10-5-2023

Comparison of models with and without roadway features to estimate annual average daily traffic at non-coverage locations

Jing Wang
University of Nebraska-Lincoln, jwang96@huskers.unl.edu

Ryan DeVine
Rummel, Klepper, & Kahl, LLP

Nathan Huynh
University of Nebraska - Lincoln, nathan.huynh@unl.edu

Weimin Jin
Arcadis

Gurcan Comert
Benedict College

See next page for additional authors

International Journal of Transportation Science and Technology xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Transportation Science and Technology

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

Comparison of models with and without roadway features to estimate annual average daily traffic at non-coverage locations

Jing Wang [a], Ryan DeVine [b], Nathan Huynh [a,*], Weimin Jin [c], Gurcan Comert [d], Mashrur Chowdhury [e]

[a] Department of Civil & Environmental Engineering, University of Nebraska-Lincoln, Lincoln, NE 68583-0851, USA
[b] Rummel, Klepper, & Kahl, LLP, Richmond, VA 23223, USA
[c] Arcadis, 10205 Westheimer Road, Suite 800, Houston, TX 77042, USA
[d] Department of Computer Science, Physics, and Engineering, Benedict College, Columbia, SC 29204, USA
[e] Glenn Department of Civil Engineering, Clemson University, 216 Lowry Hall, Clemson, SC 29631, USA

ARTICLE INFO

Article history:
Received 14 July 2023
Received in revised form 9 September 2023
Accepted 5 October 2023
Available online xxxx

Keywords:
AADT
Non-coverage roads
Kriging method
Point-based model
Gaussian process regression

ABSTRACT

This study develops and evaluates models to estimate Annual Average Daily Traffic (AADT) at non-coverage or out-of-network locations. The non-coverage locations are those where counts are performed very infrequently, but an up-to-date and accurate estimate is needed by state departments of transportation. Two types of models are developed, one that simply uses the nearby known AADTs to provide an estimate and one that requires roadway features (e.g., type of median, presence of left-turn lane). The advantage of the former type is that no additional data collection is needed, thereby saving time and money for state highway agencies. A natural question and one that this study seeks to answer is: can this type of model provide equally as good or better estimates than the latter type? The models developed belonging to the first type include hybrid-kriging and Gaussian process regression model (GPR-no-feature), and the models developed belonging to the second type include point-based model, ordinary regression model, quantile regression model, and Gaussian process regression model (GPR-with-features). The performance of these models is compared against one another using South Carolina data from 2019 to 2021. The results indicate that the GPR-with-features model yields the lowest Root Mean Squared Error (RMSE) and lowest Mean Absolute Percentage Error (MAPE). It outperforms the hybrid kriging model by 6.45% in RMSE, GPR without features model by 4.25%, point-based model by 4.69%, regular regression model by 11.35%, and quantile regression model by 4.25%. Similarly, the GPR-with-features model outperforms the hybrid kriging model by 25.21% in MAPE, GPR without features model by 17.81%, point-based model by 22.26%, regular regression model by 26.36%, and quantile regression model by 21.07%.

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

* Corresponding author.
E-mail address: nathan.huynh@unl.edu (N. Huynh).

https://doi.org/10.1016/j.ijtst.2023.10.001
2046-0430/

1. Introduction

State Departments of Transportation (DOTs) are responsible for network planning, roadway maintenance, and traffic safety, among many other responsibilities. To carry out these design and analysis tasks, an essential input that is needed is the annual average daily traffic (AADT). In addition, state DOTs are required by the Federal Highway Administration, as part of the Highway Performance Monitoring System (HPMS) Program, to submit AADT data for all roads they maintain on an annual basis. The sources used by state DOTs to ob'ain AADTs include permanent or contin uous count stations (CCS) and short-term count stations. These are called coverage counts or in-network counts because counts at these locations are updated on an annual, biennial, or triennial basis. Fig. 1a shows the locations of coverage counts in South Carolina. Even though it may appear that there are many coverage locations, comparatively, there are significantly more non-coverage locations as shown in Fig. 1b. Non-coverage counts are performed very infrequently, oftentimes more than ten years since the previous count was performed. The challenge for most state DOTs is getting accurate estimates for non-coverage counts.

Getting short-term counts at non-coverage locations requires both time and money. In addition, it requires the DOT to send its personnel or contractors to the non-coverage location to set up the pneumatic tubes and counters for a 48-h data collection period. Thus, it poses a safety risk to DOT personnel and contractors. According to the New York State Department of Transportation, a short-term count costs approximately $100 (Holik et al. (2017)). In South Carolina, there are about 28,600 non-coverage roads, which means the South Carolina DOT would need to spend nearly $2.9 million to obtain accurate non-coverage counts. Hence, budget constraints make short-term counts an impractical and infeasible approach to obtaining AADTs at non-coverage locations on a regular basis.

There is a large body of work on methods to estimate AADT using short-term counts; these methods include regression models, geospatial methods, travel demand models, centrality, image processing, and machine learning methods. Among these methods, regression models are the most frequently used. The independent variables found to be statistically significant in these regression studies include, but are not limited to, number of lanes, area type, functional classification, distance to highway, population, income, percentage of adults unemployed in a household, and percentage of households below the poverty line (Yang et al. (2014), Apronti et al. (2016) and Pan (2008)). Machine learning (ML) techniques have also been used extensively to estimate AADT using nearby short-term counts. The most commonly used ML technique is Artificial Neural Network (Sharma et al. (2000) and Sharma et al. (2001)). A few studies have used Support Vector Regression, another ML technique. A much smaller number of studies explored methods to estimate AADT at non-coverage locations. Collectively, the methods used in these studies could be grouped into the following categories: multiple linear regression, kriging, machine learning, travel demand, and others. A gap in the literature is that no study has utilized quantile regression and Gaussian Process Regression (GPR) to estimate AADT at non-coverage locations. GPR is a non-parametric, machine learning method designed to solve regression and probabilistic classification problems, and it has been shown to outperform the classical regression approach (Vogel et al. (1999) and Sun et al. (2014)). Another gap in the literature is that no study has compared the performance of models that simply use the nearby known AADTs to provide an estimate for the non-coverage count to those models that require roadway features (e.g., type of median, presence of left-turn lane). The advantage of the former type of model is that no additional data collection is needed, thereby saving time and money for state DOTs. This study addresses the two identified gaps.

The objective of this study is to develop and evaluate models to estimate AADT at non-coverage locations (i.e., non-coverage counts). Two types of models are developed, one that simply uses the nearby known AADTs to provide an estimate and one that requires roadway features (e.g., type of median, presence of left-turn lane). The models developed belonging to the first type include hybrid-kriging and Gaussian process regression model (GPR-no-feature), and the models developed belonging to the second type include point-based model, ordinary multiple linear regression model, quantile regression

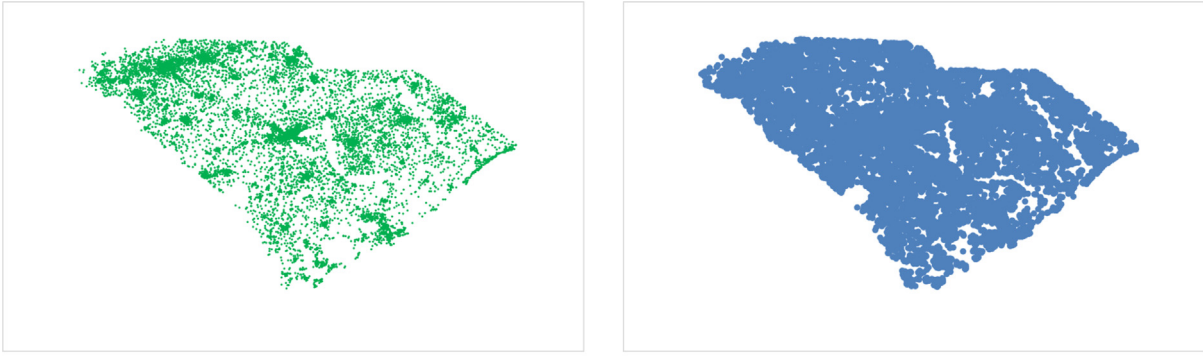

Fig. 1. Locations of count stations in South Carolina: a) coverage, and b) non-coverage.

model, and Gaussian process regression model (GPR-with-features). The performance of these models is compared against one another using South Carolina data.

The remainder of this paper is organized as follows. Section 1 provides a review of estimation approaches for non-coverage counts. Section 2 presents the methodological details of the developed models. Section 3 describes the model fitting and validation procedure. Section 4 discusses the model comparison results. Lastly, Section 5 provides a summary of the study, its limitations, and potential areas of future work.

2. Literature review

Readers are referred to the work of Baffoe-Twum et al. (2022) for a review of methods used to estimate coverage counts. The following review focuses exclusively on studies that dealt with non-coverage counts. These studies are grouped into the following categories: regression analysis, kriging, machine learning, and others. Within each category, studies are summarized in chronological order.

2.1. Regression analysis

Mohamad et al. (1998) applied multiple linear regression(MLR) to estimate non-coverage counts on local roads in Indiana. The significant independent variables were determined to be urban/rural classification, easy access to state highways, county population, and total arterial mileage of the county. Their final model was validated using measured AADTs at eight randomly selected locations with a mean square error of 16%. Xia et al. (1999) developed an MLR model to estimate non-coverage counts in urban areas of Florida. The significant variables were determined to be accessibility to non-state roads, number of lanes, land use type, functional classification, automobile ownership, and service employment. Their final model was validated using data from 40 additional locations with a Mean Absolute Percentage Error (MAPE) of 22.7%.

Seaver et al. (2000) expanded the MLR methodology to incorporate principal component analysis and cluster regression analysis. Principal component analysis was used to reduce 45 independent variables to around seven or eight, depending on the area being investigated. Cluster regression analysis was used to locate clusters with the same road type and metropolitan status (i.e., either in or out of a metropolitan statistical area, MSA). Within each cluster, an MLR model was developed to estimate AADT using the previously determined principal variables. 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. Zhao and Chung (2001) continued the work that was started in 1999 by Xia et al. (1999). Four MLR models were developed and assessed. The most promising model had five significant variables: number of lanes, functional class, regional accessibility to employment centers, employment indicator, and direct access to an expressway. Their model had an R-squared value of 0.818, compared to 0.63 obtained in the 1999 study. Chen et al. (2019) estimated uncovered local roads using generalized mixed regression with a number of lanes, population density, and distance to arterials. The authors also improved bias on uncovered roads via oversampling from these roads. The study obtained about 80% correlation with the true AADT values.

Zhao and Park (2004) were the first to investigate the use of geographically weighted multiple linear regression (GWMLR) models to estimate non-coverage counts for roads in Florida. An ordinary MLR model was developed 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 GWMLR model utilized a bi-square weighting function, and the second model utilized a Gaussian weighting function. Both GWMLR models outperformed the control model which had an R-squared value of 0.764. The bi-square model had an R-squared value of 0.8756 which was higher than the R-squared value of the Gaussian model (0.87).

Anderson et al. (2006) compared the performance of an MLR model against the travel demand method for a small urban community in Alabama. The travel demand approach is more comprehensive but it is computationally expensive. The MLR model had five significant variables: number of lanes, functional class, population, employment, and a binary variable that represents mobility. Results indicated that both approaches produced similar results; this was confirmed by using a t-test, graphical inspection, and a Nash–Sutcliffe statistic. The R-squared value of the MLR model was 0.819. Pan (2008) developed MLR models to estimate AADTs on all types of roads in Florida. The state of Florida was broken into three categories based on population (low, medium, and high), and for each of these categories, two models were developed. One model was developed for the state/county highways and another model was developed for local roads. It was observed that the state/county highway model outperformed the local model for all three population areas. Moreover, it was observed that the low population models (MAPE of 31.99% and 46.69%) outperformed the medium (MAPE of 65.01% and 65.35%) and high population models (MAPE of 46.81% and 159.49%).

Pulugurtha and Kusam (2012) investigated the effect of buffer sizes for the GWMLR method. 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 buffer 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 quasi-likelihood under the

independence model criterion (QIC) was used to assess the models, and for this metric the smaller the QIC value the better the model. The authors found that segmenting the study area into groups based on functional class allowed for better accuracy. The variables that were found to be significant in the functional class-based models are urban classification, number of lanes, speed limit, upstream road speed limit, downstream road speed limit, downstream cross street number of lanes, population, manufactured housing (land use), and rural district.

Lowry et al. (2012) developed an MLR model to estimate non-coverage counts on rural roads to include a new parameter called the connectivity importance index. The connectivity importance index was 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. The final model had functional class, number of lanes, and connectivity importance index as significant variables and an R-squared value of 0.72. Yang et al. (2014) proposed a new variable selection procedure for MLR called smoothly clipped absolute deviation penalty (SCAD). This selection procedure selected significant independent variables and estimated regression coefficients in one step. The SCAD selection procedure was then compared to backward and forward variable selection procedures. The significant variables were found to be number of cars (obtained from satellite image), number of lanes, housing units, median income, percentage below poverty line, and car intensity (obtained from satellite image). The authors concluded that backward and SCAD selection procedures resulted in the same R-squared value (0.6954), while both outperformed the forward variable selection procedure (0.6423).

Apronti et al. (2016) developed an MLR model to predict non-coverage counts 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. Before the log transformation was applied to AADT the model's R-squared value was 0.44, and after the log transformation, the model's R-squared value improved to 0.64. When the MLR model was validated an R-squared value of 0.69 was obtained. The authors concluded since the model's R-squared values for the training and validation data sets are similar, the developed model is not biased. Staats (2016) developed MLR models to estimate non-coverage counts on local roads in Kentucky. The state of Kentucky was split into three geographic areas using highway districts to account for geographic and socioeconomic variability. A model was developed using counts from probe vehicles, residential vehicle registrations, 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 were developed. The rural models were those with AADT values that ranged between 20 and 1,000; a road is not considered rural if the AADT is above 1,000. For the rural models, the MAPE was found to be between 61% and 87%, while the MAPE for the urban models ranged between 354% and 1,956%.

Doustmohammadi and Anderson (2019) presented a Bayesian Regression model for estimating low-volume AADT for roadways in Alabama. The model was developed using socio-economic factors as independent variables, including nearby population, number of households in the area, employment in the area, population to job ratio, and access to major roads. The statistical accuracy was calculated using a Percent Root Mean Square Error (%RMSE). Using this metric, the Bayesian Regression model was found to outperform the linear regression model in predicting AADT of low-volume roadways, particularly, those with at least 250 vehicles per day.

Das and Tsapakis (2020) compared the performance of two statistical models (MLR and generalized linear model) against three machine learning models (random forest, support vector machine, and K-nearest neighbor) in estimating low-volume AADT. The data used for the analysis, from the state of Vermont, included traffic count data on low-volume roadways, U.S. Census data (population per sq. mile), American Community Survey data (household per square mile), Longitudinal Employer-Household Dynamics data, and distance to major highways. The best-performing machine learning model was found to be random forest which had a higher R^2 value than the statistical models. The best-performing random forest model outperformed the MLR model by 45% to 0.77%. According to the study, significant factors were population and work employment density. The studies commonly showed that machine learning methods performed better than the traditional linear and generalized regression models.

Pulugurtha and Mathew (2021) developed ordinary least square (OLS) and geographically weighted regression (GWR) models to estimate AADT on local roads in North Carolina. Their models used land-use data (e.g., single-family residential units, multi-family residential units, agriculture land use, and commercial areas), population density, road density, and the number of local road traffic count stations available in each county as independent variables. The model estimation results indicated that road density, AADT at the nearest non-local road, and land use variables have a significant influence on local road AADT. The GWR model was found to outperform the OLS regression model, and thus, was used subsequently to estimate AADT at non-coverage locations. The prediction error was found to be higher in urban areas and counties with a smaller number of local road traffic count stations.

2.2. Kriging

Eom et al. (2006) were the first to utilize kriging to estimate AADT for non-freeway facilities. Multiple theoretical semivariograms were investigated, including Gaussian, exponential, and spherical. Both exponential and spherical models provided more accurate AADT estimates for urban and rural areas when compared to traditional regression model estimates; WLS had a mean square prediction error (MSPE) of 2.91, REML achieved an MSPE of 2.86, and OLS achieved an MSPE of 3.12. Wang and Kockelman (2009) divided roadways in the state of Texas into two categories, interstate highway, and principal arterial, and developed a kriging model for each category. The authors found that the kriging models performed well for

roads that have an AADT greater than 1,000 and that the principal arterial model overestimated the AADT on roads that had low traffic volumes. Selby and Kockelman (2013) compared kriging to GWMLR. It was found that the kriging model outperformed the GWMLR model by 3 to 8% in absolute error on average. The authors also investigated the use of Euclidian distance instead of network distance to see the effects on the model's error. No sizable difference in error was found between these two methods.

2.3. Machine learning

Sharma et al. (2001, 2000) applied Artificial Neural Networks (ANNs) to estimate AADT on low-volume roads in Alberta, Canada. Their ANNs were based on a multilayered, feedforward, and back-propagation design for supervised learning. The AADT estimation errors resulting from various durations and frequencies of counts were analyzed by computing average and percentile errors. The results of their study indicated that using two 48-h counts is better than other frequencies (one or three) or durations (24-h or 72-h) of counts. The authors noted that "the 95th percentile error values of about 25% for the neural network models compare favorably with the values reported in the literature for low-volume roads using the traditional factor approach.".

Castro-Neto et al. (2009) investigated the use of support vector regression with data-dependent parameters to estimate AADT on roads in Tennessee. A comparison of support vector regression model with data-dependent parameters, Holt exponential smoothing model, and ordinary least squares regression model was performed for urban and rural roads. It was found that the support vector regression model with data-dependent parameters outperformed the Holt exponential smoothing model (MAPE of 2.26% compared to 2.69%) and the ordinary least squares regression model (MAPE of 3.85%). Sun and Das (2015) utilized a modified support vector regression (SVR) method to estimate AADT on non-state roads in Louisiana. A variety of models was explored and SVR was found to be the best and its estimates are sufficiently accurate for transportation planning and traffic safety studies. The authors noted that their SVR model tends to underestimate AADT for roadways with AADT higher than 1,500 and they recommended that parish-specific models be developed. Sfyridis and Agnolucci (2020) proposed a methodology to combine machine learning and standard statistical methods to estimate AADT on all roads in England and Wales. They first applied a clustering algorithm to take into account (dis) similarities among count locations and their surroundings and group points with similar characteristics. Then, they applied three models, namely ordinary MLR, Random Forests (RF), and SVR for each cluster and validated the results. The SVR 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 ordinary MLR model which produced a MAPE ranging from 2% to 325%.

Zhang and Chen (2020) compared the performance of neural networks and random forests to each other and the ordinary MLR. Their study was the first to use probe vehicle data as an independent variable in the estimation model, as well as the "betweenness" centrality (BC) measure. BC is a measure that quantifies the number of shortest paths that pass through a given link and is used in complex network theory to determine the hierarchical importance of network links (Zhang and Chen (2020)). The random forest model was found to yield the best performance, with an R-squared of 0.92 and a median APE of 25.2%. Moreover, incorporating both probe vehicle data and BC in the models boosted models' accuracies by 30%–37% for all roads and 23%–43% for lower functional class (5–7) roads compared to models that used only socio-demographic and roadway characteristic factors.

2.4. Others

Another method used to estimate AADT is based on a node's centrality measure. 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. Lowry et al. (2012) used the centrality method to estimate AADT on roads in Moscow, Idaho. Stress centrality was used as the independent variable in a regression model. The model was validated using an out-of-sample data set with an R-squared value of 0.95. A key advantage of this method is that it requires minimal data collection and can be easily executed using a geographic information system. Keehan (2017) investigated the use of origin–destination centrality which includes internal-internal, internal-external, and external-external to estimate AADT on roads in South Carolina. These three parameters were then combined with three additional parameters, functional class, speed limit, and number of lanes, to produce an MLR model. The significant variables were found to be internal-internal centrality, external-external centrality, and speed limit. The final model was found to outperform the traditional travel demand model in terms of Root Mean Square Error (RMSE) and R-squared value.

The travel demand modeling method is another popular method often used to estimate coverage AADT. Wang et al. (2013) implemented the parcel-level travel demand analysis model to estimate AADT using actual traffic counts from Broward County, Florida. The model consisted of four steps: network modeling, parcel-level trip generation, parcel-level trip distribution, and parcel-level trip assignment. Different from the traditional travel demand model, which attempts to simulate the choices that travelers may make during the entire trip from the origin to the destination, the parcel level model attempts to simulate choices that travelers may make in response to the given local street system. The parcel-level model does not

include the mode choice step since transit trips and those with other modes are insignificant on local roads. Their results showed that the proposed method produces significantly lower MAPE (39%-66%) than other regression-based methods.

The simplest method used to estimate AADT is called a point-based model developed by Unnikrishnan et al. (2018). It is essentially a lookup table where the AADT can be looked up if the number of "points" or roadway features a roadway as is known. The premise of this approach is that 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. The estimated AADT is simply the median value of AADTs from roadways with the same number of points. In their work, Unnikrishnan et al. (2018) developed point-based models for four sub-regions (mountain, coast, valley-rural, and MPO), as well as local roads. For local roads, the median error ranged between −16 to 151%. A limitation of this approach is the homogenous nature of local roads, which generally have the same features in an area.

2.5. Summary

Table 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 above review indicates that while a variety of regression models have been utilized to estimate AADT, no study has investigated the effectiveness of quantile regression and Gaussian Process Regression (GPR) to estimate non-coverage counts. Also, no study has compared the performance of models that simply use nearby known coverage counts to estimate the non-coverage count to those models that require roadway features (e.g., type of median, presence of left-turn lane). The advantage of the former type of model is that no additional data collection is needed, thereby saving time and money for state DOTs. It should be noted that Ashley and Attoh-Okine (2021) were the first to apply GPR to estimate AADT. Their motivation for using GPR was to avoid dependency on predictors which may not be readily available to aid in quick and precise estimation of AADT at a particular bridge (coverage location). For this reason, their work was limited to using AADT collected at a specific bridge for both training and testing. That is, their GPR models did not include any independent variables. The

Table 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
2000	Sharma et al.	ANN and TFA	Alberta	95% percent error = 25%
2001	Sharma et al.	ANN	Alberta	95% percent error = 25%
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	MAPE = 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 = 39% −66%
2014	Lowry	C	Idaho	MdAPE = 22–29%
2014	Yang et al.	MLR	North Carolina	R^2=0.6954
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
2019	Doustmohammadi and Anderson	BRM, MLR	Alabama	% RMSE = 62.15
2020	Das and Tsapakis	MLR, GLM, RF, SVR, KNN	Vermont	R^2 = 0.45–0.77
2020	Sfyridis and Agnolucci	OLR, RF, SVR	Wales	MAPE = 2–277% RMSE = 26–58,650
2020	Zhang and Chen	OLR,RF and NN	Kentucky	R^2=0.92 MAPE = 25.2%
2021	Pulugurtha and Mathew	OLS and GWR	North Carolina	MAPE = 52.6%-120.1% MPE = −19.2%–88.3% RMSE = 111–993
2023	This paper	K,MLR, QRM, PBM and QRM	South Carolina	RMSE = 203–229

MLR = Multiple linear regression, ANN = Artificial Neural Network, TFA = Traditional Factor Approach, GWMLR = Geographically multiple linear regression, K = Kriging, TD = Travel demand, C = Centrality, EM = Emerging methods, OLS = Ordinary least square, BRM = Bayesian Regression model, GLM = Generalized linear modeling, RF = Random forest, SVR = Support vector regression, KNN = K-nearest neighbor, PBM = Point-based Model, QRM = Quantile Regression Model, SVR = Support vector regression, KNN = K-nearest neighbor, PBM = Point-based Model, QRM = Quantile Regression Model.,

performance of their GPR models was compared against the Box–Jenkins auto-regressive integrated moving average (ARIMA) models and the validation results indicated the overall MAPE for the GPR models is 3.6% compared to 44% for the ARIMA models.

This study contributes to the current body of work by comparing the effectiveness of GPR and quantile regression models against the point-based model, hybrid-kriging, and ordinary MLR model. Additionally, it extends the seminal work of Ashley and Attoh-Okine (2021) to include independent variables in the GPR models. Two types of GPR models are explored in this study, one that uses only nearby coverage counts (referred to as GPR-no-feature model hereafter) and one that uses both nearby coverage counts and roadway geometric features (referred to as GPR-with-features model hereafter).

3. Methodology

3.1. Data description

Coverage counts from 3679 local roads throughout the state of South Carolina were obtained from the South Carolina Department of Transportation (SCDOT). These counts along with their associated functional class (rural/urban), latitude, and longitude were used to develop the kriging model and GPR-no-feature model. Fig. 2a shows the locations of the coverage counts in the training data set. As part of this study, geometric features shown in Table 3 were obtained for all 3679 locations in the training data set to develop the point-based model, ordinary MLR model, quantile regression model, and GPR-with-features model. This information was obtained using Google Earth. To validate the model, counts and roadway geometric features were obtained for an additional 1024 locations. Fig. 2b shows the locations of the non-coverage counts in the testing data set. These data were collected for specific counties selected by the SCDOT. Descriptive statistics between the training and testing data sets are shown in Table 2. It can be seen that the mean AADT of the training data set (coverage counts) is much higher than that of the testing data set (non-coverage). For this reason, the AADT in the training data set was divided by a reduction factor. In this study, the reduction factor was chosen to be the integer value of the quotient: mean AADT of training data set divided by mean AADT of testing data set. Once the reduction factor was applied, any count in the training data set that was higher than the maximum value of the testing data set (5900) was considered an outlier, and thus, removed. The final sample size of the training data set after removing outliers is 3675.

3.2. Regression models

Using the collected features shown in Table 3 as independent variables, two types of regression models were developed to estimate non-coverage counts: ordinary multiple linear regression and quantile regression.

3.2.1. Ordinary multiple linear regression

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

$$y_{pred} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_iX_i \quad (1)$$

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.

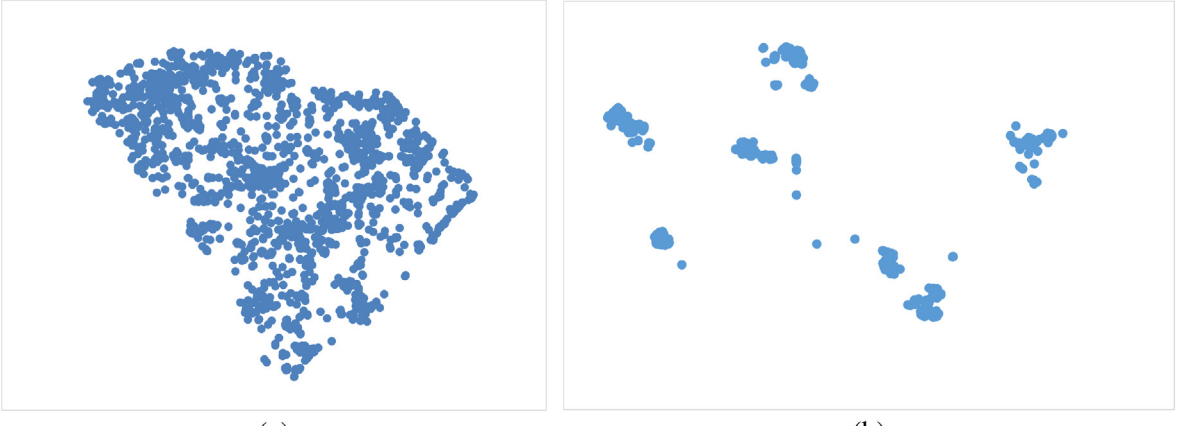

Fig. 2. Training and testing data sets: a) locations of coverage counts in training data set, and b) locations of non-coverage counts in testing data set.

Table 2
Descriptive statistics of training and testing data sets.

Dataset	Training dataset	Testing dataset
Average Value	1750	233
Minimum Value	25	25
First Quartile	300	50
Median Quartile	700	100
Third Quartile	1650	250
95% Quartile	6215	850
Maximum Value	115100	5900

Table 3
Roadway features collected for model development.

Features/ Points	Description of the Variables
Urban	"1" if the roadway segment is in an urban area, and "0" if it is in a rural area.
Double yellow line	"1" if the roadway segment has a double yellow line pavement marking, and "0" if it has no centerline pavement marking.
Median Type	"1" if the roadway segment has a median that is either flush, raised, or two-way left- turn lane, and "0" if it does not have any of these types of median.
Right-Turn Lane	"1" if the roadway segment has an exclusive right-turn lane within 1,000 feet upstream and 1000 feet downstream of the midpoint, and "0" if it does not have an exclusive right-turn lane within 1000 feet of the midpoint.
Left-Turn Lane	"1" if the roadway segment has an exclusive left-turn lane within 1,000 feet upstream and 1000 feet downstream of the midpoint, and "0" if it does not have an exclusive left-turn lane within 1000 feet of the midpoint.
Sidewalk	"1" if the roadway segment has a sidewalk on both sides within 1000 feet upstream and 1000 feet downstream of the midpoint, and "0" if it does not have a sidewalk on both sides within 1000 feet of the midpoint.
Parking Lot	"1" if the roadway segment has a parking lot (e.g., pay to park, parking lots, and parking lots for schools, shopping centers, recreational facilities, and hospitals) adjacent to it within 1,000 feet upstream and 1000 feet downstream of the midpoint, and "0" if it does not have a parking lot adjacent to it within 1000 feet of the midpoint.

3.2.2. Quantile regression model

The quantile regression model is more robust against outliers in the response variable compared to the ordinary MLR model (Li (2015)). It can be described by the following equation:

$$y_{pred} = b_0(q) + b_1(q)X_1 + b_2(q)X_2 + \ldots + b_i(q)X_i \quad (2)$$

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 50^{th} quantile for the quantile regression model.

3.3. Hybrid kriging model

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 Fig. 3. 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, $\hat{Y}(s_i)$, is determined by finding a linear combination of nearby known locations. There are

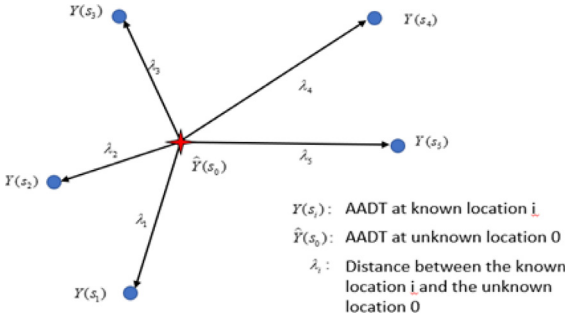

Fig. 3. Illustration of kriging assigning weights to neighbors, adapted from Smith (2016).

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.

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

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

This study adopted the approach used by Shamo et al. (2015) in applying the kriging method to estimate non-coverage counts. However, the Euclidean distance was used instead of network distance to avoid the complexity involved in calculating network distances. As mentioned in the literature review section, Selby and Kockelman (2013) found no sizeable difference between these two methods. Given a pair of latitude and longitude, the distance between the two locations follows the curvature of the earth. In this study, it was assumed that the earth's radius is constant since the distances involved are relatively short. This assumption allows for the use of the great circle distance (i.e., the shortest distance between two points on a sphere) to be calculated.

In this study, the training data set was divided into two sets. The first set (80% of training data set) was used to develop the kriging model. The second set (20% of the training data set) was used 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 use a default value (i.e., average AADT for corresponding county and functional class) instead of the kriging-predicted value. The use of a default value in conjunction with the kriging method is called "hybrid-kriging."

Compared to the popular SVM method, GPR showed great advantages in learning the kernel and regularization parameters, integrated feature selection, and generating fully probabilistic predictions.

3.4. The GPR model

To make this paper self-contained, the mathematical details of GPR is briefly provided here. Fig. 4 shows the steps followed in this study to train and test the GPR model. Following the methodology presented in Zeng et al. (2020), the prediction function of a linear model is:

$$y_* = \beta_0 X_* + \beta_1 + \varepsilon_t \tag{4}$$

where X_* indicates the matrix of test inputs, y_* indicates the matrix of test outputs, and ε_t indicates a noise term. In this study, non-coverage counts are the outputs and nearby coverage counts along with road features are the inputs. GPR assumes that the ε_t follows a Gaussian distribution with a mean of 0 and a variance of θ_n^2:

$$\varepsilon_t \sim N(0, \theta_n^2) \tag{5}$$

The marginal likelihood of the sample data can be expressed as follows.

$$p(y|X) \sim N(0, K_N + \theta_n^2 I) \tag{6}$$

where K_N represents the covariance matrix for the training set, I is the identity matrix, θ_n^2 is the variance of the noise term, and the noise terms are assumed to be identical, independently distributed (IID) random variables.

The predictive results follow a distribution of:

$$p(y_*|X_*, X, y) \sim N(\mu_*, \theta_n^2 I) \tag{7}$$

$$\mu_* = K_{*N}(K_N + \theta_n^2 I)^{-1} y \tag{8}$$

$$\theta_*^2 = K_{**} - K_{*N}(K_N + \theta_n^2 I)^{-1} K_{N*} \tag{9}$$

In the above equations, μ_* is the mean value of the Gaussian process posterior mean, θ_*^2 is the covariance matrices of prediction, K_{*N} represents the covariance matrix between the training and testing data sets, θ_n^2 is the variance of the noise term, X_* is the testing data set, and X is the training data set.

A kernel (or covariance function) describes the covariance of the Gaussian process random variables. Together with the mean function the kernel completely defines a Gaussian process. Five different kernels are evaluated in this study. Their equations are provided below.

Radial-basis function (RBF) kernel:

$$k(x_i, x_j) = exp\left(-\frac{d(x_i, x_j)^2}{2l^2}\right) \tag{10}$$

where, $d(x_i, x_j)$ = Euclidean distance, l = length-scale parameter and l should be positive ($l > 0$).

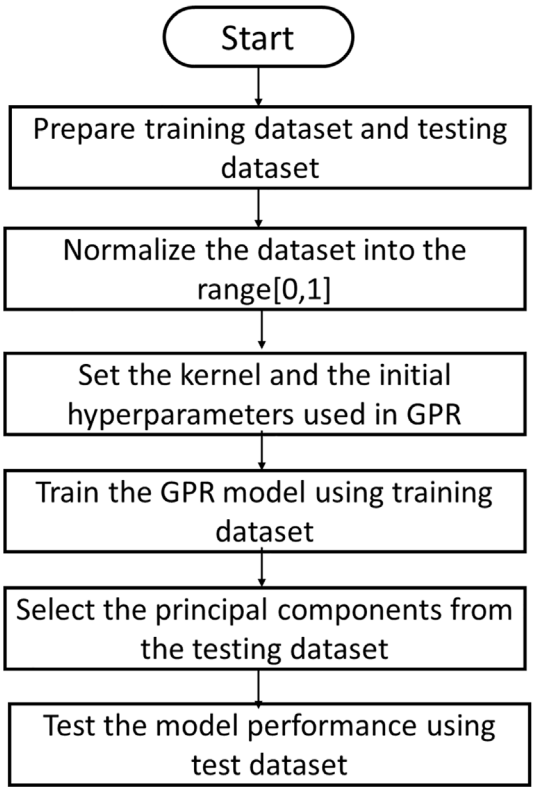

Fig. 4. GPR model development and testing procedure, adapted from Wang (2020).

Matern kernel:

$$k(x_i, x_j) = \frac{1}{\Gamma(v)2^{v-1}}\left(\frac{\sqrt{2v}}{l}d(x_i, x_j)\right)^{v} K_v\left(\frac{\sqrt{2v}}{l}d(x_i, x_j)\right) \tag{11}$$

where, $K_v\left(\frac{\sqrt{2v}}{l}d(x_i, x_j)\right)$ = the modified Bessel function, $\Gamma(v)$= the Gamma function.

$$v = \frac{1}{2}v = \frac{3}{2}v = \frac{5}{2} \tag{12}$$

Exponential-sine-squared (Exp-Sine-Squared) kernel:

$$k(x_i, x_j) = \left(1 + \frac{d(x_i, x_j)^2}{2\alpha l^2}\right) \tag{13}$$

where, l = length-scale parameter $(l, 0), p$ is periodicity parameter $(p > 0)$.

Rational quadratic kernel:

$$k(x_i, x_j) = exp\left(-\frac{2sin^2(\pi d(x_i, x_j)/p)}{l^2}\right) \tag{14}$$

where, l = length-scale parameter $(l, 0), \alpha$ = the scale mixture parameter.

Dot-product kernel:

$$k(x_i, x_j) = \theta_0^2 + x_i * x_j \tag{15}$$

where, s = parameter which controls the homogeneity of the kernel.

3.5. Point-based model

The point-based model applied in this paper is based on a study sponsored by the Oregon Department of Transportation (ODOT) (Unnikrishnan et al. (2018)). This method operates under the presumption that certain roadway features are indica-

tive of segments with higher AADTs. For example, a roadway segment with a two-way left-turn lane is more likely to have a higher AADT than one with just a double yellow line. Each feature shown in Table 3 is considered a point; the higher the number of points a roadway segment has, the higher its AADT. Although there are seven features, the double yellow line feature and median feature are in the same category. Thus, only one of these two features can have a value of 1. That is, if a roadway segment has a double yellow line, then it does not have a median. Conversely, if a roadway segment has a median, then it does not have a double yellow line. The minimum number of points a roadway segment can have is zero. In this case, it is a local road without any centerline pavement marking. The chosen seven features for this study are those that are expected to be collected as part of SCDOT's asset management system, namely, AgileAssets. Using this method, all 3679 local roads in the training data set were grouped by the number of points or features (see Table 3) they have. Within each group, the median AADT was calculated. The median AADT was then used as the predicted non-coverage count given the number of features a roadway has.

4. Results and discussion

To compare the performance of the developed models,the Mean Absolute Percentage Error (MAPE) andthe Root Mean Squared Error (RMSE) are used. Each of the developed models was validated using MAPE and RMSE. RMSE gives the square root of the average of squared differences between actual values and predicted values as shown in the following equation.

$$RMSE = \sqrt{\frac{1}{n}\sum (\hat{y}_i - y_i)^2} \tag{16}$$

MAPE measures the average magnitude of error produced between the actual values and predicted values as shown in the following equation.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{17}$$

where, $\hat{y}_i$ = predicted AADT value, y_i = actual AADT value, and n = number of observations.

4.1. Regression models

Table 4 shows the ordinary MLR 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 suggests that the presence of these features will increase the AADT. Their coefficients represent the increase in AADT. For example, the AADT is increased by 91vehicles per day (vpd) if the local road is located in an urban area versus a rural area. Similarly, the AADT is increased by 70 vpd if the 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 54 vpd.

Table 4 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 ordinary MLR 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 will add only 50 more vehicles per day compared to 91 predicted by the ordinary MLR model. If a road has none of these features, this model predicts the AADT to be 25 vpd, which corresponds to the minimum AADT the SCDOT would report to FHWA.

4.2. Hybrid-kriging model

The hybrid kriging model determined the absolute error threshold based on the calculated RMSE. 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 5 shows the effect of changing the absolute error threshold. A threshold of 90th percentile resulted in the lowest RMSE. For this reason, this threshold value was used for the hybrid kriging model. Table 5 also shows the RMSE for different radii with the absolute error threshold at the 90th percentile. As shown, a radius of 0.9 degrees resulted in the lowest RMSE. For this reason, 0.9 degrees was used for the hybrid kriging model.

4.3. Gaussian regression process model

As part of the GPR model training, the hyperparameters of each kernel were optimized. The optimization was performed by maximizing the log-marginal-likelihood. In this study, the "fmin_1_bfgs_b" function which known as the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS) was used to find optimal hyperparameters for the models. Kernels and their optimal parameters are listed in Table 6.

Table 4
Coefficients for the ordinary MLR model and quantile regression model.

Variable	Estimate	Std. Error	t-value	p-value
Coefficients for Ordinary MLR Model				
(Intercept)	54	15.45	3.135	0.022144
Urban	91	18.64	2.164	1.14E-7
Double Yellow Line	70	21.34	5.326	< 2e-16
Median Type	135	25.85	3.323	0.00081
Right-turn Lane	172	36.31	9.327	1.11E-6
Left-turn Lane	526	49.43	5.364	< 2e-16
Sidewalk	61	19.86	2.294	< 2e-16
Coefficients for Quantile Regression Model				
(Intercept)	25	2.4	12.3	0
Urban	50	4.9	11.4	0
Double Yellow Line	50	4.6	7.8	0
Median Type	50	6.5	12.1	0.0042
Right-turn Lane	200	36.1	6.5	0
Left-turn Lane	375	66.4	2.5	0
Parking Lot	50	15.6	3.9	0
Sidewalk	50	6.5	4.6	0

Table 5
Effect of absolute error threshold and Radii on Kriging model performance.

Absolute Error Threshold (percentile)	Radius (degrees)	RMSE
Effect of absolute error threshold		
95	0.9	352
90	0.9	347
85	0.9	349
Effect of radii with the absolute error threshold at 90th percentile		
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	361
90	0.8	357
90	0.9	352
90	1.0	368

Table 6
Optimized kernel w/o road features using GPR.

Kernel	Opimized kernel parameters	RMSE
Models with road features		
RBF	$316 * *2 * RBF(length\_scale = 154)$	208
Matern	$316 * *2 * Matern(length\_scale = 157, nu = 1.5)$	208
Exp-Sine-Squared	$77.4 * *2 * ExpSineSquared(lengthspace\_scale = 8.36e+03, periodicity = 108)$	203
Rational quadratic	$316 * *2 * RationalQuadratic(alpha = 1.1, length\_scale = 58.7)$	215
Dot-product	$28.9 * *2 * DotProduct(sigma_0 = 151)$	218
Models without road features		
RBF	$316 * *2 * RBF(length\_scale = 2)$	215
Matern	$316 * *2 * Matern(length\_scale = 132, nu = 1.5)$	226
Exp-Sine-Squared	$0.86 * *2 * ExpSineSquared(length\_scale = 107, periodicity = 1.36)$	212
Rational quadratic	$316 * *2 * RationalQuadratic(alpha = 1.1, length\_scale = 58.7)$	214
Dot-product	$1.97 * *2 * DotProduct(sigma_0 = 298)$	242

4.4. Point-based model

As explained previously, the coverage counts are grouped by the number of points or features the roads have in common and the median AADT in each group is used as the predicted value for the point-based model. Table 7 shows the results of the point-based model. This model indicates that if a non-coverage road has zero points then its predicted AADT is 125 vpd, and if a road with three points then its predicted AADT is 650 vpd.

Table 7
Point-based model.

Features/ Points	AADT	Description
0	125	contains none of the 7 features
1	175	contains 1 of the 7 features
2	350	contains 2 of the 7 features
3	650	contains 3 of the 7 features
4	900	contains 4 of the 7 features
5	1600	contains 5 of the 7 features
6,7	1800	contains at least 6 of 7 features

As expected, the results indicate that the more points or features a roadway segment has the higher its AADT. A local road located in an urban area will have a higher AADT than one located in a rural area due to higher population density, vehicle ownerships, and thereby higher number of vehicle trips. The presence of a median of any type suggests a safety concern at that particular location due to a higher-than-expected number of conflict points. Thus, a roadway segment with a median will have a higher AADT than one without. By design, exclusive left or right turn lanes are constructed when it is projected the roadway segment will have high traffic volume for the corresponding movements. AADT is a major criterion for constructing sidewalks in bicycle and pedestrian facility design guides used in the U.S.; the higher the AADT the higher the consideration. Lastly, the presence of a parking lot suggests there are commercial properties adjacent to the roadway. Thus, a roadway segment with a parking lot will have a higher AADT than one without.

4.5. Comparison of models performance

Table 8 shows the RMSE and MAPE of all models evaluated using the testing data set and the relative improvement of various models compared to the ordinary MLR model. For reference, the performance of the ordinary MLR model compared to the SCDOT's current method is also reported in Table 8 (22.90% improvement in RMSE and 7.48% in MAPE); the SCDOT's current method simply uses an AADT of 100 vehicles per day for rural local roads and 200 vehicles per day for urban local roads. Using the ordinary MLR model as the benchmark for evaluation of other models, it can be seen that hybrid-kriging outperformed it by 5.24%, followed by the point-based model at 6.99%, GPR-no-feature at 7.42%, quantile regression at 7.42%, and GPR-with-features at 11.35%. In Table 8, the kernel used for the listed GPR-no-feature model is Exp-Sine-Squared and the kernel used for the GPR-with-features model is also Exp-Sine-Squared. These results indicate that using the GPR-with-features model provides a significant improvement over the commonly used ordinary MLR model. The second best model is the GPR-no-feature. This model has the same RMSE as the quantile regression model but with the benefit of not having to collect additional roadway features. Compared to hybrid-kriging, which also does not require the need to collect additional roadway features, the GPR-no-feature model provides an additional 2.3% improvement in RMSE. Lastly, the relative improvement of using the GPR-with-features model over the GRP-no-feature model is 4.2%. The RMSE values of the GPR methods found in this study (≈200) are within the range of RMSE values (between 26 and 58,650) reported by Sfyridis and Agnolucci (2020) for low-volume roads using ordinary linear regression, random forest and support vector regression, and within the range between 111 and 993 reported by Pulugurtha and Mathew (2021) for local roads using ordinary least square and geographically weighted regression models; given that this study focuses on non-coverage (low-volume) roads, it corresponds to expectations that our RMSEs would be on the lower end of their reported ranges.

A relatively larger number of studies have used MAPE to evaluate the performance of their models. How this study's best MAPE value provided by the GPR-with-features model compares to other studies' best models' MAPE values is summarized in Table 9. Compared to the MLR models developed by Pan (2008), it can be seen that this study's best MAPE (53.28%) is lower than his MAPE for local roads in a large metropolitan area (159.49%) and local roads in a medium-sized urban area (65.35%), but is higher than his MAPE for local roads in rural areas (46.79%). Wang et al. (2013)'s model produced a MAPE of 52%. A possible explanation for their model's slightly better performance than the GPR-with-features model is that it was

Table 8
Comparison of models' performance.

Model	RMSE	RMSE Improvement (%)	MAPE (%)	MAPE Improvement (%)
SCDOT's current method (benchmark)	297		78.21	
Ordinary MLR	229	22.90	72.36	7.48
Model	**RMSE**	**RMSE Improvement (%)**	**MAPE(%)**	**MAPE Improvement (%)**
Ordinary MLR (benchmark)	229		72.36	
Hybrid Kriging	217	5.24	71.24	1.54
Point-based	213	6.99	68.54	6.66
GPR-no-feature	212	7.42	64.83	10.41
Quantile regression	212	7.42	67.50	6.72
GPR-with-feature	203	11.35	53.28	26.36

Table 9
Comparative analysis of MAPE values.

Authors	AADT Estimation Technique	Road Type	MAPE (%)
Pan	MLR	local roads in large urban areas	159.49
		local roads in small-medium urban areas	65.35
		local roads in rural areas	46.79
Wang et al.	TD	local roads	52
Sfyridis and Agnolucci	SVR	multiple types of roads	14.47
Pulugurtha and Mathew	OLS	local roads	52.6
Pulugurtha and Kusam	GWMLR	minor thoroughfares	25.83
Staats	MLR	local roads in rural areas	61
		local roads in urban areas	354
This paper	GPR-with-features	local roads	53.28

MLR = Multiple linear regression, TD = Travel demand, SVR = Support vector regression, OLS = Ordinary least square, GWMLR = Geographically multiple linear regression, and GPR = Gaussian process regression.

obtained using a travel demand model which required a lot more data and effort.Sfyridis and Agnolucci (2020) validated their SVR model for a number of roads and obtained an average MAPE of 14.47%. Pulugurtha and Mathew (2021) reported a MAPE of 52.6% using OLS and data in Duplin County in North Carolina. Their low MAPE values are due to the use of only 235 locations for model validation compared to 1,024 used in this study. Pulugurtha and Kusam (2012) obtained a MAPE of 25.83% using GWMLR for minor thoroughfares. Lastly, Staats (2016)'s MLR model yielded MAPE values of 61% and 354% for local roads in rural and urban areas, respectively. Collectively, it can be concluded that the MAPE of the GPR-with-features model compares favorably to previous studies.

5. Summary and conclusions

This study investigated models to provide more accurate estimates of AADT at non-coverage locations. Six different models were developed and evaluated: 1) ordinary MLR, 2) quantile regression, 3) hybrid-kriging, 4) GPR-no-feature, 5) GPR-with-features, and 6) point-based. These models were estimated using a training data set consisting of 3675 coverage counts (AADTs) and associated roadway features, and they were validated on a testing data set consisting of 1024 non-coverage counts which were collected specifically for model evaluation purposes. The results indicated that based on the RMSE of the models when applied to the testing data set, the best performing models are, from worst to best: ordinary MLR, hybrid-kriging, point-based, GPR-no-feature, quantile regression, and GPR-with-features. The GPR-with-features model outperformed the ordinary MLR model by 11.35% and it outperformed the GPR-no-feature model by 4.2%. Moreover, the GPR-no-feature model outperformed the hybrid-kriging model by 2.3%. Both GPR models, one without and one with roadway features, performed best with the Exp-Sine-Squared kernel. Moreover, the results showed that based on the MAPE of the models when applied to the testing data set, the best performing models are, from worst to best: ordinary MLR, hybrid-kriging, point-based, quantile regression, GPR-no-feature, and GPR-with-features. The GPR-with-features model outperformed the ordinary MLR model by 26.36% and outperformed the GPR-no-feature model by 17.81%, the point-based model by 22.26%, the regular regression model by 26.36%, and the quantile regression model by 21.07%. These findings suggest that GPR is an appropriate method for estimating non-coverage counts.

The practical significance of this study is that it enables the SCDOT to apply one of the developed models to obtain more accurate estimates of AADT at non-coverage locations than the current method of assuming 100 By combining for rural local roads and 200 vpd for urban local roads. Given that AADT is used in numerous applications such as roadway planning, facility deployment, highway maintenance, infrastructure improvement, and congestion mitigation, any improvement in accuracy would benefit these analyses performed by the SCDOT and its partners, including councils of governments and consultants. A previous study indicated that a 20% improvement in AADT accuracy would result in savings between $60,000 and $200,000 for a safety study (Zarei and Hellinga, 2023). Thus, when considering the multitude of studies that use AADT, the cost savings across all agencies are likely to be in the range of millions of dollars.

This study filled two research gaps in AADT estimation. Notable contributions include evaluating the effectiveness of GPR and quantile regression models against the point-based model, hybrid-kriging model, and ordinary MLR model for estimating non-coverage counts, and comparing the performance of models that use only nearby coverage counts to those that use both coverage counts and roadway features. To generalize the findings of this study, it is suggested that future work uses a sample that spans more than one state. Additionally, other independent variables should be explored to in addition to the six roadway features considered in this study.

Conflict of Interest

Dr. Nathan Huynh is an editorial board member/editor-in-chief for International Journal of Transportation Science and Technology and was not involved in the editorial review or the decision to publish this article. All authors declare that there are no competing interests.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the South Carolina Department of Transportation under the State Planning and Research grant (SPR 749).

Acknowledgement

The authors would like to thank Mr. Todd Anderson and Ms. Stacy Wise from the South Carolina Department of Transportation for their guidance and support throughout the funded study.

References

Anderson, M.D., Sharfi, K., Gholston, S.E., 2006. Direct demand forecasting model for small urban communities using multiple linear regression. Transport. Res. Rec. 1981 (1), 114–117.
Apronti, D., Ksaibati, K., Gerow, K., Hepner, J.J., 2016. Estimating traffic volume on wyoming low volume roads using linear and logistic regression methods. J. Traffic Transport. Eng. (English Ed.) 3 (6), 493–506.
Ashley, G., Attoh-Okine, N., 2021. Bayesian nonparametric approach to average annual daily traffic estimation for bridges. Transp. Res. Rec. 0361198121994591.
Baffoe-Twum, E., Asa, E., Awuku, B., 2022. Estimation of annual average daily traffic (aadt) data for low-volume roads: a systematic literature review and meta-analysis. Emerald Open Res. 4, 13.
Castro-Neto, M., Jeong, Y., Jeong, M.K., Han, L.D., 2009. Aadt prediction using support vector regression with data-dependent parameters. Expert Syst. Appl. 36 (2), 2979–2986.
Chen, P., Hu, S., Shen, Q., Lin, H., Xie, C., 2019. Estimating traffic volume for local streets with imbalanced data. Transport. Res. Rec. 2673 (3), 598–610.
Das, S., Tsapakis, I., 2020. Interpretable machine learning approach in estimating traffic volume on low-volume roadways. Int. J. Transport. Sci. Technol. 9 (1), 76–88.
Doustmohammadi, M., Anderson, M., 2019. A bayesian regression model for estimating average daily traffic volumes for low volume roadways. Int. J. Stat. Probab. 8 (1), 143.
Eom, J.K., Park, M.S., Heo, T.-Y., Huntsinger, L.F., 2006. Improving the prediction of annual average daily traffic for nonfreeway facilities by applying a spatial statistical method. Transport. Res. Rec. 1968 (1), 20–29.
Holik, W.A., Tsapakis, I., Vandervalk, A., Turner, S.M., Habermann, J. et al., 2017. Innovative traffic data qa/qc procedures and automating aadt estimation. Technical report, United States. Federal Highway Administration. Office of Safety.
Keehan, M., 2017. Annual average daily traffic (aadt) estimation with regression using centrality and roadway characteristic variables.
Li, M., 2015. Moving beyond the linear regression model: Advantages of the quantile regression model. J. Manage. 41 (1), 71–98.
Lowry, M., Dixon, M., et al., 2012. Gis tools to estimate average annual daily traffic. Technical report, National Institute for Advanced Transportation Technology (US).
Mohamad, D., Sinha, K.C., Kuczek, T., Scholer, C.F., 1998. Annual average daily traffic prediction model for county roads. Transport. Res. Rec. 1617 (1), 69–77.
Pan, T., 2008. Assignment of estimated average annual daily traffic volumes on all roads in florida.
Pulugurtha, S.S., Kusam, P.R., 2012. Modeling annual average daily traffic with integrated spatial data from multiple network buffer bandwidths. Transport. Res. Rec. 2291 (1), 53–60.
Pulugurtha, S.S., Mathew, S., 2021. Modeling aadt on local functionally classified roads using land use, road density, and nearest nonlocal road data. J. Transp. Geogr. 93, 103071.
Seaver, W.L., Chatterjee, A., Seaver, M.L., 2000. Estimation of traffic volume on rural local roads. Transport. Res. Rec. 1719 (1), 121–128.
Selby, B., Kockelman, K.M., 2013. Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. J. Transp. Geogr. 29, 24–32.
Sfyridis, A., Agnolucci, P., 2020. Annual average daily traffic estimation in england and wales: An application of clustering and regression modelling. J. Transp. Geogr. 83, 102658.
Shamo, B., Asa, E., Membah, J., 2015. Linear spatial interpolation and analysis of annual average daily traffic data. J. Comput. Civil Eng. 29 (1), 04014022.
Sharma, S., Lingras, P., Xu, F., Kilburn, P., 2001. Application of neural networks to estimate aadt on low-volume roads. J. Transport. Eng. 127 (5), 426–432.
Sharma, S.C., Lingras, P., Liu, G.X., Xu, F., 2000. Estimation of annual average daily traffic on low-volume roads: Factor approach versus neural networks. Transp. Res. Rec. 1719 (1), 103–111.
Smith, T.E., 2016. Notebook on spatial data analysis. Lecture Note.
Staats, W.N., 2016. Estimation of annual average daily traffic on local roads in kentucky.
Sun, A.Y., Wang, D., Xu, X., 2014. Monthly streamflow forecasting using gaussian process regression. J. Hydrol. 511, 72–81.
Sun, X., Das, S. et al., 2015. Developing a method for estimating aadt on all louisiana roads. Technical report, Louisiana Transportation Research Center.
Unnikrishnan, A., Figliozzi, M., Moughari, M.K., Urbina, S. et al., 2018. A method to estimate annual average daily traffic for minor facilities for map-21 reporting and statewide safety analysis. Technical report, Oregon. Dept. of Transportation. Research Section.
Vogel, R.M., Wilson, I., Daly, C., 1999. Regional regression models of annual streamflow for the united states. J. Irrigat. Drainage Eng. 125 (3), 148–157.
Wang, J., 2020. An intuitive tutorial to gaussian processes regression. arXiv preprint arXiv:2009.10862.
Wang, T., Gan, A., Alluri, P., 2013. Estimating annual average daily traffic for local roads for highway safety analysis. Transport. Res. Rec. 2398 (1), 60–66.
Xia, Q., Zhao, F., Chen, Z., Shen, L.D., Ospina, D., 1999. Estimation of annual average daily traffic for nonstate roads in a florida county. Transport. Res. Rec. 1660 (1), 32–40.
Yang, B., Wang, S.-G., Bao, Y., 2014. New efficient regression method for local aadt estimation via scad variable selection. IEEE Trans. Intell. Transp. Syst. 15 (6), 2726–2731.
Zarei, M., Hellinga, B., 2023. Method for estimating the monetary benefit of improving annual average daily traffic accuracy in the context of road safety network screening. Transport. Res. Rec. 2677 (3), 445–457.
Zeng, A., Ho, H., Yu, Y., 2020. Prediction of building electricity usage using gaussian process regression. J. Build. Eng. 28, 101054.
Zhang, X., Chen, M., 2020. Enhancing statewide annual average daily traffic estimation with ubiquitous probe vehicle data. Transp. Res. Rec. 2674 (9), 649–660.

Zhao, F., Chung, S., 2001. Contributing factors of annual average daily traffic in a florida county: exploration with geographic information system and regression models. Transport. Res. Rec. 1769 (1), 113–122.
Zhao, F., Park, N., 2004. Using geographically weighted regression models to estimate annual average daily traffic. Transport. Res. Rec. 1879 (1), 99–107.