ABSTRACT

In 2013, 346 out of a total of 616 fatal crashes in Louisiana were single vehicle crashes. In order to reduce the number of these crashes through effective crash countermeasures, safety policies, regulations and technological advancements, it is important to identify the key factors associated with single vehicle Run-Off-Road (ROR) crashes. The persistently high rate of ROR crashes in Louisiana, as well as in the overall United States, calls for research that can further benefit the on-going research in assessing the performance of roads, vehicles and humans. This research uses Multiple Correspondence Analysis (MCA), an exploratory data analysis method that associates a combination of factors based on their relative distance in a two dimensional plane, to analyze eight years (2004-2011) of single vehicle fatal crashes in Louisiana. Important contributing factors and their degree of association were measured using this method. The results revealed that drivers of lightweight trucks driving on undivided state highways, male drivers driving passenger-cars at dawn, older female drivers (65-74) driving non-passenger cars, older drivers facing hardship to yield in partial access control zones, and drivers with poor reaction time due to impaired driving were the main drivers associated with ROR crashes.

Results of the MCA can guide the selection method of crash countermeasures. The future work on the degree of association of the identified crash contributing factors can help safety management systems select the most effective and efficient crash reduction strategies.

Key words: road safety, single vehicle crashes, fatality, multiple correspondence analysis, cloud of groups, combination.

INTRODUCTION

Most single vehicle crashes are ROR crashes which are more likely than other types of crashes to result in fatalities and severe injuries [1]. In 2012 and 2013, 384 and 346 out of a total of 652 and 616 fatal crashes, respectively, were single vehicle crashes in the state [2]. Prior studies have identified roadway, vehicular and environmental factors for single vehicle crashes [3-5]. In fact, it is well known that single vehicle ROR crashes are usually caused by a combination of factors that could have come from inadequate roadway design, vehicle problems, environmental conditions and/or drivers' poor performance. The combination of factors could be spatially different, i.e. crashes occurring on highways verses intersections, and temporally different, i.e. crashes occurring in December versus those in May. 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.

Identifying crash-prone factors and combinations of factors by analyzing a large dataset is not a trivial task. The commonly used statistical inferential methods, i.e. ANOVA, and safety performance models are not capable of identifying the combination of factors simultaneously. Multiple Correspondence Analysis (MCA) is considered as an extension of Correspondence Analysis (CA) for more than two variables and is thus widely used in categorical data analyses, especially in social science and marketing research [6]. By using this technique we can visualize the patterns of the combination of crash contributing factors. MCA helps to discover the structure of categorical data by presenting complicated relationships in a simple chart that demonstrates a combination of significant variables through the reduced data dimension analysis. This method presents the correlation between the variables and their relationship to the interested resultant variable by grouping them.

The persistently high rate of fatal single vehicle crashes in Louisiana and the overall United States indicates the continuous need for research. Reducing single vehicle crashes is critical in fulfilling the state's "Destination Zero Deaths" goal and the MCA technique used in this paper will help to find out the key factors of fatal single vehicle crashes so that necessary actions can be taken to reduce crash frequencies and severities.

MCA STUDIES

I.P. Benzecri developed MCA, a statistical approach based on the CA method. MCA has since become a widely considered as the multivariate generalization of Correspondence Analysis. Since then, MCA has been considered one of the main standards of geometric data analysis (GDA) in the field of social science and marketing research. GDA is also referred to as the Pattern Recognition Method which treats arbitrary data sets as clouds of points in n-dimensional space. However, in the field of multivariate transportation data analysis, geometric methods have rarely been used. MCA is hardly utilized in traffic crash analysis. In fact, Roux and Rouanet pointed out in their book that this method, while it is a powerful tool for analyzing a full-scale research database, is still hardly discussed and therefore under-used in many promising fields [6].

Hoffmann and De Leeuw used MCA as a multidimensional scaling method to show how questions posed of categorical marketing research data may be answered with MCA in terms of significant meaningful results [7]. Fontaine was the first to use MCA analysis for a typological analysis of pedestrian related crashes [8]. The classification of pedestrians involved in crashes is divided into four major groups. The typology produced by this analysis reveals correlations between criteria, without necessarily indicating a "causal link" with the crashes. The resulting typological breakdown served as a basis for in-depth analysis to improve the understanding of these crashes and propose necessary strategies. Golob and Hensher used MCA to establish causality of nonlinear and non-monotonic relationships between socioeconomic descriptors and measures of travel behavior [9]. Factor et al. conducted a study on the systematical exploration of the homology between drivers' community characteristics and their involvement in specific types of vehicle crashes [10].

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

METHODOLOGY

Theory

For a database or a table with categorical variables, the scheme of the MCA technique can be explained by taking an individual record (in row), *i,* where three variables (represented by three columns) have three different category indicators (a_1, b_2, and c_3). 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 Figure 1: the cloud of individual records and the cloud of categories. A cloud of points is not just a simple "graphical display", it can be compared with a geographic map with the same distance scale in all directions. A geometric diagram cannot be strained or contracted along one specific dimension. Thus, a basic property of a cloud of points is known by its dimensionality. The one-dimensional cloud is a simple version whose points lie on a single line. The two-dimensional cloud is also convenient where points lie on a plane. In MCA, the clouds of categories and of individuals have the same dimensionality which is occasionally high. The full clouds are referred to by their principal dimensions (1, 2, 3, etc.) which are ranked in descending order of importance. The goal of MCA is to create a combination of groups put together from the large dataset. The flowchart of the MCA procedure is also exhibited in Figure 1. The cloud of categories and the cloud of individual records are considered as the cloud of points.

If *Q* represents the number of variables and *I* the number of records, the data matrix will be an *"I by Q"* table with all categorical values. If J_q is the number of categories for variable *q*, the total number of categories for all variables is $J = \sum_{q=1}^{Q} J_q$. To contain all categories in the data table, another data matrix is developed as *"I by J"* where each variable will have several columns to show its possible categorical values. For example, for variable drug involvement there are two columns: one for "yes" and another for "no". If an individual crash record indicates "no" drug problem in this particular crash, the "yes" column will have 0 and the "no" column will have 1. The number of categories for this variable is two.

Suppose, the number of individual records associated with category k is denoted by n_k (with $n_k > 0$), where $f_k = n_k/n$ is the relative frequency of individuals who are associated with category k. The values of f_k will generate a row profile. The distance between two individual records is created by the variables for which both have different categories. Suppose that for variable q, individual record i contains category k and individual record i' contains category k' different from k. The part of the squared distance between individual records i and i' for variable q is defined by

$$d_q^{\,2}(i,i') = \frac{1}{f_k} + \frac{1}{f_{k'}} \qquad (1)$$

Denoting Q as the number of variables, the overall squared distance between i and i' is defined by

$$d^2(i,i') = \frac{1}{Q}\sum\nolimits_{q\in Q} d_q^{\,2}(i,i') \qquad (2)$$

The set of all distances between individual records determines the cloud of individuals consisting of n points in a space whose dimensionality is L, with $L \leq K - Q$ (overall number K of categories minus number Q of variables), and assuming $n \geq L$. If M^i denotes the point representing individual i and G which is the mean point of the cloud, the squared distance from point M^i to point G is

$$(GM^i)^2 = \frac{1}{Q}\sum\nolimits_{k\in K_i} \frac{1}{f_k} \qquad (3)$$

where, K_i denotes the response pattern of individual i; that is, the set of the Q categories associated with individual record i.

The cloud of categories is a weighted cloud of K points. Category k is represented by a point denoted by M^k with weight n_k. For each variable, the sum of the weights of category points is n, hence for the whole set K the sum is nQ. The relative weight p_k of point M^k is $p_k = n_k/(nQ) = f_k/Q$; for each variable, the sum of the relative weights of category points is 1/Q, hence for the whole set the sum is 1.

$$p_k = \frac{n_k}{nQ} = \frac{f_k}{Q} \qquad with \sum\nolimits_{k\in K_q} p_k = \frac{1}{Q} \qquad and \sum\nolimits_{k\in K} p_k = 1$$

If $n_{kk'}$ denotes the number of individual records which have both of the categories *k* and k', then the squared distance between M^k and $M^{k'}$ is given by the formula

$$(M^k M^{k'})^2 = \frac{n_k + n_{k'} - 2n_{kk'}}{n_k n_{k'} / n} \quad (4)$$

The numerator is the number of individual records associating with either *k* or k' but not both. For two different variables *q* and q', the denominator is the familiar "theoretical frequency" for the cell (*k*, k') of the $K_q \times K_{q'}$ two-way table.

The actual computations in MCA are not performed on a design or indicator matrix but on the inner product of this matrix which is also known as the 'Burt Table'. The MCA product in this paper was performed by using open source statistical 'R Version 3.02' software [11]. Various R packages like 'ca', 'FactoMineR', 'ade4', 'MASS', 'homals' are available to perform the MCA technique. This study used the 'FactoMineR' package (for its convenient functions) to analyze the dataset [12]. The datasets were studied according to the individual records, the variables and the categories. In this study, the primary focus was on studying the categories, as categories represent both variables and a group of individual records.

Initial Data Analysis

To identify important contributing factors to fatal single vehicle crashes in Louisiana, eight years (2004-2011) of crash data was obtained from the Louisiana Department of Transportation and Development (LADOTD). The primary dataset was prepared by merging three different tables (the crash table, Department of Transportation and Development (DOTD) table and the vehicle table) from the Microsoft Access dataset. For any given individual crash record, there are 371 possible explanatory variables (153 from the crash table, 40 from the DOTD table and 178 from the vehicle table). Figure 2 displays the annual fatal single vehicle crashes by year in Louisiana which indicates that there was a 4% increase in these crashes between 2010 and 2011 and that the highest number of ROR fatal crashes was in 2007.

In the crash database, there are numerous variables that are redundant for this research, such as: the VIN, driver's license number, database manager's name, police report number etc. These unnecessary variables are omitted by using engineering judgment. The pre-final list of variables is primarily scanned by examining the relevance of missing values (by developing a correlation matrix) and the relevance of the distribution skew. Datasets with variables bearing over 70 percent of missing values makes the MCA plots less informative. The final dataset contains complete cases for 21 variables. The summary of the selected variable counts is displayed in Table 1A and Table 1B where the variables are grouped by:

- Human factor related (driver age, intoxication, condition of the driver, violation type, driver distraction, driver gender, driver injury)
- Crash characteristics (crash year, crash hour, day of the week, collision type)
- Roadway related (access control, alignment, lighting condition, road condition, road type, intersection, surface condition, highway type)
- Environment related (weather)
- Vehicle related (vehicle condition, vehicle type)

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: 15-24 years old, 25-34 years old, 35-44 years old, 45-54 years old, 55-64 years old, 65-74 years old and 74 plus. The other variables are nominal in nature. The number of categories in each selected variable is presented in Table 2.

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% of the crashes involved a driver with no drug intoxication, 94% of the crashes occurred on normal roadway conditions, 85% of crashes had no vehicle defects observed, and 86% 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.

Multiple Correspondence Analysis

Graphical representations are considered an easier way to perceive and interpret data because they effectively summarize large, complex datasets by simplifying the structure of the associations between variables and providing a universal view of the data [6]. Morphological maps are a way of presenting information graphically and are interpreted by looking at groupings of variables in space. 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 by forming combination, while those with different distributions are farther apart. Hence, the dimensions (axes) are interpreted by the position of the points on the map, using their loading over the dimensions as crucial indicators. A two-dimensional depiction was sufficient to explain the majority of the variance in Multiple Correspondence Analyses [13].

The eigen values measure indicates how much of the categorical information is accounted for by each dimension. The higher the eigen value, the larger the amount of the total variance among the variables loads on that dimension. The largest possible eigen value for any dimension is 1. Usually, the first two or three dimensions contain higher eigen values than others. In this analysis, the maximum eigen value in the first dimension (dim 1) was 0.18. The similarly low eigen values in each dimension indicated that the variables in the crash data are heterogeneous and all carry, to some extent, unique information which implies that reducing any of the variables might result in losing important information concerning the crash observations. The heterogeneity of the crash variables reflects the random nature of crash occurrence.

In Table 3, eigen values and percentages of variance of the first 10 dimensions are revealed. It can also be seen (Table 3) that there is a steady decrease in eigen values. Based on the calculation, the first two dimensions cover only 8.1% of the percentage of variance, and the first 10 dimensions of the 83 in the dataset cover nearly 26% of the percentage of variance.

The coordinates of the first five dimensions of ten categories and ten rows are shown in Tables 4 and 5. Besides the eigen values, the row coordinates provide information about the structure of the rows

in the analyzed table. In turn, the column coordinates provide information about the structure of the analyzed variables and their corresponding categories. Tables 4 and 5 demonstrate the first five dimension values for each variable. Large coordinate measures indicate that the categories of a variable are better separated along that dimension, while similar coordinate measures for different variables in the same dimensions indicate that these variables are related to each other. Correlated variables provide redundant information and therefore some of them can be removed. The variables with significance in two dimensions are listed in Table 6.

The key focus of MCA is to provide an insight into the dataset by using information visualization. The popular graphical R package 'ggplot2' was extensively used to produce the informative MCA plots [14]. The main MCA plot (perceptual map) is shown in Figure 3. The plots shown in Figures 4-6 are four different combinations that were selected from the MCA plot. The contribution of a category depends on data, whereas that of a variable only depends on the number of categories of that variable. The more categories a variable has, the more the variable contributes to the variance of the cloud. The less frequent a category, the more it contributes to the overall variance. This property enhances infrequent categories which is desirable up to a certain point.

The dimension description of each point figures out the main characteristics according to each dimension obtained by a factor analysis. The dominant variables in dimension 1 are: driving violation, driver condition, the primary contribution factor, driver distraction, and highway type. For dimension 2, driver condition, alcohol involvement, access control, highway type, lighting, crash hour, and driving violation are the dominant variables.

The combination selection is based on the relative closeness of the category location in the MCA plot. In Figure 3, the distribution of the coordinates of all categories is shown. This plot gives us an idea of the variable categories' positions on the two dimensional space based on their eigen values. When the categories are relatively closer they form a combination cloud. Five significant combination clouds are chosen for further explanation. Four combinations of clouds are shown in Figure 4, an extended plot of combination cloud 4 is shown in Figure 5, and combination 5 is shown in Figure 6.

Combination Cloud 1 combines older drivers (aged 54 plus), partial access control, non-alcohol, and failure to yield in a group. It indicates that in partial access control zones older drivers faced problems with failure to yield which caused fatal crashes. Combination Cloud 2 associates older female drivers aged between 65 and 74 years of age with factors like straight and hillcrest aligned roadways, and non-passenger cars. Combination Cloud 3 combines the categories like lightweight trucks, no access control, state highways, and two way roads with no physical separation. In safety literature, undivided roadways with no access control are usually considered as hazardous, so this result indicates that truck drivers are associated with more fatal crashes in undivided state highways.

Combination Cloud 4 [Figure 5] combines categories like male drivers (age 15-24, 35-44, and 55-64), no defect passenger cars, dawn, and roadway segment. In this cloud, there are other variables like day of the week, weather and surface type. As the later three variables have most of the major options in this cloud, inclusion of these three variables would be redundant in nature. This combination cloud indicates that dawn driving for male drivers in roadway segments is a significant group in fatal single vehicle crashes.

Combination Cloud 5 associates the categories like alcohol-yes, drugs-yes, road-condition-animal-in-roadway, and driver-impaired. This combination indicates that impaired driving may cause fatal crashes due to poor reaction time.

The results presented in this paper demonstrate that MCA can be used to identify significant combinationing groups that tend to increase the crash frequency of single vehicle crashes. If the crash database is more complete, MCA will generate more intersection combination clouds from the dataset in an unsupervised way. The findings of this research will be useful to the highway professional to determine the hidden risk association group of variables in single vehicle fatal crashes.

DISCUSSION

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.

This study investigated several years of single vehicle fatal crash data to determine the significant contribution factors to fatal single vehicle crashes. It should be noted that the total variance explained by selected variables is about 11%. Few interesting combination groups were identified by using the MCA method. From the analysis one of the risk groups is drivers of lightweight trucks driving on undivided state highways. While speeding is a possible concern in this regard, it is also important to improve the regularity conditions of these types of roadways. Male drivers driving passenger-cars at dawn is another group that is vulnerable to fatal single vehicle crashes. Older female drivers (65-74) are seen as risk group while they are driving non-passenger cars. Moreover, in partial access control zones older drivers facing hardship to yield are in a risk group of fatal single vehicle crashes. Impaired drivers are also in a risk group due to their poor reaction time.

CONCLUSION

In conclusion, this study uses a new method to investigate contributing factors of fatal single vehicle 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.

REFERENCES

[1] Saferoads. Website: http://www.saferoads.org/rollover. Accessed July 20, 2013.

[2] Schneider, H. Louisiana Traffic Records Data Report 2013. Louisiana State University, Baton Rouge, Louisiana, 2014.

[3] Brorsson B, Rydgren H, and Ifver J. Single-Vehicle Accidents in Sweden: A Comparative Study of Risk and Risk Factors by Age. Journal of Safety Research. 1993; 24.

[4] Kelvin KWY. Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. Accident Analysis and Prevention. 2004; 36: 333–340.

[5] Reed S, Morris A. Characteristics of fatal single-vehicle crashes in Europe. International Journal of Crashworthiness. 2012; Version 1.

[6] Roux BL, Rouanet H. Multiple Correspondence Analysis. Sage Publications, Washington D.C. 2010.

[7] Hoffman DL, De Leeuw J. Interpreting Multiple Correspondence Analysis as a Multidimensional Scaling Method. Marketing Letters. 1992; 3: 259-272.

[8] Fontaine H. A Typological Analysis of Pedestrian Accidents. Presented at the 7th workshop of ICTCT, Paris, 26-27 October, 1995.

[9] Golob TF, Hensher DA. The trip chaining activity of Sydney residents: A cross-section assessment by age group with a focus on seniors. Journal of Transport Geography. 2007; 15(4).

[10] Factor R, Yair G, Mahalel, D. Who by Accident? The Social Morphology of Car Accidents. Risk Analysis. 2010; 30(9).

[11] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org. Accessed July 20, 2013.

[12] Husson F, Josse J, Le S, Mazet J. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. R package version 1.25. http://CRAN.R-project.org/package=FactoMineR. Accessed July 20, 2013.

[13] Greenacre M, Blasius J. Multiple Correspondence Analysis and Related Methods. Chapman & Hall/CRC, FL, 2006.

[14] Wickham H. ggplot2: elegant graphics for data analysis. Springer New York, 2009.

FIGURES AND TABLES

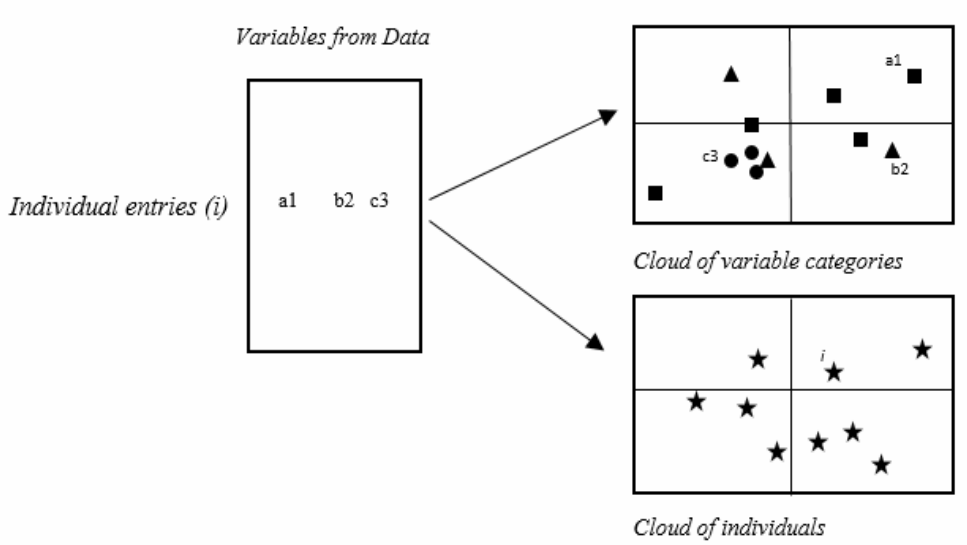

Figure 1. Data table and the two clouds of points generated by MCA with the flowchart.

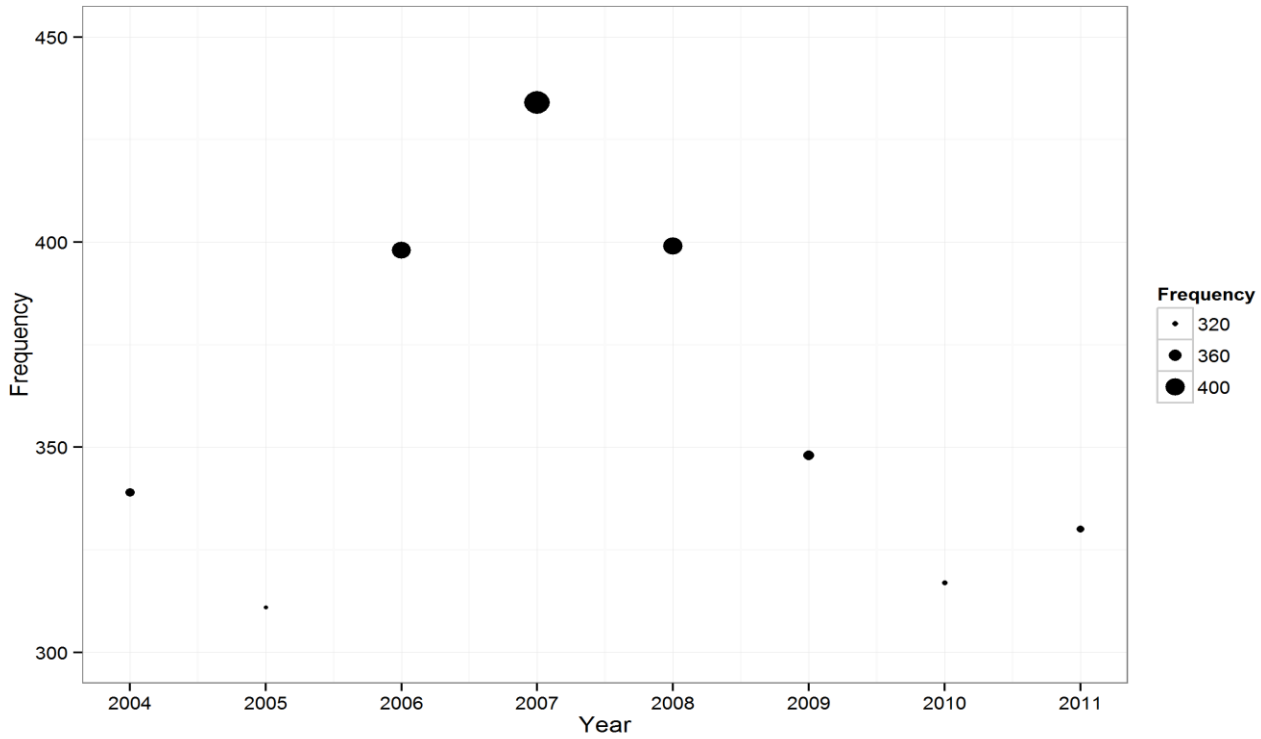

Figure 2. Fatal Single Vehicle Crashes

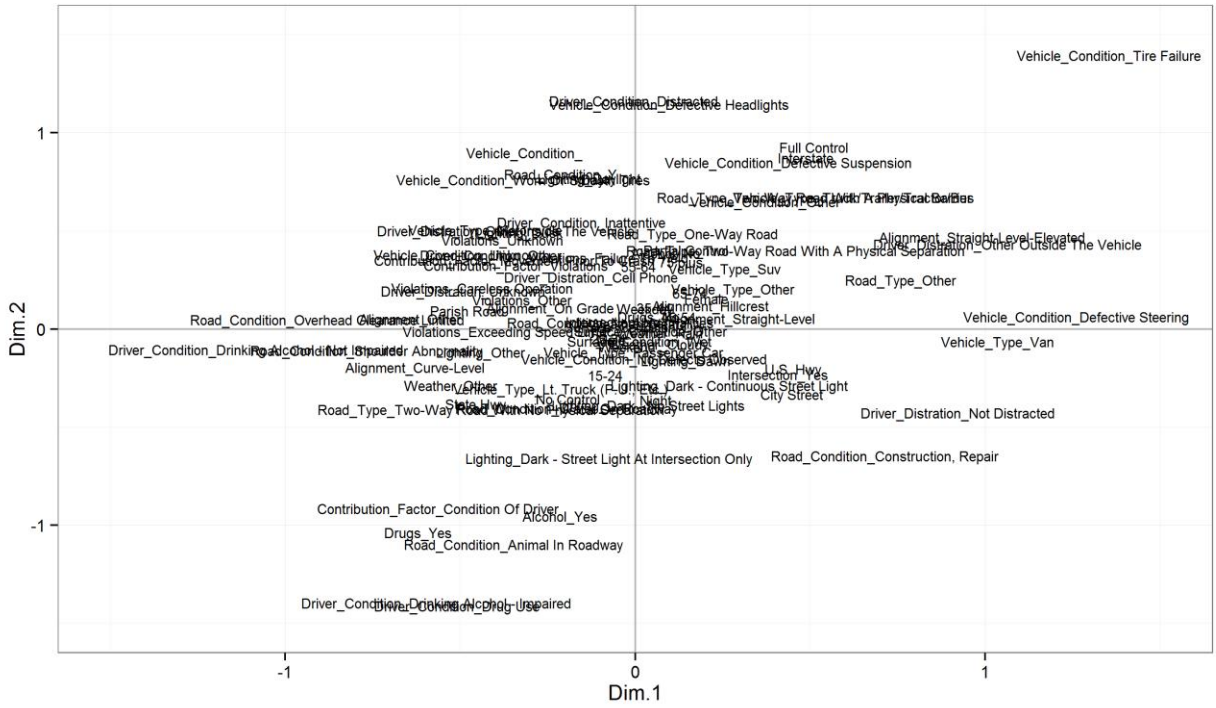

Figure 3. MCA plot for variable categories [Dim 1 (-1.5, 1.5), Dim 2 (-1.5, 1.5)].

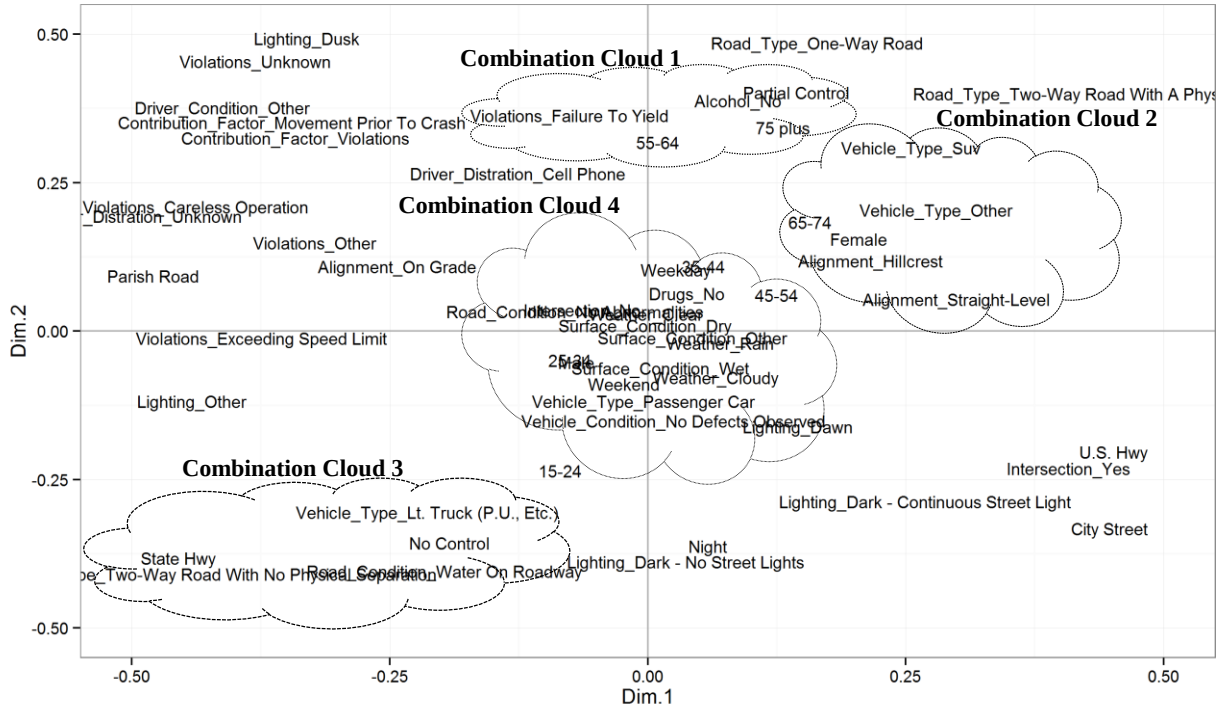

Figure 4. Combination Clouds 1-4 [Dim 1 (-0.5, 0.5), Dim 2 (-0.5, 0.5)].

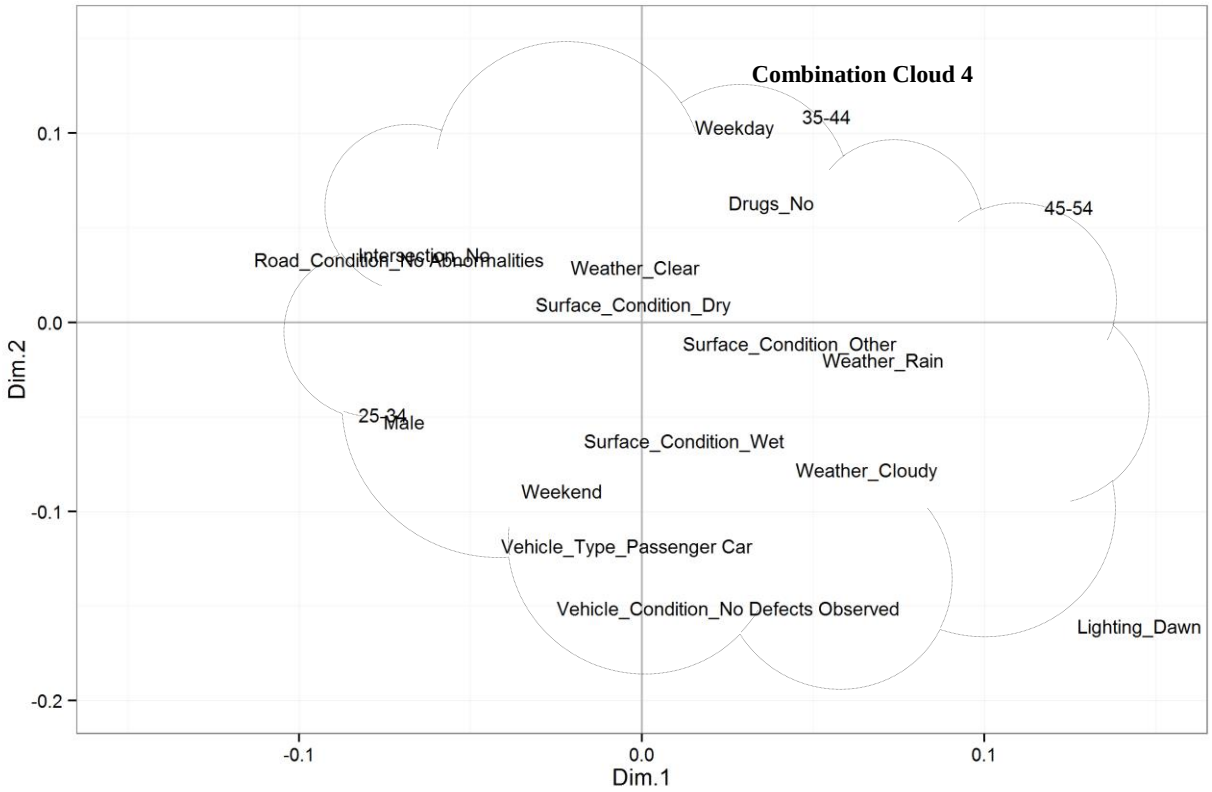

Figure 5. Combination Cloud 4 [Dim 1 (-0.15, 0.15), Dim 2 (-0.2, 0.15)].

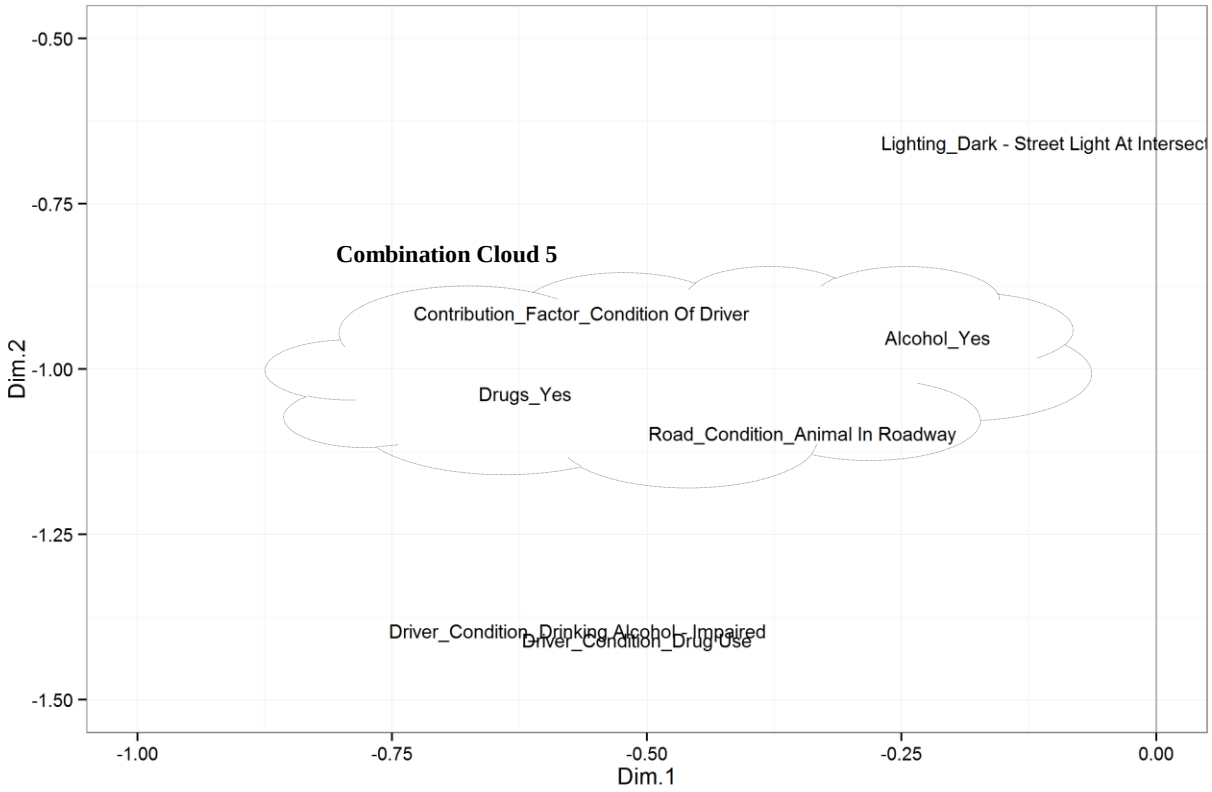

Figure 6. Combination Cloud 5 [Dim 1 (-1.00, 0.00), Dim 2 (-1.5, -0.75)].

Table 1A. Summary of the Variable Categories

Category	Frequency	Percentage	Category	Frequency	Percentage
Crash_Time			**Roadway_Condition**		
Day	358	32.17%	No Abnormalities	1,045	93.89%
Night	755	67.83%	Other	31	2.79%
Drugs			Construction, Repair	12	1.08%
No	1,049	94.25%	Shoulder Abnormality	6	0.54%
Yes	64	5.75%	Object In Roadway	5	0.45%
Alcohol			Animal In Roadway	3	0.27%
No	791	71.07%	(Other)	11	0.99%
Yes	322	28.93%	**Weather**		
Day_of_Week			Clear	823	73.94%
Weekday	516	46.36%	Cloudy	175	15.72%
Weekend	597	53.64%	Other	30	2.70%
Access_Control			Rain	85	7.64%
Full Control	285	25.61%	**Highway_Type**		
No Control	787	70.71%	City Street	4	0.36%
Partial Control	41	3.68%	Interstate	303	27.22%
Alignment			Parish Road	4	0.36%
Curve-Level	307	27.58%	State Hwy	562	50.49%
Hillcrest	22	1.98%	U.S. Hwy	240	21.56%
On Grade	106	9.52%	**Driver_Gender**		
Other	5	0.45%	Female	282	25.34%
Straight-Level	650	58.40%	Male	831	74.66%
Straight-Level-Elevated	23	2.07%	**Driver_Severity**		
Contributing_Factor			Complaint	74	6.65%
Condition Of Driver	163	14.65%	Fatal	736	66.13%
Movement Prior To Crash	125	11.23%	Moderate	68	6.11%
Other	204	18.33%	No Injury	210	18.87%
Violations	621	55.80%	Severe	25	2.25%
Lighting			**Driver_Age**		
Dark - Continuous Street Light	148	13.30%	15-24	295	26.50%
Dark - No Street Lights	501	45.01%	25-34	264	23.72%
Dark - Street Light At Intersection Only	62	5.57%	35-44	204	18.33%
Dawn	18	1.62%	45-54	177	15.90%
Daylight	363	32.61%	55-64	98	8.81%
Dusk	8	0.72%	65-74	47	4.22%
Other	13	1.17%	74 plus	28	2.52%

Table 1B. Summary of the Variable Categories

Category	Frequency	Percentage	Category	Frequency	Percentage
Road_Type			**Violations**		
One-Way Road	61	5.48%	Careless Operation	475	42.68%
Other	11	0.99%	No Violations	203	18.24%
Two-Way Road With A Physical Barrier	55	4.94%	Unknown	203	18.24%
Two-Way Road With A Physical Separation	413	37.11%	Other	83	7.46%
Two-Way Road With No Physical Separation	573	51.48%	Driver Condition	69	6.20%
Intersection			Exceeding Speed Limit	56	5.03%
No	963	86.52%	(Other)	24	2.16%
Yes	150	13.48%	**Vehicle_Condition**		
Surface_Condition			No Defects Observed	896	80.50%
Dry	956	85.89%	Unknown	129	11.59%
Other	15	1.35%	Worn Or Smooth Tires	34	3.05%
Wet	142	12.76%	Tire Failure	32	2.88%
Driver_Condition			Other	15	1.35%
Distracted	23	2.07%	Defective Headlights	3	0.27%
Drinking Alcohol - Impaired	145	13.03%	(Other)	4	0.36%
Drinking Alcohol - Not Impaired	3	0.27%	**Vehicle_Type**		
Drug Use	9	0.81%	Passenger Car	399	35.85%
Inattentive	98	8.81%	Lt. Truck (P.U., Etc.)	325	29.20%
Normal	207	18.60%	Suv	211	18.96%
Other	628	56.42%	Motorcycle	94	8.45%
Driver_Distraction			Van	23	2.07%
Cell Phone	14	1.26%	Truck/Trailer/Tractor/Bus	49	4.40%
Not Distracted	367	32.97%	(Other)	12	1.08%
Other Inside The Vehicle	25	2.25%			
Other Outside The Vehicle	13	1.17%			
Unknown	694	62.35%			

Table 2. Number of Categories in Each Variable

Variables	No. of categories	Variables	No. of categories
Crash_Time	2	Vehicle_Type	7
Contributing_Factor	3	Day_Of_Week	2
Weather	3	Road_Type	5
Violations	8	Driver_Age	7
Drugs	2	Access_Control	3
Lighting	7	Intersection	2
Highway_Type	5	Driver_Condition	7
Vehicle_Condition	10	Alignment	6
Alcohol	2	Surface_Condition	3
Roadway_Condition	13	Driver_Distraction	5
Driver_Gender	2		

Table 3. Eigen values and Percentages of Variance of the First Ten Dimensions

	Eigen value	**Percentage of variance**	**Cumulative Percentage of variance**
dim 1	0.175516866	4.336299	4.336299
dim 2	0.152233532	3.7610637	8.097363
dim 3	0.111658319	2.7586173	10.85598
dim 4	0.107289994	2.650694	13.506674
dim 5	0.098413867	2.4314014	15.938075
dim 6	0.089464642	2.2103029	18.148378
dim 7	0.085447954	2.1110671	20.259445
dim 8	0.081651997	2.0172846	22.27673
dim 9	0.076783733	1.8970099	24.17374
dim 10	0.072301528	1.7862731	25.960013

Table 4. Coordinates of Ten Random Categories

Categories	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Day	-0.1229733	0.76361165	-0.283717467	-0.73434224	0.106646578
Night	0.05831052	-0.36208341	0.134530931	0.34820467	-0.050568841
Drugs_No	0.03781442	0.06328374	-0.072056096	0.01395698	0.006015299
Drugs_Yes	-0.61980199	-1.03726	1.181044442	-0.22876364	-0.098594505
Alcohol_No	0.08738488	0.38769916	-0.238421852	0.02584315	0.066466412
Alcohol_Yes	-0.21466285	-0.95239141	0.585688462	-0.06348425	-0.163276186
Weekday	0.02699992	0.10302833	-0.124161972	-0.06206083	-0.004112785
Weekend	-0.02333662	-0.08904962	0.107315875	0.05364051	0.003554769
Full Control	0.51010777	0.9268866	1.024507953	0.16007211	-0.045563188
No Control	-0.19223862	-0.35660359	-0.371411082	-0.0762524	0.017605474

Table 5. Coordinates of First Ten Rows

Individuals	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
1	0.5441193	0.58262863	0.306731203	-0.25043933	0.09689667
2	0.1680554	0.15699736	-0.323378895	-0.30508499	-0.07106608
3	-0.3991592	-0.61429884	0.232467717	-0.2886275	0.03582543
4	-0.4189213	0.28492515	-0.162533389	-0.28865852	0.6916714
5	-0.1422736	-0.08039609	0.000350342	0.07966146	0.90260624
6	-0.4203312	0.16798321	-0.44434657	-0.44634685	-0.02466224
7	-0.3698073	-0.12410897	-0.166043145	0.06208131	0.36290678
8	-0.3135494	0.02594343	-0.081379363	0.16041872	-0.20761171
9	0.1905632	0.46308745	0.295545988	0.02478715	0.55029961
10	-0.2313237	0.09068579	-0.220361644	0.30736723	-0.47535132

Table 6. Variables in Dimensions 1 and 2 according to their Significance

Dimension 1 (categories)	R^2	p.value	Dimension 2 (categories)	R^2	p.value
Violations	0.724689	3.10E-304	Driver_Condition	0.447825	7.23E-139
Driver_Condition	0.721744	4.80E-303	Alcohol	0.369241	2.62E-113
Contributing_Factor	0.655681	3.78E-256	Access_Control	0.315864	3.19E-92
Driver_Distraction	0.447482	4.53E-141	Highway_Type	0.29089	3.20E-81
Highway_Type	0.214642	8.84E-57	Lighting	0.296896	3.45E-81
Alignment	0.190087	1.72E-48	Crash_Time	0.276491	3.79E-80
Road_Type	0.185733	3.81E-48	Violations	0.277495	9.13E-74
Roadway_Condition	0.150684	3.43E-32	Contributing_Factor	0.236809	1.08E-64
Access_Control	0.0935276	2.15E-24	Road_Type	0.181327	7.41E-47
Vehicle_Condition	0.0941033	1.99E-19	Vehicle_Condition	0.120415	4.05E-26
Vehicle_Type	0.0814227	4.26E-18	Driver_Distraction	0.0919242	3.27E-22
Intersection	0.025888	6.78E-08	Vehicle_Type	0.0919929	8.98E-21
Drugs	0.0234375	2.87E-07	Drugs	0.0656417	3.85E-18
Alcohol	0.0187583	4.52E-06	Roadway_Condition	0.0546583	9.29E-09
Driver_Gender	0.0141444	6.97E-05	Driver_Age	0.0311911	4.10E-06
Lighting	0.0196383	1.22E-03	Alignment	0.0181899	1.07E-03
Crash_Time	0.00717064	4.70E-03	Day_Of_Week	0.00917463	1.38E-03
Weather	0.00857899	2.27E-02	Intersection	0.00830842	2.34E-03
			Driver_Gender	0.00815661	2.56E-03