Data Science for Transportation (2025) 7:22
https://doi.org/10.1007/s42421-025-00135-3

RESEARCH

Novel Typology of Fatal Crash Behaviors via Natural Language Processing

Zhuo Han[1] · **Jimi Oke**[1]

Received: 7 May 2025 / Revised: 10 October 2025 / Accepted: 13 October 2025

Abstract
This study presents a data-driven framework to uncover patterns of less-studied fatal driver behaviors. We analyzed 84,170 driver behavior reports from the Fatality Analysis Reporting System (FARS) between 2021 and 2023. Using natural language processing methods, we obtained 194 distinct lemmas and quantified their importance in each report. We then applied principal components analysis to obtain 18 underlying behavioral factors. Finally, we clustered the reports via Gaussian Mixture Modeling, obtaining a novel typology of 15 behaviors. Our results reveal spatial disparities in fatal driver behaviors and provide a framework for type-specific crash mitigation efforts.

Keywords Driver behavior · Fatal crashes · Safety · Typology · NLP · PCA · GMM

Introduction

Road safety remains a critical global challenge, with approximately 1.19 million fatalities and tens of millions more injuries reported each year globally due to road traffic incidents (Global Status Report on Road Safety 2023). In the United States, 5.93 million police-reported motor vehicle crashes occurred in 2022, resulting in 1.66 million injury crashes and 39,221 fatal crashes. These incidents led to a total of 42,514 deaths and over 2.38 million injuries across all road users, including drivers, passengers, motorcyclists, pedestrians, and cyclists. Furthermore, the number of fatal crashes in the US increased by approximately 20.6% from 2015 to 2022 (2024 SafeTREC Traffic Safety Facts 2024). These statistics not only highlight the substantial human and economic costs imposed by road accidents but also emphasize the urgent need for more effective prevention measures.

Multiple interrelated factors contribute to fatal crashes. Environmental factors, including adverse weather conditions (for example, heavy rain, fog, or ice), poor road infrastructure, inadequate signage, and limited visibility, can significantly increase the risk of accidents by compromising the ability of drivers to safely navigate roads (Theofilatos and Yannis 2014). Vehicle-related issues, such as mechanical failures due to inadequate maintenance, defective brakes, tire blowouts, or malfunctioning safety equipment, also represent critical threats to roadway safety (van Schoor et al. 2001; Adanu et al. 2024). Although environmental and vehicle-related factors are crucial, driver behavior remains central, as these external conditions often exacerbate or interact closely with driver actions and decisions.

Recent research has shown that human error, such as distractions, fatigue, and impaired decision making, is a critical component in most crashes (Khattak et al. 2021). In fact, some studies indicate that driver misjudgment or inattention may account for 94% of road accidents (National Motor Vehicle Crash Causation Survey 2024). Given the overwhelming contribution of behavioral factors to road fatalities, it is urgent to accurately understand and interpret driver behaviors that lead to serious crashes. While prior work has often focused on predefined categories such as speeding or impairment (Adanu et al. 2025; Islam et al. 2024; Çelik and Oktay 2014; Høye 2020), relatively few studies have analyzed narrative crash reports to uncover and cluster the underlying behavioral patterns (Lee et al. 2023; Sayed et al. 2021).

This paper fills this gap by contributing to a deeper understanding of driver behavior in fatal crashes through

✉ Zhuo Han
zhuohan@umass.edu

Jimi Oke
jimi@umass.edu

[1] Department of Civil and Environmental Engineering, University of Massachusetts Amherst, 130 Natural Resources Road, 01003 Amherst, MA, USA

Published online: 23 October 2025

an unsupervised learning framework. In a novel effort, we apply natural language processing (NLP) techniques to extract and represent driver behavior narratives. We then employ clustering methods to identify distinct behavioral profiles, offering new insight into the types of driver errors most associated with fatal outcomes. Finally, we examine the spatial distribution of these behavior-based clusters to uncover regional trends and disparities.

Literature Review

Understanding the driver behavior relevant to fatal traffic crashes has been a focal point in transportation safety research. One of the most common behavioral factors contributing to fatal crashes is driver impairment due to alcohol and drugs (Romano and Voas 2011). Epidemiological and forensic studies have confirmed the prevalence of substance use in fatal crashes. For instance, autopsy-based investigations revealed that alcohol, narcotics (including cannabinoids and cocaine), and even mental health conditions or suicidal tendencies are disproportionately present in drivers involved in fatal incidents (Breen et al. 2018; Odoardi et al. 2023). Impaired drivers also tend to exhibit other high-risk behaviors, such as driving without a license or not wearing a seatbelt, indicating that substance use is often part of a broader pattern of risky driving (Høye 2020). Driver behavior analyses also show that impairment often overlaps with repeated violations, especially among male drivers involved in high-speed crashes (Ning et al. 2022).

Distracted and inattentive driving has emerged as a leading contributor to fatal crashes. Empirical analyses have revealed the role of distractions originating from both electronic device use and cognitive overload (Wundersitz 2019). Younger drivers are particularly susceptible to distraction induced by mobile devices, while older drivers experience more internal or cognitive distractions (Hossain et al. 2023). In-depth crash investigations found that distraction and inattention accounted for nearly a third of serious incidents, often without the involvement of other overtly risky behaviors such as speeding (Beanland et al. 2012).

Aggressive driving behaviors, including speeding and erratic maneuvers, also feature prominently in fatal crashes. Rather than occurring in isolation, these behaviors often appear alongside other violations such as unsafe lane changes, failure to obey, and disregard for traffic signals or stop signs. Aggressive violations have been shown to be among the most influential factors in fatal crashes, as demonstrated through a decision-analytic approach used to classify driver behavior by its impact on road safety (Farooq et al. 2019).

Table 1 The number of fatal crashes with reported driver behavior across different years

Year	Fatal crashes
2021	29 246
2022	28 645
2023	26 279
Total	84 170

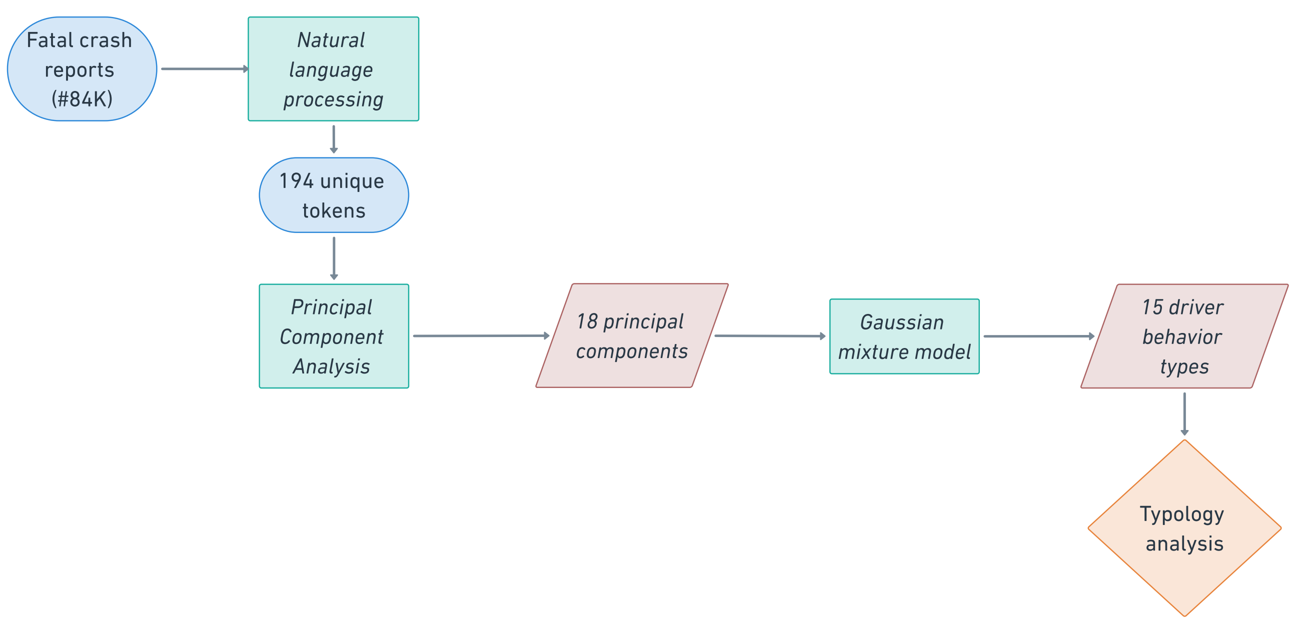


Fig. 1 Flowchart showing the research framework

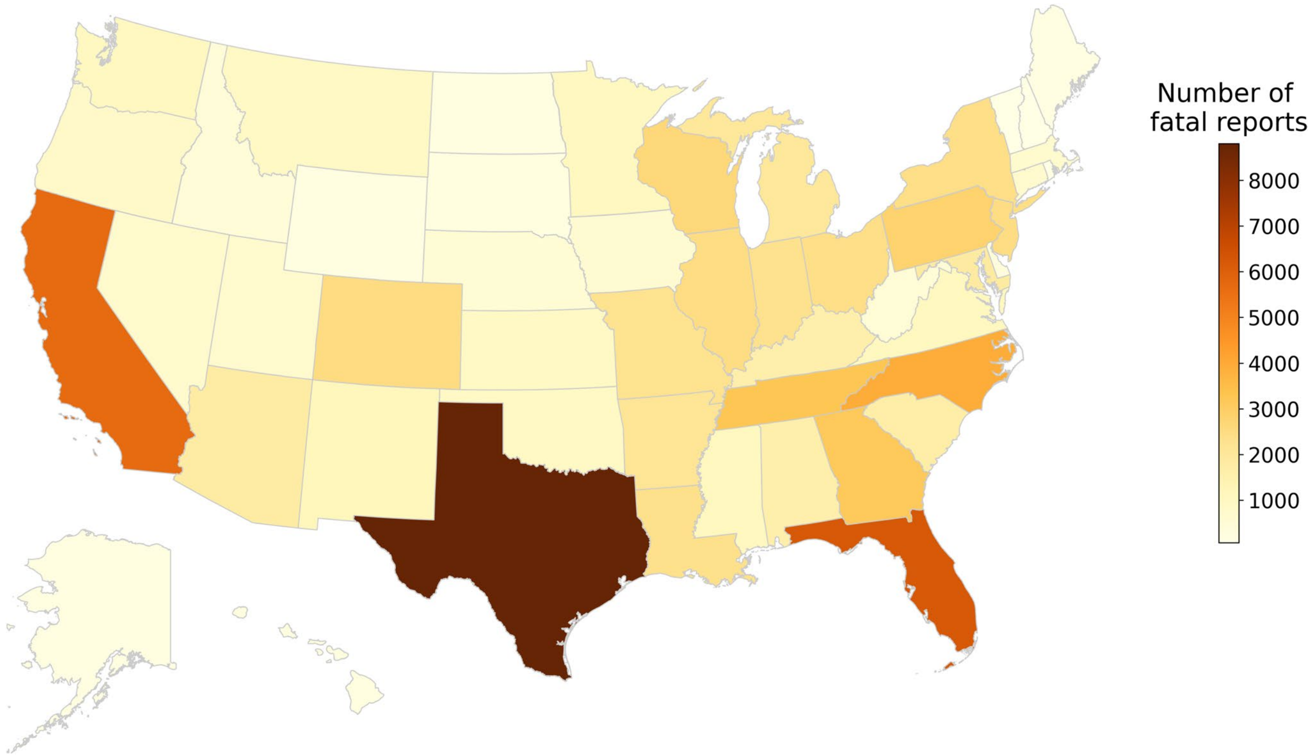


Fig. 2 Distribution of fatal crashes with reported driver behaviors across US states from 2021 to 2023

Fig. 3 Word cloud showing the vocabulary of reported driver behaviors in fatal crash reports, where word size is proportional to each term's frequency of occurrence

A wide range of techniques has been used to analyze driver behavior in fatal crashes. Statistical models, such as ordered probit and logistic regression, have long been used to examine the association between behavioral factors, such as non-compliance or distraction, and crash severity (Susilo et al. 2015; Osman et al. 2019). Naturalistic driving studies, which collect continuous in-vehicle data, have revealed that momentary distractions substantially increase the risk of crashes (Khan Rifat et al. 2024). Additional approaches include volatility-based metrics that capture erratic speed or acceleration fluctuations (Wali et al. 2020), and survey-based behavioral classification (Arafa et al. 2019).

Recent research has increasingly applied supervised and unsupervised machine learning (ML) methods to analyze driver behavior in relation to fatal crashes. Supervised approaches, including tree-based models, ensemble classifiers,

Table 2 Some examples of the driver behavior reports and the corresponding lemmas obtained

Raw Text	Lemmas obtained (separated by"I")
Driving on wrong side of two-way trafficway (intentional or unintentional)	Driving I wrong I side I twoway I trafficway I intention I unintention
Stopped in roadway (vehicle not abandoned)	Stop I roadway I vehicle I abandon
Driver has a driving record or driver's licens from more than one state	Driver I driving I record I driver I license I one I state
Overcorrecting	Overcorrecting
Failure to yield right-of-way	Failure I yield I rightofway

and deep learning architectures such as LSTM networks, have shown strong capabilities in identifying high-risk behaviors and predicting crash severity (Ghandour et al. 2020; Akin et al. 2022; Roussou et al. 2024). In the unsupervised domain, a notable study clustered fatal crashes from the Fatality Analysis Reporting System (FARS) using hierarchical and K-means methods, uncovering distinct spatio-temporal patterns based on roadway type (Li et al. 2021). Other efforts similarly applied clustering techniques to simulation and self-reported data to identify behaviorally distinct driver risk profiles (Zheng et al. 2023; Shirmohammadi et al. 2019), but these were based on limited sensor data. Recently, Natural Language Processing (NLP) (Paaß and Giesselbach 2023) has emerged as an effective computational method to extract insights from textual descriptions of events, conditions, and driver behaviors documented by law enforcement (Stanojević et al. 2013). Currently, NLP applications in driver behavior research focus primarily on classifying crash reports (Shao et al. 2024; Zou et al. 2024; Valcamonico et al. 2021). There is relatively limited exploration of NLP for driver behavior analysis.

More significantly, to date, no study has directly clustered behavioral patterns from free-text crash narratives to extract insights into the underlying behavior surrounding fatal crashes. Thus, this study aims to address this gap by using NLP methods combined with clustering algorithms to analyze textual crash reports, which offer the key advantage of revealing hidden structure in unlabeled data without relying on predefined classifications. Through this approach, we identify distinct clusters that represent patterns in fatal driver behaviors, providing deeper insight into the human factors that contribute to fatal crashes. Our clusters are derived solely from narrative text, without incorporating exogenous variables.

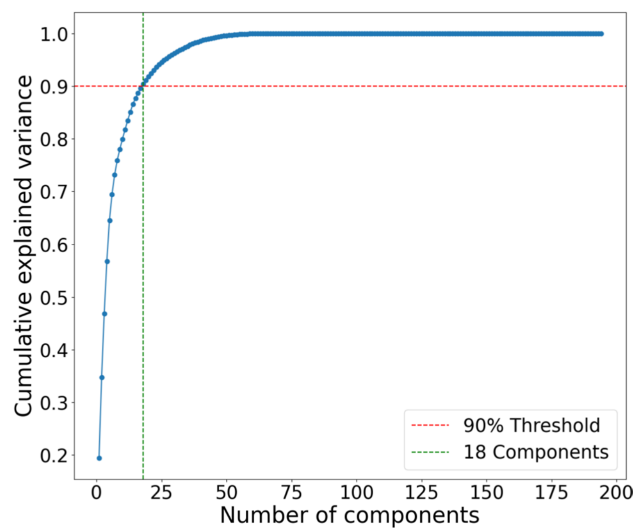


Fig. 4 Cumulative explained variance as a function of the number of PCA components. The horizontal red dashed line indicates the 90% variance threshold, and the vertical green dashed line marks the respective number of components

Methods

Research Framework

Figure 1 illustrates the framework we used in this study. First, we collected raw driver behavior data from the Fatality Analysis Reporting System (FARS) database. We preprocessed the data to clean and standardize the textual narratives through tokenization, stop-word removal, and lemmatization. The preprocessed data are then transformed into a weighted representation using the TF–IDF (inverse document frequency term frequency) approach (SPARCK JONES 1972), effectively converting textual information into a numerical matrix. After weighting, we applied Principal Component Analysis (PCA) to reduce dimensionality and interpret the underlying factors, followed by Gaussian Mixture Modeling (GMM) to cluster the documents and derive a typology of fatal crash behaviors. Finally, we also provided detailed interpretations of how the types distributed across the nation and regions.

Data

We analyzed driver behaviors that contribute to fatal crashes using data from the Fatality Analysis Reporting System (FARS) for the years 2021 to 2023. FARS is a national database maintained by the National Highway Traffic Safety Administration (NHTSA) that records all fatal motor vehicle crashes in the United States. In this database, the driver behavior associated with each fatal crash is reported as a free-text narrative.

The annual distribution of reported driver behaviors in fatal crash reports is summarized in Table 1. The extracted database comprises 199,465 fatal crash records that span 2021 through 2023. However, of these, only 84,170 contained narrative reports of driver behavior and thus served as the observations that we analyzed in this study. Figure 2 illustrates the state-wide distribution of the reports in our final dataset. Figure 3 presents a word cloud of the words found in all the reports analyzed. In addition to this narrative field, the database includes only basic identifiers, such as the crash ID and the state in which the incident occurred. In addition, we collected an urban–rural indicator for each case to support spatial analysis.

Natural Language Processing

Initially, all text was converted to lowercase to ensure consistency. Punctuation and non-alphanumeric characters were removed, preserving only letters and numbers. The text was then tokenized by splitting into whitespace to generate individual tokens (words). To focus on meaningful content and reduce redundancy, common English stopwords (e.g., and, by) and punctuation were removed. Subsequently, each token was "lemmatized", transforming words into their base forms (e.g., "driving" → "drive"). These preprocessing steps were implemented using the Natural Language Toolkit (NLTK) in Python [39]. After preprocessing, a total of $M = 194$ distinct lemmas were obtained, Table 2 presents examples comparing raw driver behavior text before and the corresponding lemmas obtained after preprocessing.

We used the term frequency–inverse document frequency (TF–IDF) method to quantify the importance of different behavioral descriptions. Let the set of N driver behavior records (documents) be represented as $\boldsymbol{\mathcal{D}} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$, and let $\boldsymbol{\Lambda} = \{\lambda_m\}_{m=1:M}$ denote the lexicon of M unique lemmas (terms) extracted from $\boldsymbol{\mathcal{D}}$.

The transformation from raw textual data to this matrix representation proceeds as follows: For each lemma $\lambda_m \in \boldsymbol{\Lambda}$ and for each document $\mathcal{D}_n \in \boldsymbol{\mathcal{D}}$, we compute the term frequency $\mathrm{tf}(\mathcal{D}_n, \lambda_m)$ as the number of occurrences of λ_m in $\mathcal{D}_n$.

To evaluate the importance of each lemma across the entire collection of documents, we calculated the inverse document frequency:

$$\mathrm{idf}(\lambda_m) = \log\left(\frac{N}{N_{\lambda_m}}\right), \tag{1}$$

where N is the total number of documents in $\boldsymbol{\mathcal{D}}$ and N_{λ_m} is the number of documents in which lemma λ_m appears.

The numeric TF–IDF weight for lemma λ_m in document $\mathcal{D}_n$ is then given by:

$$w_{nm} = \mathrm{tf}(\mathcal{D}_n, \lambda_m) \cdot \mathrm{idf}(\lambda_m). \tag{2}$$

This formula emphasizes lemmas that are particularly informative in a specific document while down-weighting those that are common across the entire set of records. Finally, we assemble the TF–IDF weights in a matrix $\mathbf{W} \in \mathbb{R}^{N \times M}$, with $N = 84,170$ and $M = 194$.

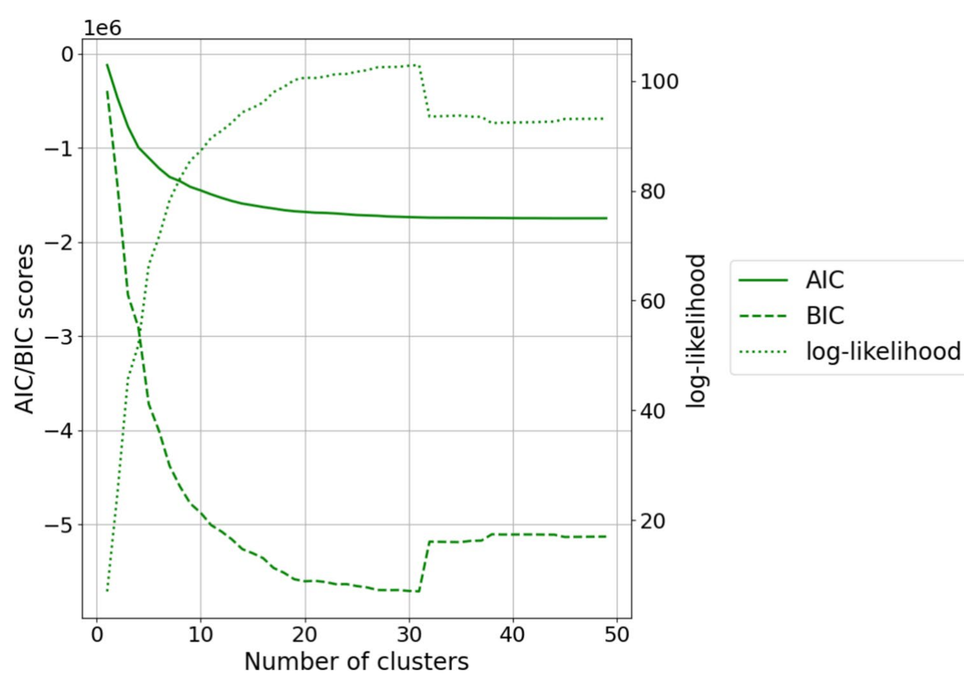


Fig. 5 Comparison of AIC, BIC, and log-likelihood values across varying numbers of GMM clusters

Dimensionality Reduction

After constructing the TF–IDF matrix $\mathbf{W}$, we reduced the dimensionality of the database to facilitate efficient processing and interpretation of the driver behavior data. We used principal component analysis (PCA), which projects the high-dimensional TF–IDF features onto a lower-dimensional subspace that retains most of the data variance. We computed the covariance matrix $\boldsymbol{C}$ from the TF-IDF matrix, which captures the pairwise relationships among lemmas throughout the database. Then we computed the eigenvalues and eigenvectors of $\boldsymbol{C}$. To reduce dimensionality, we selected the top L principal components based on a cumulative explained variance threshold, as shown in Fig. 4. The original TF–IDF matrix $\mathbf{W} \in \mathbb{R}^{N\times M}$ is then projected onto the resulting L-dimensional subspace as:

$$\mathbf{Z} = \mathbf{WV}, \tag{3}$$

where $\mathbf{V} \in \mathbb{R}^{M\times L}$ is the truncated matrix containing the top L eigenvectors. In the PCA-transformed matrix. $\mathbf{Z} \in \mathbb{R}^{N\times L}$, each row vector $z_n \in \mathbb{R}^K$ represents the driver behavior for document $\mathcal{D}_n$. We finally extracted $L = 18$ principal components from 194 unique word lemmas derived from fatal crash reports for each driver.

Clustering

We used a Gaussian Mixture Model (GMM) to cluster the records based on the PCA scores in $\mathbf{Z}$. (Fig. 14 shows the distribution of scores in each component.) GMM assumes that each observation is generated from a mixture of K multivariate Gaussian distributions. The parameter set of the GMM is given by:

$$\boldsymbol{\Theta} = \left\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right\}_{k=1}^{K}, \tag{4}$$

where π_k is the mixing coefficient with $\sum_{k=1}^{K} \pi_k = 1$, $\boldsymbol{\mu}_k$ is the mean vector, and $\boldsymbol{\Sigma}_k$ is the covariance matrix of the kth multivariate Gaussian distribution. The probability density function for z_n is then given by:

$$p(z_n \mid \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{5}$$

The model parameters ($\boldsymbol{\Theta}$) are estimated using the Expectation–Maximization (EM) algorithm. In the E-step, we compute the posterior probability (or "responsibility") that each data point z_n was generated by the kth Gaussian:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(z_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(z_n \mid \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}. \tag{6}$$

In the M-step, we update the parameters $\boldsymbol{\Theta}$ using the responsibilities:

$$\pi_k = \frac{1}{N}\sum_{n=1}^{N} \gamma_{nk}, \tag{7}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} z_n}{\sum_{n=1}^{N} \gamma_{nk}}, \tag{8}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma_{nk}(z_n - \boldsymbol{\mu}_k)(z_n - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^{N} \gamma_{nk}}. \tag{9}$$

The above steps are repeated until convergence, determined by a negligible change in the log-likelihood or the parameters.

Cluster Assignment and Model Selection

After convergence, each observation is assigned to the cluster with the highest responsibility:

$$\hat{k}_n = \arg\max_{k_n} \gamma_{nk}. \tag{10}$$

To determine the optimal number of clusters K^*, we considered parsimony, fitness and intrepretability. To quantify performance, we employed the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which balance the fit of the model against complexity. For a K-component GMM, the AIC is defined as:

$$\mathrm{AIC}_K = 2p_K - 2\log \mathcal{L}_K^*, \tag{11}$$

and the BIC is defined as:

$$\mathrm{BIC}_K = p_K \log N - 2\log \mathcal{L}_K^*, \tag{12}$$

where $\mathcal{L}_K^*$ is the maximized log-likelihood of the K-component model, N is the number of observations and p_K is the number of parameters, given by:

$$p_K = (K-1) + (K \cdot N) + K \cdot \frac{N(N+1)}{2}, \tag{13}$$

In the above equation, $K - 1$ parameters account for the mixing coefficients; $K \cdot N$ corresponds to the means of each Gaussian component; and, lastly, $K \cdot \frac{N(N+1)}{2}$ parameters represent the elements of the full covariance matrices.

To select the optimal number of clusters, we computed the AIC and BIC scores for K ranging from 1 to 50 (Fig. 5). Prioritizing parsimony and interpretability, we chose $K^* = 15$ as the first reasonably small K where AIC/BIC reach local minima and the log-likelihood shows a corresponding local maximum.

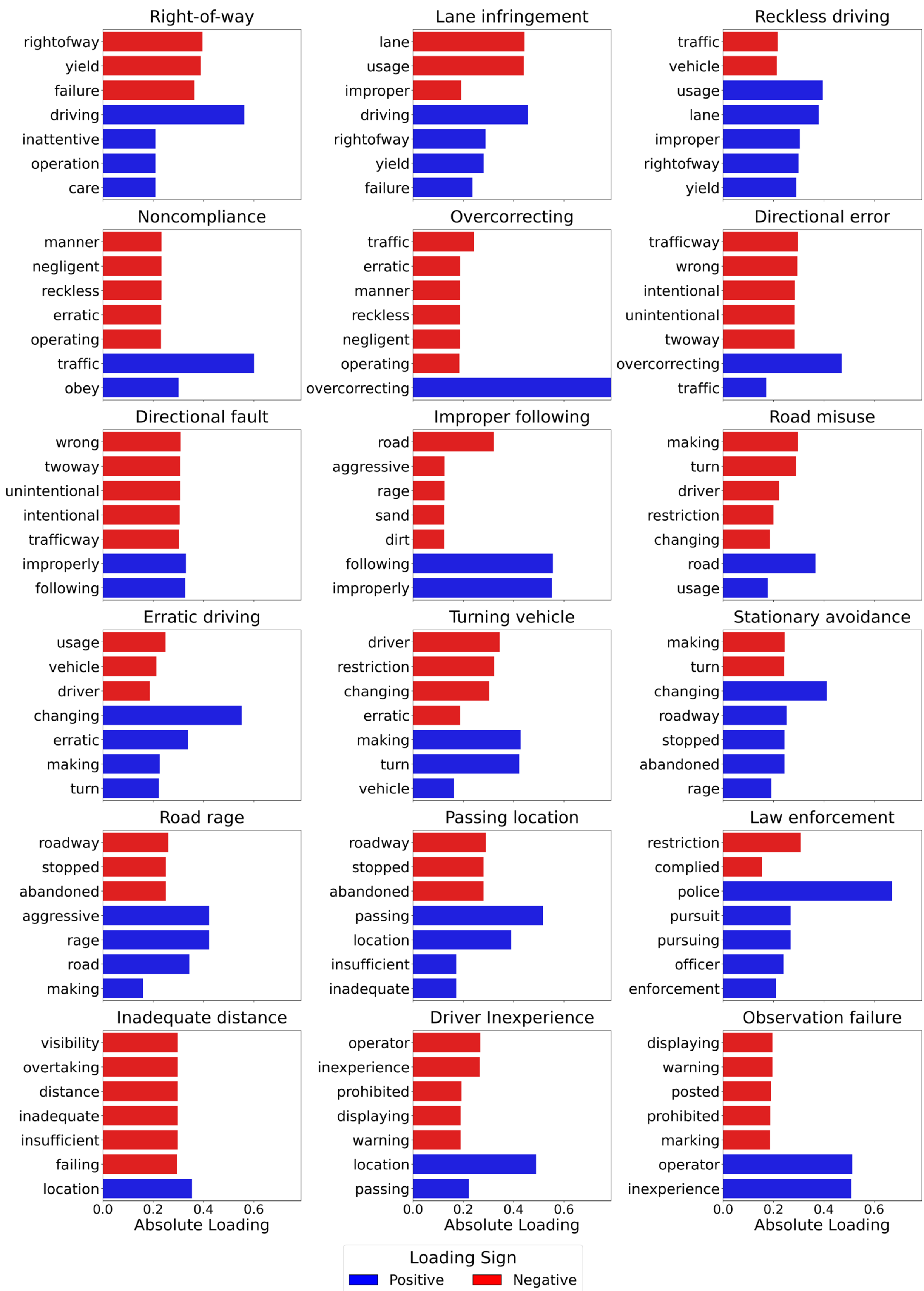


Fig. 6 Bar plots of the top 7 lemma loadings on each of the 18 extracted principal components

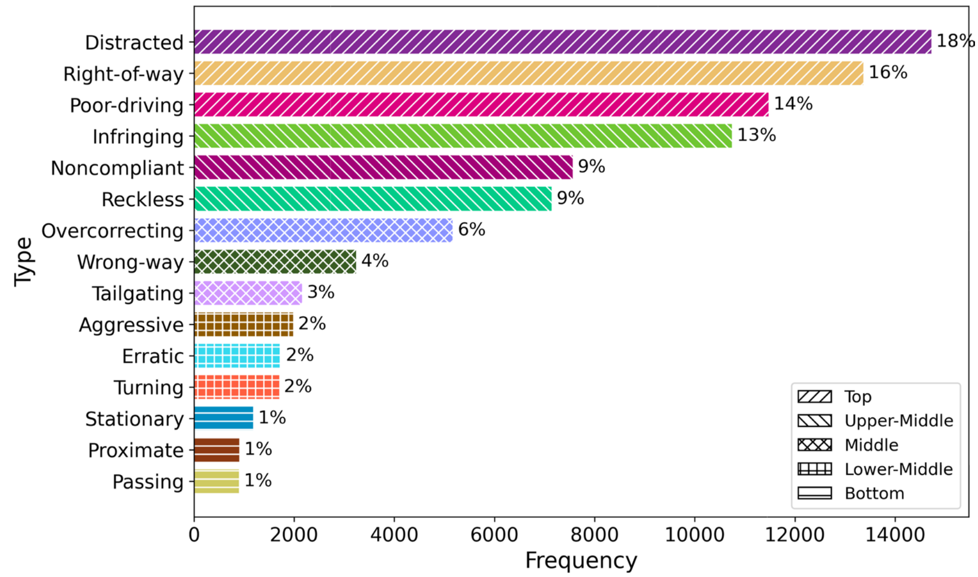


Fig. 7 Horizontal bar plot showing the frequency of GMM driver behavior types, with hatch patterns indicating their assigned frequency groups. These groups are derived from 5 equal-probability bins based on the overall distribution of behavior frequencies

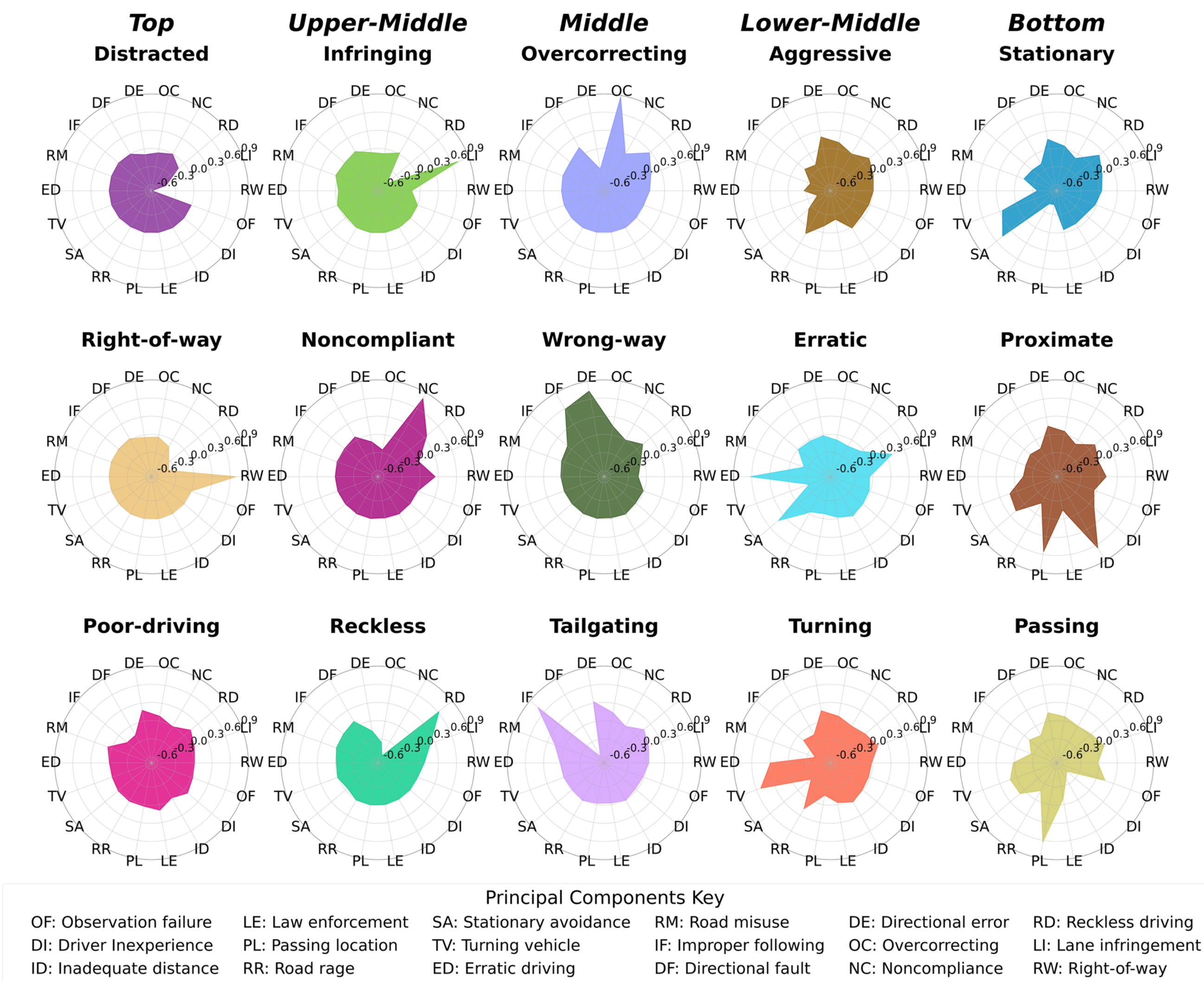


Fig. 8 Spider plots of the average component value patterns in each type

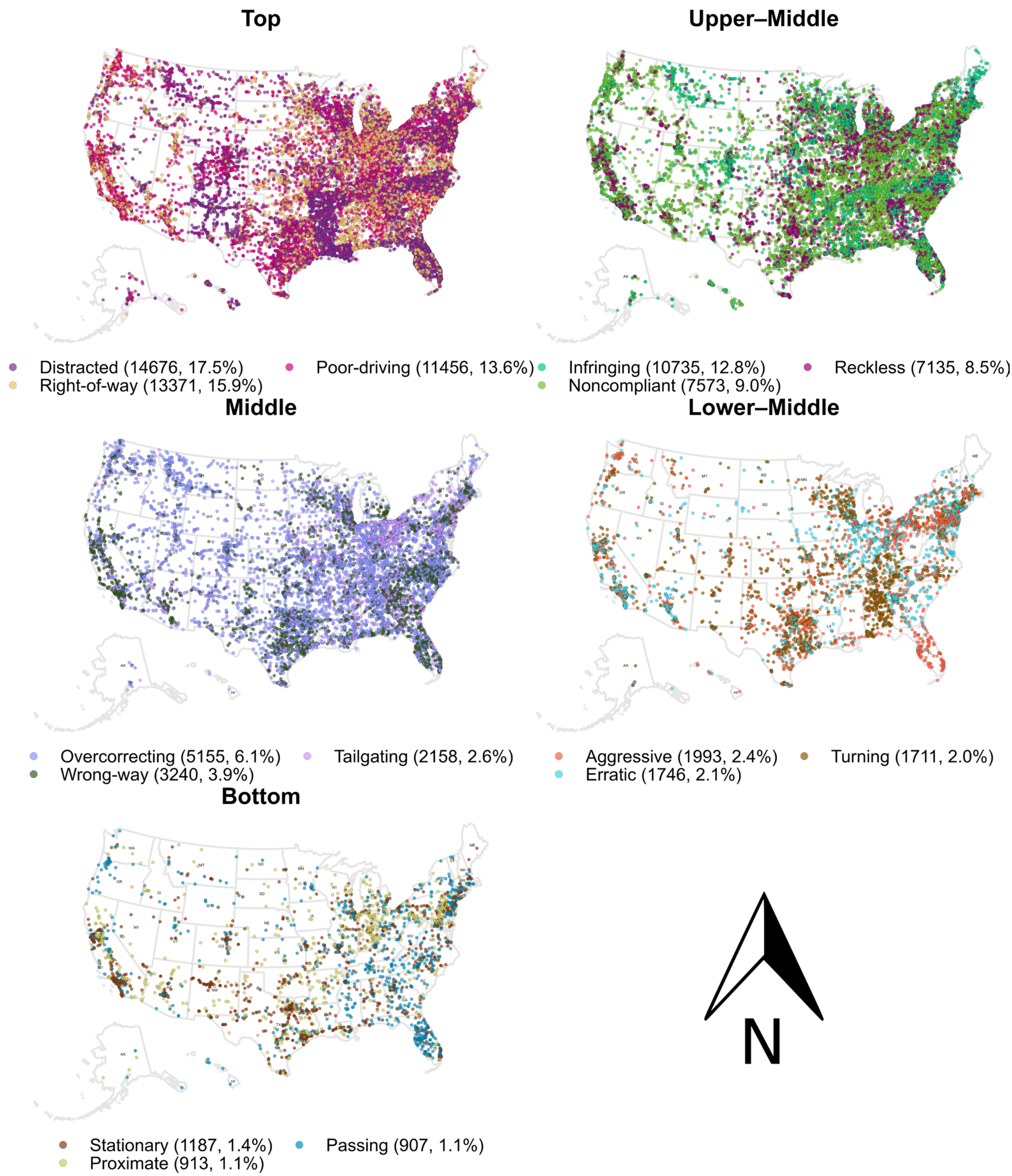


Fig. 9 Geographic distribution of driver?behavior types across the US, divided into five equal?probability bins

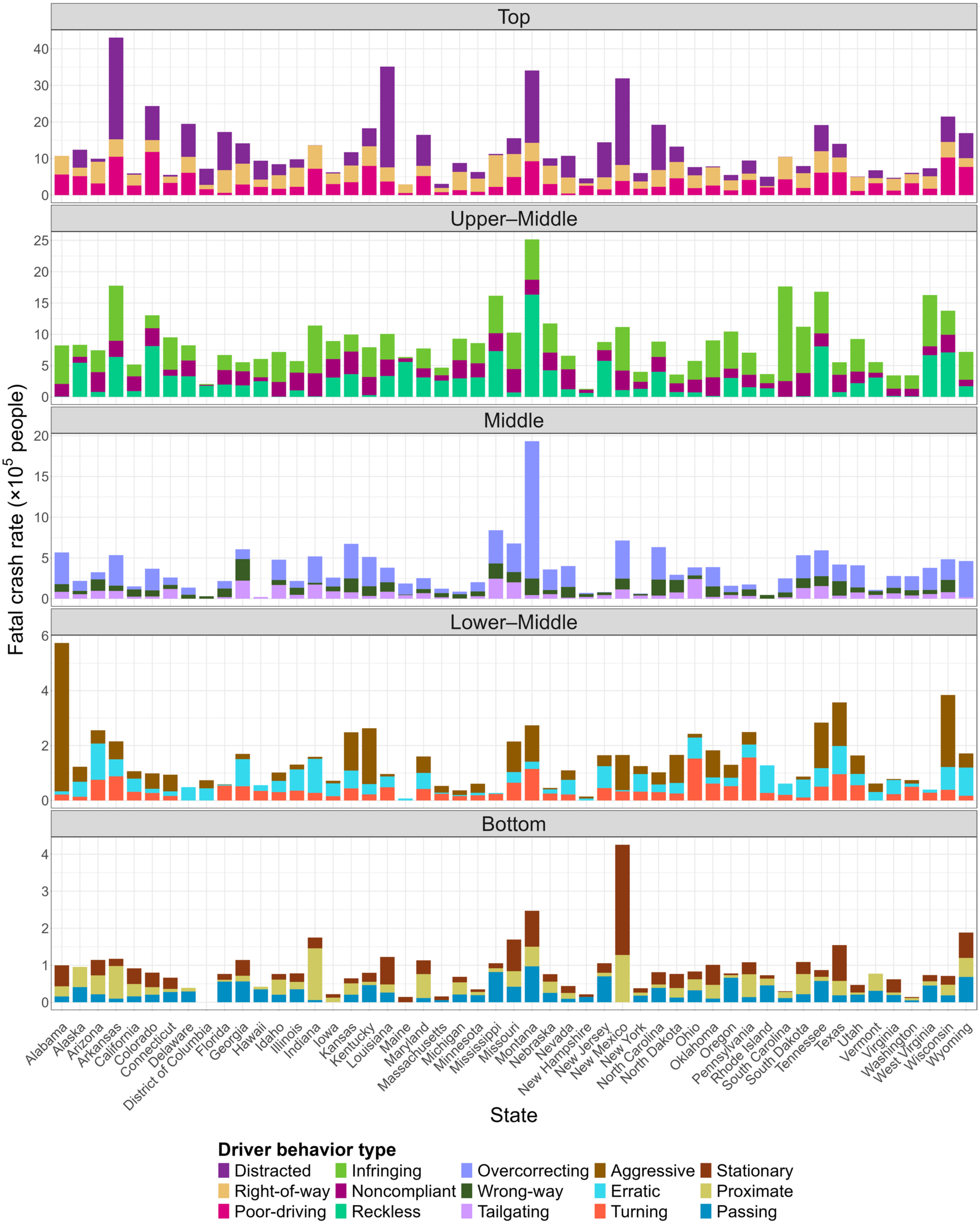


Fig. 10 State-level fatal crash occurrences per 100K population

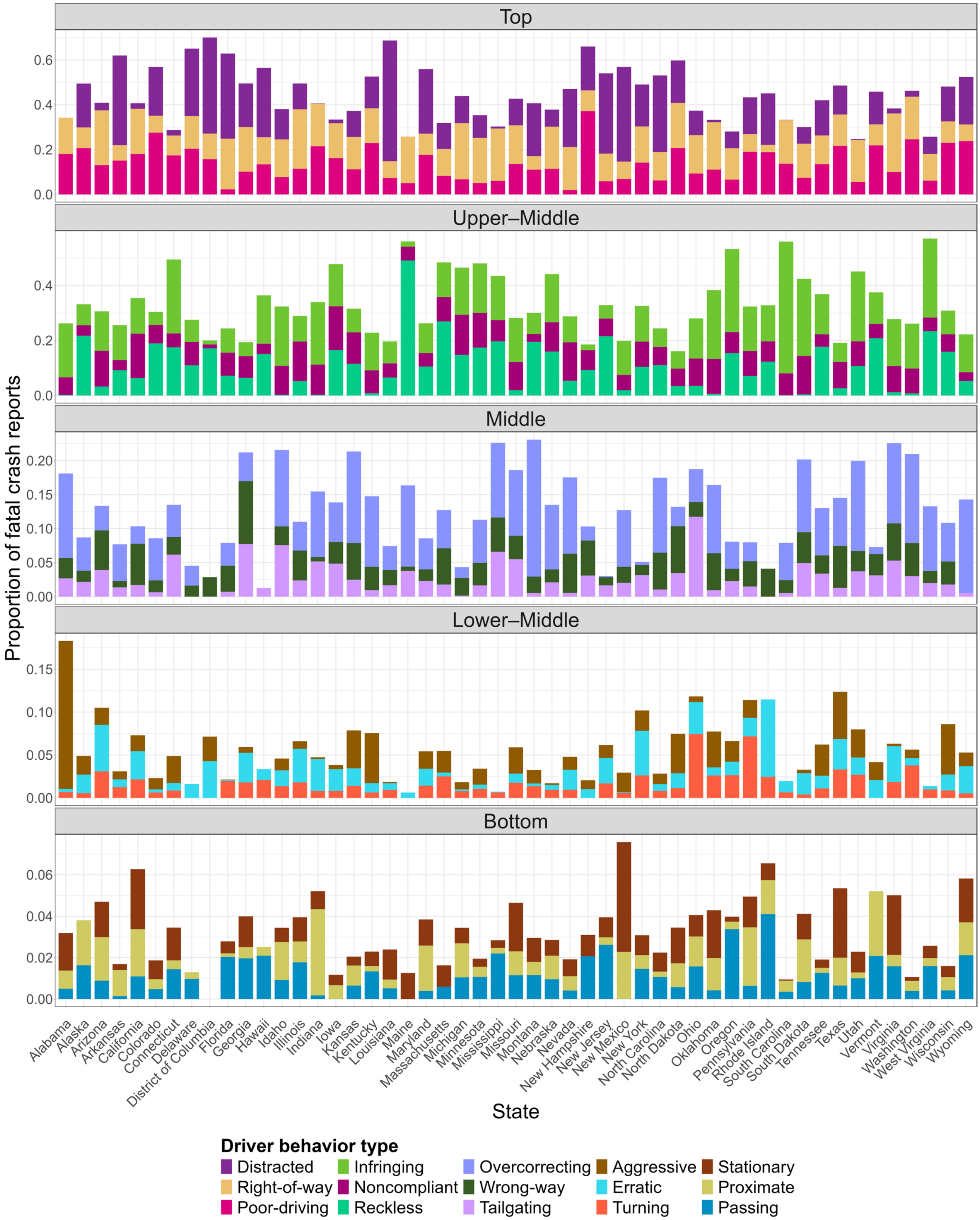


Fig. 11 State-wide fatal crash report proportion across US states categorized by driver behavior types

Results

18 Underlying Fatal Driving Behavioral Components

We extracted 18 components, which represent the main underlying factors of driver behavior in fatal crashes (beyond impairment and speed). Based on the lemma contributions, we interpreted and named each component (Fig. 6).

Right-of-way, *Lane infringement*, and *Noncompliance* reflect common rule violations, with lemma associations such as "yield," "rightofway," and "obey." Components such as *Overcorrecting*, *Erratic Driving*, and *Turning Vehicle* reflect unstable or abrupt vehicle maneuvers. Meanwhile, *Improper Following*, *Passing location*, and *Inadequate Distance* pertain to spacing and tailgating behaviors. Aggressive attitudes on the road are reflected in the *Road Rage* component, marked by lemmas, such as "rage," "aggressive," and "improper."

Several principal components reflect specific directional or positional errors, including *Directional Fault*, *Directional Error*, and *Road Misuse*, characterized by terms such as "twoway," "intentional," and "restriction." Others, such as *Stationary Avoidance* and *Observation Failure*, suggest difficulties in recognizing or responding to stationary obstacles and environmental cues, respectively.

Enforcement-related interactions are captured in *Law Enforcement*, which emphasizes terms such as "police," "officer," and "restriction." Lastly, *Driver Inexperience* highlights issues related to novice drivers, marked by lemmas such as "inexperience," "operator," and "prohibited."

15 Fatal Behavior Types

We identified 15 distinct clusters of driver behavior (hereafter referred to as behavior types) associated with fatal crashes. We further grouped these into five categories based on their frequency: "Top", "Upper-Middle", "Middle", "Lower-Middle" and "Bottom" as shown in Fig. 7. To provide deeper insights into these groups, we explored the relationship between GMM types and their underlying components shown in Fig. 8. The types maps grouped by category are shown in Fig. 9. Furthermore, we plot the distribution of fatal crash rates per 100,000 residents (per 100K) by state in Fig. 10, as well as the proportion of each type by state in Fig. 11.

Top

Driver behavior types in the "Top" group, *Distracted*, *Right-of-way*, and *Poor-driving* dominate fatal crash statistics. *Distracted* has the highest fatal crash rate impact observed in Arkansas at 27.8 per 100K and a total of 853 fatal crashes as shown in Fig. 10. This behavior is associated with multitasking, mobile device usage, and attention lapses as shown in Fig. 8. *Right-of-way* peaks in Mississippi at 8.7 per 100K. This behavior typically involves failure to yield at intersections, stop signs, or when merging, resulting in angular collisions and multi-vehicle impacts. *Poor-driving* encompasses a variety of low-skilled or careless behaviors in multiple aspects. Figure 8 shows that *Poor-driving* is moderately influenced by a wide range of factors–including *Directional Fault*, *Road Rage*,which suggests that *Poor-driving* reflects broad deficiencies in vehicle control, hazard anticipation, and real-time decision making. The elevated rate in Colorado is 11.8 per 100K and may be related to variable weather conditions and geological factors.

Upper-Middle

The "Upper-Middle" group comprises *Infringing*, *Noncompliant* and *Reckless*. *Infringing* has the highest fatal crash rates per 100K reported in South Carolina (15.1), Arkansas (8.8) and West Virginia (8.2) as shown in Fig. 10. This type of behavior includes actions such as drifting across lanes, unsafe merges, and failure to maintain lateral control, as shown in Fig. 10. *Infringing* is strongly associated with the factors related to *Lane Infringement* as shown in Fig. 8, suggesting a pattern of poor spatial awareness, possibly worsened by narrow roads, inconsistent lane markings, or driver inattention. *Noncompliant* behavior reflects disregard for traffic regulations, such as failure to obey signals, signs or right-of-way rules. It is particularly elevated in South Dakota (3.7 per 100K) and Missouri (3.7 per 100K), where broad and sparsely monitored road networks can limit enforcement visibility. Figure 8 shows this behavior type is driven by high contributions from *Noncompliance* and *Right-of-way* components, signaling a combination of intentional rule breaking and deficient vehicle positioning. *Reckless* is characterized by aggressive or impulsive actions, often at high speeds or under uncertain conditions, and is defined by high loadings in *Reckless Driving* and *Directional Fault* (Fig. 8), indicating unsafe maneuvering combined with poor situational judgment.

Middle

The behavior types in this group—*Overcorrecting*, *Wrong-way*, and *Tailgating*—reflect specific failures in maneuver control, spatial orientation, and following behavior. *Overcorrecting* is the leading type in this group, and is prominent in Montana, with a fatal crash rate of 16.9 per 100K as shown in Fig. 10. As shown in Fig. 10, the *Overcorrecting* is characterized by strong positive loadings on the principal components *Overcorrecting* and *Road Rage*, indicating that

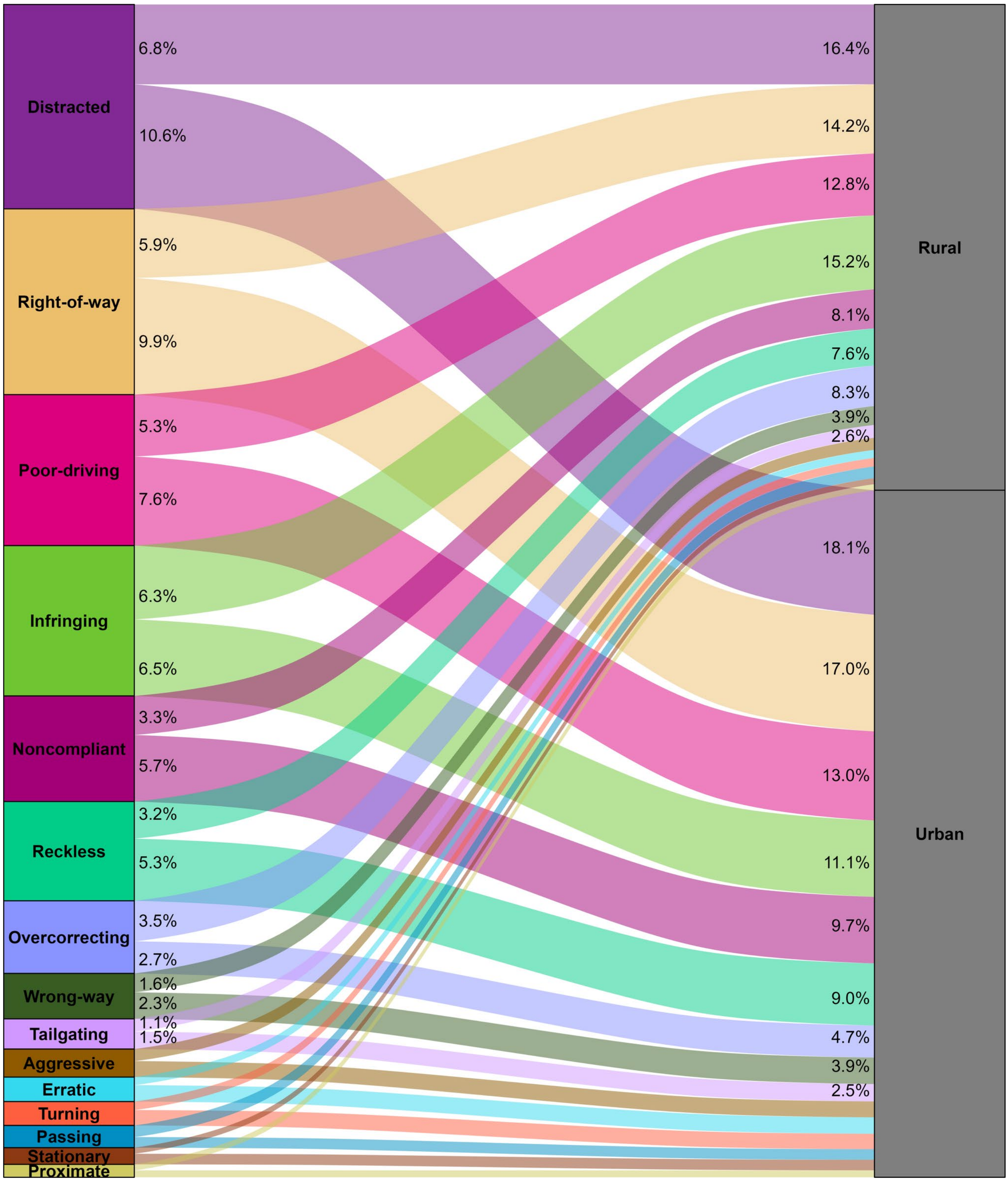


Fig. 12 Contributing driver behaviors in fatal crashes across urban and rural areas. Percentages on the left indicate absolute shares of each type based on all the observed crash reports. Percentages on the right indicate the relative proportion of each type in urban or rural locations. For clarity, annotations are not provided for the 6 least frequently occurring types

this group type reflects abrupt, emotionally driven steering responses often triggered under stress or sudden road departures. *Wrong-way* shows the highest fatal crash rate, with Georgia reporting 2.7 per 100K, accounting for 4% of all fatal crashes. This behavior is typically the result of navigational failures on highway ramps or intersections. According to Fig. 8, the *Wrong-way* type is strongly associated with *Directional Error* and *Directional Failure* components, suggesting that these incidents often stem from confusion in navigation or poor perception of the direction of traffic. *Tailgating* is common in Mississippi with the highest crash rate at 2.5 per 100K, reflecting a failure to maintain a safe following distance and often occurring under congested conditions or sudden stops. Figure 8 indicates that the *Tailgating* type is primarily influenced by *Improper Following* and *Directional Error*, pointing the driver behavior rooted in unsafe following distance and misjudgment of spatial alignment.

Lower-Middle

Within the Lower-Middle group, fatal crash behavior types (*Aggressive*, *Erratic*, and *Turning*) are shaped primarily by emotionally charged and maneuver-specific actions, which reflect situational stress, impulsive reactions, and poor execution at critical decision points such as intersections. *Aggressive* accounts for a 17.3% of all fatal crash behaviors in Alabama, 5.9% for Wisconsin and 5.8% for Kentucky as shown in Fig. 11, suggesting a pattern of road rage, speeding, tailgating, and hostile maneuvers. These behaviors are often intensified by traffic frustration, poor lane discipline, or perceived provocation (Okafor et al. 2022). The *Erratic* type is highly associated with *Erratic Driving*, *Lane Infringement*, and *Stationary Avoidance* shown in Fig. 8, revealing a behavioral pattern characterized by inconsistent vehicle control, poor lateral stability, and difficulty responding to stopped or disabled vehicles. It peaks notably in Arizona at 1.3 per 100K (Fig. 10). *Turning* reaches its highest rate in Pennsylvania at 1.6 per 100K (38 crashes), involves errors during turning maneuvers, especially at intersections.

Bottom

The Bottom category features the three least frequently occurring types. *Passing* accounts for 2% of all fatal crashes throughout the country, peaking in Montana at 1 per 100K (Fig. 10), reflecting unsafe overtaking behaviors possibly due to extensive rural two-lane road networks. Similarly, *Stationary* vehicle-related incidents peak in New Mexico at 3 per 100K, emphasizing vulnerabilities in managing roadside emergencies in sparsely populated areas. Indiana prominently experiences *Proximate* incidents (inadequate following distances) with a rate of 1.4 per 100K, possibly indicating gaps in driver experience or attentiveness.

Discussion

Infringing Is Prevalent in Northeastern and Mid-western States

Infringing stands out as a dominant contributor to fatal crashes, particularly across the Northeast and Midwest. As shown in Fig. 11, *Infringing* behaviors account for more than 25% of all fatal crash cases in states such as South Carolina, Oregon, and West Virginia. Figure 9 shows that *Infringing* crashes are more frequent in states such as Tennessee, Missouri, and Pennsylvania, where fatal incidents often arise from unsafe merging, lane drifting, and boundary violations (Fiolić et al. 2023). These behaviors are often symptomatic of limited lane discipline, complex merging environments, or inconsistent road markings (Kuo and Lord 2021).

Maneuvering Errors Widely Distributed in Areas with Limited Infrastructure

More maneuver-specific and reactive behaviors, especially *Overcorrecting*, *Wrong-way*, and *Tailgating*. On average, *Overcorrecting* accounts for 6.4% of fatal crashes in this group, with *Wrong-way* at 4.2% and *Tailgating* at 3.0%, as shown in Fig. 11. Figure 9 further reveals that these behaviors are spatially concentrated in states such as Kentucky, Indiana, and Tennessee, which have limited infrastructure (State Infrastructure Facts 2016; Penmetsa et al. 2018). The geographical distribution suggests regional exposure to conditions that increase the likelihood of reactive errors, such as limited shoulder space, rural roadway curvature, or inadequate signage at decision points [45].

Distracted Driving Dominates in Urban Areas, While Overcorrecting is a Bigger Issue in Rural Areas

Distracted driving is the most prominent behavior overall, comprising 6.8% of all fatal crashes in rural settings and 10.6% in urban areas. It is even more dominant when considering the share of distracted crashes in rural or urban areas [46]: *Distracted* accounts for 16.4% of all rural crashes and a staggering 18.1% of urban crashes. These figures suggest that distraction, often from mobile phones, in-vehicle systems, or cognitive overload, is a leading contributor to fatalities in both areas, but is especially pronounced in urban

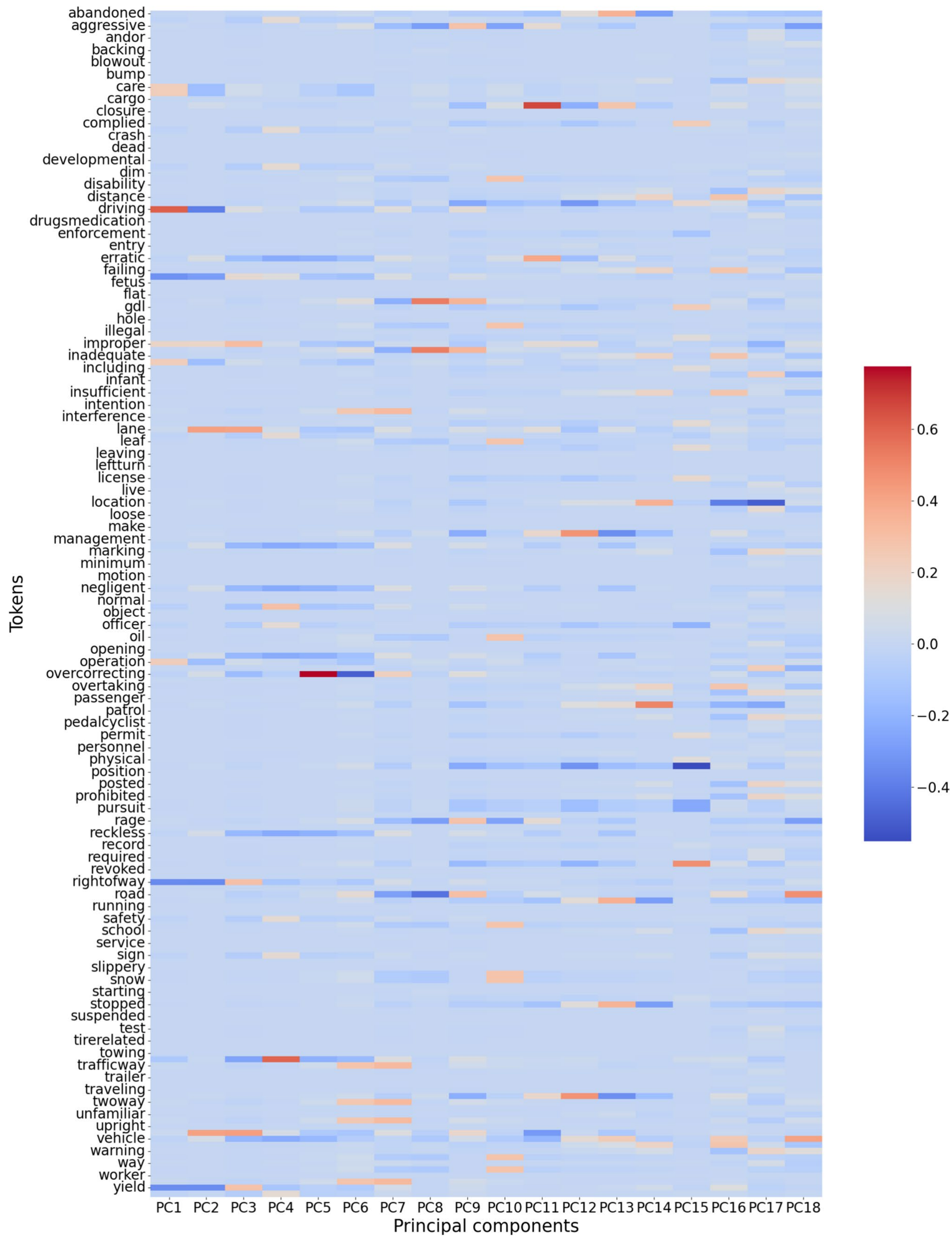
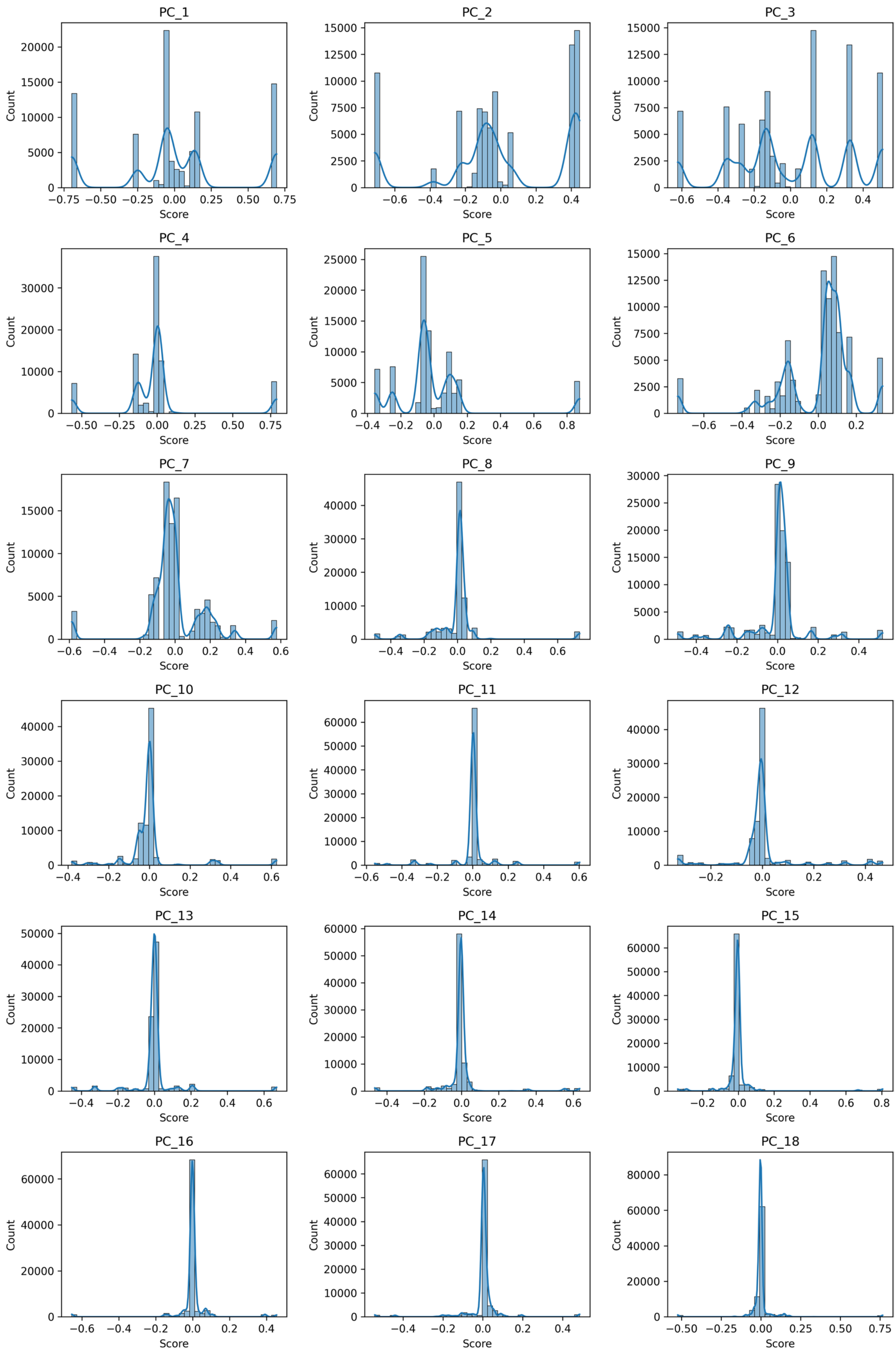


Fig. 13 Heatmap of token loadings across principal components


PC_1
PC_2
PC_3
PC_4
PC_5
PC_6
PC_7
PC_8
PC_9
PC_10
PC_11
PC_12
PC_13
PC_14
PC_15
PC_16
PC_17
PC_18
Count
Score

◂**Fig. 14** Distribution of principal component scores derived from the TF-IDF matrix of driver behavior narratives. Each subplot represents the distribution of one PC, showing distinct multimodal patterns across components

traffic environments. In contrast, certain behaviors such as *overcorrecting* occur predominantly in rural areas (8,365 incidents in rural areas versus 6,618 incidents in urban areas) (Khan et al. 2023), reflecting the driving environments and potential speed variations commonly found on rural roads.

Aggressive Driving Has Greater Impact in Rural Areas Despite Higher Urban Frequency

Aggressive driving contributes 1% to rural deaths and 1.4% to urban deaths. However, it represents a slightly larger share of total crashes within rural environments (2.5%) compared to urban areas (2.3%) as shown in Fig. 12. This pattern suggests that although absolute numbers are higher in cities, the behavioral intensity or deadliness of aggressive driving (e.g., speeding, tailgating, road rage) may be proportionally greater in rural contexts—likely due to higher speed limits, reduced enforcement, and fewer conflict-avoidance opportunities (Fatality Facts 2023 2023).

Erratic Driving Occurs Mainly in Rural Areas with Low Frequency

Erratic driving has a lower overall presence, but is more prominent in rural settings compared to its total. It represents 0.7% of all rural fatal crashes versus 0.8% in urban areas, yet accounts for 1.7% of crashes within rural settings (compared to 1.4% in urban) as shown in Fig. 12. This indicates that erratic behaviors, such as sudden swerving, inconsistent speed, or failure to react predictably, are especially risky on rural roads, where fewer cues, longer reaction distances, and fewer recovery options can turn a minor lapse into a fatal outcome (Figs. 13, 14).

Conclusions

This study provided a detailed analysis of driver behavior using natural language processing and clustering methodologies. Using a large corpus of textual crash reports from the FARS database, we identified distinct behavioral patterns that contribute significantly to fatal incidents. Through the application of TF–IDF weighting, Principal Component Analysis (PCA), and Gaussian Mixture Models (GMM), we clustered fatal driver behaviors into 15 types, each characterized by unique behavioral profiles.

Our findings indicate that behaviors *Distracted*, *Right-of-way*, and *Poor-driving* dominate national fatality statistics, particularly in urban and high-density areas. Second, maneuver-specific errors such as *Overcorrecting*, *Wrong-way*, and *Tailgating* are more prevalent in rural states with limited infrastructure, pointing to geographic disparities shaped by environmental and road conditions. Additionally, emotional and enforcement-related behaviors such as *Aggressive* and *Noncompliant* driving show distinct spatial profiles and are influenced by traffic complexity and rule adherence patterns. Importantly, our analysis sheds light on driver behaviors that go beyond well-known causes, such as speeding or impairment, revealing overlooked patterns that contribute to fatal crashes.

By identifying and spatially mapping these behavioral types, our study contributes insights to policymakers, transportation agencies and researchers. Targeted interventions can be developed to address region-specific behavioral patterns, which could reduce fatal crashes through improved road designs, stricter enforcement policies, and other actions. Future research should extend this framework to integrate temporal and real-time data, enabling predictive modeling and proactive road safety measures.

Acknowledgements The authors would like to acknowledge the support of the New England University Transportation Center (NEUTC) for funding this research project.

Author Contributions The authors confirm their contribution to the paper as follows: study conception and design: JO, ZH; data collection: ZH; methodological implementation: ZH; analysis and interpretation of results: ZH, JO; draft manuscript preparation: ZH, JO. Both authors reviewed the results and approved the final version of the manuscript.

Funding This research was funded by the US Department of Transportation via the New England University Transportation Center (Grant Number: 69A35523483). Partial support was also provided by the National Science Foundation (Grant Number: 2325956).

Data Availability The data used in this study were extracted from the Fatality Analysis Reporting System at https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars. The data is publicly accessible online. For further analysis, data and code are available in a GitHub repository at https://github.com/narslab/neutc-driver-behavior-nlp-cluster.git

Declarations

Conflict of interest The authors declare no Conflict of interest.

Ethical approval This study did not require ethical approval as it did not involve any human or animal subjects.

Supplementary Information Figure 13 presents token loadings across all principal components.

References

2024 SafeTREC Traffic Safety Facts: Distracted Driving | Safe Transportation Research and Education Center. https://safetrec.berkeley.edu/2024-safetrec-traffic-safety-facts-distracted-driving?utm_source=chatgpt.com

Adanu EK, Bullard C, Dzinyela R, Jones S (2025) Unbelted and unlicensed: an investigation of crash clusters and associated injury severities. J Transport Saf Secur. https://www.tandfonline.com/doi/full/10.1080/19439962.2025.2476996. Accessed 08 May 2025

Adanu EK, Dzinyela R, Okafor S, Jones S (2024) Injury-severity analysis of crashes involving defective vehicles and accounting for the underlying socioeconomic mediators. Heliyon 10(5):26944. https://doi.org/10.1016/j.heliyon.2024.e26944

Akin D, Sisiopiku VP, Alateah AH, Almonbhi AO, Al-Tholaia MMH, Al-Sodani KAA (2022) Identifying causes of traffic crashes associated with driver behavior using supervised machine learning methods: case of highway 15 in Saudi Arabia. Sustainability 14(24):16654. https://doi.org/10.3390/su142416654

Arafa A, El-Setouhy M, Hirshon JM (2019) Driving behavior and road traffic crashes among professional and nonprofessional drivers in South Egypt. Int J Inj Contr Saf Promot 26(4):372–378. https://doi.org/10.1080/17457300.2019.1638419

Beanland V, Fitzharris M, Young KL, Lenné MG (2012) Using in-depth crash data to assess the role of driver inattention and driver distraction in crashes. Injury Prevent 18(Suppl 1):74–74. https://doi.org/10.1136/injuryprev-2012-040580g.6

Breen JM, Naess PA, Gjerde H, Gaarder C, Stray-Pedersen A (2018) The significance of preexisting medical conditions, alcohol/drug use and suicidal behavior for drivers in fatal motor vehicle crashes: a retrospective autopsy study. Forensic Sci Med Pathol 14(1):4–17. https://doi.org/10.1007/s12024-017-9934-x

Çelik AK, Oktay E (2014) A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey. Accid Anal Prevent 72:66–77. https://doi.org/10.1016/j.aap.2014.06.010

Farooq D, Moslem S, Duleba S (2019) Evaluation of driver behavior criteria for evolution of sustainable traffic safety. Sustainability 11(11):3142. https://doi.org/10.3390/su11113142

Fatality Facts 2023: Urban/Rural Comparison. https://www.iihs.org/research-areas/fatality-statistics/detail/urban-rural-comparison

Fiolić M, Babić D, Babić D, Tomasović S (2023) Effect of road markings and road signs quality on driving behaviour, driver's gaze patterns and driver's cognitive load at night-time. Transp Res F: Traffic Psychol Behav 99:306–318. https://doi.org/10.1016/j.trf.2023.10.025

Ghandour AJ, Hammoud H, Al-Hajj S (2020) Analyzing factors associated with fatal road crashes: a machine learning approach. Int J Environ Res Public Health 17(11):4111. https://doi.org/10.3390/ijerph17114111

Global Status Report on Road Safety (2023) https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023

Hossain MM, Zhou H, Das S, Sun X, Hossain A (2023) Young drivers and cellphone distraction: pattern recognition from fatal crashes. J Transport Saf Secur 15(3):239–264. https://doi.org/10.1080/19439962.2022.2048763

Høye A (2020) Speeding and impaired driving in fatal crashes–Results from in-depth investigations. Traffic Inj Prev 21(7):425–430. https://doi.org/10.1080/15389588.2020.1775822

Islam M, Hosseini P, Kakhani A, Jalayer M, Patel D (2024) Unveiling the risks of speeding behavior by investigating the dynamics of driver injury severity through advanced analytics. Sci Rep 14(1):22431. https://doi.org/10.1038/s41598-024-73134-z

Khan Rifat MA, Kabir A, Huq A (2024) An explainable machine learning approach to traffic accident fatality prediction. Proc Comput Sci 246:1905–1914. https://doi.org/10.1016/j.procs.2024.09.704

Khan SA, Afghari AP, Yasmin S, Haque MM (2023) Effects of design consistency on run-off-road crashes: an application of a Random Parameters Negative Binomial Lindley model. Accid Anal Prevent 186:107042. https://doi.org/10.1016/j.aap.2023.107042

Khattak AJ, Ahmad N, Wali B, Dumbaugh E (2021) A taxonomy of driving errors and violations: evidence from the naturalistic driving study. Accid Anal Prevent 151:105873. https://doi.org/10.1016/j.aap.2020.105873

Kuo P-F, Lord D (2021) A visual approach for defining the spatial relationships among crashes, crimes, and alcohol retailers: Applying the color mixing theorem to define the colocation pattern of multiple variables. Accid Anal Prevent 154:106062. https://doi.org/10.1016/j.aap.2021.106062

Lee S, Arvin R, Khattak AJ (2023) Advancing investigation of automated vehicle crashes using text analytics of crash narratives and Bayesian analysis. Accid Anal Prevent 181:106932. https://doi.org/10.1016/j.aap.2022.106932

Li Y, Karim MM, Qin R, Sun Z, Wang Z, Yin Z (2021) Crash report data analysis for creating scenario-wise, spatio-temporal attention guidance to support computer vision-based perception of fatal crash risks. Accid Anal Prevent 151:105962. https://doi.org/10.1016/j.aap.2020.105962

National Motor Vehicle Crash Causation Survey (NMVCCS) (2024) National Highway Traffic Safety Administration. https://catalog.data.gov/dataset/national-motor-vehicle-crash-causation-survey-nmvccs. Accessed 08 Apr 2025

Ning H, Yu Y, Lu B (2022) Analyzing the impact of traffic violation behaviors on traffic fatal crashes using multilevel models on expressways. Math Probl Eng. https://doi.org/10.1155/2022/9917877

NLTK : : nltk.stem.WordNetLemmatizer. https://www.nltk.org/api/nltk.stem.WordNetLemmatizer.html?highlight=wordnet. Accessed 02 May 2025

Odoardi S, Biosa G, Mestria S, Valentini V, De Giovanni N, Cittadini F, Strano Rossi S (2023) Drug-impaired driving and traffic collisions: study on a cross section of the Italian population. Drug Test Anal 15(5):477–483. https://doi.org/10.1002/dta.3366

Okafor S, Adanu EK, Jones S (2022) Severity analysis of crashes involving in-state and out-of-state large truck drivers in Alabama: A random parameter multinomial logit model with heterogeneity in means and variances. Heliyon 8(12):11989. https://doi.org/10.1016/j.heliyon.2022.e11989

Osman OA, Hajij M, Karbalaieali S, Ishak S (2019) A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. Accid Anal Prevent 123:274–281. https://doi.org/10.1016/j.aap.2018.12.005

Paaß G, Giesselbach S (2023) Foundation models for natural language processing—pre-trained language models integrating media. arXiv . https://doi.org/10.48550/arXiv.2302.08575

Penmetsa P, Pulugurtha SS, Duddu VR (2018) Factors associated with crashes due to overcorrection or oversteering of vehicles. IATSS Res 42(1):24–29. https://doi.org/10.1016/j.iatssr.2017.06.001

Romano E, Voas RB (2011) Drug and alcohol involvement in four types of fatal crashes. J Stud Alcohol Drugs 72(4):567–576

Roussou S, Michelaraki E, Katrakazas C, Pooyan Afghari AP, Al Haddad C, Alam MR, Antoniou C, Papadimitriou E, Brijs T, Yannis G (2024) Unfolding the dynamics of driving behavior: a machine learning analysis from Germany and Belgium. Eur Transp Res Rev 16:40

Sayed MA, Qin X, Kate RJ, Anisuzzaman DM, Yu Z (2021) Identification and analysis of misclassified work-zone crashes using text mining techniques. Accid Anal Prevent 159:106211. https://doi.org/10.1016/j.aap.2021.106211

Shangguan Q, Fu T, Wang J, Jiang R, Fang S. Quantification of rear-end crash risk and analysis of its influencing factors based on a new surrogate safety measure. https://doi.org/10.1155/2021/5551273

Shao Y, Shi X, Zhang Y, Shiwakoti N, Xu Y, Ye Z (2024) Injury severity prediction and exploration of behavior-cause relationships in automotive crashes using natural language processing and extreme gradient boosting. Eng Appl Artif Intell 133:108542. https://doi.org/10.1016/j.engappai.2024.108542

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

SPARCK JONES K (1972) A statistical interpretation of term specificity and its application in retrieval. J Document 28(1):11–21. https://doi.org/10.1108/eb026526

Stanojević P, Jovanović D, Lajunen T (2013) Influence of traffic enforcement on the attitudes and behavior of drivers. Accid Anal Prevent 52:29–38. https://doi.org/10.1016/j.aap.2012.12.019

State Infrastructure Facts | ASCE's 2025 Infrastructure Report Card (2016)

Susilo YO, Joewono TB, Vandebona U (2015) Reasons underlying behaviour of motorcyclists disregarding traffic regulations in urban areas of Indonesia. Accid Anal Prevent 75:272–284. https://doi.org/10.1016/j.aap.2014.12.016

The Prevalence of Distraction Among Passenger Vehicle Drivers: A Roadside Observational Approach | Request PDF. https://www.researchgate.net/publication/261880535_The_Prevalence_of_Distraction_Among_Passenger_Vehicle_Drivers_A_Roadside_Observational_Approach?utm_source=chatgpt.com

Theofilatos A, Yannis G (2014) A review of the effect of traffic and weather characteristics on road safety. Accid Anal Prevent 72:244–256. https://doi.org/10.1016/j.aap.2014.06.017

Valcamonico D, Baraldi P, Amigoni F, Zio E (2021) A natural language processing method for the identification of the factors influencing road accident severity. PHM Soc Eur Confer 6(1):12–12. https://doi.org/10.36001/phme.2021.v6i1.2899

van Schoor O, van Niekerk JL, Grobbelaar B (2001) Mechanical failures as a contributing cause to motor vehicle accidents—South Africa. Accid Anal Prevent 33(6):713–721. https://doi.org/10.1016/S0001-4575(00)00083-X

Wali B, Khattak AJ, Karnowski T (2020) The relationship between driving volatility in time to collision and crash-injury severity in a naturalistic driving environment. Analytic Methods Accid Res 28:100136. https://doi.org/10.1016/j.amar.2020.100136

Wundersitz L (2019) Driver distraction and inattention in fatal and injury crashes: findings from in-depth road crash data. Traffic Inj Prev 20(7):696–701. https://doi.org/10.1080/15389588.2019.1644627

Zheng L, Li Y, Ding T, Meng F, Zhang Y (2023) Analysis of factors affecting crash under risk scenarios based on driver homogenous clustering. PLoS ONE 18(10):0293307. https://doi.org/10.1371/journal.pone.0293307

Zou G, Huang H, Zhou H, Li J (2024) Crash-causing information extraction via text-mining techniques: implementation of the Chinese state-related crash narratives. Transport Saf Environ 6(4):018. https://doi.org/10.1093/tse/tdae018

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.