Journal Home: https://journals.gmu.edu/index.php/jmms
Mason Journals ● Mason Publishing Group
Journal of Modern Mobility Systems (JMMS), Vol. 01(2020), 125-130
Document: https://doi.org/10.13021/jmms.2020.2931

Identifying Road Links and Variables Influencing the Applicability of Variable Speed Limits Using Supervised Machine Learning and Travel Time Data

Sarvani V. Duvvuri, M.S.,[a] Sonu Mathew, Ph.D.,[a] Raghuveer Gouribhatla, M.S.,[a] Srinivas S. Pulugurtha, Ph.D., P.E., F.ASCE[a,*]

[a] *The University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223; *Corresponding Author, sspulugu@uncc.edu*

Received: 2 Dec. 2020 ● **Revised** 29 Dec. 2020 ● **Accepted**: 29 Dec. 2020 ● **Published**: 29 Dec. 2020

ABSTRACT

With increasing congestion and associated challenges to manage the transportation network, intelligent transportation systems (ITS) have gained popularity due to their data-driven approach and application of advanced technologies. A variable speed limit (VSL) is a popular ITS-based solution which uses dynamic speed limit to promote harmonization along a corridor. However, not much was done in identifying road links and influencing variables for their applicability. Therefore, this paper focuses on examining road link-level data to identify road links and variables influencing the applicability of VSL signs. A multivariate cluster analysis was first used to identify potential road links susceptible to speed variation for the implementation of VSL. A supervised machine learning algorithm, forest-based classification and regression, was then used to model and examine the influence of average annual daily traffic (AADT), historical speed of the road link, and the speeds of upstream and downstream road links on the average speed of the corresponding road link. Modelling and validation were performed using data for Mecklenburg County, North Carolina, USA, for road links including all kinds of speed variation.

Keywords: Intelligent transportation systems, Variable speed limit, Supervised machine learning, Big data.

1 INTRODUCTION

The posted speed limits on roads are typically determined based on the road design, operating speed, geometry, and type of the facility [*1*]. The Federal Highway Administration (FHWA) describes traffic congestion as a direct measure of vehicle speeds [*2*], referred to as speed in this paper. A consistent and significant decrease in speeds on a road link indicates severe recurring congestion on the road link [*3*].

Dynamic message signs with variable speed limits (VSLs) is a widely explored intelligent transportation systems (ITS)-based solution to regulate the speeds on highly congested road segments, in work zone areas, during adverse weather conditions, or during incidents on a road [*4*]. The VSL control strategy also improves mobility and safety in adverse weather conditions [*5*].

The VSLs are estimated dynamically based on the traffic condition and optimized to improve the road capacity. Researchers in the past proposed various algorithms to compute the VSLs. They include simulation-based approach [*6*], cellular transmission models using bottleneck information [*7*], macroscopic simulation [*8*], algorithms like fuzzy logic with simulation-based validation [*9*], and model predictive control [*10*].

One of the most important aspects of the VSLs is the extent to which the speed limit is changed. A significant increase or decrease in the speed limits might raise a concern. Many researchers set thresholds while modelling the speed limit. Abdel-Aty et al. [*6*] used 5 mph increments for the road facilities while Hegyi et al. [*11*] considered a threshold of ± 6.2 mph to ensure safer stream performance. State agencies implementing the VSL signs used thresholds up to 7.5 mph (New Mexico), 30 mph (New Jersey), 10 mph (Washington State), or increments of 10 mph (Nevada) [*4*]. From a safety perspective, the maximum changes to the speed limit of a facility could be up to 10 mph [*6*].

The existing VSL signs use algorithms to generate the speeds needed for the corresponding time of the day and day of the week. Some of the simplest algorithms used include the display of speeds in increments of 5 mph based on the 85th percentile speeds [*12*]. Assigning the algorithm or technique to improve the traffic flow is one of the most common challenges due to its dynamic nature. Further, speeds of the upstream and downstream road links have an influence on a road link speed and should be accounted for in the VSL design process [*12*].

VSL may not be applicable to all the road links. It is important to analyse the patterns in travel times and examine the historical data when computing the speeds for VSL signs. Past research on the dynamic travel time predictions used pattern recognition using the probe data [*8*]. The VSLs from simulation models could be different from what may be observed using the field data. It is, therefore, important to identify the road links which are susceptible to higher variation in speeds using the field data.

Supervised machine learning is designed to forecast using a

training dataset and is applicable to even model non-linear relationships. It has the potential to identify the road links with a significant variation in speeds from the posted speed limits and is considered appropriate for this type of "big data" application. Therefore, the objectives of this research are to compute the variability in speeds, identify vulnerable road links, and apply a supervised machine learning algorithm to examine the influence of selected explanatory variables on speed patterns.

2 STUDY AREA, DATA, AND RESEARCH METHOD

Mecklenburg County in the State of North Carolina, USA was considered as the study area for this research. The travel time data and their corresponding network data such as the annual average daily traffic (AADT) and functional class of the road were considered for analysis.

The travel time data was obtained from Regional Integrated Transportation Information System (RITIS) with support from the North Carolina Department of Transportation (NCDOT). The data consists of raw travel times with samples collected at a 1-minute interval for each road link identified by the traffic message channel (TMC) code. The raw travel time data for March of the year 2019 during the peak period was processed using Microsoft SQL Server. The 85th percentile speed and the average speed of considered road links were computed, and the variations were examined for the corresponding analysis hour. Furthermore, data associated with the corresponding upstream and downstream road links were also considered for the analysis.

The research method adopted is two-fold. Firstly, cluster analysis was performed to identify the groups of road links with speed variations by comparing the 85th percentile and average speeds. Secondly, the influence of selected explanatory variables on the average speed of a road link was examined using forest-based classification and regression.

The K-means clustering was used in this research. The algorithm establishes thresholds to minimize the heterogeneity in speeds. It identifies the initial seeds randomly based on the number of allocated clusters, while the other seeds are typically allocated by employing a random component [*13*].

Datasets for the forest-based classification and regression analysis comprised of all the road links, road links with low-speed variation, and road links with high-speed variation. These separate datasets were considered for modelling and analysing the importance of the selected explanatory variables.

The forest-based classification and regression algorithm trains the model data [*14*], estimates the dependent variable (the average speed in this research), and helps understand the speed patterns of roads using the selected explanatory variables. The mechanism of the forest-based classification and regression includes the usage of hundreds of randomly generated trees to predict the average speed. Hence, the result from each tree contributes to the overall accuracy of the model. For higher data points, the tree-based mechanisms are suggested [*14*].

The selection of explanatory variables for analysis and modeling plays a major role in the predictability and understanding their influence on the dependent variable. All the explanatory variables are typically selected to develop a model and assess their influence on the dependent variable in forest-based classification and regression [*15*]. Therefore, the correlation between the explanatory variables was not examined in this research.

The influence of the explanatory variables is computed based on the prediction accuracy using the training dataset. For example, each decision tree in the model uses a certain portion of data to train and generate the outcomes. The remaining data is used to compute the influence and importance of each explanatory variable in predicting the dependent variable, by estimating the decrease in the prediction accuracy [*15*]. In general, a higher value indicates a higher degree of the explanatory variable's importance in the model prediction.

Modelling was performed with 80% of the data and the remaining 20% of the data were used for the validation. The functional class of the road, AADT, historical average speed of the road link, downstream road link average speed, and upstream road link average speed are considered as the selected explanatory variables.

3 RESULTS AND DISCUSSION

Data for 563 road links in the study area were considered for analysis in this research. The study area and the road links are shown as Figure 1. Tables 1 and 2 summarize the descriptive statistics (minimum, median, mean, maximum, and standard deviation) and frequency distribution of the variables considered in this research, respectively.

3.1 Cluster Analysis Results

A total of six clusters were defined by using the optimal R^2 value. The box whisker plot (Figure 2) shows the clusters along with the variations associated with the 85th percentile speed and the average speed for the analysed road links. The low-speed variation comprised of clusters with variation ranging from -7.8 mph to 4.0 mph. The remaining clusters with large variation in the speed on negative side were categorised as "high-speed variation" dataset. The spatial distribution of road links based on the defined clusters are shown in Figure 3.

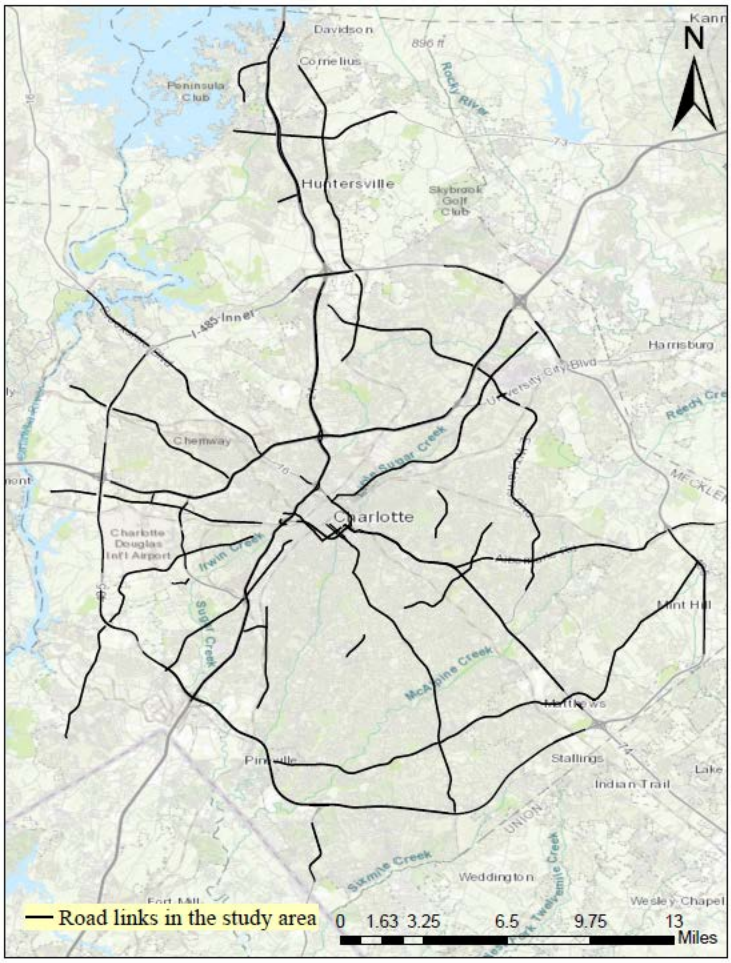

Figure 1. Study area

Table 1. Descriptive statistics of data

Variable	Min.	Median	Mean	Max.	Std. Dev.
Average speed	7.14	40.77	43.83	73.40	16.56
85th percentile speed	14.00	42.00	45.57	70.00	15.47
Difference between the 85th percentile and average speeds	-22.40	0.13	1.74	31.40	7.80
Historical average speed	5.80	40.57	43.84	73.40	16.33
Upstream average speed	7.47	40.00	43.61	73.73	16.63
Upstream reference speed	12.00	42.00	45.37	70.00	15.56
Downstream average speed	7.14	40.47	43.41	74.47	16.44
Downstream reference speed	10.00	42.00	45.33	70.00	15.37
AADT	3700	52000	73160	183000	50626

Table 2. Frequency distribution by facility type

Variable	Categories	Frequency	Percentage
Functional class	1: Interstate	242	42.98
	2: Principal Arterial - Other Freeways and Expressways	15	2.66
	3: Principal Arterial - Other	278	49.38
	4: Minor Arterial	27	4.80
	5: Major Collector	1	0.18
Number of through lanes (in both the travel directions)	2	50	8.88
	3	2	0.36
	4	255	45.29
	5	8	1.42
	6	129	22.91
	8	104	18.47
	10	11	1.95
	12	7	1.24

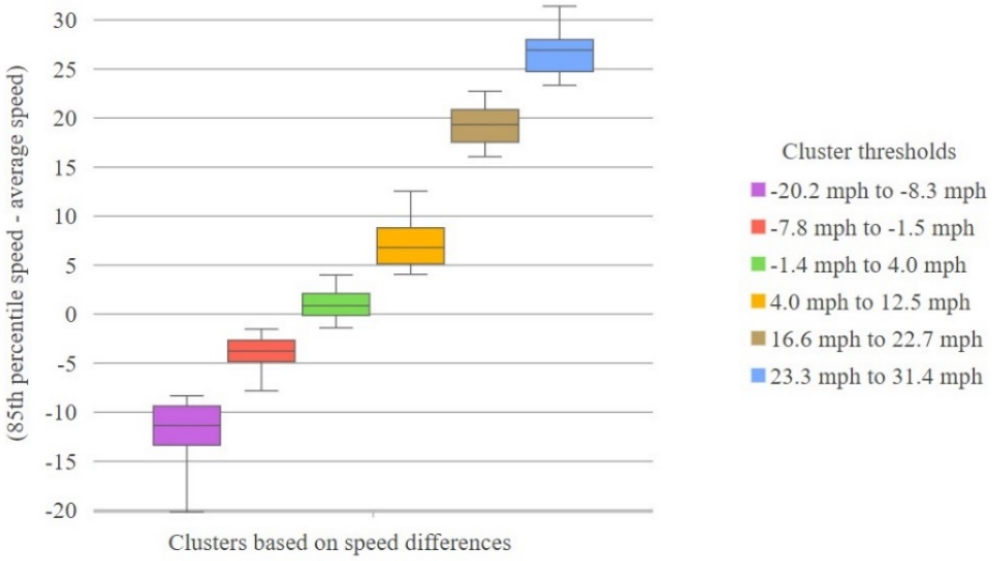

Figure 2. Multivariate cluster analysis results

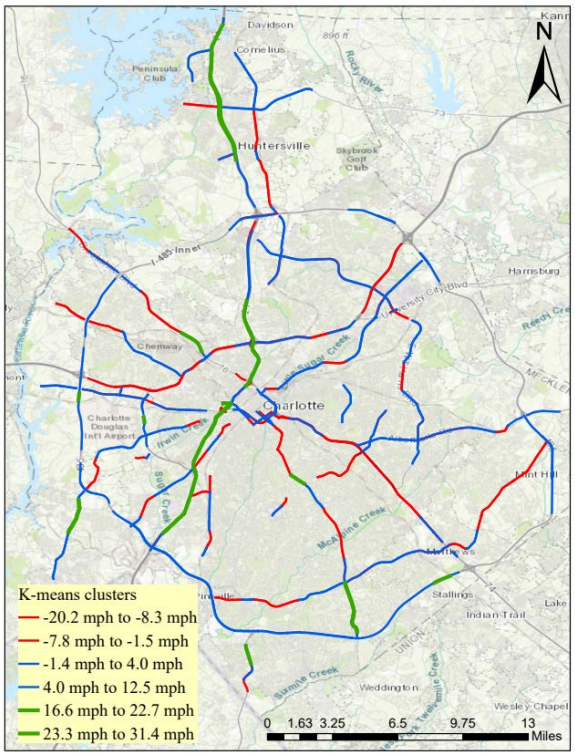

Figure 3. Spatial distribution of clusters in the study area

3.2 Classification and Regression Results

The results from the application of forest-based classification and regression using the three datasets are summarized in tables 3 and 4.

Table 3 shows the importance of the selected explanatory variables from the model results in terms of percentages. The historical average speed is the most important explanatory variable, followed by the upstream and downstream road link average speeds, in the model associated with all road links dataset. In the low-speed variation dataset-based model, the functional class followed by the historical average speed and AADT are the most important explanatory variables. However, the model

results from the dataset with high-speed variation dataset indicate that all the selected explanatory variables are important, with the historical average speed of the road link being the most important explanatory variable.

Table 3. Explanatory variables and their importance in terms of percentages

Explanatory variable	Modelling dataset		
	All data	Low-speed variation	High-speed variation
Functional class	7.57	45.05	5.61
AADT	1.25	18.86	14.59
Historical average speed	47.46	22.29	40.49
Upstream average speed	26.48	9.85	20.57
Downstream average speed	17.25	3.96	18.74

Table 4. Predictability results

Parameter / Measure	Modelling dataset		
	All data	Low-speed variation	High-speed variation
R^2	0.94	0.97	0.92
Mean percentage error (%)	-2.96	-1.38	-11.96
Mean absolute percentage error (%)	11.64	6.11	20.33
Root mean square error (in mph)	4.88	3.04	6.77

The predictability results (Table 4) from the forest-based classification and regression indicate a high R^2 value (>0.90) for all the three models (which explains the variability in each dataset). The mean percentage error varied between -1.38% and -11.96%, while the mean absolute percentage error varied between 6.11% and 20.33%. The root mean square error varied between 3.04 mph and 6.77 mph. The errors are highest for the high-speed variation dataset, followed by all the road links dataset. This could be attributed to the low sample size and/or variations in the explanatory variables.

4 CONCLUSIONS

This research explores cluster analysis and the plausible application of machine learning algorithms like the forest-based classification and regression to analyse the speed patterns on road links and assess the applicability of VSLs for congestion mitigation and transportation network management. Travel time data and selected network characteristics for road links in Mecklenburg County were considered in this research. The multivariate cluster analysis was performed to identify groups of road links with varying speeds by comparing the 85th percentile and average speeds. Datasets with all road links as well as road links with low- and high-speed variation were considered to model using the forest-based classification and regression algorithm and examine the influence of the selected explanatory variables.

The functional class of a road and AADT are the most important explanatory variables in the models associated with low- and high-speed variation datasets. However, the functional class of a road and AADT are the least important explanatory variables in the model associated with all the road links dataset. The historical average speed of the road link and upstream road link average speed are the most important explanatory variables irrespective of the dataset considered for modeling in this research.

The R^2 values are high and errors are relatively low, indicating the predictability and potential applicability of supervised machine learning algorithms for determining VSLs. The relatively high errors for high-speed variation dataset indicate that other explanatory variables and more data should be used for analysis and modeling. Furthermore, thresholds for the applicability of VSLs by area type and functional class of a road should be explored in the future.

This research proposes and illustrates the working of a method for identifying vulnerable links and implementing VSLs using travel time data and supervised machine learnig. Researching the applicability of VSLs using larger travel time datasets for even more number of links with varying road and traffic characteristics, by day of the week and time of the day, merits further investigation.

5 ACKNOWLEDGMENTS

The paper is partially based on ideas and information collected for projects funded by the United States Department of Transportation - Office of the Assistant Secretary for Research and Technology (USDOT/OST-R) University Transportation Centers Program (Grant # 69A3551747127) and NCDOT. The authors thank the NCDOT, the city of Charlotte Department of Transportation, and RITIS for providing access to the data needed for this research.

6 DISCLAIMER

This paper is disseminated in the interest of information exchange. The views, opinions, findings, and conclusions reflected in this paper are the responsibility of the authors only and do not represent the official policy or position of the University of North Carolina at Charlotte or other entity. The authors are responsible for the facts and the accuracy of the data presented herein. This paper does not constitute a standard, specification, or regulation.

REFERENCES

1. M. Papageorgiou, M. Ben-Akiva, J. Bottom, P. H. Bovy, S. P. Hoogendoorn, N. B. Hounsell, , ... and M. McDonald, "ITS and traffic management," *Handbooks in Operations Research and Management Science*, vol. 14, pp. 715-774.
2. Cambridge Systematics, "Traffic congestion and reliability: Trends and advanced strategies for congestion mitigation," Report No. FHWA-HOP-05-064, Federal Highway Administration, 2005, Retrieved from https://ops.fhwa.dot.gov/congestion_report/.

3. K. Fitzpatrick, B. Shamburger and D. Fambro, "Design speed, operating speed, and posted speed survey. *Transportation Research Record*, vol. 1523, no. 1, pp. 55-60, 1996.
4. M. Robinson, "Examples of variable speed limit applications," Speed Management Workshop at 79th Annual Meeting of Transportation Research Board, Washington, D.C., 2000.
5. M. Hadiuzzaman, T. Z. Qiu, and X. Y. Lu, "Variable speed limit control design for relieving congestion caused by active bottlenecks," *Journal of Transportation Engineering*, vol. 139, no. 4, pp. 358-370, 2013.
6. M. Abdel-Aty, J. Dilmore, and A. Dhindsa, "Evaluation of variable speed limits for real-time freeway safety improvement," *Accident Analysis & Prevention*, vol. 38, no. 2, pp. 335-345, 2006.
7. M. Hadiuzzaman, and T. Z. Qiu, "Cell transmission model based variable speed limit control for freeways," *Canadian Journal of Civil Engineering*, vol. 40, no. 1, pp. 46-56, 2013.
8. R. Yu and M. Abdel-Aty, "An optimal variable speed limits system to ameliorate traffic safety risk," *Transportation Research Part C: Emerging Technologies*, vol. *46*, pp. 235-246, 2014.
9. A. H. Ghods, A. R. Kian, and M. Tabibi, "A genetic-fuzzy control application to ramp metering and variable speed limit control," In 2007 IEEE International Conf. on Systems, Man and Cybernetics, pp. 1723-1728, Montreal, Que., Canada, 2008.
10. H. Chen, H. A. Rakha, and C. C. McGhee, "Dynamic travel time prediction using pattern recognition," In *20th World Congress on Intelligent Transportation Systems*, TU Delft, 2013.
11. A. Hegyi, B. De Schutter, and J. Hellendoorn, "Optimal Coordination of Variable Speed Limits to Suppress Shock Waves". *IEEE Transactions on Intelligent Transportation Systems*, vol. 6(1), pp. 102-112, 2005.
12. B. Katz, J. Ma, H. Rigdon, K. Sykes, Z. Huang, K. Raboy, and J. Chu, "Synthesis of Variable Speed Limit Signs," Report *No. FHWA-HOP-17-003*, Federal Highway Administration, 2017, Retrieved from https://ops.fhwa.dot.gov/publications/fhwahop17003/fhwahop17003.pdf.
13. ArcGIS Pro, Esri. Multivariate Clustering Analysis, Retrieved from https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/how-multivariate-clustering-works.htm.
14. ArcGIS Pro, Esri, Forest-based Classification and Regression, Spatial Statistics, Retrieved from https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/forestbasedclassificationregression.htm.
15. F. Degenhardt, S. Seifert, and S. Szymczak. "Evaluation of variable selection methods for random forests and omics data sets". Briefings in Bioinformatics, vol. 20(2), pp. 492-503, 2019.