**NCHRP 17-93: Updating Safety Performance Functions for Data-Driven Safety Analysis**

**Working White Paper: Development of Guidelines for Quantifying a**

**Reliable Local Calibration Factor**

**Prepared for:**

**National Cooperative Highway Research Program**

**Transportation Research Board**

**The National Academies**

**TRANSPORTATION RESEARCH BOARD OF THE NATIONAL ACADEMIES OF SCIENCES, ENGINEERING AND MEDICINE**

**PRIVILEGED DOCUMENT**

This document, not released for publication, is furnished only for review to members of or participants in the work of NCHRP. This document is to be regard as fully privileged, and dissemination of the information included herein must be approved by NCHRP.

Submitted by:

Raghavan Srinivasan and Taha Saleem

The University of North Carolina at Chapel Hill

Highway Safety Research Center

James Bonneson

Kittelson and Associates

Vikash Gayah and Kristen Kersavage

VHB, Inc.

October 2023

**ACKNOWLEDGMENT OF SPONSORSHIP**

This work was sponsored by the American Association of State Highway and Transportation Officials, in cooperation with the Federal Highway Administration, and was conducted in the National Cooperative Highway Research Program, which is administered by the Transportation Research Board of the National Research Council.

**DISCLAIMER**

This is an uncorrected draft as submitted by the contractor. The opinions and conclusions expressed or implied herein are those of the contractor. They are not necessarily those of the Transportation Research Board, the Academies, or the program sponsors.

TABLE OF CONTENTS

[Introduction 4](#_Toc158129479)

[Development of Outlier Detection Techniques 4](#_Toc158129480)

[The Empirical Setting 4](#_Toc158129481)

[Review and Evaluation of Outlier Detection Techniques 6](#_Toc158129482)

[Development of Procedures for Computing the Calibration Factor CV 12](#_Toc158129483)

[Review of Procedures for Computing the Calibration Factor CV 12](#_Toc158129484)

[Evaluation of Alternative Procedures 14](#_Toc158129485)

[Development of Guidelines for Using a CURE Plot to Assess Model Fit 15](#_Toc158129486)

[Review of Guidelines for Using the CURE Plot 15](#_Toc158129487)

[Evaluation of CURE Plot Guidelines 16](#_Toc158129488)

[References 20](#_Toc158129489)

Introduction

This paper describes the development of guidelines for quantifying a reliable local calibration factor for a crash prediction model (CPM). These guidelines are applicable to the CPMs described in Part C of the *Highway Safety Manual* (HSM) (AASHTO, 2010). However, they can also be applied to other CPMs.

The HSM provides guidelines for assembling the calibration database and computing the calibration factor. However, it does not provide guidelines for assessing the reliability of the computed calibration factor. The reliability of the calibration factor directly affects the reliability of the estimate obtained from the CPM in which the factor is used. An unreliable calibration factor will ensure that the CPM estimate is also unreliable.

This paper consists of three sections. The first section describes the development of techniques for detecting the presence of outliers in the calibration database. The second section describes the development of a procedure for computing the calibration factor coefficient of variation (CV). The third section describes the development of revised guidelines for using the cumulative residual (CURE) plot to assess the fit of the calibrated CPM.

Development of Outlier Detection Techniques

Hauer (2015) describes an outlier as “…an observation that appears to be unlikely in view of most other observations.” In the context of computing the local calibration factor, a site (e.g., intersection, segment) is considered an *extremely rare case* when its predicted average crash frequency is significantly different from its observed annual crash count, relative to all other sites in the calibration database. If this difference (i.e., “residual error”) is a result of an inaccurate prediction or incorrect observed crash count, then the site is considered an *outlier*. If one or more outliers are included in the data used to compute the calibration factor, the value of this factor will likely be biased.

There are a variety of statistical tests (e.g., Grubb’s test) and graphical techniques (e.g., box plot) that can be used to identify extremely rare cases. Many of the tests become more reliable as the residual error distribution approaches the normal distribution. However, this characteristic tends to limit their value when applied to crash count data, which tends to follow the negative binomial distribution.

As noted in the previous paragraph, there are tests and techniques that can be used to verify the “necessary” condition for outlier detection (i.e., an extremely rare case is present in the data). However, tests and techniques do not exist for verifying the “sufficiency” condition (i.e., the residual error is due to an inaccurate prediction or incorrect crash count). For this reason, there is some subjectivity in the determination of whether an observation (i.e., site) is truly an outlier.

This section consists of two subsections. The first subsection describes the process used to simulate observed crash data for the purpose of creating artificial calibration databases that can be used to evaluate alternative outlier detection techniques. The second subsection describes a statistical test and two graphical techniques for identifying extremely rare cases in a CPM calibration database.

The Empirical Setting

The evaluation of alternative outlier detection techniques was based on the use of simulated CPM calibration databases. The evaluation used these databases to compare alternatives, identified strengths and weaknesses, and, in some cases, develop needed refinements. A large number of unique calibration databases were needed to ensure that the recommended tests and techniques are robust for a wide range of realistic conditions.

A spreadsheet tool was developed to synthetically generate a calibration database containing data for a specified number of sites (ranging from 30 to 1000 sites). For each database, an AADT-only safety performance function (SPF) was established. The SPF intercept value was computed to produce a specified input overall average crash frequency for the collective set of sites. Each site was assigned a unique AADT value that ranged from 1,000 to 11,000 veh/day. The inverse dispersion parameter was a specified input value.

For each site in the database, the SPF was used to compute the predicted average crash frequency. A Monte Carlo simulation technique was then used with the gamma distribution to estimate the “observed” average crash frequency for each site (based on the predicted average crash frequency and specified overdispersion parameter).

Next, for each site in the database, a second Monte Carlo simulation technique was used with the Poisson distribution to estimate the observed annual crash count for each site (based on the “observed” average crash frequency).

Next, the following equation was used to compute the calibration factor for the calibration database:

Equation

with

Equation

Equation

where:

*C =* calibration factor to adjust the CPM for local conditions;

*No,i* = observed crash count for site *i* during the calibration period(crashes/period);

*No,i,j* = observed annual crash count for site *i* and year *j* (crashes/yr);

*Np,u,i =* predicted average number of crashes for site *i* during the calibration period and unadjusted by the calibration factor (crashes/period);

*Np,u,i,j =* predicted average crash frequency for site *i* and year *j* unadjusted by the calibration factor (crashes/yr);

*nsites* = number of sites in the calibration database (sites); and

*nca* = number of years in the calibration period (yr).

Finally, the following equation was used to compute the *adjusted* predicted average crash frequency:

Equation

where *Np,a,i,j* is the predicted average crash frequency for site *i* and year *j* adjusted by the calibration factor (crashes/yr); and all other variables are previously defined.

For all simulated calibration databases, the underlying true calibration factor is 1.0 given the manner in which the simulated database was constructed. However, any given database is unlikely to have a computed calibration factor exactly equal to 1.0 because the observed crash counts vary randomly (following the negative-binomial distribution). When several databases with *X* sites per database are created and a calibration factor is computed for each database, the average of these factors has smaller variation and converges to 1.0 as *X* increases.

The spreadsheet tool is developed such that the distribution of crashes for each site in the database is described by the negative binomial distribution. The variance of this distribution is computed using the following equation.

Equation

with

Equation

where

*V*[*X*]*i* = crash frequency variance for site *i* during the calibration period (crashes2);

*Np,a,i =* predicted average number of crashes for site *i* during the calibration period and adjusted by the calibration factor (crashes/period);

*X* = reported crash count for site *i* (crashes);

*Ls,i* = segment length for site *i* (mi) (= 1.0 if the database represents intersections);

*K* = inverse dispersion parameter (mi-1);

and all other variables are previously defined.

The overdispersion parameter, as defined and used in the HSM, is computed using the following equation:

Equation

where *ki* = overdispersion parameter for site *i* and other variables are as previously defined.

Review and Evaluation of Outlier Detection Techniques

This section describes a statistical test and two graphical techniques for identifying extremely rare cases in a CPM calibration database. The statistical test is based on the use of the negative binomial distribution to determine the probability of each site’s observed annual crash count, given its adjusted predicted average number of crashes. Extremely rare cases are associated with a very small probability of being exceeded.

Statistical Test for Identifying Extremely Rare Cases

Test Statistic

If the distribution of the average number of crashes for all sites with a given set of characteristics (e.g., AADT, lane width) can be described by the gamma distribution and the distribution of crash counts can be described by the Poisson distribution, then the distribution of crash counts at these similar sites is described by the negative binomial distribution (Hauer, 2015). The probability mass function of the negative binomial distribution is described using the following equation.

Equation

with,

Equation

where:

*P*(*No,i*) *=* probability that the observed crashes at site *i* will equal *No,i,ca*;

*pi =* probability of success for site *i*;

*kT,i* = truncated overdispersion parameter for site *i* (= Integer[*ki*]);

and all other variables are previously defined.

The Excel spreadsheet function for the negative binomial distribution is NEGBINOM.DIST(*No,i*, 1/*ki*, *pi*, TRUE for cumulative distribution or FALSE for probability mass function).

The overdispersion parameter *ki* for the negative binomial distribution can be estimated for the calibration database using various methods. If this parameter is a constant, then the method of moments or regression analysis can be used to estimate the parameter value. However, if the parameter is a function of segment length, then regression analysis should be used to estimate the inverse dispersion parameter value *K*. This parameter is then used Equation 7 is used to compute the overdispersion parameter *ki* for each site *i*.

When regression analysis is needed, one approach that can be used is to convert Equation 5 into a linear model with no intercept term. This linear model is shown in the following equation.

Equation

where *Ls,i* equals the segment length for site *i* (mi) (= 1.0 if the site is an intersection); and all other variables have been previously defined. The quantity “(*No,i* − *Np,a,i*)2” represents an estimate of the crash frequency variance *V*[*X*]. The term on the left side of the equals sign in Equation 10 represents the dependent variable *Y*. The quantity “(*Np,a,i*/*Ls,i*)” on the right side represents the independent variable *X*. The value 1/*K* represents the regression coefficient.

Several options are available for finding the best estimate of *K* in Equation 10. Minimization of the squared residuals is one option that provides a closed-form solution. It is equally applicable for quantifying an overdispersion parameter that is a constant or a function of segment length. In this regard, the following equation is used to compute least-squares estimate of the inverse dispersion parameter *K*.

Equation

with,

Equation

where all variables have been previously defined. If the sites are intersections, then the value of *Ls,i* is set equal to 1.0.

Outlier Detection Technique

To test for an extremely rare case, the cumulative form of the negative binomial distribution can be used to compute the probability of a site being associated with the observed crash count, given its predicted average number of crashes. If this probability is larger (or smaller) than a specified small value then the site is considered to be an extremely rare case.

In a calibration database of *nsites*, the one site with the largest absolute value residual error should occur with a probability *a* equal to 1/*nsites*. Thus, this site would be identified as having a cumulative probability less than *a*/2 or in excess of 1− *a*/2.

An extremely rare case would likely occur with a probability that is smaller than *a*. For this application, an extremely rare case is defined as a site whose residual error is so large (or small) that it would occur once in every other calibration database (i.e., with a probability α equal to 0.5/*nsites*). Thus, the extremely rare case is identified as having a cumulative probability that is less than α/2 or in excess of 1− α/2.

This technique can be used to verify the “necessary” condition for outlier detection (i.e., one or more extremely rare cases are present in the data). To determine if an extremely rare site is an outlier, the analyst will need to examine the associated data to determine whether one or more values are inaccurate and sufficiently similar to other sites in the database.

Evaluation of Predicted Inverse Dispersion Parameter

The predictive reliability of Equation 11 was evaluated using simulation data. The process for generating these data by simulation is described in a previous section titled The Empirical Setting. A series of calibration databases were computed for this evaluation. Each calibration database contains data for 100 intersection sites. The calibration period is 1 year. For a given true average crash frequency and inverse dispersion parameter, 200 calibration databases were developed (each database containing data for 100 sites). Equation 11 was used to compute the predicted inverse dispersion parameter for each of the 200 databases. The median predicted inverse dispersion parameter for the 200 databases was used in the evaluation.

The process described in the preceding paragraph was repeated for a range of average crash frequencies and inverse dispersion parameters. Specifically, the true average crash frequencies considered included 1, 10, and 40 crashes/site/yr and the true inverse dispersion parameter values included 1, 2, and 8. In combination, there were nine (= 3 averages × 3 parameters) unique combinations of average crash frequency and inverse dispersion parameter considered. For each of the nine combinations, 200 calibration databases were generated with each database including 100 sites.

The findings from the evaluation are shown in Figure 1. Each data point shown represents one of the nine combinations mentioned above. The dashed trend line represents an “*x* = *y*” line such that a data point would lie on this line if the predicted parameter equaled the true parameter. The solid trend line represents the best-fit trend line to the data points (using least-squares regression with no intercept). The best-fit trend line indicates that Equation 11 tends to provide a good estimate of the true parameter value; however, it may overestimate it by about 9 percent.

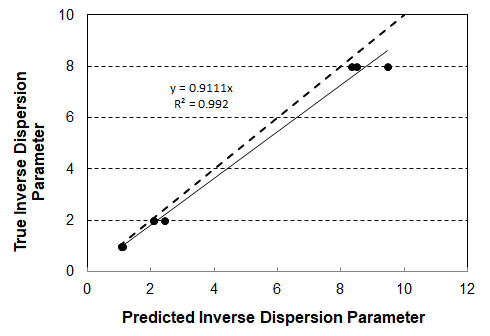


Figure . Comparison of predicted and true inverse dispersion parameter.

Graphical Techniques for Identifying Extremely Rare Cases

Two graphical techniques for identifying extremely rare cases are examined in this section. The first technique is the “Pearson residual” or standardized residual (McCullagh and Nelder, 1989) plotted against the predicted average number of crashes for the calibration period. The second technique is the CURE technique (Hauer, 2015). The CURE technique plots the cumulative sum of the residual errors against a selected independent variable. When using a CURE plot to identify extremely rare cases, the independent variable for the plot should be the predicted average number of crashes for the calibration period.

Standardized Residual Technique

This standardized residual is calculated using the following equation.

Equation

where *rp,i* is the standardized residual for site *i* and all other variables are previously defined.

The standardized residual is asymptotic to the normal distribution for databases having a larger overall average crash frequency and a smaller overdispersion parameter. For these conditions, extremely rare cases can be identified as having a cumulative normal probability that is less than α/2 or in excess of 1− α/2, where α equals 0.5/*nsites*. However, for other conditions, the standardized residual distribution is typically skewed such that the standardized residuals are not normally distributed.

In application, the standardized residual plot is constructed with each site in the calibration database shown as one data point. For a given site, the *y*-axis describes the standardized residual value and the *x*-axis describes the predicted average number of crashes for the crash period. The data points should cluster around a standardized residual value of 0.0. Points that are shown as exceptionally high (above the cluster) or low (below the cluster) on the plot are possible extremely rare cases. The likelihood of their being extremely rare increases with an increase in their vertical distance from the cluster.

Table 1 is lists some approximate lower and upper boundary values that can be used to identify extremely rare cases using standardized residuals. These boundaries were determined using simulated calibration data for a range of average-number-of-crashes and overdispersion-parameter values (the process for generating these data by simulation is described in a previous section titled, The Empirical Setting). The lower and upper boundaries listed in the table correspond approximately to the 0.1 percentile and 99.9 percentile. A standardized residual that is smaller than the lower boundary, or that is larger than the upper boundary, should be considered an extremely rare case.

Table . Standardized residual boundaries for identifying extremely rare cases.

|  |  |  |  |
| --- | --- | --- | --- |
| Average Number of Crashes () (cr/site/period) | Average Overdispersion Parameter () | Standardized Residual Boundaries | |
| Lower Boundary | Upper Boundary |
| ≥ 6 | 0.01 | -2.2 | 3.2 |
|  | 0.1 | -1.9 | 3.9 |
|  | 0.2 | -1.7 | 4.2 |
|  | 0.5 | -1.2 | 4.7 |
|  | 1.0 | -1.0 | 5.0 |
| < 6 | any | -1.1 | 4.9 |

Note: If the sites are segments with length *Ls,i* in miles, then for this table is computed as ; otherwise .

Cumulative Residual Technique

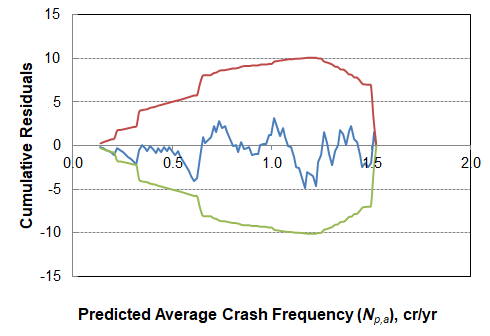
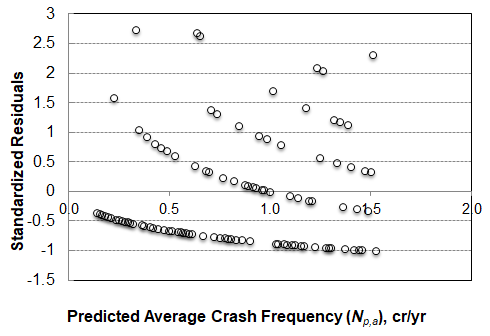
The CURE plot was developed by Hauer and Bamfo (1997) for assessing how well the specified regression model form fits the data. The plotted data points are created by sorting all sites in the database by the predicted average number of crashes for the calibration period. Then, the residual error for each observation is computed. Finally, the residual error is summed for each successive observation in the database. Thus, the cumulative residual for the first site in the sorted database equals the residual error for that site. The cumulative residual for the second site in the database equals the sum of the residual error for the first and second sites. This process continues until the cumulative residual for the last site is computed as the sum of all residual errors in the database. The mathematics for computing the cumulative residuals and for computing the associated confidence interval boundaries is described in a subsequent section.

In addition to its use as a tool for assessing the fit of the regression model, the CURE plot can also be used to identify extremely rare cases. Hauer (p. 103, 2015) indicates that when examining the CURE plot, a large vertical drop in the plotted cumulative residual is indicative of an inordinately large residual, which could be a possible outlier. A threshold value for identifying when a large vertical drop should be considered an extremely rare case is not identified by Hauer (2015).

Evaluation of Alternative Graphical Techniques

Figure 2 illustrates the two graphical techniques using simulated calibration data (as described in a previous section). The calibration database contains data for 100 intersection sites. The calibration period is 1 year.

The standardized residual plot is shown in Figure 2a. Each data point shown in this figure corresponds to one intersection. The data points tend to lie along one of several bands that extend from the upper left side of the figure to the lower right side. Each band corresponds to an integer crash count. For example, the lowest band in the figure corresponds to sites for which there were no observed crashes during the year.

***a. Standardized residual plot. b. CURE plot.***

Figure . Comparison of graphical techniques using the same data—no extreme cases present.

The intersections represented in Figure 2 have an average crash frequency of 1.0 crash/site/yr and an overdispersion parameter of 0.5. Based on the information in Table 1, the lower and upper boundaries are -1.1 and 4.9, respectively. None of the data points in Figure 2a lie below -1.1 or above 4.9 so it is concluded that no extremely rare cases are present in the calibration database.

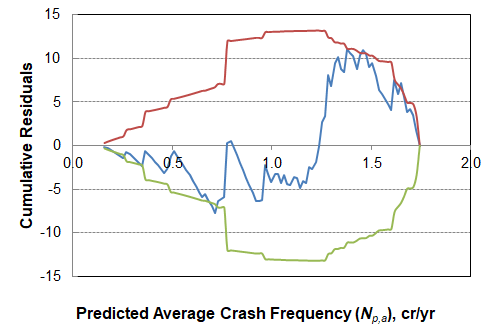
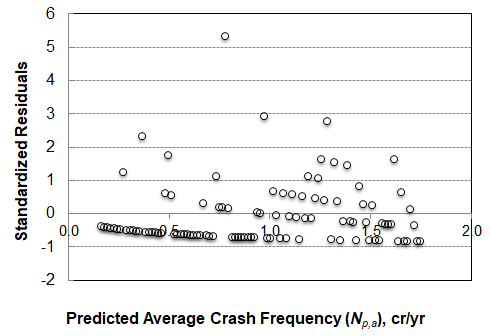
The CURE plot is shown in Figure 2b. The trends shown in this figure are based on the same data as shown in Figure 2a. The cumulative residual trend line is shown using a solid (blue) trend line that oscillates above and below the horizontal line at “0.0”. The solid (red) line above the cumulative residual line defines the 97.5 percentile upper confidence level. The solid (green) line below the cumulative residual line defines the 2.5 percentile lower confidence level. The oscillation shown in the cumulative residual trend line is expected and suggests the calibrated predictive model provides a relatively good fit to the data over the range of predicted values (i.e., no bias-in-fit is evident). Importantly, no large vertical drop is apparent at any point in the trend line, so it is concluded that no extreme cases are present.

Figure 3 also illustrates the two graphical techniques using simulated calibration data. The data have same average crash frequency and over dispersion parameter as used to create Figure 2. In fact, the data represent a second realization of the same sites where the only change is in the random numbers used in the Monte Carlo processes that underlie the simulation.

Application of the lower and upper boundaries (-1.1 and 4.9) in Figure 3a indicates there is one site that is considered extremely rare. It has a predicted average crash frequency of 0.8 crashes/site/yr. More importantly, it has a standardized residual value of 5.2 that is well in excess of the 4.9 threshold value. The data associated with this site should be further examined to determine if the site is an outlier.

The cumulative residual plot in Figure 3b has a relatively large vertical drop at 0.8 crashes/site/yr. This drop is attributable to the same site as identified in Figure 3a. However, there are other vertical drops shown in the figure (albeit smaller than that at 0.8 crashes/site/yr).

Additional realizations of the same set of sites were simulated to further explore these two graphical techniques. Based on an examination of the plots associated with each simulated database, it was concluded that the standardized residual plot provides a clearer indication of when extremely rare cases exist in a calibration database. It is less clear from the CURE plot where the important vertical drops occur and whether they are large enough to represent an extremely rare case.

***a. Standardized residual plot. b. CURE plot.***

Figure . Comparison of graphical techniques using the same data—one extreme case present.

Development of Procedures for Computing the Calibration Factor CV

The equation for computing the calibration factor (i.e., Equation 1) includes in its numerator the number of observed crashes at the sample of sites selected to represent the population of interest. The observed crash count at each site has a large random variation reflecting the many unpredictable factors that influence crash occurrence. As a result, the computed calibration factor represents a sampling statistic that is associated with some uncertainty in terms of it being a reliable indication of the true calibration factor value.

Useful statistics for assessing the reliability of the computed calibration factor are its standard error and its coefficient of variation (CV). The standard error can be used to quantify the confidence interval of the calibration factor (i.e., a range of values within which the true calibration factor value lies with a specified probability).

Bahar and Hauer (2014) have shown mathematically that the standard error of the calibration factor decreases with an increase in the total number of crashes represented in the calibration database. It also increases with an increase in the calibration factor value. As a result, the coefficient of variation (= standard error divided by calibration factor value) is a useful means of describing the calibration factor confidence interval in a consistent manner across different calibration databases.

Bahar and Hauer (2014) examined the predictive accuracy of several crash prediction models (CPMs) and their associated calibration factors. They argued that it would make “little sense” to estimate the calibration factor more accurately than the accuracy of the associated CPM. Based on this argument and their experience with HSM CPMs, they recommended that analysts should “…aim for a CV in the 0.10 to 0.15 range.”

This section consists of two subsections. The first subsection describes two procedures for estimating the standard error and the coefficient of variation of the calibration factor. The second subsection describes the findings from an evaluation of these two procedures.

Review of Procedures for Computing the Calibration Factor CV

NB-Distribution-Based Procedure

Bahar and Hauer (2014) derived equations for estimating the standard error and the coefficient of variation for the calibration factor. The standard error calculation is based on an assumed negative-binomial distribution of the observed crash data. The two statistics of interest are computed using the following equations:

Equation

and,

Equation

where:

*cv,c =* calibration factor coefficient of variation;

*sc =* standard error of the calibration factor;

and all other variables are previously defined. The calibration factor used in Equation 14 is computed using Equation 1.

Bahar and Hauer (2014) do not indicate how to acquire the overdispersion parameter *ki* used in Equation 15. The example they provided uses the parameter that is associated with the CPM being calibrated (i.e., the overdispersion parameter computed using the estimation database upon which the CPM was originally developed). In contrast, Lyon et al. (2018) recommend using the overdispersion parameter that is computed using the data in the calibration database. Hence, Bahar and Hauer (2014) imply that the parameter associated with the estimation database is preferred to the parameter associated with the calibration database while Lyon et al. clearly state that the parameter associated with the calibration database is preferred.

The overdispersion parameter associated with the estimation database has the desirable attribute of being based on a relatively large database and thus, it is a very reliable estimate of CPM fit to the sites in the regions represented in the estimation database. On the other hand, the parameter associated with the calibration database has the desirable attribute of describing the fit of the CPM in the region for which it is being calibrated. This latter attribute is more relevant for assessing the fit of a calibrated CPM and is likely the motivation for the recommendation by Lyon et al. (2018).

The evaluation of this procedure is described in a subsequent section. It is based on the approach recommended by Lyon et al. (i.e., the overdispersion parameter *ki* is computed using the data in the calibration database). For this subsequent evaluation, the inverse dispersion parameter is first computed using Equation 11. Then, the overdispersion parameter is computed with Equation 7.

Distribution-Independent Procedure

Rajabi (2017) also derived an equation for computing the standard error of the computed calibration factor. The standard error calculation is not based on an assumed distribution of the observed crash data. Rather, it is based on the standard deviation of the residual error. The standard error is computed using the following equation:

Equation

with

Equation

Equation

where *sop* is the standard deviation of the residual number of crashes (i.e., difference between the observed crash count and the predicted average crash frequency) during the calibration period, (crashes/site/period); is the average number of observed crashes per site during the calibration period, (crashes/site/period); and all other variables are previously defined. The calibration factor used in Equation 17 is computed using Equation 1. The coefficient of variation is computed using Equation 14.

Evaluation of Alternative Procedures

This section describes the findings from an evaluation of the two procedures described in the previous section. The evaluation is based on the use of simulated calibration data (the process for generating these data by simulation is described in a previous section titled The Empirical Setting).

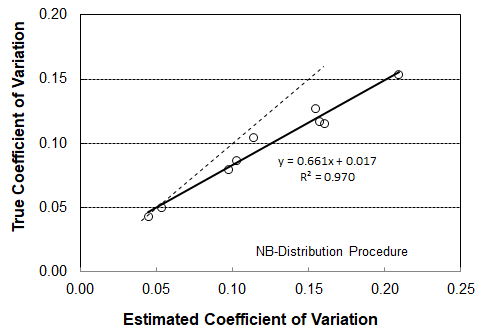
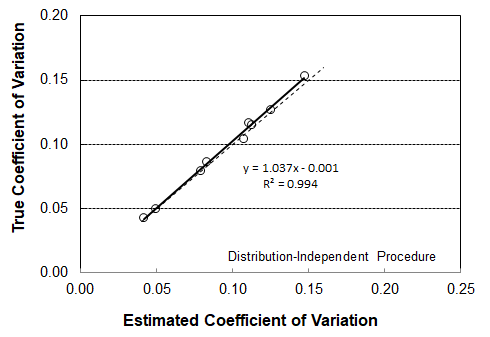
A series of calibration databases were computed for this evaluation. Each calibration database contains data for 100 intersection sites. The calibration period is 1 year. For a given average crash frequency and inverse dispersion parameter, 200 calibration databases were developed (each database containing data for 100 sites). A calibration factor was computed for each of the 200 databases. The average of these calibration factors was confirmed to equal 1.0 ± 0.01 (i.e., *C* = 1.0). The standard deviation of the calibration factors *sc* was also computed. Equation 14 was then used to compute the ground-truth estimate of the calibration factor coefficient of variation *cv,c* (= *sc*/1.0).

The two procedures described in the previous section were also used to compute, for each of the 200 databases, an estimate of the calibration factor coefficient of variation (i.e., one estimate *cv,c,15* was based on Equation 15 and one estimate *cv,c,16* based on Equation 16). The *cv,c,15* for each of the 200 databases was averaged. The *cv,c,16* for each of the 200 databases was also averaged. These two averages were each judged to be the best estimates of the calibration factor coefficient of variation for the respective procedures, *cv,15* and *cv,16.*

The process described in the preceding two paragraphs was repeated for a range of average crash frequencies and inverse dispersion parameters. Specifically, the average crash frequencies considered included 1, 10, and 40 crashes/site/yr and the inverse dispersion parameter values included 1, 2, and 8. In combination, there were nine (= 3 averages × 3 parameters) unique combinations of average crash frequency and inverse dispersion parameter considered. For each of the nine combinations, 200 calibration databases were generated with each database including 100 sites.

The findings from the evaluation of the two procedures are shown in Figure 4. Each data point shown corresponds to one of the nine combinations of average crash frequency and inverse dispersion parameter. The dashed trend line represents an “*x* = *y*” line such that a data point would lie on this line if the estimated coefficient of variation equaled the true coefficient of variation. The solid trend line represents the best-fit trend line to the data points (using least-squares regression).

The trend in the data in Figure 4a indicates that the NB-distribution procedure tends to overestimate the true coefficient of variation between 0 and 35 percent, depending on the estimated coefficient value. In contrast, the trend in the data in Figure 4b indicates that the distribution-independent procedure provides an unbiased estimate of the true coefficient of variation.



***a. NB-distribution procedure. b. Distribution-independent procedure.***

Figure . Comparison of two procedures for computing the calibration factor CV.

Development of Guidelines for Using a CURE Plot to Assess Model Fit

As noted previously, the CURE plot technique was developed by Hauer and Bamfo (1997) for assessing how well the specified regression model form fits the data used to estimate the model coefficients. Example CURE plots were previously shown in Figure 2 and Figure 3.

Hauer (2015) indicates that a model that fits the data reasonably well produces a cumulative residual trend line that “…meanders around the horizontal axis in a manner consistent with a symmetric random walk.” This type of meandering is illustrated in Figure 2b (using a blue trend line).

In contrast, a poor fit can be indicated when significant portions of the cumulative residual trend line increase and remain above the horizontal “0.0” line or decrease and remain below the horizontal “0.0” trend line. If these extended portions are well above or below the “0.0” line, it could be a result of model bias-in-fit (i.e., a systematic discrepancy between the observed and predicted values [Hauer, 2015]). This condition is suggested in Figure 3b for the range of residuals associated with a predicted average crash frequency of about 1.3 to 1.7 crashes/yr. Of course, for this figure, a claim of bias-in-fit would be incorrect given that the simulation process used to generate the data shown does not introduce bias in the model prediction.

This section consists of two subsections. The first subsection summarizes a review of the guidelines for computing and interpreting the CURE plot. The second subsection describes the findings from a review of the CURE plot guidelines and identifies some recommended revisions to this guidance.

Review of Guidelines for Using the CURE Plot

Prior to computing the cumulative residuals, the data in the calibration database are used to compute the calibration factor and the predicted average number of crashes adjusted by the calibration factor. Next, the data are sorted by the adjusted predicted average number of crashes. Then, the *m*th cumulative residual is computed from the residual error for all sites from 1 to *m*. This calculation is shown in the following equation:

Equation

where *rc,m* is the *m*th cumulative residual (where 1 ≤ *m* ≤ *nsites*) and all other variables are previously defined.

Hauer and Bamfo (1997) also derived equations for computing the upper and lower confidence limits for the cumulative residual trend line. These limits are based on an estimate of the standard deviation of the cumulative residual. These limits and the associated standard deviation are computed using the following equations:

Equation

Equation

with,

Equation

where:

*rc,m,a/2* *=* lower boundary for cumulative residual trend line based on cumulative probability of α/2;

*rc,m,1-a/2* *=* upper boundary for cumulative residual trend line based on cumulative probability of 1-α/2;

*zp* = number of standard deviations associated with a cumulative probability *p*;

*sr,m =* standard deviation of the *m*th cumulative residual (where 1 ≤ *m* ≤ *nsites*); and

all other variables are previously defined.

Hauer (2015) asserts that the distribution of cumulative residuals is approximately normal with a mean of 0.0 and a standard deviation of *sr,m*. He notes that the central limit theorem asserts that the cumulative residuals’ representation by the normal distribution improves with an increase in the number of observations in the database.

Hauer (2015) recommends establishing the upper and lower limits at *zp* equal to 2.0 standard deviations from the horizontal “0.0” axis. For normally distributed data, this *zp* value of corresponds to an α value of about 0.05 and boundary limits that define the 95th percentile confidence interval. In this manner, the functional form of a CPM is claimed to be not having “bias-in-fit over the range of the predicted crash frequency” if its cumulative residual trend line rarely goes beyond the upper and lower limits associated with a 95th percentile confidence interval (Hauer, 2015).

Lyon et al. (2018) indicate that the percentage of cumulative residuals that exceed the upper or lower boundaries of the 95th percentile confidence interval can provide a useful indication of the calibrated CPM’s fit to the calibration data. They advise that a value of 5 percent or more is an indication of a potentially poor fit of the CPM to the data.

Evaluation of CURE Plot Guidelines

The guidelines described in the previous section for computing and interpreting the CURE plot were evaluated using simulated crash data. The objective was to determine the validity of the guidance that 5 percent or more cumulative residuals outside of the upper or lower boundary are an indication of a potentially poor fit. The process for generating the simulation data is described in a previous section titled The Empirical Setting.

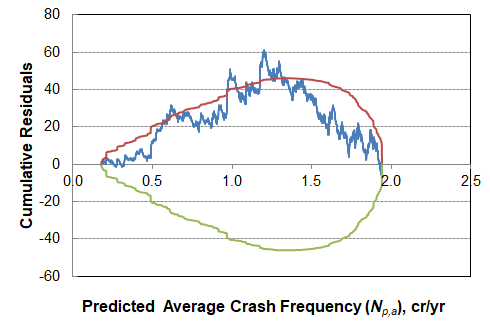
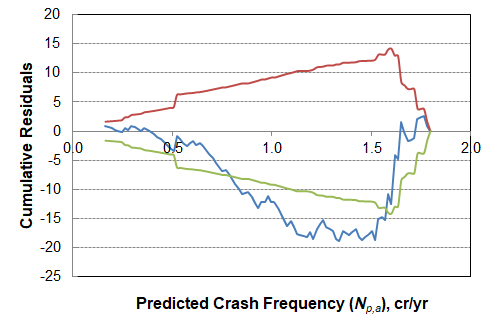
A series of calibration databases were computed for this evaluation. Each calibration database contained data for 100 intersection sites. The calibration period is 1 year. For a given average crash frequency and inverse dispersion parameter, 200 calibration databases were developed (each database containing data for 100 sites). A calibration factor was computed for each of the 200 databases. This factor was used to compute the adjusted predicted average crash frequency for each site. This prediction was then combined with the observed crash count to compute the residual error for each site, the cumulative residual, and the upper and lower boundary limits. These limits were based on a value of *zp* equal to ±2.0 standard deviations.

The process described in the preceding paragraph was repeated for a range of average crash frequencies and inverse dispersion parameters. Specifically, the average crash frequencies considered included 1, 10, and 40 crashes/site/yr and the inverse dispersion parameter values included 1, 2, and 8. In combination, there were nine (= 3 averages × 3 parameters) unique combinations of average crash frequency and inverse dispersion parameter considered. For each of the nine combinations, 200 calibration databases were generated with each database including 100 sites.

As noted in the section titled, The Empirical Setting, an SPF was used to compute the observed crash frequency for each site. The same SPF was effectively used to compute the predicted average crash frequency for each site. As a result, the process did not introduce any bias in the functional form of the CPM so the expectation was that the computed cumulative residual trend line associated with each calibration database would have 5 percent or less of the cumulative residuals outside of the upper or lower boundaries.

During the examination of the CURE plots generated by the simulation tool, it was noted that the percentage of cumulative residuals outside the upper or lower boundary varied widely from database to database. The proportion of databases with 5 percent or less of the cumulative residuals outside the boundaries ranged from 50 to 70 percent, depending on the average crash frequency and inverse dispersion parameter. As noted in the previous paragraph, the expectation was that all the databases would have 5 percent or less of the cumulative residuals outside the boundaries.

Figure 5b illustrates an example CURE plot for one of the databases where 48 percent of cumulative residuals are outside the lower boundary. Application of the guidelines described in the previous section leads to the conclusion that this CURE plot is indicative of a poor fit of the CPM to the calibration data—that the model form needs to be adjusted to provide unbiased predictions. However, this conclusion is known to be incorrect in this case because the process used to simulate the crash data ensured that there is no bias-in-fit introduced by the CPM.



***a. 1000 observations in calibration database. b. 100 observations in calibration database.***

Figure . Example databases with large percentage of cumulative residuals outside of boundaries.

These findings were further explored to determine whether they were influenced by the number observations in the calibration database. A few sets of 200 databases were created for this examination, where each database included 1000 sites. An examination of these data indicated that the proportion of databases with 5 percent or less of the cumulative residuals outside the boundaries increased to about 60 to 80 percent—still far from the expected near 100 percent value. Figure 5a was created during this evaluation to illustrate one of the cases where 15 percent of cumulative residuals are outside the upper boundary. Again, the guidance offered in the previous section would suggest that the CPM is poorly fitting the calibration data and that the model form needs to be adjusted. However, this conclusion is known to be incorrect in this case because of the process used to simulate the crash data.

The distribution of databases with a specified percentage of cumulative residuals outside the boundaries is shown in Figure 6. It is notable that 6.5 percent of the databases had more than 25 percent of the cumulative residuals outside of the boundaries. This particular figure corresponds to the set of 200 databases with an average crash frequency of 40 crashes/site/yr and an inverse dispersion parameter of 8. However, very similar distributions were found for all other combinations of average crash frequency and inverse dispersion parameter.

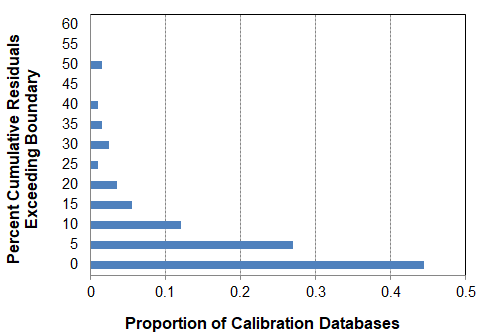


Figure . Distribution of databases by percent cumulative residuals exceeding a boundary.

The distribution of databases for each of the nine combinations of average crash frequency and inverse dispersion parameter was further examined to determine if there was some stability in the percent of cumulative residuals exceeding a boundary. This examination computed the “percent cumulative residuals exceeding” for the 95th percentile database. This computation was undertaken for each of the nine combinations. The results of this examination are listed in column 3 of Table 2. The percentages listed in this column range from 27.7 to 38.7 percent and average 31.5 percent.

As noted in the previous paragraph, the expectation was that all the databases would have 5 percent or less of the cumulative residuals outside the boundaries. The finding that this percentage is nearer to 31.5 percent was unexpected. Its cause is not totally clear. However, an examination of many simulated CURE plots where multiple cumulative residuals exceed a boundary provided some additional insight.

In general, the examination of CURE plots where multiple cumulative residuals exceed a boundary revealed that one site associated with a relatively large cumulative residual value typically caused the cumulative residual trend line to first extend beyond a boundary. However, the “percentage of cumulative residuals exceeding the boundary” was primarily based on the behavior of the cumulative residuals that occurred *just after* the one which first pushed the cumulative residual trend line beyond the boundary. These subsequent cumulative residuals would almost always lie beyond the boundary even though they were associated with sites having a relatively small residual error that were of alternating sign. These subsequent residuals are included in the “percentage of cumulative residuals exceeding the boundary” yet they were not the reason that the cumulative residual trend line is extended beyond the boundary, and, in isolation, the residual values do not indicate the presence of any bias.

Table . Percent cumulative residuals exceeding upper or lower boundary.

|  |  |  |  |
| --- | --- | --- | --- |
| True Average Number of Crashes, (cr/site/period) | True Inverse Dispersion Parameter | Average Percent Cumulative Residuals Exceeding Boundary for 95 Percent of Databases | |
| *zi* = ±2.0 | Revised *zi* Values |
| 1 | 1 | 38.7 | 5.9 |
|  | 2 | 30.7 | 5.0 |
|  | 8 | 28.8 | 4.0 |
| 10 | 1 | 33.7 | 5.0 |
|  | 2 | 31.7 | 3.5 |
|  | 8 | 27.7 | 4.0 |
| 40 | 1 | 33.8 | 3.0 |
|  | 2 | 29.8 | 5.0 |
|  | 8 | 28.8 | 5.0 |
|  | Average: | 31.5 | 4.5 |

The behavior described in the preceding paragraph is illustrated in Figure 5a at a predicted average crash frequency that starts at about 0.95 cr/yr and starts again at about 1.2 cr/yr. It is illustrated in Figure 5b starting at about 0.8 cr/yr. At each location, there is one large cumulative residual that “pushes” the trend line over the boundary, then several subsequent cumulative residuals associated with sites having a small residual error that alternate in sign. These small residual errors are not indicative of bias and they are insufficient in magnitude to pull the trend line back inside the boundary. These subsequent cumulative residuals contribute to the “percentage of cumulative residuals exceeding the boundary” yet their associated portion of the cumulative residual trend line looks like a normal “random walk” (albeit on the “wrong” side of the boundary) which, in isolation, does not indicate the presence of any bias.

Future research should investigate whether the “percentage of cumulative residuals exceeding the boundary” is a viable measure of model fit, especially as it relates to assessing the fit of a locally calibrated CPM using a database assembled for the purpose of calibration (which is inherently smaller than a database used for CPM development).

Several observations can be made based on the information in Table 2 and the guidance offered by Hauer (2015) and Lyon et al. (2018) *when using a CURE plot to evaluate the fit of a calibrated model to a calibration database wherein the number of sites and observed crash counts is relatively small* (i.e., the guidance they offer may be sound when applied to the development of a CPM using a large database). One interpretation of the information in Table 2 is as follows: “About 95 percent of all calibration databases will have 31.5 percent or less of cumulative residuals that exceed the boundaries when the underlying model fits the data.” Alternatively, it can be stated that: “The functional form of a CPM is unlikely to be creating bias-in-fit over the range of the predicted crash frequency if no more than 31.5 percent of the cumulative residuals go beyond the upper or lower boundaries. On the other hand, if the percentage of cumulative residuals that exceed the upper or lower boundaries exceed 31.5 percent, it is an indication of a potentially poor fit of the CPM to the data.”

Further examination of the trends in column 3 of Table 2 indicated that the percentages had a systematic variation with the inverse dispersion parameter value *K*. Additional simulation runs were undertaken to identify the *z* value that would support the statement, “When *z* equals a value of ±*X*, then about 95 percent of all calibration databases will have 5 percent or less of cumulative residuals that exceed the boundaries when the underlying model fits the data.”

The percentage value in column 3 of Table 2 and the inverse dispersion parameter in column 2 was rationalized to have the relationship: *z* = *a* + *b*/*K*. The constants *a* and *b* were initially established at 2.5 and 0.5 based on a preliminary examination of the simulation data. These two constants were used to compute *z*. Then, 200 databases (of 100 sites each) were simulated for each of the nine combinations of average crash frequency and dispersion parameter shown in Table 2. The specified value of *z* was used to estimate the percentage of cumulative residuals exceeding a boundary for each database. The constants *a* and *b* then were adjusted such that the new value of *z* would be more likely to yield the target 5 percent. Several iterations of this process led to convergence on the value of *a* and *b* that yield this target value. They are shown in the following equations.

Equation

Equation

where *z95,p* is the number of standard deviations associated with a cumulative probability *p* for 95 percent of all calibration databases; and all other variables are previously defined. The *z* value in Equation 23 and Equation 24 is used in Equation 20 and Equation 21, respectively. These *z* values were used in a series of simulations to yield the percentages listed in column 4 of Table 2. These percentages range from 3.0 to 5.9 percent, with an average of 4.5 percent. This average value rounds to the 5 percent value that was the intended target of the analysis.

References

AASHTO (2010). *Highway Safety Manual*. American Association of State Highway and Transportation Officials. Washington, D.C.

Bahar, G., and E. Hauer. (2014). *User’s Guide to Develop Highway Safety Manual Safety Performance Function Calibration Factors. Final Report*. NCHRP Project HR 20-7(332). Transportation Research Board, Washington, D.C.

Hauer, E. (2015). *The Art of Regression Modeling in Road Safety*. Springer International, New York.

Hauer,E., and J. Bamfo. (1997). “Two Tools for Finding what Function Links the Dependent Variable to the Explanatory Variables.” *Proceedings of the ICTCT 97 Conference*. Lund, Sweden, pp. 1-19.

Lyon, C., B. Persaud, and F. Gross. (2018). *The Calibrator—An SPF Calibration and Assessment Tool, Updated User Guide*. Report No. FHWA-SA-17-016. Federal Highway Administration, Washington, D.C.

McCullagh, P., and J.A. Nelder. (1989). *Generalized Linear Models – 2nd edition.* Chapman and Hall, Great Britain.

Rajabi, M. (2017). *Assessing the Predictability of Highway Safety Manual Crash Prediction Models in the State of South Carolina*. Dissertation 2051. Clemson University, Clemson, South Carolina.