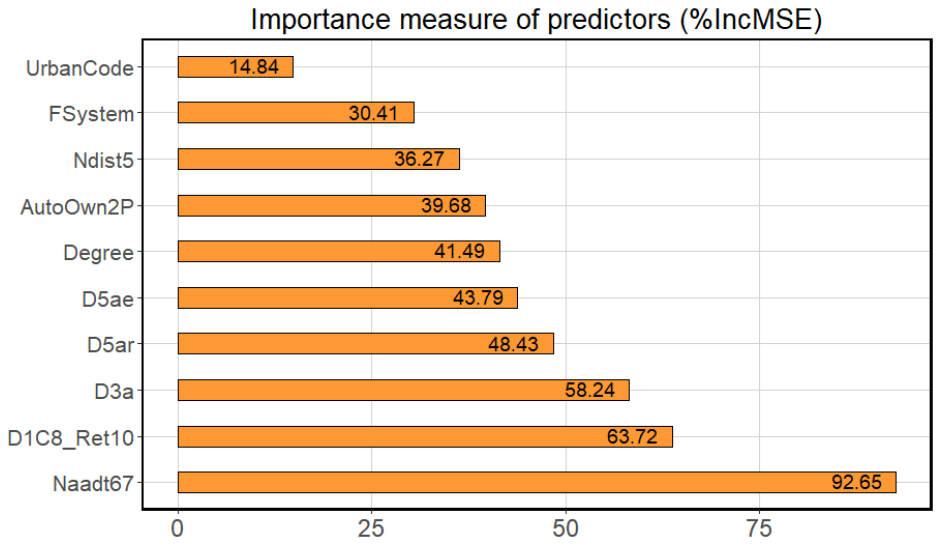

Figure 5-10 Importance measure by %MSE upon permutation

5.2.4.2 Times of splitting the root node

As for a single tree, only the significant variables can appear in the tree. Result shows that each predictor is in all 500 trees, which means all selected features are of significance. The sequence that the predictors occur from top to bottom represents the rank of importance. The first feature that splits the root node provides the most information gain. It is the most significant variable because it outperforms all other predictors regarding subdividing the whole dataset into homogeneous subsets. All 12 predictors are in all 500 trees. However, the splitting feature of the root node varies among the trees as shown from figure 5-11. Naadt67 is the most frequent one to split the root node of 129 trees, followed by AutoOwn2P occurring at the root node of 102 trees. D1C8Ret10 ranks third by splitting 82 times at the top of the tree. Other variables, including D5ae, FSystem, D5ar, UrbanCode, D3a, also become the most important feature for multiple times. Yet, Degree, Ndist5, and ThroughLane just appear several times at the top of the tree and E5ENT10 never acts as the top feature in any of the 500 trees.

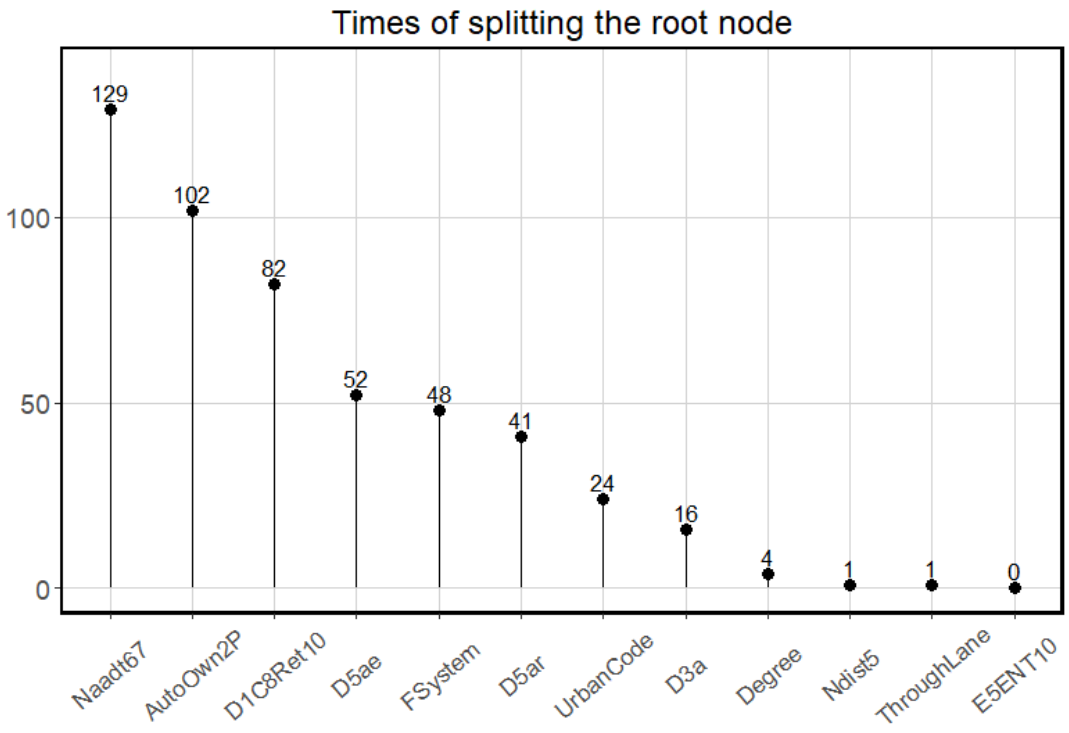

Figure 5-11 Importance measure by times of splitting the root node

5.2.4.3 Distribution of minimal depth and its mean

In addition, the sequence that features present from top to bottom of the tree represents the rank of importance. The closer to the root the feature is, the more important it is. Accordingly, the minimal depth of a given feature is defined as the number of edges along the path that connects the root node with the nearest maximal subtree of the given feature. Figure 5-12 shows the distribution of minimal depth among the 500 trees for each predictor. Naadt67 has a mean minimal depth of 1.3, which is consistent with the fact that it splits the root node most frequently, i.e. 129 times. For the left 371 trees, Naadt67 mainly serves as the secondary or tertiary feature though it is not the primary feature. Besides, it never gets five edges away from the root node. From the analysis above, E5ENT10 never splits a root node. However, its

importance is not negligible considering that it frequently occurs at the shallow part of the tree with a minimal depth of two or three in most cases. In comparison, UrbanCode seems to be the least important. Its closest maximal subtree is usually around five edges away from the root and its mean minimal depth, i.e. 4.18, ranks last. Apart from UrbanCode, FSystem is another feature whose closest position to the root still goes as far as nine edges away. Nevertheless, its mean minimal depth is as low as 2.32 because most of the time its closest maximal subtree starts at the nodes that are two or three edges away from the root. All other nine predictors seem to be important as well because in most trees their first presence is no more than four edges away from the root. In very few cases, they deep into the tree with a minimal depth of six or more. Overall, the twelve selected features performs well among the 500 trees.

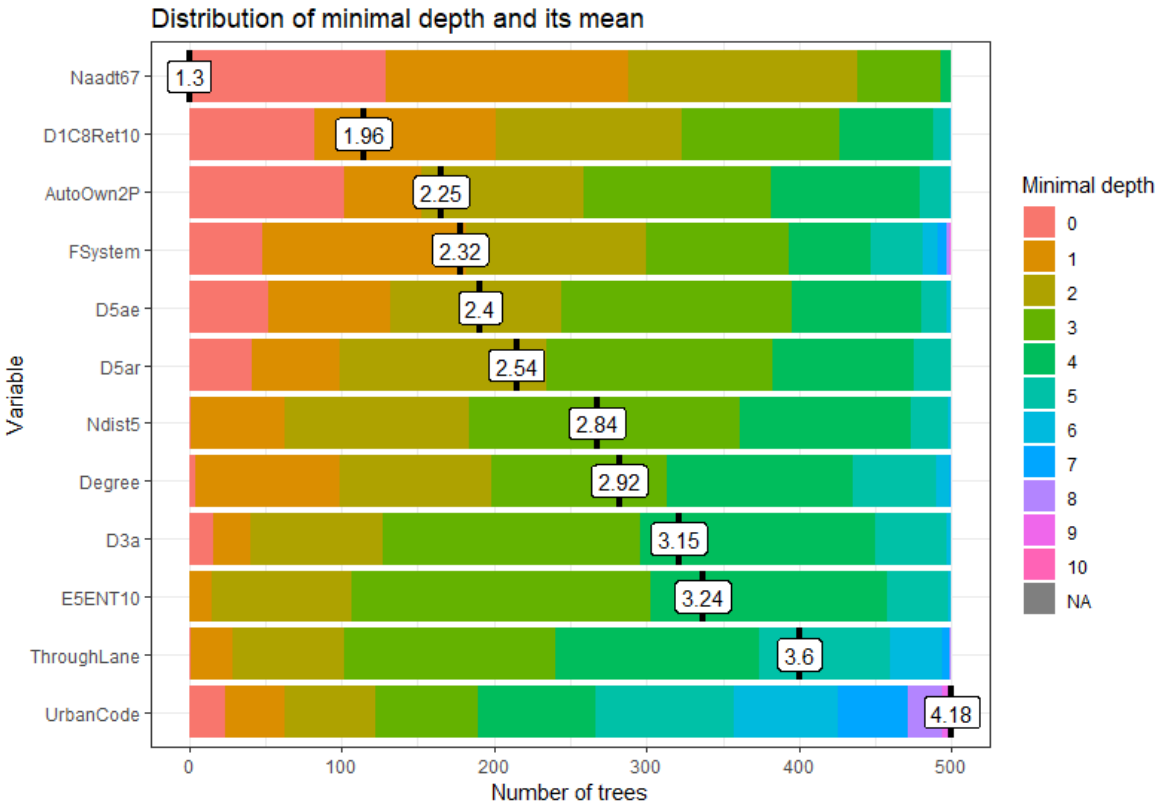

Figure 5-12 Distribution of minimal depth and mean of predicting features

5.2.4.4 Importance rank of predictors

A comprehensive investigation on variable importance is conducted from one accuracy-based method, i.e. percentage increase of MSE conditional on permuting the given predictor, and two tree-based methods, i.e. times of splitting the root node, and mean minimal depth. Considering all these aspects, the twelve predictors are given priority in order of summed importance rank as shown in the table 5-1 below.

Table 5-1 Importance rank of predictors

Rank	Predictor	Rank	Predictor	Rank	Predictor
1st	Naadt67	5th	D5ar	9th	Degree
2nd	D1C8Ret10	6th	FSystem	10th	E5ENT10
3rd	AutoOwn2P	7th	D3a	11th	UrbanCode
4th	D5ae	8th	Ndist5	12th	ThroughLane

5.2.5 Interactions among the predictors

Furthermore, the multidimensional analysis of variable importance indicates that all twelve features occur in each of the 500 trees but the roles they play differ from tree to tree. Inspecting the interactions among predictors promotes understanding the relationship between predictors and the response variable AADT. Conditional minimal depth is a proposed measurement for investigating the interactions between predictors. By definition, the conditional minimal depth of a given predictor X1 conditional on another given predictor X2 is the minimal depth of X1 in the closest maximal subtree of X2 minus one. If no node in the X2 maximal subtree splits using X1, then the mean depth of maximal subtrees of X2 in the forest is set as the conditional minimal depth of X1 conditional on X2. Features, such as Naadt67, D1C8Ret10, and AutoOwn2P, frequently split the root node and dominate in the upper part of the tree. Setting these features as the conditioning variables contributes to a better knowledge of inner interactions within the forest. Six conditioning variables were selected by the order of priority: Naadt67, D1C8Ret10, AutoOwn2P, D5ae, D5ar, and FSystem. There are 72 interactions in total with twelve predictors conditioning on the six selected predictors. 36 most frequent interactions are plotted in figure 5-13 from left to right. One interaction of X1 conditioning on X2 is denoted as X1|X2. In general, most variables become more important when conditioning on some other variables. There are nine

non-conditioning variables, four of which are not the selected conditioning variables: E5ENT10, D3a, Ndist5 and Degree. Even though they are relatively unimportant for the whole trees, they become important when conditioning on other variables. E5ENT10 and D3a are two evident examples. The unconditional mean minimal depth of E5ENT10 is 3.24. However, when conditioning on D1C8Ret10, its mean minimal depth is only 1.82. This means that E5ENT10 becomes quite important after D1C8Ret10 splits a node. Besides, such conditional importance varies with different conditioning variables. E5ENT10 is less important when conditioning on AutoOwn2P with a mean minimal depth of 2.57. The red line presents the minimum of conditional mean minimal depth across all interactions. The interaction of Naadt67 | Naadt67 has a mean minimal depth of 0.9, which means that after Naadt67 splits a node it is followed by Naadt67 once again in many instances. Among the six conditioning variables, only FSystem does not act as a non-conditioning variable, suggesting that it is more important without conditioning on others. Yet, the interactions with FSystem being the conditioning variable are very frequent. As for D5ae, only D3a becomes more important when conditional on it while D1C8Ret10 and D5ae becomes less important.

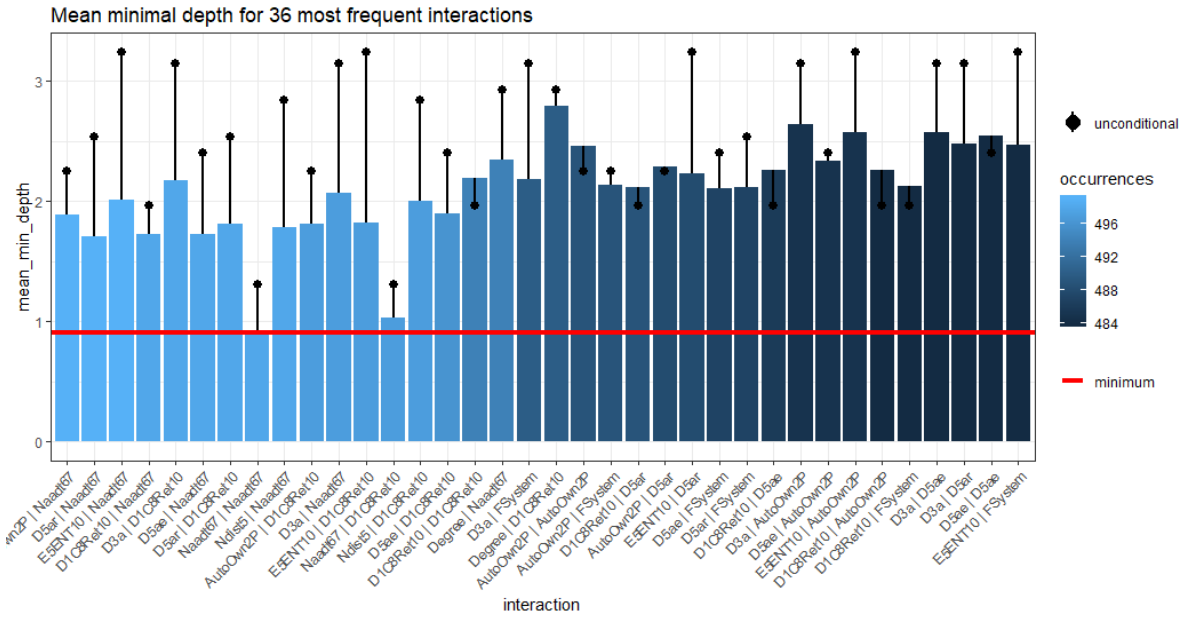

Figure 5-13 Interactions among predictors

5.3 Validation

For comparison purpose, the spatial lag model is included as a benchmark. Five accuracy measures in total are calculated for OLS, SLM, ANN, and RF. The estimated AADT values are transformed back to the real value. Results in table 5-2 show that Random Forest performs the best in each accuracy measure, which is followed by the Artificial Neural Network. Machine learning algorithms produce noticeably better AADT estimations than OLS and SLM in terms of all accuracy measures. Although SLM is excellent for its handling spatial dependence issue, the estimation result of it still shows a very high mean absolute percentage error (MAPE), i.e. 1.13 and it provides limited improvement when compared with OLS. There is only a 5% decrease of MSE, an 8% increase of RSQ, and a 2% increase of RMSE. When comparing the two machine learning algorithms with SLM, the estimation results are notably improved. The estimation result of RF shows a 57% decrease of MAPE, a 37%

decrease of MAE, a 29% decrease of MSE, a 16% decrease of RMSE, and a 10% increase of RSQ. The estimation result of ANN shows a 48% decrease of MAPE, a 28% decrease of MAE, a 20% decrease of MSE, an 11% decrease of RMSE, and an 8% increase of RSQ. These performance measures demonstrate that Random Forest and Artificial Neural Network make better estimations than Spatial Lag Model and Ordinary Least Squares model and Random Forest outperforms Artificial Neural Network in all accuracy measures.

Table 5-2 Accuracy measures of RM, ANN, SLM, and OLS

Model	**MSE =** $\frac{\sum(Y-Y^*)^2}{n}$	**RSQ =** $1-\frac{\sum(Y-Y^*)^2}{\sum(Y-\bar{Y})^2}$	**RMSE =** $\sqrt{\frac{1}{n}\sum(Y-Y^*)^2}$	**MAE =** $\frac{\sum\lvert Y-Y^*\rvert}{n}$	**MAPE =** $\frac{\sum\lvert\frac{Y-Y^*}{Y}\rvert}{n}$
RF	1379990 (-29%)	0.81 (+10%)	1174.73 (-16%)	545.80 (-37%)	0.49 (-57%)
ANN	1566888 (-20%)	0.79 (+8%)	1251.75 (-11%)	623.25 (-28%)	0.59 (-48%)
SLM (benchmark)	1957185	0.73	1398.99	866.42	1.13
OLS	2050093 (+5%)	0.67 (-8%)	1431.82 (+2%)	849.25 (-2%)	1.13 (0%)

Chapter 6: Conclusion

This thesis studies the estimation of annual average daily traffic on NFAS roads in USA at national level. Great efforts are made step by step to refine each procedure.

A deep data mining of built-in environment features for predicting inputs is the first major part. Instead of directly applying the variables used by previous studies, which is what most studies usually do, this study analyzes a long list of features (87 in total) and compares their strength of linearity and non-linearity with AADT. These 87 features are from three perspectives: on-road and off-road features, network centrality measures based on social network theory, and neighboring traffic characteristics through Spatial dependence analysis. Specifically, the built-in environment factors analyzed in this study covers demographics, employment, density, land use diversity, urban design, transit service, destination accessibility, network centrality, and influences from neighboring traffic. As of now, this is the most comprehensive study on built-in environment factors in terms of the potential predictive power of estimating AADT. Through relationship analysis, either linear or non-linear correlation, 12 out of 87 features are selected as the modeling inputs based on statistical tests. By referencing the relationship analysis results, more features can be included after lowering the threshold. Results show that all of the 12 selected features play an important role in estimating AADT. This is indicated by multiple variable importance measures after machine learning models are trained. This part of work provides an informational guidance for researchers to select useful features for AADT estimation. Besides, the data used for feature selection are two public domain databases, i.e. Smart Location

Database and Highway Performance Monitoring System data. Benefiting from the nationwide coverage and good structure of these two databases, an extensive and widespread application of the method becomes feasible and flexible from small geographical units such counties, census tracts, etc. to large study areas such as State and national level.

Modeling through machine learning is the second part of work. Instead of a simple application of machine learning algorithms, the trained model is demystified from multiple perspectives such as the inner structure after training, the importance measures of predictors, the associations between each predictor and AADT, and the interactions among input features. First, both ANN and RF, two popularly used machine-learning algorithms for prediction, are used for AADT estimation. To increase the robustness of artificial neural network modeling, the ensemble theory, a core structure of random forest, is applied by building up a group of artificial neural networks. Estimation results are more reliable and stable for this assembling structure. Final estimation results show that both ANN and RF perform well in terms of accuracy. A spatial lag model is built as a benchmark model. Significant improvements in all five accuracy measures including MSE, RSQ, RMSE, MAE, and MAPE can be seen when ANN and RF are compared with the spatial lag model. For example, RF shows a 57% decrease of MAPE and ANN shows a 28% decrease of MAE in comparison with the benchmark model. Additionally, RF performs better than ANN in all accuracy measures. Second, the mysterious mask of machine learning algorithms, usually named as black box algorithms is unveiled largely. How input features interact with AADT are analyzed through partial dependence plots. Not only the positive or negative

correlation are depicted but also the sensitivity of each input feature to AADT is presented. This enhances the understanding of how predictors act on the AADT. Besides, how the neurons transfer or interplay with each other in the layers of neural network and the importance rank of input features are visualized. Both the strength of interaction and sign (i.e. positive or negative) along the links between neurons are clearly presented as well. For random forest modeling, multiple variable importance measures are utilized, including the percentage increase in MSE upon permuting a given feature, number of root nodes that the given feature splits, and mean minimal depth. All selected predictors show their importance from different aspects. To further uncover the interactions among the features, conditional mean minimal depth is analyzed for each predictor, which shows that some features depend on the presence of other features to make a difference. It is implied that feature selection should also value the combinations of some features.

References

[1] Rossi, R., M. Gastaldi, G. Gecchele, and S. Kikuchi. Estimation of Annual Average Daily Truck Traffic Volume: Uncertainty Treatment and Data Collection Requirements. *Procedia - Social and Behavioral Sciences*, 2012. 54: 845-856.

[2] Sharma, S. C., B. M. Gulati, and S. N. Rizak. Statewide Traffic Volume Studies and Precision of AADT Estimates. *Journal of Transportation Engineering*, 1996. 122: 430-439.

[3] Zhao, F., and S. Chung. Contributing Factors of Annual Average Daily Traffic in a Florida County: Exploration with Geographic Information System and Regression Models. *Transportation Research Record: Journal of the Transportation Research Board*, 2001. 1769: 113-122.

[4] Xia, Q., F. Zhao, Z. Chen, L. Shen, and D. Ospina. Estimation of Annual Average Daily Traffic for Nonstate Roads in a Florida County. *Transportation Research Record: Journal of the Transportation Research Board*, 1999. 1660: 32-40.

[5] Mohamad, D., K. C. Sinha, T. Kuczek, and C. F. Scholer. Annual Average Daily Traffic Prediction Model for County Roads. *Transportation Research Record: Journal of the Transportation Research Board*, 1998. 1617: 69-77.

[6] Zhao, F., and N. Park. Using Geographically Weighted Regression Models to Estimate Annual Average Daily Traffic. *Transportation Research Record: Journal of the Transportation Research Board*, 2004. 1879: 99-107.

[7] Eom, J., M. Park, T. Heo, and L. Huntsinger. Improving the Prediction of Annual Average Daily Traffic for Nonfreeway Facilities by Applying a Spatial Statistical Method. *Transportation Research Record: Journal of the Transportation Research Board*, 2006. 1968: 20-29.

[8] Shamo, B., E. Asa, and J. Membah. Linear Spatial Interpolation and Analysis of Annual Average Daily Traffic Aata. *Journal of computing in civil engineering*, 2015. 29: 04014022(1)-04014022(8).

[9] Geol, P. K., M. R. McCord, and C. Park. Exploiting Correlations between Link Flows to Improve Estimation of Average Annual Daily Traffic on Coverage Count Segments: Methodology and Numerical Study. *Transportation Research Record: Journal of the Transportation Research Board*, 2005. 1917: 100-107.

[10] Azad, A. K., and X. Wang. Prediction of Traffic Counts Using Statistical and Neural Network Models. *Geomatica*, 2015. 69: 217-284.

[11] Duddu, V. R., and S. S. Pulugurtha. Principle of Demographic Gravitation to Estimate Annual Average Daily Traffic: Comparison of Statistical and Neural Network Models. *Journal of Transportation Engineering*, 2013. 139: 585-595.

[12] Tsapakis, L., W. H. Schneider, and A. P. Nichold. A Bayesian Analysis of the Effect of Estimating Annual Average Daily Traffic for Heavy-duty Trucks Using

Training and Validation Data-sets. *Transportation Planning and Technology*, 2013. 36: 201-217.

[13] Selby, B., and K. M. Kockelman. Spatial Prediction of Traffic Levels in Unmeasured Locations: Applications of Universal Kriging and Geographically Weighted Regression. *Journal of Transport Geography*, 2013. 29: 24–32.

[14] Karlaftis, M.G., and E.I. Vlahogianni. Statistical Methods Versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights. *Transportation Research Part C*, 2011. 19: 387-399.

[15] Smith, K. A., and J. N.D. Gupta. Neural Networks in Business: Techniques and Applications for the Operations Researcher. *Computers & Operations Research*, 2000. 27: 1023-1044.

[16] Sharma, S., P. Lingras, F. Xu, and P. Kilburn. Application of Neural Networks to Estimate AADT on Low-volume Roads. *Journal of Transportation Engineering*, 2001. 127: 426-432.

[17] Pulugurtha, S. S., and P. R. Kusam. Modeling Annual Average Daily Traffic with Integrated Spatial Data from Multiple Network Buffer Bandwidths. Transportation Research Record: *Journal of the Transportation Research Board*, 2012. 2291: 53-60.

[18] Zhang, L., J. Hong, A. Nasri, and Q. Shen. How Built Environment Affects Travel Behavior: A Comparative Analysis of the Connections between Land Use and Vehicle Miles Traveled in US Cities. *The Journal of Transportation and Land Use*, 2012. 5: 40-52.

[19] Heo, T. Y., M. S. Park, J. K. Eom, and J. S. Oh. A Study on the Prediction of Traffic Counts Based on Shortest Travel Path, *The Korean Journal of Applied Statistics*, 2007. 20: 459-473.

[20] Selby, B., and K. Kockelman. Spatial Prediction of AADT in Unmeasured Locations by Universal Kriging. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., 2011.

[21] Zarei N., Ghayour M.A., Hashemi S. (2013) Road Traffic Prediction Using Context-Aware Random Forest Based on Volatility Nature of Traffic Flows. In: Selamat A., Nguyen N.T., Haron H. (eds) *Intelligent Information and Database Systems. ACIIDS 2013. Lecture Notes in Computer Scienc*e, vol 7802. Springer, Berlin, Heidelberg.

[22] B. Hamner, "Predicting Travel Times with Context-Dependent Random Forests by Modeling Local and Aggregate Traffic Flow," *2010 IEEE International Conference on Data Mining Workshops*, Sydney, NSW, 2010, pp. 1357-1359.

[23] Ulrik Brandes, (1, 2) A Faster Algorithm for Betweenness Centrality. Ulrik Brandes, *Journal of Mathematical Sociology* 25(2):163-177, 2001.

[24] Shaw, M. E. (1954). Group structure and the behavior of individuals in small groups. *The Journal of Psychology: Interdisciplinary and Applied*, 38, 139–149. https://doi.org/10.1080/00223980.1954.9712925.

[25] Moran, P. A. P. (1950). "Notes on Continuous Stochastic Phenomena". *Biometrika*. 37 (1): 17–23. doi:10.2307/2332142. JSTOR 2332142.

[26] A. Paluszynska (2017). Structure mining and knowledge extraction from random forest with applications to The Cancer Genome Atlas project, University of Warsaw, Soba, Poland). Retrieved from https://rawgit.com/geneticsMiNIng/BlackBoxOpener/master/randomForestExplainer_Master_thesis.pdf.

[27] Garson, G.D. 1991. Interpreting neural network connection weights. *Artificial Intelligence Expert*. 6(4):46–51.

[28] Goh, A.T.C. 1995. Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*. 9(3):143–151.

[29] T. Klakto, T.U. Saeed, M. Volovski, S. Labi, J.D. Fricker, K.C. Sinha. 2017. Addressing the Local-Road VMT Estimation Problem Using Spatial Interpolation Techniques. *Journal of Transportation Engineering.* 143(8). doi: 10.1061/JTEPBS.0000064