Accident Analysis and Prevention 93 (2016) 160–168

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

# Sample-size guidelines for recalibrating crash prediction models: Recommendations for the highway safety manual

Mohammadali Shirazi[a], Dominique Lord[a,*], Srinivas Reddy Geedipally[b]

[a] Zachry Department of Civil Engineering, Texas A&M University, College Station TX 77843, United States
[b] Texas A&M Transportation Institute, College Station, TX 77843-3135, United States

ARTICLE INFO

*Article history:*
Received 1 January 2016
Received in revised form 8 April 2016
Accepted 9 April 2016
Available online 13 May 2016

*Keywords:*
Crash prediction models
Calibration factor
Sample size
Monte-Carlo simulation

ABSTRACT

The Highway Safety Manual (HSM) prediction models are fitted and validated based on crash data collected from a selected number of states in the United States. Therefore, for a jurisdiction to be able to fully benefit from applying these models, it is necessary to calibrate or recalibrate them to local conditions. The first edition of the HSM recommends calibrating the models using a one-size-fits-all sample-size of 30–50 locations with total of at least 100 crashes per year. However, the HSM recommendation is not fully supported by documented studies. The objectives of this paper are consequently: (1) to examine the required sample size based on the characteristics of the data that will be used for the calibration or recalibration process; and, (2) propose revised guidelines. The objectives were accomplished using simulation runs for different scenarios that characterized the sample mean and variance of the data. The simulation results indicate that as the ratio of the standard deviation to the mean (i.e., coefficient of variation) of the crash data increases, a larger sample-size is warranted to fulfill certain levels of accuracy. Taking this observation into account, sample-size guidelines were prepared based on the coefficient of variation of the crash data that are needed for the calibration process. The guidelines were then successfully applied to the two observed datasets. The proposed guidelines can be used for all facility types and both for segment and intersection prediction models.



## 1. Introduction

Crash prediction models are used to estimate or predict the number of crashes and evaluate roadway safety. Part C of the first edition of the Highway Safety Manual (HSM) (AASHTO, 2010) provides crash prediction models, or what is often referred to as safety performance functions (SPFs), for roadway segments and intersections for four facility types: rural two-lane roads, rural multilane highways, urban and suburban arterials, and more recently freeways and interchanges. All the HSM prediction models were fitted and validated with data collected from a few selected numbers of states. Therefore, since crash frequency and its dispersion vary substantially from one jurisdiction to next, it is essential to calibrate SPFs when they are applied to a new jurisdiction. In other words, calibration is a tool to account for the differences in factors that were not considered or cannot be considered in the SPF development, such as weather, driver behavior or reportability criteria, between jurisdictions into predictive models. Consequently, two options are available to the analyst: (1) developing a jurisdiction-specific model for the facility that is being analyzed or (2) calibrating the HSM pre-fitted models to the jurisdiction conditions. The detailed SPF calibration procedure is presented in appendix A of Part C of HSM. In this procedure, the calibration factor (C-factor) is eventually calculated as the ratio between the total number of observed crashes ($N_{obs}$) and the total number of predicted crashes ($N_{pre}$) (Eq. (1)), and is applied to the facility SPF as a scalar term.

$$C = \frac{\Sigma N_{obs}}{\Sigma N_{pre}} \tag{1}$$

The first version of the HSM recommends a one-size-fits-all sample size for the calibration procedure. It requires crash data collected from 30 to 50 randomly selected sites with a minimum of 100 crashes per year. However, this recommended sample size is not fully supported by documented studies. For sites with low crash history, collecting 100 crashes at 30 or 50 sites could be difficult to perform (Xie et al., 2011). On the other hand, for most facilities, this minimum recommendation may not provide desirable results (Banihashemi, 2012; Alluri et al., 2016). The later issue,

* Corresponding author.
*E-mail addresses:* alishirazi@tamu.edu (M. Shirazi), d-lord@tamu.edu (D. Lord), srinivas-g@tti.tamu.edu (S.R. Geedipally).

http://dx.doi.org/10.1016/j.aap.2016.04.011
0001-4575/

in addition to the fact that no documented study supported the HSM sample-size recommendation, inspired researchers to investigate the quality of the C-factors that are derived based on the HSM recommendation. Sensitivity analyses on C-factors derived from different sample sizes were documented in several studies to assess the HSM one-size-fits-all sample size recommendation. It has been reported that not only the HSM one-size-fits all recommendation is inappropriate, but is also insufficient to obtain the desirable accuracy in most cases.

Despite efforts that have been placed into proposing new sample-size guidelines to recalibrate SPFs, two shortcomings were identified in previous studies. First, in most studies, it was assumed that the C-factor derived from the dataset in hand is the "ideal" (true or unbiased) calibration factor. However, the ideal calibration factor is not known beforehand when empirical data are used. Consequently, the corresponding sample-size guidelines could involve potential biases or errors. For example, any changes to the empirical data or resampling of the data may also change the C-factor that was initially identified as ideal. Furthermore, even the complete dataset used for recalibration purposes may not be large enough (e.g., data collected by a city transportation agency) to obtain a reliable estimation of the ideal C-factor. These issues are overcome in the current study by conducting extensive simulations. Using simulation, a true calibration factor can be determined, and then assess if the proposed sample-size is large enough to achieve a calibration factor that is close to the true value. Second, sample-size recommendations that were proposed in previous studies are based on a specific dataset usually collected at the state level. Therefore, given the fact that the characteristics of different roadways vary substantially, it is likely that these recommendations do not provide desirable results when applied to a new jurisdiction. In order to overcome this issue, the current study documents recommendations that are based on the crash data characteristics (i.e., the coefficient of variation—ratio of the standard deviation to the mean of the crashes) that will be used for calibrating predictive models. Therefore, agencies will be able to select a sample size that represents the characteristics of the crash data for the type of facility analyzed.

The objectives of this paper are consequently (1) to examine the required sample size based on the characteristics of the data that will be used for the calibration or recalibration process; and, (2) propose revised guidelines. The objectives were accomplished using simulation and two observed datasets.

## 2. Background

Following the release of the HSM, several states have attempted to develop state-specific calibration factors for different types of facilities. Oregon (Xie et al., 2011) was one of the pioneering states that developed state-specific calibration factors. In recent years, calibration factors were generated for other states, such as Utah (Brimley et al., 2012), Illinois (Williamson and Zhou, 2012), Alabama (Brown et al., 2014), Missouri (Mehta and Lou, 2013) and Maryland (Shin et al., 2014). Several studies have noted that the recalibration of predictive models is a time-consuming task in addition to problems associated with the collection, readiness and completeness of the data.

As stated above, as a general guideline, the HSM recommends estimating the calibration factors using at least 100 crashes per year collected randomly at 30–50 sites. However, the one-size-fits-all sample size recommendation needs to be reviewed given the fact that different roadway types have different levels of homogeneity and the minimum sample size is a function of the population homogeneity (Alluri et al., 2016). Taking this into account, several researchers have attempted to evaluate the HSM recommendation and, consequently, proposed new guidelines.

Banihashemi (2012) reviewed the HSM sample-size recommendation by performing a sensitivity analysis on C-factors derived from samples with different sizes. The author used a dataset collected in Washington State and performed a sensitivity analysis for three types of facilities: rural two lane roads, rural multilane highways, and urban and suburban arterials. The author first found the calibration factor that was derived from the available dataset and referred to it as the ideal (true) calibration factor. Then, for each selected sample size, 10 samples were generated randomly and their corresponding C-factors calculated. Next, assuming that the estimated measures followed a normal distribution, the quality of each sample size was quantified by measuring the probability that the calibration factor falls within 5% or 10% (depending on the desired accuracy) of the ideal calibration factor. The sample size that ensures the estimated calibration factor falls within 10% of the ideal calibration factor with a reasonable probability was recommended in the new guidelines. The study showed that the HSM 30- to 50-site criterion was too small to derive a reliable C-factor for most roadway types.

Alluri et al. (2016) used data collected in Florida to determine the minimum sample size that results in a reliable calibration factor for the same three types of facilities described above. A similar procedure as the one proposed by Banihashemi (2012) was used to assess the C-factors and estimate the minimum sample size. In the study, for each given sample size, 30 subsets of data were generated and the corresponding C-factors calculated. The analysis showed that not only the HSM generalized sample size guideline is not appropriate, but this criterion was also insufficient to acquire the desired accuracy. The recommendations provided in the paper are based on the criterion that, with a high probability, the calibration factors lie within 10% of the ideal factor. However, for cases where sufficient data are available and a higher accuracy is desired, the recommendations based on 5% of the ideal factor were provided as well. The recommended minimum sample size of reaching the 5% accuracy almost doubles compared to the recommendations of achieving the 10% accuracy.

Trieu et al. (2014) performed a sensitivity analysis on calibration sample size to evaluate and critique the accuracy of the HSM sample-size guideline for undivided two-lane urban arterial roadways. Given different percentages of a complete dataset, the samples were resampled from the complete dataset for 500 iterations.[1] Then, C-factors for each size-group were classified based on their errors from the ideal C-factor in 5% increments. It was noted that as the sample size increased, C-factor observations with high error range decreased. It was also observed that for samples generated from 50% (or more) of the complete dataset, all C-factors fall within 10% of the ideal C-factor. The paper concluded that the current HSM sample-size criterion may not yield a reliable C-factor. The authors then analyzed the Annual Average Daily Traffic (AADT) distribution for a group of C-factors that were generated with a sample size of 37 sites (the sample size that satisfies the HSM criterion). The results showed that the AADT distribution could influence the C-factor reliability.

Since roadway and vehicle characteristics as well as driver behavior continuously changes over time, crash prediction models quickly become outdated. Because fitting a new model requires a lot of data and is a time-consuming and expensive task, it is essential

[1] Note: albeit this paper referred its method to as a Monte Carlo simulation, it seems that samples were obtained directly from the original dataset, and not from a distribution derived from the data. Monte Carlo simulation, however, is referred to as parametric sampling methods which samples are generated from parametric distributions.