Develop Agency SPF

From SafetyAnalystWiki

Contents

[hide]

Safety Performance Functions

This document provides guidance on the development of safety performance functions (SPFs) for use with the *SafetyAnalyst* software. SPFs are provided in *SafetyAnalyst* and automatically calibrated using each agency's data, so it is not necessary for agencies to develop their own SPFs. However, since some agencies may prefer to implement *SafetyAnalyst* using SPFs developed with their own agency's data; this memo provides guidance on the appropriate procedures for SPF development. SPFs are regression relationships between target accident frequencies and traffic volumes that can be used to predict the expected long-term accident frequency for a site. SPFs are used in the EB methodologies that estimate the safety performance of sites in several of the analytical tools provided with *SafetyAnalyst*. Thus, SPFs are essential to the functioning of those tools.

What SPFs Are Needed

Within *SafetyAnalyst*, SPFs are needed for three types of sites (roadway segments, intersections, and ramps) and for several subtypes of those site types. Site type and subtypes are based upon roadway characteristics and other site characteristics. The site types and subtypes for which SPFs are needed are as follows:

Roadway Segments

- Rural two-lane highway segments
- Rural multilane undivided highway segments
- Rural multilane divided highway segments
- Rural freeway segments—4 lanes
- Rural freeway segments—6+ lanes
- Rural freeway segments within an interchange area—4 lanes
- Rural freeway segments within an interchange area—6+ lanes
- Urban two-lane arterial segments
- Urban multilane undivided arterial segments
- Urban multilane divided arterial segments
- Urban one-way arterial segments
- Urban freeway segments—4 lanes
- Urban freeway segments—6 lanes

- Urban freeway segments—8+ lanes
- Urban freeway segments within an interchange area—4 lanes
- Urban freeway segments within an interchange area—6 lanes
- Urban freeway segments within an interchange area—8+ lanes

Intersections

- Rural three-leg intersections with minor-road STOP control
- Rural three-leg intersections with signal control
- Rural four-leg intersections with minor-road STOP control
- Rural four-leg intersections with all-way STOP control
- Rural four-leg intersections with signal control
- Urban three-leg intersections with minor-road STOP control
- Urban three-leg intersections with signal control
- Urban four-leg intersections with minor-road STOP control
- Urban four-leg intersections with all-way STOP control
- Urban four-leg intersections with signal control

Ramps

- Rural diamond off-ramps
- Rural diamond on-ramps
- Rural parclo loop off-ramps
- Rural parclo loop on-ramps
- Rural free-flow loop off-ramps
- Rural free-flow loop on-ramps
- Rural direct or semidirect connection ramps
- Urban diamond off-ramps
- Urban diamond on-ramps
- Urban parclo loop off-ramps
- Urban parclo loop on-ramps
- Urban free-flow loop off-ramps
- Urban free-flow loop on-ramps
- Urban direct or semidirect connection ramps

For each site subtype, SPFs are needed both for total (TOT) (i.e., all accident severity levels combined) and fatal-and-injury (FI) (i.,e., all accidents in which a fatality occurred and all accidents in which a personal injury of any severity level occurred) accidents. No SPFs are needed to estimate the frequency of fatal and severe injury (FS) accidents and property-damage-only (PDO) accidents. For FS accident frequencies, estimates are calculated within *SafetyAnalyst* by using the FI SPF multiplied by the proportion of FS accidents out of all FI accidents. Similarly, when PDO accident frequency values are needed, they are determined within *SafetyAnalyst* as the difference between TOT and FI accident frequencies.

Functional Form of SPFs

The SPFs needed for *SafetyAnalyst* predict accident frequency as a function of annual average daily traffic (ADT) volume alone. For roadway segments and ramps, the independent variable representing traffic volume is the ADT of the roadway segment or ramp. For intersections, two independent variables represent traffic volume, the ADTs of the two intersecting roads (classified as the major and minor road, where the major road is typically defined as the road with the higher ADT).

The dependent variable of the SPFs (i.e., the variable whose value is predicted by the model) is accidents per mile per year for roadway segments and ramps and accidents per year for intersections.

The functional form for roadway segment SPFs is:

$$\kappa = e^{\alpha} \times ADT^{b} \quad (1)$$

where:

κ = predicted number of target accidents per mile per year

ADT = average annual daily traffic volume (veh/day) for the roadway segment for both directions of travel combined

Since κ is expressed in terms of target accidents per mile per year, Equation (1) is equivalent to:

$$\mathbb{N} = e^{\alpha} \times ADT^{b} \times L \quad (2)$$

where:

N = predicted number of target accidents per site per year

L = length (in miles) of the roadway segment

The functional form for intersection SPFs is:

$$\kappa = e^{\alpha} \times MajADT^{b} \times MinADT^{c} \quad (3)$$

where:

κ = predicted number of target accidents per intersection per year

MajADT = average annual daily traffic volume on the major road (veh/day) for both directions of travel combined

MinADT = average annual daily traffic volume on the minor road (veh/day) for both directions of travel combined

The functional form for ramp SPFs is:

$$\kappa = e^{\alpha} \times ADT^{b} \quad (4)$$

where:

κ = predicted number of target accidents per mile per year

ADT = average annual daily traffic volume (veh/day) for the ramp

Since κ is expressed in terms of target accidents per mile per year, Equation (4) is equivalent to:

$$\mathbb{N} = e^{\alpha} \times ADT^{b} \times L \quad (5)$$

where:

N = predicted number of target accidents per site per year

L = length (in miles) of the ramp

In all three equations, a, b, and c represent the regression parameters that are estimated from the available data.

Since the default SPFs provided within *SafetyAnalyst* were developed using data from multiple states, a calibration procedure is included as part of the data import and preprocessing procedures. Yearly calibration factors are calculated within *SafetyAnalyst* during the data import and preprocessing procedures using an agency's own accident and traffic volume data and the default SPFs provided with *SafetyAnalyst*. The calibration factors are intended to account for differences in accident patterns in different geographical areas that are not directly addressed by the SPFs and provide accident predictions that are comparable to the estimates that a highway agency would obtain from SPFs developed using its own accident records system. The yearly calibration factor for a given year and site subtype is calculated as the ratio of the sum of observed accidents for all sites for a specific site subtype to the sum of the predicted accidents for the same sites using the ADT and accident count values for that year. When an agency develops its own SPFs for use within *SafetyAnalyst*, the yearly calibration factors for use with these agency defined SPFs are by definition 1.0.

Data Needs for Development of SPFs

All of the data needed for the creation of SPFs is the same data that is needed to operate and use the *SafetyAnalyst* software. As a result, it is highly recommended that an agency import their data into *SafetyAnalyst*, run the preprocessing programs on the data, then export the data for use in the development of SPFs. Alternatively, users may independently create the necessary data as described in the remaining part of this section.

The creation of SPFs requires data on the location information, roadway characteristics, traffic volumes, and accidents. Each of these types of data and their processing requirements are described next.

The roadway characteristics, which should be the current characteristics at a site, are used to define the site types and subtypes identified in the first section of this document. The logic and data used to create the subtypes in *SafetyAnalyst* out of agency data are described in the Site Subtype Assignment document on the *SafetyAnalyst* Wiki. Please note that *SafetyAnalyst* does this assignment automatically with agency data when importing the data. However, users will have to assign site subtypes themselves when using their own data to develop SPFs. Additionally, segment length data is needed for roadway segments and ramps. Sites with missing length should not be used to develop SPFs. Finally, roadway segments may also require some additionally processing based on their geometrics.

Regression models for roadway segments rely on observed accidents at a site. As roadway accidents occur infrequently over a period of time, statistically significant regression models for roadway segments are often only obtained by considering roadway segments of a certain minimal length, usually between 0.04 and 0.1 miles. For some agencies to obtain valid models, smaller roadway segment records may need to be combined into longer roadway segments to meet that minimum length; such longer segments should be as homogeneous as possible with respect to key variables such as: area type, terrain, functional class, number of through lanes, auxiliary lanes, lane width, median type, median

width, shoulder type, shoulder width, access control, driveway density, ADT, posted speed limit, two-way vs. one-way operation, bikeway, and interchange influence area.

Traffic volume data should be supplied for each calendar year for which accident data are used in *SafetyAnalyst*. For roadway segments, traffic volume should be for both directions, as the SPFs predict accidents for both directions. Some states keep separate records for traffic volumes and accidents in each direction of travel on divided highways. If a state develops an SPF for one direction of travel on a divided highway in the following form:

$$\kappa = e^{\alpha} \times ADT^{b}_{oneway} \quad (6)$$

It should be converted to a two-directional SPF for use in *SafetyAnalyst* as follows:

$$\kappa = 2(0.5)^{b} \times e^{\alpha} \times ADT^{b} \quad (7)$$

where value of the intercept term would be entered into *SafetyAnalyst* as $2(0.5)^{b}e^{a}$.

For intersections, yearly ADT are needed for both the major and minor roads, where the major road is generally defined to be the road with the highest ADT. If two major-or minor-road legs of an intersection have different ADTs, the higher of the two ADT values should be used. Years for which ADT are missing can be estimated by interpolation or extrapolation. Sites that are missing ADT information for all years should be excluded from SPF development.

Location information is needed for sites to match/merge accidents records to the sites. Accidents should be assigned to only one site and be related to the type of site. For example, roadway segments should not be matched with any intersection-related or ramp-related accidents. Accidents that occur at an intersection (within the curbline limits) or that are classified as intersection-related and occur within 250 ft of an intersection should be assigned to the intersection. Accidents should be assigned to roadway segments and ramps by matching the location information. If an accident occurs on the point between two roadway segments then it should be assigned to either the beginning segment or ending segment so that all similarly situated accidents are assigned in the same way. If an agency has difficulty in following these rules for accident assignment exactly because of data limitations, they should be followed as closely as possible.

Accident data should also include information on the severity of the accident. All reported accidents, including fatal, injury, and property-damage-only accidents should be included in the development of the total accident SPFs. The accident severity data should then be used to identify which accidents should be considered in the development of fatal and injury SPFs. Accident data should be provided for a minimum of three years (preferably five) up to a maximum of ten years. The historical accident data should include whole calendar years for each year of data.

Statistical Assumptions and Software

SPFs are usually developed with negative binomial regression analysis, but can be developed with Poisson regression analysis, depending on the relationship between the mean and variance in the data. When accident frequency data have the same mean and variance, then accident frequency follows a Poisson distribution. Alternatively, when the accident variance exceeds the mean, or the data are overdispersed, then accident frequency follows a negative binomial (NB) distribution. In fact, the variance of a NB distribution is a function of its mean (i.e., the mean plus a dispersion parameter multiplied by the square of

the mean). In this way, when the dispersion parameter nears zero, the NB distribution approaches the Poisson distribution. Since most accident frequencies are overdispersed, NB regression is typically used. However, Poisson regression is an acceptable substitute if the dispersion parameter is near zero.

The parameters of these distributions can be indirectly estimated using a generalized linear model to obtain the model regression coefficients shown in the functional form section of this document. Generalized linear models are extensions of conventional linear models where the dependant variable is related to the linear independent variables through a nonlinear link function and the dependent variable is generated from a distribution function in the exponential family. Several commercially available statistical software packages offer generalized linear model procedures that can be used to estimate the regression coefficients. In particular, the use of the procedure, PROC GENMOD[1], of the statistical package SAS will be described in this section.

In the GENMOD procedure, the model regression coefficients are estimated by the method of maximum likelihood using a ridge-stabilized Newton-Raphson algorithm. The asymptotic normality of maximum likelihood estimates is used to obtain test of significance of the parameters and goodness of fit measures for the models. To perform an analysis in this procedure, several data items will need to be specified: dataset, dependent variable and distribution, independent variable(s), link function, and offset factor.

The dataset used by the procedure should be organized as one record per site and contain the site subtype, dependent variables, independent variable(s), and offset factor. The minimum number of sites needed for model convergence is dependent on a number of accidents occurring at the sites. Site types that experience relatively high accident frequencies, such as urban intersections, will require fewer sites and/or fewer years of accident history. In contrast, rural roadway segments generally need more sites and/or more years of accident history.

Poisson and negative binomial regression are performed with a log link function, which is sometimes referred to as loglinear modeling. A log linear relationship between the accidents and ADT ensures that predictions from the fitted model are positive.

The dependent variables, or variables whose values are predicted by the model, are the number of TOT or FI accidents that are predicted to have occurred at each site during the history period. Consequently, for SPF development in *SafetyAnalyst* there should be two accident variables in the dataset, one containing the count of all (i.e., TOT) accidents occurring at the site during the history period and one containing a similar count of fatal and injury (i.e., FI) accidents. However, individual models are to be created for each dependent variable.

Traffic volume(s) are the only independent variable(s) considered in the SPF forms shown in Equations 1 through 3. As shown in these equations, these variables have a nonlinear relationship with predicted accidents. As a result, the natural logarithm of these variables should be used in the modeling procedure rather than the actual values. There does not need to be a placeholder variable in the data for the intercept, as it is automatically provided by the software.

The offset factor, or scale factor, serves to normalize the predicted accidents to a per mile per year basis, since accident frequencies, not accident rates, are used in the model. For roadway segments and ramps, the natural logarithm of the length of the segment multiplied by the number of accident calendar years is included as a scale factor. Intersections use the

natural logarithm of the number of accident calendar years as a scale factor so predicted accidents are given on a per site per year basis.

Once the dataset has been assembled for regression analysis, and the usual data quality checks for outliers and model assumptions are conducted, the following SAS code may be used to generate the SPF model coefficients:

PROC GENMOD; BY SiteSubtype; MODEL TotAcc=logADT / LINK=Log DIST=NEGBIN OFFSET=logLengthYrs;

In this specification, TotAcc and logADT are the dependent and independent variables (respectively) defined in the dataset, logLengthYrs is the offset value defined in the dataset, DIST= option specifies the negative binomial distribution, LINK= option specifies log-linear regression model, and the BY SiteSubtype option creates a separate model for each change in value of the SiteSubtype variable defined in the dataset.

Several other options are available that could be specified in the above statements to control the algorithm and display additional statistics (e.g., type I tests, type III tests, confidence intervals, etc.). In particular, using the combined options of DIST=NEGBIN SCALE=0 NOSCALE, can test for overdispersion in a Poisson model. Overdispersion is assessed by testing whether the negative binomial dispersion parameter is equal to zero (i.e., when the negative binomial distribution is equivalent to the Poisson distribution).

Model convergence, goodness-of-fit, and statistical significance of the coefficients can be assessed with the output automatically generated by the software. While all of the details of these components cannot be described in this document, some basic guidelines on excluding models for use in *SafetyAnalyst* are provided in the remainder of this section.

Coefficients from models that do not converge should not be used in *SafetyAnalyst*. Nonconvergence of the Newton-Raphson algorithm can occur for some datasets. Poor performance can be the result of a number of factors: ill-conditioned data (e.g., data that are extremely large or extremely small), a non-positive definite Hessian matrix can indicate linear dependencies among the parameters, model misspecification, or violations of the error assumptions. However, the most probable explanation for nonconvergence is lack of event data (i.e., small accident counts). Asymptotic or large sample inference used by this procedure, which is based on maximizing the likelihood function, may not be appropriate in situations where the total number of accidents in any site subtype group is small. For those groupings, exact Poisson regression is a more statistically valid approach to get regression estimates and p-values. However, this capability is not currently provided in this procedure, rather LogXact 4.1 software should be used.

There are several goodness-of-fit statistics available by the procedure. However, the preferred fit statistics used by transportation researchers are usually calculated from output generated by the procedure rather than the procedure itself. For example, the Freeman-Tukey R^2 coefficient[2] (RFT^2) has been presented in Appendix J of the User Manual for the SPFs that have been developed for *SafetyAnalyst*. This goodness-of-fit statistic is not available within the software procedure and must be calculated outside of the procedure. As observed in Appendix J, selection of models for use with *SafetyAnalyst* was not limited by the value of this goodness-of-fit statistic, particularly if it was the only available model. However, an agency developing their own SPFs may compare and consider this goodness-of-fit statistic when choosing between their own SPFs or the ones provided by *SafetyAnalyst*.

Coefficient estimates provided by the procedure should be assessed by their magnitude and direction as well as their statistical significance. For example, a negative coefficient for ADT, indicating accidents decrease as ADT increase, is probably never appropriate. Also, the significance level used to assess coefficient estimates is more relaxed in transportation research. A significance level of 10% is generally used to assess coefficient estimates for ADT on roadways, ramps, and the major road of intersections. However, a significance level of 20% is usually considered for minor road ADT on intersections.

Finally, the estimate of the dispersion parameter should always be positive. That is, a negative value for the dispersion parameter should indicate that there are problems with the model, even though no warnings are issued by the SAS procedure. However, if the value is close to zero, then the model should be re-tried assuming a Poisson distribution. Since the EB-methodology used in *SafetyAnalyst* requires a value for the dispersion parameter, SPFs generated from Poisson distributions should assume a small positive value, like 0.3.

References

[1] http://support.sas.com/documentation/cdl/en/statug/59654/HTML/default/statug_genmod_sect010.htm

[2] Fridstrom, L., et al. "Measuring the Contribution of Randomness, Exposure, Weather, and Daylight to the Variation in Road Accident Counts," Accident Analysis and Prevention (1995), Vol. 27, pp. 1-20.