IATSS Research xxx (2020) xxx

Contents lists available at ScienceDirect

IATSS Research

Research Article

Extremely serious crashes on urban roadway networks: Patterns and trends

Subasish Das [a,*], Anandi Dutta [b]

[a] Texas A&M Transportation Institute, 1111 RELLIS Parkway, Room 4414, Bryan, TX 77807, United States of America
[b] The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249-0667, United States of America

ARTICLE INFO

Article history:
Received 31 October 2019
Received in revised form 23 January 2020
Accepted 24 January 2020
Available online xxxx

Keywords:
Urban roadways
Extremely serious crashes
Taxicab correspondence analysis
Contributing factors
Dimension reduction

ABSTRACT

Extremely serious traffic crashes, defined as having a death toll of two and greater than two, have become a serious safety concern on urban roadways in Louisiana. This study examined the different contributing factors of these crashes to determine significant trends and patterns. We collected traffic crash data from Louisiana during the period of 2013 to 2017 and found that a total of 72 extremely serious crashes (around 2% of all traffic fatalities) occurred on Louisiana urban roadway networks. As crash data contain an enormous list of contributing factors, there was an issue of 'more features than data points' in solving the research problem. Most of these variables are categorial in nature. We selected a dimension reduction tool called Taxicab Correspondence Analysis (TCA) to investigate the complex interaction between multiple factors under a two-dimensional map. Findings of the study reveal several key clusters of attributes that show patterns of association between different crash attributes. The conclusions of this study are exploratory, and the results can help in better visualizing the association between key attributes of crashes. The findings have potentials in designing suitable countermeasures to reduce extremely serious crashes.

1. Introduction

Effectively incorporating roadway safety into transportation planning, design, and operation requires robust safety prediction models that can quantitatively predict the safety performance of roadways. Various safety prediction models have been developed, including the models introduced in the first edition of Highway Safety Manual by American Association of State Highway Transportation Officials (AASHTO). These models were intended to predict crashes at disaggregate levels and they require detailed input data and complex application procedures, which can function as effective decision-making tools in roadway designing and the operation stage. The past decade has witnessed a great advancement in modeling crash count and severity distribution for different types of roadways, mainly with statistical models and occasionally with machine learning algorithm. Newer statistical and machine learning methods used in crash modeling are data mining techniques, multivariate analysis, spatiotemporal modeling, and random parameters to empirical Bayesian and full-Bayesian hierarchical approaches. The key objective of most of the methods was to establish satisfactory statistical relationships between the number of crashes at roadway segments or intersections as accurately as possible.

While many studies have examined the key contributing factors that influence the perils of fatal and serious crashes, most of the studies have focused on common traffic crashes rather than crashes with extreme (at least two or more fatalities in a crash event) number of fatalities and injuries. Xu et al. conducted a study on extreme crashes in China using crash data for five years (2009–2013). They applied rules mining to identify the patterns of extreme crashes [1]. A similar U.S. study has not been conducted yet. This calls for an investigation on the patterns and associations of the contributing factors of extreme crashes to get a better understanding of the interdependency between these factors.

The identification of the key contributing factors for different types of crashes is a major task in the highway safety analysis. Taxicab Correspondence Analysis (TCA) is a dimensionality reduction method. It is used to describe the significance of variable categories from a high dimension dataset by showing the co-occurrence of groups of variables. TCA is referred to as the pattern recognition method that treats arbitrary data sets as a combination of points in a larger dimensional space. In comparison to parametric estimation, it uniquely simplifies complex data into knowledge extraction. In TCA analysis, the objective is to analyze the associations between multiple variables rather than the more traditional characterization of associations between a set of predictor

[*] Corresponding author.
E-mail addresses: s-das@tti.tamu.edu (S. Das), anandi.dutta@utsa.edu (A. Dutta).
Peer review under responsibility of International Association of Traffic and Safety Sciences.

https://doi.org/10.1016/j.iatssr.2020.01.003

variables and a single response variable of interest (i.e., number of crashes). In the past few years, correspondence analysis has been gaining popularity among researchers. When data sets are sparse, the degree of sparsity in a data set is defined as the percentage of zero abundances. Three kinds of potential outliers may be identified in sparse sets: rare observations, zero-block structure, and relatively high valued cells. As a sturdy-robust-resistant variation of correspondence analysis, TCA can handle any abundance of data in traffic safety engineering and generate satisfactory meaningful results in the presence of outliers. To determine the relationship between the variables and their significance, we used five years (2013–2017) of extreme fatal crash data on urban roadway network in Louisiana. The purpose of the present study is to evaluate the patterns and nature of extreme crash occurrences using TCA. The findings of this study could help authorities determine efficient and effective safety countermeasures to reduce these crashes.

2. Methodology

2.1. Taxicab correspondence analysis (TCA)

Correspondence Analysis (CA) was first introduced by Jean-Paul Benzécri [2]. If readers desire more detail, there are several books that contain further information about CA [2–6]. Choulakian recently proposed an improved version of CA in a series of papers [7–9]. In recent years, many researchers have used both CA and TCA to examine the complex nature of crash attributes [10–14]. The short description on the theorical concept of TCA is mostly based on the studies of Choulakian [7–9].

CA is based on Euclidean distance, whereas TCA is based on a distance that is known as Manhattan city block, or taxicab distance. Consider the coordinates of two points, $X = (x_1,x_2,\ldots.\,.,x_n)$ and $Y = (y_1,y_2,\ldots.\,.,y_n)$ and a vector $\boldsymbol{v} = (v_1,v_2,\ldots.\,.,v_n)$ to evaluate these distances. The distances can be written as:

$$\begin{aligned}&\mathit{Euclidean\ Distance} = ED(X,Y)\\&= \sqrt{\textstyle\sum_{i=1}^{n}(x_i - y_i)^2}\left[\text{with } L_2 \text{ Norm} = \|v\|_2 = \sqrt{\textstyle\sum_{i=1}^{n}(v_i)^2}\right]\end{aligned} \tag{1}$$

$$\begin{aligned}&\mathit{Taxicab\ Distance} = TD(X,Y)\\&= \textstyle\sum_{i=1}^{n}|x_i - y_i|\left[\text{with } L_1 \text{ Norm} = \|v\|_1 = \textstyle\sum_{i=1}^{n}|v_i|\right]\end{aligned} \tag{2}$$

The concepts of CA and TCA lie within the concept of singular value decomposition (SVD). We can delineate the real matrix A as $M\Lambda^{1/2}N'$, with Λ being the diagonal matrix of the real non-negative eigenvalues of AA', M , the orthogonal matrix of the corresponding eigenvectors, and N, the matrix of eigenvectors of $A'A$ (with constraints $M'M = I$ *and* $N'N = I$).We can associate the SVD with the reconstruction formula of a k-rank matrix:

$$a_{ij} = \sum_{i=1}^{k}\sqrt{\lambda_\alpha}\, m_{i\alpha} n_{i\alpha} \tag{3}$$

Choulakian [7] proposed to build the SVD solution through a recursive optimization technique. In TCA, another method of measuring distance (L_∞ metrics) is considered for analysis. We can determine the distance by by $max_{i\in(1,n)}|x_i - y_i|$ [with L_∞Norm$=\|v\|_\infty = max_{i\in(1,n)}|v_i|$]. We can find the complete solution by recursively applying the optimization problem on the residuals. This is known as Taxicab Singular Value Decomposition (TSVD). The association between the rows and columns of N can by summarized by statistics, χ^2, which measure the relationship's departure from the independence. As the independence is estimated by $N_0 = nT_0 = nrl'$, we can calculate the departure from independence by

$$\begin{aligned}\chi^2 = n\theta^2 &= n\sum_i\sum_j\left(\frac{(t_{ij} - t_{i.}t_{.j})^2}{t_{i.}t_{.j}}\right)\\&= n\sum_i\sum_j\frac{d_{ij}^2}{t_{i.}t_{.j}}\,[\text{with } (r-1)\times(l-1) \text{ degrees of freedom}]\end{aligned} \tag{4}$$

We can also term the TCA method as the Taxicab SVD of the data table, $D = T - rl'$; it considers the table's profiles, respectively $R = D_r^{-1}D$ for the rows and $L = D_l^{-1}D$ for the columns. The solution is recursive; thus, we can consider the residuals from the previous factors for each new step. We can define the reconstruction formula as following:

$$T = p_r p_c' + \sum_{\alpha=2}^{k}\frac{1}{\lambda_\alpha}B_\alpha C_\alpha' \tag{5}$$

After introducing elements, the formula becomes: $t_{ij} = t_{i.}t_{.j} + \sum_{\alpha=2}^{k}\frac{1}{\lambda_\alpha}B_{i\alpha}C_{j\alpha}$. By transformation, we get: $n_{ij} = nr_il_j(1+\sum_{\alpha=2}^{k}\frac{1}{\lambda_\alpha}b_{i\alpha}c_{j\alpha})$.

3. Data description

To achieve the research objectives, we acquired five years (2013–2017) of traffic crash data from Louisiana Department of Transportation and Development (LADOTD). This database contains crash information in several different relational databases. Three main databases (crash, vehicle, and roadway inventory) were used to develop the required dataset. The crash identification numbers are later merged with crash and roadway inventory information to prepare complete data. Altogether, 73 crashes on urban roadways involve two or more fatalities, which are defined as extremely serious crashes. In total, 150 fatalities were involved in these 73 crashes. These crashes also involved 75 severe to complaint injuries from a total of 268 occupants.

After removing non-pertinent information (for example, crash hour and daylight conditions show approximately similar information, therefore, instead of using both variables one variable can be used), a more precise database was prepared based on key contributing factors. The variable section method used the research findings from past studies. Table 1 lists the descriptive statistics of the seventeen selected variables. The weekend is over-represented compared to weekday extreme crashes. Daytime represents nearly 41% of all extreme crashes in the urban networks of Louisiana. The extreme crashes occurred more on multilane roadways. In collision types, head-on crashes have a higher percentage. The other three influential collision types are right angle, left turn, and rear-end. Two-way roadways, especially roadways with no physical separation, seem more prone towards extreme crashes on urban networks. Crashes on state highways represent around 52% percent of all extreme crashes. Approximately 60% of crashes occurred on roadways with posted speed limits between 40 and 55 mph. Multiple vehicle crashes are overrepresented compared to single vehicle crashes (87% versus 13%). Front damage of the vehicles contributed to 62% of all crashes. Extreme crashes occurred on roadways with pavement markings, which are key traffic control tools. Around 55% of the crashes happened in the dark with different lighting conditions. A large percentage of crashes occur in business areas and open country areas. Violations contribute to 75% of these crashes. Airbags are not deployed for 27% of these crashes. Around 51% of the driver injury type is fatality. Alcohol is involved in about 68% of these crashes. The mass value (proportion of the attributes among all attributes) of the top twenty key attributes is listed in Table 2. The findings are mostly in line with the findings from Table 1.

4. Modeling results and discussions

It is important to note that the identification of hidden trends from the crash dataset is central to make intuitive interventions in terms

Please cite this article as: S. Das and A. Dutta, Extremely serious crashes on urban roadway networks: Patterns and trends, IATSS Research, https://doi.org/10.1016/j.iatssr.2020.01.003

Table 1
Key variables.

Attribute	Percentage	Attribute	Percentage
DayW (Day of Week)		**TrafficControl (Traffic Control Condition)**	
FSS (Friday to Sunday)	57.33	White DL	36.67
MTWT (Monday to Thursday)	42.67	Yellow DL	19.33
Lighting (Lighting Condition)		Yellow NPL	18.67
Daylight	40.67	No Control	4.67
Dark No Light	28	Other	20.67
Dark Light	17.33	**Damage (Damage of Vehicle)**	
Dark Light at Int.	10	Front	62
Dusk/Dawn	4	Other	22
Lane (Number of Lanes)		Rear	16
Multi	55.33	**Contri (Key Contributing Factor)**	
Two	44.67	Violations	75.33
Colli (Collision Type)		Driver Cond.	14.67
Head-On	30	Prior Move.	8
Right Angle and Left Turn	24.67	Road Cond.	2
Rear End	19.33	**Alc (Alcohol Involvement)**	
Single Veh. (Vehicle)	16.67	No	32
Other	9.33	Yes	68
Loca (Locality Type)		**DrAge (Driver Age)**	
Business	44.67	17–30	32.67
Open Country	23.33	31–50	34.67
Residential	19.33	> 50	24.67
Other	12.67	Unk (Unknown)	8
Road (Roadway Type)		**DrAirbag (Driver Airbag Condition)**	
Two-Way No Sepa (Separation)	56.67	Deployed	52.67
Two-Way Sepa (Separation)	40.67	Not Deployed	26.67
One-Way	2.67	Unk (Unknown)	20.67
Hwy (Highway Type)		**DrInjury (Driver Injury Type)**	
Interstate	22.67	Fatal	50.67
State Hwy	52	Severe	4.67
U.S. Hwy	25.33	Moderate	10.67
PSL (Posted Speed Limit)		Complaint	10
< 40 mph	12	No Injury	24
40–55 mph	60	**PriorMove (Prior Movement before Crash)**	
> 55 mph	28	Go Straight	50
NumV (Number of Vehicles involved)		Crossed CL	16.67
Single	12.67	Ran Off Road	13.33
Multi	87.33	Properly Parked	5.33
		Other	14.67

Table 2
Top attributes with higher mass values.

Attribute	Mass	Attribute	Mass
NumV_Multi	0.05137	DrInj_Fatal	0.02980
Contri_Violations	0.04431	PriorMove_Go.Straight	0.02941
Alc_Yes	0.04000	Lane_Two	0.02627
VehDamage_Front	0.03647	Loca_Business	0.02627
PSL_40.55	0.03529	DayW_MTWT	0.02510
DayW_FSS	0.03373	Light_Daylight	0.02392
Road_Two.Way.No.Sepa	0.03333	Road_Two.Way.Sepa	0.02392
Lane_Multi	0.03255	TraffControl_White.DL	0.02157
DrAirbag_Deployed	0.03098	DrAge_31.50	0.02039
Hwy_State.Hwy	0.03059	DrAge_17.30	0.01922

of policy improvement and countermeasure selection. The all-in-one TCA plot (see Fig. 1), in the form of biplot or two-dimensional plot, provides a general overview of the location of the attributes. A broad category difference can be found by analyzing the presence of the variables with reference to either the x-axis or y-axis. The closer positions (co-ordinates of the attributes in the TCA plot) indicate the similarity of co-occurrences of these attributes. If an attribute is distant from other clusters or a group of points, it can be considered as a unique or rare data point in the data structure. Fig. 1 shows six major clusters based on the positions of the attribute co-ordinates. Fig. 1 shows that the positions of some of the attributes (for example, open country road, crossing centerline as the prior event, one-way roadway, moderate driver injury) are not within any clusters. The attributes indicate that crashes associated with these attributes are rare compared to other clustered groups. The analysis provided here is based on the first two axis values. The results show the first two planes and explain 70% of the total variance in the data. An explanation of these clusters is provided below:

4.1. Clusters based on attribute locations

4.1.1. Clusters in the upper right

4.1.1.1. Cluster 4. The attributes in this cluster are head-on collision, dark with no lighting, fatal driver injury, alcohol involvement, age group 17–30 or above 50, single vehicle crash, weekend, and violation as the key contributing factors. These attributes describe impaired single vehicle crashes that occurred mostly at night time. The crash shows two age related exposure groups: young drivers and older drivers. Most of these single vehicle crashes occur with driver fatalities. Safety education and effective enforcements are required to avoid alcohol-involved extreme crashes on urban networks.

4.1.2. Clusters in the upper left

4.1.2.1. Cluster 1. The attributes in this cluster are interstate roadway, posted speed limit above 55 mph, roadways with a white dashed line, and two-way roadways with separation. This cluster indicates urban extreme crashes on high functional class (i.e., interstate or principal arterial) roadways.

4.1.2.2. Cluster 2. The attributes of this cluster are multi-vehicle crash, drivers with complaint injury, dark with lighting condition, and drivers aged from 31 to 50 years old. These crashes are not single vehicle crashes (they involved one or more other vehicles), and they occurred at dark with lighting. This cluster mostly highlights the association of multiple-vehicle crashes with nighttime crashes and an age exposure group (31–50 years old).

4.1.3. Clusters in the lower left

4.1.3.1. Cluster 3. The attributes of this cluster are daylight or dark with intersection lighting, weekday, business locality, no injury of the drivers, non-alcohol involvement, airbags are not deployed, U.S. highways, right angle/left turn and rear-end collision. This cluster indicates that non-impaired extreme crashes are associated with non-fatal driver injuries. Most of these crashes are intersection related.

4.1.3.2. Cluster 6. The attributes of this cluster are posted speed limit lower than 40 mph, and prior movement is properly parked. The other two attributes are unknown driver age and an absence of traffic control devices. A closer look at the crash data of these crashes shows that one of the crashes involves pedestrians. Appropriate policies are required to provide guidance on on-street parking so that these crashes can be avoided.

4.1.4. Clusters in the lower right

4.1.4.1. Cluster 5. The attributes of this cluster are posted speed limit 40 to 55 mph, residential roadways, state highway, two lane roadways with no physical separation, single vehicle crash, and run-off-road (ROR) crash. This cluster indicates ROR crashes on two lane urban networks with no physical separation. Providing physical separation can be considered as a potential countermeasure for reducing these crashes.

Please cite this article as: S. Das and A. Dutta, Extremely serious crashes on urban roadway networks: Patterns and trends, IATSS Research, https://doi.org/10.1016/j.iatssr.2020.01.003

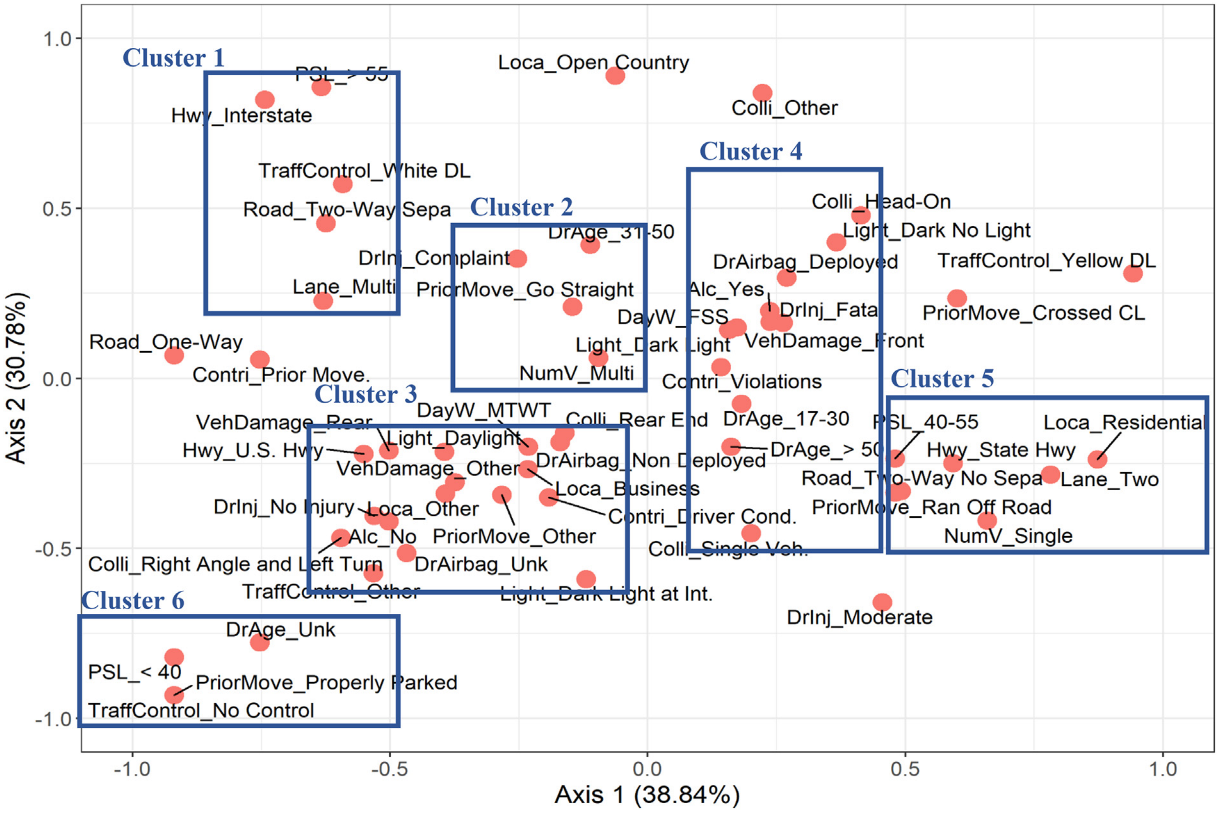

Fig. 1. TCA plot for the variable categories.

4.1.5. Clusters based on row (crash) locations

Fig. 2 shows the TCA plot for each individual crash event. Crashes with similar properties are closer in positions. Four example cases (shown in blue circles) are described below:

- Cr34 [‘CrXX’ indicates crash id] (two vehicle crash with two fatalities and four injuries) and Cr47 (two vehicle crash with two fatalities and one injury) have the same co-ordinates.
- Cr61 (two vehicle crash with two fatalities and two injuries), Cr11 (three vehicle crash with two fatalities and three injuries), and Cr53 (four vehicle crash with two fatalities and four injuries) have the same co-ordinates.
- Cr45 (two vehicle crash with two fatalities and two injuries), Cr35 (one vehicle crash with two fatalities), and Cr42 (one vehicle crash with five fatalities and one injury) have the same co-ordinates.

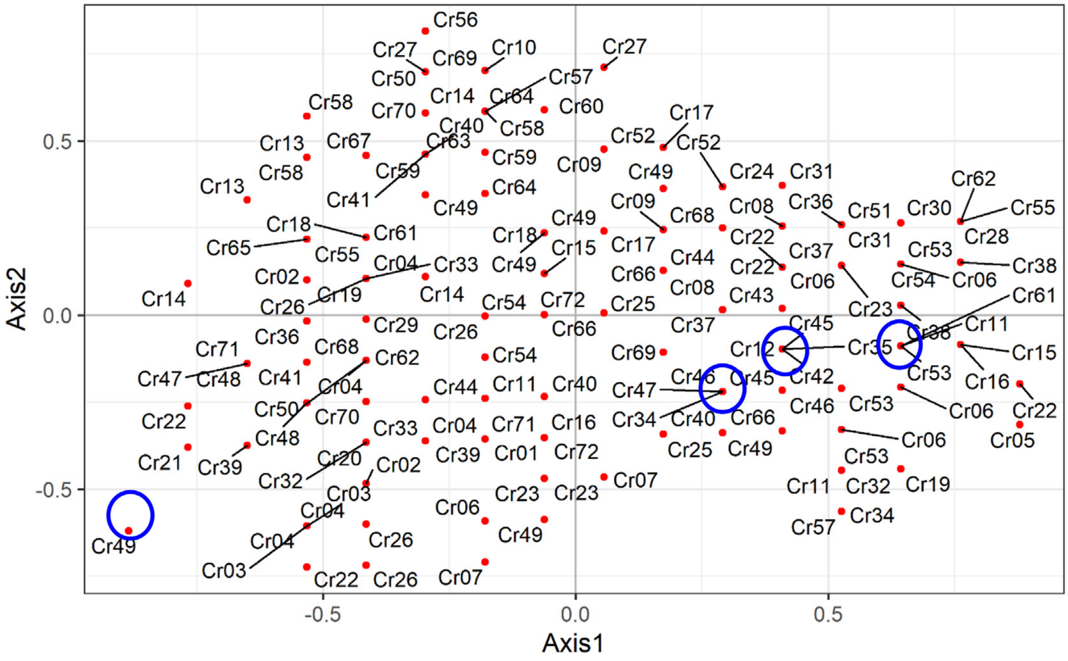

Fig. 2. TCA plot for the crash events.

Please cite this article as: S. Das and A. Dutta, Extremely serious crashes on urban roadway networks: Patterns and trends, IATSS Research, https://doi.org/10.1016/j.iatssr.2020.01.003

- Cr49 is on the lower left side of the plot. This crash involves seven roadway users with three fatalities and two injuries. As this crash is unique from the other extreme crashes, the TCA plot can show its uniqueness by positioning it distant from other extreme crash events. The crash report of this crash is described as follows: *"Vehicle #1 was initially eastbound on Airline Highway approaching the intersection with S. Carrollton Avenue. As Vehicle #1 crested the overpass, for reasons unknown, Vehicle #1 crossed the center line into the opposing lanes of travel. Vehicle #1 then continued drifting to the left until it side-swiped the left concrete guard rail. Vehicle #1 continued uncorrected, along the guard rail until the guard rail ended at which time Vehicle #1 careened to the left into a group of three pedestrians and one bicyclist, then through a fence into a private parking lot crashing into five parked vehicles before coming to a stop. One of the pedestrian victims and the bicyclist were pronounced dead on the scene. Two other pedestrians were transported to University Hospital where one was pronounced dead and the other listed in critical condition. The driver of vehicle #1 was also transported to University Hospital by EMS unit 3221."*

5. Conclusions

Many research efforts on conventional crash data have been conducted to understand the key contributing factors that influence the frequency and severity of urban traffic crashes. This knowledge is useful in determining suitable countermeasures. Extreme crashes in urban networks were not investigated in depth in the earlier studies. Moreover, the conventional way of associating the effect of a single factor on the response variable is not enough to characterize the complex nature of a crash occurrence.

We demonstrated that TCA would be a viable tool in analyzing complex categorical data in search of meaningful associations between categorical factors. Conventional statistical modeling requires supervised data (clear definition of explanatory and response variables) and prior assumptions. In TCA, the objective is to show the co-occurrence of the categories in a low dimensional space (in this study, two-dimensional) where proximity in the space indicates the similarity of the attributes. This method helps in understanding diverse variable categories and produces visual results from the key associations. As the extreme crash data on urban networks has a limited number of cases, removing entries with noise would make a small dataset smaller. Applying TCA offers the advantage of removing noise (by representing the data in low dimensional spaces) without reducing the dataset. This feature helps to describe the significant associations between the categories of a complex dataset like extreme crashes on urban networks.

We determined several key patterns from extreme crashes. Some of the patterns are: alcohol impaired crashes with higher driver fatalities for two age groups (younger and older drivers), multi-vehicle crashes at intersection associated with an age group (31–50 years), parking related pedestrian crashes, non-deployed airbag related right angle or left turn crashes, and ROR crashes on urban two lane with no physical separation. Prioritization of certain key association groups, as well as suitable countermeasures listed in this study, can help the policymakers that are developing different policy initiatives to reduce the fatalities and injuries due to extreme crashes.

Our study is not without limitations. First, cluster analysis of individual crash has been introduced but not explored in depth. Second, this study provides limited information on the potential countermeasures that will be more effective in reducing extreme traffic crashes, because little research has been conducted yet. Third, the groups of confluence factors are developed based on the first plane (with two dimensions), which represents a 70% variance of the complete database. The current limitations can offer various directions for future research in this domain.

Disclaimer

The contents of this paper reflect the views of the authors and not the official views or policies of the LADOTD.

Acknowledgement

We like to thank Dr. Xiaoduan Sun for the crash data. We also like to thank two anonymous reviewers for their thoughtful feedback and suggestions.

References

[1] C. Xu, J. Bao, C. Wang, P. Liu, Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China, J. Saf. Res. 67 (2018) 65–75.
[2] J.P. Benzećri, Correspondence Analysis Handbook, CRC Press, 1992.
[3] B. Le Roux, H. Rouanet, Multiple Correspondence Analysis, 163, Sage, 2010.
[4] F. Murtagh, Correspondence Analysis and Data Coding with Java and R. Chapman and Hall/CRC, 2005.
[5] M. Greenacre, J. Blasius, Multiple correspondence analysis and related methods, Chapman and Hall/CRC, 2006.
[6] J. Hjellbrekke, Multiple Correspondence Analysis for the Social Sciences, Routledge, 2018.
[7] V. Choulakian, Graph partitioning by correspondence analysis and taxicab correspondence analysis, J. Classif. 30 (2013) 397–427.
[8] V. Choulakian, L1-norm projection pursuit principal component analysis, Comput. Stat. Data Anal. 50 (2006) 1441–1451.
[9] V. Choulakian, Taxicab correspondence analysis, Psychometrika 71 (2006) 1–13.
[10] S. Das, X. Sun, Association knowledge for fatal run-off-road crashes by multiple correspondence analysis, IATSS Res. 39 (2) (2016) 146–155.
[11] S. Adele, S. Tréfond-Alexandre, C. Dionisio, P. Hoyau, Exploring the behavior of suburban train users in the event of disruptions, Transp. Res. F 65 (2019) 344–362.
[12] R. Baireddy, H. Zhou, M. Jalayer, Multiple correspondence analysis of pedestrian crashes in rural illinois. Transportation research record, Journal of the Transportation Research Board 2019.
[13] S. Das, K. Jha, K. Fitzpatrick, M. Brewer, T.H. Shimu, Pattern identification from older bicyclist fatal crashes, Transp. Res. Rec. 30 (2019).
[14] M. Jalayer, M. Pour-Rouholamin, H. Zhou, Wrong-way driving crashes: a multiple correspondence approach to identify contributing factors, Traffic Injury Prev. 19 (1) (2018) 35–41.

Please cite this article as: S. Das and A. Dutta, Extremely serious crashes on urban roadway networks: Patterns and trends, IATSS Research, https://doi.org/10.1016/j.iatssr.2020.01.003