Factor Association using Multiple Correspondence Analysis in Vehicle-pedestrian Crashes

Subasish Das (Corresponding author)
PhD. Student
Systems Engineering
University of Louisiana
Lafayette, LA 70504
Email: sxd1684@louisiana.edu
Phone: 225-288-9875
Fax: 337-739-6688

Xiaoduan Sun, Ph.D., and P.E.
Professor
Civil Engineering Department
University of Louisiana
Lafayette, LA 70504
Email: xsun@louisiana.edu
Phone: 337-482-6514
Fax: 337-739-6688

Word Count: 6,000 including 4 figures and 5 tables

Submitting to the 94th TRB Annual Meetings for Presentation and Publication under
Pedestrians Research (ANF10)

ABSTRACT

In the U.S., around 14% of total crash fatalities are pedestrian related. In 2011, 4,432 pedestrians were killed and 69,000 pedestrians were injured in vehicle-pedestrian crashes in the U.S. Vehicle-pedestrian crashes have become a key concern in Louisiana due to the high percentage of fatalities in recent years. In 2012, pedestrians accounted for 17% of total crash fatalities in the state. This research uses Multiple Correspondence Analysis (MCA), an exploratory data analysis method used for detecting and representing underlying structures in a categorical data set, to analyze eight years (2004-2011) of vehicle-pedestrian crashes in Louisiana. Pedestrian crash data is best represented as transactions of multiple categorical variables, so MCA would be a unique choice to determine the relationship of the variables and their significance. The findings indicated several non-trivial focus groups such as drivers with high occupancy vehicles, female drivers in bad weather conditions, and drivers distracted by mobile phone use. Other key associated factors were hillcrest roadways, dip/hump aligned roadways, roadways with multiple lanes, and roadways with no lighting at night. Male drivers were seen to be more inclined towards severe and moderate injury crashes. Fatal pedestrian crashes were correlated to two-lane roadways with no lighting at night. This method helped to measure significant contributing factors and degrees of association between the factors by analyzing the systematic patterns of variations with categorical datasets of pedestrian crashes. The findings from this study will be helpful for transportation professionals to improve the strategy of counter-measure selection.

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

INTRODUCTION

New policies tend to encourage safer and more effective travel for all roadway users in order to make transportation systems more sustainable and efficient. In 2011, 4,432 pedestrians were killed and 69,000 pedestrians were injured in vehicle-pedestrian crashes in the United States [1]. Improvement of pedestrian safety is one of the top-most priorities in the American Association of State Highway and Transportation (AASHTO) Strategic Highway Safety Plan (SHSP) [2].

A traffic crash is considered a rare, random, multifactor event always preceded by a state in which one or more roadway users fail to cope with the current environment. Any individual crash is the outcome of a series of events. Although each individual crash is unique in nature, there exists a common occurrence of a few features in several individual crashes [3]. One of the most important tasks in highway safety analysis is the identification of the most significant factors that are related to crashes. Multiple Correspondence Analysis (MCA) is a unique method to present the relative closeness of the categorical variables from any dataset. The traditional hypothesis testing is designed to verify *a priori* hypotheses regarding relationships between variables, but MCA is used to identify systematic relationships between variables and variable categories with no *a priori* expectations. The main scope of MCA is that it uniquely simplifies complex data and extracts significant knowledge from the information in the data that assumption-based statistical data analysis fails to collect. Moreover, it has a specific feature similar to multivariate treatment of the data through concurrent considerations of multiple categorical variables that would not be detected in a series of pair-wise comparisons of the variable. As pedestrian crash data can be represented as transactions of multiple categorical variables, MCA would be a good option to determine the relationship of the variables and their significance.

The vehicle-pedestrian crash statistics from Louisiana call for instant and advanced solutions to alleviate safety concerns for the pedestrians. The objective of this study was the application of MCA on vehicle-pedestrian crashes to (1) identify the relative closeness of the key association factors, (2) find important nontrivial association between the key factors and (3) provide intuitions to select better countermeasures to improve pedestrian safety. Improving pedestrian safety is crucial to accomplishing the state's 'Destination Zero Deaths' goal and the MCA method used in this paper will help find the relative closeness of the key association factors so that necessary actions can be taken to improve the strategy of pedestrian safety.

LITERATURE REVIEW

MCA has been popular in French scientific literature and has obtained a high level of development and use. Although less used in English scientific literature, the method has received increasing attention recently in the field of social science and marketing research. I.P. Benzecri made MCA, a multivariate statistical approach based on the correspondence analysis (CA) method that is popular among scientists. MCA, one of the main standards of geometric data analysis (GDA), is also referred to as the Pattern Recognition Method which treats arbitrary data sets as combination of points in *n*-dimensional space. However, in the field of multivariate traffic safety data analysis, geometric methods have rarely been used. MCA is hardly utilized in crash analysis. In fact, Roux and Rouanet pointed out 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 [4].

Fontaine was the first to use MCA for a typological analysis of pedestrian-related crashes [5]. The classification of pedestrians involved in crashes was divided into four major groups. The typology produced by this analysis revealed 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 [6]. 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 [7]. Das and Sun used the MCA method to analyze eight years (2004-2011) of single-vehicle fatal crashes in Louisiana in order to identify the important contributing factors and their degree of association [8].

Existing literature reveals an extensive variety of contributing factors in vehicle-pedestrian crashes. The key variables associated with vehicle-pedestrian crashes according to the earlier related studies are: higher speed limit (30 mph or over) [9, 10], absence of lighting at night [11], pedestrian visibility [12, 13], and certain age groups [14-15].

After performing a careful investigation on the closely associated research, it is found that a detailed study on the relative closeness of the key association factors of vehicle-pedestrian crashes in the U.S. has not yet been performed. This study attempts to determine the significant combinations of the variables for vehicle-pedestrian crashes by applying MCA which could help state agencies determine effective and efficient crash counter-measures.

METHODOLOGY

Theory of MCA

The mathematical theory development for MCA is complex in nature. We don't need to define response and dependable variables. It requires the construction of a matrix based on pairwise cross-tabulation of each variable. For a table with qualitative or categorical variables, 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). MCA can generate the spatial distribution of the points by different dimensions based on these three categories. It produces two combinations of points as shown in Figure 1: the combination of individual transactions and the combination of categories [4]. A combination of points can be compared with a geographic map with the same distance scale in all directions. A geometric diagram cannot be strained or contracted along a particular dimension. Thus, the basic property of any combination of points can be known from its dimensionality. Usually the two-dimensional combination is convenient for investigating the points lying on the plane. The complete combinations are generally referred to by their principal dimensions which are ranked in descending order of significance. MCA aims to create a combination of groups put together from a large dataset. The conventional MCA procedure is exhibited in Figure 1.

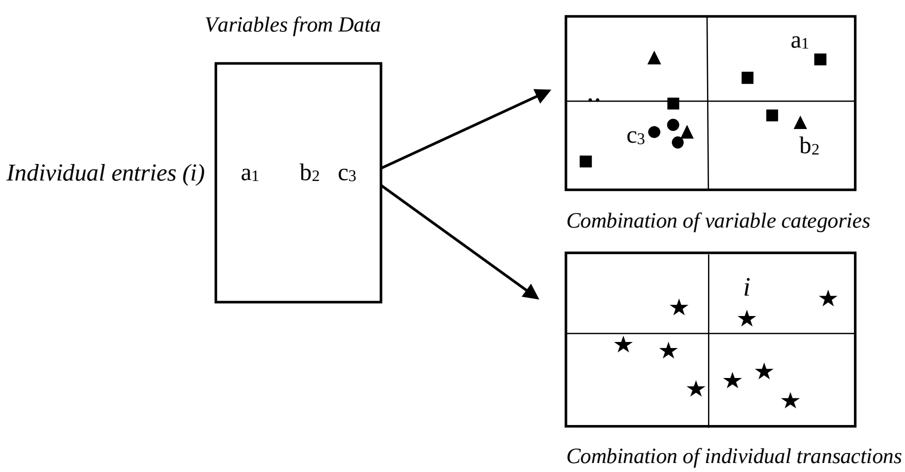

FIGURE 1 MCA method.

First we need to consider *P* as the number of variables and *I* as the number of transactions. The matrix will look like *"I multiplied by P"*, a table for all categorical values. If T_p is the number of categories for variable *p*, the total number of categories for all variables is, $T=\sum_{p=1}^{P}T_p$. We will generate another matrix *"I multiplied by T"* where each of the variables will have several columns to show all of their possible categorical values.

Now we need to consider category *k* associates with various individual records which can be denoted by n_k ($n_k > 0$), where $f_k = n_k/n$ = relative frequency of individuals associated with *k*. The values of f_k will create a row profile. The distance between two individual records is created by the variables for which both have different categories. For variable *p*, individual record *i* contains category *k* and individual record i' contains category k' which is different from *k*. The part of the squared distance between individual records *i* and i' for variable *p* is

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

The overall squared distance between *i* and i' is

$$d^2(i,i')=\frac{1}{P}\sum_{p\in P}d_p{}^2(i,i') \tag{2}$$

The set of all distances between individual records determines the combination of individuals consisting of *n* points in a space. The dimensionality of the space is *L*, where $L\leq K-P$. We assume that $n\geq L$. If M^i denotes the point representing individual *i* and *G* is the mean point of the combination, the squared distance from point M^i to point *G* is defined as

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

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

The cloud of categories is a weighted combination 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 *nP*. The relative weight w_k for point M^k is $w_k = n_k/(nP) = f_k/P$; for each variable, the sum of the relative weights of category points is 1/P, hence for the whole set the sum is 1.

$$w_k=\frac{n_k}{np}=\frac{f_k}{p}\qquad with\ \sum_{k\in K_q}w_k=\frac{1}{p}\qquad and\ \sum_{k\in K}w_k=1$$

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

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

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

The actual computations in MCA are performed on the inner product of this matrix known as the 'Burt Table'. The MCA calculations and two-dimensional plot visualizations in this paper were performed by using open source statistical 'R Version 3.02' software [16]. The authors used the 'FactoMineR' package for its usage convenience to analyze the dataset [17]. The datasets were studied according to the variables and their categories. More emphasis was given to studying the categories, as categories represent both variables and a group of individual records.

Descriptive Data Analysis

To achieve the study objectives, this study used state maintained vehicle-pedestrian crash data compiled from 2004 through 2011 in the state of Louisiana. 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. The pedestrian dataset was merged again with this merged dataset to get a complete profiling of the pedestrian related crashes.

In the crash database, there are numerous variables that are not pertinent to this research, such as: the VIN, driver's license number, database manager's name, police report number, etc. In order to focus on the meaningful analysis, a set of key variables were selected, such as: the roadway geometrics (alignment and lighting), collision type, environmental factors (weather), driver-related factors (driver gender, age and condition), number of occupants, and pedestrian-related factors (pedestrian gender, age, condition and severity). The variable section method used the research findings of the previous related research with engineering judgment.

An initial analysis indicated 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 roadways with straight and level alignment, 76% of the crashes were single-vehicle crashes, and 78% of crashes were single-occupant crashes. From Table 1, it is seen that 61% of pedestrians involved in crashes were male, which was higher than the general trend (around 50 to 55% of traffic crashes involved male drivers in Louisiana). The not-too-skewed variables include collision type, pedestrian injury, and lighting condition.

Multiple Correspondence Analysis

MCA can be explained as a graphical representation by producing a solution in which most associated categories are plotted close together and unassociated ones are plotted far apart. Graphical representations help to perceive and interpret data easily as they effectively summarize large, complex datasets by simplifying the structure of the associations between variables and providing a universal view of the data [4]. Points (categories) that are close to the mean value 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 combinations, while those with different distributions are plotted some distance apart. Hence, the dimensions are interpreted by the positions 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 [18].

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 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 0.24. 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 2, eigen values and percentages of variance of the first 10 dimensions are revealed. It can also be seen that there is a steady decrease in eigen values. The first principal axis explained 5.4% of the principal inertia, the second principal axis explained 4.7 % (hence 10.10% in total), and none of the remaining principal axes explained more than 4.7%. As the first plane (with dimensions 1 and 2) represented the largest inertia, only its results were presented and discussed.

The coordinates of the first five dimensions for the top ten categories are shown in Table 3. The variables with significance in two dimensions are listed in Table 4. 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

TABLE 1 Distribution of Rainy Weather Crashes by Key Variables

Categories	Frequency	Percentage	Categories	Frequency	Percentage
Alignment (Align.)			**Pedestrian Injury (Ped. Inj.)**		
Straight-Level	10750	93.45%	Fatal	801	6.96%
Curve-Level	360	3.13%	Severe	902	7.84%
On Grade	174	1.51%	Moderate	3877	33.70%
Dip, Hump	9	0.08%	Complaint	4156	36.13%
Hillcrest	64	0.56%	No Injury	1767	15.36%
Unknown (Unk.)	146	1.27%	**Number of Occupants (Num. Occ.)**		
Light			One	9021	78.42%
Daylight	6272	54.52%	Two	1626	14.14%
Dark - No Street Lights	1442	12.54%	Three	535	4.65%
Dark - Street Light	3231	28.09%	Four	164	1.43%
Dusk, Dawn	358	3.11%	Five or more	126	1.10%
Unknown (Unk.)	200	1.74%	Unknown (Unk.)	31	0.27%
Collision			**Number of Lanes (Num. Lanes)**		
Single Vehicle	4825	41.95%	Two	1571	13.66%
Rear End	466	4.05%	Four	2102	18.27%
Right Angle	799	6.95%	Six	432	3.76%
Right Turn	75	0.65%	Eight	16	0.14%
Sideswipe	493	4.29%	No Info.	7382	64.17%
Left Turn	209	1.82%	**Driver Distraction (Dr. Distract)**		
Head-On	185	1.61%	Not Distracted	5888	51.19%
Unknown (Unk.)	4451	38.69%	Outside Vehicle	406	3.53%
Weather			Cell Phone	83	0.72%
Clear	8770	76.24%	Inside Vehi.	158	1.37%
Abnormal	2590	22.52%	Electronic Device	10	0.09%
Unknown (Unk.)	143	1.24%	Unknown (Unk.)	4958	43.10%
Pedestrian Gender (Ped. Gender)					
Female	3738	32.50%			
Male	6958	60.49%			
Unknown (Unk.)	807	7.02%			

** In the parenthesis, the coded name of the variables is mentioned.*

other. Correlated variables provide redundant information and therefore some of them can be removed. The categories with significance in two dimensions are listed in Table 5. The most discriminant variables for dimension 1 are: weather, alignment, and lighting; regarding dimension 2 the most discriminant variables are: pedestrian injury, pedestrian gender and lighting. By observing the relative closeness of the variables, it is found that the number of lanes, types of collision, driver distraction and number of occupants are closer in the two dimensional space. A more detailed exploration of the variable categories

TABLE 2 Inertia Values for Top Ten Dimensions

Dimensions	Eigenvalue	Percentage of Variance	Cumulative Percentage of Variance
dim 1	0.2349	5.4197	5.4197
dim 2	0.2030	4.6836	10.1032
dim 3	0.1837	4.2394	14.3426
dim 4	0.1346	3.1060	17.4487
dim 5	0.1302	3.0038	20.4525
dim 6	0.1261	2.9091	23.3616
dim 7	0.1223	2.8228	26.1844
dim 8	0.1196	2.7608	28.9452
dim 9	0.1179	2.7210	31.6661
dim 10	0.1172	2.7038	34.3700

TABLE 3 Location of Top Ten Categories in First Five Dimensions

Category	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Align_Curve-Level	-0.4630	0.5324	1.1099	-0.7356	0.7293
Align_Dip, Hump	0.2036	-0.5090	-0.5219	0.6853	0.2727
Align_Hillcrest	-0.1604	0.4902	1.6002	2.2154	-0.6616
Align_On Grade	-0.5258	0.5053	1.3161	0.9319	1.1687
Align_Straight-Level	-0.0590	-0.0714	-0.0751	0.0034	-0.0384
Align_Unk	6.1688	3.1579	0.5568	-0.5575	-0.0906
Light_Dark - No Street Lights	-0.6664	0.9346	1.2191	-0.4113	0.4166
Light_Dark - Street Light	-0.0593	0.0052	0.0581	0.8377	-0.5543
Light_Daylight	0.0262	-0.3069	-0.3174	-0.3048	0.2143
Light_Dusk, Dawn	-0.1468	0.0490	-0.1905	-0.1736	-0.2926

TABLE 4 Significance of Key Variables on the First Plane

MCA Dimension 1			MCA Dimension 1		
Variable	**R^2**	**p.value**	**Variable**	**R^2**	**p.value**
Weather	0.5333	0.00E+00	Ped.Inj.	0.5246	0.00E+00
Light	0.5289	0.00E+00	Ped.Gender	0.4393	0.00E+00
Align	0.4973	0.00E+00	Light	0.2891	0.00E+00
Ped.Inj.	0.1415	0.00E+00	Weather	0.1486	0.00E+00
Ped.Gender	0.1298	0.00E+00	Align	0.1456	0.00E+00
Collision	0.0881	8.01E-225	Collision	0.1286	0.00E+00
Num.Lanes	0.0837	3.03E-216	Num.Lanes	0.1260	0.00E+00
Dr.Distract	0.0720	2.39E-183	Num.Occ	0.0216	4.01E-52
Num.Occ	0.0391	6.45E-97	Dr.Distract	0.0033	4.02E-07

TABLE 5 Significance of Key Categories on the First Plane

MCA Dimension 1			MCA Dimension 2		
Category	**Estimate**	**p.value**	**Category**	**Estimate**	**p.value**
Weather_Abnormal	-1.0697	0.00E+00	Ped.Inj._Unk	-1.0316	0.00E+00
Weather_Clear	-1.0611	0.00E+00	Ped.Gender_Unk	-0.7621	0.00E+00
Light_Dark - No Street Lights	-0.7455	0.00E+00	Align_Dip, Hump	-0.5375	3.88E-06
Align_On Grade	-0.6719	1.33E-106	Weather_Clear	-0.5341	0.00E+00
Align_Curve-Level	-0.6415	2.63E-129	Weather_Abnormal	-0.5018	1.76e-311
Align_Hillcrest	-0.4949	4.81E-33	Light_Daylight	-0.4443	0.00E+00
Light_Dusk, Dawn	-0.4937	6.04E-227	Align_Straight-Level	-0.3404	7.96E-38
Light_Dark - Street Light	-0.4513	0.00E+00	Collision_Right Turn	-0.3071	1.01E-12
Align_Straight-Level	-0.4457	1.58E-91	Light_Dark - Street Light	-0.3037	3.33E-245
Light_Daylight	-0.4099	0.00E+00	Light_Dusk, Dawn	-0.2840	5.65E-61
Ped.Inj._Fatal	-0.3765	9.13E-154	Num.Lanes_Unk	-0.2384	1.05E-14
Num.Occ_Four	-0.3579	1.90E-24	Num.Occ_One	-0.2114	8.66E-36
Align_Dip, Hump	-0.3185	9.12E-04	Dr.Distract_Cell Phone	-0.2110	1.13E-05
Num.Occ_Three	-0.3141	4.05E-38	Num.Lanes_Six	-0.2028	3.67E-14
Num.Occ_Five or more	-0.3072	2.43E-15	Collision_Left Turn	-0.1798	1.95E-11
Num.Occ_Two	-0.2630	2.67E-39	Num.Occ_Five or more	-0.1766	1.19E-06
Ped.Gender_Male	-0.2398	1.72E-253	Num.Occ_Three	-0.1519	2.50E-11
Dr.Distract_Not Distracted	-0.2309	5.75E-17	Ped.Inj._No Injury	-0.1214	4.16E-44
Ped.Gender_Female	-0.2118	2.79E-171	Collision_Rear End	-0.1140	2.64E-09
Collision_Single Vehicle	-0.1926	1.72E-62	Dr.Distract_Inside Vehi.	-0.0975	1.29E-02
Num.Lanes_Two	-0.1926	4.76E-14	Num.Occ_Two	-0.0817	1.32E-05
Num.Occ_One	-0.1732	6.92E-22	Align_On Grade	-0.0806	2.82E-02
Dr.Distract_Inside Vehi.	-0.1656	4.72E-05	Num.Occ_Four	-0.0776	1.81E-02
Dr.Distract_Cell Phone	-0.1445	3.76E-03	Align_Curve-Level	-0.0684	3.11E-02
Ped.Inj._Severe	-0.1097	2.29E-16	Light_Dark - No Street Lights	0.1150	2.04E-27
Dr.Distract_Outside Vehicle	-0.0995	2.61E-03	Ped.Inj._Complaint	0.1220	1.10E-109
Ped.Inj._Moderate	-0.0897	4.07E-29	Collision_Head-On	0.1234	1.28E-05
Ped.Inj._Complaint	-0.0665	2.60E-17	Collision_Unk	0.1271	1.84E-20
Ped.Inj._No Injury	0.0807	1.26E-10	Ped.Inj._Moderate	0.1641	8.76E-187
Num.Lanes_Six	0.1085	2.33E-04	Num.Lanes_Two	0.2096	1.71E-19
Collision_Unk	0.1250	1.16E-16	Ped.Inj._Severe	0.2432	1.84E-148
Num.Lanes_Unk	0.1807	1.01E-07	Num.Lanes_Eight	0.2591	2.15E-03
Ped.Gender_Unk	0.4517	0.00E+00	Dr.Distract_Electronic Device	0.2996	1.19E-02
Ped.Inj._Unk	0.5617	0.00E+00	Ped.Gender_Female	0.3348	0.00E+00
Dr.Distract_Electronic Device	0.6067	9.25E-07	Collision_Single Vehicle	0.3601	1.01E-248
Num.Occ_Unk	1.4155	2.29E-102	Ped.Gender_Male	0.4273	0.00E+00
Light_Unk	2.1003	0.00E+00	Ped.Inj._Fatal	0.6236	0.00E+00
Weather_Unk	2.1307	0.00E+00	Num.Occ_Unk	0.6992	4.46E-30
Align_Unk	2.5724	0.00E+00	Light_Unk	0.9170	0.00E+00
			Weather_Unk	1.0358	0.00E+00
			Align_Unk	1.1144	8.02E-136

will help more in discovering the underlying structure of the variables. The values from Table 5 indicate that dimension 1 and 2 both were governed by environmental and geometric variable categories. However, the highest estimate for dimension 2 was found for categories of pedestrian injury and gender.

RESULTS AND DISCUSSION

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. Figure 2 shows the

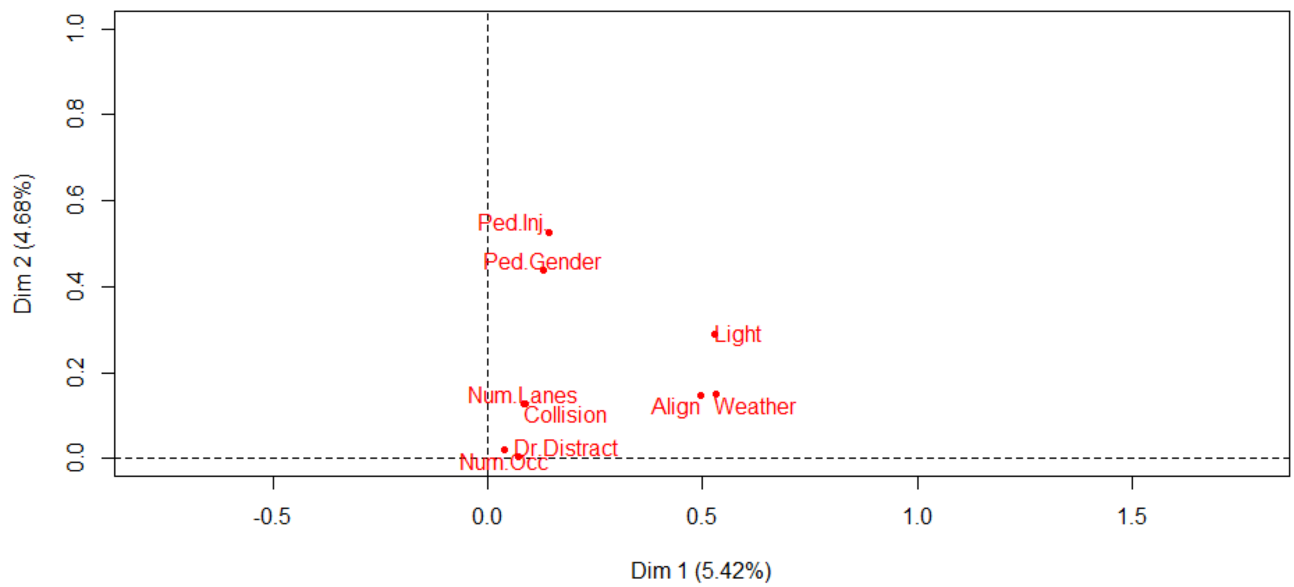

FIGURE 2 MCA plot for the variables.

relative closeness of all listed variables. 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 along with FactoMineR [19]. The combination selection was based on the relative closeness of the category location in the MCA plot. In the principle MCA plot, the distribution of the coordinates of all categories is shown [Figure 3]. This plot explores the positions of the variable categories on the two dimensional space according to the corresponding eigen values. When the categories are relatively closer they form a combination cloud.

The plots shown in Figures 4(a-d) are six different combinations that were selected from the MCA plot. Combination Cloud 1 combines a wider variety of variable categories: hillcrest aligned four-lane roadways, single vehicle collisions, severe and moderate pedestrian injuries, number of occupants 2 and 3, and male pedestrians. It indicates that hillcrest aligned four-lane roadways were prone to crashes with moderate and severe pedestrian injury. It also indicates that larger occupancy vehicles were often responsible for single vehicle-pedestrian crashes on this specific type of roadway. Combination Cloud 2 associates male pedestrians with moderate injury crashes while the number of occupants in the vehicles was two. It indicates that car occupancy has some role in pedestrian-related crashes. Combination clouds 3 and 4 seem rather insignificant because of their position near the center. However, the findings are interesting. Combination cloud 3 associated several factors: complaint injury of female pedestrians, dawn or dusk, abnormal weather and night time crashes in roadways with lighting. This non-trivial finding indicates a specific scenario for female pedestrians. Combination cloud 4 combines a few factors: clear weather, single occupant, six-lane straight-level aligned roadways, head-on collisions and driver distraction due to outside events. This finding is also non-trivial in nature. It specifically indicates a

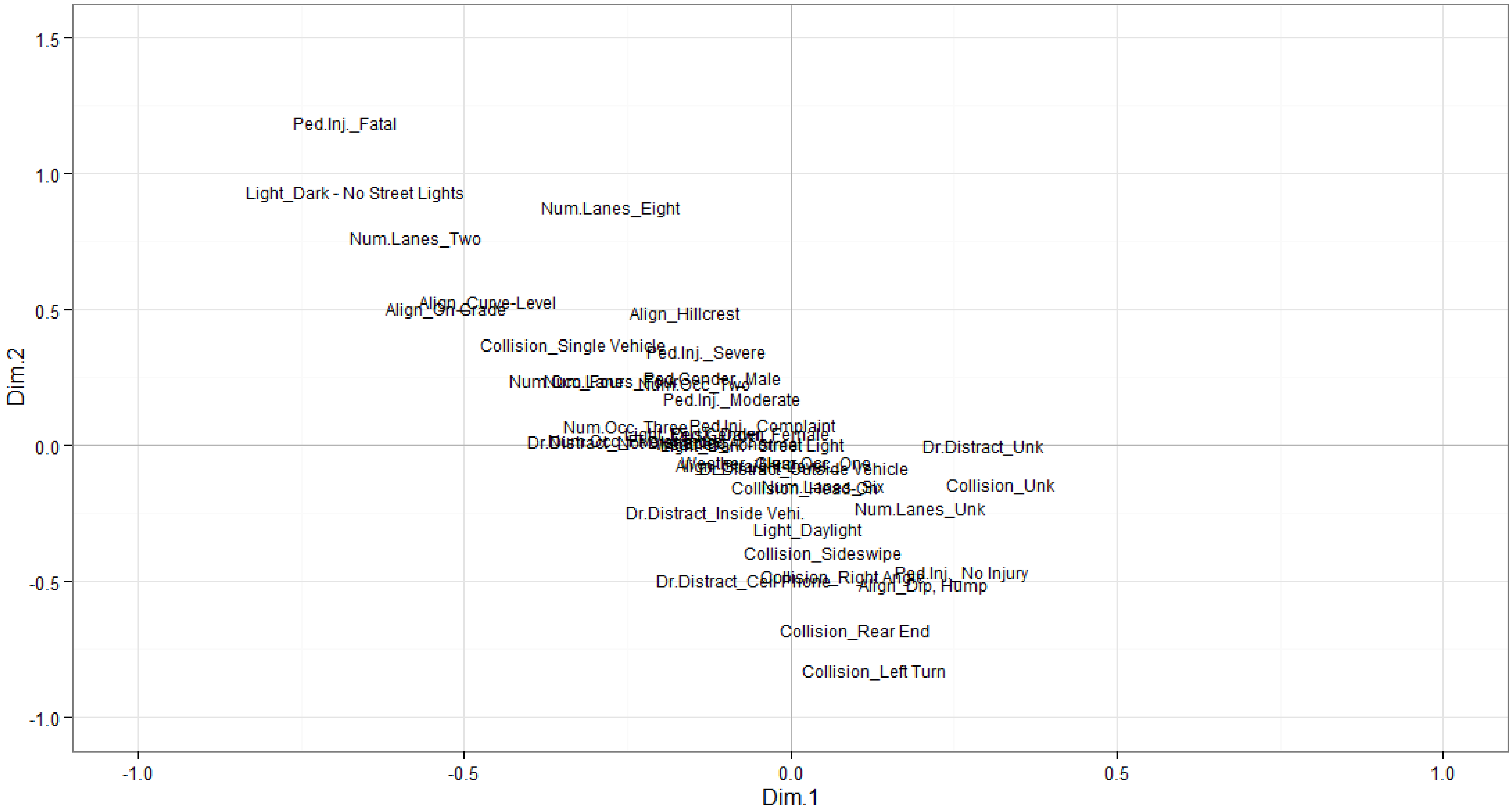

FIGURE 3 Principle MCA plot for the variable categories.

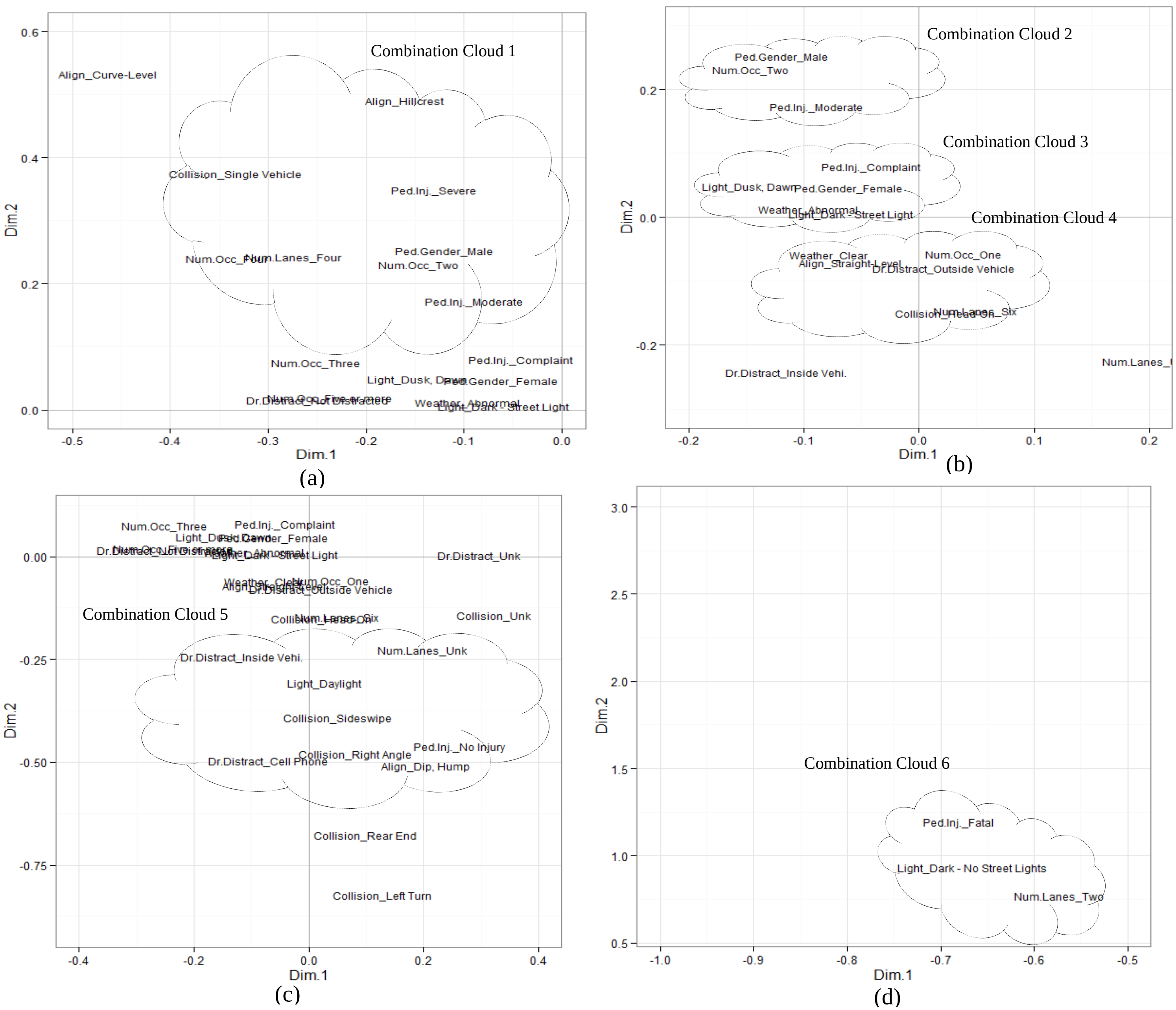

FIGURE 4 MCA plot for the variable categories.

particular roadway type where distraction happened due to an outside event. Moreover, the crashes involved head-on collisions which also implies involvement with other vehicles. Combination cloud 5 also associates different variable categories: driver distraction from mobile or inside equipment, daytime right angle and sideswipe crashes, dip/hump roadways with unknown information on lanes, and PDO pedestrian crashes. This combination indicates cell phone involvement in dip/hump aligned roadways. Combination cloud 6 associates three variable categories: fatal pedestrian crash, nighttime crash, and two-lane roadways with no lighting. It indicates that absence of lighting at night is a significant factor for pedestrian traffic severity. This cloud clearly indicates one major focus group on roadway geometrics.

The results presented in this paper demonstrate that MCA would be a good option to extract significant knowledge from pedestrian crash data. One of the limitations of the paper is that the findings were based on the two dimensional plane which explained only ten percent of inertia of the data. Explanations on more dimensions would process more knowledge extraction which was not performed in this study. As the initial variable selection was based on the previous research, exploration on other interesting variables was not performed. A more in-depth investigation on the appropriate variables would be a future scope of this research which would help explain a higher percentage of inertia of the data. If the crash database is more complete, MCA will generate more significant combination clouds from the dataset in an unsupervised way. The findings of this research will be helpful to traffic safety professionals in determining the hidden risk association group of variables in fatal single-vehicle crashes.

CONCLUSION

Conventional parametric models contain their own model assumptions and pre-defined underlying relationships between dependent and independent variables and assumption violation will lead to the model producing erroneous estimations. The MCA method, a non-parametric method, identifies systematic relationships among variables and variable categories with no *a priori* assumptions. Moreover, it uniquely simplifies large complex data and represents important knowledge from the dataset. PCA or SOMs are popular tools to describe numerical data; however MCA is a good option for exploratory data analysis for the categorical nature of vehicle-pedestrian crash occurrences.

The key focus of this study was to illustrate the applicability of MCA in identifying and representing underlying knowledge in large datasets of vehicle-pedestrian crashes. The findings indicate that MCA helps to cover multiple and diverse variable categories, showing if a relationship exists and how variable categories are related by producing analytical and visual results. Our study identified the groups of drivers and pedestrians as wells as geometric and environmental characteristics that are correlated to vehicle-pedestrian crashes. The findings revealed a few non-trivial risk groups from the analyzed dataset. The key combination groups are-

- Severe and moderate male pedestrian crashes on hillcrest aligned four-lane roadways associated with single-vehicle collision, and high occupancy vehicles (occupancy= 2 and 3).
- Moderate male pedestrian crashes when the occupancy of the vehicle is two.
- Complaint female pedestrian crashes associated with dawn or dusk, abnormal weather and night-time crashes in roadways with lighting.
- Head-on collisions on six-lane straight-level aligned roadways associated with single occupant, clear weather, single occupancy, and driver distraction happened due to outside events.
- PDO pedestrian crashes on dip/hump roadways due to drivers' distraction from mobile phones accompanying with daytime right angle and sideswipe crashes, and unknown information on lanes.
- Fatal pedestrian crashes on two-lane roadways with no lighting at night. It implies that pedestrian behavior in darkness is a continuing traffic safety issue.

In particular the ability of MCA to deal with multidimensional data makes it particularly useful for exploring the factors influencing crash occurrences. The findings from this research shed light on the pattern recognition of vehicle-pedestrian crashes and expose new aspects in pedestrian safety and also point to potential future research considering more variables and large datasets from multiple states. The findings may seem trivial in places, but the findings are based on extensive data exploration method to execute statistically significant valid combination groups. So, the jurisdictions can take appropriate actions on the strategies for the combination groups. Crashes dominated by human factors can be scrutinized by exploring the current law and safety education system. Modifications in laws can be made to make the drivers and pedestrian less vulnerable to crashes. Associated geometric features can be examined for the safety performance and improvement can be done accordingly.

REFERENCES

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

[2] Strategic Highway Safety Plan. A Comprehensive Plan to Substantially Reduce Vehicle-Related Fatalities and Injuries on the Nation's Highways. AASHTO, Washington, D.C., 2005.

[3] Montella, A. Identifying Crash Contributory Factors at Urban Roundabouts and Using Association Rules to Explore Their Relationships to Different Crash Types. Accident Analysis and Prevention, 43(4), pp. 1451–1463, 2011.

[4] Roux B., and Rouanet H. Multiple Correspondence Analysis. Sage Publications, Washington D.C. 2010.

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

[6] 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).

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

[8] Das S, and Sun X.Yair G, Mahalel, D. Exploring Clusters of Contributing Factors for Single-Vehicle Fatal Crashes Through Multiple Correspondence Analysis. TRB 93rd Annual Meeting Compendium of Papers, Washington DC, 2014.

[9] Davis, G. Relating severity of pedestrian injury to impact speed in vehicle–pedestrian crashes: simple threshold model. Transportation Research Record: Journal of the Transportation Research Board, 1773 (1), pp. 108–113, 2001.

[10] Sze, N., and Wong, S. 2007. Diagnostic analysis of the logistic model for pedestrian Injury severity in traffic crashes. Accident Analysis and Prevention, 39 (6), pp. 1267–1278, 2007.

[11] Moudon, A., Lin, L., Jiao, J., Hurvitz, P., and Reeves, P. The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. Accident Analysis and Prevention 43 (1), pp. 11–24, 2011.

[12] Ulfarsson, G., Kim, S., and Booth, K. Analyzing fault in pedestrian–motor vehicle crashes in North Carolina. Accident Analysis and Prevention 42 (6), pp. 1805–1813, 2010.

[13] Sullivan, J., and Flannagan, M. Differences in geometry of pedestrian crashes in daylight and darkness. Journal of Safety Research 42 (1), pp. 33–37, 2011.

[14] Roudsari, B., Mock, C., Kaufman, R., Grossman, D., Henary, B., and Crandall, J. Pedestrian crashes: higher injury severity and mortality rate for light truck vehicles compared with passenger vehicles. Injury Prevention 10 (3), pp. 154–158, 2004.

[15] Eluru, N., Bhat, C., and Hensher, D. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis and Prevention 40 (3), pp. 1033–1054, 2008.

[16] 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.

[17] Husson F, Josse J, Le, S., and 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 21, 2014.

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

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