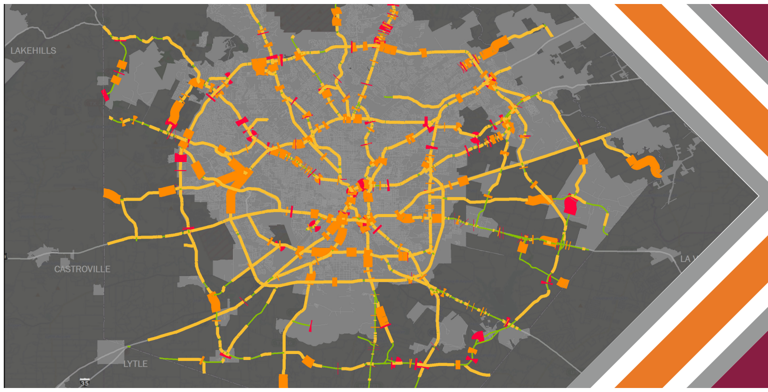
SAFE-D
SAFETY THROUGH DISRUPTION

Project Brief

Analyzing Highway Safety Datasets:

Simplifying Statistical Analyses from Sparse to Big Data



Safe-D Project 01-001:
Big Data Methods for
Simplifying Traffic Safety Analyses

Principle Investigator:
Dominique Lord
dlord@civil.tamu.edu
http://vtti.vt.edu/utc/safe-d/

Problem Statement

Three characteristics of transportation safety data make analysis challenging: (1) the large number of zero observations, (2) the rare occurrence of crash events, and (3) large datasets.

Zero Observations: Modeling crash data with many zero observations requires two critical precautions:

> *Assembling and formatting data*. Finding a balance in aggregation is a critical task in data preparation. Disaggregated data may result in excess zero observations, and the traditional negative binomial (NB) model may not be appropriate. Too much aggregation may result in loss of information.
>
> *Selecting an appropriate distribution*. Selecting the most appropriate distribution plays a crucial role in safety analyses. Often, the comparison of distributions (or models) is accomplished during the post-modeling phase, using measures such as goodness-of-fit statistics. These metrics are neither easy to compute nor practically attainable in some instances when many alternatives exist and/or the analyst deals with big data or excess zero observations. Most importantly, these metrics do not provide any intuition into why one distribution is preferred over another or the logic behind the model selection (goodness-of-logic).

Rare Occurrence of Crash Events: NB regression is a fundamental statistical tool for traffic safety modeling. A limited number of crashes and/or imbalanced data can lead to finite sample bias (i.e., biased regression parameter estimation).

Large Datasets: Big datasets are becoming more prevalent with the use of naturalistic data. Due the size and complexity of these datasets, data storage, processing, and modeling have become challenging.

Objectives, Methods, and Data

The objective of this study was to provide guidelines and tools for the analysis of highway safety data characterized by excess zero responses, rare events, and big data. The objective was addressed for each analytical challenge as follows.

Zero Observations:
We developed the following:

- Guidelines for aggregating data over time and space.
- Heuristics to determine when the Poisson-lognormal (PLN) is preferred over the NB model.
- Heuristics to determine when the Negative Binomial Lindley (NB-L), which is a recommended model when the dataset contains excess zero responses and/or high dispersion, is preferred to the NB model.

Rare Occurrence of Crash Events:

- We propose bias adjustment for more accurate estimation of the safety impact of a risk factor.
- We developed a decision-adjusted modeling framework for predicting crash risk.

Large Datasets:

- We utilized cluster analysis methods to classify data into groups with similar characteristics.
- We created predictors using cluster analysis to potentially produce insight or reduce the number of random variables (i.e., dimension).

The study objectives were accomplished using simulation and the analysis of a naturalistic dataset. The dataset used in the analysis contains information collected from 2012 to 2015 on 5,238 short road segments. There were 32,298 crashes for 10,894,920 passing vehicles, resulting in an average crash rate of 2.96×10^(-3) crashes/passing vehicle. The original dataset had 37 variables. This dataset was used for both the second and third study objectives.

Recommended Guidelines

- For datasets that have a large percentage of zero responses (50% or above):
    - When the percentage of zeros is higher than 70%, aggregate the data only if the change in the coefficient of variation (CV) of all variables when data are aggregated compared to the disaggregated data is less than 7%.
    - When the percentage of zeros is less than 70%, aggregate the data only if the change in CV of all variables when data are aggregated compared to the disaggregated data is less than 4%.

- When data aggregation is not possible:
    - Select the NB-L over the NB when the skewness is greater than 1.92, independent of the number of zero responses.
    - The selection of the PLN over the NB is governed by the percentage of zeros and the kurtosis. The boundaries are presented in Figure 1.

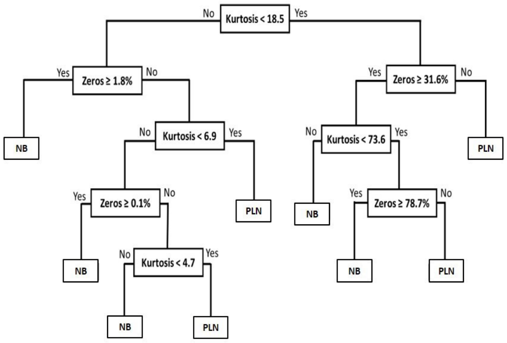


Figure 1. Heuristic to Select a Model between the NB and PLN Distributions.

- For imbalanced datasets or those with a small number of crashes:
    - Use bias-correction procedure to reduce errors with the estimation of the coefficients. Bias adjustment should be performed when the number of crashes is less than 50 in any stratum.
    - For rare-event prediction, use a decision-adjusted framework, which will provide better predictive power.

- To discover hidden patterns in the data, especially when the dataset is large, apply cluster analysis to create new predictors to potentially produce insight or reduce dimension.