Accident Analysis and Prevention 152 (2021) 106012

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Characterizing phone usage while driving: Safety impact from road and operational perspectives using factor analysis

Xiaoqiang Kong (Jack)[a], Subasish Das [b,*], Hongmin Zhou (Tracy)[c], Yunlong Zhang [d]

[a] Texas A&M University, 3135 TAMU, College Station, TX 77843-3135, United States
[b] Texas A&M Transportation Institute, 1111 RELLIS Parkway, Bryan, TX 77807, United States
[c] Texas A&M Transportation Institute, 701 N. Post Oak Road, Suite 430, Houston, TX 77024, United States
[d] Zachry Department of Civil & Environmental Engineering, 3136 TAMU, College Station, TX 77843-3136, United States

ARTICLE INFO

Keywords:
Phone use while driving (PUWD)
Driving distraction
Factor analysis
Dimension reduction method
Data mining
Traffic safety

ABSTRACT

Phone use while driving (PUWD) is one of the most crucial factors of distraction related traffic crashes. This study utilized an unsupervised learning method, known as factor analysis, on a unique distracted driving dataset to understand PUWD behavior from the roadway geometry and operational perspectives. The results indicate that the presence of a shoulder, median, and access control on the relatively higher functional class roadways could encourage more PUWD events. The roadways with relatively lower speed limits could have high PUWD event occurrences if the variation in operating speed is high. The results also confirm the correlations between the frequency of PUWD events and the frequency of distracted crashes. This relationship is strong on urban roadways. For rural roadways, this correlation is only strong on the roadways with a large amount of PUWD events. The findings could help transportation agencies to identify suitable countermeasures in reducing distraction related crashes. Moreover, this study provides researchers a new perspective to study PUWD behavior rather than only focus on drivers' personalities.

1. Introduction

Distraction related crashes are normally considered "preventable" but occur often (Bingham, 2014; Kahn et al., 2015). In 2018, a total of 2,841 traffic fatalities were associated with distracted driving (National Highway Traffic Safety Administration or NHTSA, 2020). Phone use while driving (PUWD) behavior is one of the most common factors of distracted driving (Laberge-Nadeau et al., 2003; Lamble et al., 2002; McEvoy et al., 2006). Many transportation agencies invested significant effort to mitigate PUWD behavior. These efforts include passing a law to restrict PUWD behavior in a moving vehicle and numerous local defensive driving education programs. In the U.S., 48 states have banned text messaging while driving (IIHS, 2019). The NHTSA dedicates resources to educate and promote defensive driving, such as collaborating with local schools (NHTSA, 2020).

PUWD behavior gradually captures researchers' attention. However, PUWD behavior is a complex psychological behavior driven by many factors, including drivers' personality, environmental factors, and roadway operational factors. One of the main challenges is the difficulty of obtaining the real-world data that could reflect the real PUWD behavior to perform the research. Only in particular experiments, participants allow researchers to install cameras in their vehicles to record their PUWD behavior. However, in these experiments, the number of participants is often limited and possibly biased. In some states, the crash reports might contain some information about if this accident is related to PUWD behavior. There is evidence showing this kind of case is often under-reported, and drivers who were involved in distraction-affected crashes were reluctant to admit the PUWD behavior (National Safety Council, 2013; Regev et al., 2017). An alternative approach to collect the PUWD data is through simulators (Chen et al., 2020). The main argument towards using data from simulators is that the data cannot reflect real-world situations due to the complex nature of the PUWD behavior. Many researchers have investigated efforts to understand PUWD behaviors at the driver level (Atwood et al., 2018; Papadimitriou et al., 2019). A gap exists in understanding the real world PUWD behavior from roadway geometrical and operational characteristics.

This study utilized a large PUWD data dataset (pseudonymized), which was originated from a private data service provider. The data collection process is based on a smart phone application that promotes

* Corresponding author.
E-mail addresses: X-Kong@tti.tamu.edu (X. Kong), s-das@tti.tamu.edu (S. Das), h-zhou@tti.tamu.edu (H. Zhou), yzhang@civil.tamu.edu (Y. Zhang).

https://doi.org/10.1016/j.aap.2021.106012

defensive driving without being distracted by the phone. The phone application could collect PUWD data from users who have downloaded the application and granted access to their vehicle movement data. The application encourages defensive driving by rewarding drivers with points for each minute of driving at a speed over 10 mph without interacting with their phones. Rewarded points can be redeemed in many local businesses. This rewarding process will be interrupted if the driver uses a phone while driving, also called a PUWD event. The study site of this research was Texas. The researchers integrated all PUWD events with the Texas road inventory as well as the distracted crash count on each road segment in the road inventory from the crash database of Texas. The aim of this research is to investigate the relationship between PUWD behavior and road geometrics such as shoulder, median, and number of lanes, and operational road characteristics such as annual average daily traffic (AADT), peak hour factor, and access control. Moreover, this research also investigates the relationship between the frequency of PUWD behavior and the distracted crash that occurred on the same road segment.

It might be intuitive to think PUWD behavior is a self-choice. There should be more investment into defensive driving education and law enforcement on roadways to solve the problem. The main hypothesis of this study is that PUWD behavior is more than just a self-choice. The combination of roadway features and operational features might encourage PUWD behavior. This massive real-world data empowers this research to identify the possible patterns that might encourage PUWD behavior, rather than assume the impulsive drivers are the only reason for PUWD events. This research could also offer a new perspective that investing efforts from geometrical and operational perspectives could also help to reduce the PUWD events.

2. Literature review

It is estimated that 9.7 percent of drivers were using either handheld or hands-free phones while driving in 2018. This translates to over one million drivers of PUWD behaviors at a typical daylight moment (NHTSA, 2019). Studies found clear demographical characteristics in PUWD behaviors. Younger drivers tend to have more PUWD, especially texting, behaviors than older drivers (Gras et al., 2007; Hoff et al., 2013; NHTSA, 2019). Male drivers are more likely to talk or text on the phone while driving than female drivers (Backer-Grøndahl and Sagberg, 2011; Brusque and Alauzet, 2008; Chen, 2007; Hallett et al., 2012). A key factor of distracted driving is the attitude and willingness toward distracted driving associated with their sociodemographic profile along with the perceived risk of distracted driving (Qi et al., 2020).

There is abundant literature on the impact of PUWD on driving performance and safety. PUWD impairs drivers' capability to sustain suitable speed, throttle control, and lateral position of the vehicle. It also can weaken drivers' visual search patterns, reaction time, and decision-making criteria (Brace et al., 2007). Distraction from visual-manual interaction with phone usage (e.g., reading emails on smartphones) can increase standard deviation of speed and lateral lane positions compared with cognitive distraction (e.g., conversing on phones) (Onate-Vega et al., 2020). Drivers tend to reduce speed when they are using handheld devices (Choudhary and Velaga, 2017; Liu and Lee, 2006), but this compensatory behavior is not necessarily observed when hands-free devices are used (Alm and Nilsson, 1995; Törnros and Bolling, 2006). Increasing car following headways is another compensatory behavior associated with PUWD (Jamson et al., 2004; Li et al., 2019; Yannis et al., 2014). Nevertheless, it is generally believed that the detrimental effects of PUWD on driving performance increase crash risk significantly. Depending on the type of phone usage, PUWD has been reported to increase crash risk by a range from 2.2 times to 23 times compared with undistracted driving (Choudhary and Velaga, 2017; Dingus et al., 2016; McEvoy et al., 2005; Redelmeier and Tibshirani, 1997; Strayer et al., 2006). Texting has a higher risk than talking on the phone while driving (Dingus et al., 2016; Stavrinos et al., 2015). Many studies found no significant difference between using handheld and hands-free devices in terms of the negative impact of PUWD on driving performance (Claveria et al., 2019), although a majority of drivers believe that hands-free PUWD is safer (National Safety Council, 2014).

A majority of these existing studies focused on the impact of PUWD but fall short in providing insights into the cause of PUWD due to the intrinsic limitations of methodologies adopted by these studies. These study methods include surveys or interviews (Brusque and Alauzet, 2008; Korpinen and Pääkkönen, 2012; Violanti and Marshall, 1996), field observations (Bommer, 2018; Foss et al., 2009; Goodwin et al., 2012; NHTSA, 2019), driving experiments using a simulator (Alm and Nilsson, 1995; Holland and Rathod, 2013; Strayer et al., 2006) or a vehicle in the field (Al-Darrab et al., 2009; Lamble et al., 1999; Rosenbloom, 2006), NDS using an in-vehicle data acquisition system (DAS) (Dingus et al., 2016; Fitch et al., 2015; Foss and Goodwin, 2014; Hickman and Hanowski, 2012; Schneidereit et al., 2017), and crash related studies (Das and Sun, 2015). Surveys or interviews based studies obtained drivers' responses about PUWD to identify demographical driver characteristics and to establish links between PUWD and crash risks indirectly. The self-reported PUWD data is possibly biased as the respondents may feel intimidated in providing accurate responses regarding their PUWD behaviors. It is also challenging to quantify external factors, such as roadway geometry or traffic conditions, causing PUWD behaviors from the survey data. The field observation studies found limited trends of PUWD tendency based upon the geographic and traffic conditions of the field locations and the time of the observations. This type of study requires a large number of observation samples to yield statistically significant findings and thus would probably be very costly. Driving experiments directly revealed the impact of PUWD on driving performance, but PUWD behaviors were almost always associated with the designed exposures, not the triggered results in the experiments. It would be a challenge to naturally trigger PUWD behaviors in an experimental setting regardless participants' knowledge about the experiment's purpose. The NDS captured PUWD behaviors in the context of their everyday lives and provided detailed information from safety-critical events (SCEs). However, the existing NDS datasets may have been underutilized in terms of capturing PUWD behaviors not related to any SCEs and extracting roadway or operational information as causal factors preceding PUWD behaviors, likely due to the extensive sizes of the datasets. Additionally, NDS datasets may lack sufficient diversity in roadway geometric and operational conditions to study contributing factors for PUWD behaviors. All of the above study methods suffer more or less from the small sample size issue.

While NDS data are widely used to investigate driver behaviors (Grimberg et al., 2020; Kong et al., 2020; Schneidereit et al., 2017), smartphones are gaining interest as tools to detect driver PUWD behaviors inside the vehicle (Papadimitriou et al., 2019; Wang et al., 2016) and prevent distracted driving (Oviedo-Trespalacios et al., 2020, 2019). Functioning as the accelerometer, the gyroscope, the magnetometer, and the GPS to sense driver behaviors and vehicle dynamics, smartphones are considered as a cost-effective alternative for collecting NDS data, compared with the in-vehicle DAS (Grimberg et al., 2020). Some studies have shown promising results from using smartphone data to link limited roadway and operational conditions with PUWD (Papadimitriou et al., 2019) and harsh vehicular maneuvers (Petraki et al., 2020).

Nevertheless, the literature on external factors, such as road and operational characteristics that cause PUWD is quite limited. Existing studies extensively supported the claim that the PUWD behavior results in a higher speed variation (Bowden et al., 2019; Onate-Vega et al., 2020; Oviedo-Trespalacios et al., 2018). One study (Papadimitriou et al., 2019) analyzed driver smartphone usage data and shed light on PUWD occurrence scenarios. Drivers are more prone to PUWD on divided highways or residential roadways but are less likely to use phones while driving on undivided highways. Drivers are less likely to use phones when driving at a higher average speed (55 mph or higher) compared

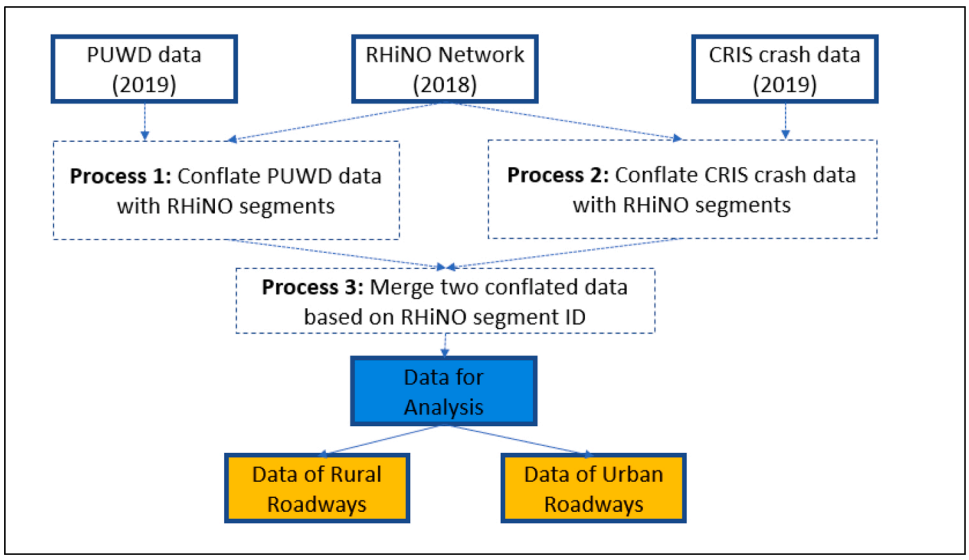

Fig. 1. Data conflation framework.

with a lower average speed. Driving speed in excess of the speed limit is a strong indicator of the tendency not to use phones on rural roadways or roadways with a higher functional class, but it may not necessarily be the case on urban roadways. Hash driving maneuvers (e.g., sudden acceleration or braking) are also negatively correlated with PUWD behaviors. PUWD tends to happen more in morning peak hour but less in afternoon peak hours. Two other studies (Papantoniou et al., 2020; Sharda et al., 2019) using naturalistic driving studies (NDS) data found that drivers show a greater proclivity to use a cell phone or tablet on longer trips (30 min or longer) than shorter trips. Several studies found that talking (rather than talking or dialing) on the phone does not impact drivers' behavior (Onate-Vega et al., 2020; Oviedo-Trespalacios et al., 2018; Young, 2018). However, there exists a gap in investigating the impact of more detailed roadway geometry and operational features on PUWD using smartphone data. Also, to the best of the authors' knowledge, this is the first attempt to establish a correlation between naturalistic PUWD events and distracted driving crash frequency on a large scale of the roadway network.

3. Methodology

3.1. Data preparation

3.1.1. Data sources

There are three critical data sources

- PUWD data
- Traffic crash data from the Crash Records Information System (CRIS).
- Roadway inventory data from the Road-Highway Inventory Network Offload (RHiNO).

3.1.1.1. PUWD data (2019). The PUWD data is pseudonymized data provided by a private data service provider. This vendor collected data through a phone application in an effort to reduce distracted driving behaviors. The users who downloaded the applications authorize the vendor to access their phone mobility data in exchange for reward points, which can be redeemed as coupons at partner businesses. The application records two types of mobility data: driving data and distraction data. For driving data, the application starts to record users' locations with temporal information when the vehicle's moving speed above 10 mph in about one data point per every minute frequency. For distraction data, the application records a geospatial point with temporal information when the drivers use their phones. This research will only focus on distraction data.

Texas is the study site for this research. The distraction dataset of android users from Texas is extracted for analysis between January 1, 2019, and December 31, 2019. After eliminating these abundant PUWD events with NULL or 0 mph driving speed, this distraction dataset contains 5,262,930 distraction events where the PUWD events occurred from 27,524 unique users.

3.1.1.2. Crash data from CRIS (2019). The Texas crash database is managed by the Texas Department of Transportation's (TxDOT's) Crash Records Information System (CRIS), which is open to the public. All the crash events in Texas in 2019 are extracted. In the CRIS database, there is one attribute describing whether the crash event is phone use related. Since PUWD behavior is one of the most common factors of distracted driving (Laberge-Nadeau et al., 2003; Lamble et al., 2002; McEvoy et al., 2006), the following analysis considers phone use related crashes as distraction-affected crashes.

3.1.1.3. Roadway inventory data RHiNO (2018). TxDOT's Roadway Highway Inventory Network Offload (RHiNO) database contains the roadway inventory data of Texas. The database contains detailed information about roadway geometrical information such as functional class, shoulder type, and number of lanes, as well as general operation level information such as AADT and truck percentage. At the time this research was conducted, the newest RHiNO version is RHiNO 2018; thus, RHiNO 2018 is used in this research. To increase the reliability of this research, road segments of less than 0.1 miles were excluded.

3.1.2. Data conflation process

Fig. 1 shows the flow of the data preparation process of conflating PUWD data and crash data on the RHiNO network.

3.1.2.1. Process 1: conflate PUWD data with RHiNO segments. The PUWD database contains the geospatial and temporal information of each PUWD event. This process utilized the Postgres spatial database to perform the conflation. The spatial join function from the PostGIS library is used. All spatial points of PUWD events are snapped to the nearest RHiNO roadway segments. One problem is that RHiNO only contains on-system roadways in Texas. The PUWD events that occurred far away from these on-system roadways will still locate their nearest on-system roadways, but the truth is these PUWD events occurred on

Table 1
Sample Data.

RHiNO_key	PUWD	seg_length	AADT	num_lanes	ru_f_syste	Distracted_CC
0002-06_94.465_95.918_76116	52	1.453	16229	4	R1	2
0002-06_95.918_97.371_76117	59	1.453	16229	4	R1	0
0002-06_97.371_98.824_76118	62	1.453	16229	4	R1	0
0002-07_100.2995_101.775_78156	69	1.4755	16229	4	R1	1
0002-07_101.775_103.729_77787	101	1.954	16229	4	R1	5
0002-07_103.729_105.683_77788	84	1.954	16229	4	R1	0
0002-07_105.683_106.614_75683	35	0.931	16052	4	R1	0
0002-07_98.824_100.2995_78155	64	1.4755	16229	4	R1	5

Table 2
Variables Descriptions.

Variable Name	Description	Data Type	Levels	Data Source
PUWD	total count of 2019 PUWD events per mile	numerical	–	Private Data Service Provider
AADT	average annual daily traffic	numerical	–	RHiNO
med_type	median type	categorical	no_med curbed_med positive_barrier unprotected_med	RHiNO
med_wid	median width	numerical	–	RHiNO
num_lanes	number of lanes	numerical	–	RHiNO
lane_width	lane width	numerical	–	RHiNO
k_fac	peak hour factor	numerical	–	RHiNO
aces_ctrl	access control	categorical	full_ctrl none_ctrl partial_ctrl	RHiNO
s_type_i	inside shoulder type	categorical	no_shoulder_i shoulder_exists_i	RHiNO
s_type_o	outside shoulder type	categorical	no_soulder_o shoulder_exists_o	RHiNO
s_wid_i	inside shoulder width	numerical	–	RHiNO
s_wid_o	outside shoulder width	numerical	–	RHiNO
ru_f_syste	functional classification	categorical	R1/U1 (rural/ urban interstate) R2/U2 (rural/ urban freeway and expressway) R3/U3 (rural/ urban principal arterial) R4/U4 (rural/ urban minor arterial) R5/U5 (rural/ urban major collector) R6/U6 (rural/ urban minor collector) R7/U7 (rural/ urban local)	RHiNO
PSL	posted speed limit	numerical	–	RHiNO
spddif	posted speed limit minus the average speed of all PUWD events on this segment	numerical	–	Private Data Service Provider
Distracted_CR	total distracted crash count in 2019 per mile	numerical	–	CRIS

roadways that are not in the RHiNO database. Many rounds of testing are conducted to avoid inaccurate conflation, and then, any points further than 30 m away from the nearest road segment were excluded. The conflated dataset contains PUWD events and corresponding roadway segments. To fulfill the purpose of this research, the total number of PUWD events (distraction events) on each RHiNO segment can be obtained by simple aggregation.

3.1.2.2. Process 2: conflate crash data with RHiNO segments. The process of conflating the crash data with the RHiNO segments is similar to process 1. The crash data provide geospatial coordinates: longitude and latitude. The final conflated file contains crashes and corresponding RHiNO segments. As mentioned above, in the crash dataset, there is an attribute described if the crash event is distraction-affected. The total distraction-affected crash events can be obtained by aggregation based on this distraction-affection attribute.

3.1.2.3. Process 3: conflate two conflated datasets. After conflating three datasets into two separate files, these two files can be merged by the common key: RHiNO_key. The RHiNO_key is generated in the data process as a unique key of each road segment by combining route number, milepost of the beginning and end of the segment. The final dataset (see Table 1 for sample data format and Table 2 for the list of variables used in this study) contains unique road segments with their corresponding total number of PUWD events, AADT, and total distracted crash counts. To fulfill the purpose of this study, an investigation into PUWD behavior is necessary. As a data-driven study, even with massive data points, a large amount of RHiNO segments were not covered by the users. However, these road segments with no PUWD events do not necessarily suggest that the geometric and operational factors of the segment considered in this study do not encourage PUWD behavior. The inclusion of these segments may compromise the purpose of this study. Thus, road segments with zero PUWD events were not included in this study. Meanwhile, the records with missing values are eliminated. The following table is a sample of the final data. The final data are divided into two data files. One file is for the urban roadways, and another one is for the rural roadways. The motivation to separate the dataset is that the PUWD behavior would be different on these two roadway types. Intuitively, urban roadways have more traffic control devices such as traffic lights, and more traffic and rural roadways have better traffic conditions with fewer traffic interruptions. These two different scenarios could lead to different PUWD behaviors. The final urban data file contains 16,410 urban roadway segments and 3,173,901 PUWD events. The final rural data file contains 18,816 rural roadway segments with 1,155,623 PUWD events.

3.2. Variable selection

In the prepared dataset, 16 variables are selected to perform unsupervised learning. The variable "RHiNO_key" is the unique identification for each RHiNO segment available for analysis. In these 16 variables, the variable "PUWD" is the total amount of PUWD events of each segment from the private data service provider, and the variable "Distracted_CC" is the total amount of distracted crash count of each segment from the

Table 3
Count and Proportions of Qualitative Variables (Rural).

Variable	Category	Rural Case 1		Rural Case 2		Rural Case 3	
		Count	Prop	Count	Prop	Count	Prop
aces_ctrl	full_ctrl	1374	21.23	432	6.39	100	1.79
aces_ctrl	partial_ctrl	366	5.66	174	2.58	39	0.70
aces_ctrl	none_ctrl	4732	73.11	6151	91.03	5448	97.51
s_type_i	shoulder_exists_i	6017	92.97	6217	92.01	4856	86.92
s_type_i	no_soulder_i	455	7.03	540	7.99	731	13.08
s_type_o	shoulder_exists_o	6103	94.3	6288	93.06	4899	87.69
s_type_o	no_soulder_o	369	5.7	469	6.94	688	12.31
ru_f_syste	R1	1289	19.92	398	5.89	96	1.72
ru_f_syste	R2	85	1.31	34	0.5	4	0.07
ru_f_syste	R3	2711	41.89	2158	31.94	1242	22.23
ru_f_syste	R4	1199	18.53	1753	25.94	1695	30.34
ru_f_syste	R5	1117	17.26	2160	31.97	2144	38.37
ru_f_syste	R6	69	1.07	250	3.7	402	7.20
ru_f_syste	R7	2	0.03	4	0.06	4	0.07

Table 4
Descriptive Statistics of Quantitative Variables (Rural).

Attribute	mean	sd	min	max	IQR
Rural Case 1: PUWDR > 40 (n = 6,472)					
DistractedCR	1.3	3.9	0.0	63.0	0.7
PUWDR	223.5	354.7	40.0	3910.3	174.7
AADT	14543.1	14608.7	0.0	104785.0	14606.8
k_fac	9.7	1.6	0.0	29.2	1.6
spddif	4.3	8.8	−59.4	51.5	8.6
med_wid	20.8	31.2	0.0	280.0	40.0
num_lanes	3.1	1.1	0.0	6.0	2.0
PSL	65.3	11.0	0.0	85.0	20.0
lane_width	12.1	1.3	0.0	31.0	0.0
s_wid_i	7.5	4.2	0.0	32.0	6.0
s_wid_o	11.2	7.3	0.0	48.0	15.0
Rural Case 2: PUWDR > 7 & PUWDR < 41 (n = 6,757)					
DistractedCR	0.7	2.8	0.0	107.4	0.0
PUWDR	20.0	9.8	7.0	41.0	16.4
AADT	6481.3	6989.9	70.0	51270.0	5802.0
k_fac	10.2	1.8	5.3	29.0	1.8
spddif	4.7	9.1	−32.2	56.0	9.3
med_wid	12.7	30.8	0.0	455.0	0.0
num_lanes	2.6	0.9	2.0	6.0	2.0
PSL	64.5	10.7	20.0	80.0	20.0
lane_width	12.0	1.4	8.0	30.0	0.0
s_wid_i	6.6	4.0	0.0	32.0	6.0
s_wid_o	8.5	6.3	0.0	38.0	7.0
Rural Case 3: Rural Case 3: PUWDR < 8 (n = 5,587)					
DistractedCR	0.4	1.5	0.0	40.5	0.0
PUWDR	3.5	2.1	0.5	8.0	3.5
AADT	3208.0	3779.4	12.0	59812.0	3063.0
k_fac	10.8	2.0	5.6	30.0	2.2
spddif	4.7	11.2	−38.0	56.0	12.0
med_wid	4.2	19.1	0.0	455.0	0.0
num_lanes	2.2	0.7	2.0	6.0	0.0
PSL	64.2	9.5	25.0	85.0	20.0
lane_width	11.8	1.3	6.0	28.0	1.0
s_wid_i	5.4	3.8	0.0	24.0	6.0
s_wid_o	6.1	5.0	0.0	35.0	7.0

CRIS dataset. The rest variables that describe the geometrical and operational characteristics of each road segment are from the RHiNO dataset.

3.3. Summary statistics

The datasets developed for rural and urban areas have been divided into three broad groups based on the k-means clustering of PUWD measures per mile or PUWD rate (denoted as PUWDR). This initial clusters have been conducted as there is a wide range of PUWDR in both rural and urban databases. Factor analysis, without considering the subgroup effect in the data, may produce biased findings. The rural dataset was divided into the following three groups:

- Rural Case 1: PUWDR > 40 (n = 6,472)
- Rural Case 2: PUWDR > 7 & PUWDR < 41 (n = 6,757)
- Rural Case 3: PUWDR < 8 (n = 5,587)

The urban dataset was divided into three groups:

- Urban Case 1: PUWDR > 160 (n = 5,790)
- Urban Case 2: PUWDR > 30 & PUWDR < 161 (n = 5,100)
- Urban Case 3: PUWDR < 31 (n = 5,520)

Tables 3 and 4 provide descriptive statistics of the qualitative and quantitative variables for the rural datasets, respectively. In Table 3, the rural data shows that PUWD behavior mostly occurred at road segments without access control, segments with shoulder presence, and higher facility types. Table 4 presents the descriptive statistics of the quantitative variables. For rural case 1, the average PUWD event count per mile is 223.5, and the distracted crash count per mile is 1.3. For rural cases 2 and 3, the average PUWD event count drop to 20 and 3.5, respectively, which are expected after the whole dataset has been divided into 3 cases based on the number of total PUWD event counts. Interestingly, the decrease of the total distracted crash count is also observed as the total PUWD event count drops. The descriptive statistics show a positive association between AADT and PUWD events. It is obvious as higher traffic volume increases the exposure leading to higher PUWD events. Moreover, wider lane width, median and shoulder and more number of lanes often associate with higher AADT.

Tables 5 and 6 provide descriptive statistics of the qualitative and quantitative variables for the urban datasets, respectively. In Table 5, the urban data show the segments with PUWD events occur mostly on roadways with no access control. However, the second dominant group is the roadways with full access control. This table also shows that the urban dataset has similar findings like the rural dataset. The three cases are divided based on the number of total PUWD event counts. The statistics for three cases state that as the average PUWD count decreases, the number of distracted crash counts, AADT, median width, number of lanes, and shoulder width also decrease.

3.4. Factor analysis of mixed data

Presence of both quantitative (integer or numeric) and qualitative (categorical) variables in a dataset is known as mixed data. Factor analysis is the most conventional method to tackle this kind of data problem. The common approach is to transform the quantitative variables into qualitative variables by breaking down their ranges into several clusters, so the generated new data structure can be analyzed in the correspondence analysis (CA) framework. CA has been widely used

Table 5
Count and Proportions of Qualitative Variables (Urban).

Variable	Category	Case 1		Case 2		Case 3	
		Count	Prop	Count	Prop	Count	Prop
aces_ctrl	full_ctrl	2381	41.12	1599	31.35	441	7.99
aces_ctrl	partial_ctrl	183	3.16	151	2.96	131	2.37
aces_ctrl	none_ctrl	3226	55.72	3350	65.69	4948	89.64
s_type_i	shoulder_exists_i	4470	77.2	3912	76.71	3819	69.18
s_type_i	no_soulder_i	1320	22.8	1188	23.29	1701	30.82
s_type_o	shoulder_exists_o	4765	82.3	4142	81.22	4053	73.42
s_type_o	no_soulder_o	1025	17.7	958	18.78	1467	26.58
ru_f_syste	U1	1508	26.04	931	18.25	144	2.61
ru_f_syste	U2	873	15.08	669	13.12	297	5.38
ru_f_syste	U3	2402	41.49	2314	45.37	2580	46.74
ru_f_syste	U4	785	13.56	868	17.02	1614	29.24
ru_f_syste	U5	222	3.83	304	5.96	868	15.72
ru_f_syste	U6	0	0.00	11	0.22	17	0.31
ru_f_syste	U7	0	0.00	3	0.06	0	0.00

Table 6
Descriptive Statistics of Quantitative Variables (Urban).

Attribute	mean	sd	min	max	IQR
Urban Case 1: PUWDR > 160 (n = 5,790)					
DistractedCR	7.9	14.2	0.0	210.9	9.1
PUWDR	1137.5	1419.9	160.1	17109.4	1030.3
AADT	58406.3	60683.3	0.0	330096.0	57980.0
k_fac	9.7	1.2	0.0	19.4	1.3
spddif	8.2	8.8	−36.5	44.0	11.5
med_wid	19.1	33.1	0.0	500.0	28.0
num_lanes	4.7	2.0	0.0	14.0	2.0
PSL	56.3	10.7	0.0	80.0	15.0
lane_width	12.4	1.4	0.0	31.0	0.0
s_wid_i	8.8	7.5	0.0	54.0	10.0
s_wid_o	11.8	8.5	0.0	48.0	16.0
Urban Case 2: PUWDR > 30 & PUWDR < 161 (n = 5,100)					
DistractedCR	4.4	8.0	0.0	129.3	6.0
PUWDR	76.7	35.6	30.0	160.9	56.5
AADT	36305.2	40032.1	0.0	252420.0	31539.3
k_fac	9.8	1.3	0.0	23.5	1.3
spddif	7.0	9.0	−41.3	54.4	10.8
med_wid	16.6	29.7	0.0	500.0	30.0
num_lanes	4.2	1.6	0.0	13.0	1.0
PSL	55.4	12.0	0.0	85.0	20.0
lane_width	12.5	1.8	0.0	33.0	0.0
s_wid_i	7.9	6.8	0.0	46.0	9.0
s_wid_o	10.9	8.2	0.0	40.0	17.0
Urban Case 3: PUWDR < 31 (n = 5,520)					
DistractedCR	3.1	5.9	0.0	88.1	4.0
PUWDR	12.8	8.4	0.5	31.0	13.7
AADT	15136.8	15556.0	98.0	213306.0	13510.0
k_fac	10.0	1.7	6.4	30.0	1.8
spddif	8.2	10.5	−62.5	51.3	12.9
med_wid	10.5	30.6	0.0	500.0	3.0
num_lanes	3.5	1.3	1.0	12.0	2.0
PSL	51.0	10.7	0.0	85.0	10.0
lane_width	12.6	1.9	8.0	31.0	0.3
s_wid_i	5.6	5.3	0.0	44.0	9.0
s_wid_o	7.3	6.8	0.0	40.0	10.0

in transportation safety studies in recent years (Das and Sun, 2016; Tsala et al., 2020; Yıldırım et al., 2019; Jalayer et al., 2018; Das et al., 2018). A very simplified version of 'factor analysis of mixed data' is described below which is mostly based on Pagès' book (Pagès, 2014).

Consider that we have I individual data entries with individual i entry is associated with a weight p_i so that $\sum_i p_i = 1$. We can assume that the individuals carry the same wights so that $p_i = 1/I \ \forall i$.

K_1 represents standardized (reduced and centered) quantitative variables $\{k = 1,K_1\}$ and Q represents qualitative variables $\{q = 1,Q\}$ in which the q th variable depicts K_q categories $\{k_q = 1,K_q\}$. Here, the total number of categories can be expressed as $\sum_q K_q = K_2$ by denoting p_{k_q} as the proportion of individuals possessing category k_q. Assume that $K = K_1 + K_2$ the total number of quantitative variables and indicator variables.

Consider $\mathbb{R}^I$ as the space of functions on I. The diagonal metric of the weights of the individuals (denoted as D) is endowed with this space:

$$D(i,j) = \begin{cases} 0 \ if \ j \neq i \\ p_i \ if \ j = i \end{cases}$$

One can explain the influence of a variable by the subspace dimensions it produces. We can express in space $\mathbb{R}^I$:

- A quantitative variable by a vector associated with an inertia of 1.
- A qualitative variable with K_q by K_q vectors by generating a subspace area E_q with having a dimension of $K_q - 1$.

4. Results

As mentioned earlier, the study is designed based on rural and urban roadways separately. This section provides results and discussion for these two facility types in separate sections.

4.1. PUWD behavior on rural roads

Table 7 lists the eigenvalue and percentage of variance measures for three rural cases. Measures for the top five dimensions are listed in this table. For rural case 1, the top two dimensions (axes) explain about 35.0

Table 7
Eigenvalue and Percentage of Variance by Top Five Dimensions (Rural).

Dimension or Axis	Case 1		Case 2		Case 3	
	Eigenvalue	Percentage of variance	Eigenvalue	Percentage of variance	Eigenvalue	Percentage of variance
Dim 1	5.2825	25.1547	4.9214	23.4352	4.7033	22.3966
Dim 2	2.0576	9.7979	2.1740	10.3525	2.3771	11.3194
Dim 3	1.7283	8.2301	1.7140	8.1621	1.6280	7.7522
Dim 4	1.3809	6.5758	1.3155	6.2642	1.2939	6.1613
Dim 5	1.1934	5.6829	1.1617	5.5317	1.2473	5.9398

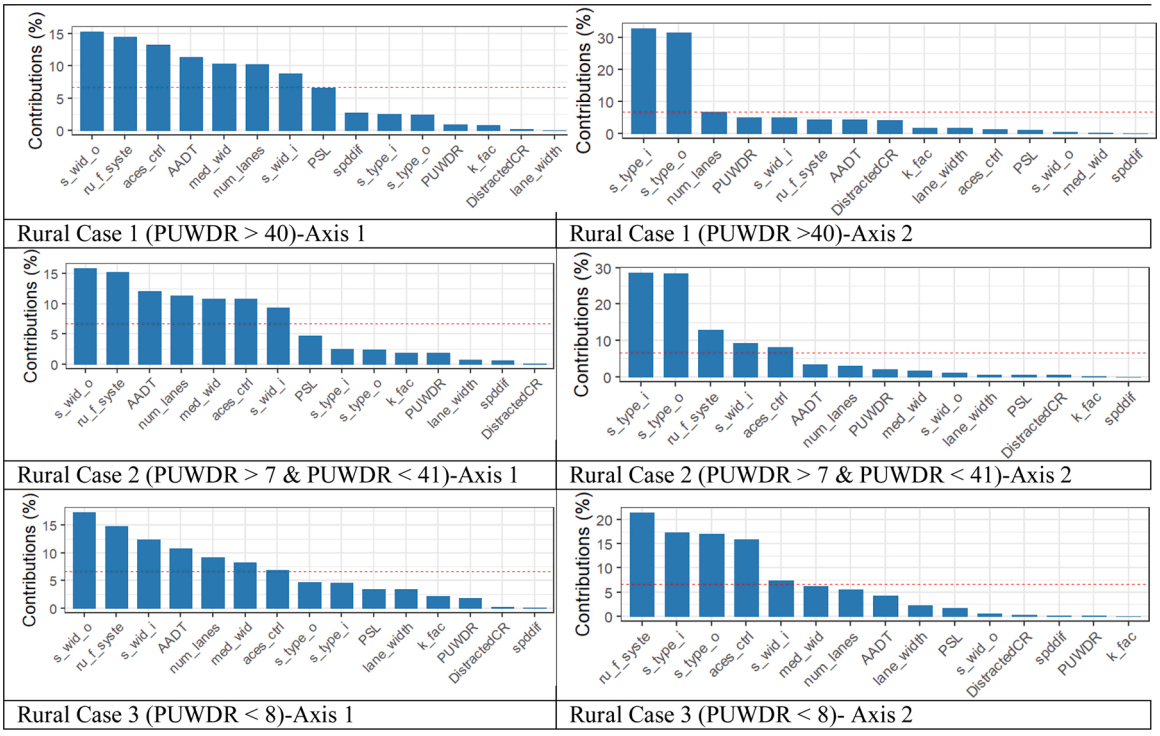

Fig. 2. Variable Contributions (Rural).

Table 8
Contribution and cos^2 Measures by First Plane (Rural).

Variable	Case 1		Case 2		Case 3	
	Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2
Contribution on each dimension						
AADT	**11.3311***	4.2603	**12.0687**	3.4229	**10.7740**	4.2595
med_wid	**10.3532**	0.2442	**10.8095**	1.6016	8.2510	6.1852
num_lanes	10.2744	**6.7935**	11.3420	2.9724	**9.2057**	5.5166
lane_width	0.0320	1.6457	0.7037	0.6098	3.3614	2.1891
k_fac	0.7702	1.7733	1.8972	0.0681	2.1868	0.0027
s_wid_i	8.7909	**4.8533**	9.2836	**9.1933**	**12.4100**	**7.3662**
s_wid_o	**15.3476**	0.4539	**15.8909**	1.1456	**17.3316**	0.5712
PSL	6.5837	0.9588	4.6955	0.5549	3.4121	1.7122
spddif	2.7080	0.0664	0.6508	0.0197	0.1431	0.1478
DistractedCR	0.2176	4.1358	0.0362	0.5132	0.1796	0.2204
PUWDR	0.9214	**5.0362**	1.8479	1.9979	1.8545	0.0769
aces_ctrl	**13.2946**	1.1662	**10.7683**	**8.0848**	6.9028	**15.9505**
s_type_i	2.4719	**32.8468**	2.4524	**28.6528**	4.5320	**17.3052**
s_type_o	2.4256	**31.4868**	2.3749	**28.3298**	4.6377	**17.0020**
ru_f_syste	**14.4777**	4.2789	**15.1785**	**12.8333**	**14.8178**	**21.4945**
cos^2 (Quality representation on the first plane)						
AADT	**0.3583**	0.0077	**0.3528**	0.0055	**0.2568**	0.0103
med_wid	**0.2991**	0.0000	**0.2830**	0.0012	**0.1506**	0.0216
num_lanes	**0.2946**	**0.0195**	**0.3116**	0.0042	**0.1875**	0.0172
lane_width	0.0000	0.0011	0.0012	0.0002	0.0250	0.0027
k_fac	0.0017	0.0013	0.0087	0.0000	0.0106	0.0000
s_wid_i	0.2156	**0.0100**	**0.2087**	**0.0399**	**0.3407**	**0.0307**
s_wid_o	**0.6573**	0.0001	**0.6116**	0.0006	**0.6645**	0.0002
PSL	0.1210	0.0004	0.0534	0.0001	0.0258	0.0017
spddif	0.0205	0.0000	0.0010	0.0000	0.0000	0.0000
DistractedCR	0.0001	0.0072	0.0000	0.0001	0.0001	0.0000
PUWDR	0.0024	**0.0107**	0.0083	0.0019	0.0076	0.0000
aces_ctrl	**0.2466**	0.0003	0.1404	**0.0154**	0.0527	**0.0719**
s_type_i	0.0171	**0.4568**	0.0146	**0.3880**	0.0454	**0.1692**
s_type_o	0.0164	**0.4197**	0.0137	**0.3793**	0.0476	**0.1633**
ru_f_syste	0.0975	0.0013	0.0930	**0.0130**	0.0810	**0.0435**

Note: *Top five highest values are shown in bold numbers.

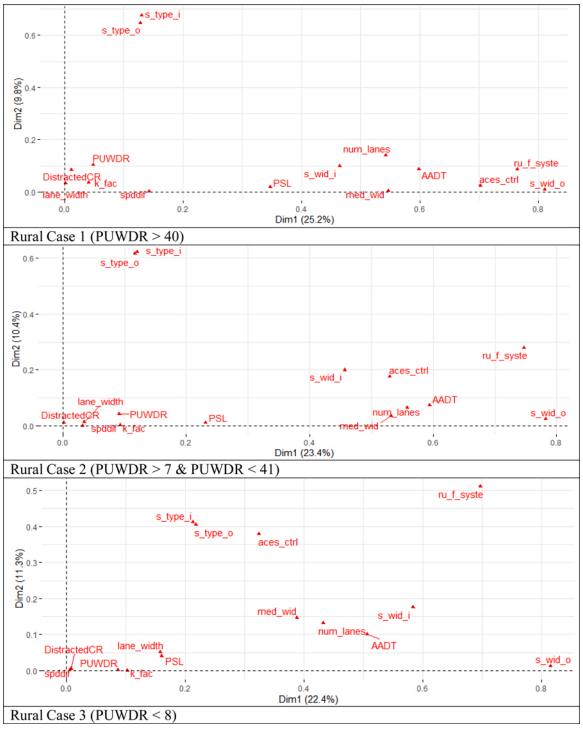

Fig. 3. Coordinates of Variables (Rural).

% of the variance. For the other two cases, the top two dimensions explain about 33.8 % and 33.7 % variance, respectively. For all three cases, the top 5 dimensions explain more than 53 % of the variance. The eigenvalue measures are high in dimension 1 compared to other dimensions. The first plane (based on dimension 1 and dimension 2) explains the highest variance when compared with other planes, the current analysis is limited to the first plane only.

Fig. 2 illustrates the distribution of the variables by the top two dimensions. Shoulder width, rural functional system, number of lanes, and AADT are the top five variables in the rural cases. It is common that rural roadways with a shoulder may encourage drivers to use the phone because the existence of a shoulder could provide the driver with a sense of security. Without a shoulder, drivers could drive more cautiously. Less lanes can be associated with lower traffic volume, which is common for rural roadways. Less challenging roadways are occasionally associated with driver inattention, which may result in PUWD occurrences. Access control is the third most important factor in the rural case 1 (PUWD $>$ 40) category, and it is not as significant for the other two categories. In this case, full access control means less traffic interruption. Without the concern of sudden interruption, the drivers feel protected and become less cautious while driving, which could explain why the higher functional class roadways with access control are the location where the most PUWD events occur.

For dimension 2, shoulder type is the most significant variable in all three cases. It is also found that the distracted crash count is positioned in the top variables in dimension 2, specifically for the rural case 1. It indicates that the distracted crash count has some association with high PUWD occurrences. Some studies found that phone use related crashes were under-reported. From this finding, we can indicate that the frequency of PUWD events on certain roadways can be associated with the frequency of distracted crashes.

Table 8 lists the contribution and $\cos^2$ measures of the variables for each of the two dimensions. Contribution measures indicate the variable contribution for each of the dimensions. For rural case 1, the top five contributions are associated with outside shoulder width, roadway functional class, access control, AADT, and median width. The number of lanes is the sixth variable contributing the most on dimension 1, which has a similar contribution as the median width. For dimension 2, the top five contributions are associated with inside shoulder type, outside shoulder type, number of lanes, total PUWD event count, and inside shoulder width. It is obvious that the patterns of the contributions for the other two cases are also similar for the top five contributions on the same dimensions. Distracted crashes show a higher contribution to dimension 2 than dimension 1. PUWD shows a higher contribution to dimension 2 for case 1.

The summation of $\cos^2$ measures on the first plane indicates the quality representation. Variables with larger values contribute a relatively larger portion to the total distance, which indicates the components are more important for that case (Abdi and Williams, 2010). For rural case 1, top five variables with high quality representation are outside shoulder width (total $\cos^2 = 0.6573$), inside shoulder type (total $\cos^2 = 0.5240$), outside shoulder type (total $\cos^2 = 0.4906$), AADT (total $\cos^2 = 0.3583$), median width (total $\cos^2 = 0.2991$), number of lanes (total $\cos^2 = 0.2946$), and access control (total $\cos^2 = 0.2446$). For the other two cases, outside or inside shoulder width is also the variable with high representation on the first plane.

Fig. 3 is the biplot of all variables. There are generally 3 clusters in each plot developed for each rural case. The general interpretation is the closer the locations then the closer association of co-occurrences. Cluster 1 is associated with shoulder types. Cluster 2 is associated with shoulder width, median width, AADT, number of lanes, access control, post speed limit, and functional class. Cluster 3 is associated with total PUWD counts per mile, distracted crashes per mile, speed difference lane width, and peak hour factors. There exists some difference in these plots. In rural case 1, the posted speed limit is in cluster 2, and this variable joined cluster 3 in case 2 and 3. The association between factors in each cluster are strong and close for Case 1. The magnitude of these associations becomes smaller (distance higher) as the number of PUWD events goes down in the other two cases.

Fig. 4 contains the biplots of the categorical variables. It is important to note that the categorical variables are strictly related to roadway function class and other roadway or geometric properties. The results are needed to be interpreted based on the geometric and functional potteries of these roadways. The clusters and the locations of the variable categories (for example, R1 is a variable category of the 'roadway functional class' variable) indicate the functional properties of these roadways vary in terms of other operational and exposure quantitative variables. For example, on the left-hand side of the plots, the cluster contains R1, R2, and full access control. The full access control characteristic is normally associated with interstate and principal arterial roadways (R1 and R2), which is obvious.

Meanwhile, partial access control is slightly further from this cluster, and no access control is way further from this cluster. That indicates roadways with R1 and R2 functional classes are mostly with full access controls. It is also obvious to observe that the cluster on the right-hand side of the plots containing features like the absence of outside shoulder, absence of inside shoulder, and R7 functional class (local roadways) is the most further from the R1/R2 cluster. The reason is that the interstate highway with full control and local roadways with no shoulder are two groups of variables that are not likely to co-exist. The biplot shows that the presence of no access control, no shoulder, R5, R6, and R7 are on the opposite side of full access control (based on the Y-axis). The patterns of the clusters are similar across three cases.

Fig. 5 shows the correlation plot of the quantitative variables for three rural cases. The number of lanes, AADT, median width, and shoulder width are the most significant quantitative variables, which are associated with PUWD events per mile. The number of lanes and AADT are correlated. The arrows have a similar direction and similar length. It was also found that the contribution of distracted crash counts goes lower as the PUWDR goes lower. A similar trend can be found for speed difference or variation.

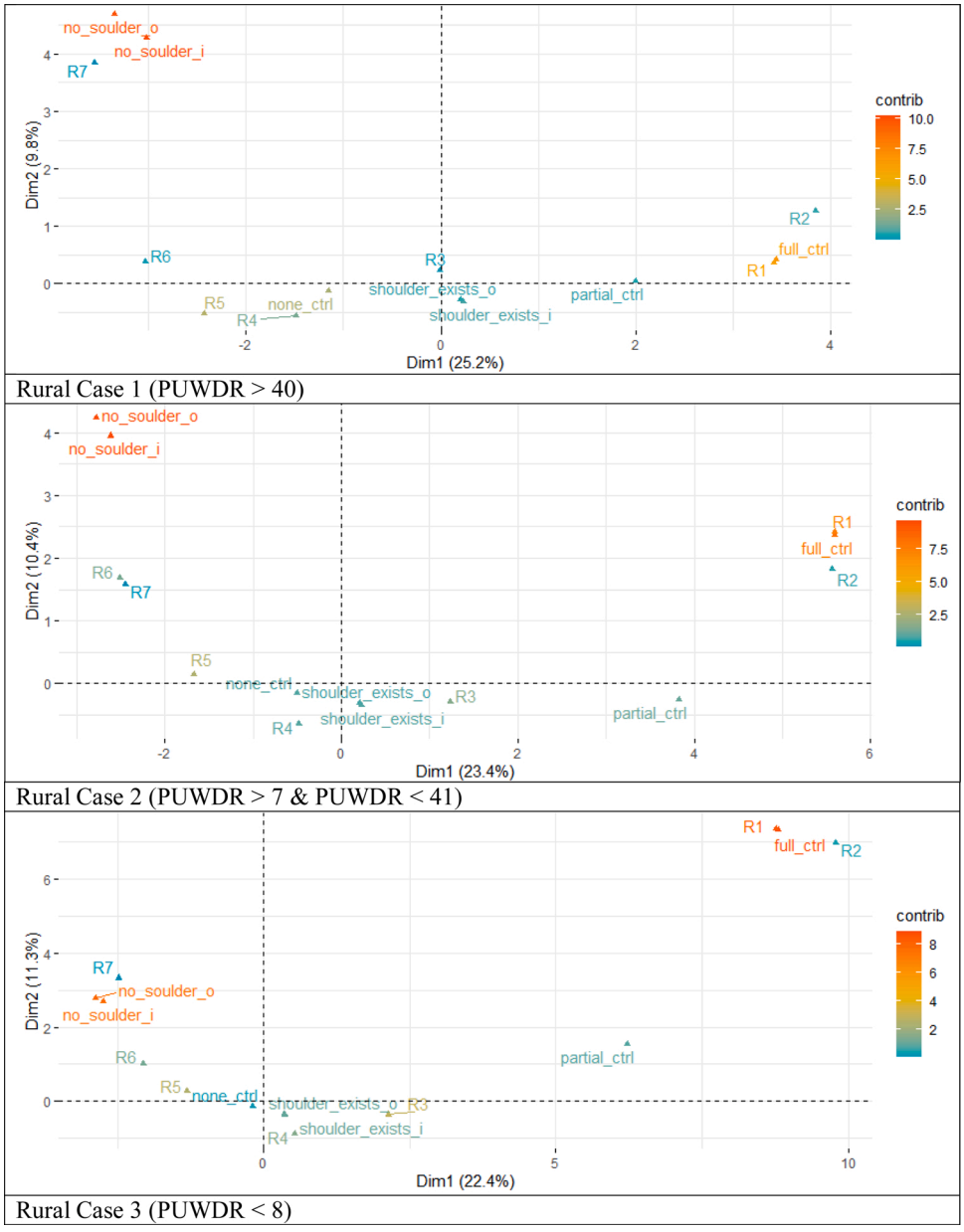

Fig. 4. Co-ordinates of Categorical Variables (Rural).

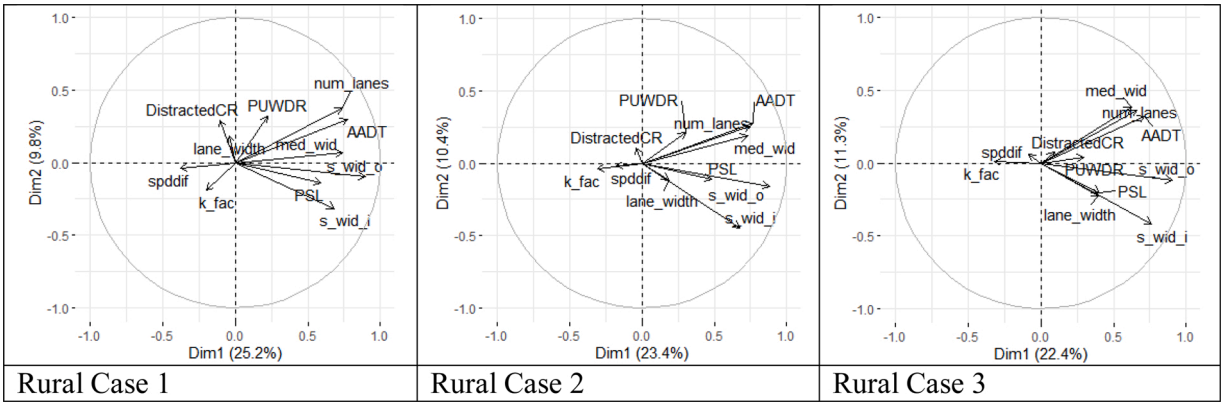

Fig. 5. Co-ordinates of Quantitative Variables (Rural).

Table 9
Eigenvalue and Percentage of Variance by Top Five Dimensions (Urban).

Dimension or Axis	Urban Case 1		Urban Case 2		Urban Case 3	
	Eigenvalue	Percentage of variance	Eigenvalue	Percentage of variance	Eigenvalue	Percentage of variance
Dim 1	5.5939	29.4415	5.4254	25.8354	4.2765	21.3827
Dim 2	2.0791	10.9425	2.0679	9.8470	2.6900	13.4498
Dim 3	1.4402	7.5799	1.3643	6.4966	1.5053	7.5264
Dim 4	1.2488	6.5724	1.3055	6.2167	1.2452	6.2258
Dim 5	1.1217	5.9035	1.1569	5.5092	1.2088	6.0438

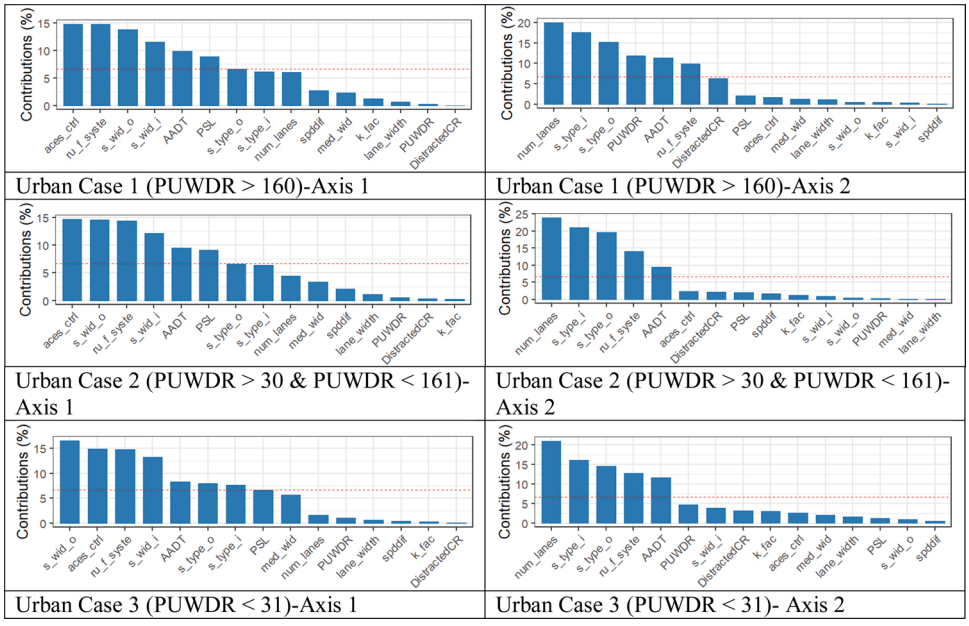

Fig. 6. Variable contributions (Urban).

4.2. PUWD behavior on urban roads

Table 9 lists the eigenvalue and percentage of variance measures for three urban cases. Measures for the top five dimensions are listed in this table. For urban case 1, the top 2 dimensions explain 40.4 % of the variance. For the other two cases, the top two dimensions explain 35.7 % and 34.8 % variance, respectively. For all three cases, the top 5 dimensions explain more than 50 % of the variance. The eigenvalue measures are high in dimension 1 compared to other dimensions. The first plane (based on dimension 1 and dimension 2) explains the highest variance when compared with other planes, the current analysis is limited to the first plane only.

Fig. 6 illustrates the distribution of the variables by the top two dimensions. Access control, functional class, shoulder width, and AADT are the top five variables in urban cases. It is common to find that roadways with access control and the presence of shoulder may intrigue more PUWD behaviors due to the roadway with access control and wide shoulder, often granting drivers a sense of security. Roadways with low AADT often encourage drivers to use their phones while driving because drivers generally become less cautious while fewer vehicles around. The speed limit variable has the sixth-highest contribution to dimension 1. Comparing with rural cases, the speed difference poses a higher contribution to urban cases. Except for rural case 1, the other two rural cases have the speed difference contributing the least among all variables. The speed difference is the speed difference between the average speed of the vehicles of all PUWD event of the road segment and the posted speed limit of that segment. In urban settings, more traffic and more interruptions like traffic lights are expected. The speed variation is higher than on the rural roadways. More PUWD events on urban roads are triggered by this speed difference caused by interruptions, such as congestions.

For dimension 2, the number of lanes and shoulder type are the most significant variables in all three cases. It is also found that the distracted crash count is positioned in the top variables in dimension 2 across three cases. This finding is aligned with the finding in the rural cases – the distraction related crash count has clear associations with high PUWD occurrences.

Table 10 lists the contribution and $\cos^2$ measures of the variables for each of the two dimensions. Contribution measures indicate the variable contribution for each of the dimensions. The summation of $\cos^2$ measures on the first plane indicates the quality representation. The quality representation in Table 10 mostly implicates similar importance of each component on each case at each dimension. However, the posted speed limit standout for the first dimension of case 1. That indicates the post speed limit contributing to the PUWD behavior significantly in terms of quality representation. This could echo with existing research, which indicates that the PUWD events have mostly occurred on roadways with high-speed limits, and the frequency of PUWD event increases in urban settings where the speed variation is relatively high (Iio et al., 2020).

Fig. 7 shows 3 clusters. Variables in the same cluster indicate the close association of co-occurrences. The cluster at the left-hand side corner has total PUWD event counts, distracted crash count, lane width, median width, peak hour factor, and speed difference. This cluster firstly indicates the possible association between the PUWD events and distracted crash events, and secondly points out the speed difference/variation could be a driving cause for the PUWD behavior and distracted crash occurrence. The cluster on the top side contains the number of lanes, inside shoulder type, and outside shoulder type. The cluster on the

Table 10
Contribution and cos^2 Measures by First Plane (Urban).

Variable	Case 1		Case 2		Case 3	
	Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2
Contribution on each dimension						
AADT	**9.9107**	**11.3105**	**9.5444**	**9.4451**	**8.2700**	**11.6192**
med_wid	2.3812	1.3377	3.4363	0.1507	5.6797	2.0751
num_lanes	6.0241	**20.0135**	4.4373	**24.0229**	1.6451	**21.0661**
lane_width	0.6905	1.1676	1.1755	0.1338	0.6505	1.6797
k_fac	1.3015	0.4870	0.3012	1.2292	0.2603	3.0500
s_wid_i	**11.6028**	0.3430	**12.1039**	0.9308	**13.2952**	3.8481
s_wid_o	**13.8561**	0.5393	**14.6140**	0.4777	**16.5938**	0.9346
PSL	8.8856	2.0471	9.1047	1.9849	6.6842	1.2309
spddif	2.6966	0.1456	2.1421	1.7764	0.3817	0.4762
DistractedCR	0.0005	6.3474	0.4149	2.2378	0.0969	3.1122
PUWDR	0.2322	**11.8260**	0.5606	0.3198	1.0338	4.6587
aces_ctrl	**14.8334**	1.7557	**14.7169**	2.4045	**14.8715**	2.6447
s_type_i	6.2009	**17.5764**	6.4340	**21.1224**	7.7096	**16.1539**
s_type_o	6.6219	**15.2273**	6.6038	**19.6769**	8.0335	**14.5886**
ru_f_syste	**14.7619**	9.8759	**14.4104**	**14.0870**	**14.7943**	**12.8619**
cos^2 (Quality representation on the first plane)						
AADT	**0.3074**	**0.0553**	**0.2681**	**0.0381**	**0.1251**	**0.0977**
med_wid	0.0177	0.0008	0.0348	0.0000	0.0590	0.0031
num_lanes	0.1136	**0.1731**	0.0580	**0.2468**	0.0049	**0.3211**
lane_width	0.0015	0.0006	0.0041	0.0000	0.0008	0.0020
k_fac	0.0053	0.0001	0.0003	0.0006	0.0001	0.0067
s_wid_i	**0.4213**	0.0001	**0.4312**	0.0004	**0.3233**	0.0107
s_wid_o	**0.6008**	0.0001	**0.6287**	0.0001	**0.5036**	0.0006
PSL	**0.2471**	0.0018	**0.2440**	0.0017	0.0817	0.0011
spddif	0.0228	0.0000	0.0135	0.0013	0.0003	0.0002
DistractedCR	0.0000	0.0174	0.0005	0.0021	0.0000	0.0070
PUWDR	0.0002	**0.0605**	0.0009	0.0000	0.0020	0.0157
aces_ctrl	**0.3443**	0.0007	**0.3188**	0.0012	**0.2022**	0.0025
s_type_i	0.1203	**0.1335**	0.1219	**0.1908**	0.1087	**0.1888**
s_type_o	0.1372	**0.1002**	0.1284	**0.1656**	**0.1180**	**0.1540**
ru_f_syste	0.1705	0.0105	0.1019	**0.0141**	0.0801	**0.0239**

right-hand side corner contains functional class, access control, and shoulder width. The posted speed limit and AADT belongs to this cluster for urban case 1 and 2. The AADT belongs to the top-side cluster, and the posted speed limit belongs to the left-hand side cluster for the urban case 3.

The biplots in Fig. 8 presents the detailed correlations among the categorical variables. The findings for urban cases are similar to the rural cases. The highest functional class U1 and U2 are clustered with full access control, which is obvious. Another cluster at the left-hand side (no shoulder) has the longest distance with the cluster with the full access control variable.

Fig. 9 shows the correlation plot of the quantitative variables for three urban cases. The number of lanes, AADT, median width, shoulder width, and posted speed limit are the most significant quantitative variables, which are associated with PUWD events. The number of lanes and AADT are correlated, as demonstrated in rural cases. Both distracted crash count and total PUWD event count are on the positive side of the second dimension. The correlation between them is clear. The contribution of distracted crash count remains the same when the PUWD event count goes lower. It is also interesting to observe the contribution of the speed difference and the posted speed limit becomes less when the total PUWD event count goes lower.

5. Findings

5.1. Rural cases

From the roadway geometrical perspective, rural cases are dominated by the shoulder width, median width, and the number of lanes. The existence of shoulder and median and wide shoulder and median could encourage PUWD behavior since these geometric features provide a safety buffer for drivers, which grants drivers a sense of security. In the biplots, the number of lanes is in the same cluster with a high functional class (R1-interstate highway and R2-freeway) and full access control. From an operational perspective, rural roadways with high functional class, fewer traffic and less interruptions due to the access control could also be a combination that triggers PUWD behavior. Posted speed limit and speed difference presented a relatively high contribution among all factors in case 1, where PUWDR > 40, but not in other two rural cases. Speed difference, also called speed variation, is a variable that describes the roadway operational features. Higher speed variation generally means the roadway is frequently congested. This finding proves that congestion often encourages drivers to use phones. In other words, drivers are more likely to use phones while driving under a congested situation. The study also found in rural case 1, the distracted crash count has a medium contribution. Moreover, the biplots repeatedly show the distracted crash count and total PUWD event are in the same cluster, which indicates the co-occurrence of these two variables. This finding indicates the hidden correlations between the amount of PUWD events and reported distracted crash counts. This correlation is obvious in the biplots, especially for rural cases 1, where the most PUWD events occurred. The co-ordinate plots also suggest the quality of co-occurrence of the distracted crash count diminishes when the number of PUWD events goes lower (from case 1 to case 3). This indicates that roadways with a large amount of PUWD events often have higher distracted crash occurrences. However, this relationship may not as strong in the roadways with fewer PUWD events. On the roadways with less PUWD events, the distracted crashes could be triggered by many other factors rather than the phone using behavior.

5.2. Urban cases

From the geometrical perspective, shoulder width, shoulder type, number of lanes, and functional class dominate the top five contributions across three urban cases. From an operational perspective, access control, AADT, and posted speed limit present above-average

Fig. 7. Co-ordinates of Variables (Urban).

contributions as well. Access control has the highest contribution of urban case 1 and urban case 2, and it has the second-highest contribution for urban case 3. In urban settings, access control is the most critical factor that grants drivers the feeling of safety, since it limits the possible interruptions and drivers become less cautious. The PUWD behavior becomes even more frequent if the roadway has a wide shoulder. The biplots also demonstrate that roadways with functional class U1 and U2, which are interstate highway and freeways, often co-exists with the full access future feature and presence of shoulder. These roadways are the place where the most PUWD events occurred. The biplot also suggests that local roadways with no shoulder presence and no access control could be the places where PUWD events happen. These locations often have constant traffic interruptions from congestions and traffic control devices such as traffic lights. These interruptions lead to a slow movement, which would encourage phone using behaviors. The distracted crash count shows the strong quality of representation in co-ordinate plots across three urban cases. On urban roadways, the distracted crash count and total PUWD events are positively correlated. More PUWD events implicate a higher probability of more distracted crashes.

5.3. *Comparisons between rural and urban cases*

There are similarities and differences in the impacts of geometrical and operational variables on the frequency of PUWD events in the urban and rural roadway settings. The geometrical features such as shoulder, median, number of lanes, and operational features like AADT, access control, and posted speed limit dominate both urban and rural cases. In general, roadways with relatively higher functional classes, full access control, wide shoulder width, and median width encourages PUWD behavior. Meanwhile, roadways with relatively low-speed limits and high-speed variations could also have a high frequency of PUWD event occurrences. The differences between urban and rural cases are also worthy of mention. Access control is the most contributed variable in urban cases, but it only found being listed in the top five contributors in rural case 1. This finding indicates that access control provides a strong sense of security for drivers in urban roadways. It also found that speed

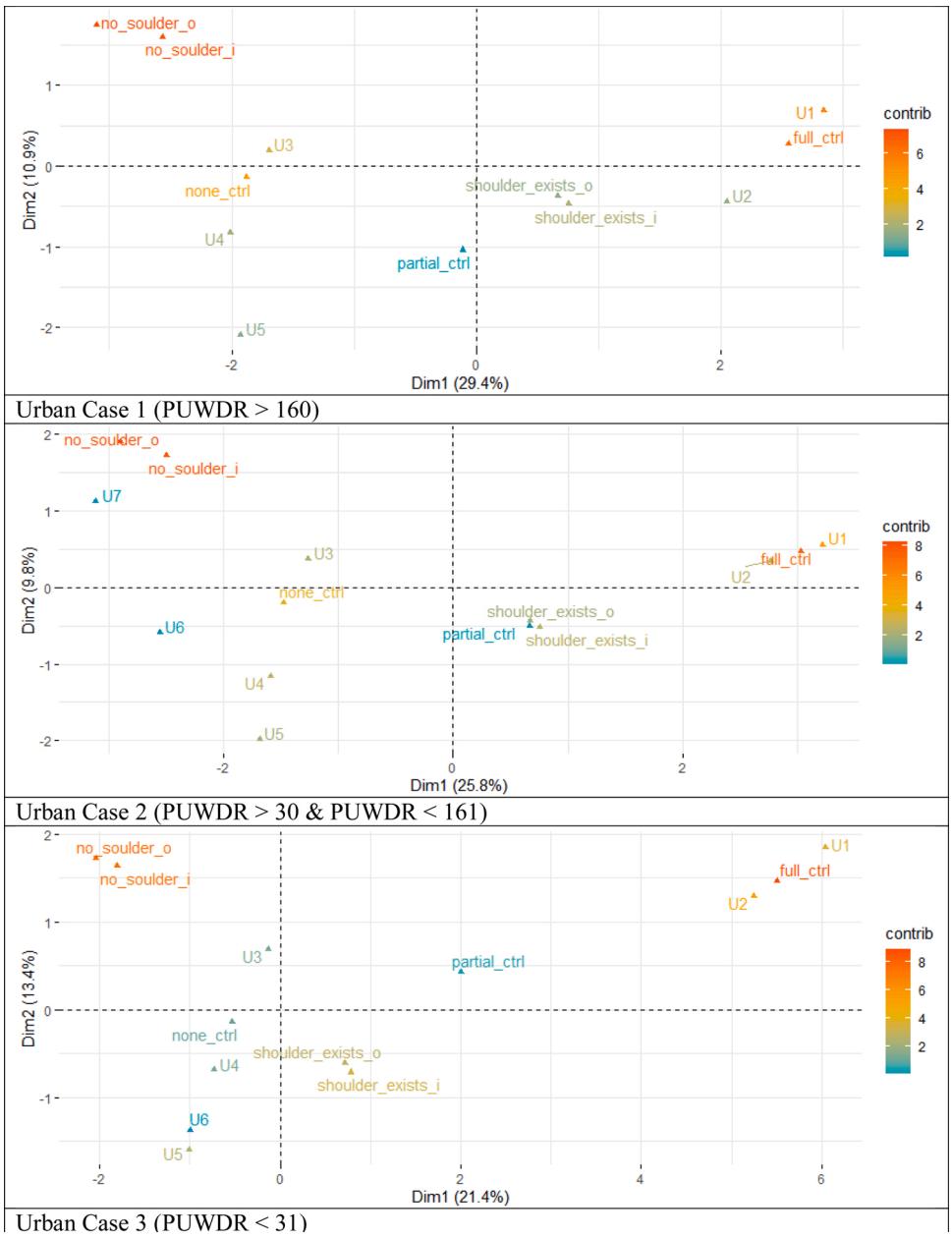

Fig. 8. Co-ordinates of Categorical Variables (Urban).

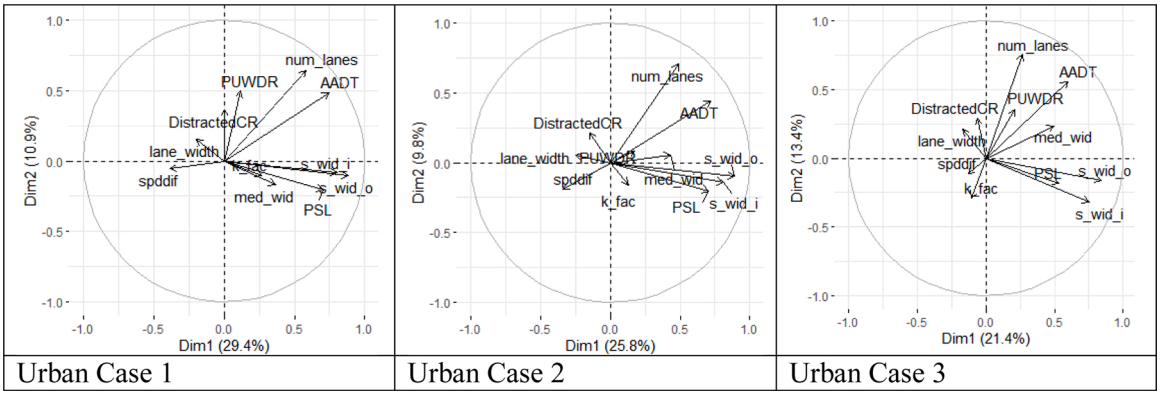

Fig. 9. Co-ordinates of Quantitative Variables (Urban).

variation on urban roadways plays a bigger role than on the rural roads in terms of triggering the PUWD behavior. The results also demonstrate the strong correlations between distracted crashes and PUWD behavior on the urban roadways. The magnitude of this relationship decays on the rural roadways while the PUWD event count drops. The rural roadways with a low frequency of PUWD events could be these rural roadways with relatively low functional class and less traffic in the rural area. The crash frequency in these areas is generally low due to the nature of low traffic. Thus, the frequency of distracted crashes is even lower. That could explain why these rural cases with less PUWD events do not show a strong correlation between PUWD events and distracted crashes.

6. Conclusions

This study utilized an unsupervised learning method, known as factor analysis, on a unique distracted driving dataset with distraction-related events to understand PUWD from the road and operational perspectives. The analysis is conducted through the cases on urban and rural roadways separately. In this way, the results identified the most influential factors for these roadways separately. Then, more interesting insights have been brought out through the comparison between the findings from urban and rural cases.

In summary, the presence of shoulder, median, higher speed limit and access control on the roadways with higher functional class rural or urban roadways could encourage more PUWD events. The study also found the roadway with a relatively low-speed limit could also have high occurrences of the PUWD event if the speed variation was high, in other words, if that road segment was congested. The results also confirmed the correlations between the frequency of PUWD events and the frequency of distracted crashes. This relationship is strong on urban roadways. For rural roadways, this correlation only strong on the roadways with a large amount of PUWD events. The relationship is not strong if the frequency of the PUWD event is low.

This research is the first of its kind to understand the PUWD behavior from the road and operational perspectives. It provides several intuitive insights into understanding PUWD behaviors, especially in the time of all smartphones getting smarter and national surveys showing the increasing amount of PUWD events. Moreover, it is proven that more PUWD events lead to more distracted crashes. The findings of this research could help transportation agencies to identify appropriate countermeasures to prevent the preventable distracted crashes causing by the PUWD behavior. For example, more visible signs and law enforcement should be placed at these urban roads with full access control and wide shoulder and medians, if these urban roadways already have higher distracted crash occurrences comparing with other urban roadways. Additionally, the roadways with high-speed variations and also being identified as high distracted crash locations could be the roads that need more attention from transportation agencies. The countermeasures could improve traffic conditions or more strict law enforcement.

The current study has several limitations. First, even this unique dataset could provide valuable insights into PUWD behavior, which has never been revealed before, but we have to admit that this unique dataset might not well represent the elder population. The reason is the PUWD data were collected through a phone application. It would be reasonable to assume that the elder population has less incentive to download this application for the reward points. Second, the current study used the most recent roadway inventory data of Texas. But there is still many missing information in the dataset. The study would reveal more interesting findings if the road inventory data could be more complete. For example, there is one attribute in the RHiNO named "traffic-signal-type," indicating if the intersection on this segment is uncoordinated fixed time or coordinated real-time adaptive. This would help us reveal the benefit of having adaptive signalized intersections on reducing PUWD events. Unfortunately, the data of this column are missing. Third, even with the massive number of users and data points, many RHiNO road segments are not covered by the users in the database, these road segments without PUWD events are excluded. This may pose some concerns. With additional data available in the future, this research may incorporate more RHiNO segments or all segments to generate improved results. Future research could focus on incorporating PUWD data into Safety Performance Function (SPF) models. The safety models for distraction-affected crashes are lacking due to the availability issue of the data source. As these novel data resources become available, building SPF models for distraction-affected crashes is possible.

Disclaimer

The contents of this paper reflect the views of the authors and not the official views or policies of the Texas Department of Transportation (TxDOT) or the private data service provider.

Author contribution statement

The authors confirm the contribution to the paper as follows: study conception and design: X. Kong, and S. Das; data collection: X. Kong; analysis and interpretation of results: X. Kong, S. Das, and H. Zhou, draft manuscript preparation: X. Kong, S. Das, H. Zhou, and Y. Zhang. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Compting Interest

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

Acknowledgments

The authors thank the private data service provider for supplying this valuable dataset. The authors like to thank two anonymous reviewers for their excellent suggestions.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2 (4), 433–459.
Al-Darrab, I.A., Khan, Z.A., Ishrat, S.I., 2009. An experimental study on the effect of mobile phone conversation on drivers' reaction time in braking response. J. Safety Res. 40 (3), 185–189.
Alm, H., Nilsson, L., 1995. The effects of a mobile telephone task on driver behaviour in a car following situation. Accid. Anal. Prev. 27 (5), 707–715.
Atwood, J., Guo, F., Fitch, G., Dingus, T.A., 2018. The driver-level crash risk associated with daily cellphone use and cellphone use while driving. Accid. Anal. Prev. 119, 149–154.
Backer-Grøndahl, A., Sagberg, F., 2011. Driving and telephoning: relative accident risk when using hand-held and hands-free mobile phones. Saf. Sci. 49 (2), 324–330.
Bingham, C.R., 2014. Driver distraction: a perennial but preventable public health threat to adolescents. J. Adolesc. Health 54 (5), S3–S5.
Bommer, W.H., 2018. Observational Study of Handheld Cellphone and Texting Use Among California driver. 2018 Summary Report. California Office of Traffic Safety, p. 12.
Bowden, V.K., Loft, S., Wilson, M.D., Howard, J., Visser, T.A., 2019. The long road home from distraction: investigating the time-course of distraction recovery in driving. Accid. Anal. Prev. 124, 23–32.
Brace, C.L., Young, K.L., Regan, M.A., 2007. Analysis of the Literature The Use of Mobile Phones While … (No. 2007:35 1401-9612). Monash University Accident Research Center, Australia.
Brusque, C., Alauzet, A., 2008. Analysis of the individual factors affecting mobile phone use while driving in France: socio-demographic characteristics, car and phone use in professional and private contexts. Accid. Anal. Prev. 40 (1), 35–44.
Chen, Y.-L., 2007. Driver personality characteristics related to self-reported accident involvement and mobile phone use while driving. Saf. Sci. 45 (8), 823–831.
Chen, Y., Fu, R., Xu, Q., Yuan, W., 2020. Mobile Phone Use in a Car-Following Situation: Impact on Time Headway and Effectiveness of Driver's Rear-End Risk Compensation Behavior via a Driving Simulator Study. Int. J. Environ. Res. Public Health 17 (4), 1328.
Choudhary, P., Velaga, N.R., 2017. Mobile phone use during driving: effects on speed and effectiveness of driver compensatory behaviour. Accid. Anal. Prev. 106, 370–378.

Claveria, J.B., Hernandez, S., Anderson, J.C., Jessup, E.L., 2019. Understanding truck driver behavior with respect to cell phone use and vehicle operation. Transp. Res. Part F Traffic Psychol. Behav. 65, 389–401.
Das, S., Sun, X., 2015. Factor association with multiple correspondence analysis in vehicle–pedestrian crashes. Transportation Research Record: Journal of the Transportation Research Board 2519, 95–103.
Das, S., Sun, X., 2016. Association knowledge for fatal run-off-road crashes by Multiple Correspondence Analysis. IATSS Research 39 (2), 146–155.
Das, S., Avelar, R., Dixon, K., Sun, X., 2018. Factor association with multiple correspondence analysis in vehicle–pedestrian crashes. Accid. Anal. Prev. 11, 43–55.
Dingus, T.A., Guo, F., Lee, S., Antin, J.F., Perez, M., Buchanan-King, M., Hankey, J., 2016. Driver crash risk factors and prevalence evaluation using naturalistic driving data. PNAS.
Fitch, G.M., Bartholomew, P.R., Hanowski, R.J., Perez, M.A., 2015. Drivers' visual behavior when using handheld and hands-free cell phones. J. Safety Res. e29–108. Strategic Highway Research Program (SHRP 2) and Special Issue: Fourth International Symposium on Naturalistic Driving Research 54, 105.
Foss, R.D., Goodwin, A.H., 2014. Distracted driver behaviors and distracting conditions among adolescent drivers: findings from a naturalistic driving study. J. Adolesc. Health 54 (5 Suppl), S50–60.
Foss, R.D., Goodwin, A.H., McCartt, A.T., Hellinga, L.A., 2009. Short-term effects of a teenage driver cell phone restriction. Accid. Anal. Prev. 6.
Goodwin, A.H., O'Brien, N.P., Foss, R.D., 2012. Effect of North Carolina's restriction on teenage driver cell phone use two years after implementation. Accid. Anal. Prev. 48, 363–367.
Gras, M.E., Cunill, M., Sullman, M.J.M., Planes, M., Aymerich, M., Font-Mayolas, S., 2007. Mobile phone use while driving in a sample of Spanish university workers. Accid. Anal. Prev. 39 (2), 347–355.
Grimberg, E., Botzer, A., Musicant, O., 2020. Smartphones vs. In-vehicle data acquisition systems as tools for naturalistic driving studies: a comparative review. Saf. Sci. 131, 104917.
Hallett, C., Lambert, A., Regan, M.A., 2012. Text messaging amongst New Zealand drivers: prevalence and risk perception. Transp. Res. Part F Traffic Psychol. Behav. 15 (3), 261–271.
Hickman, J.S., Hanowski, R.J., 2012. An assessment of commercial motor vehicle driver distraction using naturalistic driving data. Traffic Inj. Prev. 13 (6), 612–619.
Hoff, J., Grell, J., Lohrman, N., Stehly, C., Stoltzfus, J., Wainwright, G., Hoff, W.S., 2013. Distracted driving and implications for injury prevention in adults. J. Trauma Nurs. 20 (1), 31–34 quiz 35–36.
Holland, C., Rathod, V., 2013. Influence of personal mobile phone ringing and usual intention to answer on driver error. Accid. Anal. Prev. 50, 793–800.
IIHS, 2019. Cellular Phone Use and Texting While Driving Laws [WWW Document]. URL https://www.ncsl.org/research/transportation/cellular-phone-use-and-texting-while-driving-laws.aspx (accessed 11/24/2020)..
Iio, K., Guo, X., Lord, D., 2020. Examining Driver Distraction As a Function of Driving Speed: an Observational Study Using Disruptive Technology and Naturalistic Data.
Jalayer, M., Zhou, H., Das, S., 2018. Exploratory analysis of run-off-Road crash patterns. In: Alavi, A., Buttlaer, W. (Eds.), Data Analytics for Smart Cities. CRC Press, Boca Raton, FL.
Jamson, A.H., Westerman, S.J., Hockey, G.R.J., Carsten, O.M.J., 2004. Speech-based E-Mail and driver behavior: effects of an in-vehicle message system interface. Hum. Factors 46 (4), 625–639.
Kahn, C.A., Cisneros, V., Lotfipour, S., Imani, G., Chakravarthy, B., 2015. Distracted driving, a major preventable cause of motor vehicle collisions:"just hang up and drive". West. J. Emerg. Med. 16 (7), 1033.
Kong, X., Das, S., Jha, K., Zhang, Y., 2020. Understanding speeding behavior from naturalistic driving data: applying classification based association rule mining. Accid. Anal. Prev. 144, 105620.
Korpinen, L., Pääkkönen, R., 2012. Accidents and close call situations connected to the use of mobile phones. Accid. Anal. Prev. 45, 75–82.
Laberge-Nadeau, C., Maag, U., Bellavance, F., Lapierre, S.D., Desjardins, D., Messier, S., Saïdi, A., 2003. Wireless telephones and the risk of road crashes. Accid. Anal. Prev. 35 (5), 649–660.
Lamble, D., Kauranen, T., Laakso, M., Summala, H., 1999. Cognitive load and detection thresholds in car following situations: safety implications for using mobile (cellular) telephones while driving. Accid. Anal. Prev. 31 (6), 617–623.
Lamble, D., Rajalin, S., Summala, H., 2002. Mobile phone use while driving: public opinions on restrictions. Transportation 29 (3), 223–236.
Li, X., Oviedo-Trespalacios, O., Rakotonirainy, A., Yan, X., 2019. Collision risk management of cognitively distracted drivers in a car-following situation. Transp. Res. Part F Traffic Psychol. Behav. 60, 288–298.
Liu, B.-S., Lee, Y.-H., 2006. In-vehicle workload assessment: effects of traffic situations and cellular telephone use. J. Safety Res. 37 (1), 99–105.
McEvoy, S.P., Stevenson, M.R., McCartt, A.T., Woodward, M., Haworth, C., Palamara, P., Cercarelli, R., 2005. Role of mobile phones in motor vehicle crashes resulting in hospital attendance: a case-crossover study. BMJ 331, 428.
McEvoy, S.P., Stevenson, M.R., Woodward, M., 2006. Phone use and crashes while driving: a representative survey of drivers in two Australian states. Med. J. Aust. 185 (11–12), 630–634.
National Safety Council, 2013. Crashes Involving Cell Phones: Challenges of Collecting and Reporting Reliable Crash Data. National Safety Council Itasca, IL.
National Safety Council, 2014. National Safety Council Poll: 8 in 10 Drivers Mistakenly Believe Hands-free Cell Phones Are Safer Distracted Driving Awareness Month Campaign Focuses on Why Hands-free Is Not Risk-free. National Safety Council News Release.
NHTSA, 2019. Traffic Safety Facts: Driver Electronic Device Use in 2018 (Research Note No. DOT HS 812 818). U.S. Department of Transportation.
NHTSA, 2020. Distracted driving 2018 (No. DOT HS 812 926). Traffic Safety Facts. Washington, DC.
Onate-Vega, D., Oviedo-Trespalacios, O., King, M.J., 2020. How drivers adapt their behaviour to changes in task complexity: the role of secondary task demands and road environment factors. Transp. Res. Part F Traffic Psychol. Behav. 71, 145–156.
Oviedo-Trespalacios, O., Haque, M.M., King, M., Demmel, S., 2018. Driving behaviour while self-regulating mobile phone interactions: a human-machine system approach. Accid. Anal. Prev. 118, 253–262.
Oviedo-Trespalacios, O., King, M., Vaezipour, A., Truelove, V., 2019. Can our phones keep us safe? A content analysis of smartphone applications to prevent mobile phone distracted driving. Transp. Res. Part F Traffic Psychol. Behav. 60, 657–668.
Oviedo-Trespalacios, O., Truelove, V., King, M., 2020. "It is frustrating to not have control even though I know it's not legal!": A mixed-methods investigation on applications to prevent mobile phone use while driving. Accid. Anal. Prev. 137, 105412.
Pagès, J., 2014. Multiple Factor Analysis by Example Using R. CRC Press.
Papadimitriou, E., Argyropoulou, A., Tselentis, D.I., Yannis, G., 2019. Analysis of driver behaviour through smartphone data: the case of mobile phone use while driving. Saf. Sci. 119, 91–97.
Papantoniou, P., Kontaxi, A., Yannis, G., 2020. Investigating the correlation of mobile phone use with trip characteristics recorded through smartphone sensors, in: proceedings of the 5th Conference on sustainable Urban mobility. In: Presented at the 5th Conference on Sustainable Urban Mobility (Virtual CSUM2020). Thessaly, Greece, p. 11.
Petraki, V., Ziakopoulos, A., Yannis, G., 2020. Combined impact of road and traffic characteristic on driver behavior using smartphone sensor data. Accid. Anal. Prev. 144, 105657.
Qi, Y., Vennu, R., Pokhrel, R., 2020. Distracted Driving: A Literature Review. Illinois Center for Transportation/Illinois Department of Transportation.
Redelmeier, D.A., Tibshirani, R.J., 1997. Association between cellular-telephone calls and motor vehicle collisions. N. Engl. J. Med. 336 (7), 453–458.
Regev, S., Rolison, J., Feeney, A., Moutari, S., 2017. Driver distraction is an under-reported cause of road accidents: an examination of discrepancy between police officers' views and road accident reports, in: DDI2017 E-proceedings collection. The Fifth International Conference on Driver Distraction and Inattention.
Rosenbloom, T., 2006. Driving performance while using cell phones: an observational study. J. Safety Res. 37 (2), 207–212.
Schneidereit, T., Petzoldt, T., Keinath, A., Krems, J.F., 2017. Using SHRP 2 naturalistic driving data to assess drivers' speed choice while being engaged in different secondary tasks. J. Safety Res. 62, 33–42.
Sharda, S., Da Silva, D.C., Grimm, K.J., Khoeini, S., Pendyala, R.M., 2019. Modeling determinants of risky driving behaviors and secondary task engagement using naturalistic driving data. TRB's Transportation Research Circular E-C243: sHRP 2 Safety Data Student Paper Competition 2017-2019.
Stavrinos, D., Garner, A.A., Franklin, C.A., Johnson, H.D., Welburn, S.C., Griffin, R., Underhill, A.T., Fine, P.R., 2015. Distracted driving in teens with and without Attention-Deficit/Hyperactivity disorder. J. Pediatr. Nurs. 30 (5), e183–e191.
Strayer, D.L., Drews, F.A., Crouch, D.J., 2006. A comparison of the cell phone driver and the drunk driver. Hum. Factors 48 (2), 381–391.
Törnros, J., Bolling, A., 2006. Mobile phone use – effects of conversation on mental workload and driving speed in rural and urban environments. Transp. Res. Part F Traffic Psychol. Behav. 9 (4), 298–306.
Tsala, S.A.Z., Onomo, C., Mvogo, G., Ohandja, L.M.A., 2020. Elaboration of explanatory factors of accidents in Cameroon by factorial correspondence analysis. J. Transp. Technol. 10 (03), 280.
Violanti, J.M., Marshall, J.R., 1996. Cellular phones and traffic accidents: an epidemiological approach. Accid. Anal. Prev. 28 (2), 265–270.
Wang, Y., Chen, Y.J., Yang, J., Gruteser, M., Martin, R.P., Liu, H., Liu, L., Karatas, C., 2016. Determining driver phone use by exploiting smartphone integrated sensors. Ieee Trans. Mob. Comput. 15 (8), 1965–1981.
Yannis, G., Laiou, A., Papantoniou, P., Christoforou, C., 2014. Impact of texting on young drivers' behavior and safety on urban and rural roads through a simulation experiment. J. Saf. Res. 49, 25 e1-31 Proceedings of the International Conference on Road Safety (RSS2013).
Yıldırım, U., Başar, E., Uğurlu, Ö., 2019. Assessment of collisions and grounding accidents with human factors analysis and classification system (HFACS) and statistical methods. Saf. Sci. 119, 412–425.
Young, R.A., 2018. Cell phone conversation and relative crash risk update. Encyclopedia of Information Science and Technology, fourth edition. IGI Global, pp. 5992–6006.