Exploring Clusters of Contributing Factors for Single Vehicle Fatal Crashes through Multiple Correspondence Analysis

Subasish Das and Xiaoduan Sun
Department of Civil Engineering, University of Louisiana at Lafayette

Research Question

What are the key contributing factors for single vehicle fatal crashes? What's the purpose of using multiple correspondence analysis (MCA) in place of conventional parametric approaches?

Abstract

A single vehicle crash can be caused by various factors such as those related to roadway design, vehicle mechanical problems and, most importantly, the driver performance or behavior. The conventional crash analysis methods lack the ability to identify the cluster of factors simultaneously. The Multiple Correspondence Analysis (MCA) method used in this study helps to visualize the patterns of the cluster consisting of crash contributing factors for single vehicle fatal crashes. The results identify the key association factors which can guide the the selection process of crash countermeasures.

Introduction

In Louisiana, nearly 60 percent of roadway fatalities result from ROR crashes. In 2012 and 2011, 384 and 330 out of a total of 652 and 630 fatal crashes, respectively, were single vehicle crashes in the state. In fact, it is well known that single vehicle run-off-road (ROR) crashes are usually caused by a combination of factors that could have come from inadequate roadway design, vehicle problems, environment conditions and/or poor performance or behavior of the driver. The combination of factors could be spatially and temporally different. Failure to recognize the combination of these factors could possibly lead to insufficient or ineffective actions taken intended to reduce the number of ROR crashes.

The commonly used crash analysis methods and safety performance models are not capable of identifying the combination or cluster of factors simultaneously. MCA is a useful technique in exploring the structure of many categorical data because it can present complicated relationships in a simple chart that demonstrates a combination or cluster of influential variables through the reduced data dimension analysis. In summary, this technique generates the cloud of points and individual records to assess the correlation between the variables and their relationship to the interested resultant variable.

The research introduced in this paper is unique and different from previous studies and serves as a starting point to demonstrate the use of MCA to determine the significant cloud of crash contributing factors for single vehicle fatal crashes which could help state agencies develop the most efficient crash countermeasures.

Multiple Correspondence Analysis

Multiple Correspondence Analysis (MCA) is widely used in categorical data analysis, especially in social science and marketing research, and is considered as an extension of correspondence analysis (CA) to more than two variables. It is an exploratory data analysis method that can visualize the patterns of the cluster consisting of key contributing factors.

- For a database with categorical variables, the scheme of the MCA method can be explained by taking an individual row or record i as an example where three variables (represented by three columns) have three different category indicators (a1, b1, and c3).
- The spatial distribution of the points calculated by the dimensions based on these three categories would be generated by MCA. MCA yields two clouds of points as shown in the figure: the cloud of individual records and the cloud of categories.
- A cloud of points is not just a simple "graphical display", it is similar to a geographic map with the same distance scale in all directions.

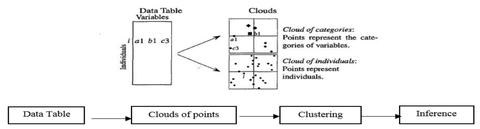

Figure 1: Clouds of points generated by MCA method.

Data Description

Eight years (2004-2011) of Louisiana crash data was used in this study.

- All variables underwent the standard pre-processing and distribution testing by examining the relevance of missing values. The final data set contains complete cases for 21 variables. Correlation of the variables are performed and reduction of the variables are also based on correlation plots.
- Some of these variables, such as drug involvement, alcohol involvement and occurrence in intersection, have logical values such as yes or no and true or false. Driver Age is a continuous variable. Since MCA mainly deals with qualitative data, the quantitative variable "age" is transformed into seven categories. The other variables are nominal in nature.
- An initial analysis indicates that some variables are highly skewed which means that a majority of crashes fall into one of the two or more categorical values. For example, 94 percentage of the crashes involved a driver with no drug intoxication, 94 percentage of the crashes occurred on normal roadway conditions, 85 percentage of crashes had no vehicle defects observed, and 86 percentage of crashes occurred on dry surface conditions.
- The non-skewed variables include alcohol involvement, day of the week, vehicle type, road type, driver age, lighting condition, and crash time.

Major Findings

The results of the MCA can guide the selection of crash countermeasures. The future work on the degree of association of the crash contributing factors can help safety management systems identify the most effective crash reduction strategies.

Cloud of points

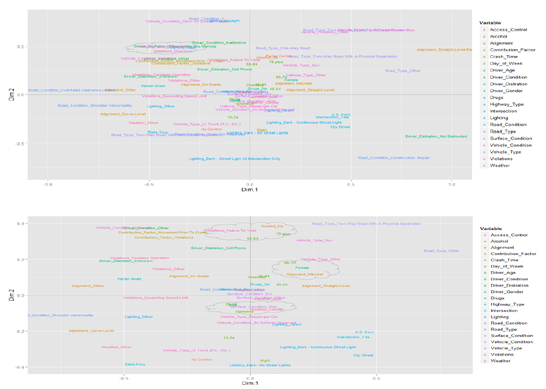

Figure 2: MCA plot for the variable categories.

MCA Factor Map

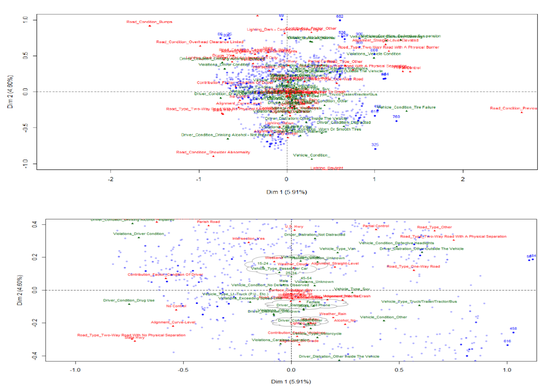

Figure 3: MCA factor map.

Discussion

Points (categories) that are close to the "mean" are plotted near the MCA plot's origin and those that are more distant are plotted farther away. Categories with a similar distribution are presented near one another, while those with different distributions are farther apart. This gives the idea of clustering of key variables associated with a particular pattern.

- The cluster selection is based on the relative closeness of the category location in the MCA plot.
- Few interesting cluster groups were identified by using the MCA method. A conclusion drawn from one of the few interesting cluster groups is that alcohol and drug impaired driving is crash prone. Another is that two-way-roads with no physical separation in the dark with no street lights is a crash-prone condition. Motorcycle drivers are seen in the same group where dusk and outside distraction are present.
- Failure to yield is found as a particular limitation in older sport utility vehicle (suv) drivers. Another cluster shows that young male passenger car drivers are more crash-prone on wet surfaces during weekends. It is also found that older female drivers (65-74) are observed as crash-prone in hilly aligned zones. Moreover, older drivers in general are seen to have difficulties driving at dusk on weekdays and are also seen to be involved in crashes due to failure to yield.

Conclusion

All of the parametric regression models contain their own model assumptions and pre-defined underlying relationships between dependent and independent variables. If these assumptions are violated, the model could lead to erroneous estimations. MCA, a non-parametric approach without any pre-defined underlying relationship between the dependent and the predictor variables, has been widely employed in social sciences and marketing research for large sets of categorical data analysis.

In conclusion, this study uses a new method to investigate contributing factors to single vehicle fatal crashes which examines the crash attributes (variables) using the MCA technique. At a theoretical level, it answers recent calls to investigate into the actual on-site mechanisms of fatal crashes using the MCA method. At an empirical level, the findings presented here shed light on the pattern recognition of traffic crashes and expose new facets in the current crash analysis. Further research can be done by applying joint correspondence analysis and other non-parametric approaches in order to find the dominant contributing factors.

UNIVERSITY OF LOUISIANA Lafayette