

Estimating AADT on Non-Coverage Roads

FINAL REPORT

Prepared by:

Nathan Huynh, Ph.D.
Jing Wang, M.S.
Ryan DeVine, M.S.

Department of Civil and Environmental Engineering
University of South Carolina

Mashrur "Ronnie" Chowdhury, Ph.D.
Weimin Jin, Ph.D.
Pronab Biswas

Department of Civil Engineering
Clemson University

Gurcan Comert, Ph.D.
Department of Computer Science, Physics, and Engineering
Benedict College

FHWA-SC-21-07

September 2021

Sponsoring Agencies:

South Carolina Department of Transportation

Office of Materials and Research

1406 Shop Road
Columbia, SC 29201

Federal Highway Administration

South Carolina Division

Strom Thurmond Federal Building
1835 Assembly Street, Suite 1270
Columbia, SC 29201

1. Report No FHWA-SC-21-07	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Estimating AADT on Non-coverage Roads		5. Report Date September 30, 2021	
		6. Performing Organization Code	
7. Author/s Nathan Huynh, Jing Wang, Ryan DeVine, Mashrur "Ronnie" Chowdhury, Weimin Jin, Pronab Biswas, and Gurcan Comert		8. Performing Organization Report No.	
9. Performing Organization Name and Address University of South Carolina Department of Civil and Environmental Engineering 300 Main St. Columbia, SC 29208		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No. SPR No. 749	
12. Sponsoring Organization Name and Address South Carolina Department of Transportation Office of Materials and Research 1406 Shop Road Columbia, SC 29201		13. Type of Report and Period Covered	
		14. Sponsoring Agency Code	
15. Supplementary Notes			
16. Abstract <p>This project developed several models to obtain accurate estimates of annual average daily traffic (AADT) at non-coverage locations. The developed models include kriging, regression and point-based. The standard kriging approach is modified in this project to use a default value when its predicted value is over a user-specified threshold (referred to as “hybrid kriging model”). Specifically, when a sampled coverage location is found to have a high absolute error using the kriging method, it is assumed that the surrounding non-coverage locations will also have AADT errors if kriging is used. In such cases, the mean AADT, based on county and functional class, is used as the AADT estimate. Two types of regression models were developed: regular and quantile. The statistically significant variables in the regular regression model are <i>Urban, Single Line, Other Type Median, and Left-Turn Lane</i>. The statistically significant variables in the quantile regression model are <i>Urban, Single Line, Other Type Median, Right-Turn Lane, Left-Turn Lane, and Parking Lot</i>. The point-based model, based on the work of Portland State University and sponsored by Oregon DOT, was calibrated using SCDOT data. It predicts that the AADT is 125 vehicles per day when a non-coverage location has zero points (i.e., features), 175 with one point, 350 with two points, 650 with three points, 900 with four points, 1,600 with five points, and 1,800 with six or seven points. Comparison of these models against the use of default values shows a 21.37% improvement for the hybrid kriging model, 22.82% for the point-based model, 17.03% for the regular regression model, and 23.19% for the quantile regression model.</p> <p>To facilitate the implementation of the developed models, an Excel-based tool was created, where the hybrid kriging model serves as the primary model because it provides comparable improvement to other models, but it does not require the SCDOT to collect any additional data. The tool also allows the user to use the predicted AADT from either regression models or point-based model if the road features are available. Other configurable parameters include an absolute error threshold for when a default value should be used instead of the kriging estimate and a reduction factor to account for discrepancy between coverage counts' mean AADT and non-coverage counts' mean AADT.</p>			
17. Key Words AADT, non-coverage counts, coverage counts, hybrid kriging method, point-based model, regression model.		18. Distribution Statement No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161.	
19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. Of Pages	22. Price

DISCLAIMER

The contents of this report reflect the views of the author who is responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the South Carolina Department of Transportation or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

The State of South Carolina and the United States Government do not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because they are considered essential to the object of this report.

ACKNOWLEDGMENTS

The project team greatly appreciates the guidance and assistance from the following Project Steering and Implementation Committee members:

- Anderson, Todd (Chair)
- Wise, Stacy
- Siddiqui, Chowdhury
- Swygert, Terry
- Heaps, Meredith

EXECUTIVE SUMMARY

The objectives of this research project were to: 1) develop models to estimate annual average daily traffic (AADT) at non-coverage locations, and 2) develop a user-friendly tool that implements the models.

A literature review and state department of transportation (DOT) survey were conducted to determine the most applicable models. Findings from the literature review indicate that multiple linear regression is the most widely used method to estimate AADT due to their simplicity. Machine learning models can provide more accurate results; however, its complexity makes it difficult to implement. Travel demand methods are theoretically sound, but they are computationally expensive. Kriging is simple to implement, and it requires only the use of known nearby AADT values. The point-based model, essentially a look-up table, proposed by Portland State University researchers in a project sponsored by the Oregon DOT is simple to understand and easy to implement, but it requires collection of additional data. Multiple regression models and travel demand models also require collection of additional data.

Findings from the online State DOTs survey indicate that multiple linear regression is the most commonly used method to estimate AADT at non-coverage locations among the 17 respondents. The next two most popular methods are visual estimation and default values. This finding suggests that there is a need for an AADT estimation technique, one that is simple to implement. This inference is supported by the fact that most respondents rated their satisfaction with the current non-coverage AADT estimation technique as three or less, with one being unsatisfied and five being satisfied.

Based on the findings from the literature review and state DOT survey, the kriging, regression and point-based models were developed using a dataset consisting of 3,687 coverage counts, 2,510 were collected in 2020 and 1,177 in 2021, and 1,024 non-coverage counts, 548 were collected in 2019, 239 in 2020 and 237 in 2021. The standard kriging approach was modified in this project to use a default value when its predicted value is over a user-specified threshold. Specifically, when a sampled coverage location is found to have a high absolute error using the kriging method, it is assumed that the surrounding non-coverage locations will also have AADT errors if kriging is used. In such cases, the mean AADT, based on county and functional class, is used as the AADT estimate. The advantage of the kriging model over the point-based and regression models is that it uses only the coverage counts to make prediction; no additional information, such as land use and socio-demographic information is needed.

Two types of regression models were developed: regular and quantile. The statistically significant variables in the regular regression model are *Urban, Single Line, Other Type Median, Left-Turn Lane, Right-Turn Lane and Sidewalk*. The statistically significant variables in the quantile regression model are *Urban, Single Line, Other Type Median, Right-Turn Lane, Left-Turn Lane, Sidewalk and Parking Lot*.

The point-based model, developed based on the work of Portland State University for Oregon DOT, predicts the AADT using the median AADT of roadways that have the same number of points or features. One point is assigned for each of the following roadway features:

- In urban area
- Presence of centerline marking (i.e., double yellow line)
- Presence of median
- Presence of right-turn lane
- Presence of left-turn lane
- Presence of parking lot adjacent to the study road segment
- Presence of sidewalk

The point-based model equates to a lookup table, as shown below. A local road with none of the above features is expected to have 125 vehicle per day (vpd) and a road with six or seven features is expected to have 1,800 vpd.

Point	Predicted AADT for Local Roads
0	125
1	175
2	350
3	650
4	900
5	1,600
6 or 7	1,800

The Root Mean Square Error (RMSE) was used to evaluate the performance of the different models. Compared to the current default values, the kriging model yielded a 21.37% improvement, the point-based model yielded a 22.82% improvement, the regular regression model yielded a 17.03% improvement, and the quantile regression model yielded a 23.19% improvement.

To facilitate the implementation of the developed models, an Excel-based tool was created, where the hybrid kriging model serves as the primary model because it provides comparable improvement to other models, but it does not require the SCDOT to collect any additional data. The tool also allows the user to use the predicted AADT from either regression models or point-based model if the road features data are available. Other configurable parameters include an absolute error threshold for when a default value should be used instead of the kriging estimate and a reduction factor to account for discrepancy between coverage counts' mean AADT and non-coverage counts' mean AADT.

Based on this project's findings, it is recommended that the SCDOT consider adopting the developed Excel-based tool. A 21.37% improvement in terms of RMSE can be expected with the use of the hybrid kriging model. When roadway features are available for non-coverage roads, the SCDOT could change the configurable parameter in the tool to use estimates from the point-based model (a 1.45% improvement over kriging) or the quantile regression model (a 1.82% improvement over kriging).

TABLE OF CONTENTS

Chapter 1: Introduction	1
Chapter 2: Literature Review	3
2.1 Literature Review.....	3
2.1.1 Regression Analysis.....	3
2.1.2 Nonlinear Regression Model	7
2.1.3 Kriging	7
2.1.4 Travel Demand.....	8
2.1.5 Machine Learning	9
2.1.6 Centrality.....	10
2.1.7 Point-Based Model.....	11
2.1.8 Summary	11
2.2 State-of-the-Practice on Non-Coverage AADT Estimation	13
Chapter 3: Models Development	16
3.1 Data Description	16
3.1.1 Non-Coverage Counts Dataset.....	16
3.1.2 Coverage Counts Dataset.....	17
3.2 Models Training Dataset.....	21
3.3 Models.....	23
3.3.1 Kriging	23
3.3.2 Point-Based Model.....	32
3.3.3 Regression Models.....	32
Chapter 4: Results	34

4.1 Use of Default Value	34
4.2 Kriging Model.....	36
4.3 Point-Based Model.....	37
4.4 Regression Model	39
4.4.1 Regular Regression Model.....	39
4.4.2 Quantile Regression Model.....	39
4.5 Comparison of Models Performance	40
Chapter 5: Conclusion and Recommendations	41
5.1 Conclusion	41
5.2 Recommendations.....	41
5.3 Implementation	41
REFERENCES	46

LIST OF FIGURES

Figure 1-1 Map of coverage locations where short-term counts are regularly collected.....	1
Figure 1-2 Map of non-coverage locations where an AADT estimate is required.....	2
Figure 2-1. Non-coverage AADT estimation tool	15
Figure 3-1 Map of non-coverage count stations	17
Figure 3-2 Map of coverage count stations.....	18
Figure 3-3 Distribution of coverage counts by functional class	20
Figure 3-4 Illustration of kriging assigning weights to neighbors (Smith, 2020).....	24
Figure 3-5 Semivariogram example.....	26
Figure 3-6 Example empirical semivariogram.....	28
Figure 3-7 Comparison of different theoretical semivariogram models: (a) gaussian model, (b) exponential model, (c) spherical model, and (d) linear model	29
Figure 3-8 Sampled locations and their absolute errors.....	31
Figure 3-9 Illustration of scenario when an average AADT is used instead of kriging-predicted AADT	32
Figure 4-1 Locations of stations used for model training.....	38
Figure 4-2 Comparison of models' performance.....	40
Figure 5-1 Graphical user interface of non-coverage AADT estimation tool	42
Figure 5-2 Map of coverage counts generated by developed tool.....	43
Figure 5-3 Dialog showing value of calculated reduction factor.....	44
Figure 5-4 Configurable parameters	44

LIST OF TABLES

Table 2-1 Summary of literature review	12
Table 2-2 The method(s) being used to estimate AADT at non-coverage locations.....	13
Table 2-3 Satisfaction with current AADT method.....	13
Table 2-4 Use of tool to estimate AADT at non-coverage locations.....	14
Table 2-5 Willingness to share tool with SCDOT	14
Table 3-1 Functional class	18
Table 3-2 Summary statistics of AADT values for each functional class	20
Table 3-3 AADT statistics of coverage and non-coverage counts	21
Table 3-4 Roadway features collected for model development.....	21
Table 3-5 Descriptive statistics of AADT by area.....	22
Table 3-6 Descriptive statistics of AADT by median types	22
Table 3-7 Descriptive statistics of AADT by presence of an exclusive right-turn lane	22
Table 3-8 Descriptive statistics of AADT by presence of an exclusive left-turn lane	23
Table 3-9 Descriptive statistics of AADT by presence of a sidewalk on both sides of the roadway	23
Table 3-10 Descriptive statistics of AADT by presence of a parking lot.....	23
Table 4-1 Statistics of rural and urban AADT values for non-coverage counts.....	34
Table 4-2 Default AADT values for counties.....	35
Table 4-3 Effect of radius on kriging model performance	36
Table 4-4 Effect of absolute error threshold on kriging model performance	36
Table 4-5 Effect of including known non-coverage counts.....	37
Table 4-6 AADT prediction by the point-based model	38
Table 4-7 Regular regression model estimation results.....	39

Table 4-8 Coefficients for the quantile regression model..... 39

LIST OF ACRONYMS

AADT	Annual Average Daily Traffic
FC	Functional Class
GWMLR	Geographically Weighted Multiple Linear regression
LRS	Linear Reference System
MAPE	Mean Absolute Percentage Error
MdAPE	Median Absolute Percent Error
MLR	Multiple Linear Regression
MSA	Metropolitan Statistical Area
MSPE	Mean Squared Prediction Error
RMSE	Root Mean Squared Error
TWLTL	Two-Way Left-Turn Lane
SCAD	Smooth Clipped Absolute Deviation
SCDOT	South Carolina Department of Transportation
SVR	Support Vector Regression

CHAPTER 1: INTRODUCTION

The South Carolina Department of Transportation (SCDOT) is responsible for the planning, design, construction and maintenance of over 41,000 centerline miles of interstate, non-interstate National Highway System (NHS), non-NHS primary, Federal Aid secondary, and Non-Federal Aid secondary roads within the State. For the SCDOT to adequately perform these tasks, the agency needs to perform traffic counts on a regular basis. Specifically, traffic counts provide annual average daily traffic (AADT), which serves as an input to many transportation studies (e.g., transportation planning, traffic safety analysis, and pavement design). The SCDOT is also required to collect and report AADTs to the FHWA annually as part of the Highway Performance Monitoring System (HPMS) program.

Currently, the SCDOT has about 185 permanent count and weight-in-motion stations located throughout the entire state; these stations are primarily on interstates. Additionally, the SCDOT has about 12,000 short-term count stations. A map of these locations is shown in Figure 1-1; these locations are called *coverage* locations because traffic counts are updated on an annual, biennial, or triennial basis. At these stations, counts are collected for 48 hours, and these “short-term” counts are then converted to AADTs using expansion factors. These factors account for the day of the week and month of the year in which the short-term count was collected, as well as the number of axles per vehicle class. A short-term count requires the SCDOT to send its personnel to the roadway of interest and set up the pneumatic tubes and counters or hire a contractor at a cost. This practice is labor intensive, costly, and puts the safety of SCDOT personnel or contractors at risk.

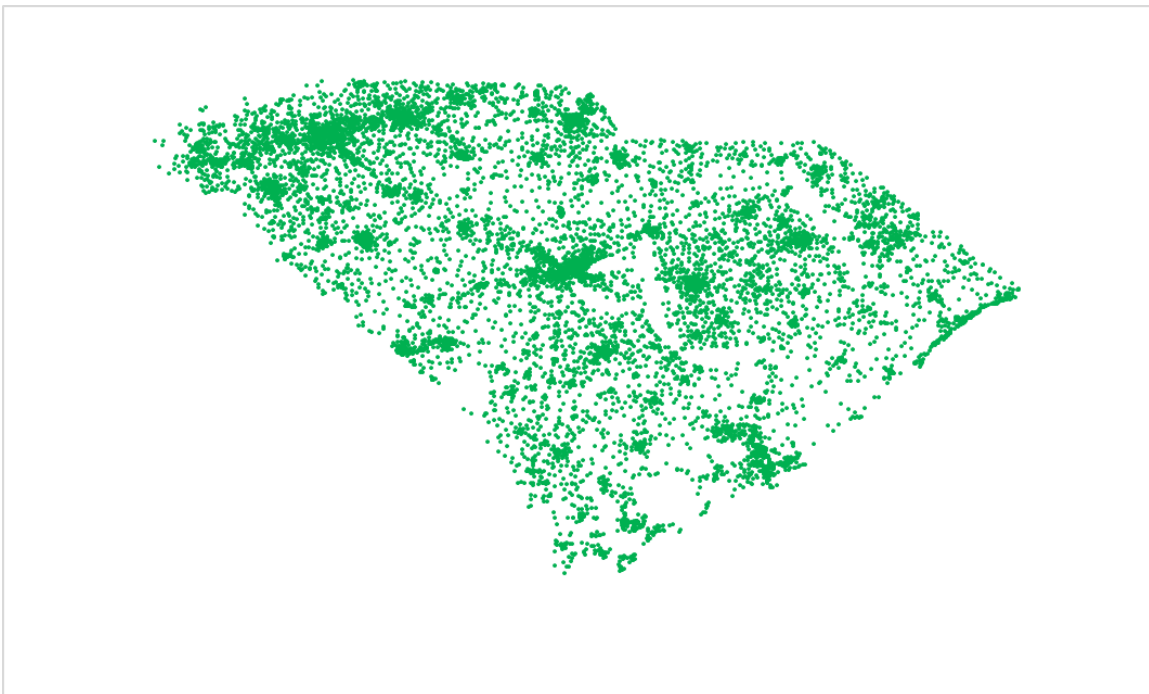


Figure 1-1 Map of coverage locations where short-term counts are regularly collected

In Figure 1-1, even though it may appear that the coverage locations cover most roadways in the state, there are significantly more roadways where the SCDOT does not collect traffic counts, and hence, does not know their actual AADTs. These locations are shown in Figure 1-2, and they are referred to as *non-coverage* because traffic counts have not been collected at these locations or it has been at least 10 years since the last time a short-term count was performed; some DOTs refer to these locations as “out-of-network.” While SCDOT performs counts at non-coverage locations as scheduling and funds permit, it is cost prohibitive to perform a short-term count at every non-coverage location in the state. Therefore, the SCDOT simply uses a default AADT value based on the roadway's functional class and area type. That is, if the roadway is a local, rural road, then a default value of 100 vehicles/day (vpd) is used. Similarly, if the roadway is an urban, local road, then a default value of 200 vpd is used. The SCDOT recognizes that these default values may not reflect the actual AADTs, and therefore, sought to improve upon current practice with this research project. The aim of this project is to provide quantitative and justifiable methods for obtaining AADT at non-coverage locations. To this end, this research project sought to: 1) develop models to estimate AADT at non-coverage locations, and 2) develop a user-friendly tool that implements the models.

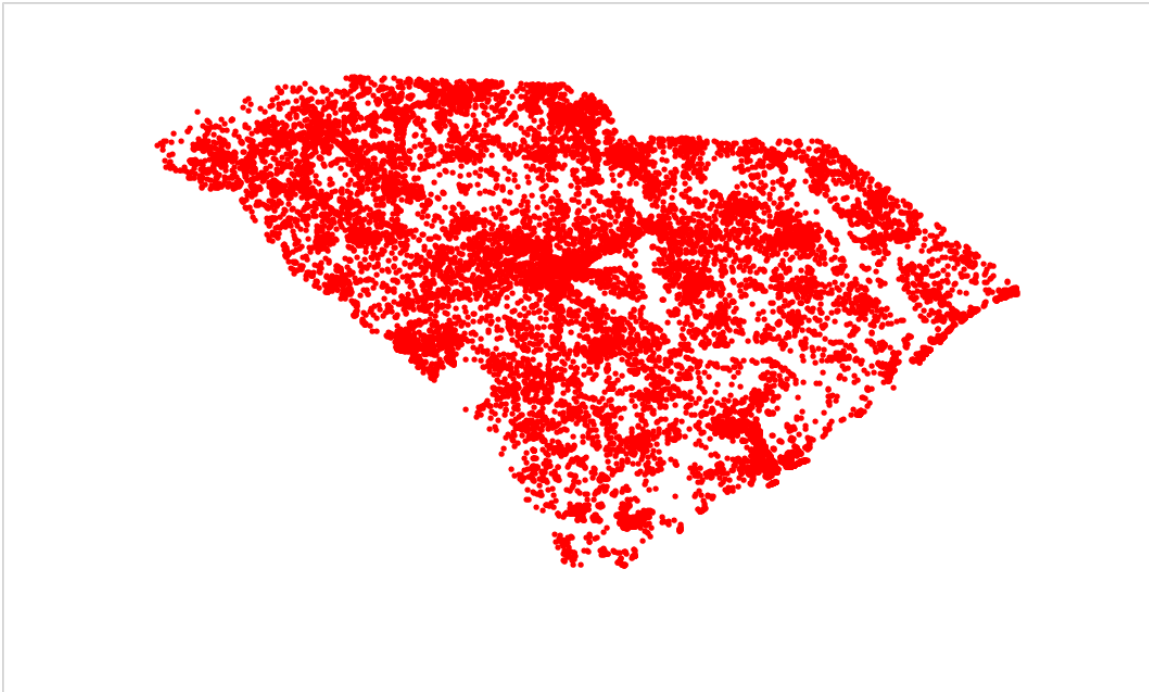


Figure 1-2 Map of non-coverage locations where an AADT estimate is required

CHAPTER 2: LITERATURE REVIEW

2.1 Literature Review

There have been many studies that focused on AADT estimation. The majority developed techniques to estimate AADT using short-term counts. A much smaller number of studies explored methods to estimate AADT using other sources of data such as land use, census, roadway and network characteristics. These studies are reviewed below and grouped into the following categories: regression analysis, kriging, travel demand, machine learning, centrality, and point-based.

2.1.1 Regression Analysis

Regression analysis is a statistical procedure used to study the linear/non-linear relationship between a dependent variable (i.e., AADT) and multiple independent variables. The commonly used regression techniques for estimating AADT are multiple linear regression, nonlinear regression, and geographically weighted regression.

2.1.1.1 Multiple Linear Regression

In 1998, Mohamad et al. (1998) applied multiple linear regression to predict the AADT of local roads in Indiana. In this model, both quantitative and qualitative variables were used to predict AADT values. The independent variables that were initially used include urban/rural classification, easy access to state highway, interstate existence, county population, total state highway mileage of county, per capita income, total households, total vehicle registration of county, total employment, total arterial mileage of county, and total collector mileage of county. Lastly, it was determined that urban/rural classification, easy access to state highway, county population, and total arterial mileage of county were the most significant variables. The final model containing with four independent significant was then validated by randomly measuring the AADT at eight new locations and using the model to predict those values. The mean square error of the validation set was 16%. The authors indicated that since the mean square prediction error of the validation set was similar to the mean square error of the test set, their model is unbiased.

In 1999, Xia et al. (1999) developed a multiple linear regression model to estimate AADT on out-of-network roads in urban areas of Florida. Using 450 count stations, their study was able to develop a large data set for multiple linear regression modeling, which had not been accomplished before. The 14 predictor variables investigated were categorized into roadway characteristics, socioeconomic characteristics, and road network connectivity measurements. Variable reduction was performed using statistical methods similar to those in Mohamad et al., 1998. After utilizing statistical tests, the authors noted that roadway characteristics were more significant than socioeconomic factors; the socioeconomic factors had little effect on AADT. The roadway characteristics considered included the number of lanes, the area land use type, and the functional classification. After removing redundant independent variables, the final model had six predictor variables including accessibility to non-state roads, number of lanes, land use type, functional classification, automobile ownership, and service employment. It was validated using data from 40 additional locations. The R-squared value for the 40 selected locations was 0.63 and the Mean Absolute Percentage Error (MAPE) was 22.7%.

Yang et al. (2014) proposed one multi-linear regression model, using the smooth clipped absolute deviation (SCAD) procedure to estimate the coefficients as well as select significant variables in one step. Data was assembled for four categories of input variables – driving behavior, roadway characteristics, satellite data, and socio-economic variables. With respect to driving behavior, the initial data considered include loading factor or the contribution of each household to roadway sections. The following variables was considered for roadway characteristics – number of lanes, length, connectivity to local roads, connectivity to high-level roads, AADT of nearest collector, and location of road. From Google map images, data on number of cars on the road and the car intensity - number of cars per unit length was extracted. The following socio-economic variables were assembled at the zip code level – population, population density, housing units, land area, water area, median income, percentage of unemployed, and percentage of people below poverty line. AADT and relevant explanatory variable data was assembled for 243 count locations in Mecklenburg County, NC, in 2007 for local subtracting the mean and dividing by the standard deviation. Two hundred out of the 243 data points were used for model calibration. The remaining 43 data points was used to validate the model. The above model obtained from SCAD variable selection procedure was found to outperform the multiple linear regression model obtained from forward stepwise regression procedure.

In 2000, Seaver et al. (2000) expanded the multiple linear regression methodology by incorporating principal component analysis and a cluster regression analysis. Starting with 45 potential parameters, principal component analysis was used to reduce the number of independent variables to around seven or eight, depending on the area being investigated. These principal variables include percent population change, median travel time, number of agricultural farms, percent of farm with 500+ acres, median household income, median time to leave for work, distance to MSA, average daily traffic, population density, unemployment rate, number of persons working outside of the county, and per capita income. The cluster regression analysis was able to locate groups with the same road type and metropolitan status (i.e., either in or out of a metropolitan statistical area, MSA). Within each cluster, a multiple linear regression was performed to estimate AADT using the previously determined principal variables. However, even with the integration of these techniques, the success of the model varied greatly. The models within an MSA achieved an R-squared value ranging from 0.46 to 0.75, and the models outside of an MSA achieved an R-squared value ranging from 0.27 to 0.94.

In 2001, Zhao and Chung (2001) continued the work that was started in 1999 by Xia et al. By 2001, the already large database of AADT count information had grown to incorporate all AADT's for state roads, the federal functional classification system, and more extensive land-use and accessibility variables. With these improvements to the database, four multiple linear regression models were developed. One model had four variables, two models had five variables, and one model had six variables. The most promising of the models had five variables, number of lanes, functional class, regional accessibility to employment centers, an employment indicator, and direct access to an expressway. It had an R-squared value of 0.818, which is a good improvement from the R-squared value of 0.63 obtained in the 1999 study.

In 2006, Anderson et al. (2006) was able to compare the multiple linear regression method to a travel demand method by focusing on a small urban community in Alabama. The travel demand method is regarded as a well-established method, but it is computationally expensive, especially

for large networks. Therefore, it was necessary to determine if multiple linear regression, which is much more time efficient, could produce comparable results. The multiple linear regression model had five independent variables: number of lanes, functional class, population, employment, and a binary variable that represents mobility. After both models were developed, it was observed that both models produced similar results. This was confirmed by using a t-test, graphical inspection, and a Nash-Sutcliffe statistic. The R-squared value for the multiple linear regression model was 0.819.

In 2008, Pan (2008) extended multiple linear regression to estimate the AADT of all roads in Florida. The independent variables that were considered included population, total lane mileage of highway, vehicle registration, personal income, retail sales, municipalities, labor force, and roadway characteristics (e.g., divided/undivided median, number of lanes, rural/urban, land use, and accessibility to freeways). The state of Florida was broken into three categories based on population (low, medium, and high population), and for each of these two models were developed. One model was developed for the state/county highways and another model was developed for local street roads. It was observed that the highway model outperformed the local model for all three population areas. However, it was also noticed that the models developed for the low population areas (MAPE of 31.99% and 46.69%) outperformed the models that were developed for the medium (MAPE of 65.01% and 65.35%) and high (MAPE of 46.81% and 159.49%) population areas.

In 2012, Lowry and Dixon (2012) integrated a multiple linear regression model into ArcGIS by using open-source Python scripts. Since most rural roads have uniform characteristics, a multiple linear regression analysis would not be able to predict AADT because there is not enough variability in the independent variables. To overcome this limitation, a new parameter called connectivity importance index was introduced. The connectivity importance index is determined by finding the shortest path between every node in the network. The number of times a node is included in a shortest path is that node's connectivity importance index. By using functional class, number of lanes, and connectivity importance index as independent variables, a multiple linear regression model was created with an R-squared value of 0.72.

In 2014, Yang et al. (2014) proposed a new variable selection procedure for multiple linear regression called smoothly clipped absolute deviation penalty (SCAD). This selection procedure was able to select significant independent variables and estimate regression coefficients in one step, instead of being split into two different procedures. The SCAD selection procedure was then compared to backward and forward variable selection procedures. The following variables were determined to be significant: number of cars in a satellite image, number of lanes, housing units, median income, percentage below poverty line, and car intensity in a satellite image. It was observed that backward and SCAD selection procedures resulted in the same R-squared value (0.6954), while both outperformed forward variable selection (0.6423).

In 2016, Apronti et al. (2016) developed a multiple linear regression model to predict AADT values in Wyoming. The final model utilized pavement type, access to primary or secondary roads, agricultural cropland, agricultural pastureland, industrial areas, and population in the census block group as independent variables. Using the Box-Cox transformation, it was determined that a log transform of AADT would enhance the multiple linear regression. Before the log transformation

was applied the R-squared value was 0.44, and after the log transformation was applied the R-squared value was 0.64. Also, after the log transformation of AADT, the errors appeared constant and the residuals appeared normally distributed, again showing the benefit of a log transformation. Lastly, when the multiple linear regression model was validated, the R-squared value was 0.69. Similar test and validation R-squared values implied that the model was not biased.

In 2016, Staats (2016) utilized probe counts in multiple linear regression to predict AADT values on local roads in Kentucky. The state of Kentucky was split into three geographic areas by using highway districts, which was done to account for geographic and socioeconomic variability. Then a model was developed using counts from probe vehicles, residential vehicle registration, and curve rating as independent variables for each of the three areas. For each of the three areas investigated, a rural model and an urban model was developed. The rural models were developed by using only AADT values that ranged between 20 and 1,000. This was chosen because a road is not considered rural if the AADT is above 1,000. This limitation was not imposed on the urban model. For the rural models, the MAPE ranged from 61% to 87%, while the MAPE for the urban models ranged from 354% to 1,956%.

2.1.1.2 Geographically Weighted Multiple Linear Regression

Geographically weighted multiple linear regression models account for dependencies and correlations between variables based on geographic locations. Geographically weighted multiple linear regression models are increasingly becoming popular in transportation applications due to their ability to better capture geographical variations.

In 2004, Zhao and Park (2004) were one of the first to investigate geographically weighted multiple linear regression (GWMLR) models for use to estimate AADT. Studying roads in Florida, an ordinary multiple linear regression model was created to serve as a control, and the same parameters were then used in two geographically weighted models. The parameters utilized included the number of lanes, regional accessibility to employment centers, population size, employment size, and direct access to expressways. The first model utilized a bi-square weighting function, and the second model utilized a Gaussian weighting function. Both GWMLR models outperformed the control model (R-squared value of 0.764), while the bi-square model (R-squared value of 0.8756) outperformed the Gaussian model (R-squared value of 0.8700). This improvement over ordinary MLR shows the necessity of utilizing spatial variation when predicting AADT values.

In 2012, Pulugurtha and Kusam (2012) improved upon GWMLR by investigating multiple bandwidths to estimate off-network characteristics. Both negative binomial and Poisson weighting distributions were investigated, and it was observed that the negative binomial weighting distribution outperformed the Poisson weighting distribution. It was also observed that an appropriate bandwidth varies with the functional class being investigated. For freeways/expressways a five-mile buffer was appropriate, while a three mile buffer was appropriate for major thoroughfares and a two mile buffer was appropriate for minor thoroughfares. A model was developed for the entire study area and additional models were developed for each functional class. The entire study area model had the following predictor variables: urban classification, freeways or expressways, major thoroughfares, number of lanes,

population, manufactured house, and innovative. The quasi-likelihood under the independence model criterion (QIC) was used to assess the models, and for this metric a smaller value is optimal. The entire study area model's QIC was found to be 61.43 for the negative binomial weighting and 1,945 for the Poisson weighting. The functional class-based models included the following predictor variables: urban classification, number of lanes, speed limit, upstream link speed limit, downstream link speed limit, downstream cross street link number of lanes, population, manufactured house, and rural district. Based on a drop in QIC, it was observed that segmenting the study area into groups based on functional class allowed for better accuracy. This improvement was also observed in the ordinary MLR models.

2.1.2 Nonlinear Regression Model

Nonlinear regression techniques assume that the AADT or the logarithm of AADT can be predicted as a nonlinear function of independent land use, socio-economic, and demographic variables.

In 2018, Chang and Cheon (2019) proposed a methodology to estimate AADT based on vehicle GPS data, also known as probe data, in South Korea. The methodology (KWPC) uses a locally weighted power curve to transform the k nearest probe counts to AADT. The number of nearest probe counts, k , was calibrated by using the elbow method. The KWPC model was then compared to multiple linear regression, geographically weighted multiple linear regression, and kriging. The KWPC model had the lowest MAPE (7.5%), followed by multiple linear regression (9.5%), then GWMLR (10.5%), and lastly kriging (42.5%).

2.1.3 Kriging

Kriging is a popular geostatistics method originally used in the mining industry for predicting ore reserves. The AADT at location s is determined based on a function of a deterministic trend $\mu(s)$ and an error $\epsilon(s)$ as follows:

$$Z(s) = \mu(s) + \epsilon(s)$$

The error terms are assumed to be spatially correlated. There are three different types of kriging depending on the nature of the assumption in describing $\mu(s)$. In simple kriging, the trend is assumed to be a known constant. In ordinary kriging, the trend is assumed to be an unknown constant. In universal kriging, the trend is assumed to be a function of independent variables. A semivariogram function is used to capture the spatial correlations. The three commonly used functions in AADT estimation are exponential, spherical, and gaussian.

In 2006, Eom et al. (2006) were at the forefront of utilizing kriging to estimate AADT at nonfreeway facilities. Multiple theoretical semivariograms were investigated, including Gaussian, exponential, and spherical. A theoretical semivariogram model was fitted to the experimental semivariogram by two approaches, weighted least squares (WLS) and restricted maximum likelihood (REML), with ordinary least squares (OLS) acting as a benchmark. The best semivariogram model for the weighted least squares was the spherical model while the best model for restricted maximum likelihood was the exponential model. It was observed that both methods

provided more accurate AADT estimations in both urban and rural areas when compared to traditional regression estimates; WLS had a mean square prediction error of 2.91, REML achieved an MSPE of 2.86, and OLS achieved an MSPE of 3.12. This shows that kriging can be utilized to estimate AADT values more accurately than MLR, without drastically increasing the complexity of the method.

In 2009, Wang and Kockelman (2009) improved upon the use of kriging by breaking the state of Texas into two different models, one for interstate highways and another for principal arterials. Once theoretical semivariograms were computed for each road type, it was observed that interstate highways had a higher nugget effect and range when compared to the principal arterial road class. It was also observed that the interstate highway developed model resulted in a median error of 33%. This was due to the kriging method overestimating the AADT values on roads that had low traffic volumes. It was observed that the model performed well for roads that have an AADT greater than 1,000. However, the overestimation could have occurred because the model was being implemented on interstate highways, meaning that a highway with a low AADT value would be an outlier compared to the other AADT locations.

In 2013, Selby and Kockelman (2013) compared kriging to GWMLR, and then investigated utilizing Euclidian distance instead of network distance. After developing the models, it was observed that kriging outperformed the GWMLR model by 3 to 8% in average absolute error. Following this, Euclidian distance and network distance were compared to see the effects on the model's error. There was no sizable difference in error between using Euclidian or network distance. This means that the time costly work of finding network distances can be exchanged with simple Euclidian distance.

In 2015, Shamo et al. (2015) comprehensively investigated different kriging techniques and different semivariogram models to estimate AADT on roads in Washington. The different kriging techniques that were investigated included simple kriging, ordinary kriging, and universal kriging, while the semivariogram models that were investigated included spherical, exponential, and Gaussian. The models were developed using traffic count data from different years - 2008, 2009, and 2010. The best fitting semivariogram model was not consistent for each kriging technique or year. In 2008, the exponential model was the best choice for all techniques, while in 2009 it was the spherical model. In 2010, the spherical model was used for simple and ordinary kriging, but the exponential model was used for universal kriging. It was also observed that simple kriging was the best model in 2008 and 2009, but ordinary kriging was the best model in 2010. Lastly, it was noticed that the RMSE was not constant for each year. In 2008 the RMSE ranged from 56.48% to 59.01%, while in 2009 it ranged from 94.49% to 95.31%, and in 2010 it ranged from 82.54% to 84.15%. This lack of consistency between the best performing semivariogram model shows the necessity in comparing all semivariogram models whenever new data is available, instead of relying on one model.

2.1.4 Travel Demand

Travel demand-based approaches mimic the four-step travel demand forecasting process. Instead of obtaining volume on links from the fourth step (i.e., traffic assignment), the modified approach produces AADTs.

In 2009, Zhong and Hanson (2009) put forward the four-step travel demand modeling approach to estimate the missing AADT information for low-class roads in York County and Beresford regions in New Brunswick, Canada. The quick response method from the National Cooperative Highway Research Program was used for trip generation, attraction, and balancing. Trip distribution was performed using a gravity model with a gamma function based on distances used for estimating impedances. The stochastic user equilibrium model was used for traffic assignment. The summation of traffic volumes for all trips was then used to estimate the daily traffic volume, which was then the AADT estimate. Traffic counts were then used to compare the estimations to actual results. Arterial highways had an average error of 9%, while collector highways had an average error of 44% and local highways had an average error of 174%. It was observed that these errors could be the result of traffic not being distributed to local or rural roads during the trip distribution process.

In 2013, Wang et al. (2013) proposed an updated travel demand method to predict the AADT of local roads in Florida. The proposed travel demand model was based on parcel level trip generation, distribution, and assignment. The parcel level model accounts for driver's response to a given local street system, while the traditional model would try to predict a driver's choices for an entire origin destination trip. The parcel level model was compared to a typical regression model. The typical regression model resulted in a MAPE of 211%, while the proposed parcel level model resulted in a MAPE of 52%. This is an improvement when compared to the previous study by Zhong and Hanson in 2009 and is caused by a mechanism similar to the improvement seen in other methods when a study area is broken into different regions.

2.1.5 Machine Learning

Machine Learning is an artificial intelligence technique which relies on pattern recognition algorithms. Two types of machine learning techniques have been applied to AADT estimation: support vector regression and decision trees.

In 2009, Castro-Neto et al. (2009) investigated the use of support vector regression with data-dependent parameters to predict AADT values on Tennessee roads. A comparison between support vector regression with data-dependent parameters, Holt exponential smoothing, and ordinary least squares regression was conducted by using Tennessee DOT data. After applying the different models to urban and rural roads, it was observed that the support vector regression with data-dependent parameters performed better than Holt exponential smoothing (MAPE of 2.26% compared to 2.69%), which performed better than the ordinary least squares regression (MAPE of 3.85%).

In 2015, Sun and Das (2015) utilized a modified support vector regression (SVR) method to estimate AADT on non-state roads in Louisiana. Using total population, total jobs, distance from interstate, and distance from a major US highway as independent variables, the SVR models were developed. Eight parishes in Louisiana were selected as the validation set for the analysis. Two SVR models were developed for each parish, one for rural areas and another for urban areas. For the rural models, the percent of samples that had an error of less than 100 ranged from 64% to 84%, while the percent of samples that had an error less than 100 for the urban area ranged from 63% to 100%.

In 2020, Sfyridis and Agnolucci (2020) integrated clustering with regression modeling to predict AADT on all roads in England and Wales. Since the predictor variables were both numeric and categorical, the K-prototype algorithm was used for clustering. Utilizing the elbow method, it was determined that the optimum number of clusters was five. The regression modeling was performed by ordinary multiple linear regression, random forest, and support vector regression. After using 80% of the test data for model development, the models were compared on the remaining 20%. The support vector regression model produced a MAPE ranging from 2% to 277% and was comparable to the random forest method, which produced a MAPE ranging from 2% to 288%. Both methods outperformed the multiple linear regression method, which produced a MAPE ranging from 2% to 325%.

2.1.6 Centrality

Centrality based methods rely on a node's centrality measure to predict the node's AADT value. There are multiple forms of centrality, but each form is a measure of how popular or utilized a node is. For example, stress centrality is the number of times a node is included in the shortest distance between every node pair. If a node has a high stress centrality, then multiple shortest paths go through that node, implying its popularity. Another common form of centrality is closeness centrality, which is based on the distance between every node. A node with high closeness centrality will be close to multiple nodes, which implies that node's popularity.

In 2014, Lowry (2014) studied the use of centrality for AADT estimation in Moscow, Idaho. Stress centrality was used as the form of centrality and is equal to the number of times a link would be used if someone traveled the shortest distance between every node pair. This was then modified by limiting the set of nodes to only origin-destination pairs and applying multipliers based on the land use type of the origin and destination nodes. The modified stress centrality was then implemented in an ordinary least squares regression and a robust regression that used a transformed AADT value. The calibration median absolute percent error (MdAPE) for the ordinary least squares model was 34% while the validation MdAPE was 22%. The calibration MdAPE for the robust model was 28% and the validation MdAPE was 29%. Lastly, the number of AADT observations being utilized was varied from 10 to 350, and the MdAPE for each number of observations was determined. It was observed that having over 100 observations made the validation and calibration MdAPE similar, which means that having over 100 observations made the model less biased.

In 2017, Keehan (2017) studied the applicability of a centrality measure to predict AADT values on roads in South Carolina. Origin-destination centrality was investigated, which includes internal-internal, internal-external, and external-external. These three parameters were then combined with three additional parameters, functional class, speed limit, and number of lanes, to produce a multiple linear regression model. It was determined that internal-internal centrality, external-external centrality, and speed limit would be the parameters in the final regression model. The final regression model was then compared to the traditional travel demand model, and it was observed that the regression model outperformed the travel demand model in terms of RMSE and R-squared value. The number of count stations used for input was also varied, and it was observed that using 60% or more of the count stations resulted in similar median absolute percent error

values. Therefore, the number of count stations used can be reduced by 40% without a loss of accuracy.

2.1.7 Point-Based Model

In 2018, Unnikrishnan et al. (2018) implemented a point-based model to estimate the AADT of roads in Oregon. The idea behind this method is to predict the AADT based on the number of “points” or roadway features a roadway has. The fewer the number of features a roadway has (e.g., left-turn lane, two-way left-turn lane, parking lot), the less traffic it is likely to carry, and vice versa. In their work, the point-based model assigns a region a set of roadway features that are each worth one point. The number of features that a roadway has will be the number of points associated with that roadway. The number of points that the roadway has was then related to the estimated AADT of that roadway. The region studied was divided into four separate areas, each having its own model and set of roadway characteristics. For local roads, the median error ranged from -16 to 151. The limitation of this model is the homogenous nature of local roads, which generally have the same features in an area. This means that the features will not vary enough in an area to reflect the trends in AADT.

2.1.8 Summary

Table 2-1 provides a summary of the studies reviewed. The study technique, study area, and reported error are shown. The error values are intended to provide a reference or benchmark for this study.

The following conclusions can be made from the above review.

- The performance of a methodological approach depends on the scope of the application (e.g., statewide vs small urban area) and data availability. In general, the recommendation is to develop models customized to various regions (e.g., urban vs rural, north vs south) rather than rely on a single statewide model.
- Multiple linear regression is the simplest technique for AADT estimation. This has led to numerous studies investigating the use of MLR, which makes it a very well documented method, and many advancements have been made through its utilization. Spatial regression models appear to perform better than multiple linear regression but are more complex to calibrate and may have transferability issues. Spatial approaches have not been tested or shown to perform well when transferred to other areas.

Table 2-1 Summary of literature review

Year	Author(s)	AADT Estimation Technique	Study Area	Reported Error
1998	Mohamad et al.	MLR	Indiana	MSE=16%
1999	Xia et al.	MLR	Florida	MPE=20%
2000	Seaver et al.	MLR	Georgia	R ² =0.27-0.94
2001	Zhao and Chung	MLR	Florida	R ² =0.818
2004	Zhao and Park	GWMLR	Florida	R ² =0.8756
2006	Anderson et al.	MLR	Alabama	R ² =0.819
2006	Eom et al.	K	North Carolina	MSPE=2.86
2008	Pan	MLR	Florida	MAPE=32-159%
2009	Castro-Neto et al.	SVR	Tennessee	MAPE=2.26%
2009	Wang and Kockelman	K	Texas	Median percent error=33%
2009	Zhong and Hanson	TD	New Brunswick	Average error=9-174%
2012	Lowry and Dixon	MLR	Idaho	R ² =0.72
2012	Pulugurtha and Kusam	GWMLR	North Carolina	MAPE=26-35%
2013	Selby and Kockelman	K	Texas	MPE=-6.5-3.9%
2013	Wang et al.	TD	Florida	MAPE=52%
2014	Lowry	C	Idaho	MdAPE=22-29%
2014	Yang et al.	MLR	North Carolina	R ² =0.6954
2015	Shamo et al.	K	Washington	RMSE=56.48-95.31
2015	Sun and Das	SVR	Louisiana	Percent within 100=63-100
2016	Apronti et al.	MLR	Wyoming	R ² =0.64
2016	Staats	MLR	Kentucky	MAPE=61-87%
2017	Keehan	C	South Carolina	R ² =0.8292
2018	Chang and Cheon	EM	Ulsan City	MAPE=7%
2018	Unnikrishnan et al.	EM	Oregon	Median error=-16-151
2020	Sfyridis and Agnolucci	SVR	Wales	MAPE=2-277%

MLR=Multiple linear regression, GWMLR=Geographically multiple linear regression, K=Kriging, SVR=Support vector regression, TD=Travel demand, C=Centrality, and EM=Emerging methods

- Kriging is similar to geographically weighted multiple linear regression in terms of complexity, but it has the advantage of not requiring additional data, unlike other types of models. This method is promising due to its simplicity and cost-effectiveness.
- Machine learning shows promise as a method for AADT estimation and has been shown to produce accurate results. However, its complexity makes it difficult to implement and it suffers from the “black-box” problem. It should be noted that the machine learning models have not been used to estimate AADT at non-coverage locations.
- Travel demand methods are theoretically sound; however, even for relatively small networks, the assignment step takes a long time to complete. Therefore, this type of models cannot be applied at the state level.

- The point-based model is simple to understand and is data driven. Also, its lookup table nature is easy to implement.

2.2 State-of-the-Practice on Non-Coverage AADT Estimation

As part of this study, an online survey was conducted to understand the state-of-the-practice in AADT estimation for non-coverage locations. The survey was distributed to other state DOTs on July 1, 2020. A total of 17 state DOTs responded to the survey.

The questions and responses are summarized below. The questions are numbered and shown in italics.

1. *Please indicate the method, technique, or procedure your agency uses to estimate AADT at non-coverage or out-of-network locations. At these locations, there is no recent history of past counts (within the last 10 years), and they are not near a station with recent counts (within the last 10 years). Check all that apply.*

Table 2-2 The method(s) being used to estimate AADT at non-coverage locations

Methods, technique, or procedure	No. of Responses	Percent of Responses
Multiple linear regression	5	27.8%
Visual estimation	4	22.2%
Geospatial method	2	11.0%
Nonlinear regression	1	5.6%
Default Values	4	22.2%
Spatial regression	1	5.6%
Travel demand	1	5.6%
Total	18	100%

As shown in Table 2-2, multiple linear regression is the most commonly used technique to estimate AADT at non-coverage locations. The next two most popular techniques are visual estimation and default value. These two methods have been shown to underestimate or overestimate the actual AADTs (Christian, 2021). These responses suggest that state DOTs do not have the manpower and resources to estimate non-coverage AADT.

2. *How satisfied are you with your current AADT estimation at non-coverage locations?*

Table 2-3 Satisfaction with current AADT method

Satisfaction Level	No. of Responses	Percent of Responses
5	2	14.3%
4	2	14.3%
3	5	35.7%
2	2	14.3%
1	3	21.4%
Total	14	100%

There were 14 responses to this question because three state DOTs indicated that they do not use any AADT estimation technique. The majority of the respondents rated their satisfaction as three or less, with one being unsatisfied and five being satisfied. This finding suggests that there is a need for an AADT estimation technique, one that is simple to implement.

3. *Is your agency using any tool to estimate AADT at non-coverage locations?*

Table 2-4 Use of tool to estimate AADT at non-coverage locations

Responses	No. of Responses	Percent of Responses
Yes	4	23.5%
No	13	76.5%
Total	17	100%

The majority of the respondents (76.5%) indicated that they do not use any tool to estimate AADT at non-coverage locations. This finding suggests that there is a lack of resources made available to state DOTs to accomplish this task.

4. *Would you be willing to share your tool with the SCDOT?*

Table 2-5 Willingness to share tool with SCDOT

Responses	No. of Responses	Percent of Responses
Yes	4	100%
No	0	0%
Total	4	100%

Four state DOTs responded to this question and all four indicated that they were willing to share the tool with the SCDOT. The methods used by these tools are default values based on function class and rural/urban classification, default values based on mobility index, geospatial, and linear regression. The tool that uses linear regression was deemed most appropriate for the SCDOT. A screenshot of this tool is shown in Figure 2-1. Upon further examination, it was found that the technique(s) used by the tool is similar to what the project team had intended to implement. Given that the goal of this project is to utilize easy to implement techniques and to provide the SCDOT with an easy-to-use tool, it was decided that a custom Excel-based tool would best meet the need of the SCDOT.

Edit Analysis Parameters [Close]

Use this form to edit the parameters for use in the current analysis.

Matrix Files

In: [] ... Out: [] ...

Regression Analysis

Minimum counts per record: [2] Multi-year projection span: [20] Min R-squared: [0] %

	<u>Default</u>	<u>Alternate</u>
Number of years' data to use:	[15]	[15]
Outlier detection:	[None] ▾	[None] ▾
Threshold (std dev/pct points to use)	[]	[]
Regression curve type:	[Linear] ▾	[Linear] ▾
Growth Percent Ranges (by AADT range):	[Edit]	

Estimation and Forecasting

Forecast method: [Median Actual AADT] ▾ Backup estimation method: [Straight line] ▾

When a group average growth factor is not available for a record:

Do not calculate estimated or forecast counts Use: [0.00] % growth

Use the global average growth factor Use the "nearest" group growth factor

Notes:

[Ok] [Cancel]

Figure 2-1. Non-coverage AADT estimation tool

CHAPTER 3: MODELS DEVELOPMENT

This chapter is composed of four parts. The first part provides information about the coverage and non-coverage datasets. The second part explains how the training dataset was prepared. The third part presents mathematical details underlying the developed models.

3.1 Data Description

3.1.1 Non-Coverage Counts Dataset

Prior to this project, the SCDOT did not have a list of non-coverage locations and did not have a procedure to identify them. The following procedure was developed in conjunction with the project steering committee to identify the stations:

1. Group road segments in each county into two categories, red and green, as follows. If there is a recent count (within the last 10 years) on a segment, then it is considered “green.” If there is not a recent count, then it is considered “red.”
2. Remove “red” segments that are less than 0.2 miles long and classified as dead ends.
3. Remove “red” segments that are classified as church, school, or cemetery driveways.
4. If there is a road that is comprised of both “red” and “green” segments, make the entire road “red.”
5. Remove “green” segments.
6. Combine connecting “red” segments and break up segments that are longer than five miles in a rural area or longer than two miles in an urban area into two segments.
7. The midpoints of the remaining segments are the locations that will be considered non-coverage count stations.

After the non-coverage stations were determined, the SCDOT provided a file that contained all of the required attributes for each location. These attributes include: a unique ID, latitude, longitude, and functional class. A map of the non-coverage locations is shown in Figure 3-1. Note that there are significantly more non-coverage locations than coverage. More than 90% of the non-coverage locations are urban (FC 18) and rural (FC 9) local roads. For this reason, it was decided in consultation with the committee that the to be-developed models should focus on predicting AADT of only urban (FC 18) and rural (FC 9) local roads at non-coverage locations.

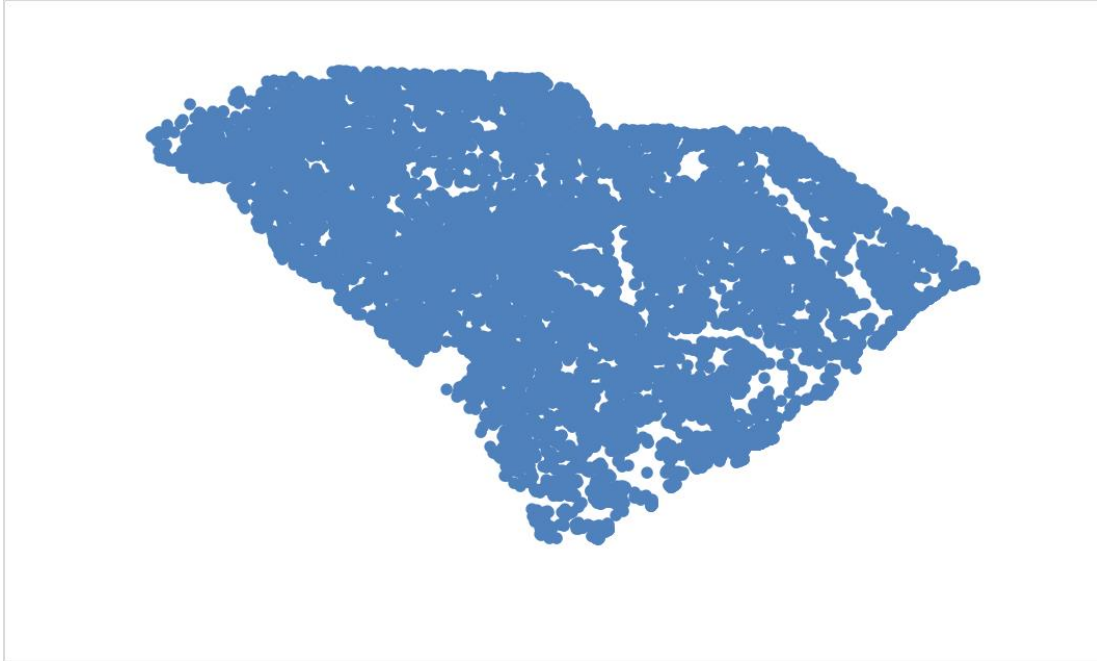


Figure 3-1 Map of non-coverage count stations

3.1.2 Coverage Counts Dataset

Two files were combined to obtain the necessary information for the coverage counts. The first file is the count station shapefile, which contained the AADT value, latitude, longitude, county, and Linear Reference System (LRS). The second file is the functional classification shapefile, which contained the functional class, latitude, longitude, county, and LRS attributes. These two files were joined by the LRS values. The coverage counts dataset comprised of seven attributes: station ID, AADT, latitude, longitude, functional class, county and LRS. A map of coverage counts' locations based on their latitudes and longitudes is shown in Figure 3-2.

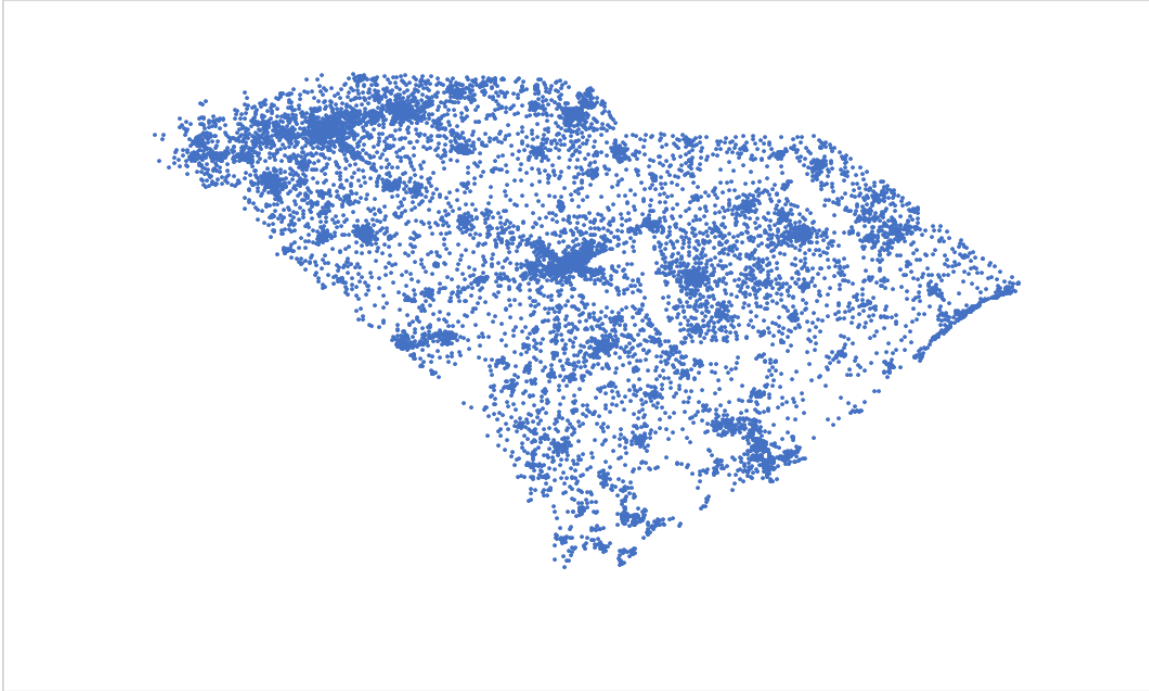


Figure 3-2 Map of coverage count stations.

Each of the attributes provided for the coverage counts is explained in the following.

- The ID is a seven-digit identifier that is unique to each station. The first two digits are the code for the county that the station is located in, and the remaining five digits denote the station number. For example, a station with an ID of 0200232 indicates that it is in county 02, which corresponds to Aiken County, and its station number is 00232, meaning it is the 232nd coverage station.
- The provided AADTs were not obtained from permanent count stations. They were obtained from short-term counts and expansion factors. In years when short-term counts were not collected, the current year's AADTs were estimated by multiplying the previous year's AADTs by a growth factor.
- The latitudes and longitudes are the GPS coordinates of the count stations. The latitudes and longitudes were converted to decimal degrees to facilitate computations.
- The functional class specifies the type of road, with each having a corresponding number as shown in Table 3-1. There are three major functional classes; arterial, collector, and local. The arterial group is divided into principal arterials and minor arterials, and the collector group is divided into major and minor arterials. The principal arterial group is further subdivided into interstates, freeways/expressways, and other. Each functional class is also divided into two groups: urban and rural.

Table 3-1 Functional class

Functional Classification			Functional Class Number	
Arterial	Principal Arterial	Interstate	Urban	1
			Rural	11

Functional Classification			Functional Class Number	
		Freeways & Expressways	Urban	6
			Rural	12
		Other	Urban	2
			Rural	13
	Minor Arterial	Urban	3	
		Rural	14	
Collector	Major Collector	Urban	4	
		Rural	15	
	Minor Collector	Urban	5	
		Rural	16	
Local		Rural	9	
		Urban	18	

- The county attribute contains the county name where the count station is located.
- The LRS is an 11-digit number that is used to describe a count station. The first two digits represent the county number, the next two represent the route type, the next five represent the route number, and the last two represent the route auxiliary. In addition, N or E is attached to the end to indicate the direction of the route (i.e., north and south or east and west).

Preliminary analysis of the coverage dataset showed that the number of count stations per functional class is not evenly distributed. Their distributions are shown in Figure 3-3. The unbalance number of counts per functional class could lead to a biased dataset where the estimated AADT is weighted more toward those with higher counts. Moreover, there is a large variation in the minimum, median, mean, maximum, and standard deviation of counts across the different functional classes as shown in Table 3-2. Note that the average AADT of an urban freeway is 14 times higher than that of an urban local road (i.e., 21,542 vs. 1,534). For these reasons, only the urban (FC 18) and rural (FC 9) local roads are retained in the coverage counts dataset when used to predict AADT of non-coverage locations.

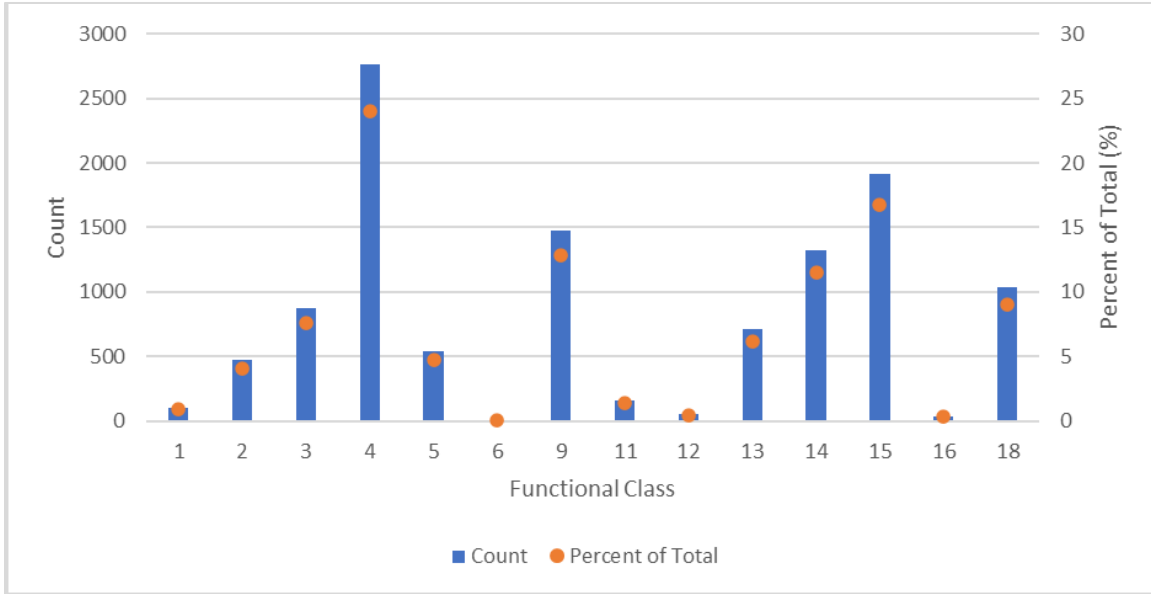


Figure 3-3 Distribution of coverage counts by functional class

Table 3-2 Summary statistics of AADT values for each functional class

Functional Class Number	Minimum AADT	Median AADT	Average AADT	Maximum AADT	Standard Deviation
1	225	42,800	42,990	120,200	17,059
2	250	6,000	8,470	111,200	9,076
3	25	3,500	4,884	56,200	5,187
4	25	1,050	1,943	63,800	3,393
5	25	325	743	11,900	1,381
6	7,400	22,800	21,542	44,600	10,884
9	25	550	1,534	80,500	4,483
11	125	73,800	74,475	176,500	33,729
12	1,000	26,300	28,533	60,200	14,966
13	175	18,000	20,627	97,900	12,962
14	75	8,100	10,472	61,000	8,109
15	25	2,600	3,977	38,800	4,103
16	75	1,200	2,224	12,700	2,571
18	25	1,250	2,746	83,000	6,347

3.2 Models Training Dataset

From the initial data exploratory analysis and findings, it was determined that AADTs from coverage urban (FC 18) and rural (FC 9) local roads will be used to predict AADT of urban (FC 18) and rural (FC 9) local roads at non-coverage locations. Table 3-3 shows a summary of the statistics between the two datasets. Even though only urban (FC 18) and rural (FC 9) local roads are considered in both datasets, the mean AADT of coverage counts is much higher than that of non-coverage counts. This discrepancy was resolved by first removing outliers from the coverage counts dataset, and second, by applying a reduction factor to the coverage counts. The reduction factor is determined by dividing the mean AADT of the coverage counts by the mean AADT of the known non-coverage counts and taking the integer value of this quotient. Once this reduction factor is determined, any coverage count higher than the maximum value of non-coverage counts times the reduction factor was removed. The remaining coverage counts were then divided by the reduction factor. The revised counts were then used by the kriging model to determine its optimal parameters (i.e., Euclidean distance, bin, nugget, range, partial sil, and weighted coefficients). Initial testing found that all models' performance improved if the training dataset was supplemented with known non-coverage data. For this study, 50% of the known non-coverage dataset was used to supplement the coverage dataset in evaluating the models' performance, and the same percentage is used by the Excel-based tool if the known non-coverage data is provided by the user (see Section three of Chapter five for additional discussion).

Table 3-3 AADT statistics of coverage and non-coverage counts

Counts	Coverage Counts	Non-coverage Counts
Average Value	1,750	233
Minimum Value	25	25
First Quartile	300	50
Median Quantile	700	100
Third Quartile	1,650	250
95% Quantile	6,215	850
Maximum Value	83,000	5,900

To develop the regression models and point-based model, several roadway features were collected, including the area where the roadway segment is located, its median type, the presence of an exclusive right-turn lane, the presence of an exclusive left-turn lane, the presence of a sidewalk on both sides of the roadway segment, and the presence of a parking lot. Table 3-4 shows the variables considered. The variable "Urban" is determined by the functional class provided by the SCDOT. Other remaining variables were collected using Google Earth.

Table 3-4 Roadway features collected for model development

Variable	Description of the Variable
Urban	"1" if the roadway segment is in an urban area; otherwise, "0"
Single Line	"1" if the type of the median of the roadway segment is a single line; otherwise, "0"
Other Type Median	"1" if the type of the median of the roadway segment is flush, raised, or two-way left turn lane (TWLTL); otherwise, "0"

Variable	Description of the Variable
Right-turn Lane	“1” if an exclusive right-turn lane on the roadway segment is present 1,000 feet upstream and downstream of the midpoint
Left-turn Lane	“1” if an exclusive left-turn lane on the roadway segment is present 1,000 feet upstream and downstream of the midpoint; otherwise, “0”
Sidewalk	“1” if a sidewalk on both sides of the roadway segment is present 1,000 feet upstream and downstream of the midpoint; otherwise, “0”
Parking Lot	“1” if a parking lot (e.g., pay to park, parking lots, and parking lots for schools, shopping centers, recreational facilities, and hospitals) adjacent to the roadway segment is present 1,000 feet upstream and downstream of the midpoint; otherwise, “0”

Table 3-5 to Table 3-10 present descriptive statistics of AADT by each roadway feature. These results indicate that urban local roads (FC 18) have a higher mean and median AADT than rural local roads (FC 9) (Table 3-5). Roads with flushed, raised, or TWLTL median have a higher mean and median AADT than undivided roads (Table 3-6). Roads with an exclusive right-turn lane have a higher mean and median AADT than roads without it (Table 3-7). Roads with an exclusive left-turn lane have a higher mean and median AADT than roads without it (Table 3-8). Roads with a sidewalk have a higher mean and median AADT than roads without it (Table 3-9). Roads with a parking lot have a higher mean and median AADT than roads without it (Table 3-10). These findings correspond to tuition in that these features are added as a result of forecasted or realized demand.

Table 3-5 Descriptive statistics of AADT by area

Area	Number of Observations	Mean AADT	Median AADT	Min AADT	Max AADT
Rural	2,339	185	75	4	13,900
Urban	1,851	350	164	4	19,183

Table 3-6 Descriptive statistics of AADT by median types

Median types	Number of Observations	Mean AADT	Median AADT	Min AADT	Max AADT
Undivided	1,347	103	61	4	2,200
Single line	1,885	265	136	4	3,886
Other types (e.g., flush, raised, TWLTL)	958	463	136	4	19,183

Table 3-7 Descriptive statistics of AADT by presence of an exclusive right-turn lane

Exclusive right-turn lane	Number of Observations	Mean AADT	Median AADT	Min AADT	Max AADT
Not present	3,974	230	107	4	19,183

Exclusive right-turn lane	Number of Observations	Mean AADT	Median AADT	Min AADT	Max AADT
Present	216	783	593	36	5,557

Table 3-8 Descriptive statistics of AADT by presence of an exclusive left-turn lane

Exclusive left-turn lane	Number of Observations	Mean AADT	Median AADT	Min AADT	Max AADT
Not present	3,852	200	100	4	19,183
Present	338	920	700	46	5,557

Table 3-9 Descriptive statistics of AADT by presence of a sidewalk on both sides of the roadway

Sidewalk	Number of Observations	Mean AADT	Median AADT	Min AADT	Max AADT
Not present	3,619	225	100	4	19,183
Present	571	470	243	4	5,557

Table 3-10 Descriptive statistics of AADT by presence of a parking lot

Parking lot	Number of Observations	Mean AADT	Median AADT	Min AADT	Max AADT
Not present	2,762	193	79	4	19,183
Present	1,428	384	221	4	5,557

3.3 Models

3.3.1 Kriging

Given a set of n data points with known information, the goal of kriging is to determine an estimate at an unknown location, which is shown in Figure 3-4. The known locations are represented by $Y(s_i)$, where s_i is a position vector that describes the location i . Since there are n known locations, i is in the range of 1 to n . The unknown location is represented by s_0 , and the estimate at that unknown location, $\hat{Y}(s_0)$, is determined by finding a linear combination of nearby known locations. There are multiple methods that use a linear combination of nearby known locations, but what makes kriging unique is its use of geostatistical methods to estimate weights to use for each utilized location. The weights are described by λ_i , which corresponds to location i .

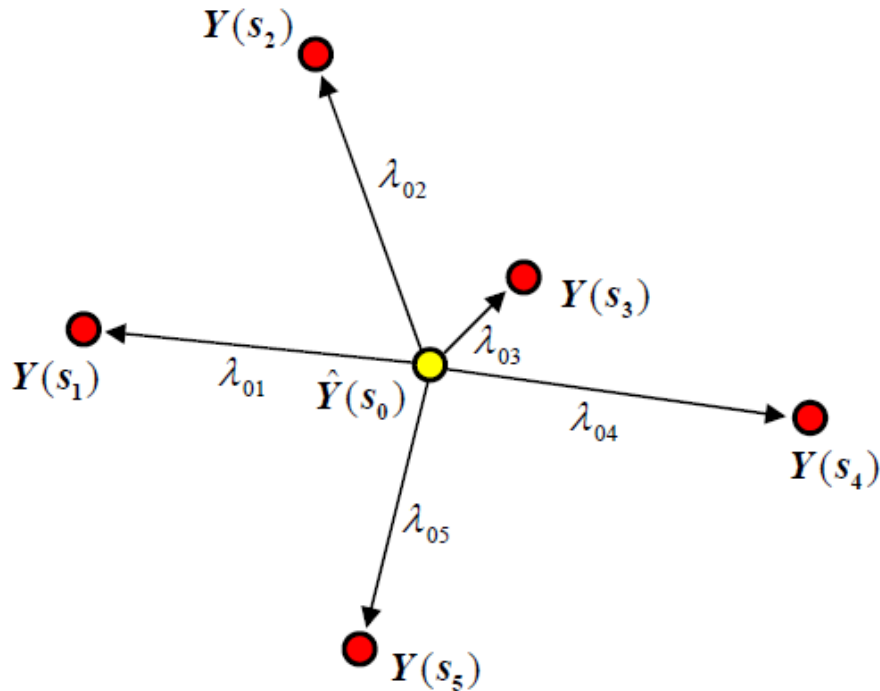


Figure 3-4 Illustration of kriging assigning weights to neighbors (Smith, 2020)

Kriging makes an estimation at an unknown location, $\hat{Y}(s_0)$, by using a linear combination of known values, $Y(s_i)$. This can be represented by the following equation.

$$\hat{Y}(s_0) = \sum_{i=1}^{n_0} \lambda_i * Y(s_i) \quad (3-1)$$

It is important to note that not every known coverage count/location will be utilized for the estimation, and therefore, the summation does not go to n , which is the number of locations, but instead goes to n_0 , which is the number of utilized neighbors for location s_0 . Kriging utilizes geostatistical methods to determine the weights, λ_i . There are multiple methods that can be used to determine these weights. A common approach is to use the inverse distance weighting. With inverse distance weighting, the weight for a location is based on the reciprocal of its distance from the unknown location. This method places a higher weight on known coverage counts/locations that are closer to the unknown location. In general, this makes sense, but it can cause problems if there is an outlier that is close to the unknown location. Kriging solves this problem by using geostatistical methods.

3.3.1.1 Semivariogram

Kriging uses the covariance between locations to determine how much weight should be given to each utilized neighbor. To calculate the covariance, the semivariogram is used. A semivariogram describes the relationship between the squared difference of two locations and the distance between them. There are three key concepts shown in a semivariogram. The first is called the nugget and refers to the squared difference at a distance of zero. Since the squared distance is not

zero at this location, it implies that the measuring the AADT at a location multiple times will result in different values. This is reasonable because there is variability in AADT measurement, and the nugget is a representation of that variability. The next concept is the range, and that is the distance where the semivariogram goes from increasing to remaining constant. Physically, this implies that after a certain distance the new AADT value that is measured can only have a maximum difference from the original location, and the distance required for this to take place is the range. Lastly, the partial sil is the difference between the maximum squared difference and the nugget. This is just a representation of the maximum squared difference between two locations. Instead of the partial sil, the sil could be used, which is the sum of the nugget and partial sil and is equal to the maximum squared difference. Usually, the partial sil is used because it allows the effects of the nugget to be taken into account. Instead of reporting the maximum squared difference, which is equal to the sil, the partial sil reports the maximum squared difference minus the nugget. An example semivariogram is shown below in Figure 3-5 to illustrate these concepts. In Figure 3-5, the nugget is equal to 0.2, the range is equal to 0.5, the partial sil is equal to 0.8, and the sil is equal to 1.0.

The procedure for constructing a semivariogram is as follows. Imagine recording the AADT at a location on a roadway, and then moving a distance d away from the starting location. Now, measure the new AADT value and compute the squared difference between them. By doing this for multiple distances, a plot could be developed that looks similar to that shown in Figure 3-5. However, there is not only one location that is a distance d from the original position. A circle with radius d could be drawn around the original point, and any value on that circle would be a distance d from the original location. Therefore, for the squared difference at a distance d to be represented by one value, an average is taken of all squared differences a distance d from the original location. When performing this on actual data, every pair of locations will have a distance between them and their squared AADT difference is calculated. It is not likely for there to be many pairs of locations that have the exact same distance between them. Therefore, binning is used to determine the average squared difference for a bin, γ , and the distance that represents that bin is its midpoint. This set of squared differences and distance is referred to as an empirical semivariogram. To calculate the squared difference between the distances of an empirical semivariogram, a semivariogram model is fitted to the empirical semivariogram.

There are multiple semivariogram models that can be fitted to an empirical semivariogram. Four of the most commonly used in the literature are the Gaussian model, exponential model, spherical model, and linear model. These models have the same three parameters, the nugget, range, and partial sil. Therefore, the theoretical models could be fitted to the empirical semivariogram by optimizing the parameters such that they minimize some error criteria. After determining which model best fits the empirical semivariogram, the squared difference can be calculated by inputting the distance between two locations and the optimized parameters.

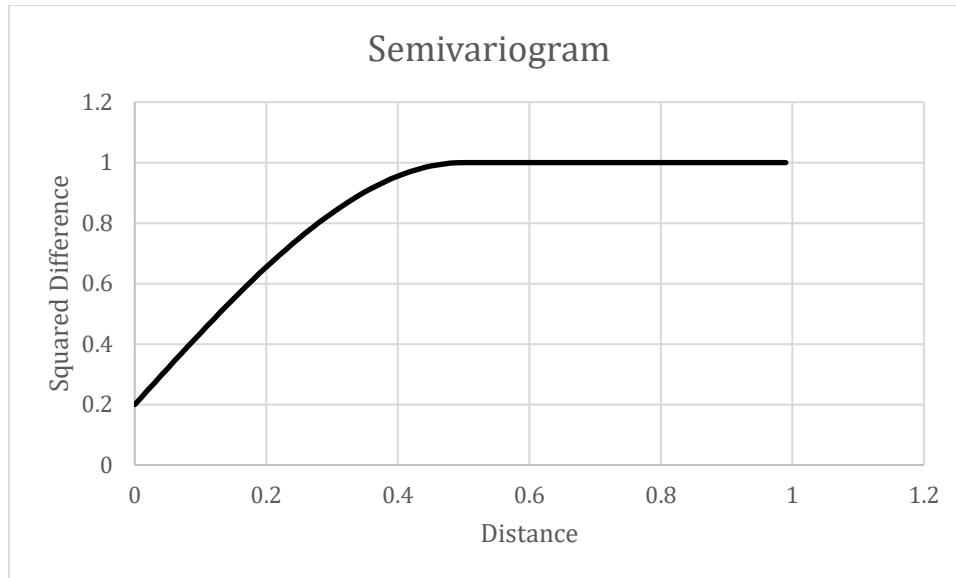


Figure 3-5 Semivariogram example

3.3.1.2 Weighted Determination

The weights utilized in the kriging method are determined by utilizing the semivariogram. The semivariogram is related to another important function called the covariogram shown the following equation (Equation 3-2). From the relationship between the covariogram and semivariogram, it is evident that the covariogram represents the relationship between the similarity between two locations to the distance between them. As the semivariogram increases, the covariogram decreases. Therefore, the covariogram starts at a maximum value and then decreases to a minimum.

$$C(d) = nug + ps - \gamma(d) \quad (3-2)$$

Since the covariogram represents the similarity between two locations, it can be used to calculate the covariance between two locations. Using the covariogram, the following matrices could be determined. The first is a matrix of covariances between the unknown location and every utilized neighbor, c_u . The second is a matrix of covariances between every pair of utilized neighbors, c_k . Using these matrices, the optimum weights are determined with the following equation.

$$\lambda = c_k^{-1}c_u \quad (3-3)$$

After determining these weights, the AADT at the non-coverage location could be determined.

3.3.1.3 Implementation of Kriging Model

To develop the semivariogram for the kriging model, the first step is to read the coverage count data into a matrix. The latitude is read into a vector, x , longitude is read into y , the logarithm of AADT is read into z , functional class is read into FC , and unique ID is read into IDs . Next, the distance between every coverage count station, as well as the square difference in AADT between every coverage count station, is calculated and stored in a matrix called *data*.

It is important to note how the distances between two locations were calculated. First, a Euclidean distance was used instead of a network distance because of the complexity involved in calculating network distances and the lack of increased accuracy over Euclidean distance (Selby & Kockelman, 2013). However, even calculating the Euclidean distance between two pairs of latitude and longitude can be complex. Given a pair of latitude and longitude, the distance between the two follows the curvature of the earth. Since the earth's radius is not constant, to determine the exact distance would be improbable. A simplifying assumption would be to assume that the earth's radius is constant. Doing so would allow the calculation of the great circle distance, which is the shortest distance between two points on a sphere. The first step is to calculate the angle between the pairs of latitude and longitude, $\Delta\sigma$, using Equation 3-4, where all angles are in radians, and then the great circle distance could be calculated using Equation 3-5 .

$$\Delta\sigma = \arccos(\sin(x_1) \sin(x_2) + \cos(y_1) \cos(y_2) \cos(y_2 - y_1)) \quad (3-4)$$

$$d_{gc} = r_e \Delta\sigma \quad (3-5)$$

A problem with this formulation is that the pairs of latitude and longitude are close enough to cause rounding errors. Also, calculating the angle between each pair and then the corresponding great-circle distance for approximately 2,500 locations would require much computational power and time. To simplify these calculations, instead of using a linear distance such as miles, the distance in terms of degrees is used.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (3-6)$$

After computing the distance, using Equation 3-6, the squared difference between every pair of coverage count stations' AADT values is determined with Equation 3-7.

$$\delta = (AADT_i - AADT_j)^2 \quad (3-7)$$

The matrix of distances and squared differences is then sorted so the smallest distance is at the lowest index and each following distance is the next smallest. In other words, the matrix is sorted from smallest to largest by distance. This sorting is needed to be performed efficiently, because if there are n known data points, calculating the distance and squared difference between every pair of coverage count stations would result in $(n^2 - n)/2$ distances and squared differences. Since n is approximately 2,500, the resulting matrix will have approximately 3,125,000 rows. Sorting this in $O(n^2)$ time would take too long for the Excel tool to be used practically. Therefore, to sort this matrix in an adequate amount of time, the quicksort algorithm is used.

The empirical semivariogram is then developed by choosing a number of bins and calculating the average squared difference for each bin. This average squared difference is equal to the empirical semivariogram at the midpoint of the bin. The average squared difference and bin midpoint are stored in a matrix called *hist*. A plot of the empirical semivariogram is shown below.

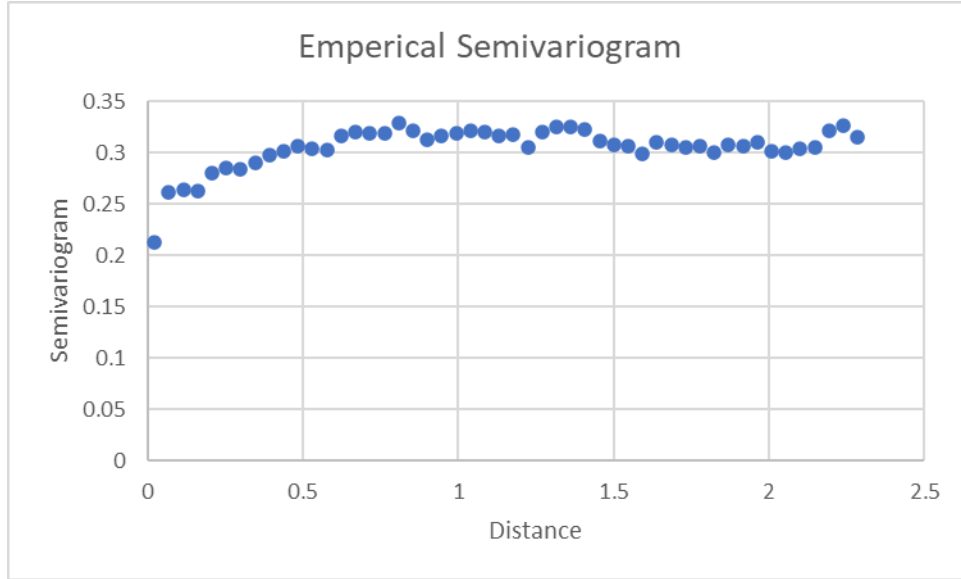


Figure 3-6 Example empirical semivariogram

After defining the empirical semivariogram, the next step in the subroutine is to fit a semivariogram model to the empirical semivariogram. There are four semivariogram models that are fitted to the empirical semivariogram by minimizing the sum of the squared errors. These included a Gaussian semivariogram, exponential semivariogram, spherical semivariogram, and linear semivariogram. These models are represented with the following set of equations.

$$\gamma_g(d) = nug + ps * \left(1 - \exp\left(-\frac{d}{r}\right)^2\right) \quad (3-8)$$

$$\gamma_e(d) = nug + ps * \left(1 - \exp\left(-\frac{d}{r}\right)\right) \quad (3-9)$$

$$\gamma_s(d) = \text{Min}\left(nug + ps * \left(1.5\left(\frac{d}{r}\right) - 0.5\left(\frac{d}{r}\right)^3\right), nug + ps\right) \quad (3-10)$$

$$\gamma_l(d) = \text{Min}\left(nug + d\left(\frac{ps}{r}\right), nug + ps\right) \quad (3-11)$$

Excel's Solver function is used to adjust the nugget, partial sill, and range parameters to minimize the sum of squared error for each theoretical semivariogram model. An example of each optimized semivariogram model is shown in Figure 3-7.

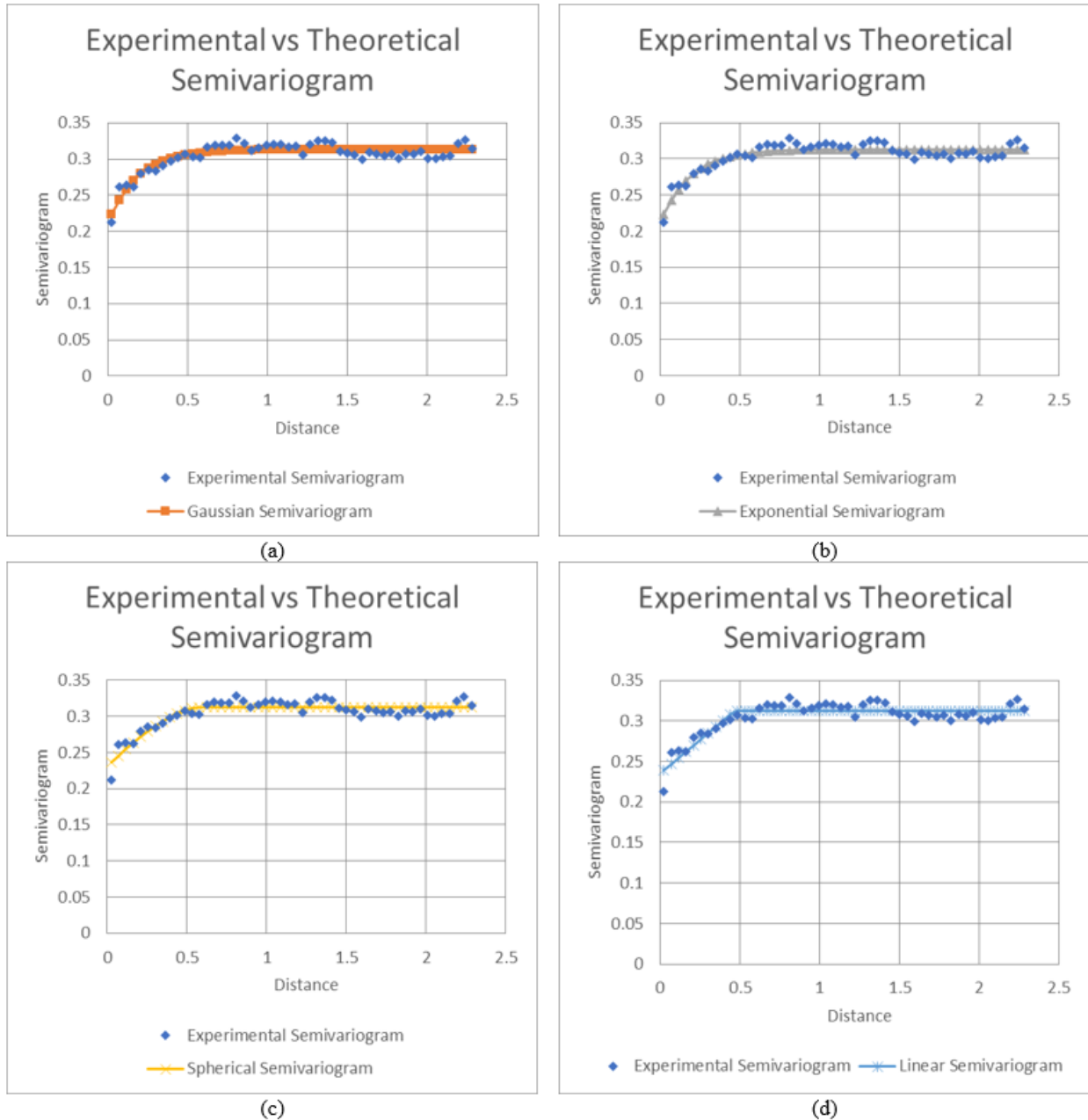


Figure 3-7 Comparison of different theoretical semivariogram models: (a) gaussian model, (b) exponential model, (c) spherical model, and (d) linear model

After optimizing each model, the model with the lowest sum of squared error is chosen as the best model.

The following procedure is used to calculate the AADT of a sampled dataset, which is a subset of the coverage count dataset. This procedure is then performed for every sampled dataset to calculate all unknown AADT values. The first step in the procedure is to determine the coverage count locations that would be utilized by the kriging model. This is done by calculating the Euclidean distance between the sampled dataset location and a coverage location. If the distance is less than the range of the optimized model from step one, that coverage count station would be

considered a neighbor of the sampled dataset location. This is repeated for the distances between the sampled dataset location and every coverage location. Next, the distances between the sampled dataset location and its neighbors are then sorted, using the quicksort algorithm, and the smallest N neighbors are utilized in the kriging model. If there is less than N neighbors, then all of the neighbors are utilized.

After determining the utilized neighbors, two covariance matrices are required to determine the weights for the utilized neighbors. The first covariance matrix, c_k , contains the covariance between each pair of utilized neighbors. In c_k , the element in row i and column j is the covariance between utilized neighbors i and j . The second covariance matrix, c_u , contains the covariance between the non-coverage location and each utilized neighbor. In c_u , the element in row i is the covariance between the non-coverage location and utilized neighbor i . The covariance between two locations is determined by utilizing the covariogram. After determining the model semivariogram, the covariance is calculated with Equation 3-12. Therefore, to determine the covariance between two locations, only the model semivariogram and the distance between those two locations are required.

$$C(d) = nug + ps - \gamma(d) \quad (3-12)$$

Once the two covariance matrices are determined, the kriging weights are determined using Equation 3-13.

$$\lambda = c_k^{-1}c_u \quad (3-13)$$

After calculating the kriging weights, each weight is normalized with respect to the sum of absolute value of the weights as shown in Equation 3-14. It is observed that the kriging weights would sum to unity, but each individual weight's value would range drastically. For example, one weight could be 0.13 while the next weight could be -23. The estimated AADT values are sensitive to these large weights because they could cause dramatic overestimation, to the point where Excel would show an error stating that the value is too large to be calculated. Therefore, it is necessary to normalize the weights to prevent such overestimation.

$$\lambda_{norm,i} = \frac{\lambda_i}{\sum_{i=1}^N abs(\lambda_i)} \quad (3-14)$$

After obtaining the normalized kriging weights, the sampled dataset AADT estimate is determined by applying each normalized weight to its respective AADT value, which is shown in Equation 3-15.

$$AADT = \sum_{i=1}^N \lambda_i * AADT_i \quad (3-15)$$

After calculating the estimated AADT, it is rounded to the nearest whole number. If it is less than 25 vpd, it is rounded up to 25 since that is the minimum AADT the SCDOT would use. If the

AADT is between 25 and 500, it is rounded to the nearest 25, between 501 and 2,000, it is rounded to the nearest 50, and if it is over 2,000, it is rounded to the nearest 100. This procedure is used to estimate the AADT of a sampled dataset location. It is then repeated for every non-coverage location.

After determining all AADT of the sampled coverage locations, their absolute errors are ranked from the smallest to largest. An example is shown in Figure 3-8.

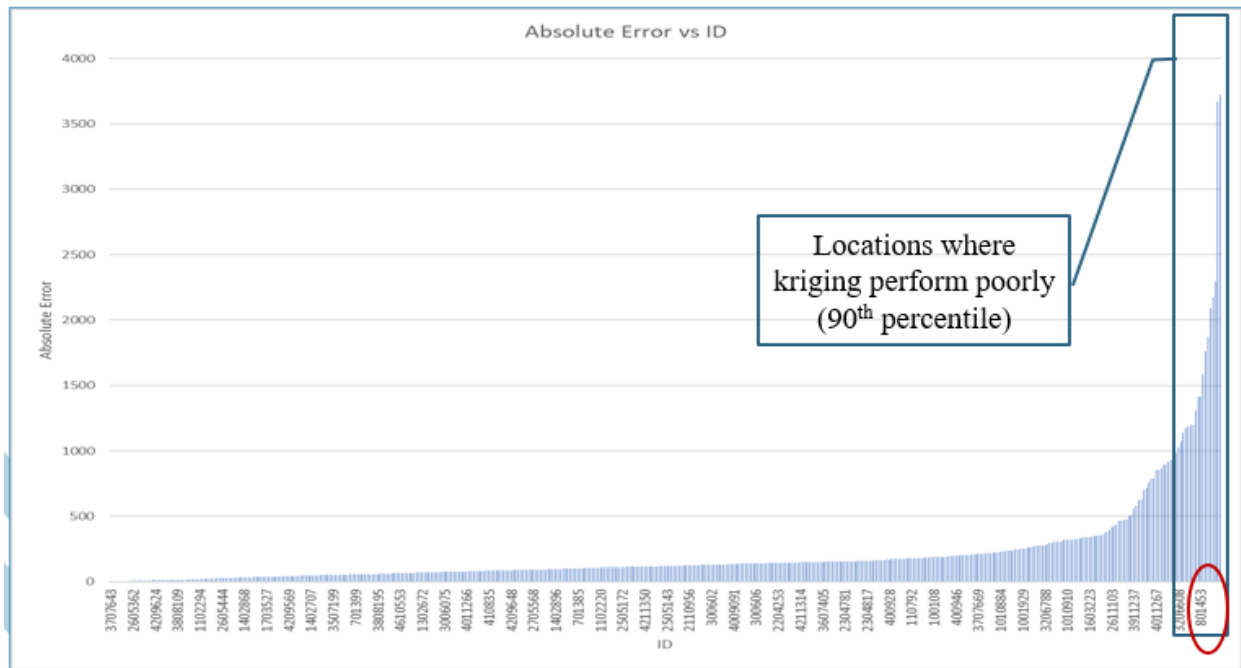


Figure 3-8 Sampled locations and their absolute errors

If the absolute error of the sampled location is above the user-specified error percentile, all non-coverage locations within 0.9 degrees of that coverage location will take on the mean AADT of all coverage counts in that county and corresponding functional class. Figure 3-9 illustrates that all non-coverage locations within 0.9 degrees of the coverage station 801453 will use the average AADT instead of the kriging-predicted value. For all other non-coverage locations, their AADTs will use the kriging predicted values. The term “hybrid kriging” is used hereafter to refer to the approach implemented, where the average AADT is used in place of the kriging-predicted value when there is evidence that a coverage location may not have an accurate AADT.

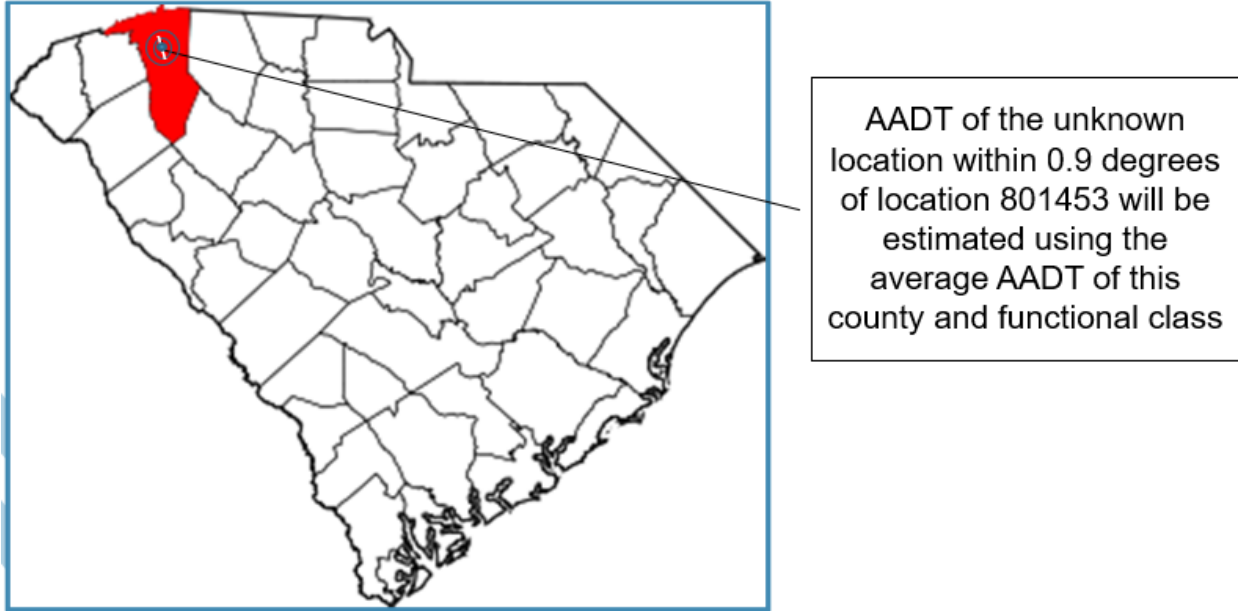


Figure 3-9 Illustration of scenario when an average AADT is used instead of kriging-predicted AADT

3.3.2 Point-Based Model

The point-based model used in this study is based on a study conducted by the Oregon Department of Transportation (Unnikrishnan, Figliozzi, Moughari, & Urbina, 2018). It uses the median AADT as the predicted AADT based on the number of points the roadway has; points are roadway features, some of which are shown in Table 3-4. Roadway features are collected on 4,701 urban (FC 18) and rural (FC 9) local roads, 4,189 of which are used to develop the model and 512 are used to validate the model. These roads are then grouped by the number of points they have. Within each group, the median AADT is calculated. The median AADT is used as the predicted value.

3.3.3 Regression Models

3.3.3.1 *Regular Regression Model*

The regular regression model explores the relationship between a scalar response and one or more explanatory variables. The standard form of the regular regression model is as follows:

$$y_{pred} = b_0 + b_1X_1 + b_2X_2 + \dots + b_iX_i \quad (3-16)$$

where y_{pred} is the predicted or expected value of the dependent variable, X_1 through X_i are distinct independent or predictor variables, b_0 is the value of Y when all the independent variables (X_1 through X_i) are equal to zero, b_1 and through b_i are the estimated regression coefficients.

3.3.3.2 Quantile Regression Model

The quantile regression model is more robust against outliers in the response variable compared to the regular regression model. The general quantile regression model can be described by the following equation.

$$y_{pred} = b_0(q) + b_1(q)X_1 + b_2(q)X_2 + \dots + b_i(q)X_i \quad (3-17)$$

where, y_{pred} is the predicted or expected value of the dependent variable, X_1 through X_i are i distinct independent or predictor variables, b_0 is the value of Y when all the independent variables (X_1 through X_i) are equal to zero, b_1 and through b_i are the estimated regression coefficients associated with q^{th} quantile. This study used the 50th quantile for the quantile regression model.

CHAPTER 4: RESULTS

To compare the performance of the developed models, the Root Mean Squared Error (RMSE) is used. RMSE gives the square root of the average of squared differences between actual values and predicted values as shown in the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - x_i)^2}{n}} \quad (4-1)$$

where,

y_i = predicted value of the i^{th} observation,

x_i = observed values of the i^{th} observation,

n = number of observations

4.1 Use of Default Value

The SCDOT currently uses default values to estimate AADT at non-coverage locations based on their functional class. The current default value for rural local roads (FC 9) is 100, and the current default value for urban local roads (FC 18) is 200. For this project, the SCDOT collected counts on 1,024 non-coverage urban (FC 18) and rural (FC 9) local roads between late 2019 and early 2021. The AADT statistics for these roads are shown in Table 4-1. These results indicate that the current default values are lower than the actual AADT values.

Table 4-1 Statistics of rural and urban AADT values for non-coverage counts

Statistics	Rural Local Roads (FC 9)	Urban Local Roads (FC 18)
Size	320	704
Mean	170	262
Minimum Value	25	25
First Quartile	50	75
Median Value	125	125
Third Quartile	250	275
Maximum Value	2,200	5,900

Since coverage counts are readily available, they can be used to obtain the default values. That is, a statewide average could be obtained annually for rural (FC 9) and urban (FC 18) local roads and then divide those values by the reduction factor; a factor of six is found based on the provided coverage and known non-coverage datasets. This approach would yield a default value of 154 for rural local roads (FC 9) and 175 for urban local roads (FC 18). Compared to the SCDOT's current default values, use of these values would provide a 2.11% improvement in terms of RMSE based on the validation set. Instead of using a default value for the entire state, a default value could also be used for each county; these values are shown in Table 4-2 and they are calculated by taking county averages and then dividing them by six. Using county-based default values would provide a 2.63% improvement over the current values in terms of RMSE.

Table 4-2 Default AADT values for counties

County	Default Rural (FC 9) AADT	Default Urban (FC 18) AADT
ABBEVILLE	84	115
AIKEN	265	265
ALLENDALE	31	30
ANDERSON	275	185
BAMBERG	114	114
BARNWELL	182	182
BEAUFORT	375	425
BERKELEY	621	1,025
CALHOUN	77	75
CHARLESTON	790	600
CHEROKEE	275	325
CHESTER	175	75
CHESTERFIELD	100	211
CLARENDON	105	105
COLLETON	205	170
DARLINGTON	120	265
DILLON	178	340
DORCHESTER	530	810
EDGEFIELD	100	381
FAIRFIELD	164	222
FLORENCE	500	290
GEORGETOWN	300	326
GREENVILLE	325	495
GREENWOOD	350	110
HAMPTON	50	75
HORRY	620	563
JASPER	190	190
KERSHAW	275	260
LANCASTER	253	260
LAURENS	342	164
LEE	60	63
LEXINGTON	365	623
MARION	182	135
MARLBORO	83	155
MCCORMICK	60	55
NEWBERRY	161	295
OCONEE	155	404
ORANGEBURG	127	405
PICKENS	351	433

County	Default Rural (FC 9) AADT	Default Urban (FC 18) AADT
RICHLAND	725	691
SALUDA	130	130
SPARTANBURG	322	370
SUMTER	200	630
UNION	70	282
WILLIAMSBURG	81	215
YORK	325	440

4.2 Kriging Model

The implemented hybrid kriging model allows the user to specify the absolute error threshold. When a sampled coverage location has an absolute error above this threshold, then all non-coverage locations within a certain radius of that coverage station will use a default value. The default value is the mean AADT based on county and functional class. Table 4-3 shows the RMSE for different radii with the absolute error set at 90th. As shown, a radius of 0.9 degrees resulted in the lowest RMSE. For this reason, 0.9 degrees is used in the tool for estimation.

Table 4-3 Effect of radius on kriging model performance

Absolute Error Threshold (percentile)	Radius (degrees)	RMSE
90	0.1	375
90	0.2	375
90	0.3	373
90	0.4	371
90	0.5	369
90	0.6	367
90	0.7	367
90	0.8	361
90	0.9	357
90	1.0	362
90	1.1	368

* Size of training dataset is 3,677; size of testing dataset is 1,024.

Table 4-4 shows the effect of changing the absolute error threshold. A threshold of 90th percentile resulted in the lowest RMSE. For this reason, it is set as the default value. However, the user can change it to whatever value is deemed appropriate.

Table 4-4 Effect of absolute error threshold on kriging model performance

Absolute Error Threshold (percentile)	Radius (degrees)	RMSE
95	0.9	352

Absolute Error Threshold (percentile)	Radius (degrees)	RMSE
90	0.9	347
85	0.9	349

* Size of training dataset is 3,677; size of testing dataset is 1,024.

The implemented hybrid kriging model also allows the user to supplement the coverage counts data with known non-coverage counts. It can be seen in Table 4-5 that when the training dataset is supplemented with known non-coverage data, it resulted in an improvement in RMSE from 19.79% to 21.37% over the current default value method.

Table 4-5 Effect of including known non-coverage counts

Dataset	Coverage counts only	Coverage counts with non-coverage counts
Size of training dataset	3,677	4,189
Size of testing dataset	1,024	512
Threshold (percentile)	90	90
Reduction factor	6	6
RMSE	304	215
Improvement over current default value method	19.79%	21.37%

4.3 Point-Based Model

Features from a total of 4,189 urban (FC 18) and rural (FC 9) local roads are collected; they include all locations from the list of coverage stations and all locations from the list of known non-coverage stations. Figure 4-1 shows the locations of these roads. As explained previously, the coverage counts are grouped by the number of points or features they have in common and the median AADT in each group are used as the predicted value. The features collected at unknown non-coverage locations are used to validate the point-based model. Table 4-6 shows the results of the point-based model using SCDOT data. It is essentially a lookup table. That is, given a number of points a road has, there is a corresponding predicted AADT. For example, a road with zero points is predicted to have an AADT of 125 vpd, and a road with three points is predicted to have an AADT of 650 vpd.

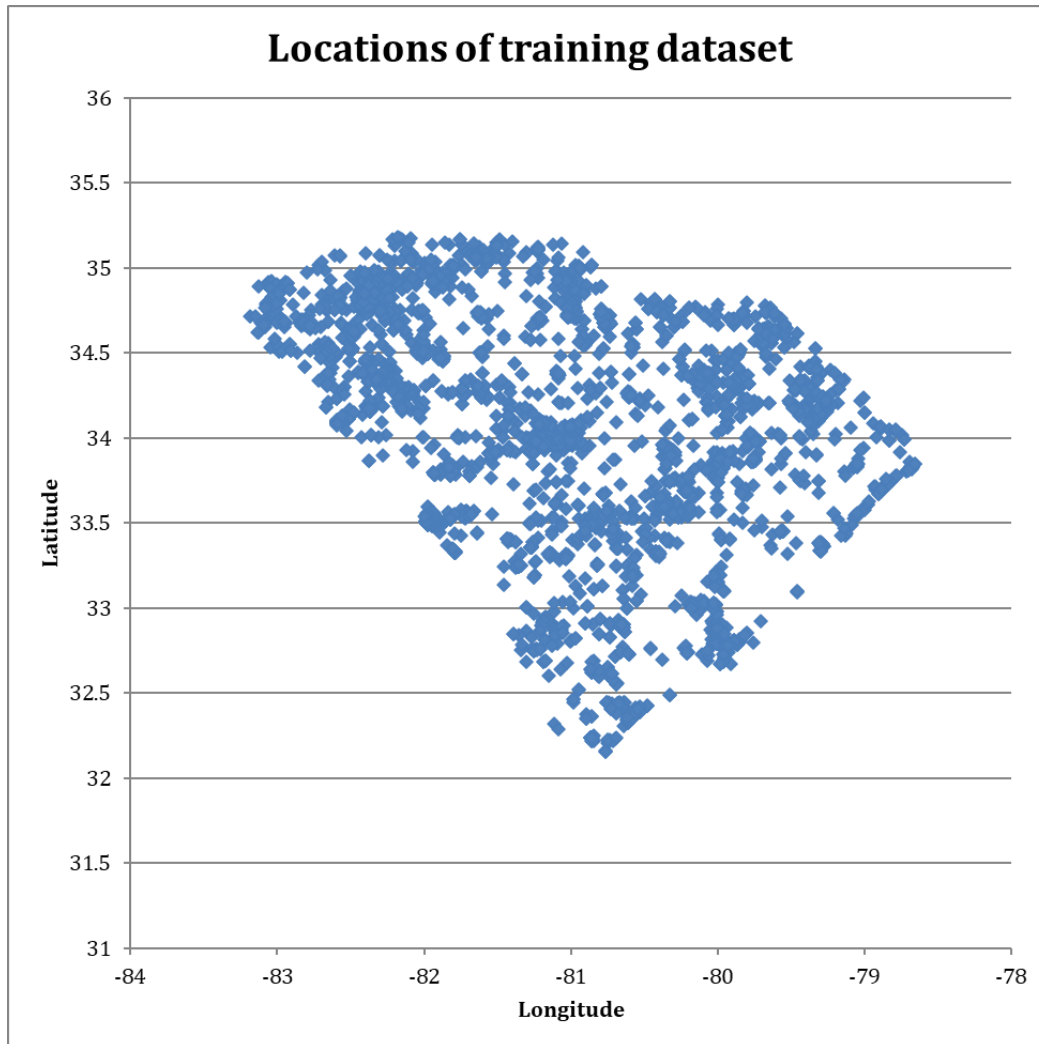


Figure 4-1 Locations of stations used for model training

Table 4-6 AADT prediction by the point-based model

Point	Predicted AADT	Description of Each Point Level
0	125	A roadway segment contains none of the seven variables shown in Table 3-4
1	175	A roadway segment contains one of the seven variables shown in Table 3-4
2	350	A roadway segment contains two of the seven variables shown in Table 3-4
3	650	A roadway segment contains three of the seven variables shown in Table 3-4
4	900	A roadway segment contains four of the seven variables shown in Table 3-4
5	1,600	A roadway segment contains five of the seven variables shown in Table 3-4

Point	Predicted AADT	Description of Each Point Level
6/7	1,800	A roadway segment contains at least six of seven variables shown in Table 3-4.

4.4 Regression Model

4.4.1 Regular Regression Model

Table 4-7 shows the regular regression model estimation results. Only the statistically significant variables are shown. To be statistically significant at the 0.05 significance level, their t-values need to be greater than 1.96 or less than -1.96. This implies that their p-values must be less than 0.05, which can be verified in the last column. All variables have a positive sign, which suggest that the presence of these features will increase the AADT. Their coefficients represent the increase in AADT. For example, the AADT is increased by 110 if the urban (FC 18) or the rural (FC 9) local road is located in an urban area versus rural area. Similarly, the AADT is increased by 113 if the urban (FC 18) or the rural (FC 9) local road has a double yellow line versus no centerline marking. If a road has none of these features, it is estimated to have an AADT of 40; using the SCDOT rounding procedure, it would be rounded to 50.

Table 4-7 Regular regression model estimation results

Variable	Estimate	Std. Error	t-value	p-value
(Intercept)	40	18.47	2.143	0.032134
Urban	110	19.19	5.724	1.11E-08
Double Yellow Line	113	21.31	5.314	1.13E-07
Other Type Median	249	25.43	9.784	< 2e-16
Right-turn Lane	158	47.03	3.35	0.000816
Left-turn Lane	539	39.71	13.582	< 2e-16
Sidewalk	66	28.55	2.296	0.021745

4.4.2 Quantile Regression Model

Table 4-8 shows the quantile regression model estimation results. Only the statistically significant variables are shown (i.e., those with p-values < 0.05). Similar to the regular regression model, all coefficients are positive. However, their coefficients are different. For example, this model predicts that a non-coverage road located in an urban area adds only 50 more vpd compared to 110 predicted by the regular regression model. Parking lot is found to be statistically significant in this model. Its presence is estimated to add 50 more vpd to the AADT. If a road has none of these features, this model predicts the AADT to be 25, which corresponds the minimum AADT the SCDOT would report.

Table 4-8 Coefficients for the quantile regression model

Variable	Estimate	Std. Error	t-value	p-value
(Intercept)	25	2.2	13.2	0
Urban	50	4.8	10.4	0

Variable	Estimate	Std. Error	t-value	p-value
Double Yellow Line	75	4.9	13.0	0
Other Type Median	50	5.5	7.6	0
Right-turn Lane	275	36.7	7.8	0
Left-turn Lane	450	65.3	6.9	0
Sidewalk	25	12.8	2.8	0.0049
Parking Lot	50	6.5	7.6	0

4.5 Comparison of Models Performance

Figure 4-2 shows the performance of the different models in terms of RMSE using a validation dataset consisting of 512 non-coverage locations. As shown, using the current default values resulted in an RMSE of 276. Using the hybrid kriging model reduced the RMSE to 217, a 21.37% improvement in terms of RMSE. The point-based model yielded an improvement of 22.28% compared to the current default value method, whereas the regular regression model yielded a 17.03% improvement, and the quantile regression model yielded a 23.19% improvement.

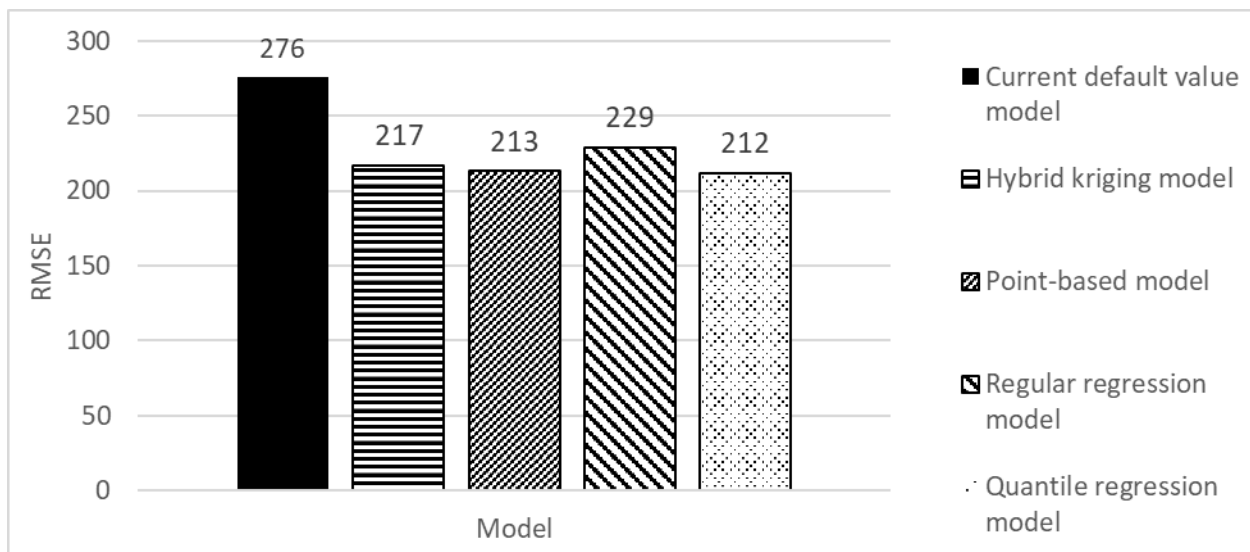


Figure 4-2 Comparison of models' performance

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

Currently, the SCDOT assigns a default value based on the functional class of the non-coverage road. If the road is a rural local road, then the default value is 100 vpd, and if the road is an urban local road, then the default value is 200 vpd. If the SCDOT practice requires the continual use of the default value method, then it is found that it would be better in terms of RMSE if a statewide average AADT is used. By using a default value of 154 vpd for rural local roads (FC 9) and 175 vpd for urban local roads (FC 18), it would lower the RMSE from 379 to 371. Instead of using statewide default value, having a separate default value for each county as shown in Table 4-2 would lower the RMSE further to 369. In summary, using statewide average value would lead to a 2.11% improvement, whereas using a county-based average value would lead to a 2.63% improvement over the current default values.

This project identified suitable methods to estimate AADT at non-coverage locations in terms of ease of implementation and accuracy. These methods include kriging, point-based model, regular regression model, and quantile regression model. The kriging model was selected as the primary model because it leverages existing coverage counts and does not require the SCDOT to collect additional data. Other models were also developed to complement the kriging model. Compared to the SCDOT's current default value method, the hybrid kriging model yielded a 21.37% improvement, the point-based model yielded a 17.03% improvement, the regular regression model yielded a 17.03% improvement, and the quantile regression model yielded a 23.19% improvement. The use of the point-based model, regular regression model and quantile regression model requires the collection of roadway features: location (urban or rural), presence of centerline marking (double yellow line), presence of median, presence of right turn lane, presence of left turn lane, presence of parking lot adjacent to the study road segment, and presence of sidewalks.

5.2 Recommendations

Based on this project's findings, it is recommended that the SCDOT consider adopting the developed Excel-based tool. A 21.37% improvement in terms of RMSE can be expected with the use of the kriging model. When roadway features are available for non-coverage roads, the SCDOT could change the configurable parameter in the tool to use estimates from the point-based model (a 1.45% improvement over kriging) or the quantile regression model (a 1.82% improvement over kriging).

5.3 Implementation

An Excel-based tool was developed as part of this project to assist the SCDOT in utilizing the developed models. Figure 5-1 shows a screenshot of the user interface. Running the tool simply involves clicking on the buttons in the sequence indicated and providing the necessary data files. Sample data files are provided along with the Excel-based tool which has VBA codes embedded.

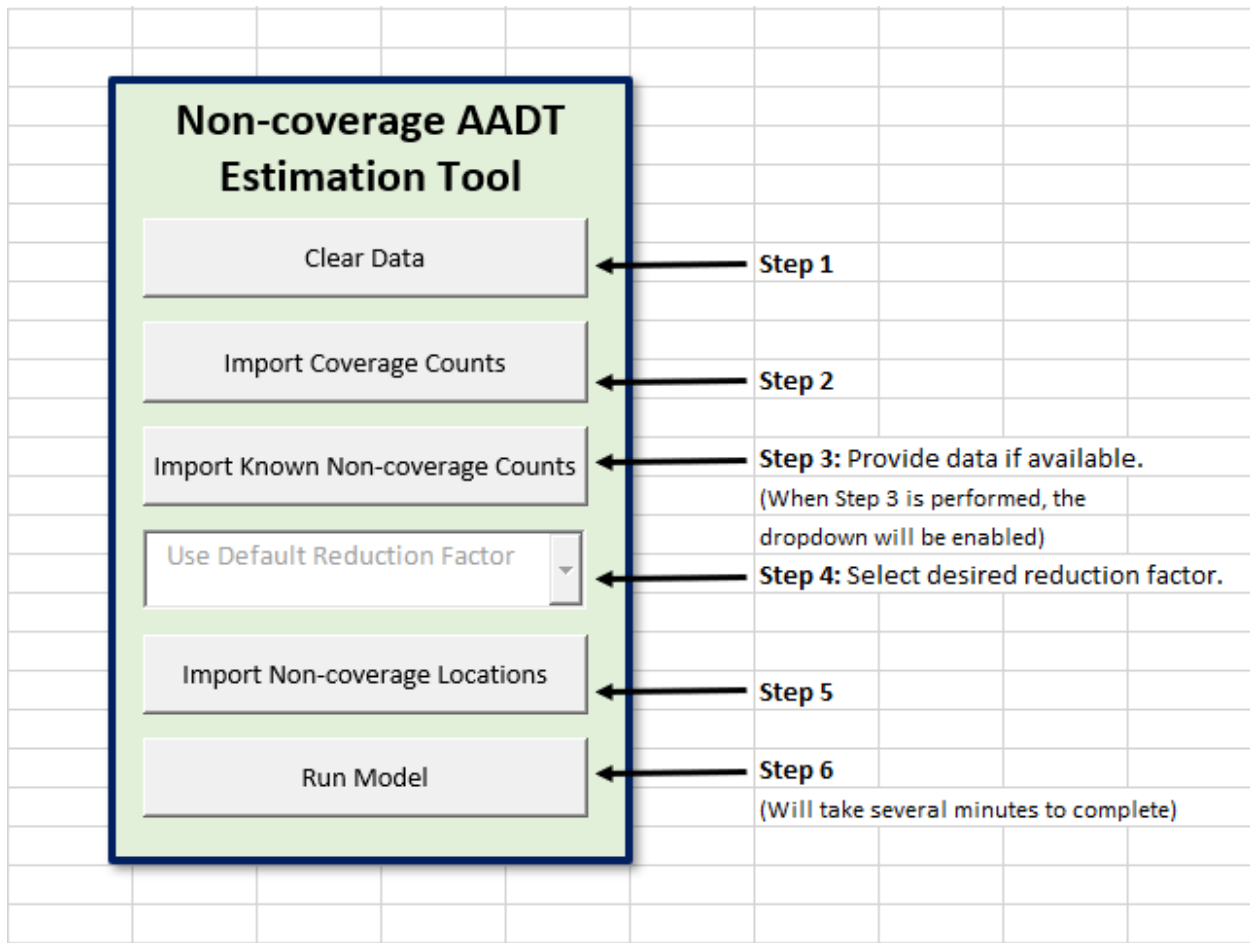


Figure 5-1 Graphical user interface of non-coverage AADT estimation tool

The following explains the steps involved in running the program.

- Step one: Click on the “Clear Data” button. As the name implies, this step clears all the data in various worksheets such as coverage counts, known non-coverage counts, and unknown non-coverage counts.
- Step two: Click on the “Import Coverage Counts” button. The user will be prompted to select a file from the user’s computer using standard Windows File Dialog. Upon successful reading of the file, a dialog box will be displayed informing the user that the data has been loaded successfully into the “Coverage Counts” worksheet. A map of the coverage counts’ location will be generated based on the stations’ latitudes and longitudes as shown in Figure 5-2.

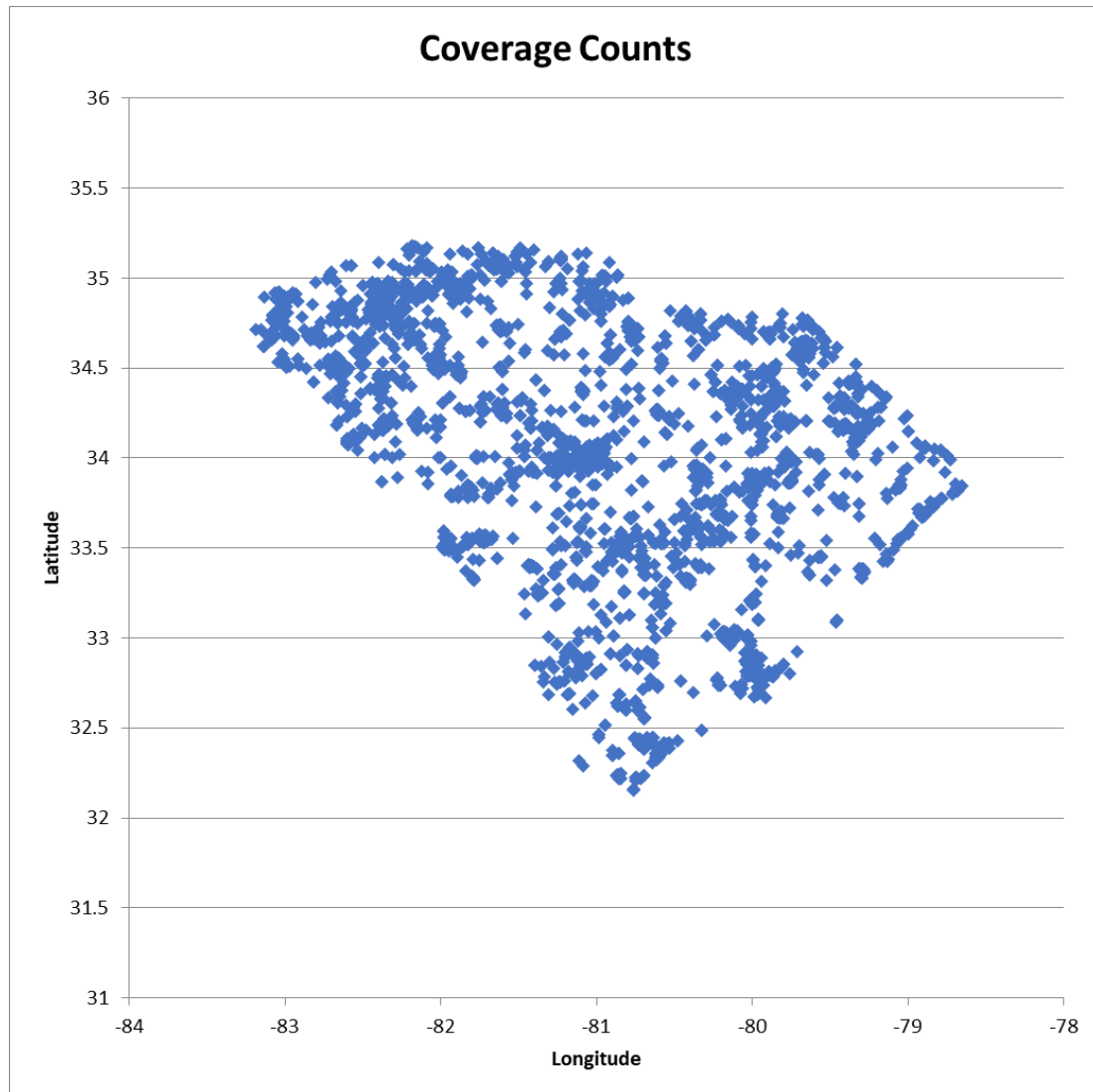


Figure 5-2 Map of coverage counts generated by developed tool

- Step three: Click on the “Import Known Non-coverage Counts” button. This step is optional and should only be executed if the SCDOT has collected counts from non-coverage locations. The user will be prompted to select a file from the user’s computer using standard Windows File Dialog. Upon successful reading of the file, a dialog box will be displayed informing the user that the data has been loaded successfully into the “Known Non-Coverage Counts” worksheet. A map of the known non-coverage counts’ location will be generated based on the stations’ latitudes and longitudes. If Step three is performed, the dropdown box will be enabled.
- Step four: Select desired reduction factor from the dropdown box. Users has the option to use the default reduction factor entered on the parameter worksheet (discussed below) or use the reduction factor calculated based on the provided data. If the latter option is selected, the tool will display the calculated reduction factor as shown in Figure 5-3.

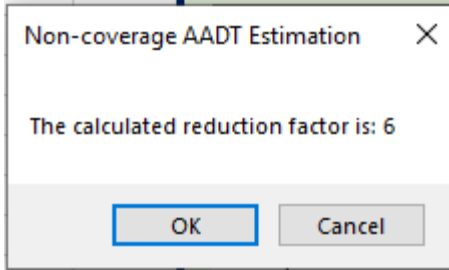


Figure 5-3 Dialog showing value of calculated reduction factor

- Step five: Click on “Import Non-coverage Locations” button. Similar to previous steps, the user will need to provide the appropriate data file, and the tool will provide a message indicating successful reading of the file, and it will generate a map showing the locations of the non-coverage locations.
- Step six: Click on “Run Model” button. This step executes a series of VBA subroutines that implement the kriging model. The run time depends on the number of locations provided. It takes a couple of minutes to complete when the coverage and non-coverage counts are less than 5,000.

The tool allows the user to change three parameters as shown in Figure 5-4.

1. Absolute error threshold (percentile). The default is 90th percentile. The decimal form of the percentile should be specified. For example, 0.8 should be entered if the desired threshold for switching from kriging-predicted to default value is 80th percentile.
2. Complimentary model. The default is “0” which means the hybrid kriging model will be used to predict AADT. If “1” is specified, then the point-based model will be applied to those stations with the provided road features. The AADT of the remaining stations will be predicted by the hybrid kriging model. Similarly, if “2” is selected, then the regular regression model will be applied and if “3” is selected then the quantile regression model will be applied to those stations with the provided road features. Note that columns F to L in the “Non-Coverage AADT Estimation” worksheet must have “0” or “1” (where “1” indicates true) if “1”, “2” or “3” is specified for this parameter.
3. Default reduction factor. The default is “6.” The user has the option to use this value or have the tool calculate the reduction factor from the data.


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q		
		Non-Coverage AADT Estimation Tool (version 0.1P)																	
Configurable Parameters																			
Absolute Error threshold (percentile):				0.9	(enter in decimal form)														
Complimentary model:				0	(0 refers to hybrid method only; 1 refers point-based model; 2 refers regular regression model; 3 refers Quantile regression model)														
Default reduction value				6															

Figure 5-4 Configurable parameters

On the *parameters* worksheet, in addition to the ability to change the three parameters discussed above, the user can also change the predicted AADT for the point-based model and coefficients of

the regular regression and quantile regression models. The updating of the regression models' coefficients should only be done if the models are re-estimated.

REFERENCES

- Anderson, M., Sharfi, K., & Gholston, S. (2006). Direct Demand Forecasting Model for Small Urban Communities Using Multiple Linear Regression. *Transportation Research Record*, 114-117.
- Apronti, D., Ksaibati, K., Gerow, K., & Hepner, J. (2016). Estimating Traffic Volume on Wyoming Low Volume Roads Using Linear and Logistic Regression Methods. *Journal of Traffic and Transportation Engineering*, 493-506.
- Attoh-Okine, G. A. (2021). Bayesian Nonparametric Approach to Average Annual Daily Traffic Estimation for Bridges. *Transportation Research Record*, 1-13.
- Barnett, J. S. (2015). *On the Estimation of Volumes of Roadways: An Investigation of Stop-Controlled Minor Legs*.
- Castro-Neto, M., Jeong, Y., Jeong, M., & Han, L. (2009). AADT Prediction Using Support Vector Regression with Data-Dependent Parameters. *Expert Systems with Applications*, 2979-2986.
- Chang, H., & Cheon, S. (2019). The Potential Use of Big Vehicle GPS Data for Estimations of Annual Average Daily Traffic for unmeasured Road Segments. *Transportation*, 1011-1032.
- Chowdhury, M., Huynh, N., Khan, S., Khan, M., Brunk, K., Torkjazi, M., . . . Keehan, M. (2019). *Cost Effective Strategies for Estimating Statewide AADT*. Washington: FHWA.
- Christian, M. (2021, 4 26). *South Carolina's population increases more than 10%*. Retrieved from https://scnow.com/news/local/south-carolinas-population-increases-more-than-10/article_34b9ab40-a6ce-11eb-80ba-3ff7ecd2846d.html
- Das, S. (2021). Traffic volume prediction on low-volume roadways: a Cubist approach. *Transportation planning and technology*, 93--110.
- Dixon, K. a. (2012). *Calibrating the future highway safety manual predictive methods for Oregon state highways*. Oregon: Oregon. Dept. of Transportation. Research Section.
- Ebden, M. (2015). Gaussian processes: A quick introduction. *arXiv preprint arXiv:1505.02965*.
- Eom, J., Park, M., Heo, T., & Huntsinger, L. (2006). Improving the Prediction of Annual Average Daily Traffic for Nonfreeway Facilities by Applying a Spatial Statistical Method. *Transportation Research Record*, 20-29.
- Erhunmwunsee, P. (1991). Estimating Average Annual Daily Traffic Flow from Short Period Counts. *Institute of Transportation Engineers*, 61(11), 23-30.

- Federal Highway Administration. (2018). *Traffic Data Computation Method Pocket Guide*. Washington: FHWA.
- Holik, W. A. (2017). Innovative traffic data QA/QC procedures and automating AADT estimation. *Technical Report United States. Federal Highway Administration. Office of Safety*.
- Holik, W., Tsapakis, I., Vandervalk, A., Turner, S., & Habermann, J. (2017). *Engagement of Local Agencies in Traffic Volume Collection and Random Sampling Procedures*. Washington: FHWA.
- Islam, S. (2016). *Estimation of Annual Average Daily Traffic (AADT) and Missing Hourly Volume using Artificial Intelligence*. Clemson: Clemson University.
- Jiang, Z., McCord, M., & Goel, P. (2006). Improved AADT Estimation by Combining Information in Image and Ground Based Traffic Data. *Journal of Transportation Engineering*, 523-530.
- Keehan, M. (2017). *Annual Average Daily Traffic (AADT) Estimation with Regression Using Centrality and Roadway Characteristic Variables*. Clemson: All Theses.
- Khan, S., Islam, S. K., Dey, K., Chowdhury, M., Huynh, N., & Torkjazi, M. (2018). Development of Statewide Annual Average Daily Traffic Estimation Model from Short Term Counts: A Comparative Study for South Carolina. *Transportation Research Record*, 55-64.
- Lam, W. H. (2000). Estimation of AADT from short period counts in Hong Kong—A comparison between neural network method and regression analysis. *Journal of Advanced Transportation*, 249--268.
- Lam, W., & Xu, J. (1999). Estimation of AADT from Short Period Counts in Hong Kong - A Comparison Between Neural Network Method and Regression Analysis. *Journal of Advanced Transportation*, 34(2), 249-268.
- Lowry, M. (2014). Spatial Interpolation of Traffic Counts Based on Origin-Destination Centrality. *Journal of Transport Geography*, 98-105.
- Ma, X. S. (2019). Spatial interpolation of missing annual average daily traffic data using copula-based model. *IEEE Intelligent Transportation Systems Magazine* 11, no. 3, 158-170.
- Michael, L., & Michael, D. (2012). *GIS Tools to Estimate Average Annual Daily Traffic*. University of Idaho.
- Mohamad, D., Sinha, K., Kuczek, T., & Scholer, C. (1998). Annual Average Daily Traffic Prediction Model for County Roads. *Transportation Research Record*, 69-77.

- Pan, T. (2008). *Assignment of Estimated Average Annual Daily Traffic Volumes on All Roads in Florida*. University of South Florida.
- Phillips, G., & Blake, P. (1980). Estimating Total Annual Traffic Flow from Short Period Counts. *Transportation Planning and Technology*, 169-174.
- Pulugurtha, S., & Kusam, P. (2012). Modeling Annual Average Daily Traffic with Integrated Spatial Data from Multiple Network Buffer Bandwidths. *Transportation Research Record*, 53-60.
- Sakib Mahmud Khan, S. I. (2017). Development of Statewide AADT Estimation Model from Short-Term Counts: A Comparative Study for South Carolina. *arXiv preprint arXiv:1712.01257*.
- Seaver, W., Chatterjee, A., & Seaver, M. (2000). Estimation of Traffic Volume on Rural Local Roads. *Transportation Research Record*, 121-128.
- Selby, B., & Kockelman, K. (2013). Spatial Prediction of Traffic Levels in Unmeasured Locations: Applications of Universal Kriging and Geographically Weighted Regression. *Journal of Transport Geography*, 24-32.
- Sfryidis, A., & Agnolucci, P. (2020). Annual Average Daily Traffic Estimation in England and Wales: An Application of Clustering and Regression Modelling. *Journal of Transport Geography*, 1-17.
- Sfryidis, A. a. (20220). Annual average daily traffic estimation in England and Wales: An application of clustering and regression modelling. *Journal of Transport Geography* , 102658.
- Shamo, B., Asa, E., & Membah, J. (2015). Linear Spatial Interpolation and Analysis of Annual Average Daily Traffic Data. *Journal of Computing in Civil Engineering*, 1-8.
- Sharma, S., Lingras, P., Liu, G., & Xu, F. (2000). Estimation of Annual Average Daily Traffic on Low-Volume Roads. *Transportation Research Record*, 103-111.
- Sharma, S., Lingras, P., Xu, F., & Kilburn, P. (2001). Application of Neural Networks to Estimate AADT on Low-Volume Roads. *Journal of Transportation Engineering*, 426-432.
- Sharma, S., Lingras, P., Xu, F., & Liu, G. (1999). Neural Networks as Alternative to Traditional Factor Approach of Annual Average Daily Traffic Estimation from Traffic Counts. *Transportation Research Record*, 24-31.
- Smith, T. (2020, January 7). *Notebook on Spatial Data Analysis*. Retrieved from ESE 502: <http://www.seas.upenn.edu/~ese502/#notebook>

- Staats, W. (2016). *Estimation of Annual Average Daily Traffic on Local Roads in Kentucky*. University of Kentucky.
- Sun, X., & Das, S. (2015). *Developing a Method for Estimating AADT on all Louisiana Roads*. Baton Rouge: Federal Highway Administration.
- Tang, Y., Lam, W., & Ng, P. (2003). Comparison of Four Modeling Techniques for Short-Term AADT Forecasting in Hong Kong. *Journal of Transportation Engineering*, 129(3), 271-277.
- Tang, Y., Lam, W., & Ng, P. (2003). Comparison of Four Modeling Techniques for Short-Term AADT Forecasting in Hong Kong. *Journal of Transportation Engineering*, 271-277.
- Transportation, S. C. (2021, July 2). Retrieved from TRAFFIC COUNTS: <https://www.scdot.org/travel/travel-trafficdata.aspx>
- Unnikrishnan, A., Figliozzi, M., Moughari, M., & Urbina, S. (2018). *A Method to Estimate Annual Average Daily Traffic for Minor Facilities for MAP-21 Reporting and Statewide Safety Analysis*. Portland: Federal Highway Administration.
- Wang, T., Gan, A., & Alluri, P. (2013). Estimating Annual Average Daily Traffic for Local Roads for Highway Safety Analysis. *Transportation Research Record*, 60-67.
- Wang, X., & Kockelman, K. (2009). Forecasting Network Data: Spatial Interpolation of Traffic Counts from Texas Data. *Transportation Research Record*, 100-108.
- Xia, Q., Zhao, F., Chen, Z., Shen, D., & Ospina, D. (1999). Estimation of Annual Average Daily Traffic for Nonstate Roads in a Florida County. *Transportation Research Record*, 32-40.
- Yang, B., Wang, S., & Bao, Y. (2014). New Efficient Regression Method for Local AADT Estimation via SCAD Variable Selection. *IEEE Transactions On Intelligent Transportation Systems*, 2726-2731.
- Zhao, F., & Chung, S. (2001). Contributing Factors of Annual Average Daily Traffic in a Florida County. *Transportation Research Record*, 113-122.
- Zhao, F., & Park, N. (2004). Using Geographically Weighted Regression Models to Estimate Annual Average Daily Traffic. *Transportation Research Record*, 99-107.
- Zhong, M., & Hanson, B. (2009). GIS-Based Travel Demand Modeling for Estimating Traffic on Low-Class Roads. *Transportation Planning and Technology*, 423-439.