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

**Working White Paper: Guidelines on Maintaining Standardized Database and Prioritizing Input Data Elements**

**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.

Introduction

Practitioners are often faced with the decision on what input data elements to prioritize for collection for inclusion in the calibration datasets. There are two issues to consider here:

* Maintaining the calibration database over time include unified procedures for checking the currency of database variables and criteria for adding or dropping sites. Agencies should also be encouraged to retain all calibration databases that are assembled to support future updating activities.
* Prioritizing the collection of input data elements by need, safety influence, and cost to collect, where more data collection effort is expended toward including high priority data and lower priority data may be defaulted.

The rest of this document is structured as follows. The next two sections provide a brief background of these issues and a summary of the approach using random forest technique for prioritizing the collection of input data elements (Saleem *et al.*, 2020 and Alluri *et al.*, 2014). This is followed by a brief overview of the prediction models from the 1st edition of the HSM for rural roads. An overview of the data that were collected is discussed next. This is followed by a summary of the results based on the approach described in Saleem *et al.* (2020) and Alluri *et al.* (2014) and the resulting guidance.

Background

The calibration database needs two main categories of data. One category includes the data elements that serve as input values to the crash prediction models (also known as safety performance functions (SPFs)). The second category is the observed crash data. Data for both categories are needed for each site represented in the calibration database.

* Input Data Elements

The input data elements needed in the calibration database include all site characteristics that are used to apply the SPF of interest. The number of input data elements for a given model is dependent on the number of variables in the SPF and CMFs that comprise the predictive model equation. These data elements typically include traffic characteristics, geometric design elements, and traffic control features. Documentation describing the development of the candidate SPF should be consulted to determine the specific definition of each input variable to ensure that it is counted, measured, or computed in a manner that is consistent with the variable’s use in the SPF selected for calibration. This information is provided in Section 5 of each chapter in Part C of the 1st edition of HSM if the subject SPF is from the HSM.

* Crash Data

The calibration database must include all target crashes that are observed (i.e., reported) at a site of interest during the calibration period. If the database is being used to calibrate two or more SPFs for a common region, facility, and site type, then the database will need to include the target crashes for each SPF of interest.

The crash attributes needed from each crash report include crash location, date and time, intersection-relationship, severity, and crash type. These data are needed for each crash that occurs at a site of interest during the calibration period. The crash location and intersection-relationship data are used to assign crashes to the sites included in the calibration database. The date of the crash is used to verify that the crash occurred during the calibration period.

Each crash must be assigned to the appropriate site. Documentation describing the development of the candidate SPF should be consulted to determine the specific criteria used to assign a crash to a site. This information is provided in Section 5 of each chapter in Part C of the 1st edition of HSM if the subject SPF is from the HSM.

The data assembly process entails the identification, acquisition, reduction, and organization of the data needed for calibration of the SPF of interest. The analyst should conduct a jurisdiction-specific data assessment for the purpose of identifying both existing data and data that may need to be acquired. A work plan for data assembly that is developed through this assessment can be a useful basis for estimating the resources required for the calibration project.

The steps associated with this process are described in the following list.

* Identify the data elements needed for calibrating the CPM of interest.
* Identify the sources of data within the jurisdiction-specific databases for the desired combination of region, facility type, and site type.
* Identify the missing data elements and establish a process to acquire them.

A complete data assembly process should only need to be performed the first time that calibration is performed for a given region, facility type, and site type. For model updating in subsequent years, the same sites may be used again with the traffic volume and crash data updated to reflect the new calibration period. Data describing the geometric design elements and traffic control devices will only need to be updated for those sites that have had a corresponding change in design or devices in the intervening years.

Random Forest Technique

Tree based models including classification and regression trees (CART) and random forests have been shown to hold strong potential for road safety analyses (Persaud, 2001). Tree-based methods are particularly effective in making predictions of expected crash frequency and have the potential to inform specifications that are part of more traditional modeling approaches through identifying the “most predictive” right-hand-side variables and uncovering informative relationships between left-hand-side and right-hand-side variables (Saleem *et al.*, 2020).

Tree-based methods are a set of machine-learning and data-mining procedures that use the form of a binary tree and act as predictive models that map values of a dependent variable or response variable as a function of key explanatory variables. There are two types of tree analyses: a classification tree where the dependent variable is categorical and a regression tree where the dependent variable is continuous. The output of these analyses is a tree that shows the most predictive variable at the top that branches off into combinations of variables that best predict the outcome variable (Saleem *et al*., 2020).

Breiman and Cutler (2013) developed the random-forest algorithm, which works within the framework of CART. With random forests, instead of having one tree, multiple trees are produced using a resampling method, and the aggregate results are then combined. Breiman and Cutler believed that a single decision tree may not reveal all variables that contribute to the dependent/target variable and that the contributions of some predictive independent variables can be masked by other independent variables. Random forests can help identify predictors that may not appear in the output of a single classification or regression tree but, nevertheless, are highly related to the target variable.

The percentage increase in mean squared error (MSE) with the removal of a variable from the random-forest model are commonly displayed using random forests (Saleem *et al*., 2020).

Saleem *et al.* (2020) presented case studies using the random forest technique to identify contributing factor for various focus crash and facility types. Similarly, Alluri *et al.* (2014) presented case studies using the random forest technique to identify and rank important variables for inclusion in calibration datasets.

Overview of Rural Segment Prediction Models in 1st Edition of the HSM

The following sections provide an overview of the prediction models from the 1st edition of the HSM for rural two-lane roads and rural multi-lane divided roads.

Rural Two-Lane Roads

For rural two-lane, two-way undivided roads, the predictive model from the 1st edition of the HSM is as follows:

Where, is the predicted average crash frequency for an individual roadway segment for a specific year, is the predicted average crash frequency for base conditions for an individual roadway segment, is the calibration factor for roadway segments of a specific type developed for a particular jurisdiction or geographical area, and to are crash modification factors (also called adjustment factors).

The base condition SPF for predicted average crash frequency for rural two-lane, two-way roadway segments is as follows:

Where, is the average annual daily traffic volume, and is the length of the roadway segment in miles.

The overdispersion parameter () is estimated as a function of the length of the roadway segment:

The base conditions for the SPF are as follows (note that a CMF is available for each of these 12 base conditions, and they can be modified based on actual conditions):

* Lane width of 12 feet
* Shoulder width of 6 feet
* Paved shoulder
* Roadside hazard rating of 3
* Driveway density of 5 driveways per mile
* No horizontal curvature
* No centerline rumble strips
* No passing lanes
* No two-way left turn lanes
* No lighting
* No automated speed enforcement
* Grade of 0%

Rural Four-Lane Divided Roads

For rural four-lane, divided roads, the predictive model from the 1st edition of the HSM is as follows:

Where, is the predicted average crash frequency for an individual roadway segment for a specific year, is the predicted average crash frequency for base conditions for an individual roadway segment, is the calibration factor for roadway segments of a specific type developed for a particular jurisdiction or geographical area, and to are crash modification factors (also called adjustment factors).

The base condition SPF for predicted average crash frequency for rural two-lane, two-way roadway segments is as follows:

Where, is the average annual daily traffic volume, is the length of the roadway segment in miles, and and are regression coefficients.

The overdispersion parameter () is estimated as a function of the length of the roadway segment:

Where, is the regression coefficient used to determine the overdispersion parameter.

For total crashes, the regression coefficients are as follows:

The base conditions for the SPF are as follows (note that a CMF is available for each of these 5 base conditions, and they can be modified based on actual conditions):

* Lane width of 12 feet
* Right shoulder width of 8 feet
* Median width of 30feet
* No lighting
* No automated speed enforcement

Summary of Data

Recent data from North Carolina (2016 – 2019) were compiled as part of two recent projects (NCHRP Project 17-72: Update of Crash Modification Factors for the Highway Safety Manual, and NCDOT Project 2020-27: Updated and Regional Calibration Factors for Highway Safety Manual Prediction Models 2016-2019).

Table 1 shows the sum of mileage and the average AADT (by roadway type). Tables 2 and 3 show the AADT ranges and the total observed crashes (by roadway type). Note that in the Tables, rural two-lane roads are denoted by Rural 2U, and rural four-lane divided roads are denoted by Rural 4D.

Note that for the purpose of this initial guidance, the random forest technique was used to identify and rank important variables for rural road segments only. Data is available for urban segments and rural/urban intersections in North Carolina from the above-mentioned recent efforts. We plan to use this data when refining the draft guidelines presented here in Phase 2.

Table . Sum of Mileage (by Roadway Type)

|  |  |  |  |
| --- | --- | --- | --- |
| **Roadway Type** | **State** | **Sum of Mileage** | **Average AADT** |
| Rural 2U | North Carolina | 732.74 | 2030.55 |
| Rural 4D | North Carolina | 197.27 | 14395.03 |

*Table 2. Summary of Yearly Volume Data (by Roadway Type)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Roadway Type** | **AADT** | **2016** | **2017** | **2018** | **2019** |
| Rural 2U | Minimum | 60 | 70 | 65 | 50 |
| Rural 2U | Maximum | 35000 | 42925 | 50850 | 58775 |
| Rural 2U | Average | 1965 | 2021 | 2052 | 2083 |
| Rural 4D | Minimum | 2800 | 2500 | 2700 | 2700 |
| Rural 4D | Maximum | 34000 | 32000 | 34000 | 38000 |
| Rural 4D | Average | 13842 | 14100 | 14579 | 15059 |

*Table 3. Total Yearly Observed Crashes (by Roadway Type)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Roadway Type** | **Crash Type** | **2016** | **2017** | **2018** | **2019** |
| Rural 2U | Total | 715 | 735 | 717 | 756 |
| Rural 4D | Total | 646 | 745 | 762 | 758 |

Results

Maintaining Standardized Database

The first step towards maintaining a standardized database is to identify the data elements needed for calibrating the SPF of interest. This can be accomplished by reviewing the SPF of interest and identifying the data elements needed to apply the base condition models and the relevant CMFs. Tables 4 and 5 list the data elements needed for calibrating the SPFs for Rural 2U and Rural 4D. These tables also list the value types for each data element of interest as well as the units / string to be collected.

Table . Data Elements Needed for Calibrating Rural 2U SPFs

|  |
| --- |
| **Rural Two-Lane Roads** |
| **Data Element** | **Value Type** | **Units / String** |
| Geographic Location | Coordinates | Latitude & Longitude |
| Segment Length | Number | miles |
| Traffic Volume | Number | AADT |
| Lane Width | Number | ft |
| Shoulder Width | Number | ft |
| Shoulder Type | *String* | *Composite / Gravel / Paved / Turf* |
| Lengths of Horizontal Curves and Tangents | Number | ft |
| Radii of Horizontal Curves | Number | ft |
| Presence of Spiral Transition | *String* | *Present / Not Present* |
| Superelevation Variance for Horizontal Curves | Number | ft/ft |
| Percent Grade | *String* | *<= 3% / 3% to 6% / >6%* |
| Driveway Density | Number | driveways/mile |
| Presence of Centerline Rumble Strips | *String* | *Present / Not Present* |
| Presence of Passing Lanes | *String* | *Present / Not Present* |
| Presence of TWLTL  | *String* | *Present / Not Present* |
| Presence of Lighting | *String* | *Present / Not Present* |
| Roadside Hazard Rating | *String* | *Scale 1 - 7* |
| Presence of Automated Speed Enforcement | *String* | *Present / Not Present* |
| Number of Through Lanes | Number | Number of Lanes |
| Roadway Division Type | *String* | *Divided / Undivided* |

Table . Data Elements Needed for Calibrating Rural 4D SPFs

|  |
| --- |
| **Rural Four-Lane Divided Roads** |
| **Data Element** | **Value Type** | **Units / String** |
| Geographic Location | Coordinates | Latitude & Longitude |
| Segment Length | Number | miles |
| Traffic Volume | Number | AADT |
| Lane Width | Number | ft |
| Median Width | Number | ft |
| Shoulder Width | Number | ft |
| Shoulder Type | *String* | *Composite / Gravel / Paved / Turf* |
| Presence of Lighting | *String* | *Present / Not Present* |
| Presence of Automated Speed Enforcement | *String* | *Present / Not Present* |
| Roadway Division Type | *String* | *Divided / Undivided* |

A complete data assembly process should only need to be performed the first time that calibration is performed for a given region, facility type, and site type. Changes in design practice, driver behavior, and vehicle crashworthiness over time require the periodic updating of SPFs. As such, the calibration databases should be maintained over time with updated traffic volume and crash data as well as recording changes to data describing the geometric design elements and traffic control devices.

Prioritizing Input Data Elements

This section summarizes the results based on the approach described by Saleem *et al.* (2020) and Alluri *et al.* (2014). Figures 1 – 2 and Tables 6 – 7 present the random forest developed for total crashes on Rural 2U and Rural 4D using data from North Carolina.



Figure .Rural 2U Total Crash Random Forest

Table . Rural 2U Percentage Increase in Mean Squared Error

|  |  |
| --- | --- |
| **Data Element** | **% Increase in MSE** |
| Segment Length | 125.99 |
| AADT | 90.52 |
| Driveway Density | 34.95 |
| Lane Width | 31.36 |
| Percent Grade | 30.08 |
| Shoulder Width | 27.50 |
| Shoulder Type | 17.80 |
| Presence of CLRS | 14.17 |
| Curvature | 8.28 |
| Presence of Lighting | 2.48 |
| Presence of TWLTL | 1.92 |
| Presence of Passing Lane | 0 |
| Roadside Hazard Rating | 0 |
| Presence of ASE | 0 |

*\*CLRS = Centerline Rumble Strips, TWLTL = Two-Way Left Turn Lane, ASE = Automated Speed Enforcement*



Figure . Rural 4D Total Crash Random Forest

Table . Rural 4D Percentage Increase in Mean Squared Error

|  |  |
| --- | --- |
| **Data Element** | **% Increase in MSE** |
| Segment Length | 75.25 |
| AADT | 53.93 |
| Median Width | 24.78 |
| Shoulder Type | 22.28 |
| Shoulder Width | 20.99 |
| Presence of Lighting | 0.43 |
| Lane Width | 0 |
| Presence of ASE | 0 |

*\* ASE = Automated Speed Enforcement*

The vertical axis in the random forest plots displays variables, and the horizontal axis shows percent increase in MSE (%IncMSE). Higher values of %IncMSE imply that a variable is a stronger predictor of crash frequency. Typically, the strongest predictors are shown at the top of the plot. For example, Figure 1 illustrates that segment length is the strongest predictor of total crashes on Rural 2U roads in North Carolina. Taking segment length out of the analysis can potentially increase the MSE of predictions by approximately 126 percent.

It should also be noted that some variables have a %IncMSE of 0. This means that though the data for the variable has been collected, there was no variations ion the collected data. For e.g., all Rural 2U and Rural 4D segments for which data were collected in North Carolina had no automated speed enforcement.

Discussion and Recommended Guidance

Based on the results from Figures 1 – 2 and Tables 4 – 7, the research team recommend the following guidance to identify input data elements as those of primary, secondary, and lesser importance (similar to the bins used by Alluri *et al*. 2014).

* Primary importance – %IncMSE of greater than 50%.
* Secondary importance – %IncMSE of between 20% and 50%.
* Lesser importance – %IncMSE of less than 20%.

Tables 8 – 9 lists the data elements for Rural 2U and Rural 4D roads in North Carolina in order of their importance.

 Table . Ranking of Input Data Elements – Rural 2U

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Element** | **Primary Importance** | **Secondary Importance** | **Lesser Importance** | **Cannot be Classified** |
| Segment Length | X |  |  |  |
| AADT | X |  |  |  |
| Driveway Density |  | X |  |  |
| Lane Width |  | X |  |  |
| Percent Grade |  | X |  |  |
| Shoulder Width |  | X |  |  |
| Shoulder Type |  |  | X |  |
| Presence of CLRS |  |  | X |  |
| Curvature |  |  | X |  |
| Presence of Lighting |  |  | X |  |
| Presence of TWLTL |  |  | X |  |
| Presence of Passing Lane |  |  |  | X |
| Roadside Hazard Rating |  |  |  | X |
| Presence of ASE |  |  |  | X |

*\*CLRS = Centerline Rumble Strips, TWLTL = Two-Way Left Turn Lane, ASE = Automated Speed Enforcement*

Table . Ranking of Input Data Elements – Rural 4D

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Element** | **Primary Importance** | **Secondary Importance** | **Lesser Importance** | **Cannot be Classified** |
| Segment Length | X |  |  |  |
| AADT | X |  |  |  |
| Median Width |  | X |  |  |
| Shoulder Type |  | X |  |  |
| Shoulder Width |  | X |  |  |
| Presence of Lighting |  |  | X |  |
| Lane Width |  |  |  | X |
| Presence of ASE |  |  |  | X |

*\*ASE = Automated Speed Enforcement*

Alluri *et al*. (2014) developed random forests using Florida data to rank the data elements by importance. Tables 10 – 11 list the data elements for Rural 2U and Rural 4D roads in Florida in order of their importance as identified by Alluri *et al*. (2014).

Table . Ranking of Input Data Elements – Florida Rural 2U by Alluri et al. (2014)

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Element** | **Primary Importance** | **Secondary Importance** | **Lesser Importance** |
| Segment Length | X |  |  |
| AADT | X |  |  |
| Driveway Density |  | X |  |
| Shoulder Width |  | X |  |
| Roadside Hazard Rating |  | X |  |
| Lane Width |  | X |  |
| Shoulder Type |  | X |  |
| Presence of Lighting |  |  | X |
| Presence of Passing Lane |  |  | X |
| Presence of TWLTL |  |  | X |
| Presence of Short Four-Lane Sections |  |  | X |
| Presence of CLRS |  |  | X |
| Presence of ASE |  |  | X |

*\*CLRS = Centerline Rumble Strips, TWLTL = Two-Way Left Turn Lane, ASE = Automated Speed Enforcement*

Table . Ranking of Input Data Elements – Florida Rural 4D by Alluri et al. (2014)

|  |  |  |
| --- | --- | --- |
| **Data Element** | **Primary Importance** | **Secondary Importance** |
| Segment Length | X |  |
| AADT | X |  |
| Shoulder Width |  | X |
| Median Width |  | X |
| Presence of ASE |  | X |
| Presence of Lighting |  | X |
| Lane Width |  | X |

*\*ASE = Automated Speed Enforcement*

Data elements of primary importance will be those that are needed to apply base condition HSM SPFs, whereas data elements of secondary or lesser importance are those needed to properly apply the CMFs to the actual conditions.

Based on the availability of funding, agencies should prioritize collecting data for data elements of primary and secondary importance, with an option of using default base conditions for data elements of lesser importance. However, if the calibration dataset is being collected for the first time, it would be recommended that a complete data assembly is undertaken to collect all required data elements in a standardized form. Agencies can then, in subsequent years, prioritize the updating of elements based on availability of funding and the importance bins.

It should be noted that the data elements of primary importance are the same for both North Carolina and Florida, however, data elements of secondary and lesser importance vary. This is expected due to changes in local conditions across states. To overcome this issue, in Phase 2, the research team will try to obtain ready to use data from other states to help further refine and develop unified national guidance on prioritizing input data elements for collection.

Furthermore, in Phase 2, the research team will also refine the guidelines for maintaining the calibration database over time (including possibly maintaining a national repository/clearinghouse of calibration datasets from all states) and discuss possible procedures for checking the currency of database variables and criteria for adding or dropping sites.

References

Alluri, P., Saha, D., Liu, K., and Gan, A. (2014). *Improved Processes for Meeting the Data Requirements for Implementing the Highway Safety Manual (HSM) and SafetyAnalyst in Florida*. Florida Department of Transportation.

Breiman, L., and Cutler, A. (2013). “Random Forests” (website). Accessed January 19, 2018. <http://www.stat.berkeley.edu/~breiman/RandomForests>.

HSM (2010). “Highway Safety Manual.” American Association of State Highway and Transportation Officials. Washington, D.C.

Persaud, B. (2001). *NCHRP Synthesis 295: Statistical Methods in Highway Safety Analysis: A Synthesis of Highway Practice*. National Cooperative Highway Safety Research Program. Transportation Research Board.

Saleem, T., Porter, R.J., Srinivasan, R., Carter, D., Himes, S., and Le, T. (2020). *Contributing Factors for Focus Crash and Facility Types*. Report FHWA-HRT-20-052, Federal Highway Administration – U.S. Department of Transportation.