Multivariate Analyses for Infrastructure-Based Crash-Prediction Models for Rural Highways

Haneen Farah, Abishai Polus and Moshe A. Cohen

Abstract

Crash-prediction models can be used to assess road safety during highway planning and design. The main objective of this study is to develop an Infrastructure Coefficient that reflects the overall safety level of a highway and can be used as an independent variable in a crash-prediction model.

Infrastructure is defined as the highway and its geometric features. It includes the road alignment, road-side environment, sight-distance along the highway, presence of guardrails, number of access-points, roadway consistency alignment, lane and shoulder width and percentage of access points with a speed-change lane. These geometric features measure the overall quality of the highway.

Two different infrastructure coefficients are developed and calibrated by two different statistical methods. The infrastructure coefficient developed by using the Principal Component Analysis (PCA) method consists of 11 infrastructure characteristics, and that developed by using the Analytic Hierarchy Process (AHP) method consists of 5 infrastructure characteristics. For each highway section, a value reflecting its infrastructure quality was calculated according to each of the infrastructure coefficients developed.

The results showed a significant correlation between highway infrastructure quality and crash rates. Based on the infrastructure coefficients and crash records, two crash-prediction models are developed. It is suggested that these models can be used to evaluate the safety level of existing or planned highways.

INTRODUCTION

Crashes usually result from a combination of four contributing elements – the driver, the road, the vehicle, and the environment. Drivers are often involved in crashes because of their own errors, but also because they are affected by a combination of highway and/or vehicle elements. It is certainly not only the driver who bears responsibility for the occurrence of crashes. Henderson (1971) suggested that focusing too much on the driver as the cause of a crash often masked the ability to see other causes that could reduce crash rates and crash severity.

Crash-prediction models enable highway engineers to provide an estimate of expected crash frequency as a function of traffic volume and roadway infrastructure characteristics over a highway segment. Development of such estimates is a critical component in the consideration of safety in highway planning and design. Because of this wide variety of applications and important practical implications, crash modeling has attracted considerable research interest over the past four decades. Most of the previous work done on the development of crash prediction models concentrated on different regression methods, such as simple linear regression and generalized linear models, including Poisson and negative binomial regressions.

Joshua and Garber (1990) used simple linear regression and Poisson regression, with traffic and geometric characteristics as independent variables, to estimate truck accident rates. . Miaou et al. (1992) proposed a Poisson regression model to establish empirical relationships between truck crash rates and highway geometric and traffic data. Hadi et al. (1993), using data from the Florida Department of Transportation's Roadway Characteristics Inventory (RCI) system, developed a negative binomial (NB) regression model for accident rates on various types of rural and urban highways with different traffic levels.

Vogt and Bared (1998) employed both Poisson and negative binomial regressions to develop crash-prediction models for both two-lane road sections and three-legged and four-legged intersections. Later, Daniel et al. (2002) developed Poisson and negative binomial accident-prediction models for truck crashes on Route 1 in New Jersey. Karlaftis and Golias (2002) examined the relationship among rural (two-lane and multi-lane) road geometric characteristics, accident rates, and their prediction, using a rigorous non-parametric statistical method known as a hierarchical tree-based regression (HTBR).

Mayora and Rubio (2003) developed a negative binomial multivariate crash-prediction model for the Spanish national network's two-lane rural roads. Zhang and Ivan (2005) employed negative binomial generalized linear models (GLIM) to evaluate the effects of a roadway's geometric features on the incidence of head-on crashes on two-lane rural roads in Connecticut. Polus et al. (2005) developed a crash-prediction model that related crash rates to an Infrastructure Coefficient (IC) by using Principal Component Analysis. This infrastructure coefficient was a linear weighted combination of several infrastructure characteristics .

None of the previous models, except for Polus et al. (2005), developed an infrastructure coefficient that summarized and incorporated various infrastructure characteristics and then used it as an independent variable in the models. All previous crash-prediction models offered several infrastructure characteristics as independent variables. Therefore, the significance of the present study is its inclusion of various infrastructure characteristics in a

representative aggregated infrastructure coefficient, which is then used as an independent variable in the crash-prediction model.

Two different approaches were chosen: (1) Principal Component Analysis (PCA) and (2) Analytic Hierarchy Process (AHP). The uniqueness of both proposed approaches, and the Infrastructure Coefficients (IC) developed, is that each incorporates most features, although not all of them, of a highway's infrastructure. Developing an infrastructure coefficient has likely significant benefits in assessing whether specific highways are potentially dangerous because of their infrastructure characteristics alone. Transportation agencies could also use the proposed coefficient when evaluating several design alternatives in order to select an alternative that potentially results in lower crash rates. The advantage of creating an "Infrastructure Coefficient" is that it enables the aggregation of several physical and geometric characteristics of any highway into a single measure that can be used for comparisons and fast estimates of road quality and safety. As shown in this paper, the higher the Infrastructure Coefficient (IC) of a highway, the better is its resulting safety level.

METHOD

Data Collection

This study focused on two-lane rural highways in northern Israel. Most rural roads in Israel are undivided two-lane highways. We randomly selected twenty-five road segments. Most segments in the study were several kilometers long, the average length being 7.4 km. All segments selected connect two major intersections although typically there were several minor intersections in between.

The data base for this study consisted of two parts: crash data and infrastructure data on the road segments selected. For each highway, we measured 11 infrastructure parameters (detailed in Table 1); these included, among others, topography, lane and shoulder widths, road-side hazardousness (depending on the proximity of adjacent trees, rigid obstacles such as rocks and steep ditches), shoulder drop-off at the end of the shoulder (i.e., difference in height elevation between the paved shoulder and the unpaved road-side), number of access points per unit length, number of access points with acceleration and deceleration lanes, length of no-passing zones (considered a surrogate variable for sight-distance), length of road segment where a guardrail was required according to existing guidelines vs. length of highway where a guardrail was actually provided, and road consistency.

The consistency-parameter value was calculated using a model developed by Polus et al. (2005). Consistency was determined by the amount of variability in the operating speed of cars and trucks along a two-lane highway. The consistency model is presented in *Equation 1*:

$$IC = [2.808 \cdot \exp^{(-0.278 \cdot Ra \cdot \sigma)}] \cdot \exp^{(-0.01 \cdot ACT)} \qquad (1)$$

Where:
IC = integrated consistency of a highway segment
Ra = normalized area bounded by the speed profile of cars and the average operating speed, defined in *Equation 2* (m/sec.)
σ = standard deviation of car speeds, defined in *Equation 3* (m/sec.)
A_{CT} = normalized area bounded by the speed profiles of cars and trucks (m/sec).

The first measure was the relative normalized area (per unit length), bounded by the average speed profile and the average speed line as shown in *Equation 2*:

$$Ra = \frac{(\Sigma a_i)}{L} \qquad (2)$$

Where:
Ra = relative area (m/sec.) measure of consistency
$\sum a_i$ = sum of i areas bounded by the speed profile and the average operating speed (m^2/sec.)
L = entire segment length (m.)

The second measure of consistency was σ, the standard deviation of speeds along the highway segment; it is calculated as shown in *Equation 3*:

$$\sigma = \left\{ \frac{(V_j - V_{avg})}{n} \right\}^{0.5} \qquad (3)$$

Where:
σ = standard deviation of operating speeds (km/h)
V_j = operating speed along the j^{th} geometric element (tangent or curve) (km/h)
V_{avg} = average weighted (by length) operating speed along a highway segment (km/h)
n = number of geometric elements along a section (km/h)

Figure 1 presents the example of a road segment and its speed profile. Further discussion of these models is provided by Polus et al. (2005).

The horizontal and vertical alignment variables were collected from "as-built" plans. All other infrastructure characteristics were obtained directly from field measurements.

The consistency parameter was found to be related to crash rates (R-square=0.55) on two-lane rural highways as shown in *Figure 2*.

The crash data was obtained from files of the Israel Central Bureau of Statistics for five consecutive years, 1997 through 2001. We eliminated from the data set highway segments that had significant infrastructure changes (e.g., widening and paving of shoulders, construction of long guardrails, or intersection control and channelization changes) during the five years for which the crash data were collected.

The data included crash numbers and average daily traffic volumes, from which crash rates were determined. All crashes in the data set involved human casualties (light, serious, and fatal crashes); damage-only crashes were excluded from the dataset. It was not possible to conduct the different statistical analyses for each severity level or for each type of accident separately, since the number of crashes over the five-year period was not sufficient. The data set consisted of 1,035 crashes on the 25 highway segments. This study assumed that the relationship between crash numbers and traffic volume was linear, for two reasons: (1) after testing, no significant difference was found between the linear and parabolic relationships (see *Figure 3*); (2) the Average Daily Traffic (ADT) on most segments was below 10,000. Other studies have shown that a non-linear relationship starts at higher traffic volumes. Therefore, the volumes in this study were assumed to be on the linear portion of the

otherwise generally non-linear relationship. As a result of this linearity, further analyses were conducted with crash-rate data without the need for additional adjustments to the impact of traffic volume.

The Road-Side Score (RSS) that was developed in the present work (detailed in *Table 2*) is based on the most pertinent features of the road side, such as shoulder width, slope at the edge of the shoulders, presence of rigid obstacles, and the existence of a drop-off at the edge of the shoulder. RSS ranged from 1 for very dangerous road sides to 7 for very safe road sides.

The RSS measure was developed based on, although not identical to, the roadside-hazard ratings developed by Zegeer et al. (1988). These ratings characterize the accident potential for roadside types. The Zegeer roadside ratings were later incorporated into the crash-prediction model developed as part of the Interactive Highway Safety Design Model (Harwood et al., 2000).

Developing the Infrastructure Coefficient by Principal Component Analysis

In order to identify highways with similar crash trends and to relate this trend to infrastructure, a two-dimensional plot of the "overall infrastructure" characteristics against crash rates was required. Data on 11 infrastructure characteristics was collected for each highway. In order to meaningfully reduce this number to two principal characteristics (i.e., two dimensions), a statistical approach termed Principal Component Analysis was employed. Principal Component Analysis (PCA), developed in the 1930s (Hotelling, 1933), is discussed extensively in textbooks on multivariate statistical methods.

The 25 highway segments, characterized by the 11 infrastructure variables, can be described by 25 points in 11 dimensional spaces. Principal Component Analysis (Cooley and Lohnes, 1962) provides a method of reducing dimensionality to a visible number of dimensions, in this case from 11 to 2 dimensions. A two-dimensional plot was chosen for the following reasons: (a) Two-dimensional plot provides more perceptible display of the two clusters of poor and good roads than the three-dimensional plot; (b) the amount of variability explained by two components (a two-dimensional plot) was found to be 58% and that explained by three components (a three-dimensional plot) was 68%. This increase in the variability explained is less significant than the additional benefit resulting from better perception of the two dimensional plot.

Principal Component Analysis computes the "distance" between each pair of points in the 11 dimensions. The distance may be zero if all 11 infrastructure components have the same value; the value increases with increased variability between components.

The goal of Principal Component Analysis is to find 25 points in two dimensions (xi, yi) for highway i, such that the distances between these points are as similar as possible to the distances computed with the original 11 dimensions. Therefore, the distance between points (highways) in the two-dimensional plot represents the degree of similarity between infrastructure components: the closer the points, the more similar the highways. It is clear that axis rotation and shifting do not change the distances between the points. Since the data contains several variables with different units, we normalized all data designated for use in this analysis. The two-dimensional plot that we obtained from the PCA method after varimax

rotation is shown in *Figure 4*. The number next to each data point is the road number indicated in Israel.

It can be seen from *Figure 4* that highways formed two groups: (1) "lower crash-rate roads" (i.e., safe roads with a crash rate equal to or less than 0.25 crashes per million vehicle-km) and (2) "higher crash-rate roads" (i.e., dangerous roads with crash rates greater than 0.25 crashes per million vehicle-km). The aggregation of roads in *Figure 4* is based only on infrastructure characteristics regardless of their crash statistics, which are attached to the roads after the aggregation. This means that we differentiated higher crash-rate roads from lower crash-rate roads only by their infrastructure elements. This finding shows the significant contribution of infrastructure elements to crashes.

In order to test for the significance of the clustering of higher and lower crash-rate highways as shown in *Figure 4*, Fisher's exact test was applied (Conover, 1999) to the 2X2 contingency table (see *Table 3*). This is a statistical significance test used in the analysis of categorical data when sample sizes are small. As shown in *Figure 4*, highways formed two groups; the purpose of Fisher's exact test is to examine whether the two groups of highways are clustered by chance. The p-value of the test calculated according to the hypergeometric distribution is $4.89*10^{-6}$; thus the odds of random clustering of good highways above the M-M line (see *Figure 4*) are approximately 1 to 700,000. These odds are so small that one must reject the hypothesis that the two groups of highways are clustered by chance. In other words, there is a statistically significant association between a highway's crash rate and its belonging to the group of poor or good infrastructure highways.

Figure 4 also reveals that lower crash-rate roads are not as dispersed as are higher crash-rate roads; i.e., the dispersion of the former is much lower than that of the latter. The reason for this finding is that safe roads are typically well built; that is, all their infrastructure characteristics are well built In contrast, poor roads are often characterized by several sub-standard infrastructure features that greatly contribute to crashes.

The vertical axis, Y, in *Figure 4* actually represents road-infrastructure qualities. The "score" that a highway receives during the Principal Component Analysis on the Y-axis is, in fact, its infrastructure coefficient (IC_{PCA}), which represents the overall infrastructure characteristic of that highway.

The Infrastructure Coefficient (IC_{PCA}) is given in *Equation 1* as:

$$IC_{PCA} = -0.094 + 0.7045 \times LW - 0.6894 \times NPZ + 0.1329 \times TOP + 0.1138 \times RC + 0.1253 \times SW + 0.0108 \times SDR + 0.0365 \times RSS + 0.1370 \times G\text{-}R + 0.0421 \times AP + 0.1998 \times SCL + 0.0137 \times GRR/GRE \quad \ldots\ldots\ldots\ldots\ldots (4)$$

Where:

LW = Lane width (m.)
NPZ = Percentage of highway with a no-passing zone(%)
TOP = Topography
RC = Road consistency
SW – Shoulder width (m.)
SDR –Shoulder drop-off (cm.)

RSS – Road-Side Score (note *Table 2*)
GR – Percentage of highway with a guardrail (%)
AP – Number of access points per km. (points/km)
SCL – Percentage of access points with a speed-change Lane (%)
GRR/GRE – Percentage guardrail required vs. existing guardrail (%)

Note that one of the infrastructure elements is road consistency; this needs to be calculated separately according to the Polus et al. (2005) model as was detailed earlier.

The importance of the IC_{PCA} coefficient is that roadway engineers can rank the different roadway segments according to the resulting IC_{PCA}, which represents the overall infrastructure characteristic of a segment and its proneness to crashes.

Developing the Infrastructure Coefficient by the Analytic Hierarchy Process (AHP)

The Analytic Hierarchy Process (AHP), first developed by Thomas Saaty (1980), is a mathematical decision-making technique that incorporates both qualitative and quantitative factors.

The main purpose of using an Analytic Hierarchy Process (AHP) in this study was to rank roadway-infrastructure characteristics according to their contribution to safety. This was done by attributing a specific weight to each infrastructure characteristic. These weights are determined by the AHP method. The Infrastructure Coefficient (IC_{AHP}) for a specific road segment can be calculated by multiplying the weight of each infrastructure characteristic by its appropriate infrastructure-characteristic value for the specific segment and adding up the products. Roadway segments with high IC_{AHP} values represent a relatively good quality of roadway design (with low crash rates); segments with low IC_{AHP} values represent a relatively poor quality design (with high crash rates).

For the purpose of this analysis and the statistical method used, we preferred to convert the actual physical dimension of each element (e.g., lane width, percentage of intersections with speed-change lanes, etc.) to categorical variables. We grouped the physical dimensions into ranges, separated by thresholds, and substituted the values with a score for each range. Low scores (such as 1) represented a poorly designed, seemingly dangerous infrastructure element, and higher scores (e.g., 7 for the road-side characteristics) an apparently safe and well-designed infrastructure element. These elements received a surrogate nominal numerical score that represented the attributes of the infrastructure and its relative risk to drivers. Scores for 10 of the 11 infrastructure elements are presented in *Table 4*, while *Table 2* presents the road-side scores.

Some thresholds that we established in order to allocate the different infrastructure characteristics to representative ranges were based on engineering and common-sense judgment. Others were set by dividing the whole domain into an equal number of ranges. For example, shoulder width was categorized into four ranges. The first range-category included all highway segments with a shoulder width that was less than or equal to 0.9 m; this threshold was set, since part of the car would intrude into the through lane when a driver decides to stop on the shoulder; for example, in emergency situations. This is due to the fact that this shoulder width is less than the average width of a car. The second category contains shoulder widths of between 0.9 m and 1.8 m; in this case, most of the car's width would be within the shoulder; however, the shoulder width is still not enough to give a driver sufficient

space to remain solely on the shoulder for a repair if needed. The third category (1.8 m – 2.4 m) provides enough space for both the car and the driver's movement around the car; however, it is not enough for trucks. Lastly, category four (2.4 m – 3.0 m) provides sufficient shoulder width for trucks to safely park clear of the traffic lane. Shoulder drop-off is categorized into two levels. The first category includes highway segments with shoulder drop-offs of less than 5 cm. In this case, run-off-the-road instances generally do not cause loss of control. When the shoulder drop-off is greater than 5 cm, running off the shoulders onto a road-side area will in most cases result in a serious crash.

A similar approach was used to set the thresholds of road consistency, topography, and lane width. Thresholds of the remaining infrastructure characteristics were set by dividing the variables ranges into equal-size bars.

In order to use the Analytic Hierarchy Process, it is important to understand the relative safety importance of each infrastructure characteristic.

Several regression analyses of the correlation between crash rates and each infrastructure parameter were conducted prior to the analysis by the AHP method. For example, *Figure 2* shows the relationship between crash rate and road consistency, and *Figure 5* the relationship between crash rate and lane width. It can be seen that as road consistency improves and as lane width widens, crash rates decrease. The purposes of the preliminary analyses were (1) to investigate the relationships between individual parameters and crash rates (the trends observed agreed with engineering judgment and the results of previous studies); (2) to choose the infrastructure characteristics with the most significant relationship to road crash rates. *Table 5* shows the infrastructure characteristics chosen for the construction of the Infrastructure Coefficient (IC_{AHP}), in descending order of importance to safety. As shown, 5 of the 11 infrastructure characteristics were chosen. Most of the remaining infrastructure characteristics, such as shoulder width, percentage of highway with a guardrail, shoulder drop-off, and topography, were actually indirectly included in the road-side scale developed (see *Table 2*) and taken into account in *Table 5.* The rest of the infrastructure characteristics were not found to be significantly correlated to crash rates and, therefore, were excluded from further analysis.

It was difficult to identify a sufficiently large pool of experts in highway-safety design in order to obtain a reliable ranking of the relative importance to safety of each infrastructure characteristic. Therefore, the approach adopted was to determine the importance of each element based on the R-square and t-test results of the regression relationships found in the preliminary analysis. For example, because road consistency explains about 55% of the crash-rate variance (see *Figure 2*), which was the highest among the infrastructure characteristics selected and the most significant according to the t-test result (-5.34), it was considered in the analysis to be the most important road characteristic for safety and, therefore, given a rank of 5 (*Table 5*). In contrast, number of access points per kilometer is the least important to safety because of its lower R-square in the same analysis and lower t-test result (1.69). It, therefore, received a rank of 1 (*Table 5*).

In the AHP, it is necessary to construct a matrix of pairwise comparisons for the infrastructure characteristics. The pairwise comparisons describe the relative safety importance of each two infrastructure characteristics. To do this, Saaty's scale (1980), which helps to determine pairwise judgments, was used. Saaty's scale consists of 7 levels, in which the mid-level equals 1, indicating that the two variables compared are of the same importance

under a chosen criterion. The highest level suggests that if objective i is much more important than objective j, then the pairwise-judgment value equals 8. However, if objective i is much less important than objective j, then the pairwise-judgment value equals 1/8. Therefore, the possible values in descending order are as follows: 8, 4, 2, 1, 0.5, 0.25, and 0.125.

In this analysis, the objectives are the 5 infrastructure characteristics chosen; and the criterion according to which the objectives are compared is road safety. For example, when comparing road consistency and access points/km, there is a difference of 4 ranks between these characteristics (which is the maximum difference between any two infrastructure characteristic ranks – note *Table* 5) because road consistency (Rank=5) is more important to safety than are access points/km (Rank=1). When judged according to Saaty's scale, road consistency is much more important to safety than are access points/km; therefore, aij=8 (see *Table 6*). Based on *Table 5* and Saaty's scale, we constructed the matrix of pairwise comparisons of infrastructure characteristics – Matrix "A" – which is presented in *Table 6*.

In Matrix "A," the number in the i_{th} row and j_{th} column gives the relative importance of the infrastructure feature in the i_{th} row compared with the infrastructure feature in the j_{th} column. The problem that remains is to map a set of weights, $W_1,. . ., W_n$, from Matrix "A" for the objectives $O_1, O_2,. . , O_n$ (infrastructure characteristics) following an understanding of how the pairwise comparisons a_{ij} convert to weights W_n.

In this case, the largest eigenvalue of Matrix "A," shown in *Table 6*, is 5.096, resulting in a consistency index of 0.024, which is considered to be sufficiently close to 0. The corresponding eigenvector of the weights (normalized so that they add up to 1) is presented in *Table 7*.

Now, the Infrastructure Coefficient (IC_{AHP}) can be computed for each roadway segment, and the 25 roadway segments ranked by multiplying each of the weights from *Table 7* by the appropriate infrastructure-characteristic value for each roadway segment and then summing up the results. This is done using *Equation 5*:

$$IC_{AHP} = 0.26 \times LW + 0.09 \times NPZ + 0.45 \times RC + 0.15 \times RSS + 0.05 \times AP \quad \text{..... 5}$$

We defined all parameters as presented in *Equation 4*. The coefficients of *Equation 5* are the calculated weights, taken from *Table 7*.

It is important to remember that for the purpose of this analysis and the statistical method used, it was preferable to convert the actual physical dimension of each element (lane width, percentage of intersections with speed-change lanes, etc.) to categorical variables. Furthermore, since the coefficients of *Equation 5* are actually normalized weights, it was necessary to present the infrastructure characteristics in terms of the nominal variables of an equal number of categories (in our case, five categories) in each scale. Low scores on these scales indicate poor infrastructure quality (e.g., narrow lane width, bad consistency, etc.), and high scores a good infrastructure quality.

CRASH-PREDICTION MODELS

By relating IC to crash rates, it is possible to estimate the safety level of a new or existing roadway based on its infrastructure components. This is important when assessing various alternatives and conducting an economic evaluation, when it is necessary to allocate funds to

the most cost- and safety-efficient projects. Alignment characteristics, then, could be converted to safety levels by using this IC coefficient.

The relationship between crash rates (CR, in crashes per million vehicle kms.) and the infrastructure coefficients developed by PCA and AHP are shown in *Equation 6* and *Equation 7*:

$$CR_{(PCA)} = 0.98 \times exp^{-0.403 \times (IC_{PCA})} \ldots\ldots\ldots\ldots\ldots (6)$$
$$R^2 = 0.46$$

$$CR_{(AHP)} = 1.00 \times exp^{-0.401 \times (IC_{AHP})} \ldots\ldots\ldots\ldots\ldots (7)$$
$$R^2 = 0.56$$

The relationships between crash rate and the infrastructure coefficient according to both methods are presented in *Figure 6*. For each highway section, three values are calculated: the IC according to the PCA model (*Equation 4*); the IC according to the AHP model (*Equation 5*); and crash rates. Based on these calculated values, data points are plotted in *Figure 6*, in which each highway appears twice. The number next to each data point is the road number. The calibrated models based on both statistical methods are also presented in *Figure 6*.

The relationships in *Equation 6* and *Equation 7* are a function of the infrastructure coefficients developed and presented in *Equation 4* and *Equation 5*. Equations 6 and 7 can be used to predict crash rates on new or existing two-lane highways, based on their infrastructure elements.

The linear correlation coefficients between each two infrastructure features were examined. Some infrastructure features were strongly correlated; for example, lane width with road consistency (0.75), road-side score with lane width (0.65) and shoulder widths with lane width (0.85). Some of these correlations were expected based on engineering judgment. For example, more use would be made of guardrails in an area with mountainous terrain, which also has less design consistency and more no-passing zones. Furthermore, roads with high-standard design elements often, although not always, have good quality elements in all their geometric features, and these are correlated. These correlations, however, may preclude the use of the models presented (*Equations 6* and *7*) to identify the exact contribution of a specific individual element to expected crash rates. Nevertheless, the use of the models to estimate the crash rates of roads based on their infrastructure coefficient is still valid. Other possible effects are included in the "error term" as is often done in regression analysis. It is not claimed that they do not exist, however they are not considered in these models. The Analytic Hierarchy Process method, discussed earlier, was used to identify the parameters that contribute the most, and relative weights were given to the most important infrastructure parameters; i.e., those that most reflect a relative importance to safety. These weights are shown in *Table 7*.

An example of two hypothetical roads for comparing the two crash-rate prediction models (Equations 4, 5, 6, and 7) is presented in *Table 8*. The table shows the crash rates predicted for a typical good vs. poor infrastructure. In this example, the predicted crash rate for the poor infrastructure is higher than the predicted crash rate for the good infrastructure according to both models. As can also be seen from *Table 8*, different predicted crash rates were obtained by the two models. This difference stems from the different IC values calculated for the hypothetical roads. In the PCA model, the calculated ICs (3.86 vs. 2.30) are based on the exact value of the infrastructure characteristics. In the AHP model, the

calculated ICs (4.89 vs. 1.64) are based on the score values of the categorical infrastructure variables.

In order to compare the two methods and determine the Infrastructure Coefficient, we conducted a correlation analysis of the IC results of the two methods. The results indicate a relatively high and significant correlation (R-square of 0.85). It was found that there is a noteworthy similarity between the two models (notice *Figure 6* – the two lines are very close to each other). Part of this similarity is caused by the fact that the AHP method was based on the correlation results (R-square values presented in *Table 5*), an approach that is similar to that used by the PCA method. Nevertheless, since there is significant similarity between the two crash-prediction models, we recommend the model developed by the AHP method, which explains crashes based on 5 independent variables instead of 11 variables. This model is simpler to use, and it even explains the crash-rate variability better than does the crash-prediction model developed by the PCA (R-square 56% vs. 46%).

SUMMARY, CONCLUSIONS, AND FURTHER RESEARCH

The purpose of this research was to develop an Infrastructure Coefficient (IC) that represents the overall characteristic of a highway and to develop models by two different methods that correlate this IC with crash rates on two-lane rural highways. The two statistical methods used: Principal Component Analysis (PCA) and Analytic Hierarchy Process (AHP).

The two Infrastructure Coefficients developed succeeded in distinguishing between lower crash-rate roads and higher crash-rate roads by the difference in their overall infrastructure characteristics.

Furthermore, these Infrastructure Coefficients enable highway planners and safety auditors to predict crash rates based on the infrastructure features of the entire highway. These coefficients can be used when evaluating several alternatives for a new highway or when rehabilitating and upgrading existing highways in order to improve their overall safety features.

Further research should concentrate on the following: (1) validation of the models by increasing the database of rural two-lane highways; (2) development of crash-prediction models based on Infrastructure Coefficients for other types of highway facilities (e.g., freeways, intersections); (3) further evaluation of potential correlations between infrastructure coefficients and their impact on crash prediction; and (4) development of crash-prediction models that account for vehicle and human characteristics in addition to infrastructure features.

REFERENCES

CONOVER, W.J. (1999). *Practical Nonparametric Statistics* (3rd ed.). John Wiley, New York.

COOLEY, W.W. and LOHNES, P.R. (1962). *Multivariate Procedures for the Behavioural Sciences.* Wiley, New York.

DANIEL, J., TSAI, C., and CHIEN, S. (2002). Factors Influencing Truck Crashes on Roadways with Intersections. *Transportation Research Board 81st Annual Meeting*. Washington, DC.

HADI, M.A., ARULDHAS, J., CHOW, L.F., and WATTLEWORTH, J.A. (1993). Estimating Safety Effects of Cross-Segment Design for Various Highway Types Using Negative Binomial Regression. *Transportation Research Record 1500*, p. 169.

HARWOOD, D.W., COUNCIL, F.M., HAUER, E., HUGHES, W.E., and VOGT, A. (2000). Prediction of the Expected Safety Performance of Rural Two-Lane Highways. *Publication No. FHWA-RD-99-207*, Federal Highway Administration, Washington, DC.

HENDERSON, M. (1971). *Human Factors in Traffic Safety: A Reappraisal*. Traffic Accident Research Unit, Department of Motor Transport, New South Wales, Australia.

HOTELLING, H. (1933). Analysis of a Complex Series of Statistical Variables into Principal Components. *Journal of Educational Psychology* 24, p. 417, p. 498.

JOSHUA, S.C and GARBER, N.J. (1990). Estimating Truck Accident Rate and Involvement Using Linear and Poisson Regression Models. *Transportation Planning and Technology* 15, p. 41.

KARLAFTIS, M.G. and GOLIAS, I. (2002). Effects of Road Geometry and Traffic Volume on Rural Roadway Accident Rates. *Accident Analysis and Prevention* 34(3), p. 357.

MAYORA, J.M.P. and RUBIO, R.L. (2003). Relevant Variables for Crash-Rate Prediction on Spain's Two-Lane Rural Roads. *Transportation Research Board 82nd Annual Meeting*. Washington, DC.

MIAOU, S.P., HU, P.S., WRIGHT, T., RATHI, A.K., and DAVIS, S.C. (1992). Relationship between Truck Accidents and Highway Geometric Design: A Poisson Regression Approach. *Transportation Research Record 1376*, p. 10.

POLUS, A., POLLATSCHEK, M. A., MATTAR-HABIB, C., and JARROUSH, J. (2005). An Enhanced Integrated Design-Consistency Model for Both Level and Mountainous Highways and Its Relationship to Safety. *Road and Transport Research* 14(4), p. 13.

POLUS, A., POLLATSCHEK, M. A., and FARAH, H. (2005). Impact of Infrastructure Characteristics on Road Crashes. *Traffic Injury Prevention* 6 (3), p. 240.

SAATY, T.L. (1980). *The Analytic Hierarchy Process*, McGraw-Hill, New York.

VOGT, A. and BARED, J.G. (1998). Accident Models for Two-Lane Rural Road: Segments and Intersections, *Report Number FHWA-RD-98-133.* U.S. Department of Transportation, Federal Highway Administration, Washington, DC.

ZEGEER, C. V., REINFURT, D. W., HUMMER, J., HERF, L., and HUNTER, W. (1988). Safety Effects of Cross-Section Design for Two-Lane Roads. *Transportation Research Record 1195*, p. 20.

ZHANG, C. and IVAN, J.N. (2005). Effects of Geometric Characteristics on Head-on Crash Incidence on Two-lane Roads in Connecticut. *Transportation Research Board 84th Annual Meeting*, Washington, DC.

Haneen Farah

Haneen Farah is a full-time Ph.D. candidate in the Department of Civil and Environmental Engineering at the Technion - Israel Institute of Technology. She holds B.Sc. and M.Sc. degrees from the Faculty of Civil and Environmental Engineering, Technion. Her major academic interests are road geometric design, crash-prediction models, transportation safety, and driving simulators.

Prof. Abishai Polus

Professor Polus received his B.Sc. in Civil and Environmental Engineering at the Technion - Israel Institute of Technology and the M.Sc. and Ph.D. in Transportation Engineering at Northwestern University . Currently he is Associate Dean for Graduate Studies and Research of the Civil and Environmental Engineering Faculty at the Technion. His research interests focus on traffic flow characteristics and its relationship to the infrastructure and safety. He has authored more than 70 journal publications and numerous research reports in these areas.

Prof. Moshe Cohen

Professor Cohen received his B.Sc. in Chemical Engineering and M.Sc. and D.Sc. in Operations Research at the Technion - Israel Institute of Technology. Currently he is a Professor of Industrial Engineering at the Jerusalem College of Engineering. His research interest focuses on simulation, in which area he holds a U.S. patent (pending); his SimWiz package can be downloaded from http://my.jce.ac.il/bani He also does research in operations research and safety aspects of transportation systems and publishes extensively in these areas.

Contact

Haneen Farah

Faculty of Civil and Environmental Engineering

Technion–Israel Institute of Technology

Haifa 32000, Israel.

TEL: +972-4-829-3163

FAX: +972-4-829-5708

E-mail: fhaneen@technion.ac.il

LIST OF TABLES AND FIGURES

Tables

Figures

Table 1

Main Infrastructure Characteristics of 25 Highway Segments

Road No.	Length(Km.)	Lane Width (m.)	Shoulder Width (m.)	Mean Access Points per Km. (Points/Km.)	Per cent Intersection with Speed Change Lane (%)	Per cent of Hwy with No-Passing Zone (%)	Per cent of Hwy with G-R (%) *	Per cent G-R Required vs. Existing G-R (%)	Shoulder Drop off (cm.)	Topography **	Road Side Score ***	Consistency ****	Average Daily Traffic Volume (Veh/day)
1	13.76	3.75	2.65	2.00	18%	9%	38%	55%	17.50	3	4	2.48	10380
2	9.89	3.85	2.60	2.12	7%	27%	24%	0%	7.50	3	2	2.13	21540
3	5.92	3.80	2.75	0.51	33%	15%	100%	0%	0.00	1.5	5	2.38	9700
4	7.39	3.80	2.90	0.68	7%	23%	74%	0%	5.00	1	5	2.66	9860
5	7.07	3.75	2.65	1.98	11%	28%	57%	M.D.	22.50	3	4	1.17	16240
6	6.61	3.75	2.65	1.13	19%	20%	88%	M.D.	22.50	3	5	2.41	16240
7	6.32	3.75	2.40	2.53	6%	8%	38%	0%	0.00	3	6	2.10	9700
8	5.12	3.60	2.50	1.07	47%	23%	33%	0%	0.00	3	7	0.99	23340
9	10.72	3.70	2.60	0.42	10%	7%	44%	0%	2.00	3	6	2.57	15160
10	7.43	3.65	2.50	0.87	23%	35%	63%	0%	2.00	2.5	6	2.60	16940
11	10.20	3.70	2.75	0.88	17%	33%	58%	0%	4.00	2.5	6	0.98	17260
12	7.30	3.85	2.75	0.55	17%	16%	81%	4%	0.00	1	5	2.68	15760
13	8.85	3.35	1.20	3.62	3%	15%	27%	9%	17.50	1+3	2	0.01	5320
14	12.77	3.45	1.00	1.10	25%	53%	40%	0%	0.00	1	3	0.00	9613
15	6.96	3.80	2.35	1.51	6%	40%	34%	44%	11.00	1	4	0.57	10260
16	6.72	3.40	0.70	1.93	7%	30%	17%	237%	2.50	3	1	0.75	5620
17	5.27	3.20	1.20	0.76	0%	25%	29%	M.D.	6.00	1	2	0.00	2680
18	5.17	3.45	2.00	1.26	14%	15%	27%	94%	0.00	M.D.	2	1.50	5640
19	6.00	3.65	1.20	1.58	0%	33%	25%	17%	17.50	1	2	0.44	3900
20	3.93	3.45	2.20	2.16	6%	42%	26%	46%	0.00	1	2	0.47	9160
21	9.35	3.20	0.60	1.07	0%	53%	21%	12%	0.00	1	1	0.00	2700
22	10.00	3.25	1.20	1.85	11%	9%	27%	M.D.	0.00	1+3	1	0.02	1900
23	9.00	3.65	2.65	1.94	15%	53%	77%	0%	0.00	1	5	0.19	8800
24	4.19	3.10	1.10	1.43	8%	73%	35%	127%	2.50	1	3	0.04	6720
25	6.70	3.40	1.70	1.94	19%	19%	31%	66%	5.50	1	2	0.81	2100

M.D. – Missing Data; (b) G-R – Guardrail; (c) Topography -- Mountainous (1), Hilly (2), Level (3); (d) Road-Side Score -- Note Table 2; (e) Consistency -- Poor (RC≤1.0), Moderate (1<RC≤2), Good (RC>2.0). Note: Scores of Infrastructure features are presented in Tables 2 and 4.

Table 2

Road-Side Criteria Score

Score	Road-Side Features
1	• No guardrail along most of the segment length (LGR<30%) • Shoulder width less than 0.9 m. • Shoulder drop-off greater than 0.05 m. • Rigid obstacles within 9.0 m. or less from the pavement edge • Slope of ditch steeper than 4:1; ditch more than 0.40 m. deep, no guardrail • No recovery area beyond shoulder
2	• Features as for Score 1, except that the shoulder width is more than 0.9 meter
3	• Guardrail length between 30% and 70% of the segment length • Shoulder width from 0.9-1.8 m. • Dangerous roadside features, such as rocks or cuts, cliffs, but with guardrail • Portion of road without guardrail has rigid obstacles within 9.0 m. of pavement edge or a shoulder drop-off of 0.05 m. or more or no recovery area beyond shoulder
4	• Features as for Score 3 except that the shoulder width is more than 2.4 meters
5	• Guardrail length greater than 70% of the segment length • Shoulders wider than 2.4 m. • Dangerous roadside features, such as rocks or cuts, cliffs, but with guardrail • No shoulder drop-off and recoverable road side
6	• Guardrail length is between 30% and 70% of the segment length • Shoulders wider than 2.4 m. • Moderate roadside compared to Score 5 • No shoulder drop-off and no rigid obstacles closer than 9.0 m. from pavement edge
7	• Shoulder wider than 2.4 m. • No shoulders drop-off • Rigid obstacles at a distance greater than 9.0 m. from pavement edge • Wide recovery area beyond shoulders • Side slope flatter than 4:1 • Length of guardrail 30% or less of segment length

Table 3
Fisher Exact Test for Statistical Significance of Clustering of Points in *Figure 4*

		Number of Highways that are		
		Above M-M Line	**Below M-M Line**	**Total**
Number of Highways with Crashes per Million Vehicle-Km.	Less than 0.25	9	1	**10**
	More than 0.25	0	15	**15**
	Total	**9**	**16**	**25**

Table 4

Scores for Infrastructure, Topography, and Consistency Features

Score	Shoulder Width (m.)	Percent of Highway with G-R* (%)	Number of Access Points/Km.	Percentage of G-R* Required vs. Existing G-R (a) (%)	Shoulder Drop-off (cm.)
1	≤ 0.9	0% - 20%	3.00 - 3.65	≥100%	> 0.05
2	0.9 -1.8	20% - 40%	2.35 - 3.00	50% - 100%	≤ 0.05
3	1.8 - 2.4	40% - 60%	1.70 - 2.35	0% - 50%	
4	2.4 - 3.0	60% - 80%	1.05 - 1.70		
5		80% - 100%	0.40 - 1.05		

Score	Lane Width (m.)	Top.	Percent of Highway with No-Passing Zone	Percent of Access Points with Acceleration / Deceleration Lanes	Consistency
1	3.00 - 3.30	M	≥ 60%	0% - 9%	RC≤1 (Poor)
2	3.30 - 3.60	H	45% - 60%	9% - 18%	1<RC≤2 (Moderate)
3	3.60 - 3.90	L	30% - 45%	18% - 27%	RC>2 (Good)
4			15% - 30%	27% - 36%	
5			0% - 15%	36% - 45%	

G-R- Guardrail
Top. - Topography
M - Mountainous
H- Hilly
L- Level

Table 5

Ranking, According to R-Square, of the Infrastructure Characteristics Chosen

Infrastructure Characteristic	R- Square	t-test	Ranking
Road Consistency	0.55	-5.34	5
Mean Lane Width	0.41	-3.98	4
Road-Side Score	0.25	-2.75	3
Per cent of Highway With No-Passing Zone	0.14	1.94	2
Access Points/Km.	0.04	1.69	1

Table 6

Judgment Values According to Relative Importance to Safety

	RC	LW	RSS	%NPZ	AP
RC	1	2	4	4	8
LW	0.5	1	2	4	4
RSS	0.25	0.5	1	2	4
%NPZ	0.25	0.25	0.5	1	2
AP	0.125	0.25	0.25	0.5	1

RC = Road consistency
LW = Lane width (m.)
RSS = Road-Side Score (note Table 2)
%NPZ = Per centage of highway with a no-passing zone (%)
AP = Number of access points/km. (points/km.)

Table 7

Weighting Infrastructure Characteristics According to the AHP Method

Infrastructure Characteristics	Weight
RC	0.45
LW	0.26
RSS	0.15
%NPZ	0.09
AP	0.05
Σ	1.0

Table 8
Example of Using Crash-Rate Prediction Models for Predicting Crash Rates for Typical Good and Poor Roadways

	According to the PCA Method		According to the AHP Method	
	Good Infrastructur	**Poor Infrastructure**	**Good Infrastructure**	**Poor Infrastructure**
LW	3.65 m.	3.2 m.	Score = 3	Score = 1
NPZ	10%	60%	Score = 5	Score = 1
TOP	level (3)	mountainous (1)	-	-
RC	2.8	0.45	Score = 3	Score = 1
SW	2.7 m.	1.2 m.	-	-
SDR	0 cm.	10 cm.	-	-
RSS	6	2	Score = 6	Score = 2
GR	40%	20%	-	-
AP	0.6 pts./km.	1.8 pts./km.	Score = 5	Score = 3
SCL	50%	15%	-	-
GRR/ GRE	10%	50%	-	-
	Calculated IC (Equation 4)	**Calculated IC (Equation 4)**	**Calculated IC (Equation 5)**	**Calculated IC (Equation 5)**
	3.86	2.30	4.89	1.64
	Calculated CR (Equation 6)	**Calculated CR (Equation 6)**	**Calculated C (Equation 7)**	**Calculated CR (Equation) 7**
	0.21	0.39	0.14	0.52

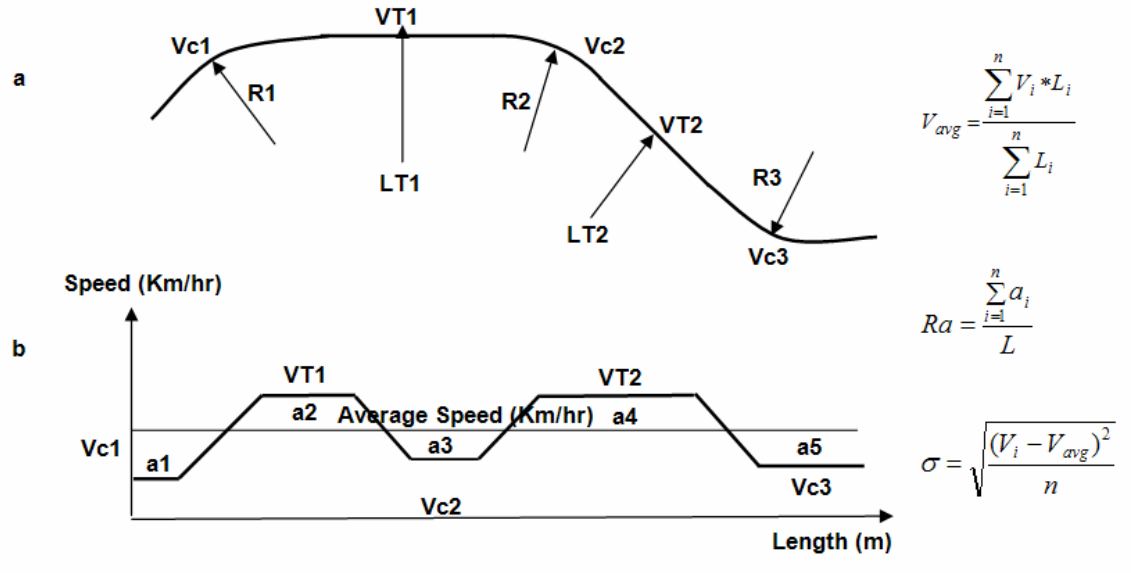


Figure 1

(a) Example of Road Segment; (b) Example of Speed Profile

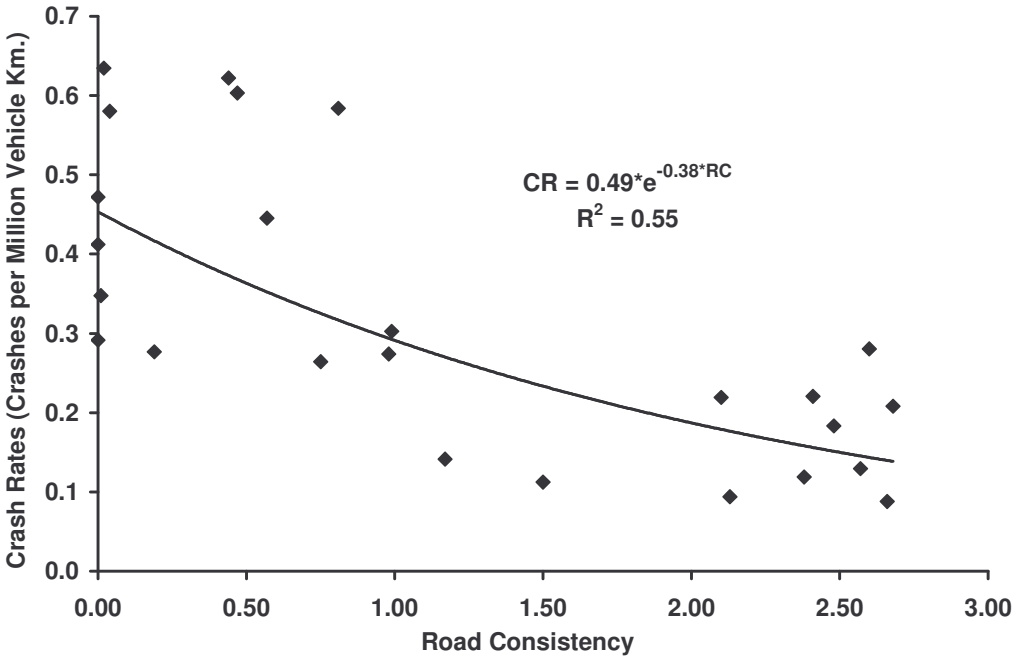


Figure 2

Crash Rates vs. Road Consistency

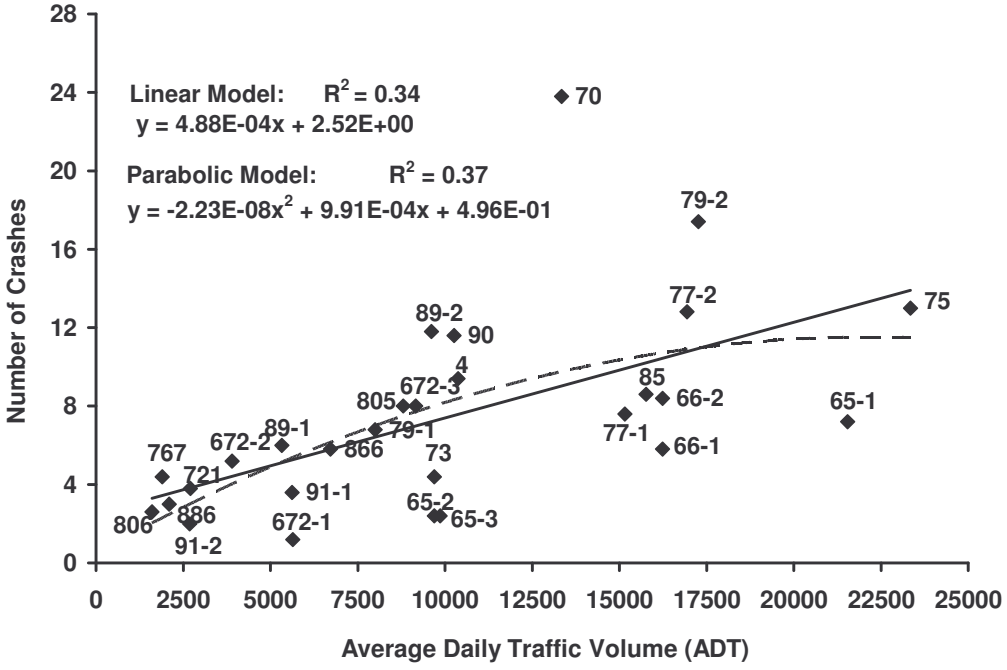


Figure 3

Relationship between Number of Crashes and ADT Volumes on all 25 Highway Segments (the number next to each data point is the road number)

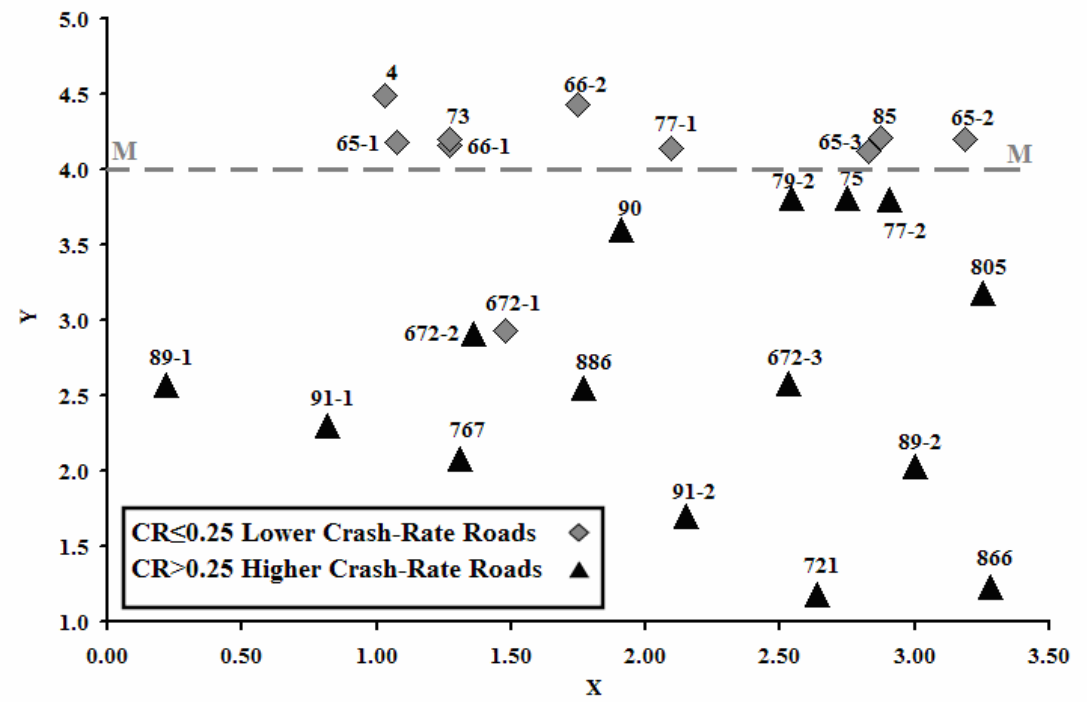


Figure 4

Road Scatter According to Two Dimensions (xi and yi represent the position of highway i based on its infrastructure similarity to other highways)

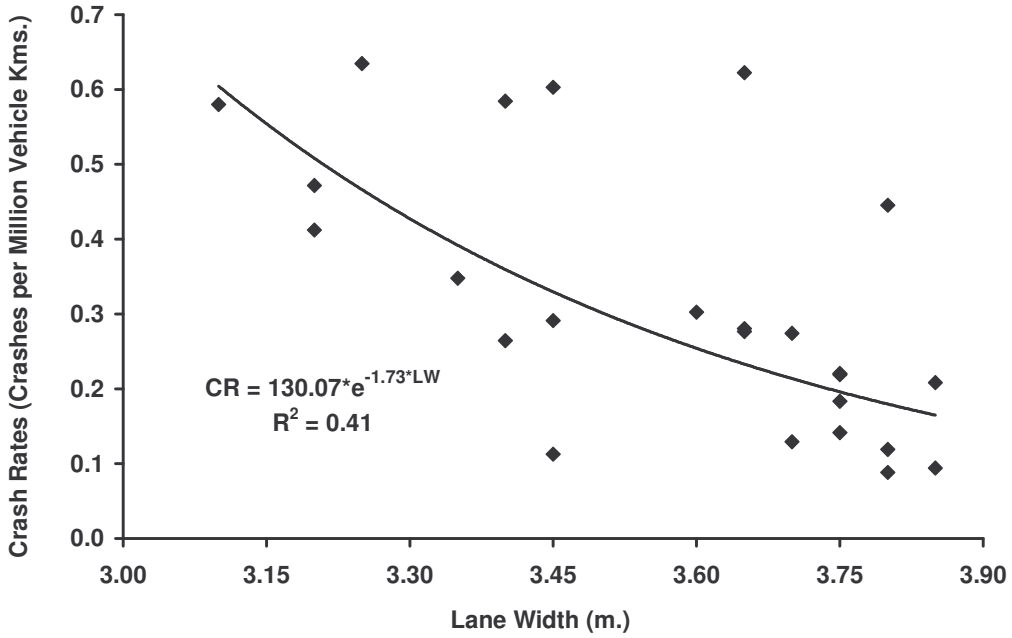
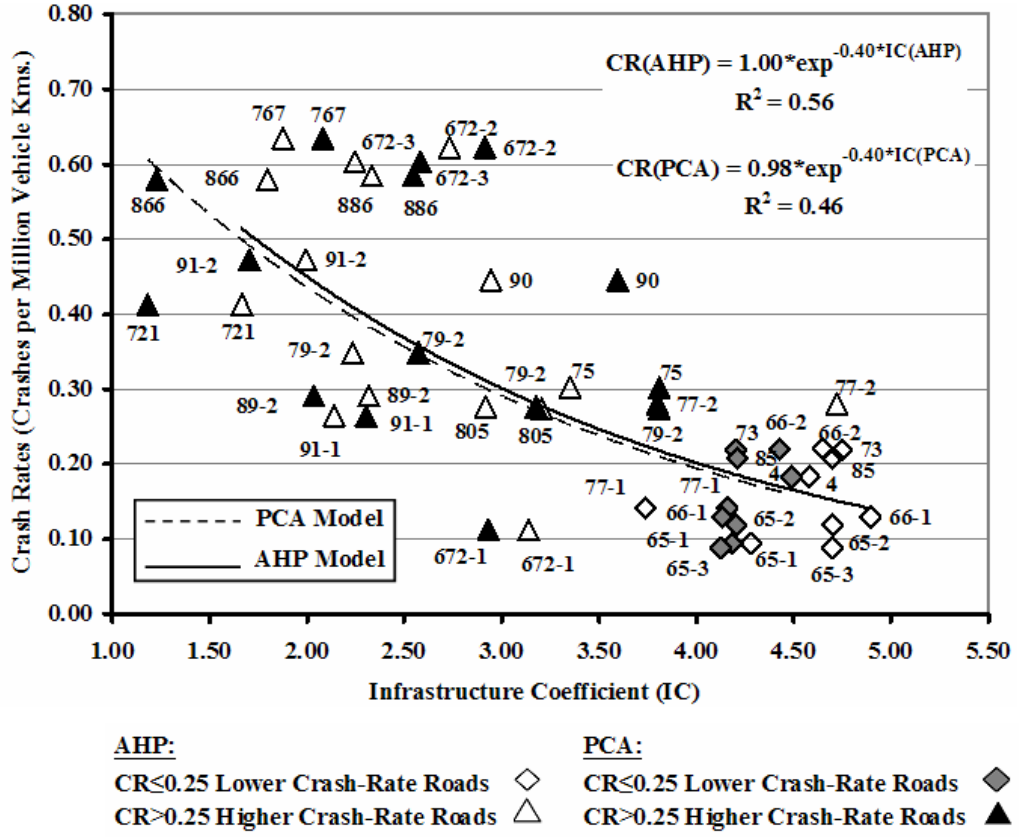


Figure 5

Crash Rates vs. Lane Width



AHP:
CR≤0.25 Lower Crash-Rate Roads ◇
CR>0.25 Higher Crash-Rate Roads △

PCA:
CR≤0.25 Lower Crash-Rate Roads ◆
CR>0.25 Higher Crash-Rate Roads ▲

Figure 6

Relationship between Crash-Rates and Infrastructure Coefficients