Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 45 (2012) 824 – 829

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Symposium on Safety Science and Technology

Analysis of freeway accident frequency using multivariate adaptive regression splines

Li-yen CHANG [a,*], Hsing-chung CHU [a], Da-jie LIN [b], Pei LUI [b]

[a]*Graduate Institute of Marketing and Logistics/Transportation, National Chia-Yi University, 580 Sin-Min Road, Chia-Yi 60054, Taiwan, China*
[b]*Department of Transportation Technology and Management, Feng Chia University, 100 Wen-Hwa Road, Taichung 40724, Taiwan, China*

Abstract

This study applies nonparametric multivariate adaptive regression splines (MARS) modeling technique to explore the effects of non-behavior factors including highway geometrics characteristics, traffic factors as well as environmental conditions on freeway accidents. 2007-2008 accident data of National Freeway 1 in Taiwan were collected for analysis. The results indicate that horizontal alignment, vertical alignment, average daily traffic volume (ADT), heavy vehicle ADT and annual precipitation have are nonlinear effects on freeway accidents.

Keywords: freeway; accident frequency; linear regression; data mining; multivariate adaptive regression splines (MARS)

1. Introduction

How to accurately predict the probability of accidents and identify the risk factors contributing to accidents have attracted considerable research interest for many decades. Following the past studies concerning to identify the contributing factors for highway accidents, this study focuses on the effects of non-behavioral factors, specifically highway geometric characteristics, traffic factors and environmental conditions, on highway accidents. Non-behavior factors not only can contribute to certain types of driver errors (e.g., speeding often occurs at downgrades) or crashes (e.g., overturn often occurs at curve segments), but accidents will be likely to occur at the same location repeatedly if the problem is not mitigated. In addition, with a better understanding of factors contributing to highway accidents, highway engineers will be able to design highways to higher safety standards.

To identify the risk factors contributing to highway accidents, statistical regression analysis, such as linear regression models, Poisson regression or negative binomial regression models, has been commonly applied to identify the relationship between accidents and risk factors. With this information, the specific locations or highway segments with higher accident frequencies can be located by the highway authorities. Mitigation or safety measures, such as illumination or enforcement, can be effectively applied. However, most regression models have their own model assumptions and pre-defined underlying relationship between response and predictor variables. If these assumptions are violated, the model could lead to erroneous estimation of accident likelihood as well as the relationship between risk factor and highway accidents. Multivariate adaptive regression splines (MARS) is a widely applied non-parametric technique that has been employed in many areas including medicine, business administration, industry, and engineering. With its advantage of exploring the complex nonlinear relationship between a response variable and predictor variables by fitting the data into a series of spline functions of the predictor variables, MARS analysis has been shown to be an effective tool particularly in dealing with prediction

* Corresponding author. Tel:+886-5-2732932; fax: +886-5-2732933.
E-mail address: liyen@mail.ncyu.edu.tw

1877-7058
doi:10.1016/j.proeng.2012.08.245

problems. However, the applications of MARS to analyze traffic safety problems have been relatively fewer. Therefore, this study adopts MARS modeling technique. The primary objective of this study is to identify the effects of non-behavior factors including highway geometrics characteristics, traffic factors as well as environmental conditions on freeway accidents.

2. Methodology

MARS is a commonly applied non-parametric modeling approach. MARS has been widely employed in the scientific fields. In contrast to well-known parametric linear regression analysis, MARS provides greater flexibility to explore the nonlinear relationship between a response variable and predictor variables by fitting the data into piecewise linear regression functions. In other words, the nonlinear relationship between a response variable and predictor variables is approximated by the use of separate regression slopes in distinct intervals of the predictor variable region. Therefore, the slope of the regression function is allowed to change from one interval to the other. The program checks all possible predictor variables as well as all possible intervals of the predictor variables to find the best fit of the data. In addition, MARS also searches for interactions between variables by checking all degrees of interactions. Because it allows for all functional forms and interactions, MARS is able to effectively track the complex data structures hidden in high-dimensional data. The general MARS function can be expressed using the following equations:

$$\hat{f}(x) = a_0 + \sum_{m=1}^{M} a_m B_m(x) \tag{1}$$

where $\hat{f}(x)$ is the predicted response, a_0 and a_m are parameters which are estimated to yield the best data fit and m is the number of basis functions included into the model. The basis function in MARS can be either one single spline function or a product of two or more spline functions for different predictor variables. The spline basis function, $B_m(x)$, can be specified as

$$B_m(x) = \prod_{k=1}^{k_m} \left[s_{km} \left(x_{v(k,m)} - t_{k,m} \right) \right]_+ \tag{2}$$

where km is the number of knots, skm takes either 1 or -1 and indicates the right/left regions of the associated step function, v(k,m) is the label of the predictor variable and tk,m is the knot location.

An optimal MARS model is developed through a two-stage forward/backward procedure. In the forward stage, MARS overfits the data by considering a great number of basis functions and all possible interactions among the predictor variables. In the backward stage, basis functions and interactions are removed in order of the least contribution to the accuracy of the fit by using the generalized cross-validation (GCV) criterion. The decrease of the calculated GVC values can be used to assess the importance of variables. The GCV be expressed as follows.

$$\mathrm{GCV(M)} = \frac{1}{N} \sum_{i=1}^{N} \left[y_i - \hat{f}_M(x_i) \right]^2 / [1 - C(\tilde{M}) / N]^2 \tag{3}$$

where N is the number of observations, $[1 - C(\tilde{M}) / N]^2$ is the penalty measure of model complexity, and $C(\tilde{M})$ is defined as

$$C(\tilde{M}) = C(M) + d * M \tag{4}$$

where $C(M)$ is the number of parameters being fit and d is another penalty factor between 2 and 4 as a typical value. More detailed description of MARS analysis and its applications can be found in Friedman et al.[1].

3. The data

National Freeway 1 is selected as the study area. National freeway 1, connecting most of the major cities and metropolitan areas, is considered as the most important transportation corridor in Taiwan. National Freeway 1 is a tolled freeway of 373 kilometers long and contains 47 interchanges and 10 mainline toll plazas.

The data used for this study include information on vehicle accidents, highway geometric design characteristics, traffic characteristics and environmental conditions of National Freeway 1. The vehicle accident data were taken from National

Traffic Accident Investigation Reports provided by the Ministry of Transportation and Communications (MOTC). The report records detailed information of each accident including the time and location of accident, involved driver's characteristics, primary causes of the accident, and injury levels of occupants. The data were obtained in a computer-ready form, which contains coded information on the reported accidents. During 2007–2008, 891 severe crashes occurred resulting in death or injury. Information regarding these 891 accidents was extracted for this study. The highway geometric design information and traffic data were supplied by the Taiwan Area National Freeway Bureau. The highway geometric design information includes vertical grade, horizontal curvature, number of lanes, and lane width, while traffic information includes ADT of various vehicle types and traffic distribution over lanes. The weather information was taken from the annual report of climatological data by the Central Weather Bureau. This report records detailed weather information of cities and towns along National Freeway 1 including pressure, temperature, humidity, precipitation, and wind speed.

With these data, the next step is to divide the study area into manageable highway sections. Two common alternatives used in previous studies for determining highway section length include the use of fixed-length sections or homogeneous sections (in terms of geometric characteristics). To avoid potential effect caused by the length of highway section, this study adopts the fixed-length approach. A detailed discussion on advantages and disadvantages between these two alternatives can be found in Shankar et al.[2]. According to this approach, the study area was first divided into fixed-length (1 km long) sections. Northbound and southbound highway sections are considered separately due to the opposite values of vertical alignment. After screening out the north and south end sections due to different operational characteristics (e.g., reduced speed limits and signal control at the end of freeway), 373 kilometers of freeway were disaggregated into 742 sections. The geometric characteristics of each highway section are determined according to the characteristic with the largest proportion. For example, a highway section has composite grade of 1% for 700 meters and 2% for 300 meters. The 1% grade was selected for the vertical alignment of this highway section. It should be noted that past studies have indicated that lane width, shoulder width and medium type have significant effects on accident occurrence. Because most of the highway sections have the same lane width and detailed information on shoulder width and median types was unavailable, this study could not examine their effects on accident frequencies. The summary statistics of these 1484 highway sections (i.e., each section produces two observations for 2007 and 2008) are presented in Table 1. The observed annual accident frequency on these highway sections ranges from 0 to 7, and the average frequency per year is 0.60.

Table 1. Sample summary of statistics of characteristics of road sections

	Minimum	Maximum	Mean	Standard deviation
Accident frequency (per year)	0	7	0.60	0.88
Degree of horizontal curve (angle, in degree, subtended by a 100 m arc, equal to 18,000/(π×Radius))	0	14.3	1.66	1.99
Vertical grade (percent)	-5.3	5.3	0	1.44
ADT/lane (vehicles)	24,329	113,380	54,972	19,723
Heavy vehicle ADT (vehicles)	3957	18412	8933	3205
Number of days with precipitation	88	208	122.42	30.15
Annual precipitation (millimeters)	1521	4064	2283.75	442.53

4. Analysis results

Eleven non-behavior predictor variables were used with the response variable of accident frequency in an attempt to identify the relationships that traffic engineers wish to understand. These eleven predictor variables include highway geometric characteristics (e.g., horizontal alignment, vertical alignment, number of lanes), traffic variables (e.g., ADT), and environmental characteristics (e.g., annual precipitation). Table 2 gives the definition of the predictor variables. To explore factors affecting freeway accident frequency, the present study developed a MARS model. Software R is used to estimate the MARS models. Table 3 summarizes the variable selection results using MARS. In a MARS model, basis functions are used to predict the effects of independent variables on accident frequency. The interpretation of MARS results is similar to but not as straightforward as that of classical linear regression models. A positive sign for the estimated coefficients for the basis function indicates increased accident frequency, while a negative sign indicates decreased accident frequency. The value of coefficient indicates the magnitude of effect of the basis function (i.e., variable effect) on the accident frequency. As the effect of each basis function, max (0, x–t) is equal to (x–t) when x is greater than t; otherwise the basis function is equal to zero. For example, for the BF1 in Table 3, if (GRADE–2.3) is greater than 0, then the value of BF1 is equal to

(GRADE–2.3) and the BF1 is equal to 0, if (GRADE–2.3) is less than 0. Therefore, if the vertical grade of a highway section is 2.5%, then the MARS model predicts the accident frequency of this highway section is increased by 0.172 (i.e., 0.858×(2.5–2.3)).

Table 2. Description of variables

Variable	Symbol	Type	Description
Accident frequency	FREQ	Count	The dependent variable
Horizontal alignment	CURVE	Continuous	Angle, in degree, subtended by a 100 m arc, equal to 18,000/(π×Radius)
Vertical alignment	GRADE	Continuous	Grade in percent
Fog zone	FOGZONE	Qualitative	1, fog zone; 0 otherwise
Interchange dummy	INTER	Qualitative	1, interchange; 0 otherwise
Toll plaza dummy	TOLL	Qualitative	1, toll plaza; 0 otherwise
Military section dummy	MILITARY	Qualitative	1, military section; 0 otherwise
Number of lanes	LANES	Count	Number of moving traffic lanes in the section
ADT (per lane) (vehicles)	ADT/LANE	Continuous	Average daily traffic volume per lane
Heavy vehicle ADT (vehicles)	HEAVY	Continuous	Average daily bus, single-unit truck and semi tractor-trailer volume
Annual precipitation (in millimeters)	P	Continuous	The total amount of precipitation in one year
Precipitation days in a year	PDAY	Continuous	Number of days with precipitation

As shown in Table 3, the first MARS model contains 15 basis functions that are single spline defined by only one explanatory variable. It can be observed that seven variables play crucial roles in determining freeway accident frequencies. These variables include vertical alignment, horizontal alignment, number of lanes, fog zone, AADT/lane, heavy vehicle ADT, and annual precipitation. The BF1, BF2, and BF3 account for the nonlinear effect of vertical alignment on accident frequency. It can be distinguished that the effect of grade on accident frequency is four-fold: (1) if the grade of the highway section is less than 2.3%, then grade has no effect on accidents (indicated by BF1); (2) if grade of the highway section is greater than 2.3% but less than 3%, the accident frequency will increase by 0.858 for 1% increase in grade (indicated by BF1 and BF2); (3) if grade of the highway section is greater than 3% but less than 5%, the accident frequency will decrease by 1.404 for 1% increase in grade (indicated by BF2 and BF3); and (4) if grade of the highway section is greater than 5%, the accident frequency will increase by 6.508 for 1% increase in grade (indicated by BF3). Grade can significantly influence vehicle operation speed particularly large trucks and buses. The effect of speed differentials is an important contributing factor for accidents. The estimated results indicate that the effect of grade on accidents can be maximal when sections with over 5 % severe uphill grade. Considering the effect of horizontal curve on accidents, the degree of horizontal curve, which is defined as the angle subtended by a 100 meter arc along the horizontal curve, is used in this study. The degree of horizontal curve can directly measure the sharpness of the curve. A larger degree indicates a sharper alignment. BF4 to BF7 show the nonlinear effect of horizontal curve on accidents and this effect is also four-fold: (1) if the degree of horizontal curve is less than 0.60, the accident frequency will reduce by 0.266 for additional increase in degree of horizontal curve (indicated by BF4); (2) if the degree of horizontal curve is greater than 0.60, but less than 7.20, the accident frequency will decrease by 0.047 for additional increase in degree of horizontal curve (indicated by BF5 and BF6); (3) if the degree of horizontal curve is between 7.20 and 8.20, the accident frequency will decrease by 0.699 for additional increase in degree of horizontal curve (indicated by BF6 and BF7); and (5) if the degree of horizontal curve is greater than 8.20, the accident frequency will increase by 0.992 for additional increase in degree of horizontal curve (indicated by BF7). The MARS result shows increased accident frequency for the sections with degree of horizontal curve greater than 8.20. This finding is expected because the maneuvers are relatively difficult at freeway segments with a sharp horizontal curve. In addition to the horizontal and vertical alignments, the positive sign of the BF8 indicates that if the number of lanes is greater than 3, the predicted accidents will increase by 0.333 for an additional traffic lane. When the number of traffic lanes increases (greater than 3 in this case), lane changing as well as the conflicts between vehicles are expected to increase.

Table 3. List of basis functions of the MARS and their coefficients

Variables	Basis Function	Coefficient
Constant		1.7247
BF1	Max(0, GRADE-2.3)	0.8579
BF2	Max(0, GRADE-3)	-1.4042
BF3	Max(0, GRADE-5)	6.5082
BF4	Max(0.6-CRUVE, 0)	-0.2662
BF5	Max(0, CRUVE-0.6)	-0.0474
BF6	Max(0, CRUVE-7.2)	-0.6987
BF7	Max(0, CRUVE-8.2)	0.9920
BF8	Max(0, LANES-3)	0.3329
BF9	Max(0, ADT/LANE-17142)	0.000038
BF10	Max(0, ADT/LANE-30264)	0.000064
BF11	Max(0, HEAVY-15805)	0.001462
BF12	Max(0, HEAVY-16989)	0.005187
BF13	FOGZONE	0.1370
BF14	Max(3015.9- P, 0)	-0.0002
BF15	Max(0, P-3015.9)	0.0019

As for the effects of traffic factors on accidents, ADT and heavy vehicle ADT are found to have effects on accidents. As indicated by BF9 and BF10, the effects of ADT on accident occurrence is three-fold: (1) if the ADT is less than 17,142 vehicles/lane then ADT have no effect on accidents (indicated by BF9); (2) if the ADT is greater than 17,142 vehicles/lane but less than 30,264 vehicles/lane, the accident frequency will increase by 0.038 for additional 1,000 vehicles in the traffic lane (indicated by BF9 and BF10); and (3) if the ADT is greater than 30,264 vehicles/lane, the accident frequency will increase by 0.046 for additional 1,000 vehicles in the traffic lane (indicated by BF10). Negative impact by the traffic is expected, because more vehicles on the road, more conflicts between vehicles are resulted. This finding is also consistent with previous finding. In addition, BF11 and BF12 show the effect of heavy vehicle on accidents. Heavy vehicles in this study include single-unit trucks, semi tractor-trailers and buses. The effect of heavy vehicle is also found to be three-fold: (1) if the heavy vehicle ADT is less than 15,805, then heavy vehicles have no effect on accidents (indicated by BF11); (2) if the heavy vehicle ADT is greater than 15,805 but less than 16,989, the accident frequency will increase by 1.462 for additional 1,000 heavy vehicles in the traffic stream (indicated by BF11 and BF12); and (3) if the heavy vehicle ADT is greater than 16,989, the accident frequency will increase by 5.187 for additional 1,000 heavy vehicles (indicated by BF12). Based on the estimated coefficient of basis functions, heavy vehicles have a nonlinear negative effect on the frequency of accidents and this effect can be greatest when the heavy vehicle ADT is greater than 17,989. Negative impact of heavy vehicles on accidents is well documented[3-4]. This finding is also expected and consistent with previous finding. As indicated by BF13 and BF14, the MARS results show the effect of precipitation is two-fold: (1) when annual precipitation is less than 3015.9mm, the accident frequency will decrease by 0.0002 with additional reduction in millimeter of precipitation and (2) when annual precipitation is greater than 3015.9mm, the accident frequency will increase by 0.0019 with additional increase in millimeter of precipitation. Drivers driving in a rainy condition receive poor visibility and require longer braking distance. Higher accident frequency is expected for highway sections with higher annual precipitation. However, the detailed data on rain intensity and time of raining are unavailable. To have a better understanding of the effect of rain on accidents, a more detailed study is suggested. Fog has been considered as an important contributor to vehicle accidents, because it can significantly reduce drivers' visibility. In Taiwan, there was a major accident caused by fog on this freeway in 1996 involving 99 vehicles and resulting in three deaths and 23 injuries. The Taiwan Area National Freeway Bureau has identified specific freeway sections as fog zones and installed more traffic safety facilities such as illumination, flashing lights, and warning signs. To identify if there is higher accident frequencies in fog zones, an indicator variable is selected for these zones. As indicated by BF15, the positive coefficient indicates that accident frequency tends to increase in fog zones. The estimated accident frequency for fog zones is 0.137 higher than that of non-fog zones.

5. Conclusions

Non-parametric MARS models were proposed to establish the empirical relationship between traffic accidents and highway geometric characteristics, traffic factors and environmental conditions. MARS is a multivariate nonparametric regression splines method that estimates complex nonlinear relationships by a series of spline functions of the predictor variables and proved to be applicable to the field of accident analysis. This represents an important methodological step in analyzing traffic accident frequency. The results obtained here, by showing that horizontal alignment, vertical alignment, average daily traffic volume (ADT), heavy vehicle ADT and annual precipitation have nonlinear effects on freeway accidents, provide helpful insight into the underlying relationship between risk factors and vehicle accidents. In terms of future work, an application of the methodological approaches used in this paper to analyze the injury severity of traffic accident would be interesting. As discussed previously, statistical models such as logit and ordered probit models were the commonly employed techniques in injury severity analysis. Further exploration by non-parametric modeling techniques might provide a better understanding of the factors that can affect the injury severity of traffic accidents. It would also be interesting to apply different data mining techniques such as association rules or random forest, to explore the factors that affect accident frequency. If more potential risk factors can be effectively uncovered, highway engineers could further promote the safety of highway operations.

References

[1] Friedman, J. H., 1991. Multivariate adaptive regression splines. The Annals of Statistics, 19(1) p. 1-141.
[2] Shankar, V. N., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accident Analysis and Prevention 27(3), p. 371-389.
[3] Chang, L.-Y., Chen, W.-C., 2005. Data Mining of Tree-based Models to Analyze Freeway Accident Frequency. Journal of Safety Research, 36(4), p. 365-375.
[4] Karlaftis, M. G., Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accident Analysis and Prevention 34(3), p. 357-365.