Turkish Journal of Physiotherapy and Rehabilitation; 32(2)
ISSN 2651-4451 | e-ISSN 2651-446X

IMPROVING THE ESTIMATION ACCURACY OF DATA TRAFFIC USING DATA MINING SPATIAL AUTOREGRESSIVE BENCHMARK MODEL

DR. G. RAMESH[1], DR. CH. MALLIKARJUNA RAO[2], DR.SRIDEVI.R[3], P. NEELIMA[4]

[1]Associate Professor, Department of Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering & Technology Hyderabad- 500090, Telangana State, India. ramesh680@gmail.com

[2]Professor, Department of Computer Science and Engineering Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad 500090, Telangana state, India. chmksharma@yahoo.com

[3]Professor, Dept of Computer Science and Engineering, K.Ramakrishnan College of Engineering, Trichy, Tamil nadu.

[4]Assistant Professor, Dept of Computer Science and Engineering, School of Engineering and Technology , SPMVV University, Tirupati, India neelima.pannem@gmail.com

ABSTRACT:

Nation-wide Annual Average Daily Traffic(AADT)data on NFAS roads across the country are destroyed. Two machine learning methods, the Artificial Neural Network and Random Forest demonstrate a substantial increase in the accuracy of estimating AADT according to five scales that is MSE, RSQ, RMse, MAE and MAPE, using a Spatial Autoregressive Model as a benchmark. An estimated AADT of 87 variables in the area of central, adjacent traffic, population, jobs, land-use diversity, density of road networks, urban design, destination access, etc. is focused on data mining from three aspects, i.e. on road and off-road, network centrality and neighbouring influences. The variable collection for estimates is promoted by aggregation of data by different buffer sizes and linearity and singletonity statistical analysis. The interplay between the variables, variable measurements of significance are extensively explored when applying machine-learning approaches not only the estimation output but also the relationship between and variable and AADT.

I. INTRODUCTION

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

According to 2016 Highway Safety Improvement Program (HSIP) Final Rule, State agencies are required to have access to AADT on all paved roads open to public travel including Non-Federal Aid–System (NFAS) roads by 2026. In traffic monitoring guide, it is stated that AADT should be reported for Highway Performance Monitoring System (HPMS). In practice, traffic data mainly come from two data programs. First is permanent monitoring of continuous traffic flow 24 hours a day and 7 days a week throughout the entire year. Second is temporary monitoring that collects short-duration traffic data several times a year or once at several years, usually based on 24-hour, 48hour, or 72-hour intervals.

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

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

Network (ANN) and Random Forest (RF) from machine learning family comes for AADT estimation. The reasons why machine learning outperforms others regarding this research topic are summarized as below.

A. Machine learning algorithms first and foremost perform well in terms of accuracy,

B. The relationship in real world among the features are usually non-linear. Sometimes it is too complicated to be modelled by mathematical models, where machine learning algorithms can handle high-order relationships,

C. Unlike parametric approaches, machine learning methods do not necessitate any statistical assumptions making it fairly flexible to be applied to data with or without a certain distribution pattern,

D. Machine learning algorithms are good at handling large-volume data just like NFAS roads in a time-efficient way. Not only state-wide estimation but also nationwide estimation becomes feasible,

E. Machine learning algorithms are tolerant of data noises, while statistical models are sensitive to the disturbances from noisy data,

F. The application of machine learning algorithms is simple and easy benefiting from many well-developed and straightforward packages,

G. With more and more techniques that demystify the inner structures of trained model, machine learning gains more and more interpretability instead of being a complete black box.

Encountered with the limited data availability of traffic data, external databases are fully utilized to help estimate AADT. To some extent, traffic volume represents the strength of daily activities, which is closely related to the social demographic features of personals and the environment characteristics such as land use pattern, urban design, accessibility, road design, etc. Accordingly, as much as possible measures of the built-in environment are thoroughly analysed to provide valuable inputs for AADT estimation. In this study, a total of 79 features from Smart Location Database (SLD), a well-structured nationwide database on built-in environment, are extracted for analysis. Additionally, the centrality of roads in the whole transportation network directly influences its capacity of delivering traffic. For example, if the shortest paths in the network are frequently passing one road, the traffic volume of this specific road should be higher than normal roads. Being enlightened from social network theory, this study makes a bold trial of employing centrality measures of road segment to help improve AADT estimation. Furthermore, the interaction among the roads regarding traffic volume is not negligible and hence spatial dependence analysis among roads initiates the involvement of neighbouring traffic features. In summary, all kinds of features from built-in environment are fully discussed to extract strong predictive factors as much as possible. Different ways of integrating spatial data are also compared to enhance the predictive power of variables. Finally, twelve predictors distinguish themselves to act as the potentially good predictors for AADT estimation. This feature exploration process provides an informational guide on future similar studies in this field.

1. Background

The procedure for AADT estimation on lower-level roads by traditional factoring methods involves three steps. First, homogeneous permanent traffic count stations are classified into multiple groups. Then short-term traffic count stations are assigned to these groups. The continuous traffic monitoring data collected by permanent traffic count stations provide all types of factors, including hourly, daily, weekly, and monthly expansion factors. These factors are used for converting the shortduration traffic counts into AADT. This method is widely used across the country for its simplicity, effectiveness, and relatively low cost. However, the traditional factoring method has many deficiencies. Summarizes error sources of factoring approach: determining the number of groups,

identifying groups of road sections, and applying wrong expansion factors [1]. Moreover, accuracy of AADT estimation based on factoring approach is very sensitive to the assignment of STTCs to PTC groups [2].

Inappropriate assignments could lead to high estimation error. Additionally, assigning STTCs has always been a difficulty of this approach. Even though assignment methods, such as agglomerative hierarchical clustering method, k-means clustering method, etc. have been proposed to improve accuracy, they all have various deficiencies [1].

Many researchers propose regression models for AADT estimation and find that factors such as population, area type (rural or urban), per capita income, and roadway characteristics, etc. are significantly correlated with AADT (3, 4, 5). Build linear regression models to estimate the AADT in Broward County, Florida using land-use and accessibility measures [3]. They find that functional class (transformed nominal variables with numeric values) and number of lanes are significant factors in the estimation of AADT. Also suggest that roadway characteristics such as number of lanes and area type are significant factors in AADT estimation[4].

Apply a geographically weighted regression (GWR) model to estimate the AADT of Broward County in Florida [6]. Compared with an ordinary linear regression model, GWR allows the parameters of regression model to be locally-unique instead of globally-uniform. Spatial nonstationarity, meaning the relationship between independents and the dependent varies across the study area, is considered in the GWR model. Improve the general regression model by incorporating spatial statistical process. Three semivariogram models (i.e. Gaussian, exponential, and spherical semivariogram) are compared for analysing the spatial autocorrelation of data points; two interpolation methods (i.e. (ordinary Kriging and universal Kriging) are compared for estimating unknown data points[7]. Apply a linear spatial interpolation to interpolating the AADT in Washington State, in which different combinations of Kriging techniques and variogram models are compared[8]. This spatial modeling stands out for its capturing the spatial relationship among data points by geostatistical procedures[24]. However, the feasibility of the interpolation method depends on not-sparsely-distributed spatial data points. Therefore, its applicability to estimating AADT on local roads at the link level lacks feasibility. [9]Demonstrate that a correlation-based method can yield better estimates than traditional methods if traffic volumes of road sections are significantly correlated with that of nearby roads. They propose a generalized-least-squares (GLS)-estimation-based method and apply it to Ohio intercity network, which is generated by Monte Carlo simulation. Even though the performance of this method in real network needs further investigation, it provides insights on the correlation issue among AADTs. Accordingly, some generalized regression models [10-12] and spatial statistical models [6-8, 13]were explored for estimating AADT.

Spatial statistical models

Spatial autoregressive models (SAM) (e.g. spatial lag model (SLM) and spatial error model (SEM)) give insights into the spatial autocorrelations in an OLS model of AADT estimation. They are more powerful techniques than GWR and they do not have the drawback like the Kriging-based method. Besides, SAM can be applied in various settings as long as spatial autocorrelations cannot be ignored in a normal regression model for AADT estimation. In this paper, spatial autoregressive models are set as the benchmark model for comparison purpose with machine learning algorithms[18][19]. There are various spatial-statistical techniques utilized as revealed in the literature to estimate the AADT on different roadway functional classifications. For example, clusters of roads with similar volume level and functional classification can be created and used to apply spatial interpolation such as Kriging, inverse distance weighting (IDW), neural neighbour (NN), and trend technique. Also, geographically-weighted regression models (GWR) have been proposed and applied for AADT estimation in a few recent studies[20][23]. GWR model assumes that land-use and demographical variables are nonstationary across space, which means the statistical properties (mean, variance, etc.) of variables are different among various locations. Therefore, the relationship between predictors and the response varies; this means that the parameters of dependent variables should not be fixed. At different locations, the influence of predictors on the response varies[17]. For example, car ownership may have a greater influence on AADT at location A than location B. Then, its estimated parameter at location A could be larger than that at location B. The GWR model allows the parameters to be locally- rather than globally-estimated to reflect the mentioned non-stationarity. Local estimations mean that the parameters are more often determined by nearby observations than farther ones. In this way, local variations can be taken into consideration when exploring the relationship between independents and the dependent. Additionally, an important assumption for the GWR model is that the error terms are independent and identically distributed with zero means and constant variance. In the GWR model, a weighted window is moved over the

data to estimate a set of parameters for each data point. Bi-square function and Gaussian function are two commonly-used weighting functions for the GWR model with one critical parameter – bandwidth.

The spatial-statistical method has several advantages. First, it considers the spatial non-stationarity of land-use and demographical variables, which is more reasonable than ordinary regression models when estimating the AADT. In addition, this methodology is economical in its data requirements because it uses existing traffic counts and does not require collection of additional count data. Moreover, spatialstatistical models can be implemented easily in standard GIS software packages that are readily available to local-road agencies. The methodology can also be updated easily in the future after the agencies receive new traffic-count data. Finally, the methodology is straightforward and does not require complex procedures. It can be transferred and adapted rather easily to jurisdictions in other states to estimate their local-road AADTs[22].

Machine learning algorithms

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

2. Data processing

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

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

Feature Selection

A. On-road and Off-road features

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

Simply speaking, either linear relationship or non-linear relationship exists between the predictors and AADT. Considering that common statistical methods are parametric-based while machine-learning methods does not necessitate a specific distribution of the variables, both Pearson Correlation Coefficient (r) and Spearman Rank-order Correlation Coefficient are employed to present the performance of all predictors. Pearson correlation measures the strength of linearity between two variables. Its coefficient falls into the range of minus 1 to 1. The closer the coefficient is to zero, the less linearity the test indicates. Spearman correlation measures the monotonicity between two variables in a non-linear way and it does not necessitate a Gaussian distribution for both variables. For this research, detecting a statistically significant monotonicity by Spearman test contributes a lot to the predictor selection whether such correlation is linear or not since machine learning algorithms applied in this study are able to give full play to theirs advantage of handling non-linear relationships. It seems like Spearman correlation test is enough for selecting variables but Pearson correlation is also necessary since Spearman test might underestimate the linear correlation. In other words, the variables that show significant linearity by Pearson test might show insignificant association by Spearman test. Thus, both tools are utilized to select predictors.

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

B. Network centrality analysis

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

1. Implementation of machine learning algorithms

1.1 Ensembling artificial neural networks

Architecture design

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

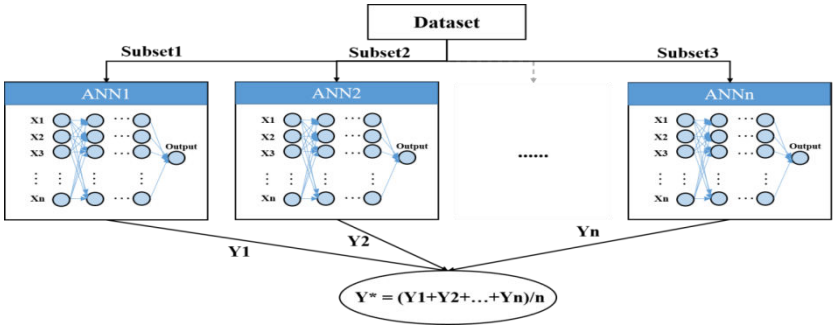

Figure 1 Structure of ensembling artificial neural networks

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

Training results and variable importance measure

The ensembling artificial neural networks consist of ten independent neural networks, each of which is trained by a random sample from the training dataset. The black links denote positive coefficients and the gray links represent negative coefficients. The strength of the linkage is presented by the width of the links. After training, all ten neural networks are shown in figure 2. The importance of each variables is shown through the color depth. The greener the node, the more important the input feature.

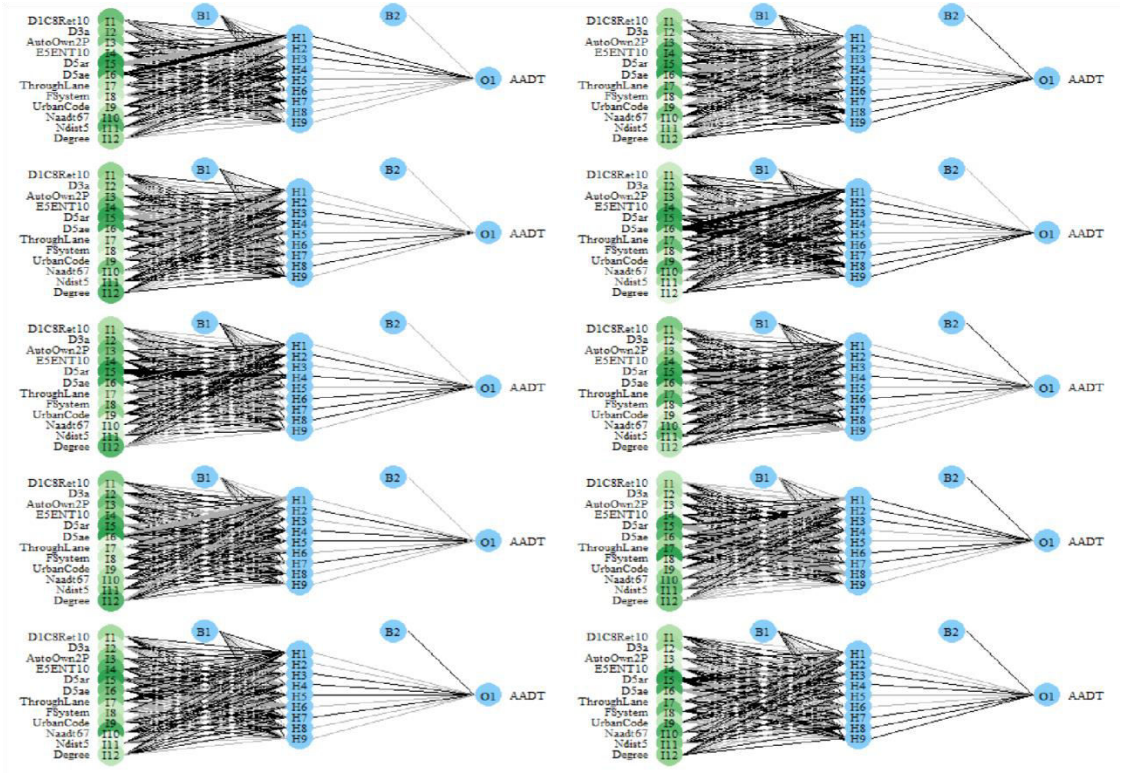

Figure 2 Architecture of artificial neural networks after training

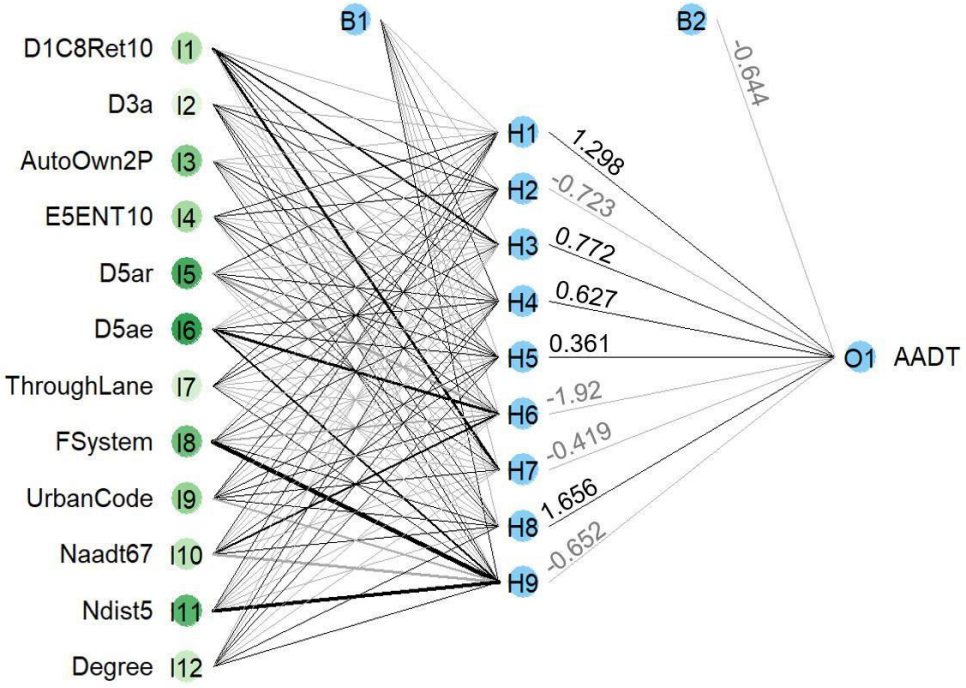

Figure 3 an example of trained neural network and importance rank of inputs

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

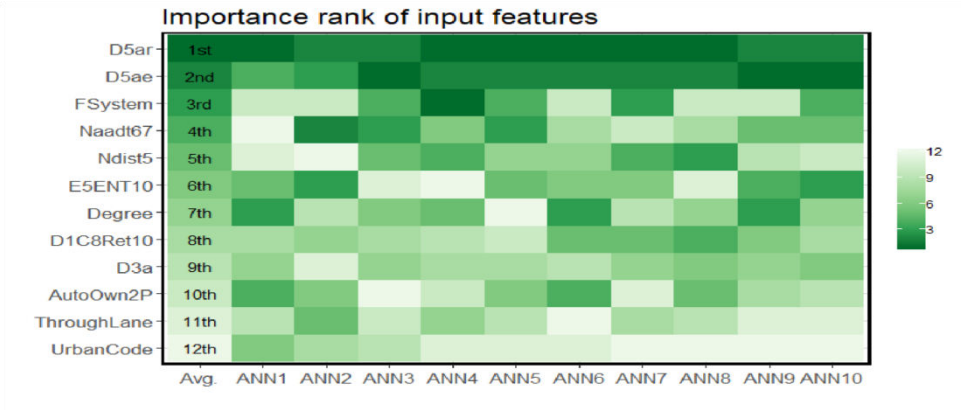

Figure 4 Importance rank of input features of ANN

Accuracy Measures

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

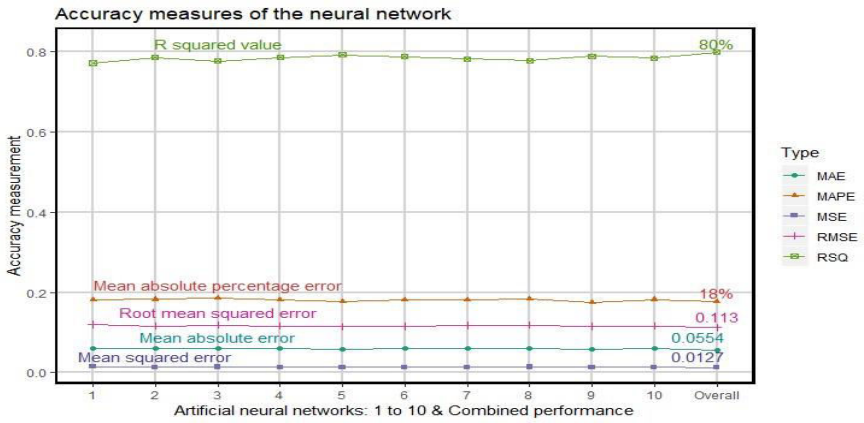

Figure 5 Performance of ensembling ANNs

1.2 Random forest

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

Architecture design

As depicted in figure 6, a random forest consists of a predefined number of decision trees – ntree, the magnitude of which is usually in hundreds. Each tree randomly extracts a portion of the original dataset in a way of bootstrap resampling (sampling with replacement). Then each tree independently makes its own estimation using randomly selected features. The number of features that each node can select is controlled by the second model parameter – mtry. Finally, a voting process takes the estimations from all trees into account and makes the final decision usually by unweighted averaging, which is a bagging process that ensembles hundreds of tree models. Random forest benefits from this bagging feature to provide a more stable and accurate estimation. Adjusting the two parameters, ntree and mtry, contributes to improving the predictive performance. Since the parameter mtry influences the accuracy of each individual tree and simultaneously determines the correlation among the trees in an opposite direction, the model is more sensitive to the mtry value.

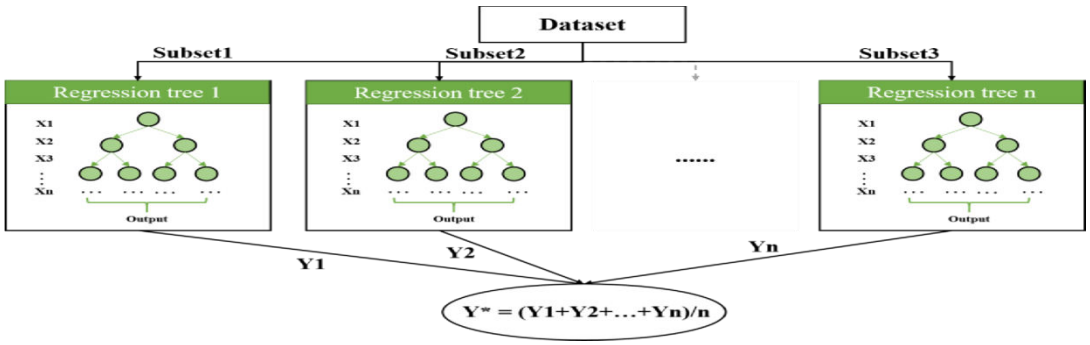

Figure 6 Architecture of random forest

Training results

The whole dataset is divided into training and testing part with a ratio of 80% to 20%. Multiple combinations of ntree and mtry are tested and theirs prediction performances are compared as shown in figure 7. It is obvious that setting mtry as three generates the best result no matter what ntree is. Besides, Mean Squared Residuals (MSE) is the minimal and R Squared value (RSQ) is the maximal when ntree is 500 and mtry equals 3. Thus, this specification of parameters is used for modeling.

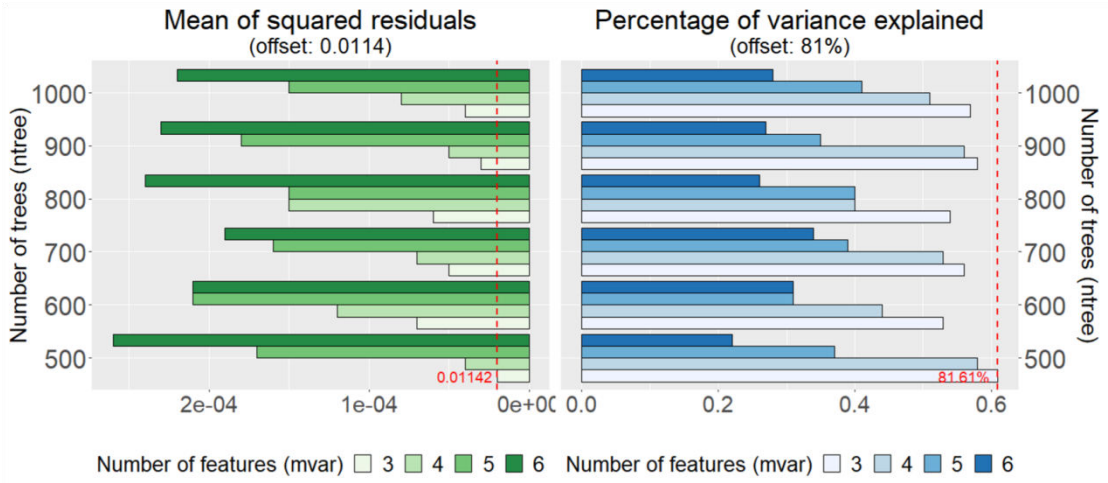

Figure 7 Training results of random forest

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

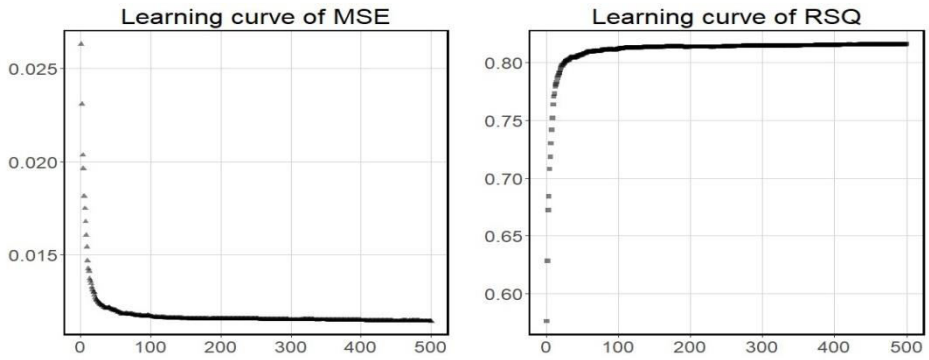

Figure 8 Learning curves on MSE and RSQ of the trained random forest

Interactions between predictors and AADT

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

UrbanCode shows that the urban code increase from 2 (small urban sections) to 3 (urban sections) brings about a larger increase of AADT compared with the change from 1 (rural sections) to 2 (small urban sections). Other predictors overall show a positive influence on AADT even though some fluctuations occur.

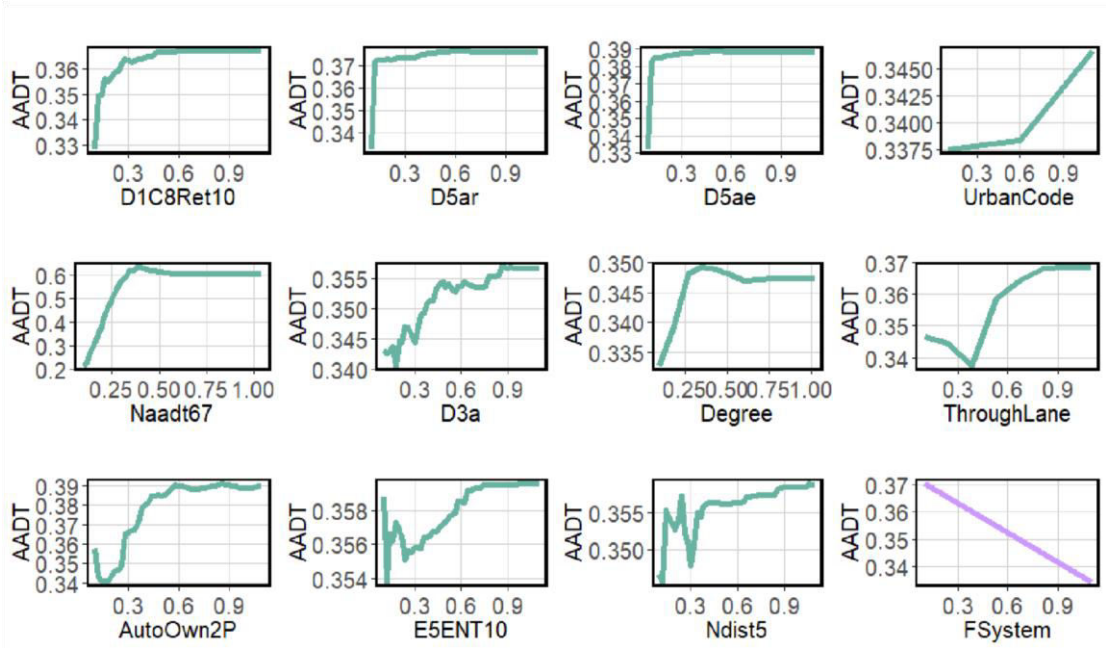

Figure 9 Partial dependence plots

Variable importance measures

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

Permutation-based measure

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

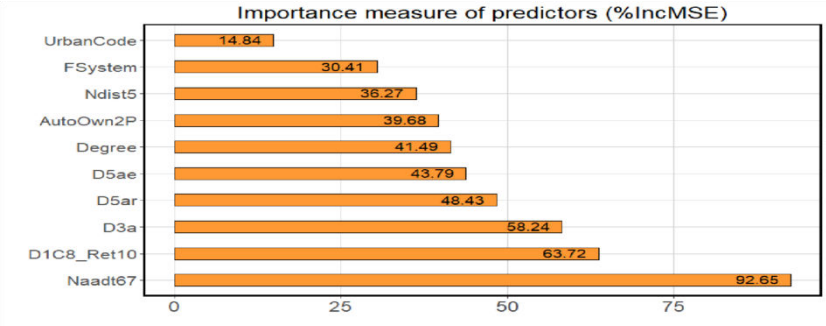

Figure 10 Importance measure by %MSE upon permutation

1.3 Validation

For comparison purpose, the spatial lag model is included as a benchmark. Five accuracy measures in total are calculated for OLS, SLM, ANN, and RF. The estimated AADT values are transformed back to the real value. Results in table 1 show that Random Forest performs the best in each accuracy measure, which is followed by the

Artificial Neural Network. Machine learning algorithms produce noticeably better AADT estimations than OLS and SLM in terms of all accuracy measures. Although SLM is excellent for its handling spatial dependence issue, the estimation result of it still shows a very high mean absolute percentage error (MAPE), i.e. 1.13 and it provides limited improvement when compared with OLS. There is only a 5% decrease of MSE, an 8% increase of RSQ, and a 2% increase of RMSE. When comparing the two machine learning algorithms with SLM, the estimation results are notably improved. The estimation result of RF shows a 57% decrease of MAPE, a 37% decrease of MAE, a 29% decrease of MSE, a 16% decrease of RMSE, and a 10% increase of RSQ. The estimation result of ANN shows a 48% decrease of MAPE, a 28% decrease of MAE, a 20% decrease of MSE, an 11% decrease of RMSE, and an 8% increase of RSQ. These performance measures demonstrate that Random Forest and Artificial Neural Network make better estimations than Spatial Lag Model and Ordinary Least Squares model and Random Forest outperforms Artificial Neural Network in all accuracy measures.

II. CONCLUSION

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

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

Modeling through machine learning is the second part of work. Instead of a simple application of machine learning algorithms, the trained model is demystified from multiple perspectives such as the inner structure after training, the importance measures of predictors, the associations between each predictor and AADT, and the interactions among input features. First, both ANN and RF, two popularly used machine-learning algorithms for prediction, are used for AADT estimation. To increase the robustness of artificial neural network modeling, the ensemble theory, a core structure of random forest, is applied by building up a group of artificial neural networks. Estimation results are more reliable and stable for this assembling structure.

Final estimation results show that both ANN and RF perform well in terms of accuracy. A spatial lag model is built as a benchmark model. Significant improvements in all five accuracy measures including MSE, RSQ, RMSE, MAE, and MAPE can be seen when ANN and RF are compared with the spatial lag model. For example, RF shows a 57% decrease of MAPE and ANN shows a 28% decrease of MAE in comparison with the benchmark model. Additionally, RF performs better than ANN in all accuracy measures. Second, the mysterious mask of machine learning algorithms, usually named as black box algorithms is unveiled largely. How input features interact with AADT are analysed through partial dependence plots. Not only the positive or negative correlation are depicted but also the sensitivity of each input feature to AADT is presented. This enhances the understanding of how predictors act on the AADT. Besides, how the neurons transfer or interplay with each other in the layers of neural network and the importance rank of input features are visualized. Both the strength of interaction and sign (i.e. positive or negative) along the links between neurons are clearly presented as well. For random forest modeling, multiple variable importance measures are utilized, including the percentage increase in MSE upon permuting a given feature, number of root nodes that the given feature splits, and mean minimal depth. All selected predictors show their importance from different aspects. To further uncover the interactions among the

features, conditional mean minimal depth is analysed for each predictor, which shows that some features depend on the presence of other features to make a difference. It is implied that feature selection should also value the combinations of some features.

REFERENCES

1. Rossi, R., M. Gastaldi, G. Gecchele, and S. Kikuchi. Estimation of Annual Average Daily Truck Traffic Volume: Uncertainty Treatment and Data Collection Requirements. Procedia - Social and Behavioural Sciences, 2012. 54: 845-856.
2. Shamo, B., E. Asa, and J. Membah. Linear Spatial Interpolation and Analysis of Annual Average Daily Traffic Aata. Journal of computing in civil engineering, 2015. 29: 04014022(1)-04014022(8).
3. Azad, A. K., and X. Wang. Prediction of Traffic Counts Using Statistical and Neural Network Models. Geomatica, 2015. 69: 217-284.
4. Duddu, V. R., and S. S. Pulugurtha. Principle of Demographic Gravitation to Estimate Annual Average Daily Traffic: Comparison of Statistical and Neural Network Models. Journal of Transportation Engineering, 2013. 139: 585-595.
5. Tsapakis, L., W. H. Schneider, and A. P. Nichold. A Bayesian Analysis of the Effect of Estimating Annual Average Daily Traffic for Heavy-duty Trucks Using Training and Validation Data-sets. Transportation Planning and Technology, 2013. 36: 201-217.
6. Selby, B., and K. M. Kockelman. Spatial Prediction of Traffic Levels in Unmeasured Locations: Applications of Universal Kriging and Geographically Weighted Regression. Journal of Transport Geography, 2013. 29: 24–32.
7. Karlaftis, M.G., and E.I. Vlahogianni. Statistical Methods versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights. Transportation Research Part C, 2011. 19: 387-399.
8. Smith, K. A., and J. N.D. Gupta. Neural Networks in Business: Techniques and Applications for the Operations Researcher. Computers & Operations Research, 2000. 27: 1023-1044.
9. Sharma, S., P. Lingras, F. Xu, and P. Kilburn. Application of Neural Networks to Estimate AADT on Low-volume Roads. Journal of Transportation Engineering, 2001. 127: 426-432.
10. Pulugurtha, S. S., and P. R. Kusam. Modeling Annual Average Daily Traffic with Integrated Spatial Data from Multiple Network Buffer Bandwidths. Transportation Research Record: Journal of the Transportation Research Board, 2012. 2291: 53-60.
11. Zhang, L., J. Hong, A. Nasri, and Q. Shen. How Built Environment Affects Travel Behaviour: A Comparative Analysis of the Connections between Land Use and Vehicle Miles Travelled in US Cities. The Journal of Transportation and Land Use, 2012. 5: 40-52.
12. Heo, T. Y., M. S. Park, J. K. Eom, and J. S. Oh. A Study on the Prediction of Traffic Counts Based on Shortest Travel Path, the Korean Journal of Applied Statistics, 2007. 20: 459-473.
13. Eom, J., M. Park, T. Heo, and L. Huntsinger. Improving the Prediction of Annual Average Daily Traffic for Nonfreeway Facilities by Applying a Spatial Statistical Method. Transportation Research Record: Journal of the Transportation Research Board, 2006. 1968: 20-29.
14. Zarei N., Ghayour M.A., Hashemi S. (2013) Road Traffic Prediction Using Context-Aware Random Forest Based on Volatility Nature of Traffic Flows. In: Selamat A., Nguyen N.T., Haron H. (eds) Intelligent Information and Database Systems. ACIIDS 2013. Lecture Notes in Computer Science, vol 7802. Springer, Berlin, Heidelberg.
15. B. Hamner, "Predicting Travel Times with Context-Dependent Random Forests by Modeling Local and Aggregate Traffic Flow," 2010 IEEE International Conference on Data Mining Workshops, Sydney, NSW, 2010, pp. 1357-1359.
16. G. Ramesh, (2020) "A Survey on NLP based Text Summarization for Summarizing Product Reviews," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 352-356, doi: 10.1109/ICIRCA48905.2020.9183355.
17. Dr. G. Ramesh (2020). A Survey on Hybrid Machine Translation, 2nd International Conference on Design and Manufacturing Aspects for Sustainable Energy (ICMED 2020), , Volume 184, August, 2020.
18. Dr. Gajula Ramesh (2020). Detection of Plant Diseases by analyzing the Texture of Leaf using ANN Classifier. International Journal of Advanced Science and Technology, 29(8s), 1656 – 1664.
19. Dr. G. Ramesh (2020). Data Storage in Cloud Using Key-Policy Attribute-Based Temporary Keyword Search Scheme (KP-ABTKS). Lecture Notes in Networks and Systems Volume 98 pp. 630–636, 2020.
20. Ramesh G. (2021) A Machine Learning Enabled IoT Device to Combat Elephant Mortality on Railway Tracks. Lecture Notes on Data Engineering and Communications Technologies, vol 59. Springer, Singapore. https://doi.org/10.1007/978-981-15-9651-3_44.
21. G. Ramesh, (2021) A Machine Learning-Based IoT for Providing an Intrusion Detection System for Security. Microprocessors and Microsystems, Volume 82, 103741. (Elsevier).
22. Natarajan, V.A., Kumar, M.S., Patan, R., Kallam, S. and Mohamed, M.Y.N., 2020, September. Segmentation of Nuclei in Histopathology images using Fully Convolutional Deep Neural Architecture. In 2020 International Conference on Computing and Information Technology (ICCIT-1441) (pp. 1-7). IEEE.
23. Sreedhar, B., Manjunath Swamy BE, and M. S Kumar. "A Comparative Study of Melanoma Skin Cancer Detection in Traditional and Current Image Processing Techniques." In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 654-658. IEEE, 2020.