SPF-R

ONLINE

USER GUIDE

CONTRIBUTIONS

The following individuals have notably contributed to this tool either by working directly on the online version, or on the original R code that powers it.

- Dr. Eric Green

- Paul Ross

- Dr. Chris Blackden

- William Staats

- Dr. Reginald Souleyrette

Developed by the Kentucky Transportation Center at the University of Kentucky. This guide was edited by Dr. Christopher VanDyke.

Special thanks to Vanasse Hangen Brustlin (VHB), the Federal Highway Administration (FHWA), and the many state DOT offices that collaborated with the Kentucky Transportation Center in the development of this tool.

ABOUT SAFETY PERFORMANCE FUNCTIONS

Since its release, the *Highway Safety Manual* (HSM) has facilitated the adoption of new approaches by highway safety professionals (*1*). Highway safety has traditionally been measured using number of crashes, crash rates, crash costs, or a combination of these metrics. High-crash locations are selected based on somewhat arbitrary rankings, or by comparing crash rates to a critical rate factor. All methods have clear disadvantages, particularly for network screening (*2*). Most notably, none of these methods account for regression to the mean or selection bias. When observed in crash data, these biases can produce misleading results when not corrected for.

Traditional crash analysis relies on crashes normalized by exposure — typically traffic volume — to create a rate. However, the use of rates erroneously assumes a linear relationship between crashes and volume (*3*). Most safety performance functions (SPFs) — models used to predict crashes based on traffic volume and other factors — find an exponential relationship between crashes and exposure (only when the exponent is 1.0 is a constant rate observed across the volume spectrum).

Network screening using the HSM's methodology redresses many disadvantages of the traditional methods. SPFs are developed to better characterize the relationship between crashes and traffic volumes as well as other variables. Empirical Bayes (EB) addresses regression-to-the-mean bias by using actual crash data and the overdispersion parameter to adjust the expected crash experience at a site. This adjusted value is a more realistic measure of a site's safety performance. More importantly, it describes the magnitude of crash reduction that could be potentially be achieved. In Kentucky, this is referred to as *Excess Expected Crashe*s (EEC). Other states use the term *potential for safety improvement* synonymously.

Description	Functional Form
Typical	$L * e^a AADT^b$
IIMS	$AADT * L * 365 * 10^{-6} * e^a$
Intersection	$L * e^a AADT\_Maj^{b1} AADT\_Min$
Shoulder	$L * e^a AADT^{b1} e^{SW×b2}$

Manual Development

- Better than a crash rate
- May be time consuming
- Large margin for error

SPF-R

- Better than doing it manually
- Not user friendly if user doesn't know how to code in R.

SPF-R Online

- Same functionality as SPF-R
- No code requirements
- No install needed, all online!

SPF DEVELOPMENT HISTORY

MANUAL DEVELOPMENT

AASHTO's *Highway Safety Manual* (HSM) provides guidance for safety analysis in highway transportation project planning and development. The HSM contains predictive methods for analyzing infrastructure improvement project alternatives and development/design, including (a) the use of national SPFs to predict crashes by severity on specific facility types and base conditions and (b) associated crash modification factors (CMFs) used to estimate the potential effects of design alternatives or changes from base conditions (4).

DEVELOPING SPFS IN R

In 2017 the Kentucky Transportation Center developed a tool using the free and open-source statistical computing and graphics environment R. Using methods outlined by the HSM, researchers created SPF-R, which automated SPF development to improve regression models. The tool gave states that had not adopted newer methodologies due to the challenge of SPF development the ability to create SPFs. SPF-R provides immediate feedback during SPF development, including goodness-of-fit measures and model visualizations. Although the FHWA Safety Toolbox now includes SPF-R, it must be downloaded locally to a computer and requires a working knowledge of executing R code.

SPF-R ONLINE

Recognizing that running code may be a barrier for some practitioners, SPF-R Online is a new web-based interface that lets users develop SPF by uploading a CSV file with certain predefined parameters. Users of this tool will encounter fewer technical challenges getting SPF-R working as the mechanics are all handled online. SPF-R Online remains free and open-source, but it significantly improves and expands accessibility.

USING SPF-R ONLINE IS EASY

Required Fields:

- Crashes
- AADT
- Length[1]

Common Fields:

- BMP
- FROM
- BEGIN_MP
- START
- BEGIN
- EMP
- TO
- END_MP
- END
- ACCIDENTS
- CRASH
- KABCO
- COUNT
- AF

PREPARING YOUR CSV

Before starting, prepare the CSV file. A CSV uploaded to SPF-R must (a) have columns with headers and (b) a least three fields: Crashes, Average Annual Daily Traffic (AADT), and Length. Without these fields SPF development is not possible. Column names are not case sensitive.

Tip: It's a good idea to wrangle or munge data before uploading it to SPF-R Online. Look for odd characters, spaces, or null values and delete them from the CSV prior to import to achieve the best results.

Columns can take any name. Common names include BMP, FROM, BEGIN_MP, START, BEGIN, EMP, TO, END_MP, END, ACCIDENTS, CRASH, KABCO, COUNT, AF (used for adjustment factors).

Note: The first row of the CSV determines the column's data type. Columns can be numeric or strings. Columns categorized as numeric that include strings can cause errors in the modeling process.

---

[1] For intersections, length is not needed and a minor road AADT should be included.

STEP 1 — UPLOAD DATA

- Once the data are prepared, navigate to https://spfr.uky.edu. When the page loads, click on the link that reads *Click here to visit the SPF-R Tool* (Figure 1).
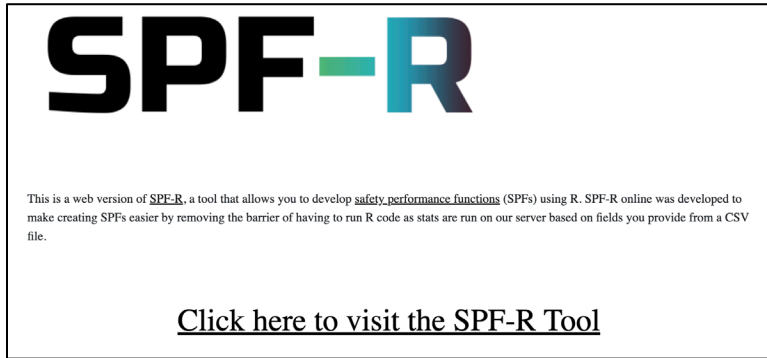

Figure 1 SPF-R Homepage

- In the box whose header reads *Step 1: Upload Data*, select the *Choose File* button (Figure 2). Locate and select the appropriate CSV file. The name of the selected CSV file appears next to the *Choose File* button.
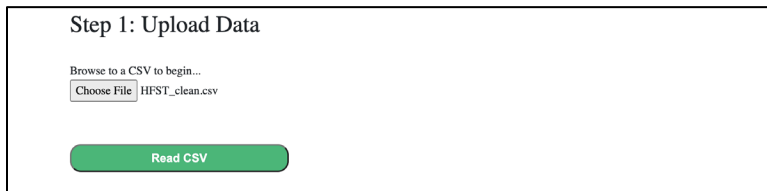

Figure 2 Uploading Data

- Click the *Read CSV* button.

STEP 2 — REVIEW AND FILTER

- Once a CSV has been uploaded, the webpage refreshes. Scroll down to the box that reads *Step 2: Review and Filter*. The box displays data from the CSV. Variable (or field) names appear at the top of the data table in bold text. If there are a large number of variables, data may extend beyond the edge of the box.

Figure 3 Display of Uploaded Data

- Because applying filters can improve model development, SPF-R gives users the option of filtering data. At the bottom of *Step 2* box, up to 10 filters can be applied. Filters are numeric (e.g., greater than, less than) or string based. SPF-R categorizes each column as numeric or string based on the first row of data. If the first row of data in a column is numeric, then the column is categorized as such, otherwise it is treated as a string (text). This is an important note as

only numeric columns can be used later in the analysis. Unwanted text in a numeric column may cause errors in the modeling process.

- After scrolling to the bottom, users see a button for *Filter 1* shaded brown, which indicates it is active. *Filter 1* is defined by entering parameters using the dropdown menus and text boxes beneath the rows of filter buttons. In Figure 4, *Filter 1* is used to select all records with average annual daily traffic (AADT) of less than 125.
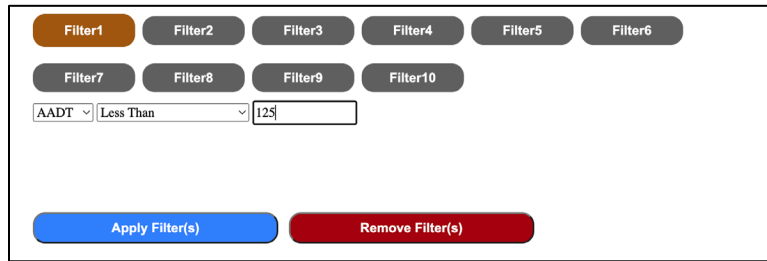
Figure 4 Applying a Filter

- Once the parameters for *Filter 1* have been entered, users can insert additional filters. To activate another filter, click on the appropriate ellipse. Once an ellipse is shaded brown it is active. In Figure 5, a filter has been applied to a column containing text, and a single route ID has been unchecked. A filter like this can be useful if there are known errors in the data that may contribute to outliers. The Select All and Clear All buttons can be used to select or deselect the checkboxes.
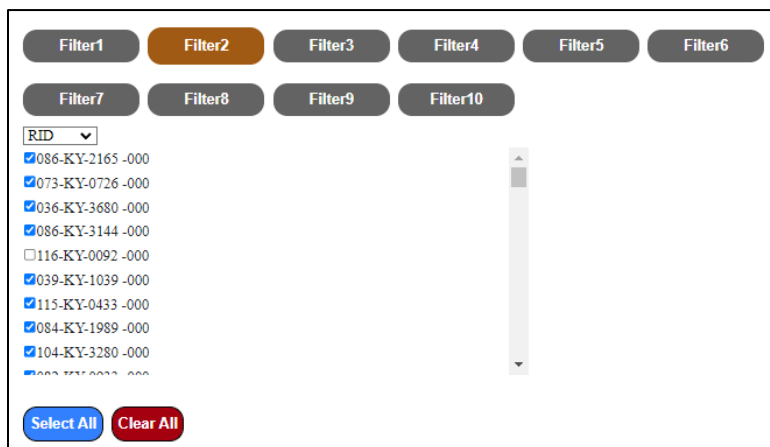
Figure 5 Applying a String Filter Based on RID

- Once all filters have been set up, scroll to the bottom of the *Step 2* box and click the *Apply Filter(s)* button (Figure 6). All filters default to AND logic. Taking the AADT and RID used in the previous two steps as an example, this means SPF-R returns segments whose AADT is less than 125 **and** route ID not equal to '116-KY-0092 -000'.
- To apply OR logic to an individual filter, uncheck the box that reads *Require this filter in addition (AND) to the other filters (uncheck if you want to make this an OR condition)*. Returning to the previous example, to identify segments whose AADT is less than 125 **or** route ID not equal to '116-KY-0092 -000', uncheck this text for both Filter 1 and Filter 2.
- After the filters have been applied, results of the query are displayed so users can check for errors. All filters are applied at once.
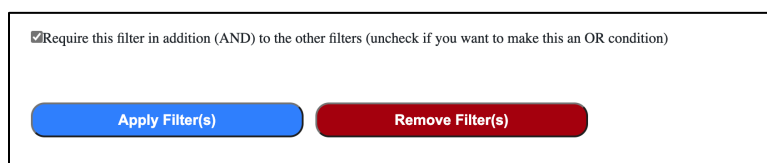
Figure 6 Selecting AND/OR Logic and Applying or Removing Filters

- Clicking *Remove Filter(s)* removes **every** filter that has been applied. Users can experiment with different values and/or logical operators. Use this to make any changes to your filters as well.
- Filters can provide a great way to improve model development. For instance, a filter can be applied to remove all segments with a shoulder width less than 3 feet. Once the model is developed, the user can remove the filter to see if the model improves or not.

STEP 3 — MAP FIELDS

- Once the filters have been applied, scroll down to the box that reads *Step 3: Map Fields*. First, define the dependent variable from the list of fields listed in the scrollable menu under the text that reads *Dependent Variable (select one)* (Figure 7).
- Because SPFs are crash prediction models that relate crash frequency to site characteristics, the dependent variable is always the number of crashes for a given segment or intersection. For some analyses, this may be total crash frequency. In other cases, users may only be concerned with the number of fatal crashes. What is imperative is choosing a dependent variable that quantifies crash numbers.
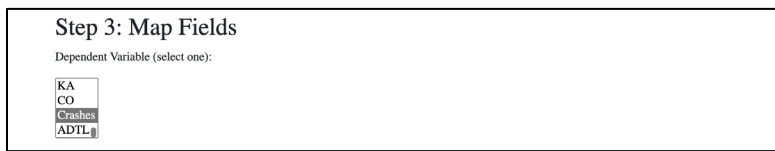
Figure 7 Choosing a Dependent Variable

STEP 4 — MODEL FORM

- Users can develop an SPF using up to 10 variables. At the top of the *Step 4: Model Form* box, choose what variables to include in the model. Any variable in the CSV can be selected (only variables categorized as numeric are shown). By default, *Variable 1* is active (as indicated by the ellipse shaded brown). To define a variable, use the dropdown menu under the rows of ellipses. In Figure 8, *Length* is assigned to *Variable 1*. To input additional variables, click on a gray ellipse that reads *<click to add>* and make an assignment using the dropdown menu.
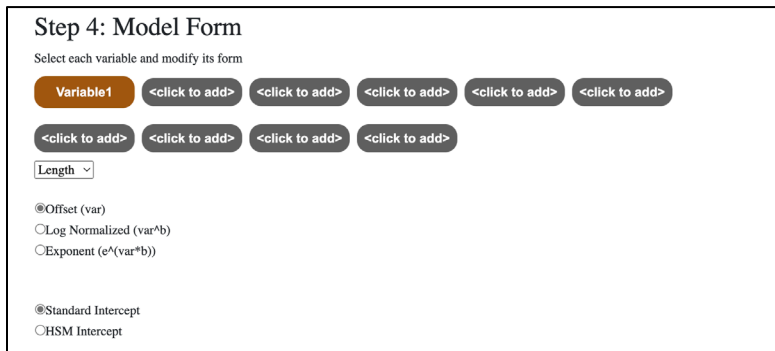
Figure 8 Defining Variables for the SPF

As a starting point, it is recommended to use the following model form for segments:

$y = L * e^a * AADT^b$ (Eq. 1)
where:

    L = the length of a roadway segment.
    AADT = average annual daily traffic
    a = regression parameter for intercept
    b = regression parameter for AADT

For intersections:

$$y = e^a * AADT_{maj}^{b1} * AADT_{min}^{b2} \text{ (Eq. 2)}$$
where:

> AADT = average annual daily traffic for major and minor approaches
> a = regression parameter for intercept
> b1/b2 = regression parameter for AADTs

- For each variable, three mathematical forms are available — Offset, Log Normalized, or Exponent. Select a model form using the appropriate radio button (Figure 8).
  - An offset variable does not generate a regression parameter. In eq 1, L is an offset.
  - Log normalized variables are used as the base with a regression parameter as the exponent. In Eqs. 1 and 2, AADT is log normalized.
  - When a variable is modeled as an exponent, it is an exponent of the natural log raised to the power of the variable (e.g., e^{AADT*b}).

Tip: Offset is typically used when the relationship between the variable and the dependent variable is suspected to be linear. Length is often used as an offset. Log normalized variables that generate a regression coefficient near 1 can potentially serve as an offset.

To decide whether to use a variable in its log normalized or exponential forms, try both and compare the results. SPF-R includes several metrics that can be used to improve models (see below).

- Variables 2 through 10 can be used as interaction terms if the previous variable is modelled as an exponent. Interaction terms can be beneficial if the user suspects an interaction between two variables. In this case both variables are multiplied by the same regression coefficient. Eq. 3 provides an example using shoulder width and lane width.

$$y = L * e^a * AADT^{b1} * e^{SW*LW*b2} \text{ (Eq. 3)}$$

where:

> AADT = average annual daily traffic for major and minor approaches
> a = regression parameter for intercept
> b1/b2 = regression parameters
> SW = Shoulder Width
> LW = Lane Width

- By default, SPF-R uses the standard intercept (e^a). Users can also apply a multiplicative factor that matches some SPFs in the HSM. In the latter case, the model form includes 365 x 10^-6. Neither option changes the model, just the magnitude of the regression coefficients. Select the appropriate intercept form by clicking the radio button next to the preferred option (Figure 8). Every time a model is revised, SPF-R presents the model in its mathematical form, which lets users check for errors.

- Users can specify crash and roadway information to include in the report SPF-R produces, including crash type, crash severity, road type, intersection type, timeframe, and other notes (Figure 9). Enter what information to include in the appropriate text boxes.

Figure 9 Model form and Request for Crash/Roadway Information

- Users have the option of selecting *Constant Over-Dispersion* or *Variable Over-Dispersion*. The most common approach is to assume a constant over-dispersion parameter. But exceptions may arise when roadway segments are modeled. To learn more about over-dispersion, see p. 17 of *the Safety Performance Function Development Guide: Developing Jurisdiction-Specific SPFs*. The Resources section links to this publication.

- Before SPF-R creates a model, choose an initial value for theta (Figure 10). The default value is 0.1. Users can adjust this up or down if model convergence issues arise. SPF-R and R represent the over-dispersion parameter as theta, which is the inverse of k. The HSM and many statistical packages use k. Larger values of theta are preferred.

Figure 10 Additional Model Options

- Users can also specify a timeout value. The default value is 60 seconds. When using larger datasets, users may want to choose a longer interval.

- Specify if the model is for intersections or segments. Developing a model for segments requires information on segment length. To specify the model type, click the radio button next to *Intersections* or *Segments (required a length field)*. From the dropdown menu, specify the numeric column to use for segment analysis (Figure 10).

- SPF-R generates a cumulative residual plot (i.e., CURE plot) when it develops an SPF. Using the dropdown menu, specify a variable from the CSV file to serve as the x-axis on the plot. Most CURE plots place AADT or length on the x-axis. A good CURE Plot has the following characteristics:

  o Oscillates around the x-axis and ends near zero.
  o Free of outliers as they can adversely affect the model parameters (e.g., large vertical jumps).
  o Cumulative residuals should rarely transgress confidence bands (blue and red dots).
  o Minimal upward or downward drifting.

  CURE plots can help identify outliers and are a good visual means of determining goodness of fit. See p. 11 of *The Calibrator: An SPF Calibration and Assessment Tool User Guide* for information on using CURE plots to assess goodness

of fit. Learn more about CURE plots in *SPF Development Guide: Developing Jurisdictional-Specific SPFs*. The Resources section links to both publications.

- To generate an SPF, click on the box that reads *Develop SPF!*
- An error message will appear if the modeling process fails. Consult the text in red to address the issue.

There was a fitting error. Try adjusting the initial theta. Or check that you have a sufficient sample size.

Develop SPF!

Figure 11 Model Error

RESULTS

- Upon successful model development, the results are displayed in a textbox in a structured text format that can be computer-read. This can be helpful if the results are to be compared to other models. A spreadsheet can easily be used to synthesize the results of several different models.

```
SPF-R results
Date run: 12/12/2022 10:52:26 AM
-------------------------------------------------------
Adjustment factor: None
Overdispersion: Constant Over-dispersion
Overdispersion=0.482668 (1/k)

a=-4.901, intercept, stderror = 1.04, statistic = -4.714, pvalue = 1.0396
b0=0.60973, var2, stderror = 0.121, statistic = 5.021, pvalue = 0.1214
-------------------------------------------------------
Sample: 2753
Mileage: 190.76 (ideally 100-200 miles)
```

Figure 12 Model Results

- Secondly, a CURE plot is displayed below the results. This plot quickly informs the user of the model quality before reading the results in more detail. A poor CURE plot (as defined above) can indicate the model should be adjusted.
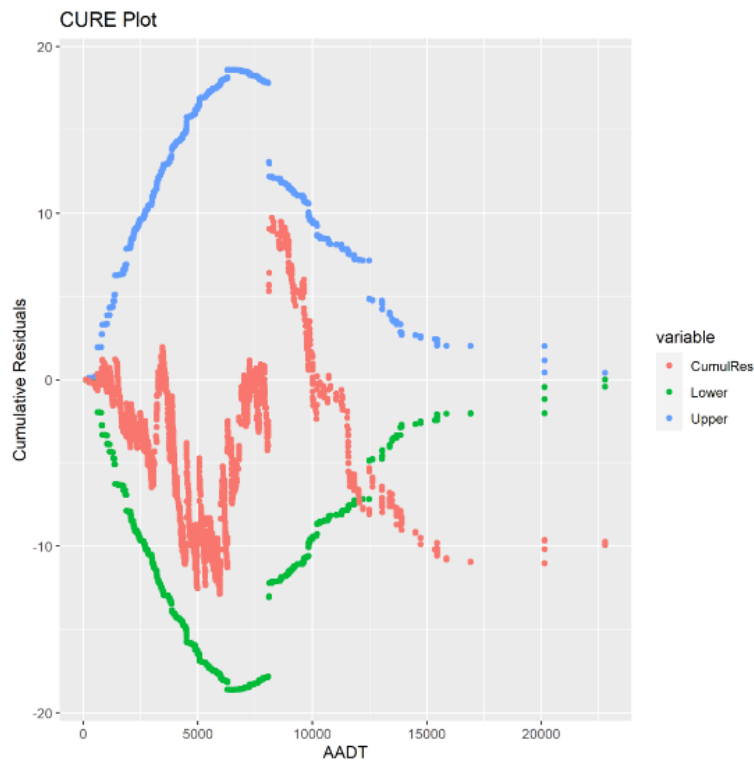

Figure 13 An Ideal CURE Plot

- Lastly, the user can export results and delete input data. Results can be exported to Excel, which includes several new fields added to the data originally imported. Results can also be displayed in an HTML report. The user can also delete the imported CSV from the server if they do not want the data stored on the SPF-R server.

Export Results:

Create Excel Export   **Create HTML Report**

If you are working with confidential data and wish to delete it from the server logs, please click the button below:
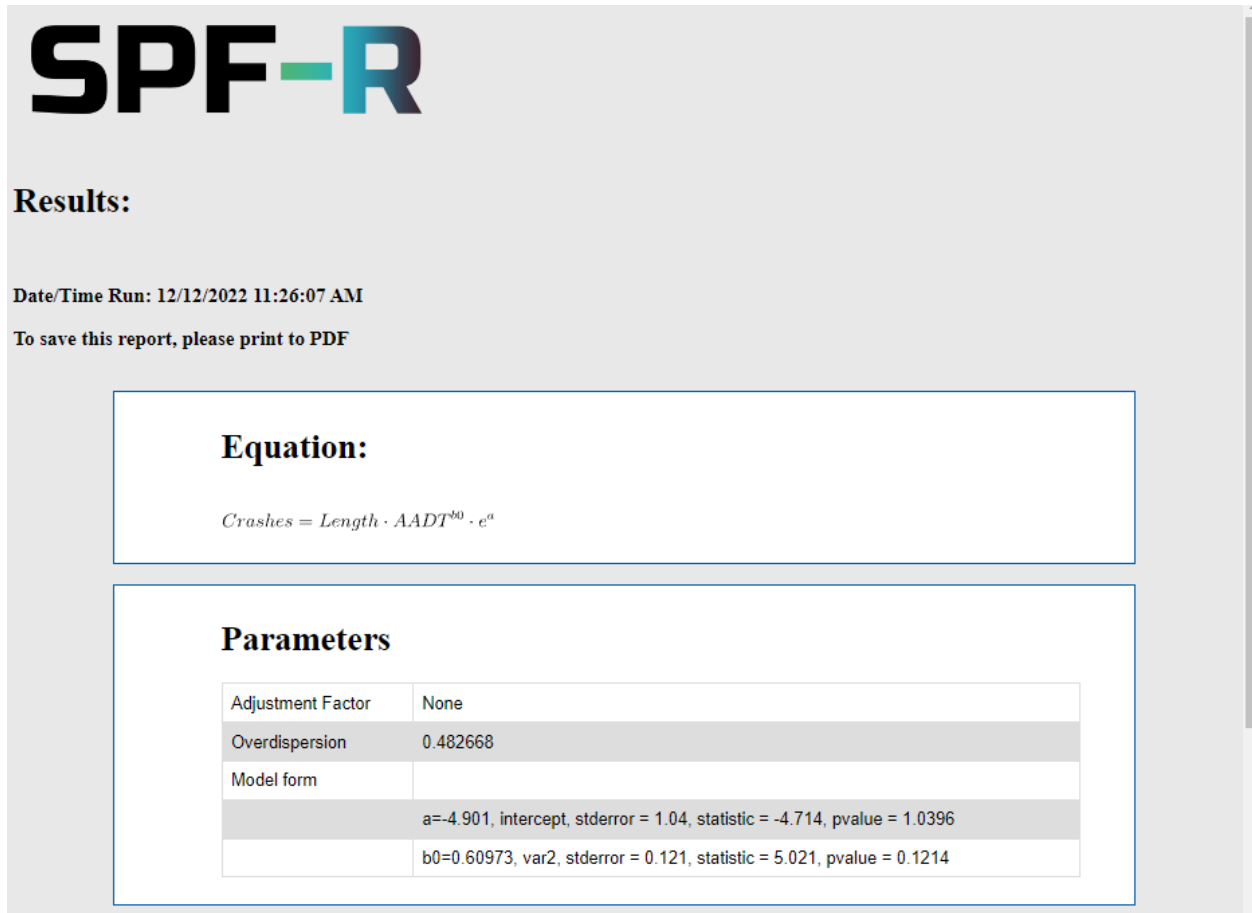
Remove my CSV from Server

Figure 14 Result Options

- The Excel report adds the fields listed in the table below. Each field is briefly defined. Refer to the R code in the results or in the *Results In Detail* section below for more information on each column. Columns defined as *an intermediate step* contain data that are used behind the scenes to complete analysis and generate visualizations.

Field	Description
Predicted	The predicted number of crashes using the resulting model. The user can use a spreadsheet to calculate the prediction using the resulting model parameter to check that the correct model form was used. The two predictions should match to several decimal points.
Residuals	An intermediate step for generating the CURE plot
CumulRes	The cumulative residuals used in the CURE plot
Squared_Res	An intermediate step
CumulSqRes	An intermediate step
SigmaSum	An intermediate step
StdDev	An intermediate step
UpperLimit	The upper limit of the CURE plot
LowerLimit	The lower limit of the CURE plot
Per_CURE	An intermediate step
Weight	The weight used in EB analysis
EB_Estimate	The EB estimate used in EB analysis
EEC	Excess Expected Crashes. This number can be used in network screening to prioritize segments or intersections. Sites with higher EECs should be considered over lower EECs. The cost of countermeasures to improve a site should be considered when selecting EECs.
CI	The standard deviation of the estimate. The actual estimate can range from (EB_Estimate – CI) to (EB_Estimate + CI). Similarly, the same range can be created for EEC.

- The HTML report combines all the results (including the CURE plot) into a single web page. A portion of the report is shown in Figure 15. The report can be printed to or saved as a PDF. Use the browser's back button to return to SPF-R.

SPF-R

Results:

Date/Time Run: 12/12/2022 11:26:07 AM

To save this report, please print to PDF

Equation:

$$Crashes = Length \cdot AADT^{b0} \cdot e^a$$

Parameters

Adjustment Factor	None
Overdispersion	0.482668
Model form	
	a=-4.901, intercept, stderror = 1.04, statistic = -4.714, pvalue = 1.0396
	b0=0.60973, var2, stderror = 0.121, statistic = 5.021, pvalue = 0.1214

Figure 15 A Portion of the HTML Report

RESULTS IN DETAIL

Text-based results are displayed immediately in a window whose contents may be copy and pasted into any application.

The following sample output provides explanatory comments in red. The results also include a CURE plot.

SPF-R results
Date run: 8/12/2022 4:40:36 PM

Adjustment factor: None – If used, the variable is listed here
Overdispersion: Constant Over-dispersion
Overdispersion=0.482668 (1/k)
Crash=Length*(AADT^b0)*(e^a) – The model form is shown here in notation similar to Microsoft Excel
a=-4.901, intercept, stderror = 1.04, statistic = -4.714, pvalue = 1.0396
b0=0.60973, var2, stderror = 0.121, statistic = 5.021, pvalue = 0.1214

Sample : 2753
Mileage : 190.76 (ideally 100-200 miles) – Suggested by the HSM
Crashes : 239 (ideally 300 crashes per year) – Suggested by the HSM
Modified RSquared : 0.147 (Higher values preferred) – Compare this value to values for other models generated with the same dataset
CDP : 11.224 (ideally less than 5%)
MACD : 19.1 (Smaller values preferred)
MAD : 0.153 (Smaller values preferred)
StdErr : 0.12
AIC : 1554.569 – Useful to compare goodness of fit only when the sample size is the same
Filter : No filter – Filters used are shown here

Variable#1, Column Name = Length, no parameter for offset – Up to 10 variables are listed
Variable#2, Column Name = AADT, b0=0.60973
Processed 2753 record(s)

All R Code: — All of the R code is shown here and can be used to run the model on your own.
library(MASS)
library(ggplot2)
library(broom)
data=read.csv('C:/pathtoyourfile/yourFileName.csv',header=T) – CSV path is unknown to the server for security reasons
data<- data[order(data$AADT),]
dep=data$Crash
var1=log(data$Length)
var2=log(data$AADT)

```
init.theta = 0.1
SPF=glm.nb(dep~offset(var1)+var2)
coef<-coefficients(SPF)
summary(SPF)
theta<-SPF$theta
out<-
cbind(data,Predicted=SPF$fitted.values,Residuals=resid(SPF,type='resp'),CumulRes=cumsum(resid(SPF,type='resp')))
curedata <- data.frame(out$Residuals)
curedata$AADT <- data$AADT
curedata$CumulRes = out$CumulRes
curedata$Squared_Res = curedata$out.Residuals^2
curedata$CumulSqRes = cumsum(curedata$Squared_Res)
curedata$SigmaSum = sqrt(curedata$CumulSqRes)
curedata$StdDev = curedata$SigmaSum*sqrt(1-curedata$CumulSqRes/sum(curedata$Squared_Res))
curedata$UpperLimit = curedata$StdDev * 1.96
curedata$LowerLimit = curedata$StdDev * (-1.96)
curedata$Per_CURE =
ifelse(curedata$CumulRes>curedata$UpperLimit,1,ifelse(curedata$CumulRes<curedata$LowerLimit,1,0))
nrow(out)
sum(out$Length)
sum(out$Crash)
ObsAvg = mean(out$Crash)
tmpTerm = sum((out$Crash-ObsAvg) ^ 2)
tmpTerm2 = sum((out$Crash-out$Predicted)^2)
(tmpTerm - tmpTerm2) / (tmpTerm - sum(out$Predicted)) # Modified RSquared
sum(curedata$Per_CURE)/length(curedata$Per_CURE)*100 #CDP
max(abs(curedata$CumulRes)) # MACD
mean(abs(out$Residuals)) #MAD
SPF$SE.theta # std err
SPF$aic # AIC
out$Weight=1/(1+out$Predicted/out$Length/SPF$theta)
out$EB_Estimate <- out$Crash*(1-out$Weight) + out$Predicted*(out$Weight)
out$EEC <- out$EB_Estimate - out$Predicted
out$CI <- sqrt((1-out$Weight)*out$EB_Estimate)
```

CONCLUSION

SPF-R Online greatly improves and simplifies the SPF development process. Building on the original SPF-R script, the online tool eliminates many of the technical barriers associated with the original process, allowing for a much larger audience to take advantage of SPF-R's capabilities. Like the original SPF-R script, SPF-R Online lets users quickly explore the effect of roadway network heterogeneity on SPF development, compare CURE plots , and compare goodness-of-fit measures to decide which SPF is most appropriate for a given dataset.

RESOURCES

The following resources offer information on SPF development and calibration.

- The *Highway Safety Manual*, First Edition
- NCHRP Project 20-7 (Task 332): *User's Guide to Develop Highway Safety Manual Safety Performance Function (SPF) Calibration Factors.*
- *SPF Decision Guide: SPF Calibration vs. SPF Development.*
  - https://safety.fhwa.dot.gov/rsdp/downloads/spf_decision_guide_final.pdf
- *SPF Development Guide: Developing Jurisdiction-Specific SPFs.*
  - https://safety.fhwa.dot.gov/rsdp/downloads/spf_development_guide_final.pdf
- SPF-R Tool (FHWA Safety Toolbox)
  - https://safety.fhwa.dot.gov/rsdp/toolbox-content.aspx?toolid=210
  - https://github.com/irkgreen/SPF-R
- *The Art of Regression Modeling in Road Safety* by Ezra Hauer
  - http://www.springer.com/us/book/9783319125282
- *The Calibrator: An SPF Calibration and Assessment Tool User Guide*
  - https://safety.fhwa.dot.gov/rsdp/downloads/fhwasa17016.pdf
- *User's Guide to Develop Calibration Factors*
  - https://safety.fhwa.dot.gov/rsdp/toolbox-content.aspx?toolid=116

REFERENCES

1. AASHTO. Highway Safety Manual, first edition. *American Association of State Highway and Transportation Officials*, Washington, DC, 2010.

2. Wu, Kun-Feng, Scott C. Himes, and Martin T. Pietrucha. Evaluation of effectiveness of the federal Highway Safety Improvement Program. *Transportation Research Record: Journal of the Transportation Research Board, No.* 2318.1, 2012: pp. 23-34.

3. Hauer, E., *On Exposure and Accident Rate*. Traffic Eng. Contr. 36 (3), 1995, pp. 134–138.

4. AASHTO - Highway Safety Manual. (2010). Retrieved 25 March 2021, from
   http://www.highwaysafetymanual.org/Pages/About.aspx

LEGAL NOTE

SPF-R was developed by the Kentucky Transportation Cabinet as a free and open-source tool in partnership with VHB, FHWA, and KYTC.