



# Research on the Relationship Between Driver Age and Traffic Risk Factors Based on K-means++ Clustering

Rongfang Zhang\*, Yufei Ma, Jingsheng Chen, Zihao He

Chang'an University, Xi'an 710064, Shaanxi, China.

**How to cite this paper:** Rongfang Zhang, Yufei Ma, Jingsheng Chen, Zihao He. (2026) Research on the Relationship Between Driver Age and Traffic Risk Factors Based on K-means++ Clustering. *Engineering Advances*, 6(1), 55-60.  
DOI: 10.26855/ea.2026.03.012

**Received:** January 29, 2026  
**Accepted:** February 28, 2026  
**Published:** March 31, 2026

\***Corresponding author:** Rongfang Zhang, Chang'an University, Xi'an 710064, Shaanxi, China.

## Abstract

This study explores the relationship between driver age and traffic risk factors through cluster analysis to support targeted traffic safety education. Based on expressway accident data collected in Xianyang, Shaanxi Province, from 2021 to 2023, the K-means++ algorithm was used to identify the risk tendencies of different driver groups. The results show that the optimal number of clusters is three. Group 1 includes young and elderly drivers and is more likely to be involved in casualty accidents, mainly related to rear-end collisions and following operation errors. Group 2 includes young and middle-aged drivers with lower casualty risk but a stronger tendency toward violations, mainly involving oblique and rear-end collisions, as well as steering and following operation errors. Group 3 also consists of young and middle-aged drivers and has the largest number of accidents, with relatively high proportions of rear-end, frontal, and side collisions. Overall, the following operation errors are the key risk factors, suggesting the need for both educational and technological interventions.

## Keywords

Traffic safety; k-means++; driver groups; traffic risks

## 1. Introduction

Traffic safety has become an important public concern with continuing urbanization and motorization. Human factors play a dominant role in traffic accidents, and driver characteristics such as age, gender, and driving experience are closely related to accident type and risk exposure [1-6]. Drivers of different age groups differ in physiological function, cognitive ability, and driving experience, which may lead to different traffic risk patterns. Therefore, examining the relationship between driver age and traffic risk factors is important for identifying age-related risks and improving traffic safety management.

Previous studies have examined the relationship between driver age and traffic risk factors from different perspectives, showing that age is associated with driving behavior, visual perception, situational awareness, and crash involvement [7-13]. However, limited attention has been paid to the interaction between age and specific traffic risk characteristics. Therefore, this study uses real traffic accident data to examine age-related traffic risks through cluster analysis.

## 2. Data and Methods

### 2.1 Data

The data for this study were derived from motor vehicle traffic accident records collected on a highway in Xianyang, Shaanxi Province, between 2021 and 2023. Cases with complete records on driver age, accident casualties, collision

type, unsafe driver behavior, and vehicle safety status were retained. After screening, 2,842 cases were initially obtained. Drivers aged 60 and above were excluded because of their relatively small proportion in the dataset, resulting in a final sample of 2,823 cases.

For quantitative analysis, collision types were coded from 1 to 6 according to severity, and unsafe driver behaviors were coded from 1 to 9, where 1-3 represent physiological and psychological factors, 4-6 represent safety awareness factors, and 7-9 represent operational errors. Descriptive statistics are shown in Table 1.

**Table 1. Descriptive Statistics of the Data**

| Characteristics                   | Total | Mean   | Std. Deviation | Minimum | Maximum |
|-----------------------------------|-------|--------|----------------|---------|---------|
| Number of casualties in accidents | 70    | 1.710  | 1.006          | 1       | 5       |
| Driver age                        | 2823  | 37.973 | 9.676          | 18      | 60      |
| Collision type                    | 2823  | 3.652  | 1.746          | 1       | 6       |
| Unsafe behaviors of drivers       | 2823  | 5.841  | 3.465          | 0       | 9       |
| Unsafe condition of vehicles      | 2823  | 0.418  | 0.501          | 0       | 3       |

## 2.2 Methods

### 2.2.1 K-means++

K-means++ is an improved version of the traditional K-means algorithm. Optimizing the selection of initial cluster centers improves clustering stability and efficiency. In this study, K-means++ was used to classify driver groups with similar traffic risk characteristics.

### 2.2.2 Distance measurement

The distance from each data point to the cluster center is measured using the Euclidean distance, denoted as:

$$d(x, \mu) = \sqrt{\sum_{i=1}^n (x_i - \mu_i)^2} \quad (1)$$

### 2.2.3 Pseudo F-statistic

The Pseudo F-statistic was used to evaluate clustering performance under different numbers of clusters. A larger value indicates better clustering performance. The formula is shown below, where  $k$  represents the number of clusters and  $n$  denotes the total number of data points.

The calculation formula is as follows:

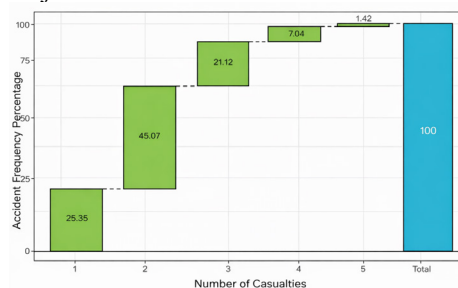
$$\text{Pseudo F-statistic} = \frac{\text{between-cluster variance}/(k - 1)}{\text{within-cluster variance}/(n - k)} \quad (2)$$

In this context,  $k$  represents the number of clusters, and  $n$  denotes the total number of data points.

## 3. Analysis of Driver Basic Information

### 3.1 The frequency distribution of accident casualties

Among the 2,823 traffic accident samples, 70 involved casualties. As shown in Figure 1, accidents with two casualties occurred most frequently, and major accidents accounted for 8.46% of the total.



**Figure 1. Frequency distribution of traffic accidents with different casualties.**

As shown in Figure 1, traffic accidents with two casualties have the highest frequency. When the number of casualties exceeds two, the frequency of accidents decreases progressively. Major accidents account for 8.46% of the total accidents.

### 3.2 Analysis of drivers' ages

Driver age was divided into nine groups at 5-year intervals, with the 18-20 group treated separately. The distribution of accident frequency and casualty frequency by age group is shown in Figure 2. Both indicators generally increased first and then decreased with age, although casualty frequency rose slightly again in the seventh and eighth age groups.

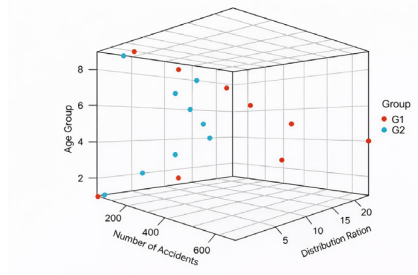


Figure 2. Frequency and casualty distribution of traffic accidents among drivers of different age groups.

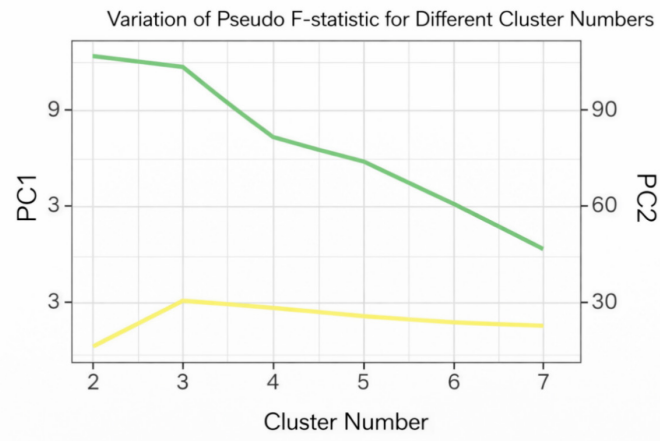
## 4. Analysis of Driver Age and Traffic Risk

### 4.1 The selection of the number of clusters

The Pseudo F-statistic was used to evaluate clustering performance for cluster numbers ranging from 2 to 7, with the proportions of traffic accidents and casualty accidents as principal components 1 and 2. As shown in Table 2 and Figure 3, Principal Component 1 decreased gradually as the number of clusters increased, whereas Principal Component 2 first increased and then decreased. When the number of clusters was 3, Principal Component 2 reached its maximum value of 18.594, while Principal Component 1 reached 10.412, close to its maximum value. Therefore, the optimal number of clusters was determined to be 3.

Table 2. ANOVA table with cluster numbers ranging from 2 to 7

| Cluster Number | Feature               | Clustering  |                   | Error       |                   | Pseudo F-Statistic |
|----------------|-----------------------|-------------|-------------------|-------------|-------------------|--------------------|
|                |                       | Mean square | Degree of freedom | Mean square | Degree of freedom |                    |
| 2              | Principal Component 1 | 130.681     | 1                 | 12.188      | 7                 | 10.722             |
|                | Principal Component 2 | 108.290     | 1                 | 25.351      | 7                 | 4.272              |
| 3              | Principal Component 1 | 83.841      | 2                 | 8.052       | 6                 | 10.412             |
|                | Principal Component 2 | 123.025     | 2                 | 6.616       | 6                 | 18.594             |
| 4              | Principal Component 1 | 59.762      | 3                 | 7.342       | 5                 | 8.140              |
|                | Principal Component 2 | 86.711      | 3                 | 5.123       | 5                 | 16.926             |
| 5              | Principal Component 1 | 46.541      | 4                 | 6.903       | 4                 | 7.305              |
|                | Principal Component 2 | 61.359      | 4                 | 3.969       | 4                 | 13.764             |
| 6              | Principal Component 1 | 29.356      | 5                 | 4.987       | 3                 | 5.862              |
|                | Principal Component 2 | 42.597      | 5                 | 2.157       | 3                 | 12.011             |
| 7              | Principal Component 1 | 24.362      | 6                 | 3.513       | 2                 | 4.339              |
|                | Principal Component 2 | 24.025      | 6                 | 1.804       | 2                 | 10.462             |



**Figure 3. Plot of the pseudo-F-statistic change of each component for different numbers of clusters.**

As shown in Table 2 and Figure 3, the Pseudo F-statistic for Principal Component 1 decreased gradually as the number of clusters increased, whereas that for Principal Component 2 first increased and then decreased. When the number of clusters was 3, Principal Component 2 reached its maximum value of 18.594, while Principal Component 1 reached 10.412, which was close to its maximum value. Considering both statistical performance and the interpretability of age-group classification, the optimal number of clusters was determined to be 3.

**4.2 Cluster results**

Based on the above analysis, the number of clusters was selected as 3, dividing the driver samples into three categories. The clustering results are shown in Table 3.

**Table 3. Driver age K-means clustering analysis results**

| Driver age group | Cluster number | Distance |
|------------------|----------------|----------|
| 1                | 1              | 1.99120  |
| 2                | 2              | 4.45084  |
| 3                | 3              | 3.96783  |
| 4                | 3              | 6.45856  |
| 5                | 3              | 2.50399  |
| 6                | 2              | 5.51767  |
| 7                | 2              | 3.02503  |
| 8                | 2              | 3.02503  |
| 9                | 1              | 1.99126  |

According to the clustering results, Group 3 includes Age Group 1 and Age Group 9; Group 2 includes Age Groups 2, 6, 7, and 8; and Group 1 includes Age Groups 3, 4, and 5. The number of samples corresponding to each cluster center is shown in Table 4.

**Table 4. The number of samples corresponding to each category for each clustering center**

| Group | The number of casualties $Z_0$ | Collision type $Z_1$ | Unsafe behavior value $Z_2$ | Unsafe state value $Z_3$ | Sample size |
|-------|--------------------------------|----------------------|-----------------------------|--------------------------|-------------|
| 1     | 0.123                          | -0.095               | -0.574                      | 0.162                    | 146         |
| 2     | -0.013                         | 0.531                | 0.169                       | -0.084                   | 1095        |
| 3     | 0.009                          | 0.942                | 0.816                       | -0.103                   | 1582        |

The K-means++ algorithm was used to divide the samples into three groups, and the probability density distribution curves for the three driver groups are shown in Figure 4.

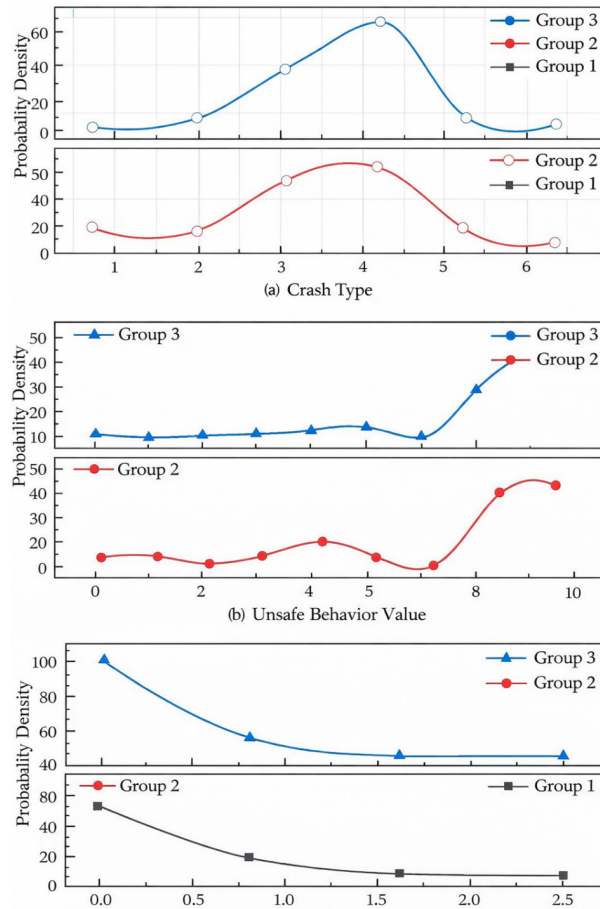


Figure 4. Probability density distribution.

### 4.3 Cluster analysis

As shown in Figure 4, the three groups exhibit distinct traffic risk characteristics. Group 1 has the highest values for  $Z_0$  and  $Z_3$  and the lowest value for  $Z_1$ . Group 2 has the lowest value for  $Z_0$ , while the other attributes remain at intermediate levels. Group 3 has the highest values for  $Z_1$  and  $Z_2$  and the lowest value for  $Z_3$ . In all three groups, unsafe driver behavior is concentrated in the operational error category.

**Analysis of Group 3:** Group 3 includes both young and elderly drivers and shows the widest age distribution. It is more likely to be involved in casualty accidents, mainly rear-end collisions caused by following operation errors. Vehicles in this group are also more likely to be in unsafe conditions. Young drivers may underestimate vehicle-related risks because of immature risk perception, whereas elderly drivers are more prone to delayed responses due to declines in sensory and reaction ability. Therefore, safe driving habits and appropriate behavioral adjustment should be emphasized for this group.

**Analysis of Group 2:** Group 2 mainly includes young and middle-aged drivers. It has the lowest casualty risk but is more prone to violations. Its main collision types are oblique and rear-end collisions, and the dominant unsafe behaviors are steering and following errors. This may reflect overconfidence in driving ability and insufficient legal awareness. Therefore, legal education and operational feedback should be strengthened for this group.

**Analysis of Group 1:** Group 1 consists of young and middle-aged drivers and has the largest number of accidents. Rear-end collisions account for the largest proportion, while frontal and side collisions are also relatively common.

Overall, operational errors, especially following errors, are prominent across the three groups and are closely related to rear-end collisions and casualty accidents. These risks are mainly associated with misjudgment of safe following distance and driver distraction. Therefore, both educational and technological measures should be adopted to reduce them.

## 5. Conclusion

Based on traffic accident data collected from a highway in Xianyang, Shaanxi Province, from 2021 to 2023, this study used the K-means++ algorithm to classify drivers into three groups according to traffic risk characteristics.

The results show that different driver groups exhibit different patterns of crash involvement and unsafe behavior. Rear-end collisions and following operation errors are the most prominent shared risk factors.

## References

- [1] Kong L-Z. Analysis of Human Factors in Causes of Traffic Accidents. *Chin. J. Saf. Sci.* 2013;23(1):28.
- [2] Bucsuházy K, Matuchová E, Zúvala R, et al. Human factors contributing to the road traffic accident occurrence. *Transp. Res. Procedia.* 2020;45:555-561.
- [3] Petridou E, Moustaki M. Human factors in the causation of road traffic crashes. *Eur. J. Epidemiol.* 2000;16:819-826.
- [4] Prasolenko O, Lobashov O, Galkin A. The human factor in road traffic city. *Int. J. Autom. Control Intell. Syst.* 2015;1(3):77-84.
- [5] Noy YI. Human factors in modern traffic systems. *Ergonomics.* 1997;40(10):1016-1024.
- [6] Zhang Y, Jing L, Sun C, et al. Human factors related to major road traffic accidents in China. *Traffic Inj. Prev.* 2019;20(8):796-800.
- [7] Fang Y-R. Experimental Study on Differences in Driving Behavior Characteristics of Different Categories of Drivers. *Saf. Environ. Eng.* 2020;27(5):204-208. doi:10.13578/j.cnki.issn.1671-1556.2020.05.030.
- [8] Cai X-Y, Tian X-Y, Weng J, et al. Visibility of Small Targets under Highway Tunnel Lighting Considering Driver Age. *Sci. Technol. Eng.* 2023;23(8):3396-3402.
- [9] Guan M-Q, Gong Q-N. Analysis of Accident Tendency of Engineering Vehicle Drivers Based on Grey Clustering Method. *Highway.* 2017;62(11):177-182.
- [10] Kim HS, Yoon DS, Shin HS, et al. Driving characteristics analysis of young and middle-aged drivers. In: *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC).* IEEE; 2016:864-867.
- [11] Scott-Parker B, De Regt T, Jones C, et al. The situation awareness of young drivers, middle-aged drivers, and older drivers: Same but different? *Case Stud. Transp. Policy.* 2020;8(1):206-214.
- [12] Ulak MB, Ozguven EE, Spainhour L. Age-based stratification of drivers to evaluate the effects of age on crash involvement. *Transp. Res. Procedia.* 2017;22:551-560.
- [13] Hu L, Bao X, Wu H, et al. A study on correlation of traffic accident tendency with driver characters using in-depth traffic accident data. *J. Adv. Transp.* 2020;2020:9084245.
- [14] Li H-R, Peng L-Q, Wu C-Z, et al. Traffic Accident Analysis on Long Downhill Sections for Heavy Freight Vehicles Based on Cluster Analysis. *Sci. Technol. Rev.* 2016;34(2):71-75.
- [15] Zhou Y-E, Gong H-F, Zhao C-X, et al. Speed-Density Model for Logistics Applicable in Mountainous Cities. *Sci. Technol. Eng.* 2021;21(4):1624-1628.
- [16] Liu S-X, Su D-L, Chi G-D, et al. Running Risk Assessment at Freeway Weaving Sections Based on Driving Simulation Experiments. *Sci. Technol. Eng.* 2021;21(2):751-757.