A Simulation Analysis to Study the Temporal and Spatial Aggregations of Safety Datasets with Excess Zero Observations

Mohammadali Shirazi, Ph.D.*
Assistant Professor
Department of Civil and Environmental Engineering
University of Maine, Orono, Maine, 04469, United States
Email: shirazi@maine.edu

Srinivas Reddy Geedipally, Ph.D. P.E.
Associate Research Engineer
Texas A&M Transportation Institute, Arlington, TX 76013, United States
Email: srinivas-g@tti.tamu.edu

Dominique Lord, Ph.D.
Professor and A.P. and Florence Wiley Faculty Fellow
Zachry Department of Civil Engineering
Texas A&M University, College Station, TX 77843, United States
Email: d-lord@tamu.edu

Words count: Text (5574 words) + 5 Tables (1250 words) = 6824 words

***Corresponding author**

ABSTRACT

Crash data are often characterized with numerous zero observations. Sometimes, the number of zero observations in the compiled dataset is directly correlated with the selected spatial and/or temporal scales. By adjusting the time and spatial scales, the number of zero responses observed in the dataset can increase or decrease. Finding a balance in aggregation is a critical task in data preparation. On the one hand, using the disaggregated data may result in having excessive zero observations, in which the traditionally used negative binomial model may not be adequate for the safety analysis. On the other hand, too much aggregation may result in loss of information. This paper documents a simulation study that aimed at determining the criteria for deciding when data aggregation is needed. The simulation study explores the information loss due to aggregation as a function of the precision or accuracy in the estimation of model coefficients. The simulation results indicate that the reduction in variability, i.e., coefficient of variation, of the independent variables upon aggregation, is an important criterion to decide on the aggregation level.

1.0 INTRODUCTION

Crash data modeling plays a crucial role in various analyses and evaluations related to the safety of transportation facilities. Sometimes, the statistical modeling of crash data is metaphorically referred to as an art, due to the important challenges the analyst may face during the analysis. The art of a reliable statistical modeling involves various critical steps, from collecting data, to cleaning the dataset, to selecting, estimating and assessing an appropriate model. While the third task in this cycle has been extensively studied in various studies in safety literature (*1- 3*), limited research has been devoted to the second task in the modeling of crash data, assembling or formatting of the collected dataset.

Data cleaning is a major modeling step across various scientific disciplines; however, there are important attributes in crash data that make this task unique for safety analysis. As such, crash data often include excessive zero observations. As documented in Lord and Geedipally (*4*), excess zero observations are often attributed to how data are assembled or formatted in spatial or temporal scales. For example, it is expected to see more zero observations in data that are aggregated weekly than monthly or yearly. Finding a balance in aggregation is a critical task in data preparation. On the one hand, using disaggregated data may result in having excessive zero observations, in which the traditionally used negative binomial (NB) model may not be adequate for the safety analysis (*4*). On the other hand, too much aggregation may result in loss of information (*5*), although it may make the NB model a better alternative.

Several research studies have encountered this issue and there is no proper guidance on whether or not an aggregated data model is better than a disaggregated data model or vice-versa (*6*, *7*). In some of the previous studies, the researchers have relied on the goodness-of-fit of models for selecting the appropriate model; however, this comparison is inadequate and by its essence incorrect, since the nature and the size of datasets for the disaggregated and aggregated data can substantially be different. As a basic but critical principle, goodness-of-fit measures are only applicable to compare different models that applied to the 'same' dataset, and shall not be used to compare the performance of similar or different models that are applied to 'different' datasets. In short, it is not the comparison of goodness of fit statistics, but the reliability of the estimated coefficients that should be considered a priori when a model based on the aggregated data is investigated versus a disaggregated model.

In this study, we address the issue related to aggregated and disaggregated data by conducting a simulation study and measuring the information loss as a function of the precision or accuracy in estimation of the coefficients. The primary objectives of this study are therefore to (1) shed insights about the importance of the aggregating or formatting process of safety data on estimated coefficients of statistical model, and (2) explore decision criteria about the temporal and spatial aggregation of crash data with many zero observations. It should be pointed out that the criteria explored in this study only apply when the observations inside the dataset can be aggregated. If they cannot, more advanced models (see *1-3*) other than the NB should be considered a priori for modeling purposes.

The remaining parts of the paper are structured as shown below. First, the impact of data aggregation on estimated coefficients of statistical models is explored. Then, a simulation study is documented to determine the criteria for deciding when the aggregation of data is preferred. Next, the results of the simulation study are examined using empirical data. In the end, the results are summarized and avenues for further research are discussed.

2.0 BACKGROUND

In safety modeling, it is a common practice to use temporal aggregation when data from a few years are aggregated to develop a cross sectional model. Study duration (usually in "years") is considered as an offset variable in the regression model. Bonneson et al. (*8*) documented that one of reasons for preferring cross sectional data modeling (i.e., aggregated data) over panel data modeling (i.e. disaggregated data) is the accuracy of annual average daily traffic (AADT) in most highway safety databases. After examining states' databases and their documentation, the authors mentioned that the segment AADT volume is frequently extrapolated by the states from partial year counts taken at temporary count stations located several miles from the subject segment, which results in accuracy implications. In addition, when a current count is not available for a segment, transportation agencies sometimes adjust the AADT volume from the last year it was counted (which could be several years prior to the selected year) or sometimes just leave the variable as missing (*8*). Consequently, it is common for a segment's AADT volume to be missing for one or more years. Pratt et al. (*6*) presented three main advantages for using cross sectional data:

- It provides a more robust predictive model than panel data when the year-to-year variability in the independent variables is largely random.
- Fewer or no observations with missing values, since some operational features may not be collected every year.
- Using cross-sectional data for model calibration will minimize the problems associated with over-representation of segments or intersections with zero crash.

However, the first point above may not be always true. Variables such as road friction, pavement markings, and retroreflective devices degrade from one year to the next and their value in the current year highly correlates with the previous year's value. In such cases, the analyst cannot make a determination about preferring cross sectional data over panel data or vice versa. Panel data modeling has its own advantages (*9*):

- From a statistical perspective, the increase in the number of observations leads to a higher degree of freedom and less collinearity, which in turn improves the parameter estimation accuracy.
- It allows researchers to test whether or not more simplistic specifications are appropriate.
- The panel models can be used to analyze some specific questions, such as change in the variable effect over time that cannot be answered with cross-sectional modeling.

To evaluate the effect of skid resistance on traffic crashes, Pratt et al. (*6*) developed statistical models with both the aggregated and disaggregated data. Data from about 40,000 rural two-lane horizontal curves for a 5-year period in Texas were used. The authors used the aggregated data because the skid number variable was missing for a few years and there was an over-representation of horizontal curves with zero crash counts. In the aggregated dataset, the dependent variable used was the sum of crashes over a 5-year period and the independent variables were averaged over the time period. However, they noticed that the skid number variable changed significantly from one year to the next and a model was developed with the disaggregated data where each year was considered as a separate observation. Table 1 indicates the modeling results for the aggregated vs. disaggregated data. Note that the temporal correlation was evaluated, but was deemed to be negligible. As it is shown in this table, the coefficient of the skid number is significantly different between the aggregated model and the disaggregated model.

TABLE 1. Cross-Sectional and Panel Parameter Estimation for Two-Lane Highway Curves (*6*)**.**

Variable	Cross-Sectional Model		Panel Model	
	Estimate	Std. Err.	Estimate	Std. Err.
Intercept	−8.169	0.154	−7.862	0.236
LN (ADT)	0.790	0.019	0.760	0.027
Curve Radius	0.461	0.038	0.356	0.050
Lane Width	−0.040	0.017	−0.064	0.025
Shoulder Width	−0.041	0.006	−0.040	0.009
Skid Number	−0.005	0.001	−0.009	0.002
Annual Precipitation	0.015	0.002	0.014	0.002

A crash modification factor (CMF) was developed from two models. The equation for the CMF is the following (*6*):

$$\mathrm{CMF}_{\mathrm{SK}} = e^{\beta(\mathrm{SK}-40)}$$

where:

$\mathrm{CMF}_{\mathrm{SK}}$ = skid number crash modification factor.
SK = skid number.
β = estimated parameter.

Figure 1 shows the comparison of the CMF from the two models. According to the cross-sectional data model, an increase of skid number by 10 units will reduce the crash frequency by 5%. However, as per the panel data models, for the same change in the skid number, the crashes reduce by 9%. It is unclear to the analyst which model provided accurate result. Although the impact of temporal aggregation was the only type of aggregation analyzed, similar observations can be observed for spatial aggregation of data.

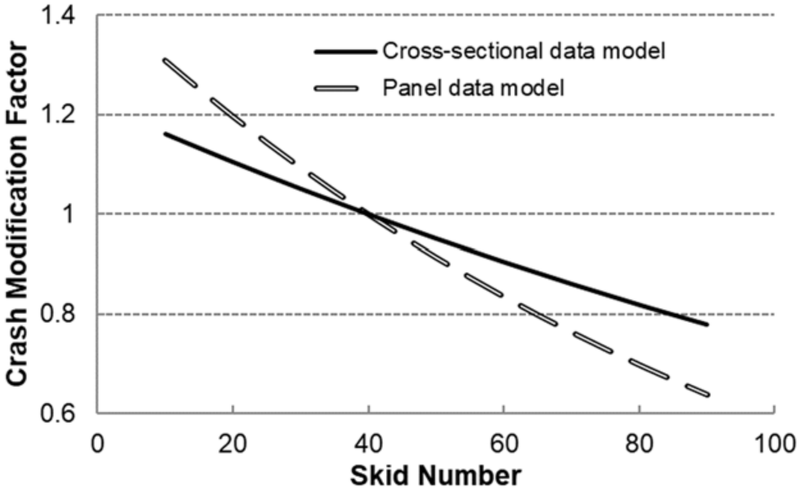
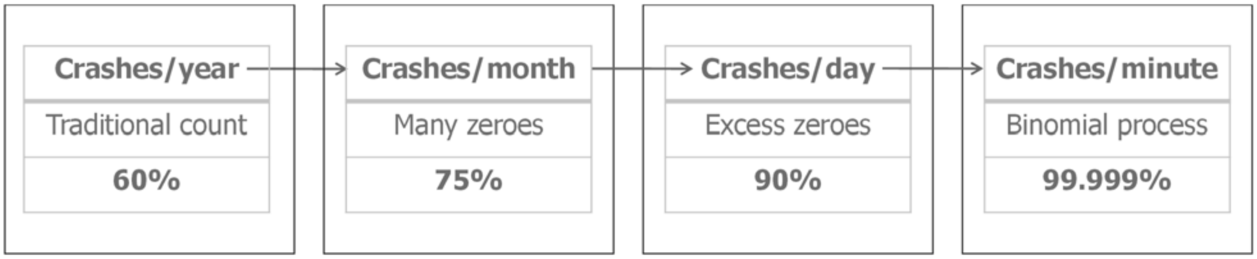


Figure 1. Skid Number Crash Modification Factor (*6*)**.**

Cafiso et al. (*7*) evaluated predictive models using different categories of spatial aggregations. One category consisted of grouping adjacent segments to reduce the number of short segments. The modeling results showed that some variables, such as the curvature change rate (CCR), provided very different estimates between the original and the aggregated segmentations, similar to what was observed above.

3.0 SIMULATION STUDY

Crash data at a site are usually defined as a count number over the space and time scales. Therefore, the number of zero observations in the compiled dataset is directly correlated with the selected spatial and/or temporal scales. By adjusting the time and spatial scales, the number of zero responses observed in the dataset can increase or decrease. For example, by changing the segment length of a site from 0.1 mile to 1 mile, the number of zero observations in the complied dataset will be reduced since the new segment will include all of the crashes on the segments now aggregated. Similarly, changing the time scale from monthly durations to yearly periods will result in reduction of number of zero responses in the dataset. This is depicted in Figure 2.



Crashes/year	Crashes/month	Crashes/day	Crashes/minute
Traditional count	Many zeroes	Excess zeroes	Binomial process
60%	75%	90%	99.999%

Figure 2. Percentage of zero responses when changing the time scale (*4*)**.**

In this section, a simulation study is executed to shed insights and explore criteria for the aggregation of data with numerous zero observations. First, the simulation protocol is introduced in detail. Next, a simulation study with several scenarios is performed to analyze highly dispersed data with different percentage of zero observations. The results of the simulation are discussed and a few recommendations are provided.

3.1 Simulation Protocol

The primary idea of the simulation study is related to the notion of information loss and the accuracy of model estimates upon aggregation of the safety data. The core idea of information loss can be explained by rehearsing the significant vs. insignificant terminologies. As a particular variable is increasingly aggregated, over the time and/or scale, its corresponded coefficient becomes less and less significant, due to the smaller variabilities observed in that variable; recursively, this setting can be continued until a final stage in which the aggregated variable becomes insignificant, and consequently no longer remains in the model. Given this notion, one may think that the disaggregated data are always preferred to the aggregated data; however, this scenario could only be impeccable if the parameters of the model could perfectly be estimated using the NB model. Although it is true in some circumstances, this situation is not viable when data include excessive zero observations. As it is widely known in safety studies, the NB model does not work well when data have numerous zero responses (*10-14*).

The opposite rationales described above creates a counter mechanism: On the one hand, aggregation of data could result in loss of information; on the other hand, the NB model could potentially estimate the parameter better when the dataset involves smaller amount of zero observations. We analyze this counter mechanism by measuring the information loss as a function of the precision or accuracy in estimation of the coefficients. Using a simulation study, we will examine different scenarios to find the aggregation decision point for different characteristics or scenarios; this stopping point or stage is referred to the situation that the coefficient of the model with the aggregated data (with smaller sample size but fewer number of zero observations) becomes less significant than the model with the disaggregated data (with larger sample size but greater zero observations).

The negative binomial distribution was used to simulate the observed crash data. The probability mass function (pmf.) function of the negative binomial distribution is structured as follows:

$$\mathrm{NB}(\mu,\varphi) \equiv \mathrm{P}(\mathrm{Y}=\mathrm{y}|\ \varphi,\mu) = \frac{\Gamma(\varphi+\mathrm{y})}{\Gamma(\varphi)\Gamma(\mathrm{y}+1)}\left(\frac{\varphi}{\mu+\varphi}\right)^{\varphi}\left(\frac{\mu}{\mu+\varphi}\right)^{\mathrm{y}}$$

where μ = mean response of observations, and φ = inverse dispersion parameter. Let us define the parameters and variables of the simulation protocol as follows:

x_{ij}^{m} = The value of the j-th covariate for i-th site at time period ‘m’.

φ^{m} = Inverse dispersion parameter at the time period ‘m’ calculated from real data.

y_{i}^{m} = simulated observation for the i-th site at each period ‘m’.

μ_{i}^{m} = mean response of the NB distribution for the i-th site at period ‘m’.

β_{j} = The true parameter for the j-th covariate (derived from a known model)

$\beta_{j}^{n^*}$= The estimated parameter for the j-th covariate at iteration ‘n’ of simulation.

The simulation is accomplished using three main steps that are summarized below:

Step 1 Initialization: In this step, the mean response of observations at each site ‘i' and time period ‘m’ is estimated using the following functional form $\mu_{i}^{m} = e^{\sum_{j=1}^{d} \beta_j x_{ij}^{m}}$. Here, the index $j = 1$ to d represents the independent variables. βs are called the “true” parameters taken from a known study.

Step 2 Simulation: In this step, crash counts are simulated using an NB distribution. First, a disaggregated dataset is created for each site and time period. Then, datasets are combined into one dataset for all time periods. Second, an aggregated dataset is created for each site i=1 to n. step 2 is repeated for “N” times (500 simulation runs for this study).

Step3 Comparison: For each simulation run, a negative binomial mode is used to estimate the coefficients of models for the disaggregated and aggregated datasets. The standard deviation of estimates from “N” runs of simulations is calculated and is used to compare the models with the aggregated data versus the disaggregated data.

The detailed steps of the simulation protocol are described below.

1. Initialization. Find the mean of crashes at each site ‘i’ as follows:

$$\mu_{i}^{m} = e^{\sum_{j=1}^{d} \beta_j x_{ij}^{m}}$$

Note: For the purpose of analysis, the x_{ij}^{m} with “NA” values are recommended to be replaced with $\min_{m}(x_{ij}^{m}) + (\max_{m}(x_{ij}^{m}) - \min_{m}(x_{ij}^{m}))/m$ in this step. However, the records with ‘NA’ values eventually should be removed in Step 2.2.1.3 and Step 2.2.2.1.

2. Simulation. Repeat the following steps for ‘N’ times:

2.1 Simulate the observation at each site $i = 1$ to n at the m-th period from the NB distribution as follows:

$$y_{i}^{m} \sim \mathrm{NB}(\mu_{i}^{m}, \varphi^{m})$$

2.2 Creating the experiment datasets.

2.2.1 Create the disaggregated dataset (D_1)

```
            2.2.1.1 Create the datasets $D^m$ at each period ‘m’, with $(y_i^m, x_{ij}^m)$ elements (where the index ‘i' denote a row and ‘j’ a column of the dataset).
            2.2.1.2 Merge the $D^m$ datasets into a single dataset $D_1$.
            2.2.1.3 Remove the records of $D_1$ that include an ‘NA’ value.
            2.2.1.4 Shuffle the records in $D_1$.
        2.2.2 Create the aggregated dataset ($D_2$):
            2.2.2.1 Find $\bar{x}_{ij} = \underset{m}{\text{mean}}\, x_{ij}^m$ (exclude $x_{ij}^m$ with the “NA” values when $\bar{x}_{ij}$ is calculated).
            2.2.2.2 Create the $D_2$ dataset with $(\sum_m y_i^m, \bar{x}_{ij})$ elements (where the index ‘i' denote a row and ‘j’ a column of the dataset).
            2.2.2.3 Shuffle the records in $D_2$.
    2.3 Refitting the simulated datasets
        2.3.1 Fit an NB GLM to $D_1$ and record the estimated coefficients in $\beta_j^{n^*}(D_1)$.
        2.3.2 Fit an NB GLM to $D_1$ and record the estimated coefficients in $\beta_j^{n^*}(D_2)$.
3. Comparison.
    3.1 For each j-th covariate, find the standard deviation of the estimated coefficients over ‘n’ iterations and denote them by $\beta_j^{std}(D_1)$ and $\beta_j^{std}(D_2)$.
    3.2 Compare $\beta_j^{std}(D_1)$ and $\beta_j^{std}(D_2)$, the one with a smaller value indicates a more reliable implementation.
```

3.2. Simulation Results

The rural two-lane horizontal curve dataset used by Pratt et al. (*6*) was obtained for the simulation. Two variables, average daily traffic (ADT) and skid number, from this dataset were used for the simulation analysis. These two variables were considered for the analysis since unlike the most safety data, the value of ADT and skid number often changed over time. The variables were collected for 5 years, in one-year durations. Only the horizontal curves that had skid number recorded for at least 3 out of 5 years were considered. In other words, for some sites, the data for skid number is missing or is incomplete. This particular feature is important for the simulation analysis, since it introduces the primary challenge of using the aggregated versus the disaggregated data. Given different scenarios, the analyst has two principle alternatives to model the data: either take the average of the available skid numbers over 5 years (say 3 out of 5), and use the results as one record, or alternatively keep the disaggregated data, but remove the records that are incomplete. In our simulation analysis, the value of the inverse dispersion parameter, for each year ‘m’ (φ^m), was directly calculated from the observed crash data. The average value of φ^m over the five years is around 0.2, which means that the data are highly dispersed.

Two major scenarios for highly dispersed data were created: 1) data that involve around 90% of zero observations and 2) data with 50% of zero observations. Then, each major scenario is divided into 7 sub-scenarios, based on year-to-year variation of the skid number. The sub-scenario (1-1) only includes records that the skid number variation from year-to-year is always less than 20%. Recursively, the sub-scenario (1-2) assumes 30% variation, the sub-scenario (1-3) assumes 40% variation, etc. The last sub-scenario (sub-scenario 1-7) includes the full data. Table 2 and Table 3 provide the results of the simulation study for different scenarios. Note that even though the ADT variable is used in the analysis, those results are not presented here because the primary focus is on the skid number variable that introduces a greater variability in the model. For each sub-scenario, the change in coefficient of variation (CV) of the skid number variable upon aggregation is also measured and shown in the Table 2 and Table 3. For instance, in sub- Scenario 1-3, the difference between the CVs of skid number in the aggregated versus the disaggregated datasets is equal to 6.8%. In other words, in this scenario, the CV or variability of skid number is reduced by 6.8% once data was aggregated.

TABLE 2. Simulation Results for Scenario with About 90% Zeros.

Scenario	True Value	Skid Number Year to Year Variation	CV_{Skid}[2] Change	Disaggregated Data (Zeros = 90.07%; Crash Mean = 0.163)		Aggregated Data (Zeros = 61.30%; Crash Mean = 1.933)	
				Mean	Std.[3]	Mean	Std.[3]
1-1	-0.005914	< = 20% (n$_1$=1570; n$_2$=6270)[1]	0.1%	-0.005590	0.004308	-0.005649	**0.003838**
1-2	-0.005914	< = 30% (n$_1$=2410; n$_2$=9602)	3.6%	-0.005939	0.003162	-0.005963	**0.002878**
1-3	-0.005914	< = 40% (n$_1$=3112; n$_2$=12368)	6.8%	-0.005854	0.002601	-0.005965	**0.002510**
1-4	-0.005914	< = 50% (n$_1$=3664; n$_2$=14528)	10.5%	-0.006039	**0.002219**	-0.006011	0.002278
1-5	-0.005914	< = 60% (n$_1$=4042; n$_2$=16047)	14.0%	-0.005944	**0.002213**	-0.005886	0.002257
1-6	-0.005914	< = 80% (n$_1$=4295; n$_2$=17083)	17.1%	-0.005827	**0.002050**	-0.005960	0.002153
1-7	-0.005914	Full Data (n$_1$=4402; n$_2$=17504)	18.7%	-0.005945	**0.001913**	-0.005898	0.002291

[1] n$_1$ = sample size of the aggregated data, n$_2$ = sample size of the disaggregated data.

[2] CV_{Skid} change denotes the change in coefficient of variation of skid number variable after aggregation.

[3] Bold numbers represent the preferred values.

TABLE 3. Simulation Results for Scenario with About 50% Zeros.

Scenario	True Value	Skid Number Year to Year Variation	CV_{Skid}[2] Change	Disaggregated Data (Zeros = 50.04%; Crash Mean = 9.82)		Aggregated Data (Zeros = 3.81%; Crash Mean = 49.15)	
				Mean	Std.[3]	Mean	Std.[3]
2-1	-0.005914	< = 20% (n_1=1570; n_2=6270)[1]	0.1%	-0.005939	0.002983	-0.005950	**0.002764**
2-2	-0.005914	< = 30% (n_1=2410; n_2=9602)	3.6%	-0.005789	0.002152	-0.006018	**0.001996**
2-3	-0.005914	< = 40% (n_1=3112; n_2=12368)	6.8%	-0.005843	**0.001639**	-0.005996	0.001662
2-4	-0.005914	< = 50% (n_1=3664; n_2=14528)	10.5%	-0.005882	**0.001586**	-0.006043	0.001644
2-5	-0.005914	< = 60% (n_1=4042; n_2=16047)	14.0%	-0.005899	**0.001422**	-0.006045	0.001484
2-6	-0.005914	< = 80% (n_1=4295; n_2=17083)	17.1%	-0.005925	**0.001401**	-0.005987	0.001503
2-7	-0.005914	Full Data (n_1=4402; n_2=17504)	18.7%	-0.005884	**0.001275**	-0.005982	0.0014620

[1] n_1 = sample size of the aggregated data, n_2 = sample size of the disaggregated data.
[2] CV_{Skid} change denotes the change in coefficient of variation of skid number variable after aggregation.
[3] Bold numbers represent the preferred values.

By comparing the standard deviation of estimates (derived from n=500 runs of simulation), one can observe that, initially, when the variability of the skid number (measured by the change in CV) is reduced at smaller rates, the model with aggregated data provides better estimates comparing to the disaggregated data. This observation is compatible to our premise that data with fewer number of zero observations fits the NB model better. However, as the variability of the skid number is reduced by higher rates, the model based on the aggregated data becomes less reliable than the one with the disaggregated dataset. This observation is also compatible to our premise of loss of information caused by too much aggregation.

Recall that the simulation study was designed to study a ‘counter mechanism’. On the one hand, aggregation could result in fewer number of zero observations that makes the NB model more reliable; on the other hand, too much aggregation will result in loss of information, and consequently erroneous estimates. Our simulation study sought to reveal the decision criteria. The decision point can be quantified by the reduction in variability of the dataset, measured by the change in coefficient of variation (CV) of the variables in the dataset once data is aggregated. For example, in Scenario 1-3, that sought to reveal the decision criteria for highly dispersed data with 90% zero observation, the change in CV of the skid number when data are aggregated is equal to 6.8%. Aggregation up to this point seems appealing as the model with aggregated data can provide better estimates than the one with disaggregated data. After this point, the model based on the disaggregated data is preferred. In that regard, it seems that a change in CV by 7% in a variable is a decision point to stop the aggregation. On the other hand, when the percentage of zero observations is small earlier, the aggregation can be stopped when the change in CV of a variable is greater than 4%. Given the simulation results, the following conservative criteria are recommended:

- When the percentage of zeros is higher 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 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%.

Although the simulation analysis was only studied for the temporal aggregation of data, similar recommendations can be generalized to the spatial aggregation. In this case, the analyst could create different aggregation scenarios based on the dataset in hand, and then measure the change in variability of each variable in the dataset, by calculating the change in their CV upon aggregation. For example, imagine that the analyst wants to combine adjacent sites with close ADT values. In that regard, multiple scenarios can be constructed, such as scenario 1: combine adjacent segments with the ADT within +10%, scenario 2: combine adjacent segments with the ADT within 20% and so on. Next, after aggregation, the reduction in the CV of variables (all variables in the model), compared to the full disaggregated data, is calculated. Once the change in CV is derived, the above recommendations can be used to pick the optimal aggregation scenario. In the next section, we demonstrate this procedure with an example.

4.0 CASE STUDIES

For the spatial aggregation case study, we obtained the Texas Interstate database that Geedipally et al. (*15*) used for identifying high-risk segments based on Fatal (K) and Incapacitating injury (A) crashes. Similar to the analysis Geedipally et al. (*15*) conducted, we aggregated the adjacent segments when the change in the ADT was less than a certain threshold and the segments were on the same highway and all other variables remain the same. We calculated the CV of the ADT in disaggregated and aggregated datasets and estimated the difference in CV values. With the aggregation, the sample size and percentage of zeros as well as the CV of ADT variable are reduced, as shown in Table 4. Since the disaggregated data had about 50% zeros, the simulation results suggest stopping the aggregation when the change in CV is above 4%. As per the simulation results, it is recommended to use the aggregation suggested in scenario 4 and stop the aggregation when the change in ADT is 25% or less between the adjacent segments.

TABLE 4. Spatial Aggregation of Interstate Segments

Aggregation Scenario	Aggregation Criteria	Number of segments	Percentage of sites with no crashes	CV_{ADT}	Change in CV_{ADT}
0	Existing	2321	54%	0.58	--
1	ADT within +10%	519	25%	0.57	2%
2	ADT within +15%	483	23%	0.56	3%
3	ADT within +20%	463	23%	0.56	3%
4	ADT within +25%	451	22%	0.56	3%
5	ADT within +50%	426	22%	0.55	5%

The example presented in the background section is used as a case study for temporal aggregation. As discussed previously, Pratt et al. (*6*) developed statistical models with both the disaggregated and temporally aggregated data to evaluate the effect of skid resistance on traffic crashes. For this analysis, two scenarios were considered, as shown in Table 5. First, we considered all sites even if the skid number variable is missing for some years. In the aggregated data, the skid number variable is the average over time but excluding the missing years' data. For example, if the data are missing for two years and available for three years, then the skid number value in the aggregated data is the average of three years. This means, in the disaggregated data, those missing years were excluded. Second, we considered only those sites where the skid number variable is available in all five years. The sample size of the disaggregated dataset is five times as that of the aggregated dataset. Since this dataset had more than 90% zeros, for the first scenario, it is recommended to use the aggregated data, as shown in the last column of Table 5, because the change in CV of skid number is 6.2%, which is the less than the 7% threshold recommended above. However, for the second scenario, disaggregated data is recommended for the model development because the change in CV is greater than the 7% threshold.

TABLE 5. Temporal Aggregation of Crashes on Horizontal Curves

Scenario	Data type	CV_{Skid}	Change in CV_{Skid}	Preferred data
I	Disaggregated	0.318	--	Aggregated
	Aggregated	0.299	6.2%	
II	Disaggregated	0.306	--	Disaggregated
	Aggregated	0.258	15.6%	

5.0 SUMMARY AND CONCLUSIONS

Crash data have unique characteristics not found with datasets used in other types of research. One of these characteristics is related to datasets with a large percentage of zero responses. This issue is often directly related to how the data is assembled or formatted. By adjusting the time and spatial scales, the number of zero responses observed in the dataset can either increase or decrease; however, too much aggregation could result in loss of information and erroneous estimates. This study performed extensive simulation analyses to study this counter mechanism, and shed insights on when aggregation of data is beneficial. When highly dispersed datasets have a large percentage of zero responses (50% or above), the recommendations are as follows:

- When the percentage of zeros is higher 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 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%.

As discussed earlier, the recommendations described above should only be applied if the data can be properly aggregated. If they cannot, then the analyst may use one of the models proposed to model data with a large percentage of zeros (*10-14*). In that regard, recent research studies presented a methodology to design heuristics for model selection based on characteristics of data, such as the percentage of zero observations in the dataset (*12, 16*). Although this paper explored useful criteria for formatting of the data with excess zero observations, further research is recommended to explore greater simulation scenarios, and to examine if statistics other than the change in the coefficient of variation of the independent variables can be used to determine when aggregated data should be used over disaggregated data.

ACKNOWLEDGMENTS

Support for this research was provided in part by a grant from the U.S. Department of Transportation, University Transportation Centers Program to the Safety through Disruption University Transportation Center (451453-19C36). [Disclaimer: The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report was funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.]

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: Mohammadali Shirazi, Srinivas Reddy Geedipally, Dominique Lord; data preparation: Mohammadali Shirazi, Srinivas Reddy Geedipally, Dominique Lord; analysis and interpretation of results: Mohammadali Shirazi, Srinivas Reddy Geedipally, Dominique Lord; draft manuscript preparation: Mohammadali Shirazi, Srinivas Reddy Geedipally, Dominique Lord. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Lord, D., & Mannering, F. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation research part A: policy and practice, 2010, 44(5), 291-305.
2. Mannering, F. L., & Bhat, C. R. Analytic methods in accident research: Methodological frontier and future directions. Analytic methods in accident research, 2014, 1, 1-22.
3. Mannering, F. L., Shankar, V., & Bhat, C. R. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic methods in accident research, 2016, 11, 1-16.
4. Lord, D., & Geedipally, S. R. Safety Prediction with Datasets Characterised with Excess Zero Responses and Long Tails. In Safe Mobility: Challenges, Methodology and Solutions (pp. 297-323). Emerald Publishing Limited, 2018.
5. Usman, T., Fu, L., & Miranda-Moreno, L. F. Accident prediction models for winter road safety: Does temporal aggregation of data matter? Transportation Research Record, 2011, 2237(1), 144-151.
6. Pratt, M.P., Geedipally, S.R., Wilson, B., Das, S., Brewer, M., and Lord, D. Pavement Safety-Based Guidelines for Horizontal Curve Safety. TxDOT 0-6932. Texas A&M Transportation Institute, College Station, TX, 2018.
7. Cafiso, S., D'Agostino, C., and Persaud, B. Investigating the influence of segmentation in estimating safety performance functions for roadway sections. Journal of Traffic and Transportation Engineering, 2018, 5(2), 129-136.
8. Bonneson, J.A., S.R. Geedipally, M.P. Pratt, D. Lord. Safety Prediction Methodology and Analysis Tool for Freeways and Intersections. 476440-1, National Cooperative Highway Research Program, Washington, DC, 2012.
9. Washington, S.P., Karlaftis, M.G., Mannering, F.L. Statistical and Econometric Methods for Transportation Data Analysis, second ed. Chapman Hall/ CRC, Boca Raton, FL, 2010.
10. Geedipally, S. R., Lord, D., & Dhavala, S. S. The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. Accident Analysis & Prevention, 2012, 45, 258-265.
11. Shirazi, M., Lord, D., Dhavala, S. S., & Geedipally, S. R. A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: characteristics and applications to crash data. Accident Analysis & Prevention, 2016, 91, 10-18.
12. Shirazi, M., Dhavala, S.S., Lord, D. and Geedipally, S.R. A methodology to design heuristics for model selection based on the characteristics of data: Application to investigate when the Negative Binomial Lindley (NB-L) is preferred over the Negative Binomial (NB). Accident Analysis & Prevention, 2017, 107, pp.186-194.

13. Shaon, M. R. R., Qin, X., Shirazi, M., Lord, D., & Geedipally, S. R. Developing a Random Parameters Negative Binomial-Lindley Model to analyze highly over-dispersed crash count data. Analytic methods in accident research, 2018, 18, 33-44.

14. Lord, D, S.R. Geedipally, F. Guo, A. Jahangiri, M. Shirazi, X. Deng. Analyzing Highway Safety Datasets: Simplifying Statistical Analyses from Sparse to Big Data, Safe-D University Transportation Center, 2019 (Forthcoming)

15. Geedipally, S., M. Martin, R. Wunderlich, D. Lord. Highway Safety Improvement Program Screening Tool. Technical Memorandum. Traffic Operations Division, Texas Department of Transportation, 2017.

16. Shirazi, M., & Lord, D. Characteristics Based Heuristics to Select a Logical Distribution between the Poisson-Gamma and the Poisson-Lognormal for Crash Data Modelling. Transportmetrica A: Transport Science, 15 (2), 2019, 1791-1803.