Factor Association with Multiple Correspondence Analysis in Vehicle–Pedestrian Crashes

Subasish Das and Xiaoduan Sun

In the United States, about 14% of total crash fatalities are pedestrian related. In 2012, 4,743 pedestrians were killed, and 76,000 pedestrians were injured in vehicle–pedestrian crashes in the United States. Vehicle–pedestrian crashes have become a key concern in Louisiana as a result of the high percentage of fatalities there in recent years. In 2012, pedestrians accounted for 17% of total crash fatalities in the state. This study used multiple correspondence analysis (MCA), an exploratory data analysis method used to detect and represent underlying structures in a categorical data set, to analyze 8 years (2004 to 2011) of vehicle–pedestrian crashes in Louisiana. Pedestrian crash data are best represented as transactions of multiple categorical variables, so the use of MCA was a unique choice to determine the relationship of the variables and their significance. The findings indicated several nontrivial focus groups (e.g., drivers with high-occupancy vehicles, female drivers in bad weather conditions, and drivers distracted by mobile phone use). The associated geometric factors were hillcrest roadways, dip or hump aligned roadways, roadways with multiple lanes, and roadways with no lighting at night. Male drivers were seen to be relatively susceptible to severe and moderate injury crashes. Fatal pedestrian crashes were correlated to two-lane roadways with no lighting at night. The MCA method helped measure significant contributing factors and degrees of association between the factors through the analysis of the systematic patterns of variation with categorical data sets of pedestrian crashes. The findings from this study will help transportation professionals improve countermeasure selection strategies.

New policies tend to encourage safer and more effective travel for all roadway users to make transportation systems more sustainable and efficient. In 2012, 4,743 pedestrians were killed and 76,000 pedestrians were injured in vehicle–pedestrian crashes in the United States (*1*). Improved pedestrian safety is one of the top priorities in the AASHTO Strategic Highway Safety Plan (*2*).

A traffic crash is considered a rare, random, multifactored event always preceded by a state in which one or more roadway users fails 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, the common occurrence exists of a few features in several individual crashes (*3*). One of the most important tasks in highway safety analysis is to identify the most significant factors that are related to crashes. Multiple correspondence analysis (MCA) is a unique method that presents the relative closeness of the categorical variables from any data set. Traditional hypothesis testing is designed to verify a priori hypotheses on 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, MCA has a specific feature similar to the multivariate treatment of the data through concurrent consideration of multiple categorical variables that would not be detected in a series of pairwise comparisons of the variable. Given that pedestrian crash data can be represented as transactions of multiple categorical variables, MCA is a good option to determine the relationship of the variables and their significance.

The vehicle–pedestrian crash statistics of Louisiana call for instant and advanced solutions to ease safety concerns for pedestrians. The objective of this study was the application of MCA on vehicle–pedestrian crashes to (*a*) identify the relative closeness of the key association factors, (*b*) find important nontrivial associations between the key factors, and (*c*) provide intuitions to select better countermeasures to improve pedestrian safety. Improvement of pedestrian safety is crucial to accomplish the state's "Destination Zero Deaths" goal, and the MCA method used in this study will help to find the relative closeness of the key association factors so that necessary actions can be taken to improve pedestrian safety strategies.

LITERATURE REVIEW

MCA has been popular in French scientific literature and thus 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 fields of social science and marketing research. Benzécri developed MCA, a multivariate statistical approach, on the basis of the correspondence analysis method that is popular among scientists. MCA, one of the main standards of geometric data analysis, also is referred to as the pattern recognition method, which treats arbitrary data sets as a combination of points in *n*-dimensional space. However, in the field of multivariate traffic safety data analysis, geometric methods rarely have been used. Roux and Rouanet pointed out that this method, while a powerful tool to analyze a full-scale research database, was hardly discussed and therefore underused 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

S. Das, Systems Engineering Doctoral Program and X. Sun, Department of Civil Engineering, University of Louisiana at Lafayette, 131 Rex Street, Lafayette, LA 70504. Current affiliation for S. Das: Texas A&M Transportation Institute, Texas A&M University System, 3135 TAMU, College Station, TX 77843-3135. Corresponding author: S. Das, s-das@tti.tamu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2519, Transportation Research Board, Washington, D.C., 2015, pp. 95–103.
DOI: 10.3141/2519-11

involved in crashes was divided into four major groups. The typology produced by this analysis revealed correlations between criteria without necessarily the indication of 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 nonmonotonic relationships between socioeconomic descriptors and measures of travel behavior (*6*). Factor et al. conducted a study of 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 8 years (2004 to 2011) of single-vehicle fatal crashes in Louisiana to identify the important contributing factors and their degree of association (*8*).

The 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 were higher speed limit (30 mph or more) (*9, 10*), absence of lighting at night (*11*), pedestrian visibility (*12, 13*), and certain age groups (*14, 15*).

After a careful investigation of the closely associated research, it was found that a detailed study of the relative closeness of the key associated factors of vehicle–pedestrian crashes in the United States had not been performed. This present study attempted to determine the significant combinations of the variable categories for vehicle–pedestrian crashes through MCA, which could help state agencies determine effective and efficient crash countermeasures.

METHODOLOGY

Theory of MCA

The mathematical theory development for MCA is complex in nature. In this method, there is no need to define response and explanatory variables. MCA requires the construction of a matrix on the basis of the pairwise cross-tabulation of each variable. For a table with qualitative or categorical variables, MCA can be explained with an individual record (in row), i, where three variables (represented by three columns) have three category indicators (a_1, b_2, and c_3). MCA can generate the spatial distribution of the points with different dimensions on the basis of these three categories. This method 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 in the investigation of the points that lie on the plane. The complete combinations in general are 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 data set.

First, P is the number of variables, and I is 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 $T = \sum_{p=1}^{P} T_p$. Another matrix will then be generated as "I multiplied by T" in which each of the variables will have several columns to show all of its possible categorical values.

Category k is considered to be associated with various individual records, which can be denoted by n_k ($n_k > 0$), where f_k is n_k/n and is equal to the 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 each has different categories. For variable p, 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 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, which consist of n points in a space. The dimensionality of the space is L, where $L \le K - P$. It was assumed that $n \ge L$. If M^i denotes the point that represents individual I, and

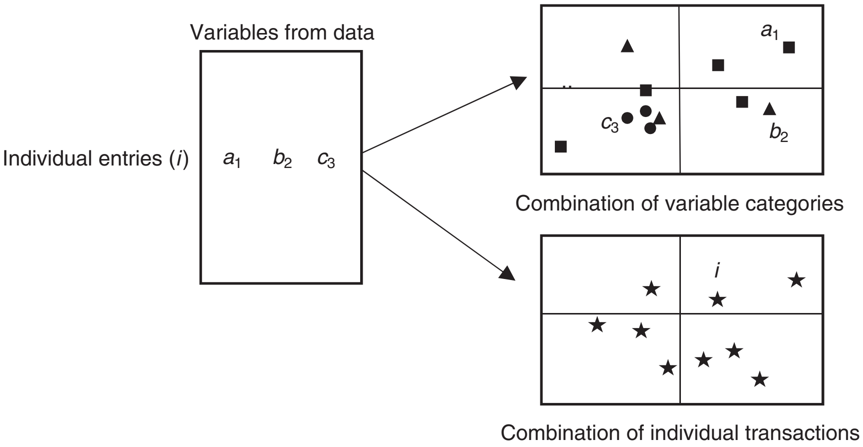


FIGURE 1 MCA method.

Z is the mean point of the combination, the squared distance from point M^i to Z can be defined as

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

where K_i is the response pattern of individual i (i.e., the set of the P categories associated with individual record i).

The cloud of categories is considered 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. Thus for the whole set K the sum is nP. The relative weight for point M^k is w_k, which equals f_k/P. For each variable, the sum of the relative weights of category points is $1/P$. Thus for the whole set the sum is 1.

$$w_k = \frac{n_k}{nP} = \frac{f_k}{P}$$

with

$$\sum_{k\in K_p} w_k = \frac{1}{P}$$

and

$$\sum_{k\in K} w_k = 1$$

If $n_{kk'}$ denotes the number of individual records that have both of the categories k and k', the squared distance between M^k and $M^{K'}$ is

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

The numerator is the number of individual records associated with either k or k' but not both. For two 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 (*4*).

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 study were performed through the use of open-source statistical R Version 3.02 software (*16*). The FactoMineR package was used for its convenience to analyze the data sets (*17*). The data sets were studied according to the variables and their categories. Emphasis was given to the study of the categories, because they represented variables and a group of individual records.

Descriptive Data Analysis

To achieve its objectives, this study used state-maintained vehicle–pedestrian crash data compiled from 2004 through 2011 in the state of Louisiana. The primary data set was prepared through the merger of three tables (i.e., crash table, Department of Transportation and Development table, and vehicle table) from the Microsoft Access data set. The pedestrian data set was merged again with this merged data set to create a complete profile of the pedestrian-related crashes.

In the crash database, numerous variables were not pertinent to this research (e.g., vehicle identification number, driver's license number, database manager's name, police report number). To focus on the meaningful analysis, a set of key variables was selected [e.g., roadway geometrics (alignment and lighting), collision type, environmental factors (weather), driver-related factors (driver gender, age, condition), number of vehicle occupants, and pedestrian-related factors (pedestrian gender, age, condition, severity)]. To select the variables, the findings of previous, related research were used in combination with engineering judgment.

An initial analysis indicated that some variables were highly skewed, which meant that most crashes fell into one of the two or more categorical values. For example, 94% of the crashes involved roadways with straight-level alignment, 76% occurred during clear weather, and 78% were single-occupant crashes. Table 1 shows that 61% of the pedestrians involved in crashes were men, a percentage that was higher than the general trend (i.e., 50% to 55% of traffic crashes involved male drivers in Louisiana). The not-too-skewed variables were collision type, pedestrian injury, and lighting condition.

MCA Explained

MCA can be explained as a graphical representation in which most associated categories are plotted close together, and unassociated ones are plotted far apart, on the basis of the calculated values. Graphical representations help make it possible to perceive and interpret data easily. These representations effectively summarize large, complex data sets through the simplification of the structure of the associations between variables, and they provide a universal and general view of the data (*4*). Points (categories) that are close to the mean value are plotted near the MCA plot's origin. Those that are more distant are plotted farther away. Categories with a similar distribution are presented near one another through the formation of combinations. Those with different distributions are plotted some distance apart. Thus the dimensions are interpreted by the positions of the points on the map, with their loading over the dimensions as crucial indicators. A two-dimensional depiction usually is sufficient to explain most of the variance in MCA (*18*).

The eigenvalues measure indicates how much of the categorical information is accounted for by each dimension; the higher the eigenvalue, the larger the amount of the total variance among the variables on that dimension. The largest possible eigenvalue for any dimension is 1. Usually, the first two or three dimensions contain higher eigenvalues than others. In this analysis, the maximum eigenvalue in Dimension 1 was 0.24. The similarly low eigenvalues in each dimension indicated that the variables in the crash data were heterogeneous. All carried to some extent unique information, which implied that a reduction in any of the variables might result in the loss of important information about the crash observations. The heterogeneity of the crash variables reflected the random nature of crash occurrences.

In Table 2, eigenvalues and percentages of variance of the first 10 dimensions are revealed. A steady decrease in eigenvalues also can be seen. The first principal axis explained 5.4% of the principal inertia, the second principal axis explained 4.7% (i.e., 10.10% in total), and none of the remaining principal axes explained more than 4.7%. Because the first plane (with Dimensions 1 and 2) represented the largest inertia, only its results were presented and discussed.

TABLE 1 Description of Key Variables

Category	Frequency	Percentage
Alignment (Align.)		
Straight–level	10,750	93.45
Curve–level	360	3.13
On grade	174	1.51
Dip, hump	9	0.08
Hillcrest	64	0.56
Unknown (Unk.)	146	1.27
Light		
Daylight	6,272	54.52
Dark—no street lights	1,442	12.54
Dark—street light	3,231	28.09
Dusk, dawn	358	3.11
Unknown (Unk.)	200	1.74
Collision		
Single vehicle	4,825	41.95
Rear end	466	4.05
Right angle	799	6.95
Right turn	75	0.65
Sideswipe	493	4.29
Left turn	209	1.82
Head on	185	1.61
Unknown (Unk.)	4,451	38.69
Weather		
Clear	8,770	76.24
Abnormal	2,590	22.52
Unknown (Unk.)	143	1.24
Pedestrian gender (Ped.Gender)		
Female	3,738	32.50
Male	6,958	60.49
Unknown (Unk.)	807	7.02
Pedestrian injury (Ped.Inj.)		
Fatal	801	6.96
Severe	902	7.84
Moderate	3,877	33.70
Complaint	4,156	36.13
No injury	1,767	15.36
Number of occupants (Num.Occ.)		
One	9,021	78.42
Two	1,626	14.14
Three	535	4.65
Four	164	1.43
Five or more	126	1.10
Unknown (Unk.)	31	0.27
Number of lanes (Num.Lanes)		
Two	1,571	13.66
Four	2,102	18.27
Six	432	3.76
Eight	16	0.14
No info.	7,382	64.17
Driver distraction (Dr.Distract)		
Not distracted	5,888	51.19
Outside vehicle	406	3.53
Cell phone	83	0.72
Inside vehicle	158	1.37
Electronic device	10	0.09
Unknown (Unk.)	4,958	43.10

NOTE: Info. = information. Coded names of variables are identified in parentheses.

The coordinates of the first five dimensions for the top 10 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 other. Correlated variables provide redundant information, and thus 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, lighting, and alignment. For Dimension 2, they are pedestrian injury, pedestrian gender, and lighting. Through observation of the relative closeness of the variables, it was found that the number of lanes, types of collision, driver distraction, and number of vehicle occupants were closer in the two-dimensional space than elsewhere. A more detailed exploration of the variable categories would be of help to discover the underlying structure of the variables. The values from Table 5 indicate that Dimensions 1 and 2 were governed by environmental and geometric variable categories. However, the highest estimate for Dimension 2 was found for the categories of pedestrian injury and gender.

TABLE 2 Inertia Values for Top 10 Dimensions

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

RESULTS AND DISCUSSION

The contribution of a category depends on data, whereas for a variable it depends only 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 relative closeness of all listed variables. The key focus of MCA is to provide an insight into the data set through information visualization. The popular graphical *R* package ggplot2 was used extensively, along with FactoMineR, to produce the informative MCA plots (*19*). The combination selection had its basis in the relative closeness of the category location in the MCA plot. In the principal MCA plot, the distribution of the coordinates of all categories is shown (Figure 3). This plot explores the positions of the variable categories in the two-dimensional space

TABLE 3 Location of Top 10 Categories in First Five Dimensions

	Coordinate by Dimension				
Category	1	2	3	4	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 First Plane

Variable	R^2	*p*-Value	Variable	R^2	*p*-Value
MCA Dimension 1			MCA Dimension 2		
Weather	.5333	0.00 E+00	Ped.Inj.	.5246	0.00 E+00
Light	.5289	0.00 E+00	Ped.Gender	.4393	0.00 E+00
Align.	.4973	0.00 E+00	Light	.2891	0.00 E+00
Ped.Inj.	.1415	0.00 E+00	Weather	.1486	0.00 E+00
Ped.Gender	.1298	0.00 E+00	Align.	.1456	0.00 E+00
Collision	.0881	8.01 E–225	Collision	.1286	0.00 E+00
Num.Lanes	.0837	3.03 E–216	Num.Lanes	.1260	0.00 E+00
Dr.Distract	.0720	2.39 E–183	Num.Occ.	.0216	4.01 E–52
Num.Occ.	.0391	6.45 E–97	Dr.Distract	.0033	4.02 E–07

TABLE 5 Significance of Key Categories on First Plane

Category	Estimate	*p*-Value	Category	Estimate	*p*-Value
MCA Dimension 1			MCA Dimension 2		
Weather_Abnormal	–1.0697	0.00 E+00	Ped.Inj._Unk	–1.0316	0.00 E+00
Weather_Clear	–1.0611	0.00 E+00	Ped.Gender_Unk	–0.7621	0.00 E+00
Light_Dark—No Street Lights	–0.7455	0.00 E+00	Align_Dip, hump	–0.5375	3.88 E–06
Align_On Grade	–0.6719	1.33 E–106	Weather_Clear	–0.5341	0.00 E+00
Align_Curve-Level	–0.6415	2.63 E–129	Weather_Abnormal	–0.5018	1.76 E–311
Align_Hillcrest	–0.4949	4.81 E–33	Light_Daylight	–0.4443	0.00 E+00
Light_Dusk, Dawn	–0.4937	6.04 E–227	Align_Straight-Level	–0.3404	7.96 E–38
Light_Dark—Street Light	–0.4513	0.00 E+00	Collision_Right Turn	–0.3071	1.01 E–12
Align_Straight-Level	–0.4457	1.58 E–91	Light_Dark—Street Light	–0.3037	3.33 E–245
Light_Daylight	–0.4099	0.00 E+00	Light_Dusk, Dawn	–0.2840	5.65 E–61
Ped.Inj._Fatal	–0.3765	9.13 E–154	Num.Lanes_Unk	–0.2384	1.05 E–14
Num.Occ_Four	–0.3579	1.90 E–24	Num.Occ_One	–0.2114	8.66 E–36
Align_Dip, hump	–0.3185	9.12 E–04	Dr.Distract_Cell Phone	–0.2110	1.13 E–05
Num.Occ_Three	–0.3141	4.05 E–38	Num.Lanes_Six	–0.2028	3.67 E–14
Num.Occ_Five or more	–0.3072	2.43 E–15	Collision_Left Turn	–0.1798	1.95 E–11
Num.Occ_Two	–0.2630	2.67 E–39	Num.Occ_Five or more	–0.1766	1.19 E–06
Ped.Gender_Male	–0.2398	1.72 E–253	Num.Occ_Three	–0.1519	2.50 E–11

(continued on next page)

TABLE 5 *(continued)* **Significance of Key Categories on First Plane**

Category	Estimate	*p*-Value	Category	Estimate	*p*-Value
MCA Dimension 1			MCA Dimension 2		
Dr.Distract_Not Distracted	–0.2309	5.75 E–17	Ped.Inj._No Injury	–0.1214	4.16 E–44
Ped.Gender_Female	–0.2118	2.79 E–171	Collision_Rear End	–0.1140	2.64 E–09
Collision_Single Vehicle	–0.1926	1.72 E–62	Dr.Distract_Inside Vehicle	–0.0975	1.29 E–02
Num.Lanes_Two	–0.1926	4.76 E–14	Num.Occ_Two	–0.0817	1.32 E–05
Num.Occ_One	–0.1732	6.92 E–22	Align_On Grade	–0.0806	2.82 E–02
Dr.Distract_Inside Vehicle	–0.1656	4.72 E–05	Num.Occ_Four	–0.0776	1.81 E–02
Dr.Distract_Cell Phone	–0.1445	3.76 E–03	Align_Curve-Level	–0.0684	3.11 E–02
Ped.Inj._Severe	–0.1097	2.29 E–16	Light_Dark—No Street Lights	0.1150	2.04 E–27
Dr.Distract_Outside Vehicle	–0.0995	2.61 E–03	Ped.Inj._Complaint	0.1220	1.10 E–109
Ped.Inj._Moderate	–0.0897	4.07 E–29	Collision_Head-On	0.1234	1.28 E–05
Ped.Inj._Complaint	–0.0665	2.60 E–17	Collision_Unk	0.1271	1.84 E–20
Ped.Inj._No Injury	0.0807	1.26 E–10	Ped.Inj._Moderate	0.1641	8.76 E–187
Num.Lanes_Six	0.1085	2.33 E–04	Num.Lanes_Two	0.2096	1.71 E–19
Collision_Unk	0.1250	1.16 E–16	Ped.Inj._Severe	0.2432	1.84 E–148
Num.Lanes_Unk	0.1807	1.01 E–07	Num.Lanes_Eight	0.2591	2.15 E–03
Ped.Gender_Unk	0.4517	0.00 E+00	Dr.Distract_Electronic Device	0.2996	1.19 E–02
Ped.Inj._Unk	0.5617	0.00 E+00	Ped.Gender_Female	0.3348	0.00 E+00
Dr.Distract_Electronic Device	0.6067	9.25 E–07	Collision_Single Vehicle	0.3601	1.01 E–248
Num.Occ_Unk	1.4155	2.29 E–102	Ped.Gender_Male	0.4273	0.00 E+00
Light_Unk	2.1003	0.00 E+00	Ped.Inj._Fatal	0.6236	0.00 E+00
Weather_Unk	2.1307	0.00 E+00	Num.Occ_Unk	0.6992	4.46 E–30
Align_Unk	2.5724	0.00 E+00	Light_Unk	0.9170	0.00 E+00
			Weather_Unk	1.0358	0.00 E+00
			Align_Unk	1.1144	8.02 E–136

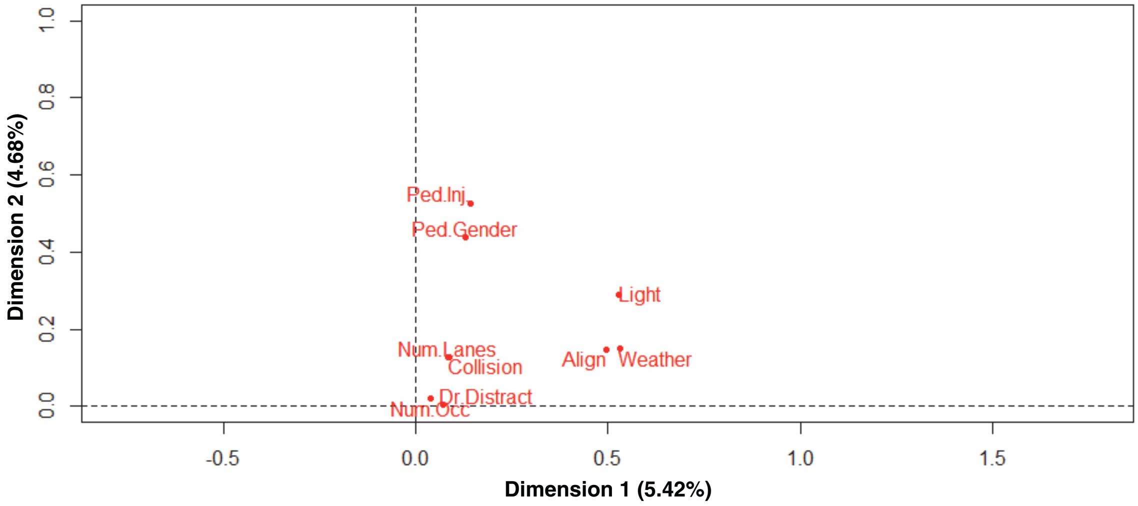


FIGURE 2 MCA plot for variables.

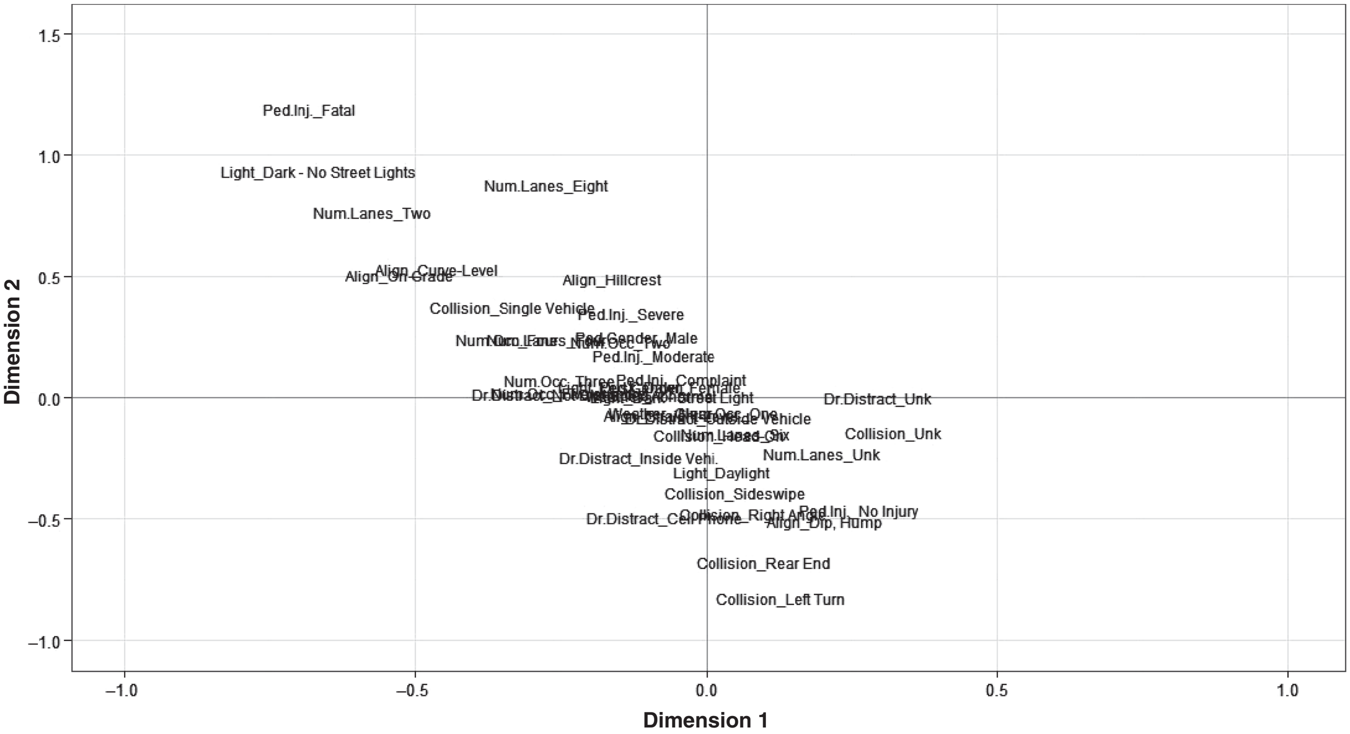


FIGURE 3 Principal MCA plot for variable categories.

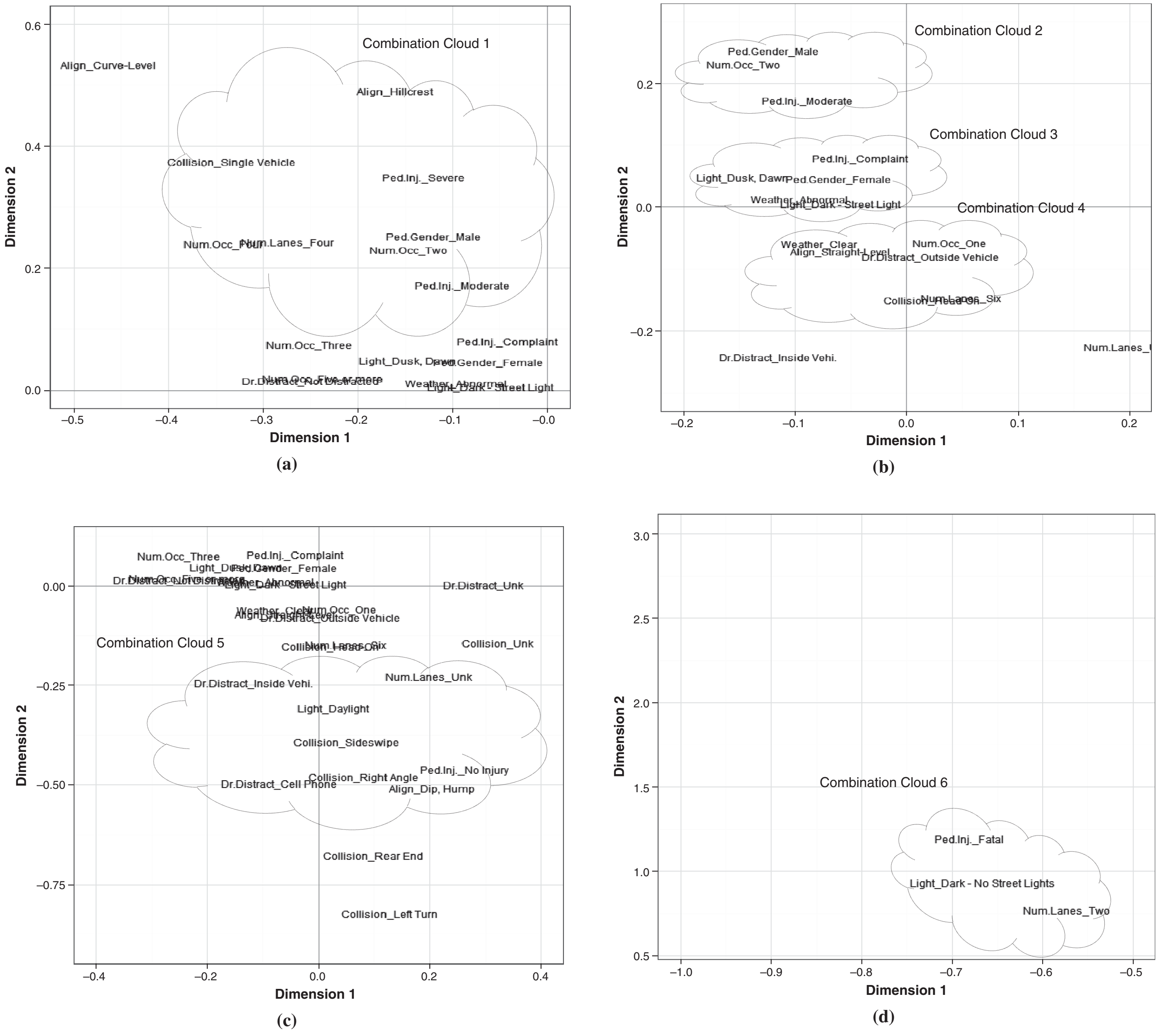
according to the corresponding eigenvalues. When the categories are relatively closer, they form a combination cloud.

The plots shown in Figure 4 are six combinations selected from the MCA plot. Combination Cloud 1 in Figure 4*a* combines a wider variety of variable categories: hillcrest aligned four-lane roadways, single vehicle collisions, severe and moderate pedestrian injuries, number of occupants two and three, 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 often were responsible for single vehicle–pedestrian crashes on this specific type of roadway. Combination Cloud 2 in Figure 4*b* associates male pedestrians with moderate injury crashes, while the number of occupants in the vehicles is two. It indicates that car occupancy has some role in pedestrian-related crashes. Combination Clouds 3 and 4 in Figure 4*b* seem rather insignificant because of their positions near the center. However, Combination Cloud 3 associated several factors: complaint injury of female pedestrians, dawn or dusk, abnormal weather, and nighttime crashes in roadways with lighting. This nontrivial finding indicated a specific scenario for female pedestrians. Combination Cloud 4 combined a few factors: clear weather, single occupant, six-lane straight-level aligned roadways, head-on collisions, and driver distraction as the result of outside events. This finding also was nontrivial in nature. It specifically indicated a particular roadway type in which distraction occurred as the result of an outside event. Moreover, the crashes involved head-on collisions, which implied the involvement of other vehicles. Combination Cloud 5 in Figure 4*c* also associates different variable categories: driver distraction from mobile or inside equipment, daytime right angle and sideswipe crashes, dip or hump roadways with unknown information on lanes, and property-damage-only pedestrian crashes. This combination indicated the impact of cell phone use in dip and hump aligned roadways. Combination Cloud 6 in Figure 4*d* 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 and clearly identifies one major focus group on roadway geometrics.

The results presented in this paper demonstrated that MCA would be a good option to extract significant knowledge from pedestrian crash data. One of the limitations of the study was that the findings had their bases in the two-dimensional plane, which explained only 10% of the inertia of the data. Explanations on more dimensions would process more knowledge extraction, which was not done in this study. Because the initial variable selection had its basis in previous research, other variables of interest were not explored. A more in-depth investigation into the appropriate variables could form the future scope of this research, which would help explain a higher percentage of inertia in the data. If the crash database is more complete, MCA will generate more significant combination clouds from the data set in an unsupervised way. The findings of this research will help traffic safety professionals to determine the hidden risk association group of variables in pedestrian crashes.

CONCLUSIONS

Conventional parametric models contain their own model assumptions and predefined fundamental relationships between response and explanatory variables, and assumption violation will lead to the model's production of erroneous estimations. The MCA method, a nonparametric 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 data set. Principal component analysis, or the self-organizing map, is a popular tool to describe numerical data,



(a)

(b)

(c)

(d)

FIGURE 4 MCA plots for variable categories.

but MCA was a good option for an exploratory data analysis of the categorical variables of vehicle–pedestrian crash occurrences.

The key focus of this study was to illustrate the applicability of MCA to identify and represent underlying knowledge in large data sets of vehicle–pedestrian crashes. The findings indicated that MCA helped to cover multiple and diverse variable categories through its showing of nontrivial relationships. The current research identified the groups of drivers and pedestrians as well as geometric and environmental characteristics that correlated to vehicle–pedestrian crashes. The findings revealed a few nontrivial risk groups from the analyzed data set. 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 of two or three);
- Moderate injury of male pedestrians when the occupancy of the vehicle is two;
- Complaint injury of female pedestrians associated with dawn or dusk, abnormal weather, and nighttime 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 from outside events;
- No injury pedestrian crashes on dip and hump roadways as a result of driver distraction from mobile phone, accompanied by daytime right angle and sideswipe crashes and unknown information about lanes; and
- Fatal pedestrian crashes on two-lane roadways with no lighting at night; this result implies that pedestrian behavior in darkness is a continuing traffic safety issue.

The capability of MCA to deal with multidimensional data makes it particularly useful to explore the factors that influence crash occurrences. The findings from this research shed light on the pattern recognition of vehicle–pedestrian crashes, exposed new aspects of pedestrian safety, and also pointed to potential research to consider more variables and large data sets from multiple states. The findings of this study might have seemed trivial in places, but the findings had their basis in an extensive data exploration method to execute statistically significant and valid combination groups so that jurisdictions could take appropriate action on safety strategies for the combination groups. Crashes dominated by human factors can be scrutinized through the exploration of the current law and safety education system. Modifications in the law may be made to make drivers and pedestrians less vulnerable to crashes. Associated geometric features can be examined for safety performance, and improvements can be made accordingly.

REFERENCES

1. NHTSA. *Traffic Safety Facts: 2012 Data. Pedestrians.* DOT HS 81 1 888. April, 2014. http://www-nrd.nhtsa.dot.gov/Pubs/811888.pdf. Accessed March 10, 2015.
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,* Vol. 43, No. 4, 2011, pp. 1451–1463.
4. Roux, B., and H. Rouanet. *Multiple Correspondence Analysis.* Sage Publications, Thousand Oaks, Calif., 2010.
5. Fontaine, H. *A Typological Analysis of Pedestrian Accidents.* Presented at 7th Workshop of International Co-Operation on Theories and Concepts in Traffic Safety, Paris, 1995.
6. Golob, T.F., and D.A. Hensher. The Trip Chaining Activity of Sydney Residents: A Cross-Section Assessment by Age Group with a Focus on Seniors. *Journal of Transport Geography,* Vol. 15, No. 4, 2007, pp. 298–312.
7. Factor, R., G. Yair, and D. Mahalel. Who by Accident? The Social Morphology of Car Accidents. *Risk Analysis,* Vol. 30, No. 9, 2010, pp. 1411–1423.
8. Das, S., and X. Sun. Exploring Clusters of Contributing Factors for Single-Vehicle Fatal Crashes Through Multiple Correspondence Analysis. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
9. Davis, G.A. Relating Severity of Pedestrian Injury to Impact Speed in Vehicle–Pedestrian Crashes: Simple Threshold Model. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1773,* TRB, National Research Council, Washington, D.C., 2001, pp. 108–113.
10. Sze, N., and S. Wong. Diagnostic Analysis of the Logistic Model for Pedestrian Injury Severity in Traffic Crashes. *Accident Analysis and Prevention,* Vol. 39, No. 6, 2007, pp. 1267–1278.
11. Moudon, A., L. Lin, J. Jiao, P. Hurvitz, and P. Reeves. 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,* Vol. 43, No. 1, 2011, pp. 11–24.
12. Ulfarsson, G., S. Kim, and K. Booth. Analyzing Fault in Pedestrian–Motor Vehicle Crashes in North Carolina. *Accident Analysis and Prevention,* Vol. 42, No. 6, 2010, pp. 1805–1813.
13. Sullivan, J., and M. Flannagan. Differences in Geometry of Pedestrian Crashes in Daylight and Darkness. *Journal of Safety Research,* Vol. 42, No. 1, 2011, pp. 33–37.
14. Roudsari, B., C. Mock, R. Kaufman, D. Grossman, B. Henary, and J. Crandall. Pedestrian Crashes: Higher Injury Severity and Mortality Rate for Light Truck Vehicles Compared with Passenger Vehicles. *Injury Prevention,* Vol. 10, No. 3, 2004, pp. 154–158.
15. Eluru, N., C. Bhat, and D. Hensher. A Mixed Generalized Ordered Response Model for Examining Pedestrian and Bicyclist Injury Severity Level in Traffic Crashes. *Accident Analysis and Prevention,* Vol. 40, No. 3, 2008, pp. 1033–1054.
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 March 10, 2015.
17. Husson, F., J. Josse, S. Le, and 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 March 10, 2015.
18. Greenacre, M., and J. Blasius. *Multiple Correspondence Analysis and Related Methods.* Chapman and Hall/CRC Press, Boca Raton, Fla., 2006.
19. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis.* Springer, New York, 2009.

The Standing Committee on Pedestrians peer-reviewed this paper.