Procedure for Developing Accident Modification Factors from Cross-Sectional Data

James A. Bonneson and Michael P. Pratt

This paper describes a procedure for developing accident modification factors (AMFs) by using a cross-sectional study. It is recognized that AMFs are most accurately derived from controlled experiments and observational before–after studies. However, the execution of experiments and before–after studies is not always practical or feasible. The procedure described in this paper is intended to be used in this situation. The procedure is demonstrated through the development of a curve radius AMF for rural two-lane highways.

This paper describes a procedure for developing accident modification factors (AMFs) by using cross-sectional data. It is recognized that AMFs are most accurately derived from controlled experiments and observational before–after studies (*1*). However, the execution of experiments and before–after studies is not always practical or feasible. For example, for an engineer who desires an AMF that describes the effect of a change in curve radius on safety, the use of before–after studies to develop a curve radius AMF that is sensitive to a wide range of radii (and possibly speeds) is probably infeasible given the cost of reconstructing horizontal curves.

The procedure described in this paper is intended to be used when controlled experiments or observational before–after studies are not practical or feasible for use in the development of an AMF. The procedure is amenable to the development of AMFs represented as constants or as functions of other variables. It is described in more detail in a report by Bonneson et al. (*2*).

This paper is divided into three parts. The first part provides a brief review of the literature on the topic of AMFs. The second part describes the proposed procedure for developing AMFs. The last part describes an application of the procedure for the development of an AMF for curve radius.

LITERATURE REVIEW

Safety Prediction Model

The expected crash frequency for a highway segment with specified attributes is computed by using a safety prediction model. The most widely used form of this model was developed by Harwood et al. (*3*). It includes a "base" model and one or more AMFs. The base model is used to estimate the expected crash frequency for a typical segment. The AMFs are used to adjust the base estimate when the attributes of the specific segment are not considered typical. The basic form of the safety prediction model is shown in Equation 1.

$$E[N] = E[N]_b \times \text{AMF}_1 \times \text{AMF}_2 \times \cdots \times \text{AMF}_n \quad (1)$$

where

$E[N]$ = expected crash frequency (crashes per year),
$E[N]_b$ = expected base crash frequency (crashes per year), and
AMF_i = AMF for geometry or traffic control variable i (i = 1, 2, . . . , n).

An AMF can be a constant or a function that represents the change in safety following a specific change in the design or operation of a segment. A figure or table is sometimes used to portray the functional relationship of the AMF (instead of an equation). An AMF represents the ratio $N_{\text{with}}/N_{\text{without}}$ where, N_{with} represents the expected number of crashes experienced by a segment with one or more specified geometric design elements or traffic control devices, and N_{without} represents the expected number of crashes that would be experienced without the specified elements or devices.

Development of AMFs

AMFs have been developed through several methods. The more widely recognized methods include the before–after study and the cross-section study (*1*). Other methods also exist. For example, an expert panel method was used by Harwood et al. (*3*) to estimate AMFs based on experience and the subjective assessment of research results. A more formalized method of combining expert judgment and previous research findings to obtain more accurate AMFs has been developed by Washington and Oh (*4*). More recently, crash-based case-control and cohort methods have been shown to be viable for estimating AMFs by using cross-section data (*5*).

The remainder of this section focuses on the cross-section study because of its direct relationship to the procedure proposed in a subsequent section. This study is a form of case-control study, but the assignment to groups is based on whether an attribute of interest exists (or does not exist). In this version, the expected crash frequency of a group of locations having a specific component of interest is compared to the expected crash frequency of a group of locations with similar characteristics, but which do not have the component. The expected crash frequency of the former group is represented as N_{with} and that of the latter group as N_{without}. Any difference in crash frequency

Texas Transportation Institute, Texas A&M University System, 3135 TAMU, College Station, TX 77843-3135. Corresponding author: J. A. Bonneson, j-bonneson@tamu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2083, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 40–48.
DOI: 10.3141/2083-05

between the two groups is attributed to the component. The ratio of these two estimates is used to compute the AMF (i.e., AMF = $N_{with}/N_{without}$). AMFs developed by using this type of cross-section study are typically represented as constants.

Statistical techniques can also be used to calibrate a regression model by using the combined database (i.e., data from locations with and without the component). With this technique, a variable in the model is used to represent the effect of differences in condition (e.g., grade, speed, lane width, etc.) on crash frequency among the locations represented in the database. The calibrated model can then be used to estimate both N_{with} and $N_{without}$. As before, the ratio of these two values is used to compute the AMF. However, these AMFs may not be precise indicators of cause and effect due to many factors, such as correlations in the model variables (*6*). AMFs developed by using this type of cross-sectional study are typically represented as functions.

PROCEDURE FOR DEVELOPMENT OF AMFS

This part of the paper describes the proposed procedure for developing AMFs. It consists of two component procedures: a segmentation procedure and a calibration procedure. Each component procedure is described separately.

Segmentation Procedure

The AMF development procedure is based on the use of matched pairs of road segments. The segment pairs (SPs) are selected such that their attributes are identical, except for differences in the attributes that are associated with the AMF input variables. By selecting pairs of matched segments, the effect of the selected attributes on safety is isolated and all other factors are controlled. To ensure that the pairs are matched as closely as possible, the segments are selected to be relatively near each other on the same roadway. Hauer (*6*) discusses the issues and challenges of using matched pairs in cross-section studies.

Figure 1 illustrates how the matched pairs are physically related for the examination of curve radius, shoulder width, and lane width. A similar approach would be used to develop AMFs for changes in other elements or devices.

As Figure 1 shows, the SPs are located on segments of roadway that are near to one another to ensure similarity in environment as well as geometry and traffic stream. For each pair, the length of one segment is adjusted such that both segments have the same length. Thus, for the curve–tangent pairs shown in Figure 1*a,* this adjustment requires using a portion of the adjacent tangent segment that is equal to the curve length.

There should be a buffer zone separating the two segments. This zone is used to ensure that crashes associated with one segment are not inadvertently placed on the other segment. Furthermore, this zone minimizes the spillover effect by which attributes of one segment have some influence on driver behavior that carries over to the other segment. As a minimum, the length of this buffer should equal the precision with which crash location is defined in the database to ensure that crashes are not mistakenly associated with the wrong segment.

Calibration Procedure

State departments of transportation (DOTs) typically maintain a database that includes geometry and traffic data that describe the state highway system. This database can contain thousands of segments from which to select the necessary SPs. However, there are relatively few pairs that have identical geometric and traffic attributes (with the exception that the one attribute of interest can change). Moreover, the length of some segments can be relatively small such that they have no recorded crashes over a period of

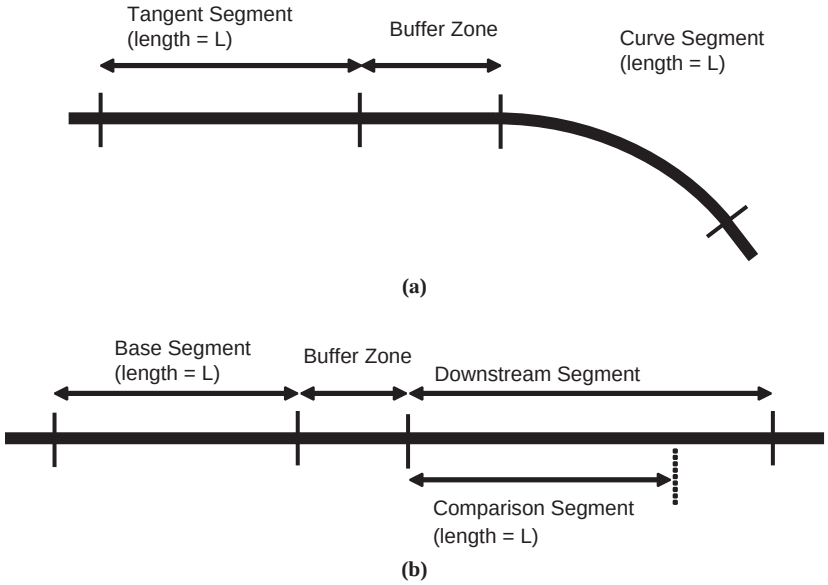

FIGURE 1 Segment pairs for AMF calibration: (*a*) for calibrating horizontal curve radius AMF and (*b*) for calibrating shoulder width and lane width AMFs (tangent segment length defined as equal to curve length; comparison segment length defined as equal to base segment length).

several years. As a result, the direct comparison of crash frequency on individual matched-SPs tends to yield little useful information.

A procedure is described in this section that overcomes the aforementioned limitation of low crash count for matched SPs yet still allows the calibration of an AMF. The procedure uses a multivariate regression model to estimate the expected crash frequency for one segment of each pair, as may be influenced by its geometric and traffic attributes. This estimate is then refined with the empirical Bayes (EB) technique developed by Hauer (*1*) to include information about the reported crash frequency for the segment. The segment for which the expected crash frequency is estimated is referred to as the "before" segment.

The second segment of each pair is considered to be the "after" segment. Its reported crash frequency is compared with the EB estimate for the before segment during AMF calibration. To facilitate this calibration, the pairs are aggregated into groups in which all members of a group have similar before values and similar after values for the geometric element of interest. For example, one group could consist of all SPs in which the before lane width is 10 ft and the after lane width is 12 ft; a second group could consist of all pairs in which the before lane width is 11 ft and the after lane width is 13 ft; and so on. The number of groups for each possible lane width combination is factorial; however, many of these combinations are irrelevant as they are rarely found on highways.

The computational steps that compose the calibration procedure are described in the following subsections.

Step 1. Assemble Segment-Pair Databases

The objective of this step is to identify SPs in the agency's roadway database. The pairs should be defined by using the methods described in the previous section and should be selected to facilitate the calibration of one specific AMF. The SPs should be identical, except that they should exhibit some variation in the input variables associated with the subject AMF. For example, if the AMF is for curve radius, then one segment should have a curve with a specified radius and the other segment should be on the adjacent tangent. If the AMF to be calibrated is for shoulder width, then the SPs should be identical except that they would not be required to have the same shoulder width.

The SP database is restructured to form two additional databases. The first database is used for regression model calibration. It is referred to as the SP regression database. In this database, each individual segment represents one observation. Thus, if there are n segment pairs, there would be $2n$ observations in the SP regression database. The second database is used for AMF calibration. It is referred to as the SP group database. In this database, the number of observations (or groups) is equal to the number of unique combinations of the before and the after values for the geometric element of interest. Three databases are produced in this step: SP, SP regression, and SP group; however, only the latter two are used in subsequent steps.

Step 2. Calibrate the Multivariate Model

The objective of this step is to calibrate a multivariate regression model by using the SP regression database. One form of this model is shown in Equation 2 (other forms are possible).

$$E[N] = \mathrm{ADT}^{b_1} L e^{\left(b_0 + \sum b_i x_i\right)} \quad (2)$$

where

$E[N]$ = expected crash frequency (crashes per year),
L = highway segment length (mi),
ADT = average daily traffic volume (vehicles per day),
b_i = calibration coefficients ($i = 0, 1, 2, \ldots, n$), and
x_i = variables describing various geometric or traffic-control-device attributes ($i = 0, 1, 2, \ldots, n$).

Two types of variables are included in the model: key variables and supplemental variables. The key variables include traffic volume, segment length, geographic region, and the input variables in the subject AMF (e.g., curve radius, lane width, shoulder width, etc.). These variables are kept in the multivariate model regardless of whether they are found to be statistically significant.

All key variables that have a continuous value, except volume and length, are converted to discrete values and included in the model as categorical variables. Discrete variables, such as the presence of a turn bay, do not require this conversion, and an indicator variable can be used (i.e., $x_i = 1.0$ if bay is present, 0.0 if it is not present). In contrast, variables that are continuous (e.g., radius, lane width, etc.) should be converted into scalar categories, with the value for each category representing a small range of values for the input variable. For example, to calibrate a lane width AMF, the lane width of each segment could be rounded to the nearest 1.0 ft and converted to scalar categories of 9, 10, 11, 12, and 13 ft. Each of these categories would have its unique indicator variable and calibration coefficient b_i in the regression model. The number of categories used for lane width would depend primarily on the range of values associated with the variable in the database and the number of observations in each category.

The advantage of using categorical variables in the regression model is that they do not impose any preconceived functional relationship on the estimates obtained from the model. If a functional relationship is used in the regression model (e.g., $b_i \times$ lane width), then this relationship could be reflected in the expected values used to calibrate the AMF. It follows that, if a functional relationship is used in the model, then it could indirectly bias the AMF calibration in Step 5. The use of categorical variables minimizes the potential for this type of bias because it does not require the specification of a function for the subject AMF input variables at this point in the process. If the subject AMF has several continuous input variables, then each variable would be converted into discrete values and inserted into the multivariate model as a categorical variable. The conversion of continuous variables into categorical variables may moderate the underlying functional relationship, especially if there are only a few categories established.

The second type of variable to include in the multivariate model is the supplemental variable (e.g., speed limit). Ideally, the regression database would contain subsets such that all included segments have identical attributes (except for the key variables) and that these attributes would be considered typical for the type of highways being considered. This approach would eliminate the need for supplemental variables. However, it is not always possible to find a sufficient number of "identical" segments while the required minimum sample size is maintained for statistical significance. In addition, supplementary variables can be used to explore interactions with key variables in the AMF model. In this case, the need for identical segments is relaxed, and a categorical variable is included in the model to account for differences in the supplemental variable among sites. A supplemental variable is kept in the model only if its category (i.e., the collective set of indicator variables) is statistically significant.

It is assumed that crash occurrence at a particular location follows a Poisson distribution and that the mean crash frequency for a group of similar segments follows a gamma distribution. In this manner, the distribution of crashes for a group of similar locations can be described by the negative binomial distribution. For highway segments, the variance of this distribution is

$$V[X] = yE[N] + \frac{(yE[N])^2}{kL} \quad (3)$$

where

$V[X]$ = crash frequency variance for a group of similar locations,
X = reported crash count for y years (crashes),
y = time interval during which X crashes were reported (years), and
k = overdispersion parameter (1/mi).

Equation 3 includes a variable for the length of the segment. As demonstrated by Hauer (*7*), this variable should be added to ensure that the regression model coefficients are not biased by exceptionally short segments.

Step 3. Estimate the Segment's Expected Crash Frequency

The objective of this step is to estimate the expected crash frequency for each before segment in the SP database. For the analysis of curve radius, the before segments are the tangent segments. The EB method developed by Hauer (*1*) is used to compute this estimate. The expected crash frequency $E[N]$ from the multivariate regression model represents the average crash frequency for all segments that have the same volume, length, and value for any other variables in the regression model. The EB method combines the expected crash frequency $E[N]$ with the observed count of crashes X for the subject segment to yield a best estimate of the expected crash frequency for that specific segment. This estimate is obtained by using the following equations:

$$E[N|X] = E[N] \times \text{weight} + \frac{X}{y} \times (1 - \text{weight}) \quad (4)$$

with

$$\text{weight} = \left(1 + \frac{E[N]y}{kL}\right)^{-1} \quad (5)$$

where $E[N|X]$ is the expected crash frequency given that X crashes were reported in y years, (crashes per year) and weight is the relative weight given to the prediction of expected crash frequency.

The variance of the expected crash frequency can be estimated by using the following equation:

$$V[N|X] = (1 - \text{weight})\frac{E[N|X]}{y} \quad (6)$$

where $V[N|X]$ is the variance of $E[N|X]$.

Equation 6 indicates that the variance of the expected crash frequency is less than that of the observed crash count. This reduction in variance is important because $E[N|X]$ is used in Step 5 as the independent variable in a regression model (as opposed to X). Weed and Barros (*8*) have shown that significant variability in the independent variable causes bias in the model's coefficients and increases its residual error.

Step 4. Assemble Group Database

The objective of this step is to assemble the SP group database by sorting the SP data into groups. Each SP is assigned to the one group in which all members have a similar before and a similar after value for the geometric element of interest. The aggregation is based on the specified geometric element (e.g., lane width, shoulder width, or curve radius) for which the AMF is being calibrated. For example, if five lane-width categories are represented in the SP group database (e.g., 9, 10, 11, 12, and 13 ft), then there can be as many as 25 (= 5 × 5) unique combinations of lane width for the before and the after conditions represented by each SP.

Two values are computed for each group. One value represents the sum of the expected crash frequency for each before segment $E[N|X]_{\text{before}}$ in the group. The second value represents the sum of the reported crash frequency for each after segment X_{after} in the group.

Step 5. Calibrate the AMF Model

The objective of this step is to calibrate the subject AMF model. This objective is achieved by using regression to relate the paired crash frequencies in the SP group database. The regression model used in this step includes the subject AMF. The form of the regression model is

$$X_{\text{after}} = c_0\, yE[N|X]_{\text{before}}\, \text{AMF}_m \quad (7)$$

with

$$\text{AMF}_m = f(c_j, x_j) \quad (8)$$

where

X_{after} = sum of reported crash frequency for after segments in a group (crashes),
$E[N|X]_{\text{before}}$ = sum of expected crash frequency for before segments in a group, (crashes per year),
c_0 = calibration coefficient,
c_j = calibration coefficient ($j = 0, 1, 2, 3, \ldots$), and
AMF_m = AMF for geometric element m, as a function of geometric variables x_j and calibration coefficients c_j.

The use of the expected crash frequency $E[N|X]$ for the before segments in Equation 7 (instead of the reported crash frequency X for the before segments) is intended to minimize the variability associated with the independent variable. This step assumes that the uncertainty associated with the coefficients in the multivariate regression model is negligible.

The AMF model represented by Equation 8 can have the form of a function, such as the following AMF for shoulder width:

$$\text{AMF}_{\text{SW}} = e^{c_1(\text{SW}_{\text{with}} - \text{SW}_{\text{without}})} \quad (9)$$

where

AMF_{SW} = shoulder width accident modification factor,
c_1 = constant describing the relationship between a change in shoulder width and a change in crash frequency,
SW_{with} = shoulder width on segment for which width is different from that of base segment (ft), and
$SW_{without}$ = shoulder width of base segment (ft).

The AMF could also be a constant. Regardless of whether it is a function or a constant, regression analysis is used to calibrate the AMF.

A Poisson distribution is assumed for the dependent variable in Equation 7 given the nature of the SP group database. The expected crash frequency $E[N|X]_{before}$ is used as an offset variable in this application. Model fit is based on maximum-likelihood methods that minimize the scaled deviance for the Poisson function. This function is

$$SD = 2\sum_i \left[y_i LN\left(\frac{y_i}{u_i}\right) - \left(y_i - u_i\right) \right] \tag{10}$$

where

SD = sum of scaled deviance for all i observations,
y_i = ith observation, and
u_i = model prediction for ith estimate.

Step 6. Assess Goodness of Fit

Several statistics are available for assessing the fit of the models developed in Steps 2 and 5. One measure of model fit is the Pearson χ^2 statistic. This statistic is calculated as

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(X_i - yE[N]_i\right)^2}{V[X]_i} \tag{11}$$

where n is the number of observations.

This statistic follows the χ^2 distribution with $n - p$ degrees of freedom, where n is the number of observations (i.e., segments) and p is the number of model variables (*9*). This statistic is asymptotic to the χ^2 distribution for larger sample sizes.

The root mean square error is a useful statistic for describing the precision of the model estimate. It represents the standard deviation of the estimate when each independent variable is at its mean value. This statistic can be computed as

$$s_e = \frac{1}{y}\sqrt{\frac{\sum_{i=1}^{n}\left(X_i - yE[N]_i\right)^2}{n-p}} \tag{12}$$

where s_e is the root mean square error of the model estimate, crashes per year.

A more subjective measure of model fit can be obtained from a graphical plot of the Pearson residuals versus the expected value of the dependent variable (e.g., $E[N]$). This type of plot provides a graphical means of assessing the predictive capability of the model. A well-fitting model would have the residuals symmetrically centered around zero over the full range of the dependent variable, most clustered near zero, with a spread ranging from about –3.0 to +3.0. The Pearson residual PR_i for segment i can be computed as

$$PR_i = \left(X_i - yE[N]_i\right)\sqrt{\frac{1}{V[X]_i}} \tag{13}$$

The scale parameter φ is used to assess the amount of variation in the observed data relative to the specified distribution (*9*). This statistic can be calculated by dividing Equation 11 by the quantity $n - p$. A scale parameter near 1.0 indicates that the assumed distribution of the dependent variable is approximately equivalent to that found in the data (i.e., negative binomial or Poisson).

Another measure of model fit is the coefficient of determination R^2. This statistic is commonly used for normally distributed residuals. However, it provides some useful interpretation when applied to data from other distributions and computed in the following manner (*10*):

$$R^2 = 1 - \frac{SSE}{SST} \tag{14}$$

with

$$SSE = \sum_{i=1}^{n}\left(X_i - yE[N]_i\right)^2 \tag{15}$$

$$SST = \sum_{i=1}^{n}\left(X_i - X.\right)^2 \tag{16}$$

where $X.$ is the average crash frequency for all n observations.

The last measure of model fit is the dispersion parameter–based coefficient of determination R_k^2. This statistic was developed by Miaou (*11*) for use with data that exhibit a negative binomial distribution. It is computed as

$$R_k^2 = 1 - \frac{k_{null}}{k} \tag{17}$$

where k_{null} is the overdispersion parameter based on the variance in the observed crash frequency.

The null overdispersion parameter k_{null} represents the dispersion in the observed crash frequency relative to the overall average crash frequency for all segments. This parameter can be obtained by using a null model formulation (i.e., a model with no independent variables but with the same error distribution, link function, and offset).

APPLICATION

This section describes an application of the procedure developed in earlier. The application is focused on the development of a horizontal curve radius AMF for rural two-lane highways.

AMF for Horizontal-Curve Radius

The AMF for horizontal-curve radius developed by Harwood et al. (*3*) is

$$AMF_{cr} = \frac{1.55L_c + \frac{80.2}{R_c} - 0.012I_S}{1.55L_c} \tag{18}$$

where

AMF_{cr} = AMF for horizontal-curve radius,
L_c = length of horizontal curve (= $I_c R_c$ / 5280 / 57.3) (mi),
I_c = curve deflection angle (degrees),
I_S = spiral transition curve presence (= 1.0 if spiral present, 0.0 if not present), and
R_c = curve radius (ft).

This AMF is compared with the AMF developed by using the proposed procedure.

Site Selection and Data Collection

The segments used for this application were obtained from the TxDOT roadway database. The curve segments had to satisfy the following criteria:

- Cross section: undivided, two through lanes, no median;
- Area type: rural;
- Minimum curve length: 0.1 mi;
- Intersection presence: no intersections;
- Shoulder type: shoulder present (no curbed cross sections);
- Shoulder width: 4 to 13 ft;
- Lane width: 11 to 13 ft; and
- Curve transition: tangent to circular curve (i.e., no spiral present).

The curve segments were initially identified by using the criteria above. Then, a nearby tangent segment was identified. The pair was retained if the characteristics of the tangent segment (e.g., cross section, area type, intersection presence, shoulder width, lane width, speed limit, etc.) matched those of the curve segment. Speed limit ranged from 55 to 70 mph in the database.

Experience with the database indicated that many of the shorter segments tended to have no reported crashes during a period of several years. This trend was also exhibited by segments that had very low traffic volume. When a short or low-volume segment is associated with one or more crashes, it can exhibit significant leverage on the regression model coefficients and increase the Pearson χ^2 statistic in a disproportionate manner relative to other segments. To address this concern, Equation 13 was used to derive the Equation 19 for computing the minimum segment exposure:

$$E_{\min} = \frac{\mathrm{PR}^2 + 2 - \sqrt{\left(\mathrm{PR}^2 + 2\right)^2 - 4}}{2 \text{ base } y} \tag{19}$$

where $E_{\min}$ is the minimum segment exposure associated with prediction ratio (PR) = 3.0 [million vehicle miles (mvm)] and base is the injury (plus fatal) crash rate (crashes per mvm).

Equation 19 is based on the conservative assumptions that $V[X]_i$ is equal to $E[N]$ and that X_i is equal to 1.0 crash. When PR is set to 3.0, the estimated crash rate base is 0.23 crash per mvm, and the time interval for the crash data y is 3 years; the resulting minimum exposure is 0.13 mvm. Equation 20 was used to compute the exposure for all candidate segments. Only those segments exceeding 0.13 mvm were included in the database.

$$E = 0.000365 \text{ ADT } L \tag{20}$$

where E is the segment exposure (mvm).

The minimum-exposure criterion defined by Equation 19 was used in the selection of sites for the SP regression database. In contrast, it was not used when the SP group database was being formed because each observation in this database represents the summation of many segments such that one segment could have undue leverage on the model.

Site Characteristics

As described earlier, two databases were assembled for the curve radius AMF calibration. In total, 3,514 segments (1,757 curved and 1,757 tangent) were identified for the SP database. They represented a total of 335.4 rural two-lane highway miles. Furthermore, 1,382 segments (691 curved and 691 tangent) were identified for the SP regression database as a subset of the SP database. The smaller number of segments in the SP regression database is due to the use of the minimum-exposure criterion described in previous section on site selection and data collection. These segments represented a total of 152.2 mi.

Data Collection

Crash data were identified for each segment by using the Texas Department of Public Safety database. Three years of crash data, corresponding to 1999, 2000, and 2001 were identified for each segment. The ADT for each of these 3 years was obtained from the Texas Reference Marker database and averaged to obtain one ADT for each segment. Crashes that were associated with intersections and driveways were identified as noncurve-related and excluded from the database.

Data Analysis

Database Summary

The crash data for each of the segment pairs are summarized in Table 1. The segments in the SP regression database were associated with 566 crashes, of which 257 occurred on the tangent segments and 309 on the curved segments. The segments in the SP database were associated with 822 crashes, of which 349 occurred on the tangent segments and 473 on the curved segments. Segment crash rates are provided in the last two columns of the table. Typical injury (plus fatal) crash rates for rural two-lane highways range from 0.2 to 0.3 crashes per mvm. The injury (plus fatal) crash rates for the tangent segments listed in Table 1 are within this range.

Model Development and Statistical Analysis Methods

This subsection describes the form of the multivariate model and the AMF calibration model. A separate formulation was used for each model and was dictated by the variables designated as inputs to the subject AMF (i.e., the AMF for horizontal-curve radius). In addition, the Texas region was also included as a model variable in the two formulations to facilitate exploration of regional influence on curve crash frequency.

The general form of the multivariate model is

$$E[N] = \mathrm{ADT}^{b_1} L e^{\left(b_0 + \sum_{2}^{6} b_i \deg_i + \sum_{7}^{11} b_j \mathrm{region}_i\right)} \tag{21}$$

TABLE 1 Crash Data Summary for Tangent-Curve Segment Pairs

Database	Segment Type	Exposure,[a] mvm	Crashes/3 Years			Crash Rate, cr/mvm	
			PDO[b]	I + F[c]	Total	I + F[c]	Total
SP regression	Tangent	209.2	100	157	257	0.25	0.41
	Curve	209.2	130	179	309	0.29	0.49
	Total:	418.4	230	336	566		
Segment pair	Tangent	274.8	136	213	349	0.26	0.42
	Curve	274.8	193	280	473	0.34	0.57
	Total:	549.6	329	493	822		

[a]mvm: million vehicle miles.
[b]PDO: property-damage-only crashes.
[c]I+F: injury plus fatal crashes.

where

b_0 = calibration coefficient corresponding to west region and 6° curve,
region_i = categorical variable for Texas region (five levels: central, northeast, north, southeast, and south), and
deg_i = categorical variable for degree of curve (= 5,730/R_c) (five levels: 0, 1°, 2°, 3°, and 4°).

The generalized modeling procedure (GENMOD) in SAS was used to automate the regression analysis. This procedure estimates model coefficients by using maximum-likelihood methods for a specified error distribution.

The general form of the AMF calibration model is

$$X_{\text{curve}} = c_0 y E\left[N \middle| X\right]_{\text{tangent}} \text{AMF}_{\text{cr}} \tag{22}$$

with

$$\text{AMF}_{\text{cr}} = 1 + \left(d_0 + \sum_1^5 d_i \text{ region}_i\right) \frac{80.2}{1.55 L_c R_c} \tag{23}$$

where d_i is the calibration coefficients (i = 1, 2, 3, . . . , n) and d_0 is the calibration coefficient corresponding to the central region. Equation 23 was derived from Equation 18 for curves without spiral transitions.

The nonlinear regression procedure (NLIN) in the SAS software was used to estimate the calibration model coefficients. Like GENMOD, this procedure also estimates model coefficients by using maximum-likelihood methods. The benefit of using this procedure is that it is not constrained by additive model terms. Rather, it can be used to evaluate complex AMF model forms such as Equation 23. The loss function associated with NLIN was specified to equal the scaled deviance for the Poisson distribution (i.e., Equation 10).

Model Calibration

Multivariate Model

The regression analysis of the multivariate model is presented in Table 2. The calibration of this model was based on injury (plus fatal) crash frequency. The variables and coefficients listed in this table correspond to those identified in Equation 21.

The quality of model fit is indicated by the statistics in the top portion of Table 2. The Pearson χ^2 statistic for the model is 1,373, and the degrees of freedom are 1,370 (= $n - p$ = 1,382 −12). As this statistic is less than $\chi^2_{0.05,\,1,370}$ (= 1,457), the hypothesis that the model fits the data cannot be rejected. R_k^2 for the calibrated model is .22. This statistic indicates that about 22% of the variability due to systematic sources is explained by the model.

The statistics for each categorical variable as a group are listed in the last two rows of Table 2. The p-value of .11 suggests that there is not a significant difference between the regions. However, the degree-of-curve and region variables are key categorical variables and were kept in the model regardless of their significance.

Calibration Model

The regression analysis of the AMF calibration model (i.e., Equation 22) revealed that Equation 23 was found to have formulation problems that resulted in a poor fit to the data. The correlation between curve length and the AMF value was found to be relatively weak. A similar lack of correlation was found for the effect of region. The revised AMF model form that reflects these findings is

$$\text{AMF}_{\text{cr}} = 1 + d_0 \left(\frac{5730}{R_c}\right)^2 \tag{24}$$

The statistics related to the AMF calibration model are shown in Table 3. The calibration coefficient c_0 in Equation 22 was not significantly different from 1.0 and was removed. The calibration is based on the frequency of injury (plus fatal) crashes. The SP group database consisted of 56 unique combinations of curve radius and segment length (as obtained from the SP database). The variables and coefficients listed in Table 3 correspond to those identified in Equation 24. The Pearson χ^2 statistic for the model is 52.1, and the degrees of freedom are 55 (= $n - p$ = 56 −1). As this statistic is less than $\chi^2_{0.05,\,55}$ (= 73.3), the hypothesis that the model fits the data cannot be rejected. R^2 for the model is .91.

The crash data assembled for the curve radius AMF analysis excluded driveway-related crashes. An examination of the database with, and without, these crashes indicated that they constitute about 20% of all segment crashes for segments located in the vicinity of a curve. Thus, Equation 24 was adjusted to reflect its

TABLE 2 Multivariate Model Statistical Description

Variable	Variable Name	Units	Minimum	Maximum
Range of Model Variables				
ADT	Segment ADT	vpd	770	17,600
L	Segment length	miles	0.10	0.82
Deg	Degree of curvature (= 5,730 / R_c)	degrees	0	6

Variable	Definition	Value	Std. Dev.	t-Statistic
Calibrated Coefficient Value				
b_0	Intercept	−7.09	1.08	−6.5
b_1	Effect of segment ADT	0.847	0.119	7.1
b_2	Effect of 0-degree curvature (i.e., tangent)	−1.280	0.477	−2.7
b_3	Effect of 1-degree curvature	−1.355	0.487	−2.8
b_4	Effect of 2-degree curvature	−1.012	0.486	−2.1
b_5	Effect of 3-degree curvature	−0.803	0.519	−1.5
b_6	Effect of 4-degree curvature	−0.522	0.643	−0.8
in b_0	Effect of 6-degree curvature	0.000		
b_7	Effect of central region	0.403	0.289	1.4
b_8	Effect of northeast region	0.179	0.282	0.6
b_9	Effect of north region	−0.089	0.378	−0.2
b_{10}	Effect of southeast region	0.524	0.269	1.9
b_{11}	Effect of south region	0.352	0.270	1.3
in b_0	Effect of west region	0.000		

NOTE:

Model Statistic	Value
R^2 (R_k^2) 0.05 (0.22)	
Scale parameter ϕ	1.00
Pearson χ^2 1,373 ($\chi^2_{0.05,\,1,370}$ = 1,457)	
Overdispersion parameter k	11.3 mi^{-1}
Observations n_o	1,382 segments (336 injury + fatal crashes in 3 years)
Standard deviation s_e	±0.17 crashes/segment/year

Category	Definition	Deg. Freedom	Chi-Square	p-Value
Categorical Variable Statistics				
Deg	Degree of curve	5	15.0	0.01
Region	Region	5	9.1	0.11

TABLE 3 AMF Calibration Model Statistical Description

Model Statistic	Value
R^2	0.91
Scale parameter ϕ	0.95
Pearson χ^2	52.1 ($\chi^2_{0.05,\,55}$ = 73.3)
Observations n_o	56 curve radius and length combinations (493 injury + fatal crashes in 3 years)
Standard deviation s_e	±0.62 crashes/segment/year

NOTE: Range of model variables—R_c = curve radius; minimum, 955 ft; maximum, 57,300 ft. Calibrated coefficient values—d_0 = effect of radius; value = 0.133, SD = 0.020, t-statistic = 6.7.

focus on non-driveway-related crashes. This adjustment was based on a weighted average technique, where the AMF_{cr} from Equation 24 (for nondriveway crashes) was weighted by 0.80 and an AMF_{cr} = 1.0 (for driveway crashes) was weighted by 0.20. The adjusted AMF is

$$AMF_{cr} = 1 + 0.106\left(\frac{5730}{R_c}\right)^2 \quad (25)$$

Sensitivity Analysis

The revised AMF model is shown in Figure 2 for a range of radii. The values obtained from the revised model are shown with a solid trend line. The values obtained from Equation 18 are shown with two dashed

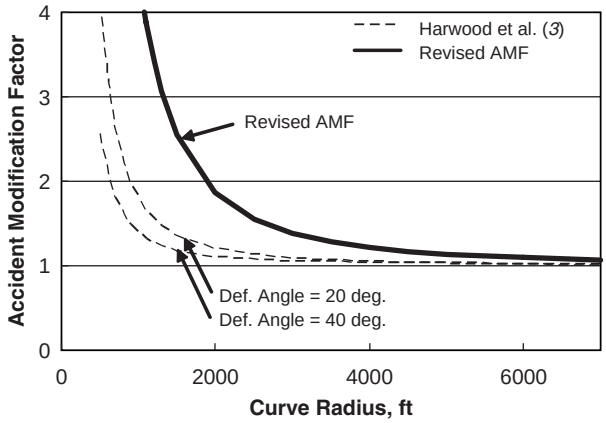

FIGURE 2 Relationship between radius and AMF value.

lines. This equation is sensitive to curve length; however, it was converted to include a sensitivity to curve deflection angle I_c instead by using the relationship between curve length and deflection angle provided in the variable definitions associated with Equation 18. This conversion was performed to facilitate a more equitable presentation of Equation 18 for the range of radii shown in Figure 2.

The revised AMF values are larger than those obtained from Equation 18, although the difference diminishes with increasing radius. One reason for this difference may be that the model developed by Harwood et al. (*3*) is based on the full range of crash severities, while that developed in this research is focused on injury (plus fatal) crashes. The difference shown suggests that curve crashes tend to be more severe. A second reason for this difference may be that the data used to develop Equation 18 were obtained from Washington State. The terrain and climate of this state are quite different from that found in Texas.

CONCLUSIONS AND RECOMMENDATIONS

A procedure for developing AMFs by using cross-sectional data was described in this paper. It is based on the use of matched SPs and appears viable when applied to databases representing large roadway systems (e.g., a state DOT database) given the likelihood of finding a large number of nearby segments with similar features. It is likely to be more difficult to apply to intersections given the tendency for nearby intersections to be dissimilar.

The procedure was demonstrated by application to data describing the geometry, traffic, and crash history for two-lane highways in Texas. It was shown to produce an AMF that describes the relationship between curve radius and crash risk. The calibrated AMF is consistent with another curve AMF previously derived by using data from Washington. However, the calibrated AMF indicates that a greater crash risk exists on the Texas curves for a given radius.

Further research is needed to determine whether there are any statistical implications (e.g., bias, loss of power) associated with the AMF estimate due to the use of the same data to calibrate the multivariate model and the AMF calibration model.

ACKNOWLEDGMENTS

The research reported here was sponsored by TxDOT as the project Incorporating Safety into the Highway Design Process. The authors recognize Elizabeth Hilton of the TxDOT for her support and guidance throughout this project.

REFERENCES

1. Hauer, E. *Observational Before–After Studies in Road Safety.* Pergamon Press, Oxford, United Kingdom, 1997.
2. Bonneson, J. A., D. Lord, K. Zimmerman, K. Fitzpatrick, and M. Pratt. *Development of Tools for Evaluating the Safety Implications of Highway Design Decisions.* FHWA/TX-07/0-4703-4. Texas Department of Transportation, Austin, Sept. 2006.
3. Harwood, D. W., F. M. Council, E. Hauer, W. E. Hughes, and A. Vogt. *Prediction of the Expected Safety Performance of Rural Two-Lane Highways.* FHWA-RD-99-207. FHWA, U.S. Department of Transportation, 2000.
4. Washington, S., and J. Oh. Bayesian Methodology Incorporating Expert Judgment for Ranking Countermeasure Effectiveness Under Uncertainty: Example Applied to At-Grade Railroad Crossings in Korea. *Accident Analysis and Prevention.* Vol. 38, No. 2, 2006, pp. 234–247.
5. Gross, F., and P. P. Jovanis. Estimation of Safety Effectiveness of Changes in Shoulder Width with Case Control and Cohort Methods. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2019,* Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 237–245.
6. Hauer, E. Cause and Effect in Observational Cross-Section Studies on Road Safety. Feb. 2005. roadsafetyresearch.com/. Accessed Oct. 2007.
7. Hauer, E. Over-dispersion in Modeling Accidents on Road Sections and in Empirical Bayes Estimation. *Accident Analysis and Prevention.* Vol. 33, 2001, pp. 799–808.
8. Weed, R. M., and R. T. Barros. Demonstration of Regression Analysis with Error in the Independent Variable. In *Transportation Research Record 1111,* TRB, National Research Council, Washington, D.C., 1987, pp. 48–54.
9. McCullagh, P., and J. A. Nelder. *Generalized Linear Models.* Chapman and Hall, New York, 1983.
10. Kvalseth, T. O. Cautionary Note About R^2. *American Statistician,* Vol. 39, No. 4, Nov. 1985, pp. 279–285.
11. Miaou, S. P. *Measuring the Goodness-of-Fit of Accident Prediction Models.* FHWA-RD-96-040. FHWA, U.S. Department of Transportation, 1996.

The Safety Data, Analysis, and Evaluation Committee sponsored publication of this paper.