

Comparison of Estimation Methodologies for Daily Traffic Count Prediction in Small and Medium Sized Communities

Mohammad Shojaeshafiei^{1,*}, Mehrnaz Doustmohammadi², Suraj Subedi¹, Michael Anderson²

¹Department of Computer Science, The University of Alabama in Huntsville, Huntsville, USA

²Department of Civil and Environmental Engineering, The University of Alabama in Huntsville, Huntsville, USA

Abstract Annual Average Daily Traffic (AADT) is a critical input to many transportation analyses, including maintenance, safety and capacity improvements. Due to cost limitations, AADT data is not typically collected for every roadway in a community. However, the necessity of having quality AADT data estimate for the purpose of making decisions is not subject to only making decisions on roadways with AADT values. This paper presents two machine learning techniques (K* Classification and Random Forest) intended to improve AADT estimation and compares the results to linear regression models. This research was conducted to identify models that can accurately estimate AADTs within a small or medium sized community. The data elements for the models use a combination of roadway and socio-economic factors near the desired count location. The models were tested using statistical tests to ensure the robustness of the models and validated to additional data collected for the community. The results of the paper indicate that the linear regression model was the best option for the communities considered in the study area, and that the machine learning models did not improve the ability to estimate AADT due to over specifying the training datasets.

Keywords AADT estimation, Machine Learning, Small Communities

1. Introduction

Annual Average Daily Traffic (AADT) is a critical input to many transportation analyses, safety assessments, maintenance schedules, capacity improvements etc. AADT is defined as the average 24-hour volume at a highway location over a full year. The amount of labor and the associated costs required to collect actual AADT data for every roadway in a community, even for a small area, are so high that it is unfeasible that any community would be able to gather AADT data for every roadway in the community. However, the importance of AADT is not diminished for roadways where the actual AADT collections are not taken, and often a comparison of data must be made between selected roadways in a community with incomplete AADT values. For example, one area where having quality AADT data is valued is in the calculation of crash rates on small and medium sized community roadways and the implementation of crash reduction factors. In order to accurately identify areas in which high crash levels occur and to successfully appropriate money wisely on potential

improvements, accurate AADT data is required.

The demand for quality AADT data on small and medium sized community roadways, coupled with the lack of available accurate data, has prompted this research to examine the use of machine learning to see if a sample of traffic counts can be used to train an instance-based classification algorithm and develop accurate estimates of AADTs. The machine learning techniques will be compared to traditional linear regression models and data transformed linear regression models that have been used previously. All of the models used in this research will predict AADT using a combination of roadway and socio-economic factors near the desired count location. The paper's objective is to determine the quality of the predicted AADTs and determine which model most accurately reflects actual AADTs in the community for transportation decision making purposes. In this study, AADTs from four cities were collected and randomly assigned into a development group and a validation group (70 percent development and 30 percent validation). Then prediction models were developed using linear regression techniques and two machine learning algorithms to determine the best option for predicting AADTs. The results of the paper indicate that simple linear regression model was superior to the transformed model and machine learning models for the data collected and used in the study.

* Corresponding author:

ms0083@uah.edu (Mohammad Shojaeshafiei)

Published online at <http://journal.sapub.org/ijtte>

Copyright © 2017 Scientific & Academic Publishing. All Rights Reserved

2. Literature Review

The collection and development of models for the development and estimation of AADT is not a novel concept [1-5]. Over the past thirty years, the focus of predicting traffic counts was to determine the best methodology for developing AADT counts from minimal data sources. The literature review summarizes several studies that have been performed to conduct AADT estimates for communities through the development of linear regression models and then presents a review of a machine learning technique that has the potential to improve count estimation by “learning” how different factors influence count values to make better predictions than linear models are capable of providing.

A report developed by Garber presented methodologies to optimize the time and cost associated with the collection of traffic count data [6]. The result of the study shows that the collecting data on Monday, Tuesday and Wednesday provided better AADT estimation due the stability of traffic during these days [6]. However, these models were based on traditional count methodologies and still required the expense associated with field data collections. An AADT prediction model for county roadways was developed over two decades ago for application in Indiana [7]. The model utilized regression analysis and 11 independent variables to predict AADT for county roads, as the Indiana Department of Transportation, like most state departments of transportation, does not monitor traffic counts on county roadways [7]. The methodology presented was designed to provide an efficacious method for highway departments to accurately estimate county roadway AADTs. A study in Florida conducted by Zhao used statistical analysis and regression to determine the contributing factors for AADT estimation [8]. The selected variables in the model included functional classification, number of lanes and access to regional employment canters, directness of expressway access, and population and employment proximal to the count location [8]. A limitation of the study was the inability to transfer the models to other urban areas [8]. An alternate methodology was proposed and utilized step-wise regression to relate socio-economic data for each county in Florida with the total roadways in the respective county [9]. While the statistical application of the stepwise regression was applauded, the overall lack of quality roadway count data to support the development of the models results in large prediction errors and the models suffered from limited applicability [9]. Developing estimated AADT for safety analysis was proposed by Wang to allow for safety studies to be performed on minor roadways where no count had been previously conducted [10]. The model developed operated similarly to a trip generation model where the parcel level data in the location near the area the count was desired was used to estimate the AADT [10]. This model was shown to have good error measurements and was shown to be applicable for safety analyses; however, there was no discussion of transferability to other locations. Using Geographic Information Systems (GIS) to collect data for

estimating AADT have been shown to be effective [11]. The study by Lowry used linear regression and variables that included number of lanes and speed limit to predict AADT [11]. The results of the report indicate that the models developed were of statistical validity to be used in practical application [11]. A paper by Doustmohammadi *et al.* presented a direct AADT estimation model the used lanes, function classification and population and employment data collected near the count location [12]. A companion paper that examined logarithmic transformations of the variables was also developed by Doustmohammadi [13].

Machine learning techniques attempt to use a set of response and independent variables to train an algorithm to better understand and see patterns in data. There were two machine learning tools used in this paper.

K-STAR (K*), which is an entropic distance measure algorithm intended to quick learn patterns and be able to predict responses using a multi-step approach [14-16]. The advantage of K* for the traffic count estimation is related to its ability to see patterns beyond linear regression, and even linear regression models with transformations [14]. K* has been used in a variety of applications for prediction models, however, this is first reviewed attempt to use this tool to support traffic count estimation.

The random forest is an ensemble approach in which the results from the multiple decision trees are combined together to predict the class of the test instance [17]. It uses the mode class output from number of decision trees to predict the final class of the test instance. A random forest model is actually a collection of decision trees each of which is created using a subset of the original training dataset [17]. The subset of the dataset used to generate a tree is randomly selected from the whole dataset. Once the forest is created for all random subsets of data, the model is said to be ready. The classification of the test instance is done by traversing each decision tree in the forest to predict the class of the instance. The mode class of the prediction obtained from each decision tree traversal is said to be the final class of the test instance.

The literature reviewed in this paper supports this effort by comparing methodologies for estimating AADTs from a variety of known sources with a new model that has the potential to provide more insightful analysis of the data through machine learning techniques. The tools identified: linear regression, transformed linear regression and machine learning will all be utilized in this work to produce the best models.

3. Data Collection

The purpose of this paper is to test different tools to develop AADT estimation models for small and medium sized community roadways or locations where traffic counts have not been conducted using new tools then previously attempted. AADT data were collected for four cities in Alabama (Florence, Gadsden, Huntsville and Montgomery).

All the cities were considered small to medium communities with population ranging from 50,000 to 200,000. A total of 235 counts were collected, of which 165 were used to develop models and train the machine learning tools and 70 were used as validation data points to determine which model best performed.

The data necessary to support the development of the models were collected and managed using ArcGIS, a commercially available desktop geographic information system (GIS) program. Within ArcGIS, a database was developed that contained roadway characteristics (number of lanes and functional classification) as well as multiple layers of socio-economic data: population obtained from the Census Department and retail and non-retail employment at specific business locations obtained from the communities serving as case study locations that were used in the long-range transportation plans recently completed by the individual communities. The actual data collected included the real AADT, total number of lanes for the roadway, roadway functional classification, population within a 0.25-mile buffer of the count location, retail employment within a 0.25-mile buffer of the count location, and non-retail employment within a 0.25-mile buffer of the count location.

To quantify the roadway functional classification for inclusion in the multiple linear regression models, the following convention was used:

- Collector road = 1,
- Minor arterial = 2, and
- Principal arterial = 3.

This convention was selected as the highest functionally classified roadway would have the greatest value and this should have the desired effect that the parameter for this variable would be non-negative. Additionally, it should be noted that freeways and interstates were explicitly not included in the study or model development because the surrounding socio-economic variable would not be relevant for these roadways because they would not have the access to the facilities due the controlled nature of the roadway system.

The response variable is:

1. Traffic volume; AADT

The predictors (independent) variables are:

2. Function classification of the road; FCLASS
3. Number of lanes; LANE
4. Population within a 0.25-mile buffer around a traffic count station; POPBUFF
5. Retail Employment within a 0.25-mile buffer around a traffic count station; RETAILEMPBUFF
6. Non-Retail Employment within a 0.25-mile buffer around the traffic count station; NONRETAILEMPBUFF.

4. Model Development

Previous work of Doustmohammadi [13] had tried to model the problem as a regression problem in which the regression parameters were used to predict the ADT value. In this paper, we tried to use a different approach of modeling the problem. Here, we have tried to model the problem as instance based classification problem and we present two different methods: K* Algorithm and Random Forest. Our assumption is that due to the lack data, the previous model of Doustmohammadi [13] was not able to properly tune the parameters of the regression model. So, we wanted to solve the problem with the help of non-parametric models such as instance based classifiers.

Multiple linear regression analysis and machine learning tools were used in this study to produce AADT forecasting models.

All the statistical analyses were conducted using Minitab and in accordance with standard statistical methodologies [18]. Regression analysis was selected as this methodology is used to predict the value of one or more responses, AADT in this study, from a set of predictors, roadway and socio-economic variables [18]. For the data collected, the AADT estimation models developed were as follows:

Linear:

$$\begin{aligned} \text{AADT} = & -10,286 + 5,918 * \text{FCLASS} + 3,436 * \text{LANE} \\ & - 1.97 * \text{NONRETAILEMPBUFF} \\ & + 5.17 * \text{RETAILEMPBUFF} \end{aligned}$$

Logarithmic Transformed:

$$\begin{aligned} \text{AADT} = & -8,827 + 4107 * \text{FCLASS} + 4,763 * \text{LANE} \\ & - 807 * \text{LN}(\text{POPBUFF}) \\ & + 788 * \text{LN}(\text{NONRETAILEMPBUFF}) \\ & - 376 * \text{LN}(\text{RETAILEMPBUFF}) \end{aligned}$$

The K* algorithm and Random Forest do not produce a model in a conventional fashion as linear regression does, but do use the initial data to train the search methodology for future data analysis.

5. Statistical Validation

Validation of the four models was conducted using a set of data collected from cities separate from the data set used to build the models.

The first test performed was a Mann Whitney U Test. This test was performed because the validation data were not normally distributed, thus requiring a non-parametric test [18]. The results of the test indicated that all models examined produced AADT estimates that were not statistically different from the actual counts. The quality of the models measured by P-value, showed a value of 0.48 for the linear model, and 0.28 and 0.27 for the machine learning models, K* and Random Forest, respectively.

Additionally, the validation of the models was conducted using an R-squared value for the validation dataset and the prediction. The values obtained from an analysis showed a value of 0.72 for the linear model, 0.68 for the logarithmic model, and 0.07 and 0.75 for the K* and Random Forest, respectfully.

To further test the models statistically, a Nash-Sutcliffe (N-S) statistic was calculated to test the model ability to accurately predict the traffic. The N-S statistic is calculated as:

$$E = 1 - (\sum(Q_o - Q_m)^2 / (\sum(Q_o - Q_{ave})^2)$$

E is the Test Statistic

Q_o is the value of actual count

Q_m is the predicted count

Q_{ave} is the average value of all the actual counts.

The Nash-Sutcliffe coefficient ranges from 1 to negative infinity and measures the accuracy of the model to the actual values and compares the results using the average of the traffic counts. For comparing multiple models, the model with the highest calculated coefficient is the model that provides the most accurate estimate of the actual data. The calculated coefficients for the different models are 0.72 for the linear model, 0.68 for the logarithmic model, 0.58 for the K* model, and 0.46 for the Random Forest model.

6. Conclusions

This paper examined the use of linear regression techniques and machine learning techniques to estimate AADT from roadway and socio-economic data near a location where an estimate of the AADT is desired. The thought behind the paper is that the machine learning techniques would provide a better estimated AADT due to the algorithms within the models and the ability to “learn” patterns that might not be linear or straightforward from classical statistic techniques.

The results of this work indicate that for the data collected and tested, the classical statistical techniques were shown to perform better than the machine learning techniques. While this result does not diminish the ability of the machine learning tools, it does emphasize a limitation of the machine learning. More accurate AADT models were developed using other, simpler techniques. One aspect that can be understood from the analysis is that the machine learning tool can over-learn the training data and once applied to a different data set, with differences, the machine learning result can be too specified and therefore make a poor estimation.

ACKNOWLEDGMENTS

This paper was made possible by funding provided by the Alabama Department of Transportation.

REFERENCES

- [1] Gecchele, G., Rossi, R., Gastaldi, M., and Caprini, A. (2011). "Data Mining Methods for Traffic Monitoring Data Analysis: A case study." *Procedia - Social and Behavioral Sciences*, 10.1016/j.sbspro.2011.08.052, 455-464.
- [2] Zhong, M. and Liu, G. (2007). "Establishing and Managing Jurisdiction-wide Traffic Monitoring Systems: North American Experiences." *Journal of Transportation Systems Engineering and Information Technology*, 10.1016/S1570-6672(08)60002-1, 25-38.
- [3] Sharma, Satish C., Brij M. Gulati, and Samantha N. Rizak. "Statewide traffic volume studies and precision of AADT estimates." *Journal of Transportation Engineering* 122.6 (1996): 430-439.
- [4] Sharma, S., Lingras, P., Xu, F., and Kilburn, P. (2001). "Application of Neural Networks to Estimate AADT on Low-Volume Roads." *Journal of Transportation Engineering*, 10.1061/(ASCE)0733-947X(2001)127:5(426), 426-432.
- [5] Zhao, Fang, and Nokil Park. "Using geographically weighted regression models to estimate annual average daily traffic." *Transportation Research Record: Journal of the Transportation Research Board* 1879 (2004): 99-107.
- [6] Garber, Nicholas J. "A Methodology for Estimating AADT Volumes from Short-Duration Counts." *Virginia Highway & Transportation Research Council*. 1984.
- [7] Dadang Mohamad, Kumares C. Sinha, Thomas Kuczek, Charles F. Scholer. "Annual verage Daily Traffic Prediction Model for County Roads, Transportation Research Board 2013 Annual Meeting. 1998.
- [8] Zhao, Fang and Soon Chung; "Contributing Factors of Annual Average Daily Traffic in a Florida County", *Transportation Research Board*, 2001.
- [9] Tao Pan. "Assignment of estimated average annual daily traffic on all roads in Florida" *University of South Florida*, 2008. J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," *Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep.* 99-02, 1999.
- [10] Wang, Tao; Gan, Albert; Alluri, Priyanka. "Estimating Annual Average Daily Traffic (AADT) for Local Roads for Highway Safety Analysis." *Transportation Research Board 2013 Annual Meeting*. 2012.
- [11] Lowry, Michael; Dixon, Michael. "GIS Tools to Estimate Average Annual Daily Traffic." *National Institute for Advanced Transportation Technology. University of Idaho*. 2012.
- [12] Doustmohammadi, M. and Anderson, M.D. "Developing Direct Demand AADT Forecasting Models for Small and Medium Sized Urban Communities" *International Journal of Traffic and Transportation Engineering*. Vol 5, No. 2. Pp 27-31. 2016. Owen, D. R. J., and Hinton, E., 1980, *Finite elements in plasticity-theory and practice*, Pineridge Press, Swansea.

- [13] Doustmohammadi, Mehrnaz, Michael Anderson, and Ehsan Doustmohammadi. "Using Log Transformations to Improve AADT Forecasting Models in Small and Medium Sized Communities." *International Journal of Traffic and Transportation Engineering* 6.2 (2017): 23-27.
- [14] Frank, Eibe, et al. "Weka-a machine learning workbench for data mining." *Data mining and knowledge discovery handbook*. Springer US, 2009. 1269-1277.
- [15] Cleary, John G., and Leonard E. Trigg. "K*: An instance-based learner using an entropic distance measure." *Proceedings of the 12th International Conference on Machine learning*. Vol. 5. 1995.
- [16] <http://weka.sourceforge.net/doc.dev/weka/classifiers/lazy/KStar.html>.
- [17] Breiman Leo, Schapire E. Robert, "Random Forests", "Kluwer Academic Publisher", *Machine Learning*, 45, 5-32, 2001.
- [18] Montgomery, Douglas, Elizabeth A. Peck, G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley ISBN-1119180171. 2012.