

# Using Artificial Neural Network to Predict Collisions on Horizontal Tangents of 3D Two-Lane Highways

Omer F. Cansiz and Said M. Easa

**Abstract**—The purpose of this study is mainly to predict collision frequency on the horizontal tangents combined with vertical curves using artificial neural network methods. The proposed ANN models are compared with existing regression models. First, the variables that affect collision frequency were investigated. It was found that only the annual average daily traffic, section length, access density, the rate of vertical curvature, smaller curve radius before and after the tangent were statistically significant according to related combinations. Second, three statistical models (negative binomial, zero inflated Poisson and zero inflated negative binomial) were developed using the significant variables for three alignment combinations. Third, ANN models are developed by applying the same variables for each combination. The results clearly show that the ANN models have the lowest mean square error value than those of the statistical models. Similarly, the AIC values of the ANN models are smaller to those of the regression models for all the combinations. Consequently, the ANN models have better statistical performances than statistical models for estimating collision frequency. The ANN models presented in this paper are recommended for evaluating the safety impacts 3D alignment elements on horizontal tangents.

**Keywords**—Collision frequency, horizontal tangent, 3D two-lane highway, negative binomial, zero inflated Poisson, artificial neural network.

## I. INTRODUCTION

THE policy on highway geometric design plays an important role for improving road safety and reducing collisions [1]. Collision prediction models for individual design elements of highways have been used in the design guide of the Transportation Association of Canada [2]. The two-lane rural highways have more severe safety problems than urban highways [3]. Regression models for estimating collision frequency for three-dimensional (3D) alignments on two-lane rural highways have been developed for horizontal curves and tangents [4, 5].

Straight-line segments of the roadways which are named horizontal tangents can be classified as independent and non-independent [6]. If the tangent is non-independent, the element sequence curve-to-curve, not the interim tangent,

controls the safety evaluation process. If the tangent is independent, the element sequences curve-to-tangent and tangent-to-curve would control the safety evaluation process.

To the effect of horizontal tangents on the safety at horizontal curves has been examined by a number of researchers. Fink et al. [7] found that the effects of approach tangent length and approach sight distance were not clear in the collision prediction model, but suggested that the adverse safety effects of long approach tangent and short approach sight distance become more pronounced on sharp curves. Brenac [8], based on a review research work in European, described the tangent length as an external factor to the safety on horizontal curves. The results showed that the collision rate on horizontal curves increases when the radius decreases and the length of the straight alignment (or the curve with radius greater than 1000 m) that precedes the horizontal curve increases.

A thorough literature review of the safety effect of grade by Hauer [9] showed that the results were inconsistent as to the effect of grade on collision frequency, while other studies found that grades under 6% have relatively little effect on the collision rate, and that collision rate increases sharply on grades of more than 6%. For combined horizontal and vertical alignments, some research has been conducted on the safety effect of their coordination and interaction, mainly focusing on the combination of curvature and grade [10, 11]. Easa and Mehmood [12] have developed an optimization model based on substantive safety for horizontal alignments on flat terrain. The model has used collision prediction models for horizontal curves and tangents, and the authors have stressed the need for developing collision prediction models for 3D alignments. Previous studies on tangents have focused more on their influences on horizontal curves that follow them. Little work has been done on the effects of the horizontal curves before and after the tangent [5].

The data used for developing the collision models for horizontal tangents on 3D alignments were prepared by You and Easa [4]. The data were obtained from the Highway Safety Information System (HSIS) which includes many variables related to collision frequency, roadway, traffic volume, and 3D alignment. Four years of data (2002 – 2005) on about 5,760 km (3,600 miles) of rural two-lane highways with legal speed limits in the range 88 – 104.6 km/h (55 – 65 mi/h) were analyzed in this study. The reader is referred to [4, 5] for more details on the data collection process.

O. F. Cansiz is Asst. Prof., Department of Civil Engineering, Faculty of Engineering, Mustafa Kemal University, Iskenderun, Hatay, Turkey, ( phone: 326-613-5600 - 4225 e-mail: ofcansiz@mku.edu.tr).

S. M. Easa is Professor with the Civil Engineering Department, Ryerson University, Toronto, Canada (phone: 416-979-5000-7868; fax: 416-979-5122; e-mail: seasa@ryerson.ca).

## II. PREPARATION OF COMBINATIONS AND VARIABLES

The method of data preparation is described in [5]. Each roadway route was subdivided into road sections with individual horizontal tangents, in which the vertical curves were determined and thus the combination types of horizontal tangents and vertical curves. This method allows exploring the interaction of combined horizontal tangent and vertical alignments. A roadway route was divided into sections with specific combinations of horizontal tangents and vertical alignment in that year when the data were collected. The same road section was considered in a different observation year as a separate road section. This process allows the year-to-year changes on highway geometric design and traffic conditions to be considered in the model.

Three combinations of horizontal tangents and vertical curves were selected, as shown in Table 1. The determination of these combination types, however, is not always straightforward because horizontal and vertical curves may overlap. This is described in Figure 1. Cases a and d involve situations where a horizontal tangent fully overlaps with a vertical curve or vice versa. Cases b and c involves partial overlapping, with only vertical point of curvature or vertical point of tangency lying on the horizontal tangent, respectively. For horizontal tangent on crest vertical curve, it was found that only the natural logarithm of *AADT*, section length *L*, and access density *AD* were statistically significant. All characteristics of the crest vertical curve (e.g. rate of vertical curvature) were not found to be significant. The smaller radius, ratio of larger radius to smaller radius, and the binary variable for the turning direction of the horizontal curves that precede and follow the tangent were insignificant. The cross-section variables were not detected as significant variables. This result was attributed to the small number of observations.

TABLE I COMBINATION OF HORIZONTAL TANGENTS AND VERTICAL CURVES

Combination Number	Combination Type	Variables
1	Horizontal Tangents Combined with Crest Vertical Curve	<i>AADT</i> , <i>L</i> , <i>AD</i>
2	Horizontal Tangents Combined with Sag Vertical Curve	<i>AADT</i> , <i>L</i> , <i>K</i> , <i>AD</i>
3	Horizontal Tangents Combined with Multiple Vertical Curves	<i>AADT</i> , <i>L</i> , <i>K</i> , <i>W</i> , <i>AD</i> , <i>R</i>

For horizontal tangent on sag vertical curve, the rate of vertical curvature *K* was found to be statistically significant in explaining the variability in collisions in addition to access density and the nature logarithm of *AADT* and section length. However, by examining their chi-square values in the NB model, the contribution for the rate of vertical curvature *K* was

very small ( $\chi^2 = 246, 132, 5$ , and  $13$  for  $\log(AADT)$ ,  $\log(L)$ , *K*, and access density *AD*, respectively). The smaller radius, ratio of larger radius to smaller radius, and the binary variable for the turning direction of the horizontal curves that precede and follow the tangent were not found to be significant. The cross-section variables were not found to be significant, either.

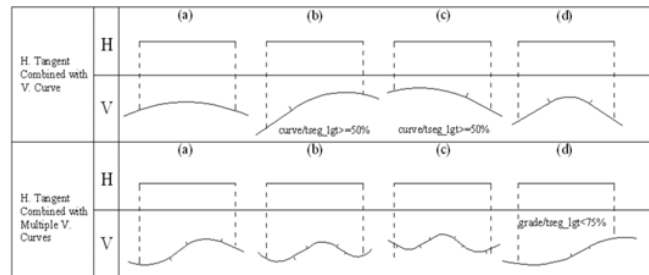


Fig. 1. Classification of horizontal tangents combined with vertical curves [5]

The results for horizontal tangent on multiple vertical curves showed that the rate of vertical curvature, total roadway width, access density, smaller curve radius before and after the tangent, and the nature log of *AADT* and section length were statistically significant variables. However, the contributions of the rate of vertical curvature *K*, total roadway width *W*, access density *AD*, and smaller curve radius *R* before and after the tangent were relatively small as determined by their chi-square values in the NB model ( $\chi^2 = 1142, 1621, 4, 6, 37$ , and  $10$  for  $\log(AADT)$ ,  $\log(L)$ , *K*, *W*, *AD*, and *R*, respectively). The negative sign of the estimated coefficient for the smaller radius *R* before and after the tangent suggests that as the smaller radius increases, collision frequency decreases.

## III. FORMATION OF STATISTICAL MODELS

The statistical models were estimated by You and Easa [5]. The variables which are related to vertical curves, grades, horizontal tangent, and cross-section, were used. The variables that contribute to collision occurrence on horizontal tangents are related to the horizontal curves before and after the tangents and the vertical alignment that overlaps with them. During the process of developing the collision prediction models, potential influencing factors that may affect collision frequency were considered in the analysis. Subsequently, a preliminary investigation to determine the best statistical technique (Poisson, NB, ZIP, or ZINB) for each alignment combination was conducted. The final models for horizontal tangents combined with vertical curves were selected.

The SAS statistical package [13] was used to estimate all models. The models were estimated using the generalized linear model (GLM). Four statistical modeling techniques were applied for each of the three combinations. The Poisson model was first estimated to identify the independent variables. The Lagrange multiplier and Wald tests were

conducted to test the overdispersion of the data. The NB, ZIP, and ZINB models were then examined. In determining whether the zero-inflated model (ZIP or ZINB) provides an improvement over the traditional Poisson or NB model, the Vuong-statistic test [14] was carried out for the ZIP versus the Poisson, the ZINB versus the NB, and the ZIP versus the NB. If this statistic does not favour any model, the model with the smallest value of the Akaike Information Criterion (AIC) was selected.

A horizontal tangent combined with vertical curves has three possible situations, which correspond to Combination 1-3 in Table 1. Collision occurrences on the highways are usually sporadic events, which are represented with no reported collisions for most of the road sections. For example, a total of 6,619 sections of horizontal tangent combined with multiple vertical curves were obtained. It is worth noting that 69.6% of the horizontal tangents experienced zero collisions. The zero-inflated regression models (ZIP and ZINB) have been widely employed to model collision counts with typical excess zeros due to their improved statistical fit in comparison with the Poisson and NB models [15]. Four statistical modeling techniques (Poisson, NB, ZIP, and ZINB) were therefore explored to fit the actual data distributions. The final three models that have the lower AIC value are then selected for combination 1-3 as follows [5]

$$Y = AADT^{0.8434} L^{0.8936} \exp(-7.1290 + 0.0569 AD) \quad (1)$$

$$Y = AADT^{0.8769} L^{0.7464} \exp(-7.1503 - 0.00032 K + 0.06446 AD) \quad (2)$$

$$Y = AADT^{0.8701} L^{0.9154} \exp(-6.7907 - 0.00008 K - 0.01218 W + 0.06782 AD - 0.00003 R) \quad (3)$$

where  $Y$  = expected collision frequency on the road section (number of collisions per year),  $AADT$  = annual average daily traffic (veh/day),  $L$  = length of horizontal tangent (mi),  $K$  = rate of vertical curvature (ft per %),  $W$  = total roadway width (ft),  $AD$  = access density (number of driveways per mile), and  $R$  = smaller curve radius before and after the horizontal tangent (ft). Table 2 shows the Vuong test and AIC values of Eqs. 1-3.

TABLE II RESULTS OF STATISTICAL MODELS FOR ALL COMBINATIONS

	Combination 1 (NB)	Combination 2 (ZIP)	Combination 3 (ZINB)
Vuong Test	-1.42	1.81	1.43
AIC	1965.2	2211.2	10470

The results of the three separate models developed for horizontal tangents combined with vertical curves show that vertical curves (sag or multiple curves) have a significant effect on collision occurrence on horizontal tangents, as

indicated by the rate of vertical curvature. The effect of crest vertical curves was found to be insignificant.

Sight distance on the independent tangent sections, which is affected by only the vertical curves, was analyzed. It was found that sight distance and  $K$  were highly correlated which is expected as sight distance equals the product of  $A$  and  $K$ , where  $A$  is algebraic difference in grade. Therefore, only the  $K$  variable was included in the models. In most tangent combinations, the results showed that collision frequency decreases as the total roadway width increases, consistent with the results of previous studies for horizontal curves. In addition, the access density was found to be a significant variable that adversely affects road safety as the access density increases.

#### IV. ARTIFICIAL NEURAL NETWORK

ANN has been successfully applied in a number of diverse fields including transportation engineering prediction models. The traffic systems' excessive variables and their complex character make it difficult to predict the results. The actual components of traffic predictive ability may be enhanced through the use of ANN analysis that is able to examine nonlinear interactions among variables. The ANN method, which enables the prediction of complex relationships and has many successes in this regard, is superior to the statistical methods. Recent research indicates that even though the statistical methods are useful in understanding the characteristic of raw data, they are not as successful as ANN method in the prediction problems. In addition, in ANN method there is no need to have any pre-assumption in their formation. This property makes the ANN method very useful in many engineering applications.

The ANN method has a few troublesome conditions in addition to its advantages. First, the ANN model works like a black box. That is, while the ANN model is trained the effects of the variables are not known explicitly. Hence even unidentified relationships exist in the model. The other disadvantage is that while the ANN model is trained there exists a state of memorizing instead of learning. Even if the performance of the training data is very well, the results of the test data may not be suitable, including adjustment of itself to the training data and decreasing the error rate to small numbers. To avoid these problems, one must arrange the training, validation, and test data appropriately and maintain the epoch number's magnitude at a reasonable level.

The ANN model consists of artificial neurons that are related functionally. Artificial neurons are like simple mathematical expressions of biological neurons. The neurons are the centers of the mathematical operations of the ANN model. A good illustration that helps understand this structure is shown in Figure 2.

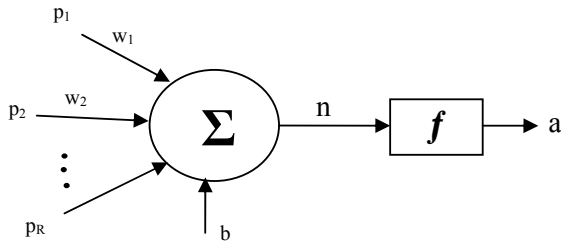


Fig. 2. An artificial neuron

The basic relationships of the ANN mode are given by

$$n = \sum_{i=1}^R w_i p_i + b \quad (4)$$

$$a = f\left(\sum_{i=1}^R w_i p_i + b\right) \quad (5)$$

where  $n$  = transfer function's input,  $R$  = number of inputs,  $w$  = weights,  $p$  = neuron's inputs,  $b$  = bias value,  $a$  = artificial neuron's output, and  $f$  = transfer function.

Neurons are related to each other by their weights. These weights are determined by training and contain the dataset's characteristic. As shown in Eq. 4,  $p$  input values are multiplied by their weights, added respectively, and the bias value is added. Thus, the first part of the artificial neuron's  $n$  value is obtained. Afterwards the  $n$  value is used in the transfer function and a value, which is artificial neuron's output, is achieved (Eq. 5).

Neurons constitute the layers that have small neuron groups before they form the ANN model. Namely, the ANN model consists of layers and the layers consist of neurons. The hidden layers exist between those layers in which there are enough neurons. There is no relationship between the neurons of the same layers. The input layer admits the new data and delivers them to all neurons of the other layers. The neurons are related with the weights between the preceding and the following layers. The data of the input layer are processed by mathematical operations in the hidden layers. Here the number of hidden layers and the number of neurons in the layers can be increased according to the complexity of the problem. The output layer takes the processed data that come from the hidden layers and produces as output the whole ANN network.

ANN method is an artificial intelligence technique because of its ability of training. The model is trained by the training set after determining the structural properties of the ANN model. The trained model is tested by the testing set. Thus, it can be understood that either the model is trained well (predicts the testing set accurately) or it memorizes the process (the performance of the model for testing set is not acceptable). During the training process the weights between the neurons are arranged for the purpose of their input and output values. The epoch number, which expresses the number of applications of the training data to the model, continuously changes its own value and each data's

application within it to decrease the errors. The weights change till reaching the determined error value, epoch number or validation fail. The training process occurs by gradually altering the weights. After the training process, the model is tested and the training and test performance is evaluated. The formula presented in Eq. 6 was used for MSE calculation.

$$MSE = \frac{1}{h} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (6)$$

where  $MSE$  = mean square error,  $y$  = real or observed value,  $\hat{y}$  = predicted value of the model, and  $h$  = number of data points.

## V. DEVELOPING OF THE ANN MODELS TO ESTIMATE COLLISION FREQUENCY

Collision frequency is estimated in this study based on supervised neural networks. It uses the multi-layer feed-forward networks which are applied frequently in ANN models. The estimated ANN models, as shown in Figure 3, have an input layer, a hidden layer, and an output layer. It is enough to have one general hidden layer to solve complex problems [16, 17]. In this study, only one hidden layer was considered enough to predict collision frequency. To determine the best ANN architecture to predict the collision frequency, different ANN models were constructed, trained, and tested. When constructing the ANN model, the number of neuron in the hidden layer was changed from 1 to 25 and the training algorithms were changed. Popular training algorithms of the Matlab program are chosen for the training of the models. They are Gradient Descent, Gradient Descent Moment, and Levenberg-Marquardth (LM) training algorithms [18]. The *Purelin*, *logsig*, and *Tansig* functions of the Matlab program is used for the transfer function of the neurons. This function has a linear character, whereas the *tansig* and *logsig* transfer functions show an S-shaped curve character. Each time, a new ANN model was obtained and its performance was compared according to the  $MSE$  value. The ANN model that has the lowest  $MSE$  was selected for the prediction of collision frequency.

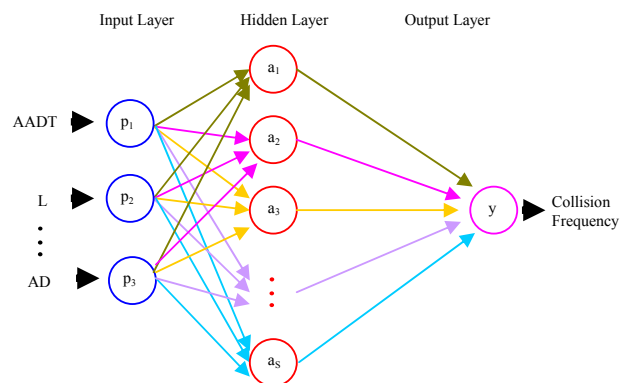


Fig. 3. A feed forward neural network

### A. Horizontal Tangent Combined with Crest Vertical Curve

To obtain the best ANN model, in which the annual average daily traffic, length of horizontal tangent, access density are used as input and the collision frequency on a horizontal tangent combined with a crest vertical curve as output, many different models were formed as shown in Figure 3. These models have different transfer functions, are trained by different training algorithms, and have different number of neurons in their hidden layers. It is observed that the best ANN model has 9 neurons in its hidden layer and is used in the LM training algorithm and the *Tansig* transfer function.

Figure 4 shows the *MSE* of the ANN model which depends on the epoch number for the training, validation, and test data sets. The best line shown in Figure 4 illustrates the start point of the overfitting. In this study, with the aid of the maximum validation failure criterion, the training process of the ANN model was terminated. In the Matlab, it was accepted that when the *MSE* values of the training and validation data sets are decreasing simultaneously, this shows that the learning process continues. However, when the *MSE* values of the training data set continues to decrease while the *MSE* values of the validation data set tend to increase, the training process may be stopped. The training also stopped when the validation error continues to increase for six iterations.

As shown in Figures 4 and 5, the training stopped at iteration 20. The best validation performance occurs by iteration 14. Figure 4 also shows that the results are reasonable because of the following considerations: the final *MSE* is small, the test and validation data set errors have similar characteristics, and no significant overfitting has occurred by iteration 14. Also changing the gradient and *mu* at the training process is seen in Figure 5. The gradient value changed between 0.0039 and 114.2843. The value of *mu* changed between 0.0001 and 1.0.

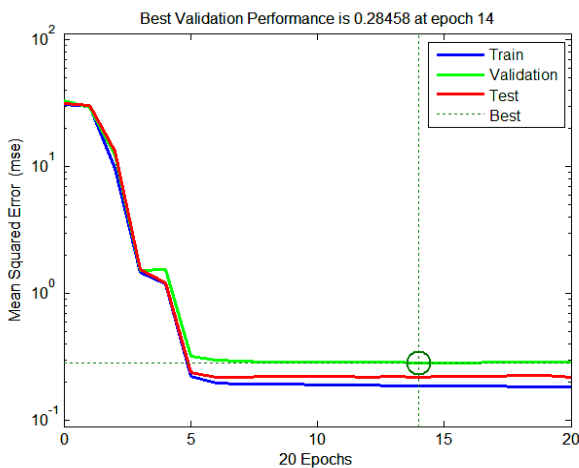


Fig. 4. Variation of MSE value of training, validation, and test datasets during the training process of ANN for Combination 1

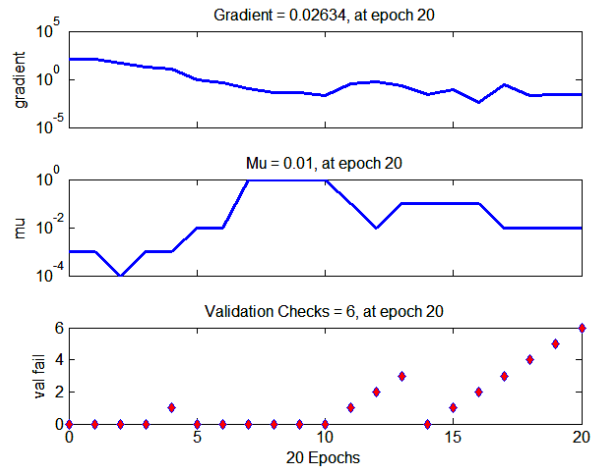


Fig. 5. Variation of parameters during the training process of ANN for Combination 1

**Building Explicit Formulation based on ANN model.** By using optimum weights and bias of ANN model explicit mathematical formulation can be obtained. The transfer functions in the hidden and output layers must also be considered in the explicit mathematical formulation [19]. Taking the annual average daily traffic, length of horizontal tangent, access density, as independent parameters into account, the collision frequency can functionally be expressed as follows:

$$CF = f(AADT, L, AD) \quad (7)$$

where *CF* = collision frequency. To derive an explicit ANN-based formulation for the collision frequency as a function of the input parameters, weights and biases of the trained ANN model were used to extract an explicit expression. This expression is given by

$$CF = \text{purelin} \left( \text{tansig} \left( \begin{bmatrix} -0.7341 & 5.4127 & -2.6342 \\ 5.0362 & 0.2602 & 2.0606 \\ -2.5357 & -15.51 & -3.6482 \\ 1.1287 & 1.3737 & -0.0292 \\ 3.6688 & -6.6603 & -3.1945 \\ 0.0276 & -4.8045 & -1.6520 \\ 3.8493 & -4.4982 & 3.6884 \\ 7.5653 & -7.9299 & -6.6505 \\ 9.9670 & -5.6009 & -7.3152 \end{bmatrix} * \begin{bmatrix} AADT \\ L \\ AD \end{bmatrix} + \begin{bmatrix} 2.9238 \\ 10.047 \\ -16.143 \\ -8.5725 \\ -7.6129 \\ -2.9647 \\ 1.6809 \\ 14.627 \\ -1.0473 \end{bmatrix} \right)^T * \begin{bmatrix} 0.2159 \\ -9.1063 \\ -0.1417 \\ 0.9354 \\ -0.1647 \\ 0.2792 \\ 0.2082 \\ -0.4598 \\ 0.0284 \end{bmatrix} + [9.2819] \right) \quad (8)$$

(Combination 1)

where *T* = transpose, *Purelin* stands for Purelin transfer function, *n* = input of the transfer function, and *Tansig* stands for tangent sigmoid transfer function which is given by

$$\text{Tansig} = \frac{2}{1 + \exp(-2n)} - 1 \quad (9)$$

It should be noted that all of the numeric variables were normalized to a range of  $[-1, 1]$  before being introduced in the ANN model using the following equation

$$y = \frac{(y_{\max} - y_{\min})(x - x_{\min})}{(x_{\max} - x_{\min})} + y_{\min} \quad (10)$$

where  $y$  = normalized value,  $y_{\max}$  = maximum value of normalized values (+1),  $y_{\min}$  = minimum value of normalized values (-1),  $x$  = value of variable,  $x_{\min}$  = minimum value of variable values, and  $x_{\max}$  = maximum value of the variable values. The minimum and maximum values of each variable are given in Table 3.

TABLE III RANGE OF THE VARIABLES FOR COMBINATION 1

Variable	Min	Max
Collision Frequency	0	7
Annual Average Daily Traffic	175	23352
Length of Horizontal Tangent	0.01	5.46
Access Density	0	25

#### B. Horizontal Tangents Combined with Sag Vertical Curve

To predicted collision frequency on horizontal tangents combined with sag vertical curve, similar models ANN models were formed. In this case, the best ANN model has 5 neurons, LM training algorithm, 12 epoch numbers, tan-sig transfer function in hidden layer and *Purelin* transfer function in output layer. Figure 6 shows the changing of gradient,  $\mu$  and validation failure of the training set according to the epoch number. The gradient value changed between 0.0126 and 16.63. The value of  $\mu$  increased from 0.001 to 1.0. Because the validation failure occurred at the iteration 18, the training process stopped and no significant overfitting has occurred by iteration 12 where the best validation performance occurs.

Figure 7 presents the *MSE* performance for the training, validation, and test datasets between the ANN output and target values according to the epoch number. As noted, the *MSE* are very close and they have the same characteristics. *MSE* consistency among the three datasets is a good evidence of how the ANN model is well trained. The explicit neural network formulations for the collision frequency derived from the proposed ANN model can be expressed as follows

$$CF = \text{purelin} \left( \text{tansig} \left( \begin{bmatrix} 0.2475 & -0.6429 & -0.6192 & -0.7695 \\ 2.0575 & -2.7034 & 0.3035 & 0.1519 \\ -0.9214 & -0.0465 & 0.6082 & 1.1251 \\ 1.6967 & 3.1825 & -1.3890 & 2.3060 \\ 0.5168 & 1.1406 & 0.4958 & 0.4363 \end{bmatrix} * \begin{bmatrix} AADT \\ L \\ K \\ AD \end{bmatrix} + \begin{bmatrix} -1.1161 \\ -2.9110 \\ 0.3732 \\ 1.4196 \\ 1.7408 \end{bmatrix} \right) + [-1.2426] \quad (11)$$

(Combination 2)

where  $K$  = rate of vertical curvature. It should be noted that all of the numeric variables were normalized to a range of  $[-1, 1]$  using Eq. 10 before being introduced the ANN model using values listed in Table 4.

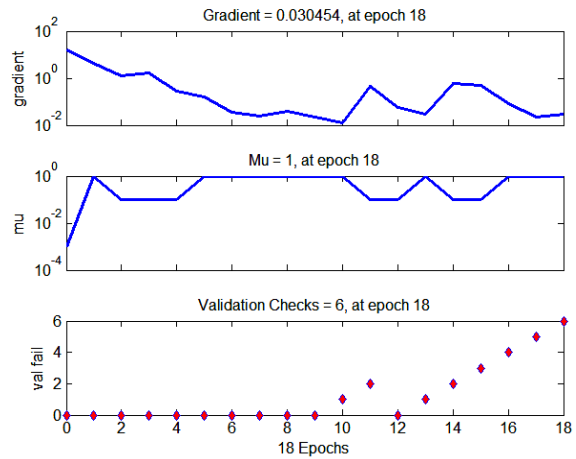


Fig. 6. Variation of parameters during the training process of ANN for Combination 2

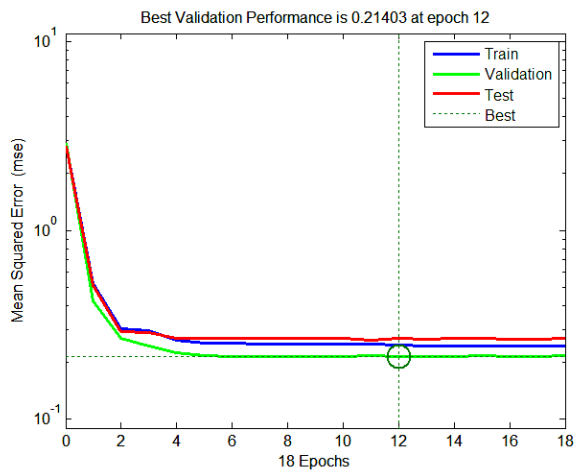
Fig. 7. Variation of *MSE* value of train, validation and test data set at training process of ANN for Combination 2

TABLE IV RANGE OF THE VARIABLES FOR COMBINATION 2

Variable	Min	Max
Collision Frequency	0	5
Annual Average Daily Traffic	175	29474
Length of Horizontal Tangent	0.01	2.13
Rate of Vertical Curvature	33.445	5000
Access Density	0	25

### C. Horizontal Tangents Combined with Multiple Vertical Curves

The best ANN model for horizontal tangents combined with multiple vertical curves was developed in a similar manner. The best ANN model has 10 neurons, LM training algorithm, 7 epoch numbers, tan-sig transfer function in hidden layer and *Purelin* transfer function in output layer, is obtained.

Figure 8 shows the gradient change,  $\mu$  and validation failure of the training set according to the epoch number. The gradient value changed between 0.0515 and 51.5598. The value of  $\mu$  increased from 0.001 to 1.0. Because the validation failure occurred at iteration 12, the training process stopped and no significant overfitting has occurred by iteration 7 where the best validation performance occurs. MSE performance of the training, validation and test sets is shown in Figure 9. The MSE values are very close. In addition, the MSE consistency among the three data sets is a good evidence of how the ANN model was well trained.

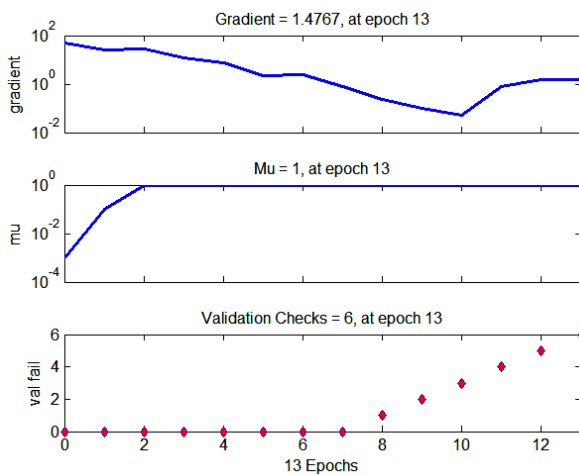


Fig. 8. Variation of parameters during the training process of ANN for Combination 3

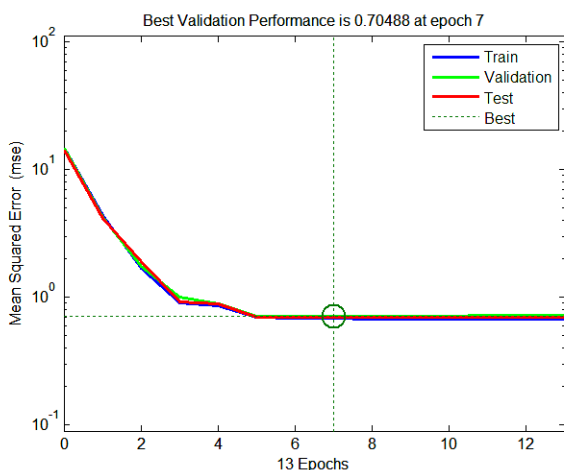


Fig. 9. Variation of MSE value of the training, validation, and test datasets during the training process of ANN for Combination 3

Similar to the previous two combinations, for Combination 3 an ANN based equation is developed by using the optimum weights, bias and transfer functions of the ANN model. The explicit neural network formulations for the collision frequency derived from the proposed ANN model is given by

$$CF = \text{purelin} \left( \text{tansig} \left( \begin{bmatrix} -0.9061 & -1.380 & 1.2545 & -0.1898 & -0.6637 & 0.6704 \\ -0.1639 & 0.8296 & -0.0960 & -0.8624 & 2.1364 & -0.3366 \\ -1.0792 & -0.2569 & 0.9582 & -0.3084 & 2.5400 & -0.4991 \\ -0.5047 & 4.4303 & 0.2413 & -0.0637 & -1.2643 & -1.2460 \\ 3.0370 & -1.5376 & 0.4458 & 1.1503 & -1.6434 & -1.9432 \\ 1.6340 & 2.1691 & 0.4702 & 0.3269 & 0.5926 & 0.0589 \\ 1.1137 & 1.2334 & -0.2179 & 0.0065 & -0.2818 & -0.3587 \\ -2.010 & 1.5363 & -1.1387 & 0.0151 & -1.6086 & -1.6605 \\ -0.1748 & 0.5936 & 0.1847 & -1.2719 & 1.7919 & 1.3399 \\ -2.7877 & 1.0427 & -0.8504 & -0.0818 & 0.5123 & -1.3159 \end{bmatrix} \begin{bmatrix} AADT \\ L \\ K \\ W \\ AD \\ R \end{bmatrix} + \begin{bmatrix} 1.4360 \\ 1.5813 \\ 0.9636 \\ 2.3873 \\ -0.3626 \\ 2.6029 \\ 2.0951 \\ -0.4127 \\ -1.6821 \\ -3.3168 \end{bmatrix} \right) + [-0.0895] \quad (12)$$

where  $W$  = total roadway width and  $R$  = smaller curve radius. Similar to the previous combinations, Eq. 10 was used to normalize the numeric variables to the range  $[-1, 1]$ . Table 5 shows the minimum and maximum values of the variables of combination 3.

TABLE V RANGE OF THE VARIABLES FOR COMBINATION 3

Variable	Min	Max
Collision Frequency	0	14
Annual Average Daily Traffic	122	26359
Length of Horizontal Tangent	0.05	16.37
Rate of Vertical Curvature	32.347	12661
Total roadway width	23	47.547
Access Density	0	17.647
Smaller Curve Radius	225	50000
Before or after horizontal tangent		

## VI. COMPARISON OF REGRESSION AND ANN MODELS

The results of the ANN models were compared with those of the regression models using MSE criteria. The comparisons for the three combinations are shown in Table 6. As noted, the statistical performances of the ANN models are better than those of the regression models for estimating collision frequency. It is clear that the ANN models have the lowest mean square error values than those of the other models. Consequently, the ANN models have better prediction capability than those of the statistical models in predicting collision frequency on horizontal tangents combined with vertical curves.



TABLE VI STATISTICAL COMPARISON OF PERFORMANCE OF THE REGRESSION AND ANN MODELS FOR ALL COMBINATIONS

	MSE	
	ANN	Regression
Combination 1	<b>0.207</b>	0.210 <sup>a</sup>
Combination 2	<b>0.245</b>	0.268 <sup>b</sup>
Combination 3	<b>0.687</b>	0.712 <sup>c</sup>

<sup>a</sup> NB model. <sup>b</sup> ZIP model. <sup>c</sup> ZINB model

## VII. APPLICATION EXAMPLES

To demonstrate how the proposed explicit ANN equations can be applied, three examples, one for each of the three alignment combinations, are presented.

### A. Example 1: Horizontal Tangent Combined with Crest Vertical Curve

The variables of the Combination 1 are  $AADT$ ,  $L$ , and  $AD$ . For the  $CF$ , the initial values of variables were used as  $AADT = 4566$ ,  $L = 0.17$ , and  $AD = 5.88$ . To input the parameters in Eq. 8, initial values of those parameters must first be normalized using Eq. 10 and the maximum and minimum values which were given in Table 3. The normalized values are as follows  $AADT = -0.621$ ,  $L = -0.941$ , and  $AD = 0.529$ . These normalized values are the values of input parameters that must be used in Eq. 8. After that, the input layer values (the preceding normalized values) are multiplied by the weights between the input and hidden layers, which are given by the first matrix in Eq. 8. This yields the first column of Table 4 which is a vector. Then the bias vector of the hidden layer (third matrix of Eq. 8) is added, giving the second column of Table 7 which is a vector and its transpose is obtained. This matrix is then input to the *Tansig* transfer function (Eq. 9) which gives a vector as shown in the third column of Table 7.

TABLE VII COMBINATION 1 IMPLEMENTATION PROCESS VALUES

First Step	Second Step	Third Step
-3.244	-0.321	-0.310
-4.965	13.082	1.000
18.106	1.963	0.961
2.257	-6.316	-1.000
5.682	-1.931	-0.959
5.380	2.415	0.984
-0.109	1.572	0.917
6.286	20.913	1.000
2.954	1.907	0.957

The resulting vector is multiplied by the weights matrix between the hidden and output layers (fourth matrix of Eq. 8). This multiplication gives a value of -10.134. After that the last bias value (9.282) is added and a value of -0.852 is obtained. Finally, the obtained value is applied to the *Pureline* transfer function and a value of -0.852 is obtained. This value is the

predicted normalized  $CF$ . The exact value of  $CF$  is obtained using the normalized equation (Eq.10). Note that -0.852 is the normalized value and the aim is to obtain the exact  $CF$  value, so an inverse process of the normalization must be applied. After this process, a value of 0.519 for  $CF$  is obtained for Combination 1.

### B. Example 2: Horizontal Tangent Combined with Sag Vertical Curve

For Combination 2,  $AADT = 6692$ ,  $L = 0.5$ ,  $K = 373.832$ , and  $AD = 4$  were selected as the initial values of the variables. These initial values were normalized using Eq. 10 and the maximum and minimum values shown in Table 4. The normalized values are calculated as  $AADT = -0.555$ ,  $L = -0.538$ ,  $K = -0.863$ , and  $AD = 0.680$ . These values are multiplied by the weights matrix which is the first matrix in the Eq. 11. This multiplication yields the first column of Table 8 which is a vector. The bias vector of the hidden layer (third matrix in Eq. 11) is then added. The obtained vector is the second column of Table 8 and its transpose is then obtained. The values of the matrix are transferred by the *Tansig* transfer function (Eq. 9) and a vector is obtained and its values are shown in third column of Table 8.

TABLE VIII COMBINATION 2 IMPLEMENTATION PROCESS VALUES

First Step	Second Step	Third Step
1.266	0.150	0.149
-0.054	-2.965	-0.995
-0.753	-0.380	-0.363
-3.023	-1.603	-0.922
-1.629	0.112	0.111

Then this vector is multiplied by the fourth matrix of Eq. 11, which represents the weights between the hidden and output layers, giving a value of 0.532. The last bias value (-1.243) is then added and a value of -0.711 is obtained. Finally, applying the *Pureline* transfer function gives a value of -0.711. The inverse normalization process results in  $CF = 0.722$  for Combination 2.

### C. Example 3: Horizontal Tangent Combined with Multiple Vertical Curve

For this example (Combination 3), the initial values of  $AADT$ ,  $L$ ,  $K$ ,  $W$ ,  $AD$ , and  $R$  were used as 23829, 0.5, 349.784, 40, 2, and 2865, respectively. Using Eq. 10 and the values in Table 5, the initial values of these variables were normalized, giving  $AADT = 0.807$ ,  $L = -0.945$ ,  $K = -0.950$ ,  $W = 0.385$ ,  $AD = -0.773$ , and  $R = 0.894$ . These normalized values were used in Eq. 12.

The input layer values are multiplied by the weights between the input and hidden layers, which is the first matrix in Eq. 12. The values demonstrating the first column of Table 5 are obtained as a vector. Then the bias vector of the hidden layer (third matrix in Eq. 12) is added and the second column of Table 9 is obtained as a vector. This vector is transposed. Then, the matrix is input to the *Tansig* transfer function (Eq.



9) to obtain a vector. The values of this vector are the third column of Table 9.

TABLE IX RESULTS OF SOME STEPS OF COMBINATION 3

First Step	Second Step	Third Step
-0.743	0.693	0.600
-2.316	-0.735	-0.626
-3.175	-2.212	-0.976
-2.755	-0.368	-0.352
6.932	6.569	1.000
-0.659	1.944	0.960
0.482	2.577	0.989
0.742	0.329	0.318
-3.951	-5.633	-1.000
-2.943	-6.259	-1.000

This vector is multiplied by the weights matrix between the hidden and output layers (fourth matrix of Eq. 12), yielding -0.638. The output bias value of -1.243 was then added giving -0.727. This value is applied to the *Pureline* transfer function and a value of -0.727 is obtained. Then using the inversion process, the predicted normalized value of *CF* is obtained as 1.912 for Combination 3.

### VIII. CONCLUSIONS

This study has explored the safety effects of horizontal tangents combined with vertical curves using neural network models. The collision prediction models were established using artificial neural network for these horizontal tangents and were compared with the existing regression models. The comparison demonstrates that, while the regression method is useful in understanding the characteristics of the raw data, the ANN method provided better results for predicting collision frequency on horizontal tangents.

The study by You and Easa [5] has identified the variables which are related to vertical curves, horizontal tangents, and cross-sections. The regression models were estimated using the significant variables for all combinations [5]. The ANN models of the present study were estimated by applying the same variables for each combination. After the ANN models were analyzed, explicit formulations for the ANN models were developed. The *MSE* values of the ANN models were 0.207, 0.245, and 0.687 for Combinations 1, 2, and 3, respectively, compared with 0.210, 0.266, and 0.712 for regression models.

These results show that the statistical performance of the ANN models in estimating collision frequency is slightly better than that of the regression models based on the *MSE* criterion. Although the ANN models are more complex, their implementation in computer applications would be as easy as regression models. Therefore, the ANN models are recommended for evaluating the safety impacts of 3D alignment elements on horizontal tangents. An attractive feature of the ANN method is that it does not require any pre-assumption in its formulation. The developed ANN models

can be used in the explicit safety evaluation of highway geometric design.

### ACKNOWLEDGEMENTS

The original data used in this study have been provided by Yusuf Mohamedshah, Manager of the Highway Safety Information System Lab of the Federal Highway Administration. The analysis to extract the relevant variables has been conducted by Qing You. Their contributions are gratefully acknowledged. This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

### REFERENCES

- [1] American Association of State Highway Transportation Officials. *A Policy on Geometric Design of Highways and Streets*. Washington, D.C., 2004.
- [2] Transportation Association of Canada. *Geometric Design Guide for Canadian Roads*. TAC, Ottawa, Ontario, 2007.
- [3] Canadian Council of Motor Transport Administrators. *2005 Annual Report - Road Safety Vision2010*. 2005. [www.tc.gc.ca/roadsafety/vision/menu.htm](http://www.tc.gc.ca/roadsafety/vision/menu.htm). Accessed July 20, 2007.
- [4] Easa, S.M. and Q.C. You. Collision Prediction Models for Three-Dimensional Two-Lane Highway Alignments: I. Horizontal Curves. *Transportation Research Record*, Journal of Transportation Research Board, 2092, 2009, pp. 48–56.
- [5] You, Q.C. and Easa, S.M. Collision Prediction Models for Three-Dimensional Two-Lane Highway Alignments: II. Horizontal Tangents. Presented at the 88th *Transportation Research Board Annual Conference*, Paper 09-3558, 2009.
- [6] Lamm, R., B. Psarianos, and T. Mailaender. *Highway Design and Traffic Safety Engineering Handbook*. McGraw-Hill, New York, 1999.
- [7] Fink, K. L. and R.A. Krammes. Tangent Length and Sight Distance Effects on Accident Rates at Horizontal Curves on Rural Two-Lane Roads. *Transportation Research Record*: Journal of the Transportation Research Board, No. 1500, 1995, pp. 162–167.
- [8] Brenac, T. Safety at Curves and Road Geometry Standards in Some European countries. *Transportation Research Record*: Journal of the Transportation Research Board, No. 1523, 1996, pp. 99–106.
- [9] Hauer, E. *Road Grade and Safety*. University of Toronto, Toronto, Ontario, 2001. [www.roadsafetyresearch.com](http://www.roadsafetyresearch.com). Accessed July 8, 2007.
- [10] Zador, P., H. Stein, J. Hall, and P. Wright. Relationship between Vertical and Horizontal Roadway Alignments and the Incidence of Fatal Rollover Crashes in New Mexico and Georgia. *Transportation Research Record*: Journal of the Transportation Research Board, No. 1111, 1987, pp. 27–41.
- [11] Harwood, D., F. Council, E. Hauer, W. Hughes, and A. Vogt. *Prediction of the Expected Safety Performance of Rural Two-Lane Highways*. Publication FHWA-RD-99-207. FHWA, U.S. Department of Transportation. [www.fhrc.gov/safety/99207.htm](http://www.fhrc.gov/safety/99207.htm). Accessed Dec. 18, 2006.
- [12] Easa, S.M. and A. Mehmood. Optimizing Design of Highway Horizontal Alignments: New Substantive Safety Approach. *International Journal of Computer-Aided Civil and Infrastructure Engineering*, 23(7), 2008, pp. 560–573.
- [13] SAS Institute Inc. *SAS/STAT® 9.1 User's Guide*. SAS Institute Inc, Cary, North Carolina, 2004.
- [14] Vuong, Q. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2), 1989, pp. 307–333.
- [15] Shankar, V., J. Milton, and F.L. Mannering. Modeling Accident Frequency as Zero-Altered Probability Processes: an Empirical Inquiry. *Accident and Analysis Prevention*, 29(6), 1997, pp. 829–837.
- [16] Cybenko, G., 1989. Approximation by Superpositions of Sigmoidal Function. *Mathematics of Control, Signals, and Systems*, 2, pp. 303–314.
- [17] Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, pp. 359–366.

- [18] MATLAB, Neural Network Toolbox. The MathWorks Inc., Natick, MA.
- [19] Cansiz, O.F. Improvements in the Forecast Model of Vehicle Crashes Deaths Formed by Artificial Neural Network. *Simulation: Transactions of the Society for Modeling and Simulation International*, doi:10.1177/0037549710370842. First published on May 26, 2010.

**Omer Faruk Cansiz** is the Asst. Prof. of the Civil Engineering at Mustafa Kemal University, Iskenderun, Hatay, Turkey. He earned his B.S. from Cukurova University in Turkey and M.S. from Mustafa Kemal University, and Ph.D. from Gazi University in 2007 (Turkey). His areas of research include accident analyses, roadside safety, finite element method, transportation energy, energy analyses, traffic safety, forecast models, and artificial intelligent methods. He has published lots of papers in these areas.

**Said M. Easa** is Professor with the Department of Civil Engineering at Ryerson University, Toronto, Ontario, Canada. He earned Ph.D. from University of California at Berkeley in 1982. His areas of research include highway geometric design, road safety, and human factors. He has published extensively in refereed journals and his work has received, from the American Society of Civil Engineers, the 2001 Frank M. Masters Transportation Engineering Award for outstanding contributions to the transportation profession and the 2005 Wellington M. Arthur Prize for best journal paper.