Transportation Research Interdisciplinary Perspectives 10 (2021) 100360

Contents lists available at ScienceDirect

Transportation Research Interdisciplinary Perspectives

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

Mining crowdsourced data on bicycle safety critical events

Subasish Das[a,*], Zihang Wei[b], Xiaoqiang “Jack” Kong[b], Xiao Xiao[b]

[a] *Texas A&M Transportation Institute, 3500 NW Loop 410, Room 315E, San Antonio, TX 78229, USA*
[b] *Texas A&M University, 3135 TAMU, College Station, TX 77843-3135 USA*

ARTICLE INFO

Keywords:
Bikemaps.org
Bicycle collision
Bicycle near miss event
Crowdsourced data
Taxicab correspondence analysis

ABSTRACT

Cycling has become a popular transportation mode for short term trips. Due to the high exposure bicycle trips, the number of collisions and near miss events has been increasing significantly. This study explores the pattern of the bicycle-related collision or near miss events by using a unique crowdsourced dataset collected from BikeMaps.org. The dataset not only contains near miss events, which are not included in the conventional state-maintained crash databases, and it also includes the psychological impact of the event on the cyclist. The taxicab correspondence analysis (TCA) results reveal patterns for bike-related collision or near miss events and associated impact on the cyclists involved. Several factors such as inclement weather, windy condition, poor lighting conditions, wet ground, loose sand, or dirt pavement are associated with the increasing probability of the collision or near-miss events. The study indicates that collision or near miss events have a greater impact on cyclists if the events occurred when cyclists already have taken extra caution while cycling. These cyclists tend to cycle less and be more careful after these events. Interestingly, the results find that frequent cyclists are not psychologically affected by collisions occurred during recreational trips. The finding of this study could help researchers further understand bike collisions/near miss events and provide better countermeasures to mitigate the frequency of bike collisions.

1. Introduction

Cycling acts not only as one of the essential travel modes in urban areas but also as a healthy means of transport (Fishman et al., 2015). Meanwhile, the willingness to cycle is limited by the perception of safety, which could be impacted by both collisions and near miss events (Dill and Voros, 2007; Sanders, 2015; Horton et al., 2016). Even when the cyclist is not injured, near miss events are still incidents that are usually experienced by cyclists (Aldred and Goodman, 2018). Aside from collisions or crashes, near miss events are another factor that could have an impact on people’s perceived risk of bicycling (Sanders, 2015). Conventional police-reported crash databases are usually used in performing safety analysis. Near miss events associated with bicyclists are rarely examined. There is a need for a comprehensive study to examine both bicycle collisions and near miss events. Generally, the definition of a near miss event is an incident in which no direct physical collision happened, but where, given a slight shift in time or position, collision could have occurred.

To improve the safety of cycling activities, research studies have been conducted aiming to design and optimize the networks’ land use and infrastructures. Some countries, such as the Netherlands, Denmark, and Germany, have improved cycling safety by looking into comprehensive network-wide planning, restriction of car ownership, and strict land-use policies (Pucher and Buehler, 2008a, 2008b). The Netherlands promotes cycling safety by giving cyclists the right of way (Schepers et al., 2014). With separated paths and intersections for bicycles and motors, the potential for bicycle-motor crashes is reduced (Schepers et al., 2017). Denmark and Germany improved cycling safety by providing enough parking facilities and integrating cycling with other public transports. Education and training are also effective methods of improving cycling safety (Pucher and Buehler, 2008b). Aside from European countries in which the trips taken by bike is a large share, some other countries, such as the UK and the US, may have a low share among total means of trips, with about 1% (Pucher and Buehler, 2008b). For those countries, instead of planning and studying to improve the infrastructure from a network view, an alternative approach is used to study the potential factors that can be associated with collisions from a local view. After the data analysis, the issues that can lead to potential risks can be addressed accordingly.

Many factors linked to cycling collisions or near misses are studied in previous research. The use of alcohol not only acts as the main risk factor in motor driving but also in cycling. Cycling-related crashes and

[*] Corresponding author.
E-mail addresses: s-das@tti.tamu.edu (S. Das), z-wei@tti.tamu.edu (Z. Wei), x-kong@tti.tamu.edu (X. “Jack” Kong), xx1991@tamu.edu (X. Xiao).

https://doi.org/10.1016/j.trip.2021.100360
Received 24 December 2020; Revised 14 February 2021; Accepted 23 March 2021

their results show that severe injuries on the head and face can be associated with cycling under the influence of alcohol (Andersson and Bunketorp, 2002). To protect the head of cyclists, the use of a helmet is suggested for cyclists in many countries. Although it may reduce the willingness to cycling, the use of helmets is significant in reducing the severity of collisions. A cycling-related injuries report in Canada shows that the use of a helmet has decreased head injuries caused by collision while cycling by 50% and 26% for adults (Dennis et al., 2013). Geometric information is also the focus of research for cycling safety. For example, a naturalistic methodology-based study collects data from cameras, GPS, inertial measurement units, and pressure sensors to study these influences. According to data fusion results, the risks of collision are high near intersections, especially when the obstacles appear at the intersection. The pavement, as another factor, is also shown to relate in the same study (Dozza and Werneke, 2014). When concerning geometric location, the appearance of other cyclists cannot be ignored. Although the speed or the position of other cyclists is concerned as a safety issue for young cyclists in their own opinions (Amiri and Sadeghpour, 2015), another bicycle or pedestrian can contribute to the collisions (Dozza and Werneke, 2014). Other than geometric information, the weather and seasons are factors studied by previous studies. A study finds that the number of collisions or near misses increases in spring and summer. This is explained by the increasing interactions of bicycles with cars in these seasons (Fyhri et al., 2017). The weather is not always considered influential. According to a study from Canada, most young cyclists can bear cold temperatures while cycling (Amiri and Sadeghpour, 2015). Finally, aside from naturalistic factors, social factors are also important. For instance, gender is studied via a multi-group structural equation in a study about risky behavior among cyclists. Age, knowledge of traffic rules, psychological distress, and risk perception shows different levels of impact in keeping positive cycling behaviors for different genders (Useche et al., 2018). Most of the previous studies reveal the strength of the association of different factors with cycling incidents. Some studies also provide evidence to show the irrelevance. The involvement of a phone is concerning, and no evidence is found to show the positive relationship between the frequency of listening to music or the use of a phone can lead to a crash. This is explained by the compensating when using portable devices (Stelling-Konczak et al., 2017).

Recently, BikeMaps.org was developed (Nelson et al., 2015) as a crowdsourced information resource. It provides various records of factors that could link to cycling collisions and near misses, such as the status of pavement, weather, visibility, and psychological impact on the cyclists. The information provided by the website serves as a crowdsourced geographic information that attracts cyclists, especially young people, to report their incidents (Ferster et al., 2017). Previous studies using the dataset from BikeMaps.org show that younger people and females tend to have a high frequency of incidents in the center of a city (Jestico et al., 2016). It is usually seen that frequency of near miss events is much higher than the collisions. The peak hour, non-intersection locations, and the location of bike facilities are more likely to be associated with collisions (Branion-Calles et al., 2017). A balanced random forest is applied to classify 21 explanatory variables from the crowdsourced reports that can lead to three levels of injuries. The object the bike collides with is found to be the most impactful factor (Fischer et al., 2020). Although statistical studies using data from BikeMaps.org have been conducted before, most of them focus on one or two aspects and reveal their several relations. Due to the availability of near miss information and some unique variables, a comprehensive study on BikeMaps.org is needed.

This paper reveals character reporting patterns among influential factors and the potential influences on collisions and near misses using a correspondence analysis (CA) technique known as taxicab correspondence analysis (TCA) to bridge the gap. CA method determines a low-dimensional depiction that optimally illustrates relationships in the form of a contingency table. The output of the method is in the form of a CA plot that illustrates the optimal representation of variable categories based on the eigenvalue measures (Benzécri, 1973; Greenacre, 2007). TCA can determine the common factors that influence the variables, so the analysis of the problem can be clearer. In the recent years, several transportation safety studies applied TCA and other CA variants to determine domain specific patterns (Das and Sun, 2015, 2016; Baireddy et al., 2018; Das et al., 2018, 2020a, 2020b, 2021a, 2021b; Ali et al., 2018; Das, 2020; Sivasankaran and Balasubramanian, 2020; Das and Dutta, 2020; Kong et al., 2021). The usefulness of this method has been applied in this study by using data from BikeMaps.org. By using TCA, this study aims to answer two research questions: 1) What are the key patterns of bicycle collisions and near crash events? 2) What are key differences between these two patterns? Findings from this study will help authorities to perform bicycle-friendly roadway design and to implement suitable countermeasures.

2. Data description and methodology

2.1. Data description

This study acquired the bicycle-related collisions and near miss event data from BikeMaps.org. (BikeMap, 2020). The data are limited to Phoenix Area in Arizona. BikeMaps.org allows cyclists or bike crash spectators to report any bike-related collisions, including near miss events. Both crash and near miss event are evaluated by the reporters. Generally, the definition of a near miss event is an incident in which no direct physical collision happened, but where, given a slight shift in time or position, collision could have occurred. The person who reports the crash is encouraged to provide the crash details such as crash location, crash time, any other information related to the incident. The full list of the attribute of the bike-related crash can be found in a paper published by Nelson et al. (2015). The dataset contains 226 bike-related collisions and near miss events from 2013 to 2020 in the Arizona area. In this dataset, 108 collision events and 118 near miss events were reported. Fig. 1 shows the locations of the bicycle collision events and near miss events to provide an overview of the spatial distribution of these reported events. Majority of the collision data occurred at downtown areas.

BikeMaps.Org collects a wide range of variables (Nelson et al., 2015). This study selected 19 variables for the final analysis. Table 1 lists the variables and categories in each variable.

2.2. Summary statistics

Table 2 summarizes the BikeMap.org variables used in the analysis. The data are separated into two groups based on crash type (collision and near miss). There are 108 recorded collisions and 118 recorded near miss events. Most of these events (91.67% for collision and 92.37% for near miss) are related to moving objects. Collisions are more likely to result in severe crashes; 35.19% of collisions reported an emergency visit injury. In comparison, 98.31% of near miss crashes reported no injury. The majority of collisions (40.74%) are involved with vehicle angle, and the majority of near miss events (37.29%) are involved with sideswipe crashes. Additionally, 73.15% of collisions and 73.73% of near miss events are reported between Monday and Thursday. For collisions, the majority (49.07%) of cyclists indicate that they will become more careful in the future. While, for near miss events, the majority (37.29%) indicate that the event had no impact on them. Approximately 31% of the cyclists that reported collisions and 66.95% of the cyclists that reported near miss events are regular cyclists. Most of the collisions and near miss events were reported during the fall, and the fewest portion of crashes were reported during summer. One possible reason for this is that the summer in Arizona is too hot for outdoor cycling. Overall, the missing variable rate for collisions is higher than that of near miss events (see Table 2).

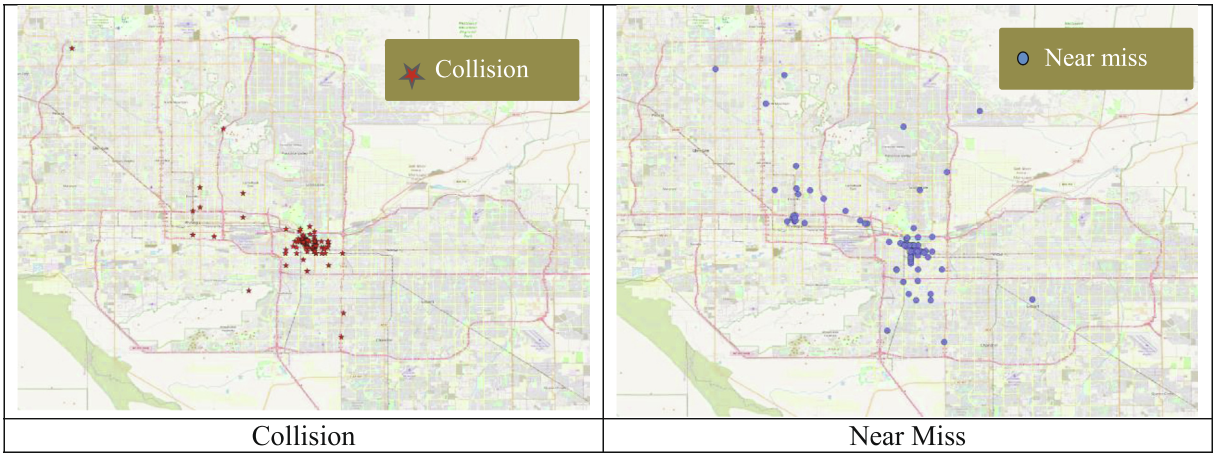


Fig. 1. Locations of collisions and near miss events.

Table 1
Selected Variables.

Variable Name	Category	Explanation	Variable Name	Category	Explanation
collision	MovingObjectVeh	Moving Object	**helmet**	Y	With Helmet
	NearCollisionMovingObjectVeh	Near Miss Moving Object		N	Without Helmet
	NearCollisionFixedObject	Near Miss Fixed Object	**pvnt**	Dry	Dry
	FixedObject	Fixed Object		Wet	Wet
	Fall	Fall		LooseSandDirt	Loose Sand Dirt
with	VehTurnRight	Vehicle Turn Right	**obs**	NoObstructions	No Obstructions
	VehTurnLeft	Vehicle Turn Left		ViewObstructed	View Obstructed
	VehPassing	Vehicle Passing*		Glare	Glare
	VehHeadOn	Vehicle Head On	**bikelight**	NL	No light
	Other	Other		B	Back light
	AnotherCyclist	Another Cyclist		FB	Front and Back
	VehicleSide	Vehicle Side*	**terrain**	Flat	Flat
	VehicleRearEnd	Vehicles’ Rear End		Uphill	Uphill
	VehAngle	Vehicle Angle		Downhill	Downhill
	Pedestrian	Pedestrian	**turning**	HeadingStraight	Heading Straight
	VehOpenDoor	Vehicle Open Door		TurningRight	Turning Right
dow	MTWT	Mon Tue Wed Thu	**age**	31_50	Between 31 and 50
	FSS	Fri Sat Sun		GT50	Great than 50
impact	Witness	Witness		19_30	Between 19 and 30
	None	None	**gen**	F	Female
	TooSoon	Too Soon to tell		M	Male
	MoreCarefulandBikeLess	More Careful and Bike Less		Other	Other
	MoreCareful	More Careful	**weather**	Clear	Clear
	StoppedBiking	Stopped Biking		Overcast	Overcast
	BikeLess	Bike Less		PartlyCloudy	Partly Cloudy
injury	Unknown	Unknown		MostlyCloudy	Mostly Cloudy
	NoInj	No Injury		Drizzle	Drizzle
	InjNoTreat	Non-threatening Injury		LightRain	Light Rain
	InjFamilyDoctor	Injury Family Doctor Needed		Light	Light
	Hospitalized	Hospitalized Injury	**season**	Spring	Spring
	EmergencyVisit	Emergency Visit Injury		Summer	Summer
purpose	Commute	Commute		Fall	Fall
	PersonalBusiness	Personal Business		Winter	Winter
	SocialReason	Social Reason	**lighting**	DawnDusk	Dawn Dusk
	ExerciseRecreation	Exercise Recreation		Day	Day
	DuringWork	During Work		Night	Night
regcy	Y	Regular Cyclist	**windspeed**	B48	Between 4 and 8
	N	Not Regular Cyclist		GT8	Greater than 8
				LT4	<4

*Note: “Vehicle Passing” means sideswiping and “Vehicle Side” means colliding directly with vehicle side.

2.3. Taxicab correspondence analysis

The difference between CA and TCA is that TCA uses a different singular value decomposition (SVD) based on taxicab norm called taxicab singular value decomposition (TSVD). TCA is similar to original CA and the only difference is that the geometry of CA is Euclidean, and the geometry of TCA is taxicab geometry which is non-Euclidean. Interested readers can consult Choulakian (2006) for the details of the mathematical concept.

Consider TSVD is a matrix Y, which is calculated in a stepwise manner. Let $v = (v_1, \cdots, v_r)^{'}$ u $=$ is an r-dimensional vector, the taxicab norm of v is $\|v\|_1 = \sum_{i=1}^{r}|v_i|$. T_c is the collection of all vectors of length c with coordinates $+1$ or -1. Thus, the total number of unique

Table 2
Frequency and percentages of key variables.

	Collision (108)		Near miss (118)			Collision (108)		Near miss (118)	
Category	**Freq**	**%**	**Freq**	**%**	**Category**	**Freq**	**%**	**Freq**	**%**
Collision Type					**Obstruction**				
Fall	4	3.70	0	0.00	Glare	2	1.85	2	1.69
Fixed Object	5	4.63	9	7.63	No Obstructions	34	31.48	88	74.58
Moving Object	99	91.67	109	92.37	View Obstructed	2	1.85	7	5.93
Missing	0	0.00	0	0.00	Missing	70	64.81	21	17.80
Collision Object					**Bike Light**				
Another Cyclist	2	1.85	3	2.54	Back Only	3	2.78	13	11.02
Other	12	11.11	2	1.69	FB	12	11.11	17	14.41
Pedestrian	0	0.00	2	1.69	No Light	24	22.22	64	54.24
Vehicle Angle	44	40.74	14	11.86	Missing	69	63.89	24	20.34
Vehicle Head On	5	4.63	20	16.95	**Type of Terrain**				
Vehicle Rear End	6	5.56	9	7.63	Downhill	4	3.70	4	3.39
Vehicle Side	25	23.15	44	37.29	Flat	37	34.26	88	74.58
Vehicle Passing	2	1.85	10	8.47	Uphill	0	0.00	5	4.24
Vehicle Turn Left	5	4.63	4	3.39	Missing	67	62.04	21	17.80
Vehicle Turn Right	7	6.48	6	5.08	**Turning**				
Vehicle Open Door	0	0.00	4	3.39	Heading Straight	36	33.33	92	77.97
Missing	0	0.00	0	0.00	Turning Right	5	4.63	5	4.24
Day of Week					Missing	67	62.04	21	17.80
Monday-Thursday	79	73.15	87	73.73	**Cyclist Age**				
Friday-Sunday	29	26.85	31	26.27	19 to 30	18	16.67	26	22.03
Miss	0	0.00	0	0.00	31 to 50	9	8.33	27	22.88
Impact					>50	4	3.70	24	20.34
Bike Less	3	2.78	0	0.00	Missing	77	71.30	41	34.75
More Careful	53	49.07	39	33.05	**Cyclist Gender**				
More Careful and Bike Less	10	9.26	9	7.63	Female	10	9.26	19	16.10
None	10	9.26	44	37.29	Male	22	20.37	61	51.69
Stopped Biking	5	4.63	0	0.00	Other	0	0.00	1	0.85
Too Soon	6	5.56	9	7.63	Missing	76	70.37	37	31.36
Witness	10	9.26	2	1.69	**Weather**				
Missing	11	10.19	15	12.71	Clear	82	75.93	85	72.03
Injury Condition					Light Rain	1	0.93	1	0.85
Emergency Visit	38	35.19	1	0.85	Light	0	0.00	1	0.85
Hospitalized	4	3.70	0	0.00	Drizzle	0	0.00	2	1.69
Injury Family Doctor	12	11.11	0	0.00	Mostly Cloudy	3	2.78	7	5.93
Injury No Treat	16	14.81	1	0.85	Overcast	2	1.85	0	0.00
No Injury	29	26.85	116	98.31	Partly Cloudy	20	18.52	22	18.64
Unknown	9	8.33	0	0.00	Missing	0	0.00	0	0.00
Missing	0	0.00	0	0.00	**Season**				
Purpose of the Trip					Fall	53	49.07	41	34.75
Commute	34	31.48	65	55.08	Spring	25	23.15	36	30.51
During Work	2	1.85	3	2.54	Summer	14	12.96	6	5.08
Exercise Recreation	9	8.33	28	23.73	Winter	16	14.81	35	29.66
Personal Business	2	1.85	13	11.02	Missing	0	0.00	0	0.00
Social Reason	2	1.85	6	5.08	**Lighting Condition**				
Missing	59	54.63	3	2.54	Dawn Dusk	24	22.22	25	21.19
Regular Cyclist					Day	69	63.89	83	70.34
Yes	33	30.56	79	66.95	Night	15	13.89	10	8.47
No	0	0.00	3	2.54	Missing	0	0.00	0	0.00
Missing	75	69.44	36	30.51	**Wind Speed**				
Helmet					4–8	43	39.81	44	37.29
Yes	22	20.37	53	44.92	>8	16	14.81	39	33.05
No	10	9.26	25	21.19	< 4	49	45.37	35	29.66
Missing	75	69.44	40	33.90	Missing	0	0.00	0	0.00
Pavement Type									
Dry	38	35.19	93	78.81					
Loose Sand Dirt	2	1.85	3	2.54					
Wet	1	0.93	2	1.69					
Missing	67	62.04	20	16.95					

vectors in T_c is 2^c. T_r is similar for column points. To find the first principle axis v_1 of r row points, v_1 should satisfy:

$$\underbrace{max}_{v\in T_c} \| Yv_1 \| = \| Yv_{11} \|$$

where Yv is the projection of r row points on v. And $\|Yv\|_1$ is the taxicab norm of the projection. λ_1 is the first taxicab principal axis dispersion measure:

$$\lambda_1 = \|Yv_1\|_1 \qquad (1)$$

f_1 are the first-row principal factor scores:

$$f_1 = Yv_1 \qquad (2)$$

λ_1 can be written as:

$$\lambda_1 = sgn({f_1}')Yv_1 \qquad (3)$$

λ_1 can be written as equation (4) by putting $u_1 = sgn(f_1) \in T_r$:

$$\lambda_1 = {u_1}'Yv_1 \qquad (4)$$

where u_1 is the first column axis of c column points of dataset Y. And the first column principle factor scores g_1 can be written as:

$$g_1 = Y'u_1 \tag{5}$$

λ_1 can also be written as follow:

$$\lambda_1 = u_1{}'f_1 = v_1{}'g_1 \tag{6}$$

From equation (1)–(5), the calculation of the first principle axis and the first principle score for both r row points and c column points (v_1, u_1, f_1, g_1) are presented. Then Wedderburn's rank-one reduction formula is applied from here to calculate the second principal axis and the principle factor score for both row and column points of dataset Y. Finally, the TSVD of a matrix Y will have the following form:

$$Y = \sum_{a=1}^{k} \frac{f_a g_a{}'}{\lambda_a} \tag{8}$$

To implement TCA to a contingency table Y, there is need to apply TSVD to a correspondence matrix P = Y/n which is similar to the original correspondence analysis. P is a correspondence matrix with marginal proportions $p_{i.}$ and $p_{.j}$. The process is to apply TSVD to P and continue the same process as above until the kth iteration.

3. Results

TCA method has been applied to two datasets (collision data, and near miss event data) to answer the research questions. Note that this study used a crowdsourced dataset. It is possible that some individuals reported inaccurate information. However, due to the nature of TCA, features are clustered together only if they have strong relationship. Thus, a few inaccurate records can hardly change the over patterns.

3.1. Results of collision data

TCA method transforms tabular numerical data into two-dimensional data visualizations commonly known as TCA plot. Figs. 2 and 3 show the TCA plots of collision data. Four plots in Fig. 1 (1st and 2nd quadrants) and Fig. 2 (3rd and 4th quadrants) represent four quadrants of a TCA plot for the purpose of clarity. Seven meaningful clusters are identified in these four quadrants. The general interpretation indicates that the closeness of the attributes are associated with co-occurrence in the dataset. First two axes explained approximately 80% (axis 1: 56.6%, axis 2: 23.2%) variance of data.

Quadrant 1 (Fig. 2)

In cluster 1, "impact_Witness" is clustered with "injury_Hospitalized," "injury_Unknown," and "weather_MostlyCloudy." As mentioned in the data description, "impact_Witness" indicates that the people who reported an event is a whiteness instead of the person involved. It is reasonable to have this category clustered with "injury_Hospitalized" and "injury_Unknown," because the witness can only report the cyclist who involved the crash is hospitalized if the victim was taking way by an emergency vehicle. Otherwise, the cyclist's injury condition would be unknown to the person who reported this crash. Cluster 2 indicates weekdays, moving vehicles, and moderate wind speed between 4mph to 8 mph are associated with bike collisions. Generally, weekdays are when most bike trips occurred, and the majority of bike collisions involve moving vehicles. These two statements can also be validated by the descriptive statistics of this research in Table 2. Thus, the cluster supports the strong association between bike collision patterns with weekday trips, colliding with moving vehicles, and windy weather.

Quadrant 2 (Fig. 2)

Cluster 3 suggests twelve categories associating with the bike collision. These categories are the movement types of vehicles collided with the bike, windy weather, seasonality, lighting conditions, trip purpose, injury severity, and psychological impact on the involved cyclists. The movements of the vehicles which crashed with the bikes in this cluster are turning right, going straight (head-on collision) and passing by (sideswipe). These types almost cover all possible movement types of a vehicle, which demonstrates that, with other categories in the cluster, the cyclist could potentially collide with the vehicle of any movement type. During poor lighting conditions, commute trips – during dusk or dawn and inclement weather – wind speed greater than 8 mph are highly associated with bike crashes. Interestingly, the cluster also associates with minor or no injuries. This might be explained by the extra caution taken by cyclists during inclement weather and lighting conditions. However, the collision during this severe weather and poor lighting conditions posed a strong influence on cyclists psychologically because the cluster shows the association between the crashes during these conditions and the impact afterward, which is "stopped biking". Note that, the "stopped biking" indicates that the crashes have posed significant psychology impacts on the cyclists. It does not necessarily means the cyclists will stop biking forever, they may resume biking in the future.

Quadrant 3 (Fig. 3)

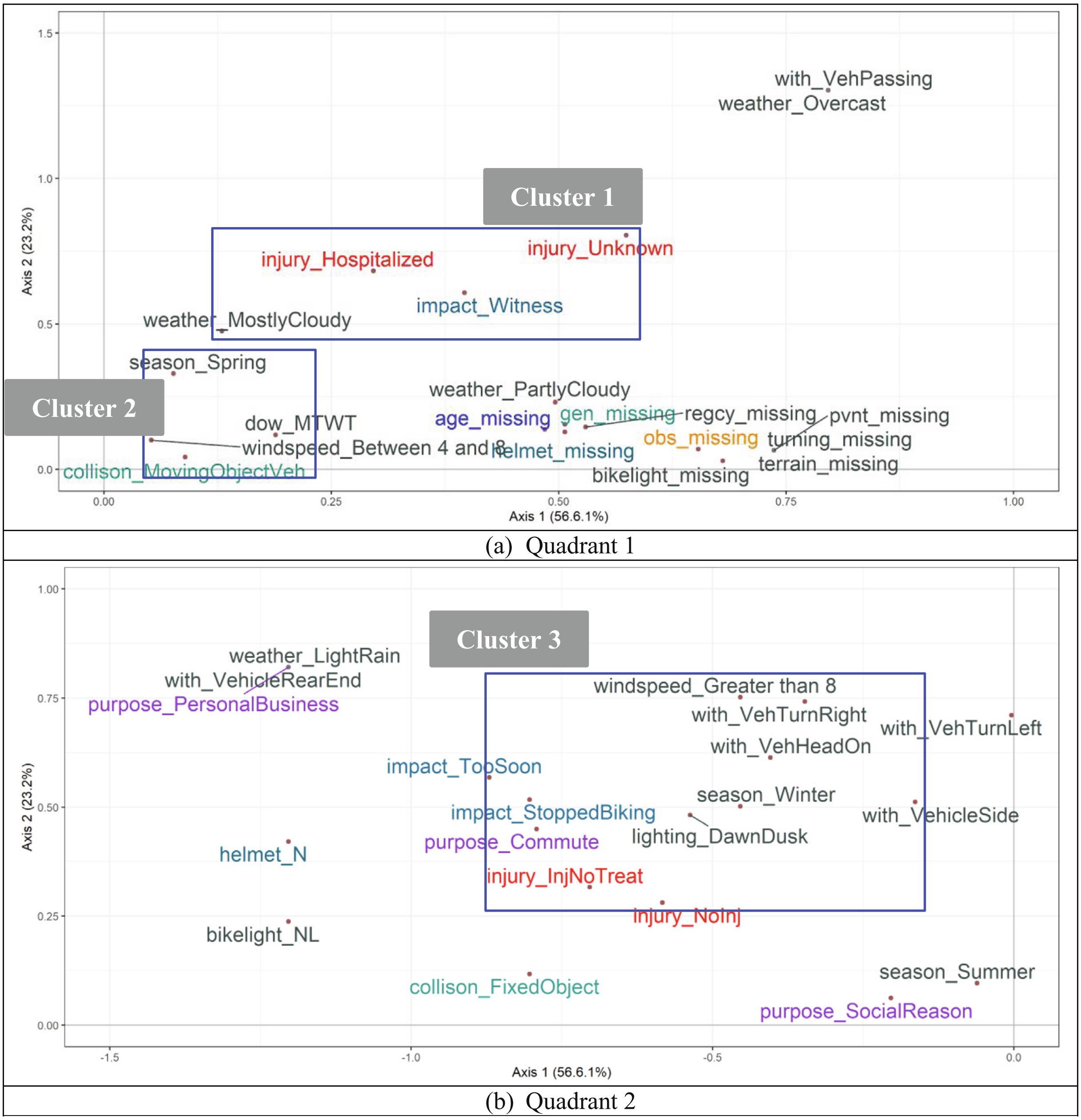
Cluster 4 is one of two clusters identified in quadrant 3. It also clusters a wide range of categories. There are three findings in this cluster. First, the glare effect is associated with the bike crash. The glare effect also co-exists with dry pavement conditions. Second, the cluster also shows that recreational cyclists are often frequent/regular cyclists, who often wear helmet and age from 19 to 50. This does not suggest that older cyclists do not exist in this category. Cyclists report the data voluntarily through websites and phone applications. The chance of having these elder cyclists reporting events on the website or through applications is relatively less. Third, this cluster also finds the collisions of these recreational trips of frequent cyclists had limited impacts on their attitude towards cycling. Cluster 5 indicates fall-off collision often occurs on the wet or loose sand/dirt pavement and usually associates with cyclists with age over 50 years old.

Quadrant 4 (Fig. 3)

This quadrant has two clusters: Cluster 6 and 7. Cluster 6 again states the possible impacts of crashes could be bike less or bike with more caution when the crashes happened during the day and non-windy weather. This might say the crash would affect cyclists more if the crash occurred in unexpected conditions with good lighting and non-windy weather. Cluster 7 associates the collision with the vehicle from an angle with relatively more severe injuries, requiring emergency visits or visiting family doctors. This finding echoes with the results in research conducted by Fischer et al. (2020), which found that collisions involving vehicle angles are more likely to cause severe injury (requiring emergency visits or family doctor). Because when vehicles are turning, cyclists are more likely to collide with a vehicle angle, these two findings are consistent. Meanwhile, the relatively more severe collision such as these requiring emergency visits often encourages cyclists to bike less and bike with more cautions.

Information from each collision (row-level analysis) can be used to develop clusters. This analysis helps understanding the subgroup effect in data. The row level cluster analysis helps to identify the association of the clusters with some key variable categories or attributes. The log odds ratio (LOR) measures of each cluster indicate the higher or lower odds of the categories in clusters (see Table 3). For the collision dataset, 13 clusters were developed based on the location of the x-axis measures of the row related data points on a two-dimensional plot (see Fig. 4). For example, 'Clus01' has two solid circles with the same x-axis and different y-axis. The size of the circle indicates the sample size of each location.

The log odds ratio (LOR) for "Clus01" near miss incidents with more cautious bicyclists can be calculated



(a) Quadrant 1

(b) Quadrant 2

Fig. 2. TCA Plots of Collision Data (1st and 2nd Quadrants).

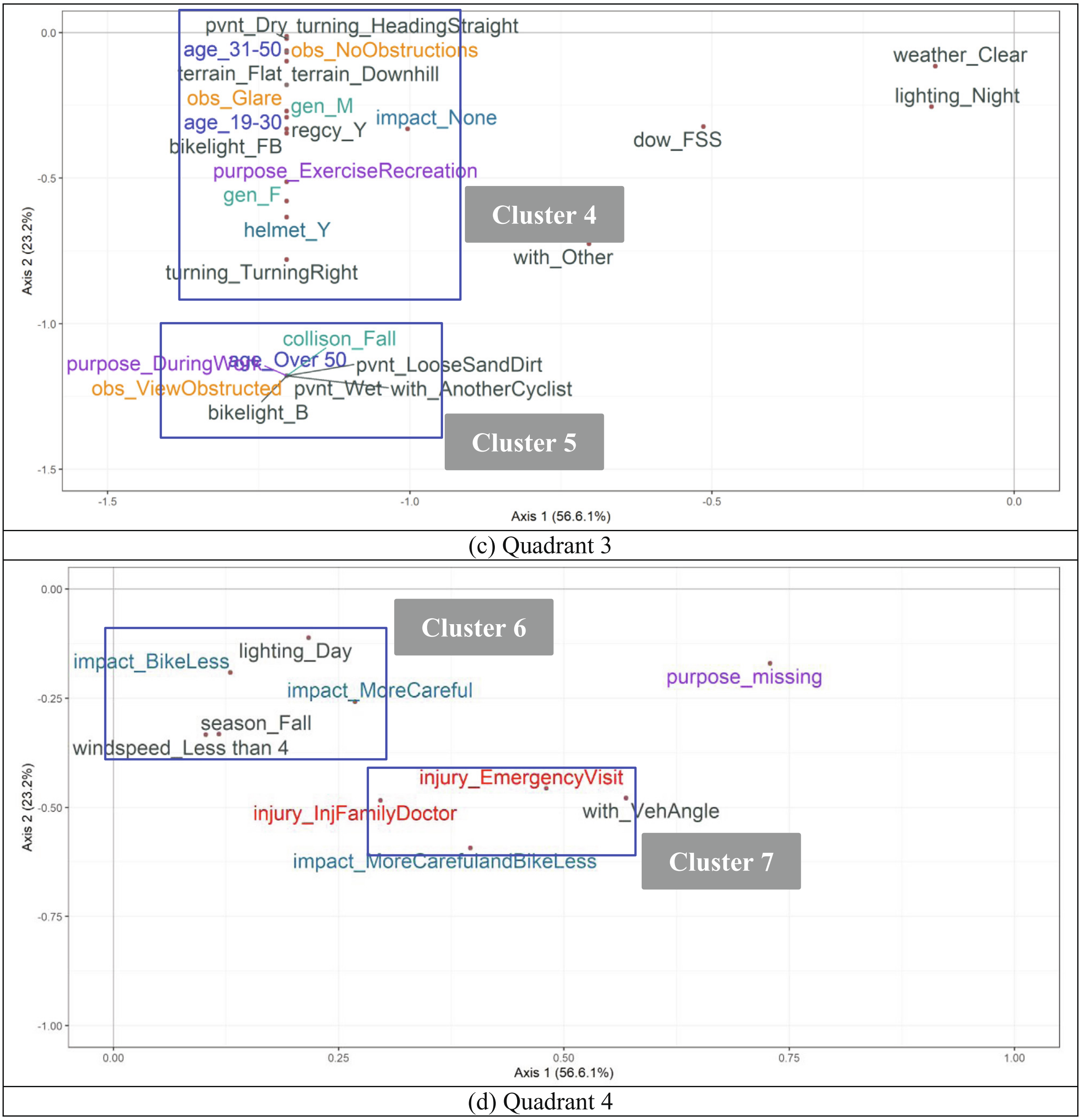
as:$LOR(MoreCautious|Others) = ln\left(\frac{4/2}{48/70}\right) = 1.07$. The interpretation of LOR can be explained as follows:

- LOR = 0, then the proportion of 'More Cautious' bicyclists in cluster s equals the proportion of 'Others' in the sample.
- LOR > 0, then the proportion of 'More Cautious' bicyclists in cluster s is greater than the proportion of females in the sample. That is, the cluster s is positively associated with 'More Cautious' bicyclists, and negatively associated with 'Others'.
- LOR < 0, then the proportion of 'Others' in cluster s is larger than the proportion of 'More Cautious' bicyclists in the sample. That is, the cluster s is positively associated with 'Others', and negatively associated with 'More Cautious' bicyclists.

To determine the risker groups, three major bicyclist related variables are considered for analysis. These variables are impact of the incident, injury condition, and weather condition. Log odds ratio (LOR) for "Clus02" collisions with more cautious bicyclists can be calculated as:$LOR(MoreCautious|Others) = ln\left(\frac{7/2}{71/37}\right) = 0.601$. It indicates that for cluster "Clus02," the proportion of cautious cyclists is larger than the proportion of cyclists with other options. For three clusters (Clus08, Clus10, Clus13), the first two (impact and injury type) LOR measures are positive. It indicates that these clusters are positively associated with bicyclists who were more likely to be involved in an injury or hospitalization-related collisions and would express that they would be more cautious and careful towards bicycling in the future. Out of 13 clusters, six clusters show positive LOR values for bicyclists who were more likely to be cautious in bicycling in the future.

3.2. Results of near miss data

Figs. 5 and 6 illustrate the TCA plots for near miss data computation. Seven clusters are identified in the four quadrants. First two axes



(c) Quadrant 3

(d) Quadrant 4

Fig. 3. TCA Plots of Collision Data (3rd and 4th Quadrants).

explained approximately 65% (axis 1: 36.1%, axis 2: 28.5%) variance of data.

Quadrant 1 (Fig. 5)

Cluster 8 shows the near miss events that are associated with a head-on vehicle, glare effect, drizzle weather, trips with the purpose of social reasons, and being more careful later. Glare effects and drizzle weather could create an uncomfortable biking environment for the cyclist and encourages the cyclist to be more careful with the surroundings, especially the head on vehicles. However, a near miss event with a head-on vehicle could have a considerable amount of impact on the cyclist, like causing the cyclists to be more careful in the future. This is likely because accidents involved with head-on vehicles usually leave a significant psychological shadow on cyclists.

Quadrant 2 (Fig. 5)

There are two clusters in this quadrant. Cluster 9 indicates that many near miss events with fixed objects occur during commute trips in poor lighting conditions. These trips often occur with female cyclists that are not wearing helmets. This may suggest that many cyclists of near miss fixed objects events are non-frequent cyclists who have relatively fewer experiences of biking in poor lighting conditions. Cluster 10 shows that many sideswipe related near miss events occur on weekdays while the cyclists are turning right. Biking during clear weather, the drivers and cyclists can easily detect each other's position. Especially during the weekday commuting time, both sides move with more caution.

Quadrant 3 (Fig. 6)

Cluster 11 depicts a decent biking environment: flat terrain, dry pavement, and the cyclists heading straight. The near miss event under

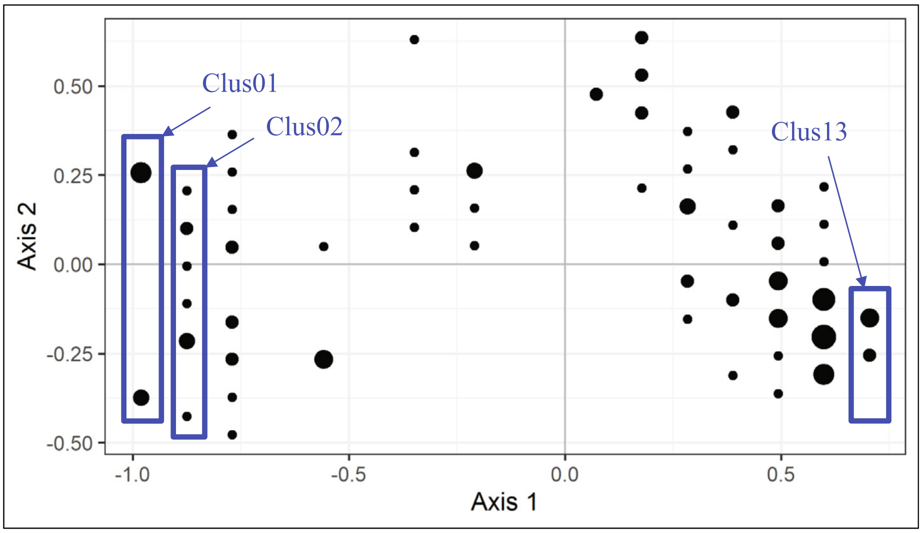


Fig. 4. Row based clusters for collision data.

Table 3
Log Odds Ratio Measures for three key variables (collision data).

Row Clusters	Impact			Injury Condition			Weather		
	More Cautious	Others	LOR	Injury related	No Injury	LOR	Clear	Inclement	LOR
Clus01	4	4	−0.652	0	5	–	8	0	–
Clus02	7	2	0.601	5	4	−0.388	6	3	−0.455
Clus03	5	6	−0.834	3	8	−1.592	10	1	1.154
Clus04	4	1	0.735	3	2	−0.205	2	3	−1.554
Clus05	2	2	−0.652	0	4	–	4	0	–
Clus06	4	1	0.735	2	3	−1.016	5	0	–
Clus07	1	1	−0.652	1	1	−0.611	0	2	–
Clus08	6	1	1.14	5	2	0.305	5	2	−0.232
Clus09	5	3	−0.141	7	1	1.335	7	1	0.797
Clus10	6	1	1.14	6	1	1.181	5	2	−0.232
Clus11	9	5	−0.064	11	3	0.688	10	4	−0.232
Clus12	14	8	−0.092	19	3	1.235	14	8	−0.589
Clus13	4	2	0.041	5	1	0.999	6	0	–
Grand Total	71	37		70	38		82	26	

this decent biking condition for the frequent/regular male cyclists could pose a relatively strong impact on them, including biking less and being more careful when biking in the future.

Quadrant 4 (Fig. 6)

Three clusters are found in quadrant 4. Cluster 12 shows the association between near miss events and nonregular cyclists' recreational trips on weekends during windy weather. It is expected that most recreational trips happen on weekends, including Friday. Cluster 13 indicates that some near miss events also occur with cyclists greater than 50 years old, turning right on wet pavement. Turning on the wet pavement could be a combination that triggers many near miss events. In cluster 14, another combination that may cause near miss events are light rain, loose sand or dirt pavement, and a non-flat terrain. The combination of these contributing factors could cause not only near miss events, but also collision events.

Information from each near miss event (row-level analysis) can be used to develop clusters. For the near miss event dataset, 13 clusters were developed based on the location of the x-axis measures of the row related data points on a two-dimensional plot (see Fig. 7).

This indicates that for cluster "Clus01," the proportion of cautious cyclists is larger than the proportion of cyclists with other options (see Table 4). For "Clus13," the first two LOR measures (impact and injury type) are positive. This indicates that these clusters are positively associated with bicyclists who were more likely to be involved in injury or hospitalization related collisions and would express that they would be more cautious and careful when bicycling in the future. Out of 13 clusters, six clusters show positive LOR values for bicyclists who were associated with near miss events and were more likely to be cautious when bicycling in the future. "Clus02" associates near miss incidents in unpleasant weather with a large negative LOR compared to others, while the other clusters contain incidents that mostly occurred in clear weather conditions.

3.4. Key findings

Several apparent relationships are found in this study. For collision events, the main findings of the dataset include:

- Inclement weather environments and poor pavement conditions could increase the probability of collisions.
- Injuries from collision events that occurred during commute trips under poor lighting conditions are often not severe because commuters were likely to be more cautious in poor lighting conditions, which generally prevented them from getting injured. However, these crashes could have a relatively strong impact on cyclists.

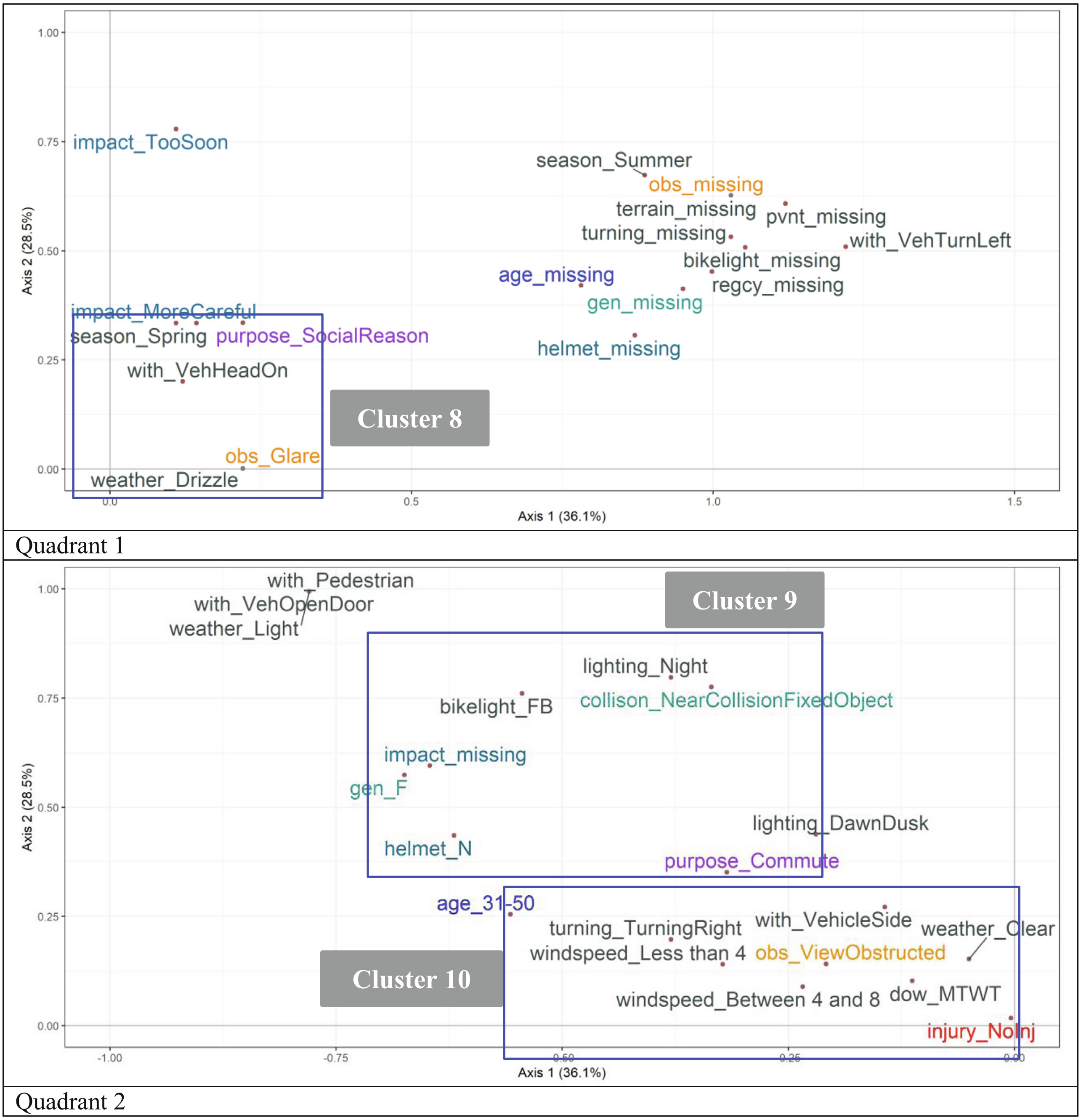


Fig. 5. TCA Plots of Near miss Data (1st and 2nd Quadrants).

For example, some of them reported stopped biking afterward. Lighting on streets and bike-friendly road design could reduce these events.

- Riders of recreational trips are mostly regular cyclists, who tend to have a helmet on. Meanwhile, the collisions that occur during these recreational trips often have less impact on cyclists. A possible explanation could be that regular cyclists are often aware of the risks of riding on the roads. Their skills and experiences could help them avoid or mitigate the magnitude of the collision, so the crashes have a lesser impact on them.
- Fall-off collisions are often related to cyclists more than 50 years old as well as wet or loose sand/dirt pavement. Safe cycling training and safety education can be helpful in reducing these events.
- Not surprisingly, the results show that cyclists who have been involved in severe crashes tend to be more careful and cycle less later.

For near miss events, the main findings are:

- Near miss events often occur during dawn and dusk when inexperienced cyclists are commuting, often biking without wearing a helmet.
- Near miss events with fixed objects often occur during commute trips under poor lighting conditions. The cyclists in these events were more likely to be female and nonregular cyclists. Lighting on streets and bicycle-friendly road design could reduce these events.

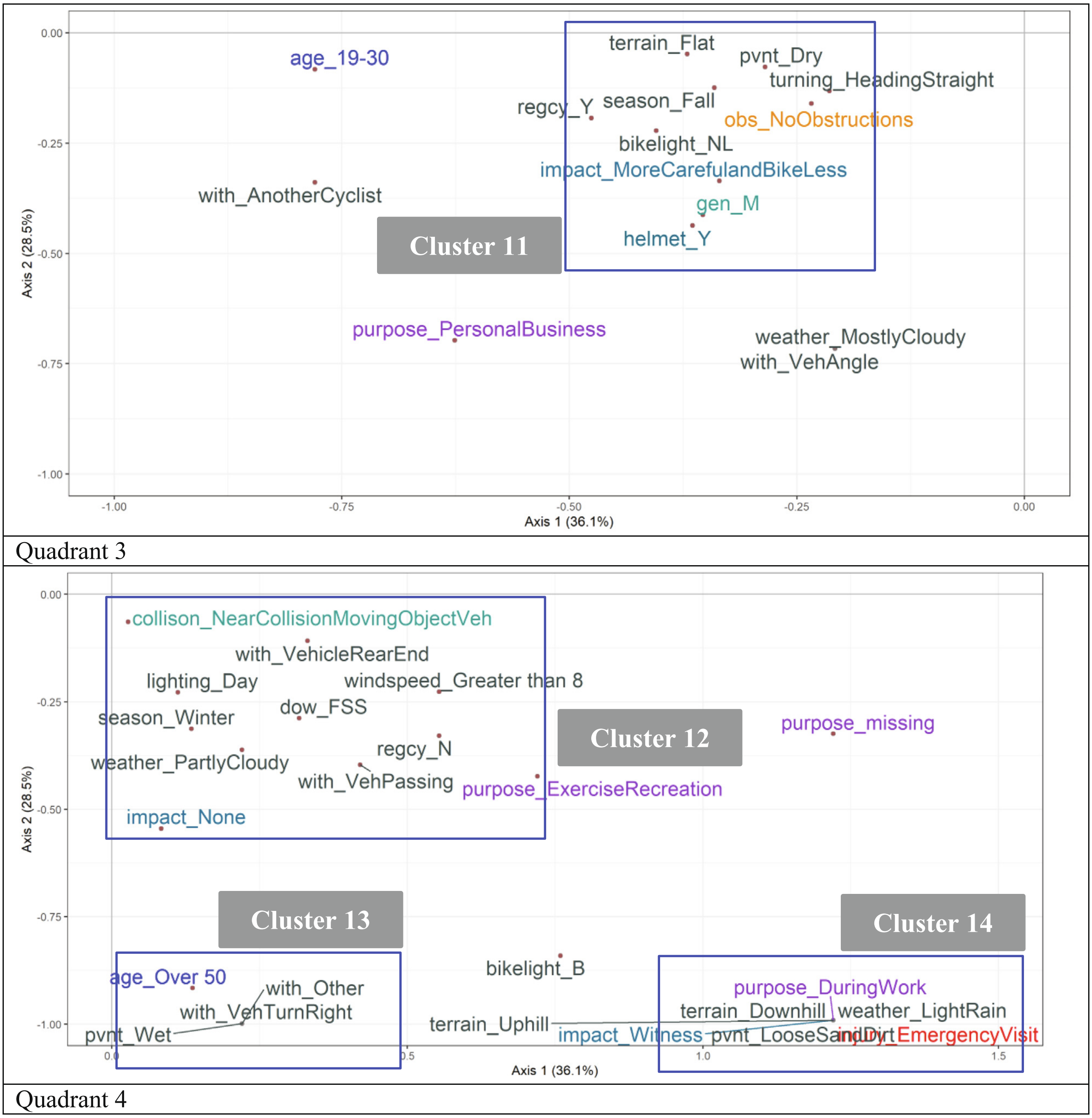


Fig. 6. TCA Plots of Near miss Data (3rd and 4th Quadrants).

- Near miss events are also likely to occur among cyclists older than 50 when the pavement was wet. A similar pattern is also found for collision events where wet pavement is also clustered with cyclists older than 50.
- Near miss events could also occur during recreational trips with nonregular cyclists during inclement weather conditions.
- Both collision and near miss events are less likely to occur in clear weather conditions.

The odds ratios show the differences of the clusters that are associated with collision or near miss events in three variables: impact of the incident, injury condition, and weather condition. When considering the variables of impact and injury for collision events, the odds ratio indicates that cyclists who have experienced injury tend to be more cautious when cycling in the future, which is consistent with the results from TCA plots. The odds ratio also shows that near miss events could encourage the cyclist to be more cautious while cycling.

3.5. Potential countermeasures

Based on the findings, several important countermeasures can be considered in preventing bicycle involved crash and near-crash events in the future. Under inclement weather condition, poor pavement conditions are more likely to be associated with traffic collisions. More investments on pavement condition can improve safety. For commute trip, poor light condition can likely to be associated with bicycle related collision events. Roadways with high number of commuters can be considered in the priority list of roadways having proper lighting condition at night. Fall-off collisions are often related to cyclists

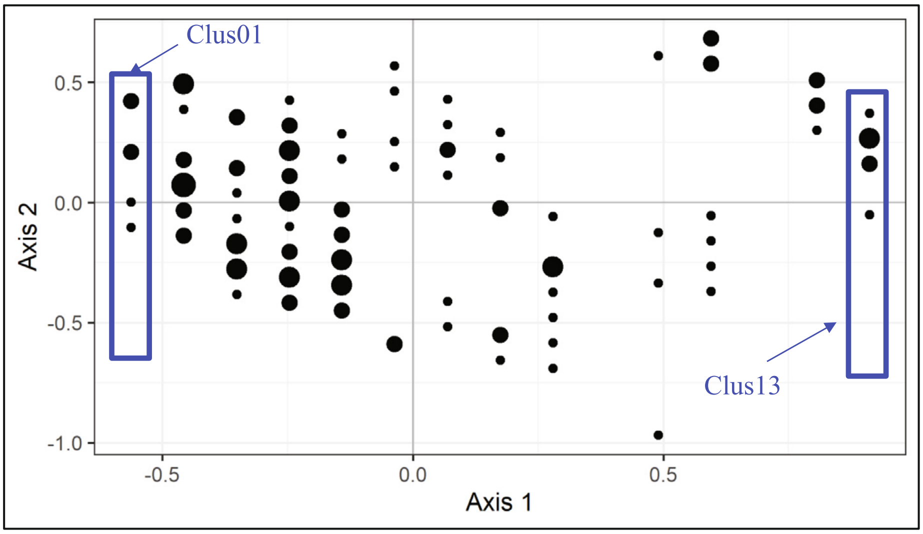


Fig. 7. Row based Clusters for Near miss Data.

Table 4
Log Odds Ratio Measures for Three Key Variables (Near miss Data).

Row Clusters	Impact			Injury Condition			Weather		
	More Cautious	Others	LOR	Injury related	No Injury	LOR	Clear	Inclement	LOR
Clus01	4	2	1.07	0	6		5	1	0.663
Clus02	3	11	−0.922	0	14		5	9	−1.53
Clus03	5	8	−0.093	0	13		10	3	0.258
Clus04	10	9	0.483	0	19		14	5	0.083
Clus05	8	6	0.665	0	14		12	2	0.846
Clus06	2	4	−0.316	0	6		4	2	−0.25
Clus07	1	6	−1.414	0	7		4	3	−0.66
Clus08	4	3	0.665	0	7		5	2	−0.03
Clus09	2	6	−0.721	1	7	2.115	6	2	0.152
Clus10	1	3	−0.721	0	4		3	1	0.152
Clus11	4	4	0.377	0	8		7	1	1
Clus12	1	4	−1.009	0	5		4	1	0.44
Clus13	3	4	0.09	1	6	2.269	6	1	0.846
Grand Total	48	70		2	116		85	33	

more than 50 years old. Safety education and training in proper gear and helmets can improve safety awareness among the elder cyclists.

Near miss events with fixed objects often occur during commute trips under poor lighting conditions. Improving the lighting condition and designing more bike friendly infrastructures can mitigate this problem. Near miss and collision events are often associated with wet pavement condition among cyclists over 50 years old. For elder citizens, it is necessary to recommend them to avoid cycling during wet surface condition.

4. Conclusions

Conventional bicycle-related collision datasets normally do not contain near miss events, which could be an important resource for researchers to understand these events and further recommend suitable countermeasures to protect cyclists. This study explores a unique dataset containing bicycle-related collisions and near miss events. The information about the impact of the collision or near miss events on cyclists could help researchers gain a profound understanding of cyclists from a psychological perspective. This study explores this unique dataset using TCA, which reveals the insightful correlations among categorical variable attributes through dimension reduction. The dataset is divided into two groups: collision and near miss events; and then fourteen meaningful clusters are identified from the two groups.

The contention of this study is that identification associations and patterns of key contributing factors can divulge realistic indication on gaps in the suite of suitable countermeasures that may contribute to bicycle crash or near miss event reduction. This study demonstrated how intuitive insights could be revealed using TCA by using a dataset collected from BikeMaps.org. This dataset is unique in comparison to conventional crash databases. Impact of future cycling behavior, helmet usage, and near crash events are some of the few features of this database. The TCA analyses suggest that both bike collision events and near miss events are more likely to occur during weekdays and when the bicyclist is traveling for commute purposes. Inclement weather, very windy, poor lighting conditions, wet ground, loose sand or dirt pavement are all factors that increase the chance of collision and near miss events. The study indicates that collision or near miss events have a stronger impact on cyclists if the events occurred when the cyclist was paying extra attention while cycling. These cyclists tend to cycle less and more carefully after these events. Interestingly, the results find that regular cyclists are not psychologically affected by collisions that occur during recreational trips.

As data-driven research, this analysis reveals several important findings. However, there are still some limitations to the research. First, the sample size of the available data is limited. It is reasonable to believe that many bicycle-related collisions are not reported on this website. The research purpose would be better served if the data were richer. Additionally, comparison with the real crash data can be insightful. However, the findings of the existing data are still reasonable and insightful due to the new information (e.g, near-crash, effect on future biking) that are associated with this data. Second, since the events were voluntarily reported, there are many missing values in the dataset. The results could be more significant if all of the information were filled by the reporters. Additionally, some of the geometric data can be manually added. As the study mainly focused on the usefulness of crowdsourced data, additional data were not included in the analysis. Third, this research only includes a dataset from the Arizona area, mostly in Phoenix. It is possible that the patterns found in this research may vary from patterns in other regions of the U.S. Future studies could perform more comprehensive data collection across different areas.

CRediT authorship contribution statement

Subasish Das: Conceptualization, Software, Validation, Formal analysis, Writing - original draft. **Zihang Wei:** Conceptualization, Software, Validation, Formal analysis, Writing - original draft. **Xiaoqiang "Jack" Kong:** Formal analysis, Writing - review & editing. **Xiao Xiao:** Formal analysis, Writing - review & editing.

Acknowledgments

The authors like to thank the developers of BikeMaps.org for supplying the data.

Funding

This study did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Aldred, R., Goodman, A., 2018. Predictors of the frequency and subjective experience of cycling near misses: findings from the first two years of the UK Near Miss Project. Accid. Anal. Prev. 110, 161–170.

Ali, F., Dissanayake, D., Bell, M., Farrow, M., 2018. Investigating car users' attitudes to climate change using multiple correspondence analysis. J. Transp. Geogr. 72 (2018), 237–247.

Amiri, M., Sadeghpour, F., 2015. Cycling characteristics in cities with cold weather. Sustainable Cities Soc. 14, 397–403.

Andersson, A.-L., Bunketorp, O., 2002. Cycling and alcohol. Injury 33 (6), 467–471.

Baireddy, R., Zhou, H., Jalayer, M., 2018. Multiple correspondence analysis of pedestrian crashes in rural Illinois. Transp. Res. Rec. 2672 (38), 116–127.

Benzécri, J.P., 1973. L'Analyse des Données. Tome 1: La Taxinomie. Tome 2: L'Analyse des correspondances. Dunod Paris.

BikeMap, 2020. BikeMaps [WWW Document]. BikeMaps. URL < https://bikemaps.org/ > (accessed 10.4.20).

Branion-Calles, M., Nelson, T., Winters, M., 2017. Comparing crowdsourced near miss and collision cycling data and official bike safety reporting. Transp. Res. Rec. 2662 (1), 1–11.

Das, S., Mousavi, M., Shirinzad, M., 2021a. Speeding related motorcycle injuries: findings from cluster correspondence analysis. J. Traffic Saf. Secur..

Das, S., Kong, X., Lavrenz, S., Jalayer, M., Wu, L., 2021b. Pattern recognition from rail grade crossing fatal crashes. Int. J. Transp. Sci. Technol..

Das, S., 2020. Identifying key patterns in motorcycle crashes: findings from taxicab correspondence analysis. Transp. A: Transp. Sci. 17 (4), 593–614.

Das, S., Ashraf, S., Dutta, A., Tran, L.-N., 2020a. Pedestrians under influence (PUI) crashes: patterns from correspondence regression analysis. J. Saf. Res. 75, 14–23.

Das, S., Islam, M., Dutta, A., Shimu, T.H., 2020b. Uncovering deep structure of determinants in large truck fatal crashes. Transp. Res. Rec.: J. Transp. Res. Board 2674 (9), 742–754.

Das, S., Dutta, A., 2020. Extremely serious crashes on urban roadway networks: patterns and trends. IATSS Res. 44 (3), 248–252.

Das, S., Avelar, R., Dixon, K., Sun, X., 2018. Investigation on the wrong way driving crash patterns using multiple correspondence analysis. Accid. Anal. Prev. 111, 43–55.

Das, S., Sun, X., 2016. Association knowledge for fatal run-off-road crashes by multiple correspondence analysis. IATSS Res. 39 (2), 146–155.

Das, S., Sun, X., 2015. Factor association with multiple correspondence analysis in vehicle-pedestrian crashes. Transp. Res. Rec.: J. Transp. Res. Board 2519 (1), 95–103.

Dennis, J., Ramsay, T., Turgeon, A.F., Zarychanski, R., 2013. Helmet legislation and admissions to hospital for cycling related head injuries in Canadian provinces and territories: interrupted time series analysis. BMJ 346. f2674.

Dill, J., Voros, K., 2007. Factors affecting bicycling demand: initial survey findings from the Portland, Oregon, region. Transp. Res. Rec. 2031 (1), 9–17.

Dozza, M., Werneke, J., 2014. Introducing naturalistic cycling data: What factors influence bicyclists' safety in the real world?. Transp. Part F Traffic Psychol. Behav. 24, 83–91.

Ferster, C.J., Nelson, T., Winters, M., Laberee, K., 2017. Geographic age and gender representation in volunteered cycling safety data: a case study of BikeMaps.org. Appl. Geogr. 88, 144–150.

Fischer, J., Nelson, T., Laberee, K., Winters, M., 2020. What does crowdsourced data tell us about bicycling injury? A case study in a mid-sized Canadian city. Accid. Anal. Prev. 145, 105695.

Fishman, E., Schepers, P., Kamphuis, C.B.M., 2015. Dutch cycling: quantifying the health and related economic benefits. Am. J. Public Health 105 (8), e13–e15.

Fyhri, A., Sundfør, H.B., Bjørnskau, T., Laureshyn, A., 2017. Safety in numbers for cyclists—conclusions from a multidisciplinary study of seasonal change in interplay and conflicts. Accid. Anal. Prev. 105, 124–133.

Greenacre, M.J., 2007. Correspondence Analysis in Practice. CRC Press, Boca Raton, FL.

Horton, D., Rosen, P., Cox, P., 2016. Cycling and Society. Routledge, Washington DC.

Jestico, B., Nelson, T., Winters, M., 2016. Mapping ridership using crowdsourced cycling data. J. Transp. Geogr. 52, 90–97.

Kong, X., Das, S., Zhou, H., Zhang, Y., 2021. Lessons learned from pedestrian-driver communication and yielding patterns. Transp. Res. Part F: Traffic Psychol. Behav..

Nelson, T.A., Denouden, T., Jestico, B., Laberee, K., Winters, M., 2015. BikeMaps. org: a global tool for collision and near miss mapping. Front. Publ. Health 3, 53.

Pucher, J., Buehler, R., 2008a. Cycling for everyone: lessons from Europe. Transp. Res. Rec. 2074 (1), 58–65.

Pucher, J., Buehler, R., 2008b. Making cycling irresistible: lessons from the Netherlands, Denmark and Germany. Transp. Rev. 28 (4), 495–528.

Sanders, R.L., 2015. Perceived traffic risk for cyclists: the impact of near miss and collision experiences. Accid. Anal. Prev. 75, 26–34.

Schepers, P., Hagenzieker, M., Methorst, R., van Wee, B., Wegman, F., 2014. A conceptual framework for road safety and mobility applied to cycling safety. Accid. Anal. Prev. 62, 331–340.

Schepers, P., Twisk, D., Fishman, E., Fyhri, A., Jensen, A., 2017. The Dutch road to a high level of cycling safety. Saf. Sci. 92, 264–273.

Sivasankaran, S.K., Balasubramanian, V., 2020. Investigation of pedestrian crashes using multiple correspondence analysis in India. Int. J. Inj. Contr. Saf. Promot. 27 (2), 144–155.

Stelling-Konczak, A., van Wee, G.P., Commandeur, J.J.F., Hagenzieker, M., 2017. Mobile phone conversations, listening to music and quiet (electric) cars: Are traffic sounds important for safe cycling?. Accid. Anal. Prev. 106, 10–22.

Useche, S.A., Montoro, L., Alonso, F., Tortosa, F.M., 2018. Does gender really matter? A structural equation model to explain risky and positive cycling behaviors. Accid. Anal. Prev. 118, 86–95.