Research Article

Identifying the principal factors influencing traffic safety on interstate highways

Aschalew Kassu[1] · Mahbub Hasan[1]

Received: 3 October 2019 / Accepted: 26 November 2019 / Published online: 30 November 2019

Abstract
This study aims at identifying the principal factors influencing fatal, nonfatal injury and non-injury traffic crashes on urban and rural interstate highway segments using a statistical approach called principal component analysis. Initially, fourteen explanatory variables including segment length, annual average daily traffic (AADT), weekday/weekend, hour of the day, urban–rural designation of the segment, median type, pavement surface condition, roadway geometric characteristics, weather, number of lanes, and drivers' age and gender, and the accident year were considered. Separate principal component analyses are performed for the three crash categories and the aggregate dataset. The results of the analyses show that, regardless of the crash categories used, seven principal components accounting for over 70% of the variances in the original datasets were retained. In addition to the overall dimensional reduction of the original dataset by 50%, the results suggest that the key variables contributing to the crashes across the categories of the accident types observed on the freeway segments are insignificant. The retained principal component loadings of the factors PC1, PC2, PC3, and PC5 revealed the fact that the number of lanes, the median type, and the AADT of the segments are highly correlated and represented by the first factor (PC1). Similarly, other interrelated factors such as the prevailing weather and the pavement surface condition (wet, dry, snow), the hour of the day and the lighting condition, the drivers' age and gender are well represented by PC2, PC3, and PC5 respectively.

Keywords Freeways · Principal component analysis · Interstate highway safety · Traffic accidents · Traffic safety · Multivariate analysis

1 Introduction

In 2016, about 1.35 million people (18.2 per 100,000 population), which is on an average of about 3700 deaths a day, lost their life due to roadway traffic crashes [1], and 20–50 million people were injured or disabled [2]. The reports pointed out that, for children and younger adults between 5 and 29 years of age, road traffic crash is the number one leading cause of death globally. For all age groups combined, the road traffic crash is ranked as the eighth leading cause of death. In the United States, the recent report [3] revealed that the overall traffic crashes in the year 2015 is ranked as the 13th cause of death. Traffic accidents are the results of several factors including the roadway geometric design elements, the type and condition of the pavement surface conditions, the prevailing environmental factors, roadway lighting, the urban–rural designation of the segment, the vehicle conditions, traffic flow characteristics and the compositions (percentage of trucks), the vehicle condition, and the human factors. The characteristics of traffic accidents and the factors influencing the crash rates also vary based on the roadway facility types [4–6]. Most of these factors consist of several independent variables and the interaction between two or more of these variables

✉ Aschalew Kassu, aschalew.kassu@aamu.edu | [1]Department of Mechanical, Civil Engineering and Construction Management, Alabama A&M University, Normal, AL 35762, USA.

may exacerbate the negative impacts. The study of the human factors aspect of the crash analysis by itself encompasses a wide range of variables. To mention a few, the human factor includes the drivers' age, gender, alcohol and drug use, fatigue, level of experience, distraction, seatbelt use, violation of traffic regulations, and so on. For simplicity and ease of interpretations of the results of the multicategorical dimension of the human factor, Shirmohammadi et al. [7] used statistical clustering analysis to identify the drivers' behavior and skills. Here, it is wise to suggest that it is impractical to obtain all the datasets across each potential predictor and incorporate in road traffic safety studies. As a result, most of the existing literature focuses on analyzing or modeling traffic accidents using a limited number of predictive variables. Understanding the effects of crash contributing factors individually, and the impacts of the interaction between various factors on traffic accidents help to develop a model with a reasonable crash predictive power for planning as well as the implementation of real-time traffic management procedures in line with the prevailing traffic conditions.

For a better understanding of the level of significance, the variability of the variables under consideration, and dimensional reduction of the data, extracting the critical and manageable factors is essential. This can easily be done by using a multivariate statistical approach, commonly known as principal component analysis (PCA). PCA adopts rigorous statistical procedures to linearly combine the larger dimension original variables into smaller and manageable dimensional factors called principal factors (PCs) [8–10]. The main objectives of using the method of principal component analysis in this study are to demonstrate the efficacy of PCA in traffic safety study, primarily in reducing the fourteen potential predictive variables into a smaller number of representative variables without losing much of the original characteristics of the variables. This is done by deriving a correlation among the large independent variables defining and creating a set of new components representing the common characteristics of the larger predictive variables with an ultimate goal of reducing the numbers of predictors without causing significant variation with the original dataset.

2 Literature review

A considerable amount of literature is available on methods to develop crash prediction models. The focus of the majority of the existing traffic safety research relies on modeling the effect of a predictive variable to predict crash rates on a specified roadway facility. Several alternative statistical modeling techniques have been proposed, including the classical Poisson, negative binomial, logistic regression, and others. However, the literature on the application of PCA in traffic safety research is limited. Golob et al. used PCA extensively [11, 12] to identify the major traffic flow variables affecting freeways accidents and study the relationships among the predictors in Southern California. In one of their work [11], twelve traffic flow variables, including traffic speed, the median volume/occupancy values in the left, interior and right lanes. To remove about half of the redundant variables, principal component analysis was utilized. They reported that the six significant components represented nearly 88% of the variances in the original information, which appears to be good. The dependent variable in the study was the type of crash likely to occur. Once the insignificant traffic flow variables are discarded, the impacts of the retained variables on three different freeway lanes (left, interior and right) as a function of four different conditions (daylight dry pavement, daylight wet pavement, night time dry pavement, and night time wet pavement) was studied. The analysis concluded that during daylight on dry and wet pavements, the most likely accidents to occur are rear-end, and multiple collisions due to lane changing. In subsequent studies, they used a similar data set (twelve independent traffic flow variables) and PCA technique to identify the key variables and develop a systematic approach to assess the effects of changes in freeway traffic flow (volume and speed) on the likely occurrence of freeway crashes in southern California.

In another work [12], the authors studied the relationship between 36 independent variables and freeway collisions. Some of the variables considered were, traffic conditions (speed, volume, occupancy), the type of freeway collisions [rear-end, sideswipe, hit object collision, number of vehicles involved, location of the lane (left, interior, right)], accident severity (injury, fatality or PDO) on two interstates (I-5 and I-405) and four State Routes in California. The data used were 6 months police-reported accident data (March to August 2001) and traffic flow data from vehicle detection system (loop detectors). The authors used PCA approach to reduce a large number of variables into only eight factors, which accounted for about 79% of the variances within the 36 variables. The eight key traffic flow variables are used for further analysis for modeling probabilities of crash characteristics, including crash severity, the type, and location of the collision. To remove the effect of multicollinearity and retain only independent variables in the models proposed to predict injury severity in the head-on crash [13] and broadside and angle collisions [14], Mercier et al. used PCA. They were able to extract the four most essential variables out of the original 14 variables associated with age, gender, use of lap and shoulder restraint, the effect of an airbag, and others. The purpose of both works was to test the hypothesis that in

the incidences where either head-on crash or broadside-and-angle collisions occur, older drivers and passengers are more likely to suffer severe injury than younger drivers and passengers.

Caliendo and Parisi [15] used PCA to identify the significant variables causing accidents in curves and tangents. The study was conducted on 51.6 km long four-lane median divided rural highway in Italy using 6 years of data (1998–2003). The dependent variable was the number of accidents per year and the independent variables considered for tangent sections were ADT, segment length, longitudinal slope, pavement friction, and design speed. For curve sections, two additional variables were included: sight distance and radius of curvature. As the numbers of the original independent variables considered in the study were too small, the application of PCA in this study did not suggest a significant reduction in the number of independent variables. In their subsequent work, a more detailed and systematic analysis of the data and modeling results were conducted using similar data sets with additional independent variables, including pavement surface condition (wet or dry), presence of junction in the tangent sections, on 46.6 km multilane road section [16]. The study section is divided into 147 sub-segments ranging in length from 0.069 to 1.695 km. The results of PCA technique are used as independent variables to develop crash prediction models using Poisson, negative binomial and negative multinomial regression models for curves and tangent sections of multilane roads in Italy. Based on the study, the three significant variables affecting crash frequency on curve sections are AADT, segment length, and radius of curvature. Whereas, the effects of sight distance, longitudinal slope, and pavement friction are statistically insignificant at a 95% confidence interval. However, for tangent sections, the presence of junction, and an increase in AADT and segment length increased in crash frequency. One of the most significant variables increasing the crash frequency to both curve and tangent sections by a factor of 2.7 and 2.3, respectively, is rain. In terms of model adequacy on the data used, the negative multinomial regression model was suggested to be preferable to both Poisson and negative binomial models. Wu et al. [17] used PCA to identify and compare the significant factors contributing to crash in four facility types on about 69 km (~ 43 mi) of four-lane congested and uncongested urban expressway in Japan, consisting of sharp curve, diverge segment, merge segment, and basic expressway segment using 3 years of data (2007–2009). Based on the report, the level of significance of the variables varies with the facility type. For uncongested flow conditions, the crash rate is higher at the merged segment than a sharp curve, diverge segment and basic segment. However, due to the restricted visibility, the sharp curve has the highest crash rate during both under congested and uncongested flow conditions.

Molla [18] analyzed several driver-related variables causing fatal crashes. He applied PCA to identify the major factors contributing to fatal crash in the United States [18]. Based on the study, nearly 100 factors were identified as driver-related variables, including sleepy, failure to take medication, speeding, driving the wrong way, flat tire, inexperienced driver, under the influence, and many others. Although some of the factors listed appears to be redundant, for instance, 'driving too fast' and 'driving in excess of speed limit', out of the large list of potential variables, the results of the PCA outputs indicated that the first principal component accounted for about 40% and the first 13 components explain 80% of the total variances. Based on the study, speeding, backing improperly and drivers' failure to observe warning or instructions are the top three causes of drivers' related traffic crash in the United States. Different from the typical results widely demonstrated in traffic safety studies, de Andrade et al. [19] applied PCA to identify the road traffic hotspot locations in Brazil; and Nagendra and Khare [20] used it to study urban intersection related traffic, emission and meteorological factors including cloud cover, humidity, pressure, rainfall, sunshine hours, temperature, visibility, wind speed and direction. More recently, Youming et al. [21] used PCA to analyze fatal traffic crashes associated with the vehicle conditions such as tire wear, failure of the vehicles' exhaust system and rim damage.

Apart from the commonly used traffic crash studies such as Poisson, negative binomial regression and logistic regression models [22–25], the application of tree-based data mining techniques like classification and regression tree (CART) [26–29], and a hierarchical tree-based regression (HTBR) [30], are also found to be suitable approaches in classification and prediction of traffic crashes. Tree-based models have been used in many disciplines but not popular in traffic safety studies. These studies [26–30] reported that the graphical display options that these models possess provide more insight about the relationship between crash variables and simplify interpretation of the results than the odds ratios used in logistic regression techniques. Recently Ye et al. [31] used a semi-non parametric (SNP) Poisson model to study crashes observed on rural highways in the State of California for 10 years ranging from 1993 to 2002. The authors suggested that the ability of the SNP Poisson model to take the unobserved heterogeneity of the crash data into account, makes it a powerful model outperforming the traditional Poisson and negative binomial models.

3 Methodology and data collection

This study used urban and rural freeway segment crash data observed in the state of Ohio within the 5 years ranging from 2010 to 2014 supplied by The U.S. Department of Transportation, Federal Highway Administration (FHWA). The data were collected on 1500 miles of Interstate highways across the state [32]. The detailed description of the original data and the procedures followed to link and merged the separate sub-files were provided elsewhere [33]. Tables 1 and 2 show the list and summary descriptive statistics of the continuous and categorical variables (N = 101,789) used in the principal component analysis of fatal, nonfatal injury and non-injury traffic crashes on four and six-lane interstate urban and rural highways. Here, we acknowledge that that speeding is an important factor influencing traffic safety; however, since the scope of the present study is limited to freeways, and the posted speed limit is predominantly 70 mph (about 113 km/hr), this variable is not included in the analysis. Several variables can be contributing factors to highway crashes. However, all the factors are not equally important, and some of them are pertinent but essentially redundant. To thoroughly analyze the data and arrive at a reasonable conclusion, there is a need to remove the non-essential and redundant variables. PCA is one such statistical approach, which is not widely used in highway safety research but can be used to identify the most significant variables to be used in developing crash predictive models, traffic management, and implementation of informed highway safety countermeasures. The model developed considering only the significant predictors identified, which account for most of the variabilities in the original data, provides a meaningful and easier interpretation of the results.

There are several potential contributing factors for the likelihood of occurrences of traffic crashes. Using all possible variables in developing traffic crash predictive models undermines the accuracy of the parameter estimates to a certain extent. Utilizing PCA overcomes the high-dimensionality of the interrelated crash variables while maintaining the inherent variability in the original dataset [20]. It is also a powerful technique to study the pattern-similarity of a wide range of predictive variables and eliminate any potential multi-collinearity of the crash variables [8–10]. Applying the PCA procedure also excludes redundant variables and extracts the key contributing factors simplifying the identification of the principal variables for an in-depth analysis of the crashes.

It can be assumed that there is no such single and comprehensive model developed to account for all potential crash predictors [16]. Some of the potential reasons could be the level of significance of the independent variables for the likelihood of crash occurrence, the difficulty associated with handling the huge dataset, the availability of the data itself, and other resources. Some of the available variables could be redundant and thus need to be removed from the model to simplify the predictive model and interpretation of the results. One such statistical technique to exclude the least significant and redundant variables from further analysis and identify the most significant crash predictors, although not widely used in traffic safety research, is Principal Component Analysis (PCA). As shown in the equation below, PCA combines variables linearly into a smaller number of components while maximizing the variances at the same time. This procedure reduces a large number of variables into a smaller number of significant variables. A component is described by a linear model [34]: $y_i = b_1x_{1i} + b_2x_{2i} + \cdots + b_nx_{ni}$, where y_i = Components, b_i = factor Loadings (coefficient of the variable), indicating the relative weight of the variable to the factor, and x_i = predictive variables. The calculated factor loadings are used to relate the variables with the components.

The suitability of PCA technique for the current study and the significances of the sampling adequacy are tested using the Kaiser–Meyer–Olkin (KMO) statistics [35] and Bartlett test of sphericity. In principle, to apply PCA on a given data, the number of observations should be at minimum ten times the number of the variables to be analyzed. To this end, as can be seen in Tables 1 and 2, the number of observations, N, across each crash category is much more than the number of variables considered in the analyses. To diffuse the factors, adjust the principal components' axes, and simplify the interpretation of the results, the commonly used factor rotation method called Varimax orthogonal rotation has been used [10]. To decide the number of components to be retained, the following guidelines were adopted [8, 10]: (a) The 70% Rule: In this case, the standard recommendation is to retain enough components so that the retained variables capture 70% of the total variances. (b) Eigenvalue rule: Eigenvalues are variances of the principal components. In PCA, a variable is considered significant if the eigenvalue of the variable is greater than one. Hence, this approach suggests retaining the components whose eigenvalues are ≥ 1. However, in our study, using the eigenvalue rule restricts the percentage of variances accounted for to less

Table 1 Descriptive statistics of the continuous variables (N = 101,789) used in the principal component analysis of the aggregate traffic crashes on four and six-lane interstate urban and rural highways

Variable	Minimum	Maximum	Mean	SD
Segment length	0.01	7.65	1.3820	1.33170
AADT	4770	1720,00	57,660.99	30,307.560

Table 2 Descriptive statistics of categorical variables used in the principal component analysis of the aggregate traffic crashes on four and six-lane interstate urban and rural highways

Variables	Category	Frequency	Percent (%)
Accident year	4 = 2014	20,956	20.6
	3 = 2013	20,166	19.8
	2 = 2012	20,258	19.9
	1 = 2011	21,370	21.0
	0 = 2010	19,039	18.7
	Total	101,789	100.0
Weekday	1 = Weekend	22,788	22.4
	0 = Weekday	79,001	77.6
	Total	101,789	100.0
Hour of the day	3 = Night (12:00 am–5:59 am)	9032	8.9
	2 = Evening (6:00 pm–11:59 pm	22,439	22.0
	1 = Afternoon (Noon–5:59 pm)	40,691	40.0
	0 = Morning (6:00 am–11:59 am)	29,627	29.1
	Total	101,789	100.0
Lighting condition	1 = Dark	31,508	31.0
	0 = Light	70,281	69.0
	Total	101,789	100.0
Urban/rural designation	1 = Rural	28,335	27.8
	0 = Urban	73,454	72.2
	Total	101,789	100.0
Median type	1 = Unprotected	51,122	50.2
	0 = Protected	50,667	49.8
	Total	101,789	100.0
Pavement surface condition	2 = Snow	9132	9.0
	1 = Wet	23,851	23.4
	0 = Dry	68,806	67.6
	Total	101,789	100.0
Weather condition	2 = Snowing	10,565	10.4
	1 = Raining	13,926	13.7
	0 = Normal	77,298	75.9
	Total	101,789	100.0
Roadway characteristics	3 = Curve-grade	8549	8.4
	2 = Curve-level	5423	5.3
	1 = Straight-grade	18,249	17.9
	0 = Straight-level	69,568	68.3
	Total	101,789	100.0
Drivers' age	2 = Older (65 and older)	6488	6.4
	1 = Mid-age (25–64)	72,614	71.3
	0 = Younger (16–24)	22,687	22.3
	Total	101,789	100.0
Drivers' gender	1 = Male	65,046	63.9
	0 = female	36,743	36.1
	Total	101,789	100.0
Number of lanes	1 = Six	52,219	51.3
	0 = Four	49,570	48.7
	Total	101,789	100.0

than 70%, which violated the 70% rule. Understanding the tradeoff between the number of principal components retained and the selection rules, to capture the minimum percentage recommended above, we used an eigenvalue ≥ 0.96, which also increased the number of factors retained by one.

4 Results and discussions

Principal Component Analysis is used to condense the original information into a smaller number of factors [10]. Accordingly, the primary focus of this study is utilizing PCA technique to identify the principal components and the critical variables influencing traffic safety on Interstate highways in the state of Ohio. Initially, the suitability of PCA technique was tested using KMO statistics and Bartlett's test of sphericity. The value of KMO ranges between 0 and 1, with a cutoff value of 0.5 and a higher value implying reasonable correlation among variables, which warrants the suitability of the PCA approach for data analysis [10]. As shown in Table 3, the Chi square statistics of Bartlett's test of sphericity for the fatal, nonfatal injury, non-injury, and the aggregate crash datasets were found to be 0.577, 0.580, 0.578 and 0.578 respectively, with a degree of freedom (df) 91. In all the four different datasets, the KMO statistics higher than the cutoff criterion indicates that our crash data are sufficient to proceed with the principal component analysis. With a 95% level of significance ($\alpha = 0.05$), the significance level (p value = 0.001) of Bartlett's test of sphericity also indicates that the statistical principal component analysis method is a valid approach. The 3-dimensional scatter plots of the component loadings (not shown) are used to understand the potential clustering of the variables across each category of the crash.

Table 4 shows the results of the principal component analysis performed on fourteen variables representing the traffic characteristics, environmental, human factors, and highway design elements for the aggregate crash data combining the three categories of crashes described above. Seven principal components accounting for 72.19% of the variances in the initial fourteen variables were extracted using the 70 percentile rule [10], which corresponds to an eigenvalue greater than or equal to 0.96. The first principal component (PC1) highly correlates the number of lanes, median type, and AADT, and marginally with the urban–rural designation of the roadway segment, and explains the highest variability followed by PC2, which explains the variability not accounted by the first PC. The

Table 3 Kaiser–Meyer–Olkin (KMO) and Bartlett's test of sphericity

	Crash categories			
	Fatal	Nonfatal	Non-injury	Aggregate
KMO	0.577	0.580	0.578	0.578
Bartlett's test of sphericity				
Approx. Chi square	8418.352	57,813.022	232,723.13	298,784.8
df	91	91	91	91
Sig.	0.000	0.000	0.001	0.000

Table 4 Rotated principal component loadings for the aggregate fatal, nonfatal injury and non-injury crash on urban and rural freeways

Variables	Principal components						
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Number of lanes	**0.862**	0.007	0.007	−0.056	0.002	−0.023	0.003
Median type	**−0.745**	−0.031	0.016	0.007	−0.019	−0.033	0.002
AADT	**0.735**	−0.023	0.017	−0.346	−0.067	−0.031	−0.006
Weather condition	0.015	**0.955**	0.003	0.014	−0.023	0.022	0.013
Pavement surface condition	0.011	**0.954**	−0.025	−0.006	−0.030	0.043	0.010
Hour of the day	0.014	−0.022	**0.902**	−0.025	−0.011	0.034	−0.011
Lighting condition	−0.017	0.003	**0.898**	−0.011	−0.025	0.007	−0.004
Segment length	−0.032	−0.005	−0.015	**0.850**	−0.017	−0.025	−0.002
Urban/rural designation	−0.265	0.014	−0.018	**0.790**	0.063	−0.026	−0.012
Drivers' age	−0.036	−0.052	0.025	−0.019	**0.739**	−0.096	0.032
Drivers' gender	0.013	0.010	−0.054	0.057	**0.739**	0.071	−0.034
Weekday	0.077	−0.071	0.203	0.114	0.045	**0.751**	0.100
Roadway characteristics	−0.086	0.139	−0.157	−0.173	−0.074	**0.650**	−0.099
Accident year	−0.006	0.024	−0.016	−0.013	−0.004	0.003	**0.989**
Percentage of total variance	16.38	13.55	12.01	8.58	7.59	7.15	6.94
Cumulative percentage	16.38	29.93	41.94	50.52	58.10	65.25	72.19

Bold values indicate the variables captured by the corresponding principal component

PC4 shows a higher correlation between segment length and the urban–rural designation of the roadway segments. The AADT of the segment also shows a moderate association with this factor. As a whole, these two principal components mainly cover traffic characteristics and the roadway design elements of the highway segments. The PC2, PC3, and PC5 reveal an obvious combination of two variables into a single factor. Here, PC2 largely represents the prevailing weather condition, which represents the categories of the weather condition (snowing, raining, and normal) as well as the pavement surface condition (snow, wet, and dry) at the time of the crashes.

Similarly, the factor represented by PC3 relates the hours of the day (morning, afternoon, evening, nighttime), and the lighting condition (dark, light) during the crash. The factor PC5 shows a higher correlation and takes into account the human factor aspect of the variables (drivers' age and gender). Interestingly, this factor does not show a considerable representation of the remaining twelve variables. Likewise, for PC2 and PC3, beyond the two categories of the variables represented by each factor, the effects of the other variables are not accounted under these components. The factor, PC6, mainly accounts for the effects of two variables categorized as weekday/weekend and the roadway characteristics (curve-grade, curve-level, straight-grade, and straight-level). The last factor, PC7, accounts for the accident year (2010, 2011, 2012, 2013, and 2014).

Tables 5, 6 and 7 reveal the results of the principal component analyses performed on three separate crash datasets representing the fatal, nonfatal injury and non-injury crash categories respectively. Similar to the aggregate crash dataset, seven principal components accounting for 72.11%, 71.85%, and 72.32% of the variances in the original 14 variables across the fatal, nonfatal injury and non-injury crash categories of the crashes respectively, are extracted. Interestingly, the first rotated loading factor (PC1) of all the crash types (fatal, non-fatal injury, and PDO) presented in Table 5, 6 and 7 showed a strong correlation with the number of lanes, AADT, and the median type of the roadway segment. Factor 3, which accounts for about 12% of the total variances in the original 14 variables, represents the two highly correlated environmental (visibility) variables relating the hour of the day and the lighting condition during the crash. Another principal component, PC5, captured the impact of human factors (drivers' age and gender) on the likelihood of occurrences of a traffic crash. As can be seen in the tables, the factors PC1, PC2, PC3, PC4 and PC5, all show similar correlation pattern across the three different crash categories. For fatal crashes (Table 5), the factor PC6 primarily represents the single weekday/weekend variable and PC7 correlated accident year and the roadway characteristics of the segment under consideration. For fatal crashes (Table 5), the rotated factor PC1 comprised of the roadway elements (number of lanes and the type of median) and the traffic volume (AADT) of the segment, which accounts for about 16.5% of the total variances.

On the other hand, the urban–rural designation of the segment, which shows a relatively lower correlation with PC1, is captured by the rotated factor PC4 with a higher correlation with the segment length and a lower correlation with a roadway characteristic (straight-level,

Table 5 Rotated principal component loadings for fatal crashes on urban and rural freeways

Variables	Principal components						
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Number of lanes	**0.847**	−0.002	0.019	−0.051	0.002	0.056	−0.046
AADT	**0.765**	−0.035	0.008	−0.277	−0.097	0.048	−0.036
Median type	**−0.699**	−0.011	−0.010	0.033	−0.085	0.088	−0.104
Weather condition	−0.012	**0.959**	−0.014	0.020	−0.007	0.015	0.013
Pavement surface condition	−0.007	**0.959**	−0.031	−0.014	−0.029	−0.007	0.020
Lighting condition	0.029	−0.022	**0.909**	0.004	0.007	0.038	0.000
Hour of the day	0.006	−0.021	**0.904**	0.010	−0.009	0.102	−0.014
Segment length	−0.047	−0.010	0.031	**0.844**	0.044	−0.061	0.010
Urban/rural designation	−0.361	0.022	−0.020	**0.724**	0.070	0.037	−0.012
Drivers' gender	0.054	−0.043	0.015	0.133	**0.754**	−0.163	−0.067
Drivers' age	−0.069	0.018	−0.041	−0.062	**0.627**	0.528	0.052
Weekday	0.007	0.005	0.135	−0.025	−0.066	**0.743**	0.023
Accident year	0.104	0.004	−0.052	0.166	−0.210	0.233	**0.778**
Roadway characteristics	−0.091	0.047	0.054	−0.284	0.253	−0.334	**0.629**
Percentage of total variance	16.46	13.46	12.00	8.07	7.75	7.32	7.04
Cumulative percentage	16.46	29.92	41.92	50.00	57.75	65.07	72.11

Bold values indicate the variables captured by the corresponding principal component

Table 6 Rotated principal component loadings for nonfatal injury crashes on urban and rural freeways

Variables	Principal components						
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Number of lanes	**0.855**	0.022	0.004	−0.051	−0.003	0.002	−0.023
AADT	**0.754**	−0.039	0.029	−0.313	−0.046	0.000	−0.056
Median type	**−0.719**	−0.029	0.022	0.026	−0.008	0.010	−0.052
Weather condition	0.016	**0.951**	−0.001	0.021	−0.024	0.013	0.011
Pavement surface condition	0.010	**0.950**	−0.032	0.001	−0.027	0.010	0.033
Hour of the day	0.002	−0.031	**0.902**	−0.010	0.003	−0.004	0.034
Lighting condition	−0.004	0.000	**0.899**	−0.021	−0.027	−0.011	−0.004
Segment length	−0.042	0.006	−0.009	**0.848**	−0.020	−0.010	−0.026
Urban/rural designation	−0.299	0.018	−0.024	**0.767**	0.068	−0.021	−0.004
Drivers' age	−0.015	0.012	0.009	−0.023	**0.765**	0.024	−0.027
Drivers' gender	−0.011	−0.052	−0.027	0.057	**0.728**	−0.027	0.007
Accident year	−0.016	0.026	−0.020	−0.032	−0.008	**0.973**	−0.002
Weekday	0.098	−0.065	0.171	0.149	0.040	0.152	**0.781**
Roadway characteristics	−0.113	0.122	−0.150	−0.212	−0.072	−0.178	**0.616**
Percentage of total variance	16.35	13.45	11.79	8.44	7.67	7.16	6.99
Cumulative percentage	16.35	29.80	41.59	50.03	57.70	64.86	71.85

Bold values indicate the variables captured by the corresponding principal component

Table 7 Rotated principal component loadings for PDO crashes on urban and rural freeways

Variables	Principal components						
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Number of lanes	**0.864**	0.004	0.007	−0.056	0.003	−0.023	0.004
Median type	**−0.752**	−0.032	0.014	0.003	−0.021	−0.026	0.000
AADT	**0.730**	−0.018	0.013	−0.353	−0.071	−0.025	−0.009
Weather condition	0.016	**0.956**	0.005	0.012	−0.023	0.025	0.013
Pavement surface condition	0.012	**0.955**	−0.023	−0.008	−0.030	0.047	0.010
Hour of the day	0.018	−0.019	**0.902**	−0.028	−0.014	0.035	−0.012
Lighting condition	−0.020	0.004	**0.897**	−0.007	−0.027	0.010	−0.001
Segment length	−0.030	−0.008	−0.017	**0.851**	−0.017	−0.025	0.001
Urban/rural designation	−0.254	0.012	−0.016	**0.796**	0.061	−0.030	−0.010
Drivers' gender	0.021	0.025	−0.064	0.058	**0.744**	0.101	−0.026
Drivers' age	−0.042	−0.069	0.029	−0.020	**0.730**	−0.129	0.026
Weekday	0.068	−0.076	0.210	0.098	0.048	**0.746**	0.079
Roadway characteristics	−0.081	0.151	−0.159	−0.158	−0.079	**0.650**	−0.075
Accident year	−0.005	0.022	−0.014	−0.009	−0.002	0.004	**0.993**
Percentage of total variance	16.38	13.61	12.07	8.63	7.56	7.14	6.93
Cumulative percentage	16.38	29.98	42.06	50.69	58.25	65.39	72.32

Bold values indicate the variables captured by the corresponding principal component

straight-grade, curve-level, and curve-grade), which is captured by PC7. The second factor, PC2, of the freeway fatal crash loadings accounting for about 13.5% of the total variance, is highly correlated with the weather (normal, snowing, raining) and the pavement surface condition (snow, wet, dry) during the crash. Factor 7 of both the aggregate dataset (Table 4) and the PDO crashes (Table 7) and Factor 6 of the non-fatal injury crashes captured a single independent variable representing the impact of the crash year. These factors do not have a considerable association with the other 13 variables in their respective crash categories. Another single variable captured by Factor 6 of the fatal crash dataset is the weekday/weekend, whereby the crash occurred. Again, the results of the PCA analysis shown in Tables 4, 5, 6 and 7 suggest the fact that the effects of human factors on traffic safety on urban and rural interstate highway segments can linearly be combined into a single factor accounting for both drivers' age

and gender (PC5). Similarly, the effects of the prevailing weather and the pavement surface condition can be combined into a single factor PC2. The lighting condition and the hour of the day where the crash is observed can be linearly combined into a factor PC3. As demonstrated above, PCA is a valuable technique to reduce the variables into more manageable dimensions. The results above demonstrate the functionality of the statistical PCA technique as a dimension reduction tool while embedding a significant aspect of the original information. This is performed by involving statistical procedures linearly combining correlated explanatory variables into uncorrelated principal components (PCs). Each of the PCs generated accounts for a certain percentage of the variability in the original data. The PCs retained can be used for further analysis, including as input variables for statistical modeling of crash prediction modeling and other studies [36, 37].

5 Conclusions

This study used 14 explanatory variables representing a wide range of potential roadway traffic safety influencing factors and analyzed the principal components for the likelihood of occurrences of crashes on interstate highways. The analysis used an aggregate crash dataset and three different categories of crashes, namely fatal, nonfatal injury, and non-injury crashes. The results of KMO statistics and Bartlett's test of sphericity strongly suggest that PCA approach for analyzing the four datasets described above is valid. Interestingly, irrespective of the crash category, over 70% of the variances in the original variables can well be captured by seven-factor, cutting down the independent variables by half. Here, it is worth mentioning that the results of the principal component analyses revealed the facts that the two human factors are linearly combined to form a single factor, and the prevailing weather condition is correlated with the pavement surface condition, and the hours of the day are correlated to the lighting condition to form their respective factors. It is also found that, except for the fatal crash category where the accident year appears to have a reasonable combination with the roadway characteristics of the segment, the effect of the accident years is mainly accounted under a single factor with an insignificant combination with the other variables. The findings of this study validate the efficacy of the principal component analysis technique for its application to traffic safety studies, including reduction of the dimensions of the data and crash predictive modeling using either principal components or principal variables identified as independent variables. The results of the study also verified the ability of the PCA in identifying the most significant independent variables so that an informed decision could be made in removing the noncritical and redundant variables. The application of PCA for data reduction in a wide range of disciplines is well established. Here, it is important to understand that the dimensional reduction procedure followed by the standard statistical PCA approach is performed through a linear transformation of the linearly correlated variables, without addressing the non-linear relationship between the original variables. This inherent limitation of the technique, makes principal component modeling and interpretation of the results challenging.

Acknowledgements The author would like to thank the U.S. Department of Transportation, Federal Highway Administration (FHWA); Highway Safety Information System (HSIS) for providing the data.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

1. Global status report on road safety (2018) Geneva: World Health Organization (WHO); 2018. License: CC BYNC-SA 3.0 IGO
2. Road Safety Facts, Association for Safe International Road Travel (ASIRT). https://www.asirt.org/safe-travel/road-safety-facts/. Accessed 22 July 2019
3. Webb CN (2018) Motor vehicle traffic crashes as a leading cause of death in the United States, 2015. Traffic Safety Facts, Crash Stats. Report No. DOT HS 812 499, National Highway Traffic Safety Administration, Washington, DC, February 2018
4. Sun X, Hu H, Habib E, Magri D (2011) Quantifying crash risk under inclement weather with radar rainfall data and matched-pair method. J Transp Saf Secur 3:1–14. https://doi.org/10.1080/19439962.2010.524348
5. Kassu A, Anderson M (2018) Analysis of non-severe crashes on two and four-lane urban and rural highways: effects of wet pavement surface condition. J Adv Transp. https://doi.org/10.1155/2018/2871451
6. Kassu A, Anderson M (2018) Determinants of severe injury and fatal traffic accidents on urban and rural highways. Int J Traffic Transp Eng 8(3):294–308. https://doi.org/10.7708/ijtte.2018.8(3).04
7. Shirmohammadi H, Hadadi F, Saeedian M (2019) Clustering analysis of drivers based on behavioral characteristics regarding road safety. Int J Civ Eng 17(8):1327–1340. https://doi.org/10.1007/s40999-018-00390-2
8. Rencher A (1995) Methods of multivariate analysis. Wiley, New York
9. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York
10. Hair J, Anderson R, Tathan R, Black W (1998) Multivariate data analysis, 5th edn. Pearson, New York
11. Golob TF, Recker WW (2003) Relationship among urban freeway accidents, traffic flow, weather, and lighting conditions. J Transp Eng 129:342–353. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:4(342)

12. Golob TF, Recker W, Pavlis Y (2008) Probablilistic models of freeway safety performance using traffic flow data as predictors. Saf Sci 46:1306–1333. https://doi.org/10.1016/j.ssci.2007.08.007
13. Mercier C, Shelley M, Rimkus J, Mercier J (1997) Age and gender as predictors of injury severity in head-on highway vehicular collisions. Transp Res Rec J Transp Res Board. https://doi.org/10.3141/1581-05
14. Mercier C, Shelley M, Adkins G, Mercier J (1999) Age and gender as predictors of injury severity in broadside and angle vehicular collisions. Transp Res Rec J Transp Res Board. https://doi.org/10.3141/1693-09
15. Caliendo C, Parisi A (2005) Principal component analysis applied to crash data on multilane roads. In: Proceedings of the of 3rd international SIIV Congress 2005
16. Caliendo C, Guida M, Parisi A (2007) A crash-prediction model for multilane roads. Accid Anal Prev 39:657–670. https://doi.org/10.1016/j.aap.2006.10.012
17. Wu Y, Nakamura H, Asano M (2013) A comparative study on crash-influencing factors by facility types on urban expressway. J Mod Transp 21(4):224–235. https://doi.org/10.1016/j.aap.2006.10.012
18. Molla MM (2016) Identification of road traffic fatal crashes leading factors using principal components analysis. Int J Res Eng Tech 05(01):79–84
19. de Andrade L, Vissoci JRN, Rodrigues CG, Finato K, Carvalho E, Pietrobon R, de Souza EM, Nihei OK, Lynch C, Carvalho MDB (2014) Brazilian road traffic fatalities: a spatial and environmental analysis. PLoS ONE 9(1):e87244. https://doi.org/10.1371/journal.pone.0087244
20. Nagendra SMS, Khare M (2003) Principal component analysis of urban traffic characteristics and meteorological data. Transp Res Part D 8:285–297. https://doi.org/10.1016/S1361-9209(03)00006-3
21. Youming T, Deliang Z, Xinyu Z, Na L (2018) Principal component analysis of fatal traffic accidents based on vehicle condition factors. In: The 2018 11th international conference on intelligent computation technology and automation (ICICTA). IEEE, Changsha, China, pp 315–317. https://doi.org/10.1109/icicta.2018.00078
22. Lord D, Mannering F (2010) The statistical analysis of crash frequency data: a review and assessment of methodological alternatives. Transp Res Part A Policy Pract 44(5):291–305. https://doi.org/10.1016/j.tra.2010.02.001
23. Savolainen PT, Mannering FL, Lord D, Quddus MA (2011) The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid Anal Prev 43(5):1666–1676. https://doi.org/10.1016/j.aap.2011.03.025
24. Mannering FL, Bhat CR (2014) Analytic methods in accident research: methodological frontier and future directions. Anal Methods Accid Res 1:1–22. https://doi.org/10.1016/j.amar.2013.09.001
25. Kassu A, Anderson M (2019) Analysis of severe and non-severe traffic crashes on wet and dry highways. Transp Res Interdiscip Perspect 2(100043):1–8. https://doi.org/10.1016/j.trip.2019.100043
26. Chang LY, Chen WC (2005) Data mining of tree-based models to analyze freeway accident frequency. J Saf Res 36(4):365–375. https://doi.org/10.1016/j.jsr.2005.06.013
27. Chang LY, Wang HW (2006) Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accid Anal Prev 38(5):1019–1027. https://doi.org/10.1016/j.aap.2006.04.009
28. Scheetz L, Zhang J, Kolassa J (2009) Classification tree modeling to identify severe and moderate vehicular injuries in young and middle-aged adults. Artif Intell Med 45(1):1–10. https://doi.org/10.1016/j.artmed.2008.11.002
29. Xu X, Šarić Ž, Kouhpanejade A (2014) Freeway incident frequency analysis based on CART method. Promet Traffic Transp 26(3):191–199. https://doi.org/10.7307/ptt.v26i3.1308
30. Karlaftis MG, Golias I (2002) Effects of road geometry and traffic volumes on rural roadway accident rates. Accid Analysis Prev 34:357–365. https://doi.org/10.1016/S0001-4575(01)00033-1
31. Ye X, Wang K, Zou Y, Lord D (2018) A semi-nonparametric Poisson regression model for analyzing motor vehicle crash data. PLoS ONE 13(5):e0197338. https://doi.org/10.1371/journal.pone.0197338
32. Nujjetty AP, Sharma S, Council FM (2015) Guidebook for state data files: Ohio., Federal Highway Administration, Highway Safety Information System (HSIS), Office of Safety and Office of Safety Research & Development, Washington, D.C., June 2015
33. Kassu A, Hasan M, Sileshi R (2018) A comparative evaluation of the effects of categorical factors on the safety of multi-lane Interstate highways. IOSR J Mech Civ Eng 15(4):56–62. https://doi.org/10.9790/1684-1504025662
34. Field A (2015) Discovering statistics using IBM SPSS statistics, 4th edn. Sage, London
35. Cerny BA, Kaiser HF (1977) A study of a measure of sampling adequacy for factor-analytic correlation matrices. Multivar Behav Res 12(1):43–47. https://doi.org/10.1207/s15327906mbr1201_3
36. Naidu VM, Prasad CSRK, Srinivas M, Sagar P (2018) Analysis of cities data using principal component inputs in an artificial neural network. Int J Traffic Transp Eng 8(3):271–281. https://doi.org/10.7708/ijtte.2018.8(3).02
37. Zhang K, Hassan M, Yahaya M, Yang S (2018) Analysis of work zone crashes using the ordered probit model with factor analysis in Egypt. J Adv Transp. https://doi.org/10.1155/2018/8570207