IDENTIFICATION OF DRIVER AND VEHICLE CHARACTERISTICS THROUGH DATA MINING THE HIGHWAY CRASH DATA

Santokh Singh

NCSA, National Highway Traffic Safety Administration
400 Seventh Street SW, Washington, DC 20590
Santokh.Singh@nhtsa.dot.gov

Abstract

A crash can be thought of as a system composed of several elements, including drivers and vehicles that continually interact with each other, while a crash database is a record of the errors attributable to different components of the crash system. Learning from mistakes (errors) is important if crashes are to be avoided. With more than one hundred variables related to the drivers, occupants, crash sites and vehicles involved in crashes, the General Estimates System database contains crucial information about the phenomena of crash occurrence. This information can be used to develop crash countermeasures at all levels, including drivers, vehicles and roadways. One of the ways to achieve this objective is to explore the data for any patterns that exist among drivers, vehicles, and roadways.

In this study, we identify driver and vehicle characteristics that contributed to their crash involvement. Preliminary analysis was conducted for selection of crash variables that were relevant to drivers and vehicles involvement in crashes. One of the data mining techniques called "principal components analysis" was further used to identify age- and gender-based groups of drivers and body types of vehicles by highlighting their relation with the crash variables. Some of the variables that were considered in this study included distraction, drinking, speeding etc. (at driver level), and vehicle contributing factors, vehicle's control and the path prior to its initial involvement in the crash (at vehicle level). This in turn helped in identifying the hidden characteristics that may have adversely influenced the driving behavior of drivers and/or running of vehicles, eventually resulting in crashes.

Key words: age, body type, characteristics, gender, patterns, principal components.

1. Introduction

Every year, motor vehicle crashes on the US roadways cause the loss of thousands of lives, in addition to the enormous cost that the economy has to bear. Based on the databases, the General Estimates System (GES) of the National Automotive Sampling System (NASS) and the Fatality Analysis Reporting System (FARS), compiled by the National Highway Traffic Safety Administration (NHTSA), of the estimated 6,323,000 police reported motor vehicle crashes that occurred in the U.S. in 2001, about one third resulted in injuries. These crashes proved to be fatal for 25,840 drivers, 10,546 other occupants, 1,313 children under age nine, and 4,882 pedestrians. In its efforts to save lives and property, NHTSA at the U.S. Department of Transportation has initiated programs, such as the Intelligent Transportation System (ITS) which focused on the Collision Avoidance Systems (CAS) to develop technologies that can assist drivers in avoiding a crash. An elaborate report on the benefits of CAS by NHTSA Benefits Group presents detailed analysis to estimate the impact of crash avoidance systems by using the best available estimates of system and driver performance. In addition, many studies were conducted to appraise the technical aspects of the issue, e.g., Martin et al. [2], Olsen and Wierwille [3], Burgett and Gunderson [4], Burgett and Miller [5], Martin and Burgett [6]. These studies mainly focus on the technical aspects of the crash avoidance issue. The present study aims at statistically investigating some of the possible factors that can be associated with drivers and vehicles contribution to the occurrence of crashes.

NHTSA compiles huge amounts of data about drivers and vehicles involved in crashes as well as other details of the crashes every year. With more than one hundred crash-related variables, these data contain useful information about crashes that can be utilized to plan efforts, such as data collection, aimed at developing crash countermeasures at all levels, including drivers, vehicles and roadways. The objective of this study is to explore the data for any patterns that exist among drivers and vehicles with respect to certain crash characteristics. This in turn can help in understanding the extent and the ways in which drivers and vehicles contribute to the occurrence of crashes. The identification of problem drivers and vehicles through these patterns can provide guidelines for future efforts in the direction of crash avoidance. We use data mining tools to recognize patterns among drivers and vehicles with respect to the factors, such as 'alcohol involvement of driver', 'speeding', ‘failure/malfunction of vehicle brakes’, etc., that typically characterize the occurrence of a crash. While the patterns among drivers can provide guidelines for the traffic rule making and driver education programs, etc., the patterns among vehicles can provide useful hints for designing a sample for data collection with the purpose of crash avoidance.

2. Objective

A crash can be considered as a system with its components as DRIVER, VEHICLE, ROADWAY, OCCUPANTS and OUTSIDE, as shown in Figure1. These components continually interact with each other to produce driving scenarios, one of which changes into a crash scenario, eventually resulting in a crash. In fact, a crash is the resultant effect of interactions between two or more of these components. Thus, there are five main sources that are likely to contribute to a crash-causing error. In this study, we will consider two of these sources: DRIVERS and VEHICLES and identify their characteristics that are likely to contribute to the emergence of crash scenarios. This in turn can identify the driver/vehicle error that possibly contributed to a crash. The basic idea behind the analysis conducted toward crash avoidance in the present study is to ‘*learn from mistakes*’.

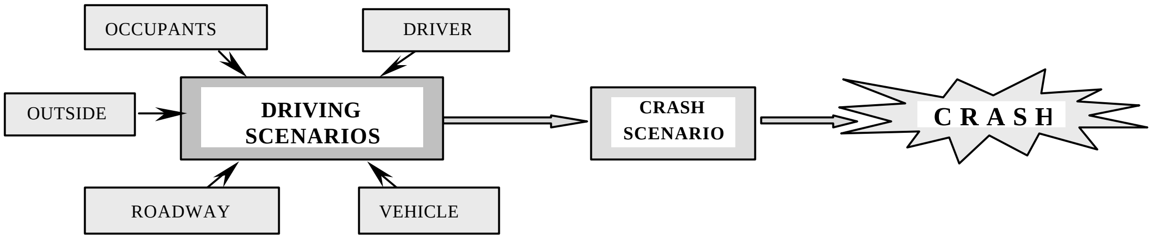


Figure 1. A crash looked at as a system

It would be ideal if we could consider all, or at least a large number, of crash contributing factors (variables) together and make inferences about the hidden structure among drivers and vehicles with respect to these variables. Unfortunately, due to the correlations that exist among some of these variables, the process of making inferences may become extremely difficult and even confusing. Data mining techniques, such as Principal Components Analysis (PCA) [7] can be used to reduce the dimension of the problem to a smaller (manageable) number of uncorrelated variables and yet consider a large number of variables of interest that may be correlated. In this way, clear and reliable inferences can be made about the hypotheses. This multivariate technique will be used for identifying driver and vehicle characteristics that most likely contributed to their crash involvement.

3. Preliminary Analysis: Selection of the Driver and Vehicle Identifying Variables

Preliminary analysis was conducted to select relevant crash-related variables from GES database that would characterize crash involved drivers and vehicles. Contingency analysis was used for testing the hypotheses of independence. Since the collection of GES data is based on three-stage sampling, the statistical software SUDAAN 8.01 was used for this purpose, which takes into account the underlying sampling design of the data being used in the analysis .

3.1. Analysis Variables for the Crash-System Component: DRIVER

Driving is a task performed by a DRIVER in a vehicle on the roadway that continually requires attention, decision-making, and use of reflexes. It is, therefore, likely that the human attributes ‘age’ and ‘gender’ play important roles in the performance of this task. In this study, we will consider groups of drivers based on these attributes and associate them with the driver-related factors that possibly contributed to their involvement in crashes. Sixteen groups were considered. Eight age groups, A1 (younger than 18), A2 (18 to 24), A3 (25 to 34), A4 (35 to 44), A5 (45 to 54), A6 (55 to 64), A7 (65 to 74) and A8 (above 74) were created. Two groups were further created from each of these groups, depending upon gender of drivers, thus resulting in sixteen groups. For the following analyses, we will denote male drivers of age group Ai as MAi, i = 1, 2, ... 8 and female drivers of age group Ai as FAi, i = 1, 2, ... 8.

As far as the selection of driver-related crash variables is concerned, contingency analysis was conducted to test association between driver’s age/sex and some of the crash variables. Several variables recorded in GES were considered. The results of contingency analysis are presented in Table 1, which shows values of χ^2, the corresponding degrees of freedom and p-values.

These results show that the variables, ‘Driver drinking in vehicle’, ‘Driver distracted by’, ‘Speed related’, ‘Pre-crash vehicle control’, ‘Corrective action attempted’ and ‘Driver maneuvered to avoid’ were strongly associated with driver’s age and gender. These variables are, therefore, appropriate for further analysis. In the subsequent discussion, these variables will be referred to, respectively, as Drinking, Distraction, Speeding, PreCrashCntrl, CorrectiveAction and Maneuver.

Table 1. Contingency analysis: Association between Age/Gender and Driver Related Crash Variables

Driver attributes	Driver-related crash variable	c^{2} (Chi-Square)	Degrees of freedom	p-value
Age/Gender	Drinking	856.10	15	0.000
	Distraction	154.74	15	0.000
	Maneuver	288.03	15	0.000
	PreCrashCntrl	618.92	60	0.000
	Speeding	716.48	15	0.000
	CorrectiveAaction	150.21	15	0.000

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

3.2. Analysis Variables for the Crash-System Component: VEHICLE

As described in Section 2, one of the other components of the crash-system is the ‘VEHICLE’. For the purpose of identification of vehicles, we will focus on vehicle body types, including: Convertible, 2-door sedan, 3-door/2-door hatchback, 4-door sedan, Station wagon, Compact utility, Large utility, Minivan, Large van, Compact pickup and Large pickup. In the subsequent discussion, these body types will be referred to, respectively, as Convertible, Sedan2, Hatchback32, Sedan4, StationWag, CompUtility, LargeUtility, MiniVan, LargeVan, CompPickup and LargePickup. There are several possible vehicle-related factors that contribute to crash involvement of a vehicle. A crash, for instance, can occur due to faulty operation/failure of vehicle components, such as brake system, power train system etc. The crash-involved vehicle’s pre-crash stability and the path followed thereafter are among other important crash-related factors. Contingency analysis was conducted to establish association between body type of vehicles and vehicle contributing factors as well as pre-crash control and pre-crash location of the vehicle. The results of contingency analysis are presented in Table 1, which shows values of χ^2, the corresponding degrees of freedom and p-values. These results show that the variables, Vehicle Contributing Factors, Pre- Crash Vehicle Control and Pre-crash Location are strongly associated with vehicle Body Type. Following these findings, in the multivariate statistical analysis that follows, we will consider Vehicle Contributing Factors, Vehicle Pre-crash Control and Vehicle Pre-crash Locations, and recognize patterns among vehicle body types with respect to these variables.

Table 2. Contingency analysis: Association between vehicle Body Type and Vehicle Related Crash Variables

Vehicle variable	Vehicle crash variable	c^{2} (Chi-Square)	Degrees of freedom	p-value
Body Type	Vehicle Contributing Factors	135.05	90	0.0015
	Pre-Crash Vehicle Control	329.44	30	0.0000
	Pre-crash Location	396.84	60	0.000

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

4. Identification of Driver and Vehicle Characteristics: Methodology

The above analysis provides broad idea about the association between driver attributes age and gender and driver-related crash variables, as well as between body types of vehicles and vehicle-related crash variables. The question “what is associated with what” needs to be answered in order to have a deeper insight into the crash phenomenon. For instance, the question “which driver age- and gender-based group is associated with drinking, or which vehicle body type is associated with pre-crash control as tracking” needs to be answered. In fact, to answer this question, what is needed is to recognize patterns among drivers and vehicles. An analysis, such as PCA, can help in identifying age- and gender-based groups of drivers and body types of vehicles by highlighting the relation among variables and groups. In the present study, PCA was used to determine which characteristics of drivers and vehicles were strongly associated with some of the factors that might have contributed to their involvement in crashes. In this way, PCA can help in identifying the hidden characteristics that may have adversely influenced the driving behavior of drivers and/or running of vehicles, eventually resulting in crashes.

4.1. Principal Components

PCA is a factor analytic technique that systematically combines variables into a set of new variables, called 'principal components' with the objective of producing maximum discrimination among groups with respect to the original variables (Appendix A). In the subsequent discussion, a principal component will be referred to simply as a 'component', which should not be confused with the component of the crash-system as used in Section 2. As variables, these components are independent of each other and can explain a certain proportion of total variance. The components are ordered according to the amount of variance they can explain (Appendix A). Thus, the first component explains the maximum variance, while the last accounts for the least. In the subsequent discussion, these components will be referred to as Component 1, Component 2, etc. The proportion of variance that is 'unexplained' by Component 1 may be explained in part by Component 2, which is independent of the first, so on and so forth. The maximum number of such components can be as large as the number of original variables. In a given situation, however, it may be a much smaller number that would suffice for making reasonable interpretations.

4.2. Interpretation of Principal Components

Interpretation of principal components is crucial to PCA. There are three basic steps involved in this process: (i) selection of dominant principal components, (ii) selection of explicative variables and groups, based on the selected components and (iii) establishing correspondence between the selected variables and groups via the dominant components. In step (i), the dominant components are chosen based on the acceptable proportion of variance (say, 80%) explained by the first few of them or using a scree plot (plot of eigen values of the correlation matrix); in this study, we will use proportion of explained variance to decide the number to be used in interpretation. By construction, a component depends on the variables that are most correlated (positive or negative) with it and hence are capable of explaining this component. Such variables can be identified in the subspace generated by a pair of components by looking at the distances of their representative points from the origin. The rule of thumb is that the farther a variable is from the origin along a PC (in either direction), the higher the correlation of the variable with this PC will be. However, for selection of explicative variables to be used in the interpretation, it is the relative contribution of a variable that is used as a criterion. Thus, Step (ii) consists of selecting those variables for interpretation that are highly correlated with the components selected in step (i) and are represented by the outermost points with respect to a component in a subspace. In fact, these are the variables that can classify groups with maximum discrimination. Choice of groups in step (ii) depends on how much contribution a group makes to the variance of a component; those groups are preliminarily selected that have above-average contribution. How much above average depends on how stringent one would like to be in this selection. The explicative groups finally selected for interpretation are the ones that are represented by the outermost points with respect to a component in the subspace. In fact, these are the groups that stand out from the point of view of their contribution to the occurrence of crashes due to certain crash characteristics. Having chosen the dominant (in terms of the variance explained) components, the variables and groups are projected onto the subspace(s) generated by pairs of PC's, as shown in Figure 2. Although all the variables and groups will be projected onto a subspace, the association between variables and groups is sought only between the explicative variables and explicative groups that are selected in step (ii). The correspondence between groups and crash variables is established using the magnitude and sign of the coordinates (to be referred to simply as coordinates in the subsequent discussion) of cases and variables: farthest positive (negative) group in projection of groups goes with the farthest positive (negative) variable in projection of variables. In this way, a pair of components can guide us in detecting relationships between crash variables and groups (of drivers, based on age and gender; and of vehicles, based on body type).

5. Identification of Drivers

Drivers contribute much more to the occurrences of crashes as compared to other sources, such as vehicles, roadways, etc., and hence need special attention in a study aimed at crash avoidance. Identification of problem drivers based on some of their characteristics can reveal the extent to which drivers contribute to the occurrences of crashes. As mentioned earlier, we will consider sixteen age- and gender-based groups of drivers: MAi, FAi, (i = 1, 2, …, 8) and identify those that contributed to crashes most due to the factors Distraction, Drinking, Speeding, PreCrashCntrl, CorrectiveAction and Maneuver.

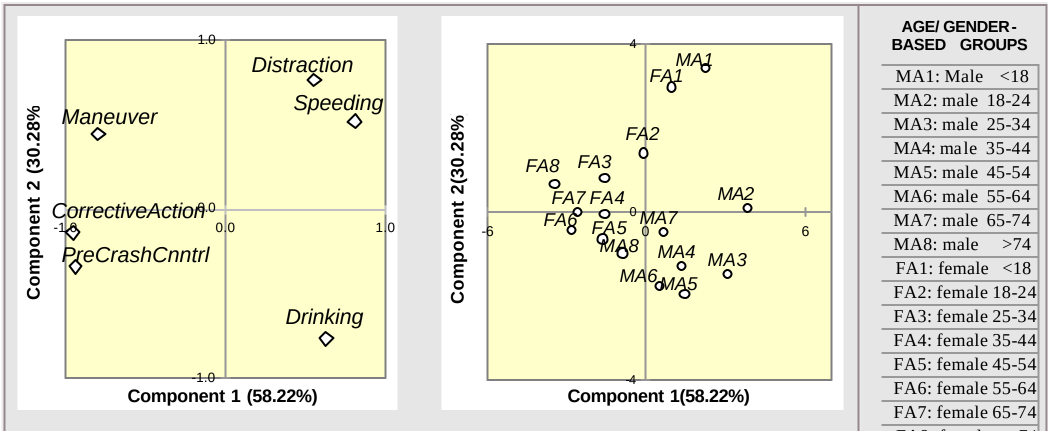
5.1. Data Preparation for Identification of Drivers

Based on GES data for the year 2001, the marginal frequency distribution of each group was evaluated over the variables Distraction, Drinking, Speeding, PrecrashCntrl, CorrectiveAction, and Maneuver. Percent frequencies across these variables for each group were used as measures of their contributions to crashes due to these factors (variables).

5.2. PCA of Driver-Related Variables for Identification of Drivers

PCA conducted for the sixteen age- and gender-based groups of drivers yielded principal components, the first two of which explained 88.5% of variance. The results of PCA discussed in this section are presented in Tables B1.1 and B1.2 (Appendix B). The coordinates of variables show that the variables PrecrashCntrl, CorrectiveAction, Speeding, and Maneuver are significantly correlated with Component 1 (correlations being -0.978, -0.936, 0.921 and 0.730, respectively). Similarly, the variables Distraction, Drinking and Maneuver are significantly correlated with Component 2 (correlations being 0.619, –0.883 and 0.620, respectively). In Figure 2(a), these correlations are represented as distances of the representative points of variables from the origin. It can be seen in this figure that Component 1 classifies variables into two types: those concerned with speeding and the ones concerned with control of the vehicle. Similarly, Component 2 classifies variables into two types: those concerned with distraction and the ones related to alcohol. PCA results further show that the variables Speeding (with +ve coordinates) and CorrectiveAction and PreCrashCntrl (with –ve coordinates) have significant (above average) contributions to Component 1. Among groups of drivers, MA2, MA3 (with +ve coordinates) and FA6, FA7, FA8 (with –ve coordinates) have significant contributions to this component. The coordinates of these variables and groups are plotted, respectively, in Figure 2(a) and Figure 2(b). Based on this information about the crash variables and age- and gender-based groups, the latter can be identified by simultaneously using their coordinates. The coordinates with respect to Component 1 as shown in Figure 2(a) and Figure 2(b) suggest that most of the drivers from MA2 and MA3 contributed to crashes due to Speeding. On the other hand, drivers from FA6, FA7 and FA8, can be closely associated with PreCrashCntrl and CorrectiveAction.

PCA results further show that the variables Distraction (with +ve coordinates) and Drinking (with –ve coordinates) have significant contributions to Component 2. Among groups, this component receives significant contributions from MA1 and FA1 (with +ve coordinates) and MA5 and MA6 (with –ve coordinates). Thus, by using Component 2 in the two projections, it can be concluded that drivers belonging to groups MA1 and FA1 contributed to crashes due to Distraction, while in case of drivers belonging to MA5 and MA6, Drinking might have been the contributing factor in their crash involvement.



AGE/ GENDER-BASED GROUPS
MA1: Male <18
MA2: male 18-24
MA3: male 25-34
MA4: male 35-44
MA5: male 45-54
MA6: male 55-64
MA7: male 65-74
MA8: male >74
FA1: female <18
FA2: female 18-24
FA3: female 25-34
FA4: female 35-44
FA5: female 45-54
FA6: female 55-64
FA7: female 65-74
FA8: female >74

Figure 2(a) Projection of driver-related variables. Figure 2(b) Projection of age- and gender-based groups

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Figure 2. Projection of driver-related variables and age- and gender-based groups on the subspace generated by Component 1 and Component 2.

6. Identification of Vehicles, Based on Vehicle Contributing Factors

Identification of vehicles based on some of their crash-related characteristics can provide better insight into the occurrence of crashes due to vehicles. With this objective in mind, the following analysis is conducted to recognize patterns among vehicles with respect to vehicle contributing factors: Brake system, Power train system, Steering, Suspension and Wheels. For the sake of brevity, these variables will be referred to, respectively, as Brake, Wheels, PoweTrain, Steering, Suspension and Wheels.

6.1. Data Preparation for Identification of Vehicles by Body Type

As mentioned earlier, eleven groups, Convertible, Sedan2, Hatchback32, Sedan4, StationWag, CompUtility, LargeUtility, MiniVan, LargeVan, CompPickup, and LargePickup were considered. Based on GES data for the year 2001, the marginal frequency distribution of each group was evaluated over the vehicle-related variables, Brake, PowerTrain, Steering, Suspension and Wheels. Percent frequencies over these variables for each group were used as measures of their contributions to crashes due to the variables under consideration.

6.2. Identification of Vehicle Body Types, Based on Vehicle Contributing factors

PCA of GES2001 data yielded principal components for groups of vehicles, the first two of which explained 74.27% variance. PCA results presented in Tables B2.1 and B2.2 (Appendix B) also show that the variables Brake, Steering and Suspension are significantly correlated with Component 1 (the respective correlations being, 0.704, 0.709 and 0.863). The variable Brakes and Steering are significantly correlated with Component 2 (the correlation being, 0.615 and 0.615, respectively). In Figure 3(a), these correlations are represented as distances of the representative points of variables from the origin.

As far as the choice of explicative variables for interpretation is concerned, the variable Suspension (with +ve coordinates) has significant contribution to Component 1. In case of vehicle body types, LargePickup and StationWag (with +ve coordinates) have significant contributions to this component. The joint interpretation of Figure 3(a) and Figure 3(b) suggests that the crash involvement of LargePickup and StationWag can be associated with Suspension.

PCA results also show that the variables Break and Steering (with +ve coordinates) and PowerTrain (with –ve coordinates) have significant contributions to Component 2. In case of vehicle body types, LargeVan (with +ve coordinates) and CompPickup and CompUtility (with –ve coordinates) have significant contributions to this component. The coordinates of variables and body types mentioned above with respect to Component 2 (Figure 3(a) and Figure 3(b)) show that LargeVan can be associated with Brake/ Steering and CompPickup and CompUtility with PowerTrain.

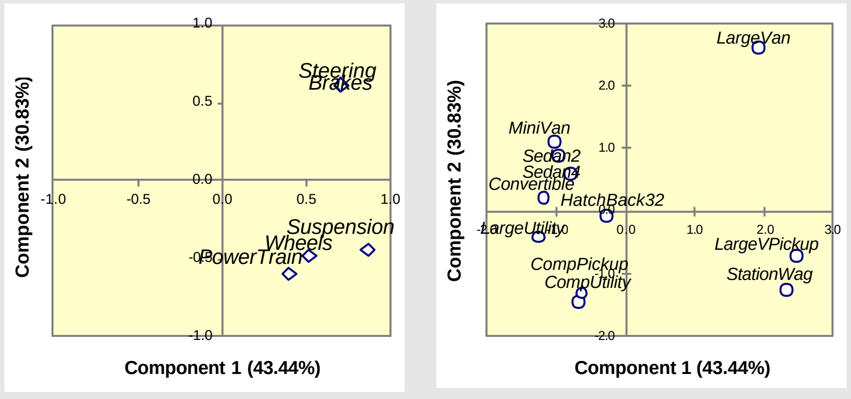


Figure 3 (a) Projection of vehicle Contributing factors. **Figure 3 (b) Projection of vehicle body types**

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Figure 3. Projection of vehicle mechanism-related variables and body types on the subspace generated by Component 1 and Component 2.

Since the first two components explained 74.27% of variance, we need to consider Component 3 to account for additional 18.84% variance, thus making up for 93.11% of the total variance. The variables PowerTrain and Wheels have significant correlations with this component (respective correlations being 0.678 and –0.688). Also, the variables, PowerTrain (with +ve coordinates) and Wheels (with –ve coordinates) have significant contributions to Component 3. In case of vehicles, StationWag and LargeUtility (with +ve coordinates) and CompUtility and LargePickup (with –ve coordinates) have significant contributions to Component 3. Joint interpretation of Figure 4(a) and Figure 4(b) brings out that StationWag and LargeUtility can be associated with PowerTrain and CompUtility; and LargePickup with Wheels.

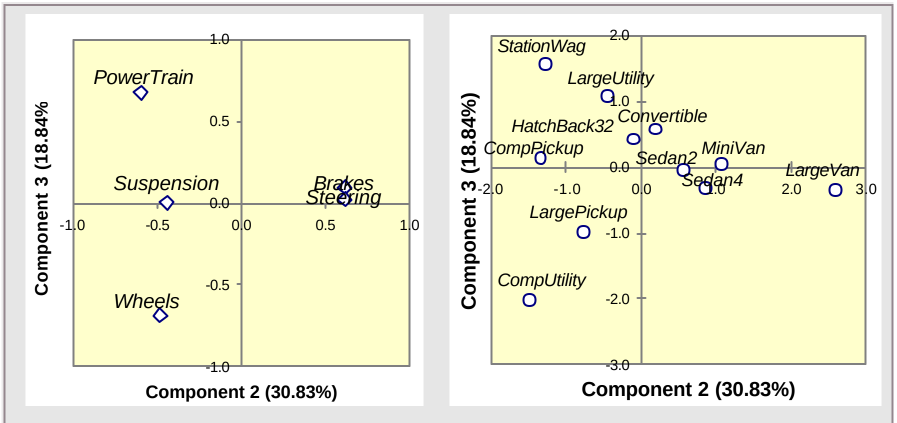


Figure 4(a) Projection of Vehicle contributing factors. **Figure 4 (b) Projection of vehicle body types.**

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Figure 4. Projection of vehicle mechanism-related variables and body types on the subspace generated by Component 2 and Component 3.

7. Identification of Vehicles, Based on Vehicle's Pre-Crash Control

The GES variable 'Pre-Crash Vehicle Control' assesses the stability of the vehicle during the period immediately prior to its initial involvement in the crash sequence. As established earlier in Section 3.2, this variable is associated with the 'Body Type' of a vehicle. In the following analysis, we will recognize patterns among vehicles with respect to this variable. In other words, we will establish correspondence between body types and the type of vehicle stability as defined by the variable Pre-Crash Vehicle Control. In the following analysis, we will consider pre-crash vehicle controls: Tracking, Skidding longitudinally, Skidding laterally-clockwise rotation and Skidding laterally-counterclockwise rotation as variables. For the sake of brevity, these variables will be abbreviated as follows: Tracking (Tracking), Skidding longitudinally (SkidLong), Skidding laterally-clockwise rotation (SkidLatClock), and Skidding laterally-counterclockwise rotation (SkidLatContClock). PCA of these variables was conducted for eleven body types using GES2001 data. The results used in the following discussion are presented in Tables B3.1 and B3.2 (Appendix B). Of the four extracted components, the first two components explained 95.6% of the total variance. Component 1 is highly correlated with Tracking, SkidLong, SkidLatClock and SkidLatContClock (the respective correlations being, 0.996, -0.895, -0.930 and -0.896). This component classifies these variables into two types: Tracking (with +ve coordinates) as opposed to SkidLatClock (with -ve coordinates), both of which have significant contributions to this component. So far as vehicle body types are concerned, MiniVan and StationWag (with +ve coordinates) and HatchBack32 (with -ve coordinates) have significant contribution to the Component 1. The configuration of variables (Figure 5(a)) and of body types (Figure 5(b)) enable us to conclude that MiniVan and StationWag are associated with Tracking, while CompPickup and HatchBack32 with some kind of skidding: SkdLong, SkidLatCounterCloc, or SkidLatClock.

Similarly, Component 2 differentiates variables into two types: SkidLong (with +ve coordinates) as opposed to SkidLatContClock (with negative coordinates). Both these variables have significant contributions to the variance of this component. Speaking about body types, LargeVan (with +ve coordinates) and CompUtility (with -ve coordinates) have significant contributions to the variance of Component 2. This classificatory information about variables and body types brings out that LargeVan can be associated with SkidLong and CompUtility with SkidLatContClock.

Component 3 can be used to identify a few more body types. This component receives significant contributions from SkidLatClock (with +ve coordinates) as opposed to SkidLatContClock (with -ve coordinates). Among body types, Sedan2 (with -ve coordinates) have significant contributions to the variance of Component 3. These two pieces of information put together, enables us to conclude that Sedan2 can be associated with SkidLatContClock.

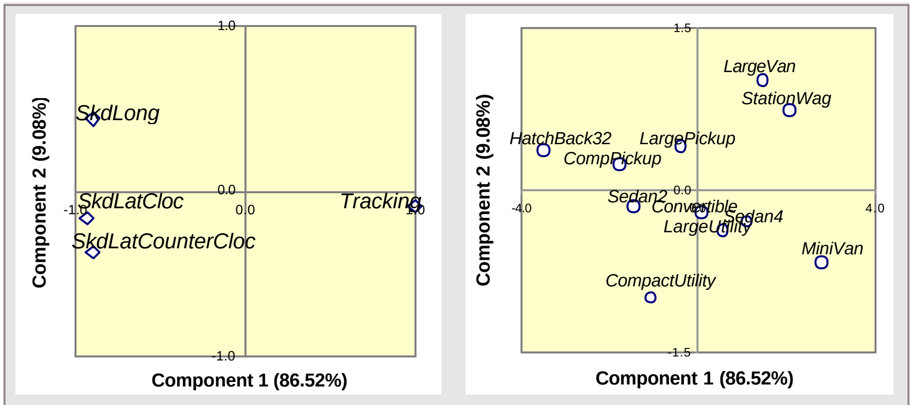


Figure 5(a) Projection of vehicle stability variables. **Figure 5(b) Projection of body types.**

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Figure 5. Projection of vehicle's driving related variables and body types on the subspace generated by Component 1 and Component 2.

8. Identification of Vehicles, Based on Vehicle's Pre-Crash Location

The GES variable 'Pre-crash Location' identifies the path of a vehicle prior to its first involvement in the crash sequence and further reports the results of its pre-crash stability coded in the GES variable 'Pre-crash Vehicle Control'. It has been established through contingency analysis in Section 3.2 that this variable, that is, the path of the vehicle prior to its first involvement in the crash is closely associated with its body type. In this section, we continue data mining so as to be more specific about this association. For instance, we will investigate which body type displayed Pre-crash Location as 'Vehicle stayed in travel lane' most in comparison with other pre-crash locations, such as Vehicle departed roadway, Vehicle remained off roadway, etc. The pre-crash locations: Vehicle stayed in travel lane, Vehicle stayed on roadway but left travel lane, Vehicle departed roadway, Vehicle remained off roadway, Vehicle returned to roadway and Vehicle entered roadway were considered as analysis variables for recognizing patterns among vehicles body types. For the sake of brevity, these variables will be abbreviated as follows: Vehicle stayed in travel lane (StTrLn), Vehicle stayed on roadway but left travel lane (StRdWyLftTrLn), Vehicle departed roadway (DepRdWy), Vehicle remained off roadway (RemOffRdWy), Vehicle returned to roadway (RetRdWy) and Vehicle entered roadway (EntRdWy). PCA of these variables for eleven body types was conducted using GES data for the year 2001. The results used in the following discussion are presented in Tables B4.1 and B4.2 (Appendix B). Of the six extracted components, the first two explained 66% of the total variance. Component 1 is highly correlated with StTrLn, DepRdWy, and EntRdWy with respective correlations -0.916, 0.944, and -0.671. Component 2 is significantly correlated with StRdWyLftTrLn, and RemOffRdWy (correlations being 0.877and -0.763, respectively).

PCA results further show that the variables DepRdWy (with +ve coordinates) and StTrLn (with -ve coordinates) have significant contributions to the variance of Component 1. Also, this component receives significant contributions from HatchBack32 and Convertible (both with +ve coordinates) and MiniVan and LargeUtility (with -ve coordinates). Using this variable and body type information, represented, respectively, in Figure 6(a) and Figure 6(b), it can be established that HatchBack32 and Convertible are associated with DepRdWy, while MiniVan and LargeUtility are associated with StTrLn.

Component 2 receives significant contributions to its variance from the variables StRdWyLftTrLn (with +ve coordinates) and RmnOffRdWy (with –ve coordinates). Similarly, body types StationWag and Sedan2 (with +ve coordinates) and LargePickup and CompPickup (with -ve coordinates) have significant contributions to the variance of this component. The configuration of variables and body types is presented in Figure 6(a) and Figure 6(b), respectively. Joint interpretation of the results presented in these figures, brings out that StationWag and Seadn2 are associated with DepRdWy, while LargePickup and CompPickup with StTrLn.

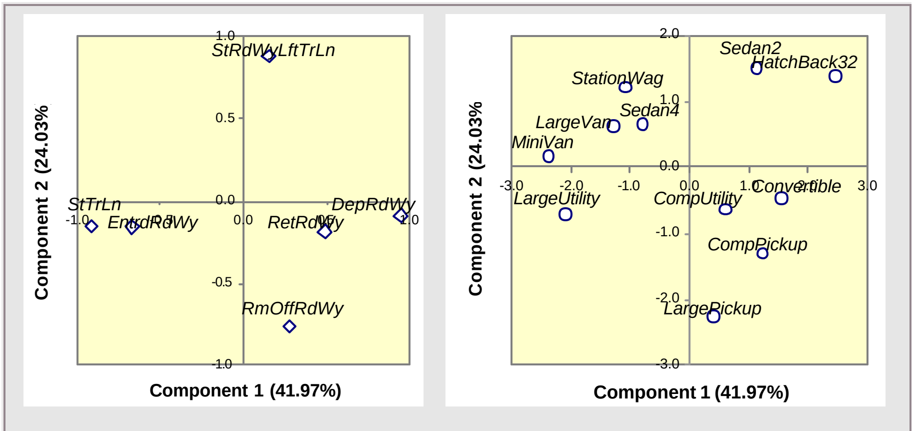


Figure 6(a) Projection of vehicle path.

Figure 6(b) Projection of body types.

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Figure 6. Projection of vehicle stability variables and body types on the subspace generated by Component 1 and Component 2.

Component 3 explains an additional 16.58% variance, thus making for 82.59% of the total variance. This component is significantly correlated with RetRdWy (correlation being 0.689). There are only two variables: RetRdWy and EntrRdWy with significant contributions to the variance of this component (both with positive coordinates) that can be used for identifying more body types. Among body types, LargeVan has significant contribution to the variance of this component and have positive coordinates. These results show that LargeVan is associated with RetRdWy and EntrRdWy.

9. Conclusions

Planned experiments in real-life driving conditions with the purpose of data collection are a must if future efforts in the direction of crash avoidance are to be effective. However, in certain situations when field experiments are not feasible, simulated experiments, too, can provide a good amount of information. In either case, the identification of driver characteristics is important for designing an efficient sample. This study shows that different age- and gender-based groups of drivers exhibit different tendencies resulting in their crash involvement. While teenage drivers' involvement can be associated with speeding, that of young drivers with drinking. In fact, by using the data mining technique PCA, we could identify most of the age- and gender-based groups with respect to the crash variables, thereby establishing specific associations between groups of drivers and their crash characteristics. This shows that the data collection aimed at crash avoidance measures must include both male and female drivers of different age groups. However, if the aim is to study only alcohol involvement of drivers, then the main focus needs to be on young drivers of both genders, while in a study aimed at speeding, the subpopulation of teenage drivers needs to be the target population.

PCA of GES 2001 data for identifying vehicles with respect to some of the crash variables, throws more light on the crash phenomenon. The results obtained in this study show that among other factors, body types of vehicles are likely to have an impact on how the vehicle stays in control immediately prior to the crash and the path it is likely to assume thereafter. In fact, through the patterns recognized among vehicle-related crash variables and vehicle body types, the body types could be identified that can be associated with the pre-crash vehicle control: tracking as opposed to those that can be associated with skidding. Similarly, through data mining, vehicle body types could be identified with respect to pre-crash location of the vehicle. Keeping in mind these facts, an efficient sample design for data collection for the purpose of crash avoidance should use two-way stratification, using driver's age/gender and vehicle body type as stratification criteria

10. References

[1] NHTSA Benefits Working Group, 'Preliminary Assessment of Crash Avoidance Systems Benefits', NHTSA Technical Report, 1994.

[2] Martin, Peter G., Burgett, August L., and Srinivasan, Gowri, 'Characterization of a Single Vehicle Road Departure Avoidance Maneuver', Paper 308, 2003.

[3] Oslen, Erik C. B. and Wierwille, Walter W.,‘ A Unique Approach for Data Analysis of Naturalistic Behavior’, Technical Report (2001-01-2518), Virginia Tech Transportation Institute, 2001.

[4] Burgett, August and Gunderson, Kirsten, ‘Crash Prevention Boundary for Road Departure Crashes – Derivation’, Research Note DOT HS 809 399, September 2001.

[5] Burgett, August and Miller, Robert J., Jr., ‘A New Paradigm for Rear-end Crash Prevention Driving Performance’, NHTSA, Research and Development, 2003.

[6] Martin, Peter G., Burgett, August, ‘Rear-end Collision Events: Characterization of Impeding Crashes’, Proc. 1st Human-Centered Transportation Simulation Conference, November 2001.

[7] Michel Jambu, Exploratory and Multivariate Data Analysis, Academic Press, Boston, 1991.

11. Appendix A. Analytical and statistical details of PCA

This appendix provides analytical and statistical details of PCA as supplement to the methodology used in Section 4.

11.1. Analytical

The mathematical technique used in PCA is an Eigen analysis in which we solve for the eigenvalues and eigenvectors of a square symmetric matrix with sums of squares and cross products. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component. The sum of the eigenvalues equals the trace of the square matrix and the maximum number of eigenvectors equals the number of rows (or columns) of this matrix.

Consider a set of p random variables $\underline{X'} = [X_1\ X_2\ \ldots X_p]$. PCA linearly combines these variables, resulting into a new set of p variables called principal components, such that each of them captures maximum possible variation in X and is independent of all other principal components. In mathematical terms, a principal component is given by

$$\underline{Y} = C\underline{\beta},$$

where C is $n \times p$ data matrix of standardized values and the p-vector of unknowns $\underline{\beta}$ is a determined by maximizing $\underline{\hat{\beta}'} C' C \underline{\hat{\beta}}$, such that the vector $\underline{\beta}$ is a normalized vector, i.e., $\underline{\beta'}\underline{\beta} = 1$. In fact, maximization of $\underline{\hat{\beta}'} C' C \underline{\hat{\beta}}$, subject to the condition $\underline{\beta'}\underline{\beta} = 1$ is done by diagonalization of $C'C$, yielding the eigenvalues $\lambda_1, \lambda_2, .., \lambda_p$ with $\underline{\hat{\beta}_1}, \underline{\hat{\beta}_2}, \ldots, \underline{\hat{\beta}_p}$ as the respective associated eigenvectors.

11.2. Statistical

One of the important properties of the principal components is their independence of each other. To prove this contention, we use the fact that if **P** is a matrix of eigenvectors calculated for a symmetric matrix **A,** then $R'AR = \mathbf{D}$ is a diagonal matrix with eigenvalues of **A** on the diagonal and, of course, zero elsewhere. Now, let U be an $n \times p$ matrix containing the p principal components. Then the sample covariance matrix of U is equal to $B' C' C \underline{B}$. Since B is the matrix of eigenvectors calculated from the symmetric matrix $C'C$, the covariance matrix $B' C' C \underline{B}$ must be a diagonal matrix with eigenvalues (variances of the components) on the main diagonal and zeros elsewhere. Hence, all covariances are zero, thereby showing that the principal components are independent of each other.

12. Appendix B. Tables showing PCA results: Contributions and Coordinates of Variables and Groups

This appendix provides results of PCA for different driver- and vehicle -related variables, used in Sections 4 thro Section 8.

Table B1.1. Contributions and Coordinates of driver-related Variables

Driver-related variables	Contributions			Coordinates		
	Component 1	Component 2	Component 3	Component 1	Component 2	Component 3
Distraction	0.136	0.228	0.262	0.730	0.619	-0.286
Drinking	0.053	0.463	0.006	0.457	-0.883	0.044
Maneuver	0.126	0.229	0.367	-0.704	0.620	0.338
PreCrashCntrl	0.244	0.011	0.030	-0.978	-0.136	-0.097
CorrectiveAction	0.224	0.002	0.332	-0.937	0.059	-0.321
Speeding	0.217	0.067	0.004	0.922	0.336	0.034

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Table B1.2. Contributions and Coordinates of age- and gender-based groups of drivers, based on driver-related variables

Driver groups	Contributions			Coordinates		
	Component 1	Component 2	Component 3	Component 1	Component 2	Component 3
MA1	15.276	26.063	14.217	2.998	2.565	0.815
MA2	23.013	0.926	5.542	3.679	-0.483	-0.509
MA3	10.371	10.932	0.525	2.470	-1.661	0.157
MA4	0.890	5.212	9.485	0.724	-1.147	0.666
MA5	1.346	13.594	1.690	0.890	-1.852	-0.281
MA6	0.026	8.812	1.293	0.124	-1.491	-0.246
MA7	0.154	0.785	28.182	-0.301	-0.445	1.148
MA8	1.758	2.273	5.717	-1.017	-0.757	-0.517
FA1	4.275	18.948	7.090	1.586	2.187	-0.576
FA2	0.383	5.223	12.635	0.475	1.148	-0.769
FA3	2.195	2.947	2.042	-1.136	0.862	0.309
FA4	2.843	0.076	0.016	-1.293	0.139	0.027
FA5	3.831	0.537	2.515	-1.501	-0.368	-0.343
FA6	10.884	0.064	3.200	-2.530	0.127	0.387
FA7	10.169	0.267	0.946	-2.446	0.259	0.210
FA8	12.585	3.342	4.904	-2.721	0.918	-0.479

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Table B2.1. Contributions and Coordinates of vehicle contributing variables

Variables	Contributions			Coordinates		
	Component 1	Component 2	Component 3	Component 1	Component 2	Component 3
Brake	0.228	0.245	0.010	0.704	0.615	0.098
Steering	0.232	0.245	0.001	0.709	0.615	0.023
Suspension	0.343	0.126	0.000	0.863	-0.441	0.001
PowerTrain	0.073	0.232	0.487	0.399	-0.598	0.678
Wheels	0.124	0.151	0.502	0.518	-0.482	-0.688

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Table B2.2. Contributions and Coordinates of vehicle body types, based on vehicle contributing factors

Groups	Contributions			Coordinates		
	Component 1	Component 2	Component 3	Component 1	Component 2	Component 3
Convertible	6.331	0.252	3.434	-1.173	0.197	0.569
Sedan2	2.769	2.184	0.046	-0.776	0.580	-0.066
HatchBack32	0.337	0.060	1.824	-0.271	-0.096	0.415
Sedan4	4.213	4.945	1.168	-0.957	0.873	-0.332
StationWag	25.110	10.416	25.589	2.335	-1.267	1.553
CompUtility	2.104	14.183	43.791	-0.676	-1.479	-2.031
LargeUtility	7.103	1.208	12.175	-1.242	-0.431	1.071
MiniVan	4.794	7.712	0.013	-1.021	1.090	0.035
LargeVan	17.182	44.004	1.321	1.932	2.604	-0.353
CompPickup	1.824	11.451	0.185	-0.629	-1.329	0.132
LargePickup	28.233	3.585	10.454	2.476	-0.743	-0.992

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Table B3.1. Contributions and Coordinates of Pre-crash control variables

Pre-crash control variables	Contributions		Variables	
	Component 1	Component 2	Component 1	Component 2
Tracking	0.287	0.023	0.996	-0.091
SkidLong	0.232	0.527	-0.895	0.438
SkidLatrlClock	0.250	0.072	-0.930	-0.162
SkidLatContClock	0.232	0.378	-0.896	-0.371

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Table B3.2. Contributions and Coordinates of vehicle body types, based on Pre-crash control variables

Body types	Contributions		Variables	
	Component 1	Component 2	Component 1	Component 2
Convertible	0.031	1.229	0.103	-0.211
Sedan2	5.996	0.649	-1.440	-0.153
HatchBack32	35.075	3.634	-3.484	0.363
Sedan4	3.607	2.486	1.117	-0.300
StationWag	12.438	14.478	2.075	0.725
CompUtility	3.300	27.496	-1.069	-0.999
LargeUtility	0.925	4.077	0.566	-0.385
MiniVan	22.795	12.357	2.809	-0.670
LargeVan	6.342	27.830	1.481	1.005
CompPickup	9.058	1.457	-1.770	0.230
LargePickup	0.433	4.307	-0.387	0.395

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Table B4.1. Contributions and Coordinates of pre-crash location variables

Pre-crash location	Contributions				Coordinates			
	Component 1	Component 2	Component 3	Component 4	Component 1	Component 2	Component 3	Component 4
StTrLn	0.333	0.017	0.088	0.054	-0.916	-0.155	0.295	-0.189
StRdWyLftTrLn	0.009	0.534	0.079	0.066	0.153	0.877	-0.281	0.210
DepRdWy	0.354	0.005	0.020	0.006	0.944	-0.087	-0.140	-0.066
RemOffRdWy	0.030	0.404	0.213	0.004	0.275	-0.763	-0.460	0.049
RetToRdWy	0.094	0.023	0.476	0.354	0.488	-0.184	0.689	0.487
EntRdWy	0.179	0.017	0.125	0.516	-0.671	-0.157	-0.352	0.588

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001

Table B4.2. Contributions and Coordinates of body types, based on pre-crash location variables

Body types	Contributions				Coordinates			
	Component 1	Component 2	Component 3	Component 4	Component 1	Component 2	Component 3	Component 4
Convertible	9.848	1.630	9.965	44.317	1.575	-0.485	-0.996	-1.721
Sedan2	5.295	15.359	6.322	0.013	1.155	1.488	-0.793	0.030
HatchBack32	24.311	13.143	3.492	7.279	2.474	1.377	0.590	0.698
Sedan4	2.362	2.828	0.568	3.964	-0.771	0.639	-0.238	-0.515
StationWag	4.413	9.900	0.101	2.043	-1.054	1.195	0.101	-0.370
CompUtility	1.562	2.897	23.321	0.068	0.627	-0.646	1.524	0.067
LargeUtility	16.903	3.761	10.127	0.125	-2.063	-0.736	-1.004	-0.092
MiniVan	22.100	0.175	27.787	3.391	-2.359	0.159	1.663	-0.476
LargeVan	6.250	2.489	9.008	30.972	-1.255	0.599	-0.947	1.439
CompPickup	6.273	12.036	5.344	1.070	1.257	-1.317	0.729	0.267
LargePickup	0.681	35.781	3.964	6.758	0.414	-2.271	-0.628	0.672

Data source: National Center for Statistics and Analysis, NHTSA, GES 2001