IATSS Research 39 (2016) 146–155

Contents lists available at ScienceDirect

IATSS Research

Research article

# Association knowledge for fatal run-off-road crashes by Multiple Correspondence Analysis

Subasish Das [a,*], Xiaoduan Sun [b]

[a] Systems Engineering, University of Louisiana at Lafayette, Lafayette, LA 70504, United States
[b] Civil Engineering Department, University of Louisiana at Lafayette, Lafayette, LA 70504, United States

ARTICLE INFO

*Article history:*
Received 15 August 2014
Received in revised form 2 July 2015
Accepted 6 July 2015
Available online 13 July 2015

*Keywords:*
ROR crashes
Multiple Correspondence Analysis
Dimensionality reduction
Cloud of combination groups

ABSTRACT

In 2013, 346 out of 616 fatal crashes in Louisiana were single vehicle crashes with Run-Off-Road (ROR) crashes being the most common type of single vehicle crash. In order to create effective countermeasures for reducing the number of fatal single vehicle ROR crashes, it is important to identify any associated key factors that can quantitatively assess the performance of roads, vehicles and humans. This research uses Multiple Correspondence Analysis (MCA), a multidimensional descriptive 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 fatal ROR crashes in Louisiana. This method measures important contributing factors and their degree of association. The results revealed that drivers of lightweight trucks, drivers on undivided state highways, male drivers in passenger-vehicles at dawn, older female (65–74) drivers in non-passenger vehicles, older drivers facing hardship to yield in partial access control zones, and drivers with poor reaction time due to impaired driving were closely associated with fatal ROR crashes.

Results of the MCA method can help researchers select the most effective crash countermeasures. Further work on the degree of association between the identified crash contributing factors can help safety management systems develop the most efficient crash reduction strategies.

© 2015 The Authors. Publishing services by Elsevier Ltd. on behalf of International Association of Traffic and Safety Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

## 1. Introduction

Most single vehicle crashes are ROR crashes, which are more likely to result in fatalities and severe injuries than typical vehicle crashes are [1]. In 2012 and 2013, respectively 384 and 346 out of a total of 652 and 616 fatal crashes in Louisiana were ROR crashes [2]. From prior studies, we know that single vehicle ROR crashes are usually caused by a combination of factors such as inadequate roadway design, mechanical problems, environmental conditions and/or drivers' poor performance [3–5]. The combination of factors could be spatially different (i.e. crashes occurring on highways versus intersections) and temporally different (i.e. crashes occurring in December versus those in May). Failure to recognize the spatial and temporal differences of those factors may lead to insufficient or ineffective actions taken 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 cannot identify the combination of factors simultaneously. Multiple Correspondence Analysis (MCA) is an extension of Correspondence Analysis (CA) for more than two variables and is widely used in categorical data analyses, especially in social sciences and marketing research [6]. By using this technique we can visualize the patterns of combined crash contributing factors. MCA helps researchers 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 creating combination clouds.

The persistently high rate of fatal ROR crashes in Louisiana and the United States indicates continuous need for research. Reducing ROR crashes is critical in fulfilling state and national safety goal and MCA will help determine the association between key factors of fatal ROR crashes so that transportation authorities can take necessary actions to reduce crash frequencies and severities.

## 2. Literature review

J.P. Benzécri developed Multiple Correspondence Analysis (MCA), a statistical approach based on Correspondence Analysis (CA). MCA is usually considered to be one of the main standards of geometric data

* Corresponding author. Tel.: +1 225 288 9875; fax: +1 337 739 6688.
*E-mail addresses:* subasishsn@gmail.com (S. Das), xsun@louisiana.edu (X. Sun).
Peer review under responsibility of International Association of Traffic and Safety Sciences.

http://dx.doi.org/10.1016/j.iatssr.2015.07.001
0386-1112/© 2015 The Authors. Publishing services by Elsevier Ltd. on behalf of International Association of Traffic and Safety Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

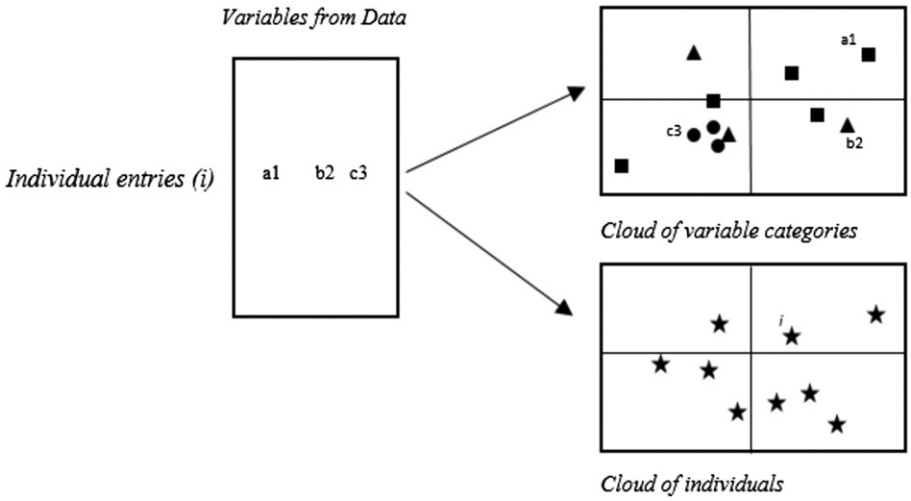


**Fig. 1.** Data table and the two clouds of points generated by MCA with the flowchart [6].

analysis (GDA) in the fields of social science and marketing research. GDA is also referred to as the pattern recognition method that treats arbitrary data sets as clouds of points in *n*-dimensional space. However, in the field of multivariate transportation data analysis, researchers rarely use geometric methods. Roux and Rouanet pointed out that this method, though it is a powerful tool for analyzing a full-scale research database, is still rarely 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 can be answered with MCA in terms of significant and meaningful results [7]. Fontaine was the first to use MCA for a typological analysis of vehicle-pedestrian crashes [8]. For Fontaine's research, the classification of pedestrians involved in crashes was 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 utilized MCA to establish causality of nonlinear and non-trivial relationships between socioeconomic descriptors and outcomes of travel behavior [9]. Factor et al. used MCA to conduct a systematic exploration of the connection between drivers' characteristics and their involvement in collision types [10]. There is a vast amount of literature on accident research and model development that, for the sake of brevity, cannot be covered in this article. The research team has compiled an extensive list of this literature in a webpage for the convenience of any interested readers [11].

The research introduced in this paper serves as a starting point to demonstrate the application of MCA to determine the significant clouds of crash contributing factors for fatal ROR crashes. The findings will help state agencies determine effective crash countermeasures.

## 3. Methodology

### 3.1. Theory

For a database or table with categorical variables, the scheme of MCA 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. As shown in Fig. 1, MCA yields two clouds of points: the cloud of individual records and the cloud of categories [6]. A cloud of points is not just a *simple graphical* display; it can be compared to a geographic map with the same scale in all directions. A geometric diagram cannot be strained or contracted along one specific dimension. Thus, a

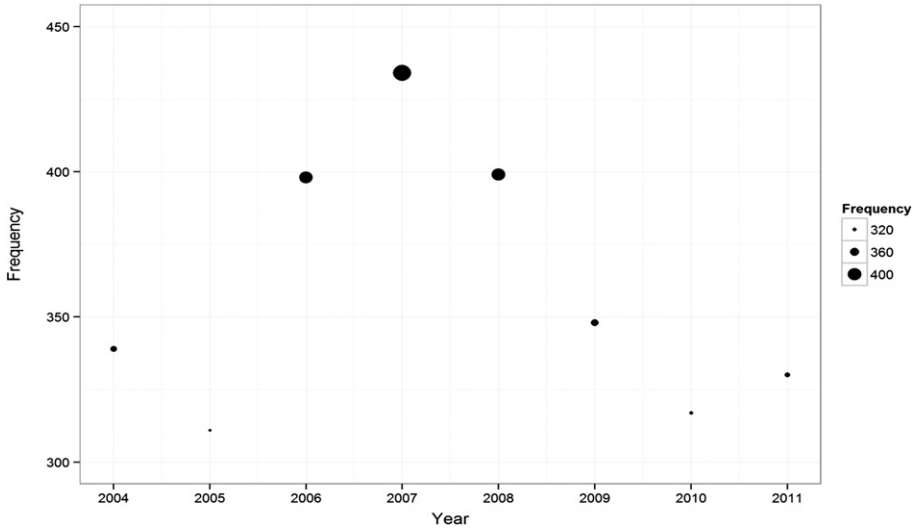


**Fig. 2.** Fatal ROR crashes.

**Table 1A**
Summary of variables.

| Category | Frequency | Percentage |
|---|---|---|
| Crash_time | | |
| Day | 358 | 32.17% |
| Night | 755 | 67.83% |
| Drugs | | |
| No | 1049 | 94.25% |
| Yes | 64 | 5.75% |
| Alcohol | | |
| No | 791 | 71.07% |
| Yes | 322 | 28.93% |
| Day_of_week | | |
| Weekday | 516 | 46.36% |
| Weekend | 597 | 53.64% |
| Access_control | | |
| Full control | 285 | 25.61% |
| No control | 787 | 70.71% |
| Partial control | 41 | 3.68% |
| Alignment | | |
| Curve-level | 307 | 27.58% |
| Hillcrest | 22 | 1.98% |
| On grade | 106 | 9.52% |
| Other | 5 | 0.45% |
| Straight-level | 650 | 58.40% |
| Straight-level-elevated | 23 | 2.07% |
| Contributing_factor | | |
| Condition of driver | 163 | 14.65% |
| Movement prior to crash | 125 | 11.23% |
| Other | 204 | 18.33% |
| Violations | 621 | 55.80% |
| Lighting | | |
| Dark—continuous street light | 148 | 13.30% |
| Dark—no street lights | 501 | 45.01% |
| Dark—street light at intersection only | 62 | 5.57% |
| Dawn | 18 | 1.62% |
| Daylight | 363 | 32.61% |
| Dusk | 8 | 0.72% |
| Other | 13 | 1.17% |
| Roadway_condition | | |
| No abnormalities | 1045 | 93.89% |
| Other | 31 | 2.79% |
| Construction, repair | 12 | 1.08% |
| Shoulder abnormality | 6 | 0.54% |
| Object in roadway | 5 | 0.45% |
| Animal in roadway | 3 | 0.27% |
| (Other) | 11 | 0.99% |
| Weather | | |
| Clear | 823 | 73.94% |
| Cloudy | 175 | 15.72% |
| Other | 30 | 2.70% |
| Rain | 85 | 7.64% |
| Highway_type | | |
| City street | 4 | 0.36% |
| Interstate | 303 | 27.22% |
| Parish road | 4 | 0.36% |
| State hwy | 562 | 50.49% |
| U.S. hwy | 240 | 21.56% |
| Driver_gender | | |
| Female | 282 | 25.34% |
| Male | 831 | 74.66% |
| Driver_severity | | |
| Complaint | 74 | 6.65% |
| Fatal | 736 | 66.13% |
| Moderate | 68 | 6.11% |
| No injury | 210 | 18.87% |
| Severe | 25 | 2.25% |
| Driver_age | | |
| 15–24 | 295 | 26.50% |
| 25–34 | 264 | 23.72% |
| 35–44 | 204 | 18.33% |
| 45–54 | 177 | 15.90% |
| 55–64 | 98 | 8.81% |
| 65–74 | 47 | 4.22% |
| 75 plus | 28 | 2.52% |

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 a convenient version where points lie on a plane. The full clouds are referred to by their principal dimensions (1, 2, 3, etc.) that are ranked in descending order of importance. The goal of MCA is to create a combination of groups from a large dataset. Fig. 1 exhibits the flowchart of the MCA procedure where the cloud of categories and the cloud of individual records are considered as the cloud of points. Since many texts detail the theory of MCA, we will describe only the basic fundamentals of the theory. Interested readers can consult the listed references [6, 12–14] and references included therein.

As shown in Fig. 1, MCA uses tables and user-defined data matrices to develop the data clouds it produces. The data matrix is an "I by Q" table with all categorical values with $Q$ representing the number of variables and $I$ indicating the number of records. The total number of categories for all variables is $J = \sum_{q=1}^{Q} J_q$ with $J_q$ as the number of categories for variable $q$. To contain all categories in the data table, another data matrix is developed as *"I by J"* where each variable has 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 contain "0" and the "no" column will contain "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 $k'$ which is different from k. The squared distance between individual records $i$ and $i'$ for variable $q$ is defined by

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

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

$$d^2(i, i') = \frac{1}{Q} \sum_{q \in Q} d_q{}^2(i, i'). \tag{2}$$

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 $L \le K - Q$ (overall number $K$ of categories minus number $Q$ of variables), and assuming $n \ge L$. If $M^i$ denotes the point representing individual $i$ and $G$ denotes the mean point of the cloud, the squared distance from point $M^i$ to point $G$ is

$$\left(GM^i\right)^2 = \frac{1}{Q} \sum_{k \in K_i} \frac{1}{f_k} \tag{3}$$

where $K_i$ denotes the response pattern of individual $i$, meaning it is the set of $Q$ categories associated with individual record $i$.

The cloud of categories is a weighted cloud of $K$ points (by category $k$, a point denoted by $M^k$ with weight $n_k$ is represented). 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. The sum of the whole set is equal to one.

$$p_k = \frac{n_k}{nQ} = \frac{f_k}{Q} \quad with \sum_{k\in K_q} p_k = \frac{1}{Q} \quad and \sum_{k\in K} p_k = 1$$

If $n_{kk'}$ indicates the number of individual records having both categories ($k$ and $k'$), then the squared distance between $M^k$ and $M^{k'}$ is

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

The numerator is the number of individual records associated with either $k$ or $k'$. For two different variables (say $q$ and $q'$), the denominator is the familiar *theoretical frequency* for the cell $(k, k')$ of the $K_q \times K_{q'}$ two-way table.

While modern machine-learning approaches like association rules mining can tackle the research problem in this study, MCA was determined to be the better choice. MCA was chosen because it is better for interpreting large datasets than conventional log-linear models are. Moreover, in MCA there is no need to consider any underlying distribution and no relationship has to be hypothesized. Also, association rules mining has limitations when it comes to selecting appropriate threshold values of support and confidence. Smaller values of support and confidence increase the number of rules immensely which makes interpreting results difficult, but larger support or confidence values may ignore import rules. Moreover, rules with a large number of item sets are difficult to interpret in association rules mining. MCA overcomes these difficulties by performing efficient dimensionality reductions and compiling results into easy-to-read plots.

The actual MCA computations are conducted on the inner product of the matrix known as the 'Burt Table'. The research team used open source statistical software *R Version 3.02* to perform the MCA technique [15]. This study used the *FactoMineR* package to analyze the dataset due to its convenient functions compared to other available packages [16]. We developed the combination clouds in MCA on both a variable level and a category level. It is important to note that categories represent both variables and a group of individual transactions.

### 3.2. Initial data analysis

To identify important contributing factors related to fatal ROR crashes in Louisiana, we collected eight years (2004–2011) of crash data from the Louisiana Department of Transportation and Development (DOTD). The primary dataset was created by merging the crash, roadway, and vehicle tables. For any given individual crash record, there are 371 possible variables (153 from the crash table, 40 from the roadway table and 178 from the vehicle table). Fig. 2, displaying the annual fatal ROR crashes by year in Louisiana, shows that there was a 4% increase in these crashes between 2010 and 2011 and that the highest number of fatal ROR crashes was in 2007. The master database created for this analysis includes all 2777 fatal crashes that occurred in the eight-year period.

The original list of relevant variables was primarily scanned by examining the relevance of missing values via a correlation matrix and the relevance of the distribution skew. This was necessary because datasets with a large number of missing values makes the MCA plots

**Table 1B**
Summary of variables.

| Category | Frequency | Percentage |
|---|---|---|
| Road_type | | |
| One-way road | 61 | 5.48% |
| Other | 11 | 0.99% |
| Two-way road with a physical barrier | 55 | 4.94% |
| Two-way road with a physical separation | 413 | 37.11% |
| Two-way road with no physical separation | 573 | 51.48% |
| Intersection | | |
| No | 963 | 86.52% |
| Yes | 150 | 13.48% |
| Surface_condition | | |
| Dry | 956 | 85.89% |
| Other | 15 | 1.35% |
| Wet | 142 | 12.76% |
| Driver_condition | | |
| Distracted | 23 | 2.07% |
| Drinking alcohol—impaired | 145 | 13.03% |
| Drinking alcohol—not impaired | 3 | 0.27% |
| Drug use | 9 | 0.81% |
| Inattentive | 98 | 8.81% |
| Normal | 207 | 18.60% |
| Other | 628 | 56.42% |
| Driver_distraction | | |
| Cell phone | 14 | 1.26% |
| Not distracted | 367 | 32.97% |
| Other inside the vehicle | 25 | 2.25% |
| Other outside the vehicle | 13 | 1.17% |
| Unknown | 694 | 62.35% |
| Violations | | |
| Careless operation | 475 | 42.68% |
| No violations | 203 | 18.24% |
| Unknown | 203 | 18.24% |
| Other | 83 | 7.46% |
| Driver condition | 69 | 6.20% |
| Exceeding speed limit | 56 | 5.03% |
| (Other) | 24 | 2.16% |
| Vehicle_condition | | |
| No defects observed | 896 | 80.50% |
| Unknown | 129 | 11.59% |
| Worn or smooth tires | 34 | 3.05% |
| Tire failure | 32 | 2.88% |
| Other | 15 | 1.35% |
| Defective headlights | 3 | 0.27% |
| (Other) | 4 | 0.36% |
| Vehicle_type | | |
| Passenger car | 399 | 35.85% |
| Lt. truck (P.U., etc.) | 325 | 29.20% |
| SUV | 211 | 18.96% |
| Motorcycle | 94 | 8.45% |
| Van | 23 | 2.07% |
| Truck/trailer/tractor/bus | 49 | 4.40% |
| (Other) | 12 | 1.08% |

**Table 2**
Number of categories for each variable.

| Variables | No. of categories |
|---|---|
| Crash_time | 2 |
| Contributing_factor | 3 |
| Weather | 3 |
| Violations | 8 |
| Drugs | 2 |
| Lighting | 7 |
| Highway_type | 5 |
| Vehicle_condition | 10 |
| Alcohol | 2 |
| Roadway_condition | 13 |
| Driver_gender | 2 |
| Vehicle_type | 7 |
| Day_of_week | 2 |
| Road_type | 5 |
| Driver_age | 7 |
| Access_control | 3 |
| Intersection | 2 |
| Driver_condition | 7 |
| Alignment | 6 |
| Surface_condition | 3 |
| Driver_distraction | 5 |

**Table 3**
Eigenvalues and percentages of variance of the first ten dimensions.

| | Eigenvalue | 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 |

less informative. By focusing on meaningful results, a set of key variables were selected for the final analysis. The variable selection method used previous research results with engineering judgment. The final dataset contains 21 variables relevant for this research. Table 1A and Table 1B enlist the summary of the selected variable counts where the variables are grouped by:

- Human factor or driver characteristics (driver age, intoxication, condition of the driver, violation type, driver distraction, driver gender, driver injury)
- Crash characteristics (crash year, crash time, 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 intersection-related crashes, 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, we transformed the quantitative variable Driver Age 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 75 plus. The other variables are nominal in nature. Table 2 lists the number of categories in each selected variable.

A preliminary analysis indicates that some variables are highly skewed, meaning that a majority of crashes fall into one of the two or more categorical values. For example, 94% of crashes involved a driver with no drug intoxication, 94% of 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, roadway type, driver age, lighting condition, and crash time.

### 3.3. Multiple Correspondence Analysis

Graphical illustrations are an easy way to perceive and interpret data because they effectively summarize large, complex datasets by simplifying the structure of the relations between variables and providing a collective view of the data [6]. Morphological maps are a better way of presenting information graphically and one can interpret them by examining the distribution of variable groupings 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 near one another in the map as groups, while those with different distributions stay farther apart. Hence, we interpret the dimensions (axes) 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 MCA [12].

**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**
Variables in dimensions 1 and 2 according to their significance.

| Dimension 1 | $R^2$ | p.Value | Dimension 2 | $R^2$ | 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 |

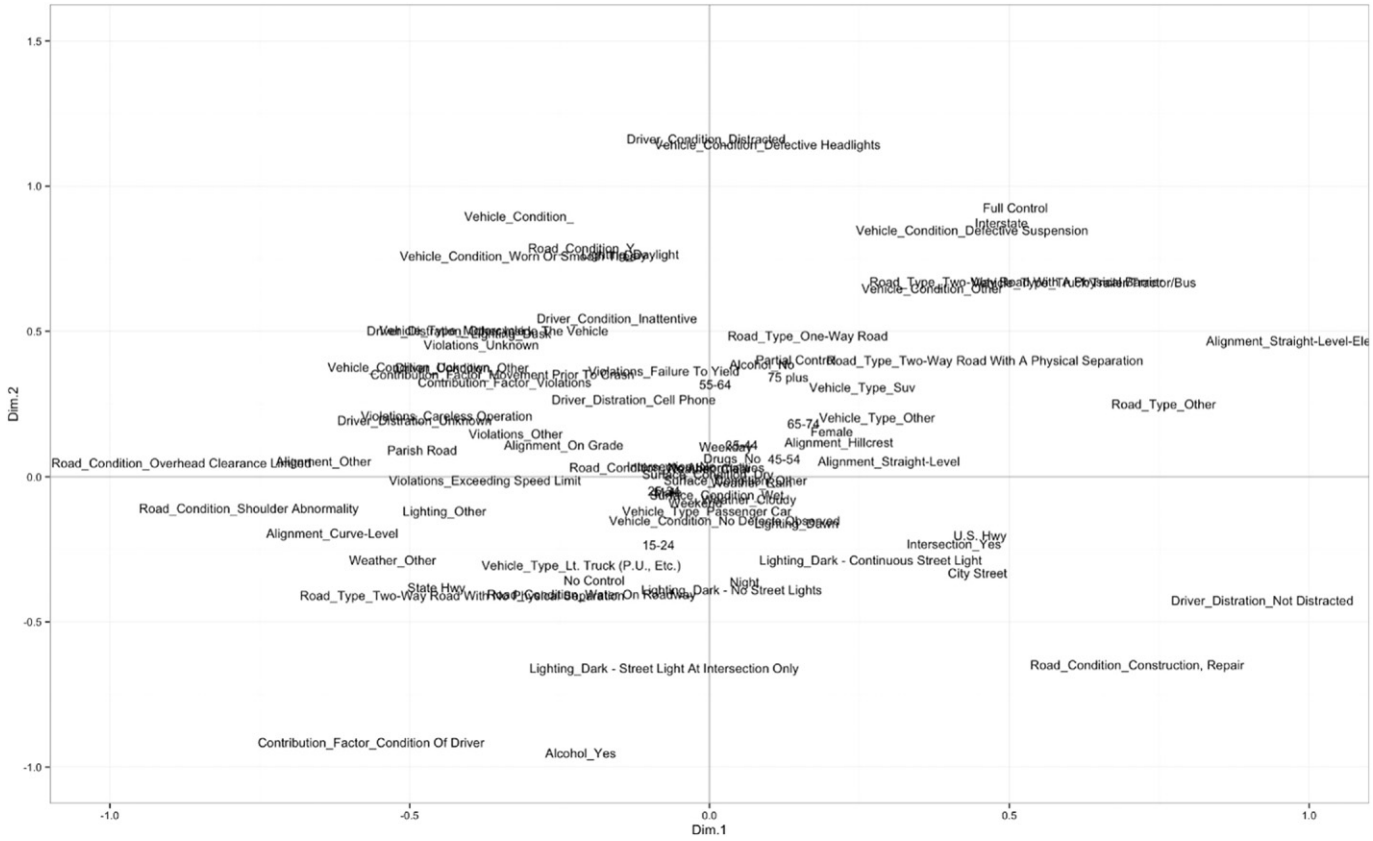


**Fig. 3.** MCA plot for variable categories.

The eigenvalue measures indicate how much each dimension accounts for categorical information. A higher eigenvalue indicates a larger total variance among the variables' loads on that dimension. The largest possible eigenvalue for any dimension is one. Usually, the first two or three dimensions contain higher eigenvalues than the others. In this analysis, the maximum eigenvalue in the first dimension (dim

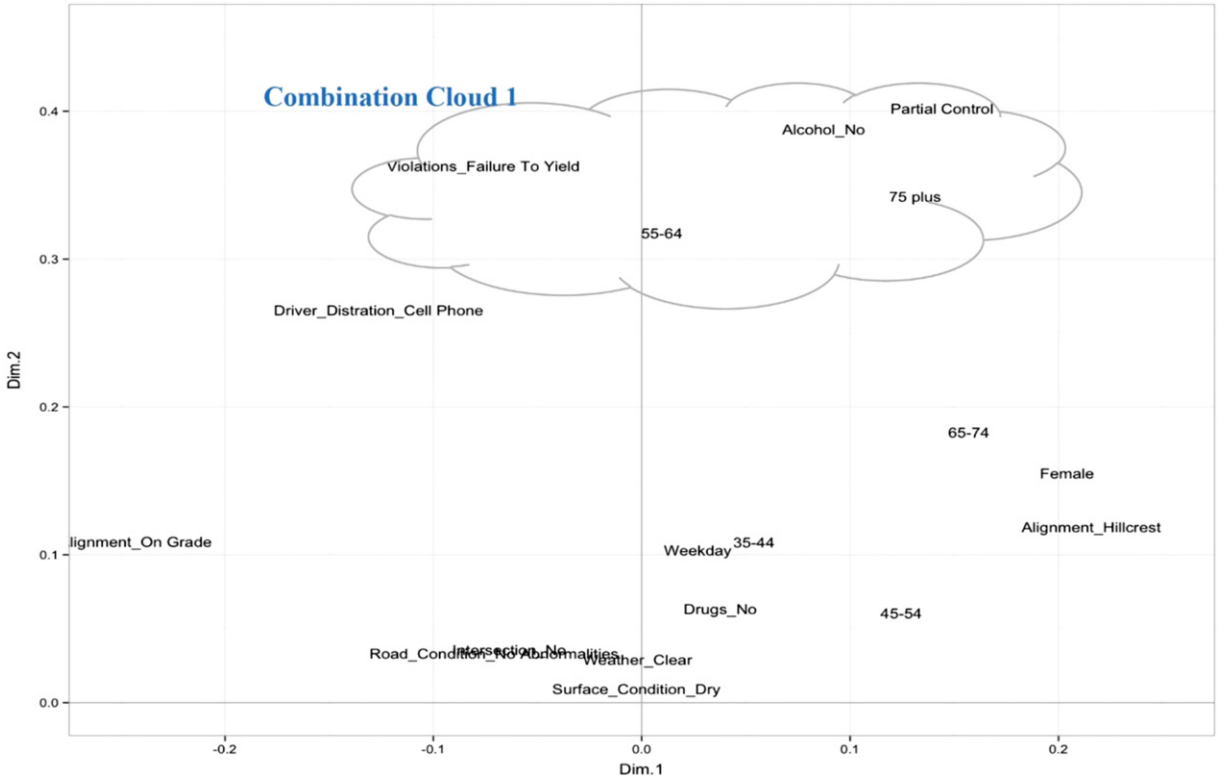


**Fig. 4.** Combination Cloud 1.

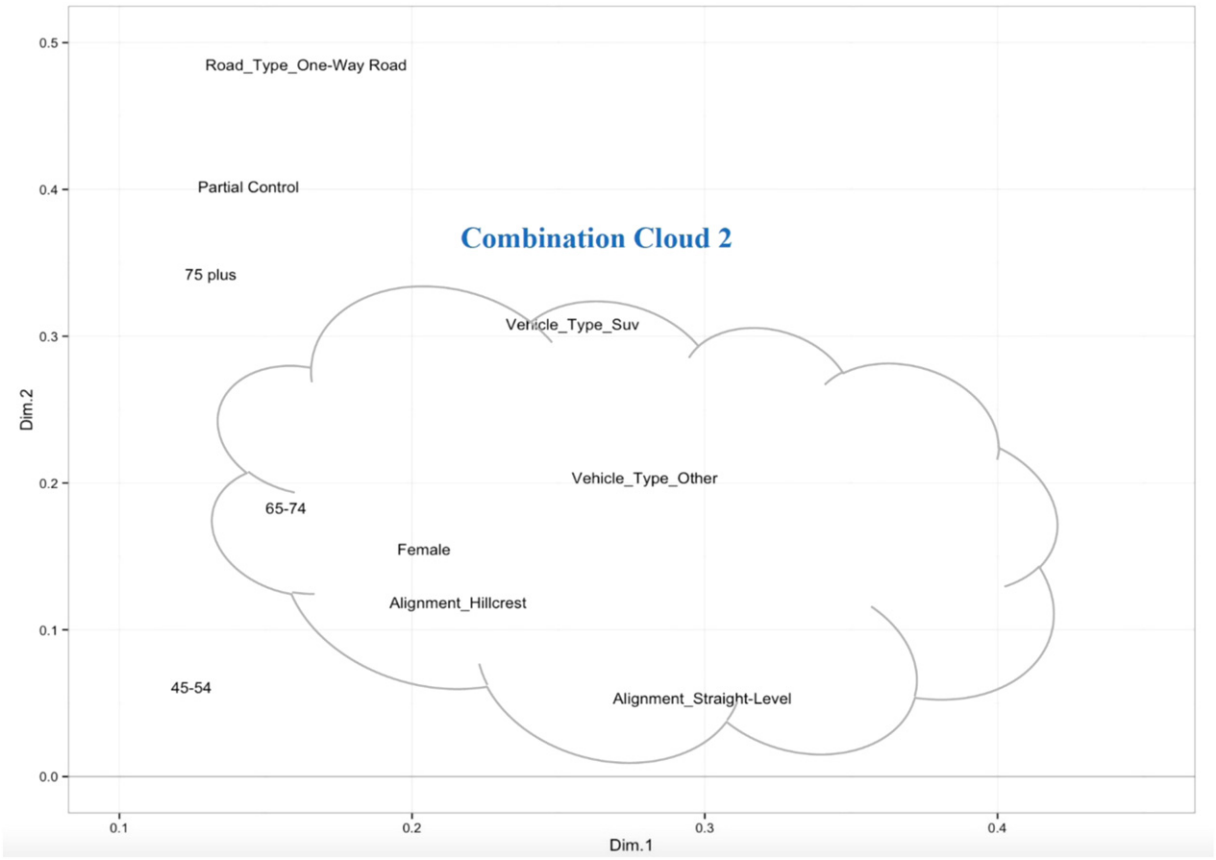


**Fig. 5.** Combination Cloud 2.

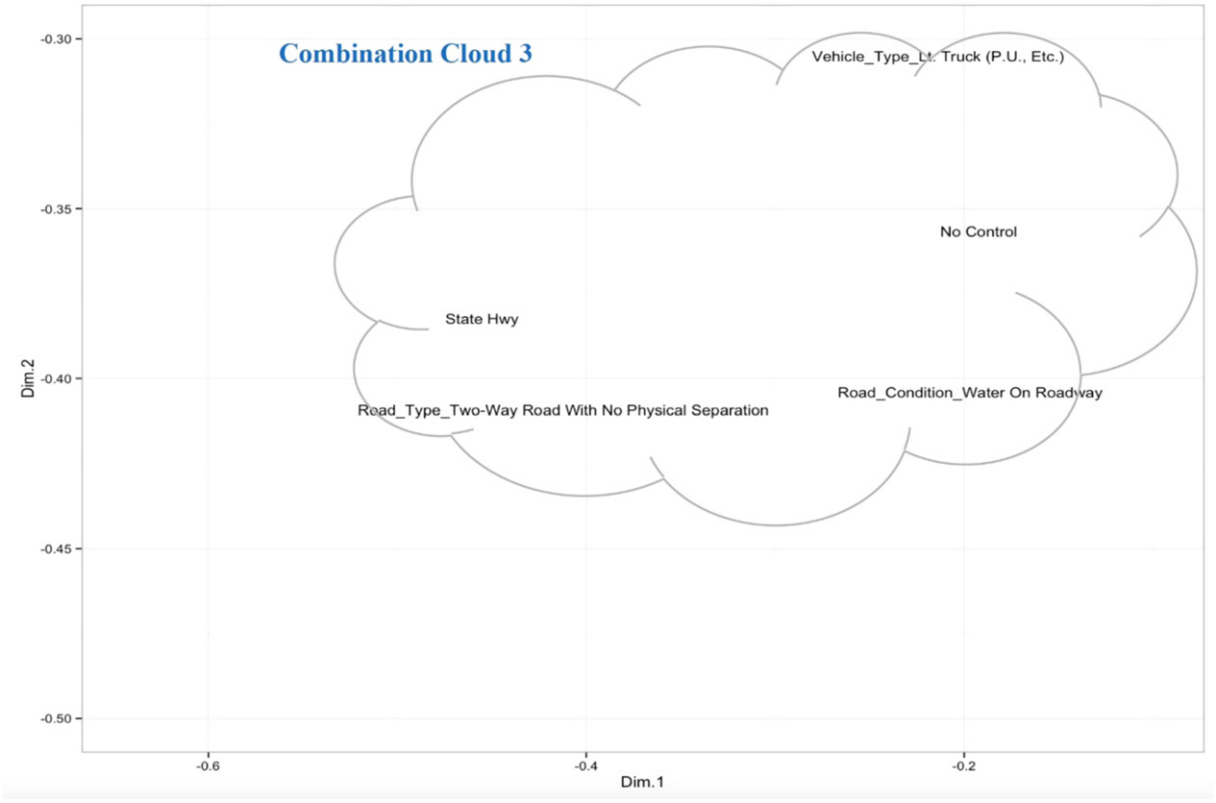


**Fig. 6.** Combination Cloud 3.

1) was 0.18. The similarly low eigenvalues in each dimension indicate 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 alludes to the random nature of crash occurrence.

As seen in Table 3, the eigenvalues of the first 10 dimensions show a steady decrease in eigenvalues. Based on the calculation, the first two dimensions cover only 8.1% of the percentage of variance, and the first 10 dimensions (out of 83 dimensions) cover nearly 26% of the percentage of variance.

Table 4 lists the coordinates of the first five dimensions of ten categories. Large coordinate measures indicate that the categories of a variable are separated along that dimension, while similar coordinate measures for different variables in the same dimension indicate the relationship between those variables. Correlated variables provide redundant information; therefore we did not consider them for combination formation. Table 5 lists the variables with significance in two dimensions.

The key advantage of MCA is that it provides insight into the dataset by using information visualization. We used popular graphical R package *ggplot2* to produce the informative MCA plots [17]. Fig. 3 illustrates the main MCA plot (perceptual map). The plots in Figs. 4–8 elaborately show different selected combinations based on their relative closeness and interestingness. The contribution of a category depends on the data, whereas the contribution 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 shows the main characteristics according to each dimension obtained by a factor analysis. The dominant variables in dimension 1 are driving violation, driver condition, driver distraction (the primary contributing factor), and highway type. The dominant variables for dimension 2 are driver condition, alcohol involvement, access control, highway type, lighting, crash hour, and driving violation.

The combination selection is based on the relative closeness of the category location in its MCA plot. Fig. 3 shows the distribution of the coordinates of all categories. This plot gives us an idea of the variable categories' positions on the two dimensional space based on their eigenvalues. When the categories are relatively close, they form a combination cloud. In this study, we chose five significant combination clouds from the MCA plot for further explanation. We did not consider combination groups with redundant information even though the relative distance was often closer. Combination clouds one to five are shown in Figs. 4 to 8.

Combination Cloud 1 combines four categories: older drivers (aged 54 plus), partial access control, non-alcohol, and failure to yield. This cloud indicates that in partial access control zones, older drivers failed to yield which caused a fatal crash. Combination Cloud 2 shows that

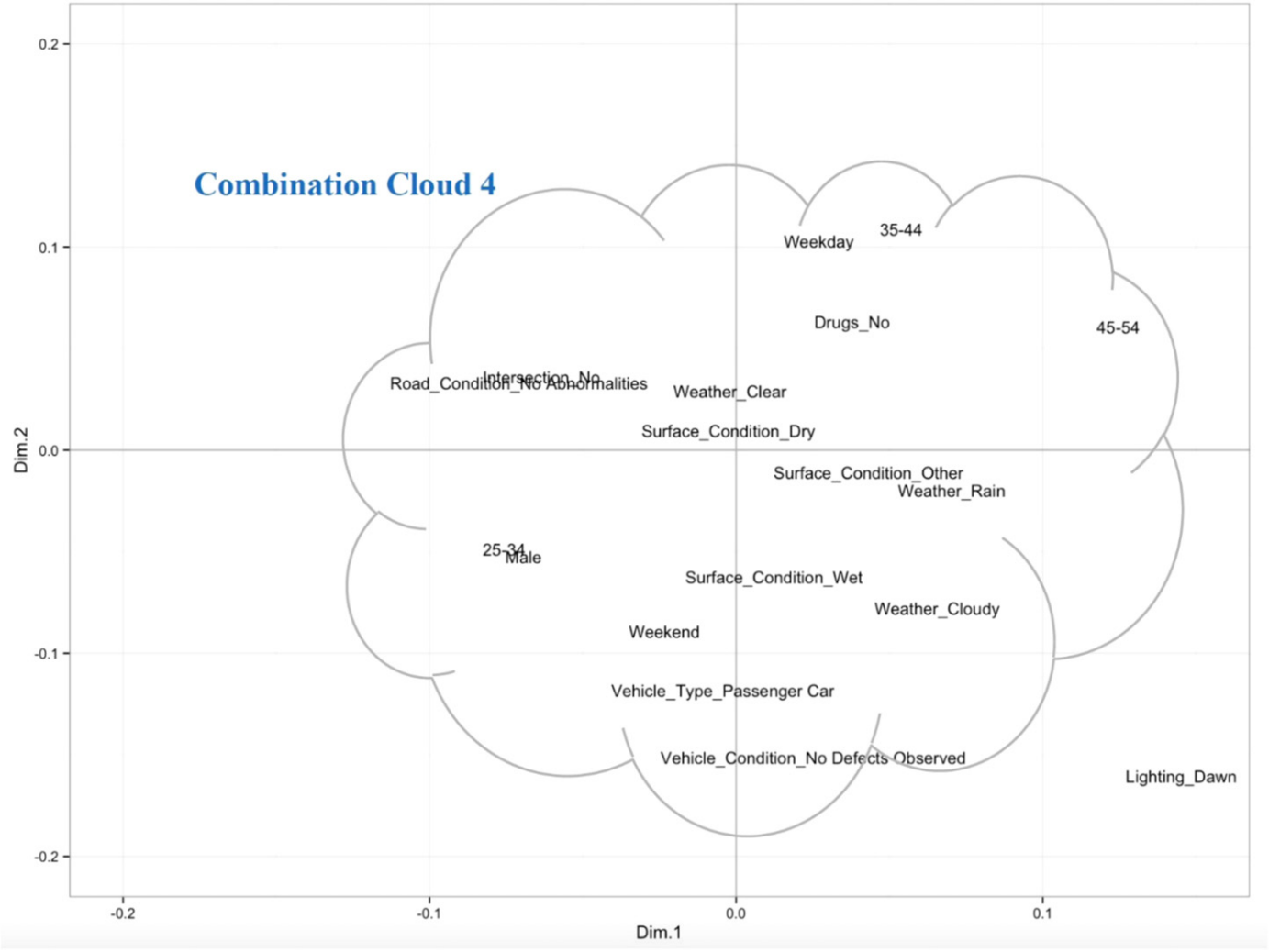


**Fig. 7.** Combination Cloud 4.


**Fig. 8.** Combination Cloud 5.

older female drivers aged between 65 and 74 are most likely to have fatal crashes on straight and hillcrest-aligned roadways while driving non-passenger cars. Combination Cloud 3 combines the categories of lightweight trucks, no access control, state highways, and two-way roads with no physical separation. From this combination cloud, we know that truck drivers on undivided state highways with no access control are more likely to have fatal crashes.

Combination Cloud 4 combines the categories: male drivers (age 15–24, 35–44, and 55–64), no-defect passenger cars, dawn, and roadway segment. This cloud indicates that for male drivers, driving at dawn on roadway segments is a significant focus group in fatal ROR crashes.

Combination Cloud 5 indicates that impaired driving may cause fatal crashes due to poor reaction time.

The results presented in this paper demonstrate that we can use MCA to identify significant combination groups that contribute to fatal ROR crashes. The authors would also like to mention that the total variance explained by the selected variables is not high (nearly 8.1% in this study). To adjust for this, we recalculated the inertia coverage by using the Burt table. The inertia of these two major axes the reached 47%. With a tidy dataset, i.e., a dataset with no missing values, the unsupervised method used in MCA can generate more interesting combination clouds. The findings of this research are useful to highway professionals in determining the nontrivial focus groups in fatal ROR crashes.

## 4. Conclusions

All parametric regression models contain their own model assumptions and pre-defined underlying relationships between response and exploratory variables. These models could lead to incorrect results due to the violation of any assumption. MCA, a widely used non-parametric approach in social sciences and marketing research, has proven to be a valuable analysis tool in roadway safety, as shown by the research presented in this paper. Without any pre-defined underlying relationships between response and explanatory variables, the research presented in this paper analyzed large sets of categorical crash data, avoiding the difficulty seen in using association rules mining.

By analyzing several years of fatal ROR crash data, the research team recognized the key association between the significant contributing factors using the MCA method. With this method, we identified a few particularly interesting variable combinations. We found that drivers of lightweight trucks on undivided state highways have a high crash risk, which may imply a speeding problem. We also found that male passenger-car drivers at dawn are vulnerable to fatal ROR crashes, and females between the ages of 65 and 74 driving non-passenger cars also have a high crash risk. Also, it was found that in partial access control zones, older drivers facing hardship to yield have a high risk for fatal ROR crashes. The MCA method was used to determine these fatal ROR crash focus groups by identifying the combination of factors for fatal ROR crashes. To reduce such crashes, safety programs should develop strategies that target to these factors simultaneously for the best results.

By performing an investigation on the fatal ROR crashes, this study has developed a methodology on the relative closeness of the key associated factors of ROR crashes. 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 show insight on the pattern recognition of traffic crashes and expose new aspects in traffic crash investigations. Further research will focus on the joint correspondence analysis and other non-parametric approaches to find the most dominating association among the contributing factors.

## References

[1] Saferoads, Website: http://www.saferoads.org/rollover (Accessed July 20, 2013).
[2] H. Schneider, Louisiana Traffic Records Data Report 2013, Louisiana State University, Baton Rouge, Louisiana, 2014.
[3] B. Brorsson, H. Rydgren, J. Ifver, Single-vehicle accidents in Sweden: a comparative study of risk and risk factors by age, J. Saf. Res. 24 (1993).
[4] K.W.Y. Kelvin, Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong, Accid. Anal. Prev. 36 (2004) 333–340.
[5] S. Reed, A. Morris, Characteristics of fatal single-vehicle crashes in Europe, Int. J. Crashworthiness 17 (6) (2012) 665-664.

[6] B.L. Roux, H. Rouanet, Multiple Correspondence Analysis, Sage Publications, Washington D.C, 2010.
[7] D.L. Hoffman, J. De Leeuw, Interpreting multiple correspondence analysis as a multidimensional scaling method, Mark. Lett. 3 (1992) 259–272.
[8] H. Fontaine, A typological analysis of pedestrian accidents, Presented at the 7th workshop of ICTCT, Paris, 26–27 October 1995.
[9] T.F. Golob, D.A. Hensher, The trip chaining activity of Sydney residents: a cross-section assessment by age group with a focus on seniors, J. Transp. Geogr. 15 (4) (2007).
[10] R. Factor, G. Yair, D. Mahalel, Who by accident? The social morphology of car accidents, Risk Anal. 30 (9) (2010).
[11] List of accident research and model development, https://dl.dropboxusercontent.com/u/13642258/acc_research1.html (Last retrieved on June 24, 2015).
[12] M. Greenacre, J. Blasius, Multiple Correspondence Analysis and Related Methods, Chapman & Hall/CRC, FL, 2006.
[13] L. Lebart, A. Morineau, K. Warwick, Multivariate descriptive statistical analysis correspondence analysis and related techniques for large matrices, Wiley Series in Probability and Mathematical Statistics, Applied Probability and Statistics, John Willey and Sons, New York, USA, 1984.
[14] S. Weller, A. Romney, Metric scaling—correspondence analysis, in: M.S. Lewis-Beck (Ed.), SAGE University Papers Series on Quantitative Applications in the Social Sciences no. 07-75, Sage, Newbury Park, CA, USA, 1990.
[15] 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).
[16] F. Husson, J. Josse, S. Le, J. Mazet, 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).
[17] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer New York, 2009.