Journal of Safety Research 72 (2020) 127–138

Contents lists available at ScienceDirect

Journal of Safety Research

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

Exploring the severity of bicycle–vehicle crashes using latent class clustering approach in India

Sathish Kumar Sivasankaran, Venkatesh Balasubramanian *

RBG lab, Department of Engineering Design, IIT Madras, Chennai 600036, India

ARTICLE INFO

Article history:
Received 23 June 2019
Received in revised form 23 October 2019
Accepted 14 December 2019
Available online 31 December 2019

Keywords:
Bicycle Crashes
Latent Class Clustering
Crash Severity
Cluster Analysis

ABSTRACT

Introduction: Bicyclists are vulnerable users in the shared asset like roadways. However, people still prefer to use bicycles for environmental, societal, and health benefits. In India, the bicycle plays a role in supporting the mobility to more people at lower cost and are often associated with the urban poor. Bicyclists represents one of the road user categories with highest risk of injuries and fatalities. According to the report by the Ministry of Road Transport and Highways (Accidents, 2017) in India, there is a sharp increase in the number of fatal victims for bicyclists in 2017 over 2016. The number of cyclists killed jumped from 2,585 in 2016 to 3,559 in 2017, a 37.7% increase. *Method*: Few studies have only investigated the crash risk perceived by the bicyclists while interacting with other road users. The present paper investigates the injury severity of bicyclists in bicycle-vehicle crashes that occurred in the state of Tamilnadu, India during the nine year period (2009–2017). The analyses demonstrate that dividing bicycle-vehicle collision data into five clusters helps in reducing the systematic heterogeneity present in the data and identify the hidden relationship between the injury severity levels of bicyclists and cyclists demographics, vehicle, environmental, temporal cause for the crashes. *Results*: Latent Class Clustering (LCC) approach was used in the present study as a preliminary tool for the segmentation of 9,978 crashes. Later, logistic regression analysis was used to identify the factors that influence bicycle crash severity for the whole dataset as well as for the clusters that were obtained from the LCC model. Results of this study show that combined use of both techniques reveals further information that wouldn't be obtained without prior segmentation of the data. Few variables such as season, weather conditions, and light conditions were significant for certain clusters that were hidden in the whole dataset. This study can help domain experts or traffic safety researchers to segment traffic crashes and develop targeted countermeasures to mitigate injury severity.

1. Introduction

Cycling may be the most eco-friendly mode of transportation that can help ease the choking of cities. It is considered to be a viable solution to free up our roads from the burgeoning number of vehicles (Behnood & Mannering, 2017). Cycling is also a popular physical and recreational activity for people (Klassen, El-Basyouny, & Islam, 2014), which provide environmental and societal benefits (Rojas-Rueda, de Nazelle, Tainio, & Nieuwenhuijsen, 2011; Xia, Zhang, Crabb, & Shah, 2013; Macmillan et al., 2014) as well as health benefits (Kelly et al., 2014; Götschi, Garrard, & Giles-Corti, 2016). This unique mode of transportation provides potential benefits such as increased physical activity, decreased stress levels, reduced environmental pollution, fuel consumption, and congestion through a healthy, affordable, and enjoyable transport option (Pucher, Buehler, & Seinen, 2011). During the past century, both developing and developed nations have undergone a rapid transition toward urbanization, which has reduced and disfavored bicycle use (Schäfer, Heywood, Jacoby, & Waitz, 2009; Koglin & Rye, 2014). Having recognized its economic and environmental benefits, transport and urban professionals encourage and promote cycling while designing transportation facilities for urban and suburban use (Nabors, Goughnour, Thomas, DeSantis, Sawyer, & Moriarty, 2012).

Bicyclists share the same infrastructure with the motorized vehicles, yet bicyclists are the more vulnerable road user in a bicycle-vehicle collision since bicycles do not provide a protective environment in the event of a crash with motorized vehicle (Vanparijs, Panis, Meeusen, & de Geus, 2015). Bicyclists showed much higher injury risk (29 times higher) and fatality risk (10 times higher) relative to car occupants (Nilsson, Stigson, Ohlin, &

* Corresponding author.
E-mail address: chanakya@iitm.ac.in (V. Balasubramanian).

https://doi.org/10.1016/j.jsr.2019.12.012
0022-4375/

Strandroth, 2017). Hence, safety challenges associated with bicyclists are an important concern for transportation safety in a vehicle-bicycle crash. In India, the bicycle plays a role in supporting the mobility to more people at lower cost and are often associated with the urban poor. Bicyclists represents one of the road user categories with the highest risk of injuries and fatalities. According to the report by the Ministry of Road Transport and Highways (Accidents, 2017), there is a sharp increase in the number of fatal victims with respect to bicyclists in 2017 compared to 2016 in India. The number of persons killed raised from 2,585 in 2016 to 3,559 in 2017, an increase of around 37.7%. In 2017, bicyclist deaths made up 2.4% of all motor-vehicle fatalities in India.

A comprehensive literature review reveals that, though there is an abundance of research studying the safety of bicyclists, limited research has investigated the factors influencing the bicyclist injury severities in India. Earlier studies have investigated the crash risk perceived by bicyclists interacting with other road users (Chaurand & Delhomme, 2013). In this current study, we examine the factors that influence the bicyclist severity in crashes that involved a bicycle and motorized vehicle. Researchers generally rely on crash data to identify possible risk factors associated with severity, but usually crash data are too heterogeneous in nature, which makes it difficult to identify underlying patterns in the dataset. The objective of this study is to explore the application of LCC in reducing the heterogeneity present in the crash data for the bicycle-vehicle crashes reported in the state of Tamilnadu, and to identify the factors that influence the cyclists' severity using binary logit models. The present study also estimates and compares the effects of different contributory factors for the crashes including whole dataset as well for individual clusters.

The next section of this paper reviews previous studies related to this topic. Section 3 describes the methodology used in this paper. Section 4 describes the dataset used in the current study. Section 5 provides the results of the analyses and discussions regarding the action that can reduce bicyclist injury severity levels in Tamilnadu. The last section provides conclusions and recommendations for future studies.

2. Literature review

Various contributing factors have been identified in bicycle-vehicle collisions in the literature. Though falls and collisions are more frequent with non-motorized users, bicyclists' collisions involving a motor vehicles have identified majorly for the reported fatalities and serious injuries (Rowe, Rowe, & Bota, 1995; Nicaj et al., 2009; Chong, Poulos, Olivier, Watson, & Grzebieta, 2010; Sze, Tsui, Wong, & So, 2011; Rosenkranz & Sheridan, 2003, Schepers et al., 2015). The factors contributing to the severity of crashes have been studied at various levels: crash characteristics (type of collision, opponent vehicle), infrastructure characteristics (road type, road condition, road signage, road separation), cyclist demographics (gender, age and behavior), and environmental characteristics (day of the week, weather conditions, season).

Kim, Kim, Ulfarsson, and Porrello (2007) explored the factors contributing to the severity of bicyclists in a bicycle-vehicle collision using a multinomial logit model (MLN) for the police reported crashes at North Carolina, USA. They found that inclement weather, darkness with no streetlights, head-on collisions, morning peak hours, vehicle speed above 30mph, intoxicated driver, truck involved, intoxicated bicyclist, or bicyclist age above 55 increased the probability of fatal crashes by more than 16 fold. Yan, Ma, Huang, Abdel-Aty, and Wu (2011) used MLN and binary logit models to perform a comprehensive analysis of bicycle-motor vehicle crash in Beijing city. The results showed that better lighting conditions and the presence of median and division tend to reduce the injury severity in road segments. Moore, Schneider, Savolainen, and Farzaneh (2011) developed standard MLN and mixed logit models to estimate the degree of influence of different characteristics on bicyclist severity at the intersections and non-intersection locations in the state of Ohio, USA from 2002 to 2008. A study by Kaplan, Vavatsoulas, and Prato (2014) estimated the severity of the cyclist injury by using a generalized ordered logit model for collisions involving a cyclist and collision partner on Danish roads from 2007 to 2011. The results identified that cyclist fragility, cyclist intoxication, speed limits above 70–80 km/hr, slippery road surface, the location of the crash on the road sections, heavy vehicle involvement, and conflicts between cyclists going straight or turning left and other vehicle going straight are the aggravating factors. Another study by Klassen et al. (2014) identified significant factors affecting bicycle-motor vehicle collision severity. Their study included the interaction between roadway and approach-control type, the existence of partial crossways and bike signs, and the cyclist gender and age. Additionally, study by Behnood and Mannering (2017) showed that factors that potentially affect the likelihood of severe injuries in bicyclist/motor-vehicle crashes include bicyclist and driver race and gender, alcohol-impaired bicyclists or drivers, older bicyclist, riding on the wrong side of the road, bicyclist not wearing helmet, and drivers unsafe driving behavior.

Statistical methods are generally used to analyze the predictors of the injury severity, such as binary logit models (Yan et al., 2011), ordered logit or probit models (Kaplan et al., 2014), mixed probit models (Moore et al., 2011; Klassen et al., 2014), multinomial logit models (Kim et al., 2007; Moore et al., 2011; Yan et al., 2011; Behnood & Mannering, 2017). Data mining techniques such as clustering, classification and association rules have also been used for crash data exploration. Prati, Pietrantoni, and Fraboni (2017) used popular data mining techniques CHAID decision tree and Bayesian networks for Italian bicycle crashes from 2011 to 2013. The important predictors for the severity of the bicycle crash found through CHAID decision tree were road type, crash type, the age of the cyclist, road signage, cyclist gender, type of opponent vehicle, month, and type of road segment. Bayesian networks identified the crash type, road type and type of opponent vehicle as the most important predictors of bicycle severity crashes in their study. Depaire, Wets, and Vanhoof (2008) applied latent class clustering for segmenting the heterogeneous crash data set into homogenous subgroups as a preliminary analysis followed by injury analysis for each cluster to discover the underlying hidden relationships. Kaplan and Prato (2013) used latent class clustering, an unsupervised probabilistic clustering approach to study the cyclist-motor vehicle crashes in Denmark, which yielded 13 distinguishable cyclist motorist, latent classes. After clustering, the distribution of injury severity within each cluster was analyzed. The LCC models have also been used to segment databases and analyze vehicle crash data (Weiss et al., 2016; de Ona, López, Mujalli, & Calvo, 2013 and Kaplan, Vavatsoulas, & Prato, 2013) and pedestrian crash data (Mohamed, Saunier, Miranda-Moreno, & Ukkusuri, 2013; Sasidharan, Wu, & Menendez, 2015; Sun, Sun, & Shan, 2019).

3. Methodology

3.1. Latent class clustering analysis

Clustering is an unsupervised machine learning method that can aid in crash segmentation process (Depaire et al., 2008; de Ona et al., 2013; Mohamed et al., 2013). It also addresses the heterogeneity issue in the crash datasets by dividing the data into homogenous subgroups or clusters (Berry & Linoff, 1997). Cluster analysis is based on heuristics, which maximizes the similarity

between in-cluster elements and the dissimilarity between inter-cluster elements. Similarity-based clustering has two approaches. One approach is a distance-based clustering algorithm to identify homogenous subsets (partitioning approach; e.g., k-means clustering). Another approach is the hierarchical clustering (e.g., ward's method, a single linkage method). However, more recently probability based clustering approach, known as Latent Class Clustering (LCC) is used for data segmentation to identify homogenous subgroups. LCC is a portioning approach that divides the data into k-cluster and found to have many advantages over k-means clustering (Hair, Anderson, Tatham, & Black, 1998; Vermunt & Magidson, 2002). The advantages of the LCC approach are: (1) it is not required to specify the number of clusters beforehand; (2) allow us to use all type of variables (categorical, nominal or metric variables and the combination of them) without a standardization process; and (3) probability classifications are made by using subsequent membership probabilities, which are estimated through maximum likelihood method.

Consider a data sample of N crashes, measured with a set of observed variables $Y_1 \ldots Y_j$, which are considered as indicators of the latent variable X. These variables form a latent class model with K classes. If each of the observed value contains a specific number of categories Y_i contains D_i levels with i = 1, 2, ... j; and if the particular latent class is specified by index k, k = 1, 2, ...K, then the basic cluster form is given by

$$P_{Y_i} = \sum_{k=1}^{K} P_{x_k} P_{Y_i/x_k} \quad (1)$$

Where P denotes the probability of obtaining response variable Y_i, P_{Y_i} is the prior probability of membership in cluster k and P_{Y_i/x_k} is the conditional probability that the randomly selected case has a response pattern Y_i, given its membership in the k class of latent variable X.

$$P_{Y_i} = \sum_{k=1}^{K} P_{x_k} \prod_{i=1}^{j} P_{Y_i/x_k} \quad (2)$$

The method of maximum likelihood is used to estimate the model parameters. Once the model has been estimated, the cases are classified into a different class by using a Bayes rule to calculate the posterior probability by

$$P_{x_k/y_i} = \frac{P_{x_k} P_{Y_i/x_k}}{P_{Y_i}} \quad (3)$$

3.2. Number of clusters identification

Pearson or likelihood-ratio chi-square statistic is tested to estimate the goodness of fit of an estimated LC model. But sometimes the p values estimated cannot be valid when the frequency table becomes very sparse and, in such cases, parametric bootstrapping can be used. Another alternative approach is to use information criteria such as Bayesian Information Criterion (BIC) (Raftery, 1986), Akaike Information Criterion (AIC) (Akaike, 1987), and Consistent Akaike Information Criterion (CAIC) (Fraley & Raftery, 1998). In terms of AIC, BIC, and CAIC, the number of clusters is determined by one that minimizes the score of these criteria. In general, the lower the values, the better the model is because it adapts better to the data. According to some researchers, BIC is considered to be better than AIC and CAIC in determining the number of clusters (Biernacki & Govaert, 1999). However, some other researchers have suggested that minimum value cannot be reached when we increase the number of clusters, especially when analyzing a large sample from traffic databases (Bijmolt, Paas, & Vermunt, 2004). During such cases, the percentage reduction in BIC can be used to determine the number of clusters.

3.3. Severity analysis using logistic regression

The present section describes the method to identify the factors that influence the bicyclist crashes in Tamilnadu. The number of fatal crashes is less when compared to serious injury, minor injury, and property damages only. Further, during the clustering of data into smaller subgroups, the number of fatal crashes in the individual cluster will result in a low number. Therefore, in the present analysis, fatal crashes are combined with the serious injuries. Finally, there were two injury severity levels considered in this analysis – fatal/serious injury crashes and minor injury/non-injury crashes. Since the crash severity outcome is two, logistic regression can be applied for the current analysis.

A binary logistic regression model is more appropriate to determine the cause of contributing factors for the whole dataset as well for the individual clusters.

$$Y = logit(P) = ln\frac{p}{1-p} = \beta X \quad (4)$$

Where P is the probability, β is the vector of parameters and X is the vector of independent variables. When the independent variable increases by one unit keeping all other variables constant, the odds increase by the factor called as odds ratio and ranges from 0 to positive infinity. This indicates the relative amount by which the odds of the outcome increase or decrease when the value of the corresponding independent variable increases by one unit (Tay, Rifaat, & Chin, 2008).

3.4. Variable selection

Crash data used in this study were extracted from the RADMS database, the Road Traffic Accident Database of government of Tamilnadu. They are the only officially available source of crash data in Tamilnadu. Data were recorded and reported by the traffic police on scene who conduct assessment and provide feedback immediately to the headquarters. RADMS dataset used for this study covered all the bicycle-related crashes reported for the most recent years (2009–2017). The complete description of the dataset is provided in the next section. The retrieved crash database contains variables that are redundant in nature. For example, few data such as state code, zip code, etc. are identified and omitted from the analysis. To focus on the prime objective of this study, the research team identified the significant factors that influence the severity of bicycle crashes from the past studies. Each data includes variables describing the conditions that contributed to the crash and injury severity. The identified variables from the literature review include bicyclist characteristics (bicyclist gender, bicyclist age and bicyclist behavior), crash-related factors (collision type, number of lanes, road separation, Intersection, road category, road conditions, posted speed limit, contributing factors), traffic-related factors (traffic movement, traffic control), vehicle (colliding vehicle category), and environment-related factors (population setting, weather condition, light condition, region, season and day of the week).

The outcome variable was the binary indicator of severity levels of bicycle-vehicle collisions. The demographic characteristics of the bicyclist were gender (male or female), age was categorized into four groups (less than 17 yrs. [very young cyclists]; 18–25 yrs. [younger cyclists]; 25–55 yrs. [middle aged cyclists]; and above 55 yrs. [older cyclists]. The other variable available was the behavioral covariate of the bicyclist at the time of the crash, which was categorized into seven categories: stationary, while crossing traffic steam, while diverging/merging, while going ahead/overtaking, turning while starting/stopping, and unknown.

Tamilnadu traffic police have divided the state into four police zones. The zones are classified as north, south, central, and west

zone. North zone includes the jurisdiction of seven districts (Chennai, Cuddalore, Kanchipuram, Thiruvallur, Thiruvannamalai, Vellore, and Villupuram). West zone includes the jurisdiction of 8 districts (Dharmapuri, Krishnagiri, Namakkal, Salem, Erode, Thiruppur, Coimbatore, and Nilgiris). South zone include the jurisdiction of 10 districts (Madurai, Theni, Dindigul, Sivagangai, Virudhunagar, Ramanathapuram, Tirunelveli, Thootukudi, Kanyakumari, and Pudukkottai). The central zone includes the jurisdiction of seven districts (Ariyalur, Nagapattinam, Perambalur, Thanjavur, Tiruchy, Karur, and Thiruvarur). The seasons were classified as summer season (March–May), winter (December–February), monsoon (June–September), and post-monsoon (October–November). To account for the missing data in the crash dataset, the extra category was added to indicate that the information was not present. (e.g., traffic movement was coded as "one-way," "two-way," and "unknown").

4. Data

The data were obtained from the crash database maintained by the state transport planning commission of Tamilnadu refereed as RADMS database. RADMS database includes information related to road accident case reports from the police headquarters located throughout the state. The information recorded in the database is regarded as reliable because Tamilnadu police personnel investigate the crash on spot and enter the data in the registry. The responsibility of data collection, data recording, and reporting lies with Additional Director General of Police (ADGP), State Traffic Planning Cell (STPC) Tamilnadu. Further, trained police officials compile the crash data all over the state using the same instruction manual. The dataset was developed by merging three different datasets (crash data, driver data, and vehicle data). The crash dataset contains all crash information regarding collision type, road characteristics, and environment characteristics. The vehicle dataset provides detailed characteristics of each vehicle involved and vehicle specific information. The driver dataset demonstrates the demographic and behavior information of each driver involved in the crash. Each crash data has a unique accident identification number that is common in different datasets to link the records together. A comprehensive dataset containing information on including bicyclists' characteristics, crash characteristics, roadway characteristics, environmental characteristics, crash contributory factors, and traffic characteristics was created Table 1 gives the overview of the descriptive statistics of bicycle crashes and all variables used for this study.

The dataset includes 9,978 bicycle-vehicle reported crashes for which injury severity levels were reported. There were four categories of outcome: fatal, incapacitating or serious injury, non-incapacitating or minor injury, and no injuries. There were 1,855 fatal crashes, 1,050 serious injuries, 6,901 minor injury, and 174 no injuries cases. In this study fatal and serious injuries were classified into a single category, minor injury and no injuries were combined together. The final proportion of categories were 2,905 (29.11%) fatal/serious crashes and 7,074 (70.88%) minor injury/no injuries.

5. Results

5.1. Latent class Clustering:

Bicycle-vehicle collisions were all clustered by using the variables displayed in Table 2. To determine the appropriate number of clusters in the final model, a different number of clusters were tested from 1 to 12. The values of BIC, AIC, CAIC were used to select the number of clusters. From Fig. 1 it is clearly identified that as the number of cluster increases, the values for all three criteria decreases. Since the value does not reach the minimum for all the tested cases, the percentage reduction in BIC values was used (Scheier, Abdallah, Inciardi, Copeland, & Cottler, 2008). The percentage reduction in the BIC values computed for the different models decreases to less than 1% from the fifth cluster onwards. The quality of clustering was also assessed by calculating entropy value, which was close to 0.8, indicating the clear separation of the clusters (McLachlan & Peel, 2000). Also the values of CAIC and AIC support division of data into five clusters. Finally, the crash dataset was divided into five clusters for further analysis.

The final model was described by the proportion of each variable in each cluster. The cluster was analyzed and the names were given to each cluster based on the variable distribution similarly to work of Depaire et al. (2008). After determining the number of clusters, the next step is to characterize each cluster. For this purpose, important categories within each cluster for each variable are identified (using the highest conditional probability obtained for the determined category of a variable given its membership in a specific cluster).

The cluster profiles are shown in Table 3. For cluster 1, the variables are a number of lanes and contributory factor for the crash. Single lane roads and fault of the driver were overrepresented in Cluster 1. Hence cluster 1 (28.89%) is referred to as "crashes occurring on the two-lane roads due to the fault of the driver." Cluster 2 (29.87%) is similar to cluster 1 for a number of lanes and contributory factor, but distinguishes with cluster 1 by the overrepresentation of median separators and highways. Cluster 2 is hence referred to as "crashes occurring on the highways in presence of median separators." Three variables are specific to cluster 3: crashes occurring on two-lane roads, population setting is rural. Cluster 3 (15.34%) is hence referred to as "crashes occurring on rural two-lane roads." For cluster 4, the variables are the presence of median separators, the number of lanes, and the bicyclist behavior was diverging/merging with the traffic flow at the time of the accident. Cluster 4 (14.29%) is referred to as "crashes occurring on a single lane in the presence of median separators when the bicyclist merges/diverges with the traffic flow." The variables specific to cluster 5 are: crashes occurring on highways, contributory factor for the crash is the fault of the driver, and the bicyclist behavior was going ahead/overtaking in the traffic at the time of the crash. Cluster 5 (11.61%) is referred to as "crashes occurring on highways due to the fault of the driver when the bicyclist going ahead/overtaking the traffic at the time of the crash." Thus clustering has segmented the whole dataset into five homogenous subgroups and was useful in identifying the variables that have an influence on injury severity. Table 3 provides an overview of the identified clusters and the size of each cluster. The number of cases in each cluster ranges between 11% and 30% of the total sample size. Cluster 2 stands out with 2,980 cases, whereas cluster 3, cluster 4, and cluster 5 are similar in size with 1100–1550 cases. Cluster 1 has 2,882 cases of the whole dataset.

5.2. Injury severity analysis using binary logit models

As the goal of the current study is to explore which factors have an influence on the occurrence of fatal/severe crashes, binary logit models were applied in which the severity of the crash was considered as the dependent variable. The results of the model developed for the whole dataset as well for the individual clusters is presented in this section. The binary logit models were estimated for the whole dataset as well for five clusters by using maximization of log likelihood method. In all models, minor injury/non-injury was considered to be the base outcome. Twenty explanatory variables were entered into the model, including day of the week, season, region, collision type, number of lanes, road separation,

Table 1
Characteristics of the bicycle-vehicle crash by severity.

		Fatal/serious crashes	Minor injuries/No-injuries	Total Crashes
Day of the week	Weekend	8.9%	20.5%	29.4%
	Workday	20.2%	50.4%	70.6%
Season	Monsoon	9.7%	24.6%	34.2%
	Post Monsoon	4.4%	11.4%	15.8%
	Summer	7.5%	17.7%	25.2%
	Winter	7.5%	17.2%	24.8%
Region	Central	5.2%	15.2%	20.4%
	North	7.8%	21.8%	29.6%
	South	8.6%	20.2%	28.8%
	West	7.6%	13.7%	21.2%
Collision Type	Head On	12.0%	28.7%	40.7%
	Hit From Rear	12.1%	30.0%	42.1%
	Hit From Side	3.0%	7.6%	10.6%
	Others	1.6%	3.5%	5.1%
	Overturning	0.2%	0.5%	0.7%
	Ran Off Road	0.2%	0.2%	0.4%
	Side Swipe	0.0%	0.2%	0.2%
	Skidding	0.1%	0.1%	0.2%
Number Of Lanes	Multiple	2.7%	3.3%	6.0%
	Single	16.2%	58.7%	74.9%
	Two	9.7%	7.8%	17.5%
	Unknown	0.5%	1.1%	1.6%
Road Separation	No	11.5%	10.9%	22.4%
	Yes	17.6%	60.0%	77.6%
Intersection	Intersection	4.5%	4.7%	9.2%
	Non Intersection	20.2%	42.9%	63.1%
	Unknown	4.4%	23.3%	27.7%
Traffic Control	No Control	5.3%	11.8%	17.1%
	Not At Junction	23.1%	57.9%	81.0%
	Present	0.7%	1.2%	1.9%
Road Category	Highways (National/State/Express)	22.2%	59.5%	81.7%
	Major District Road	1.9%	0.8%	2.7%
	Other Roads	3.4%	5.2%	8.6%
	Unknown	0.2%	0.9%	1.1%
	Village Roads	1.4%	4.5%	5.9%
Light Conditions	Dark- Lighted	4.2%	11.2%	15.4%
	Darkness - Unlighted	2.5%	4.7%	7.2%
	Daylight	17.9%	43.0%	61.0%
	Twilight	4.2%	11.2%	15.4%
	Unknown	0.3%	0.8%	1.0%
Weather Conditions	Cloudy	0.2%	0.6%	0.8%
	Fine	28.7%	69.8%	98.5%
	Mist/Fog/Smoke/Dust	0.0%	0.1%	0.2%
	Others	0.1%	0.3%	0.3%
	Rainy	0.1%	0.1%	0.2%
Road Conditions	Good	28.8%	69.9%	98.7%
	Others	0.2%	0.2%	0.3%
	Poor	0.1%	0.2%	0.2%
	Unknown	0.1%	0.7%	0.8%
Traffic Movement	One-Way	1.2%	3.2%	4.4%
	Two-Way	27.9%	67.5%	95.4%
	Unknown	0.1%	0.2%	0.2%
Posted Speed Limit	30–40	26.3%	65.6%	91.9%
	40–50	1.0%	1.9%	2.9%
	55+	1.6%	2.9%	4.5%
	<30	0.2%	0.5%	0.7%
Population Setting	Rural	23.1%	57.4%	80.5%
	Urban	6.0%	13.5%	19.5%
Contributing Factors	Aggressive/Impaired Driving	0.3%	0.4%	0.7%
	Cause Not Known	2.4%	6.3%	8.7%
	Defect In Road Condition	0.0%	0.0%	0.0%
	Drunken Driver	0.0%	0.0%	0.1%
	Fault Of Bicyclist	0.1%	0.5%	0.6%
	Fault Of Driver	26.2%	63.7%	89.9%
	Others	0.0%	0.1%	0.1%
Colliding Vehicle	Bicycle	0.1%	0.9%	1.0%
	Bus	3.9%	6.9%	10.8%
	Heavy Goods Vehicle	5.7%	7.6%	13.3%
	Lmv	8.3%	19.3%	27.6%
	Motor Cycle	8.7%	32.9%	41.6%
	Unknown Vehicle	2.4%	3.3%	5.7%
Bicyclist Gender	Female	1.0%	3.9%	5.0%

(*continued on next page*)

Table 1 (*continued*)

		Fatal/serious crashes	Minor injuries/No-injuries	Total Crashes
	Male	28.1%	66.9%	95.0%
Bicyclist Age	18–25 yrs.	1.9%	6.8%	8.7%
	26–55 yrs.	14.1%	37.2%	51.3%
	55+ yrs.	10.2%	18.8%	29.0%
	<17 yrs.	2.9%	8.1%	11.0%
Bicyclist Behaviour	Stationary	1.2%	2.4%	3.6%
	Turning	1.5%	3.6%	5.1%
	Unknown	4.1%	9.8%	13.9%
	While Crossing Traffic Stream	0.7%	1.5%	2.2%
	While Diverging/Merging	7.0%	19.0%	26.0%
	While Going Ahead /Overtaking	14.0%	33.2%	47.2%
	While Starting/Stopping	0.4%	1.5%	1.9%

intersection, traffic control, road category, light condition, weather condition, road condition, traffic movement, posted speed limit, population setting, contributing factors, colliding vehicles, bicyclist age, bicyclist gender and bicyclist behavior. Due to uncertainty issues nominal variables coded as "others and unknown" are removed. The estimated coefficients and their significance values are shown in Table 4. The examination of results depends on the statistical significance of the coefficients of the independent variables. Similar to the studies of Kim et al. (2007), Depaire et al. (2008), Mohamed et al. (2013), Sasidharan et al. (2015), Sun et al. (2019), the significance value was chosen to be at 10%. In Table 4 the variables that are statistically significant were alone presented at 10% confidence level due to space constraints. All the cluster models and the whole dataset were found to be statistically significant. We conducted our analyses with statistical package R.

The predictors in the model with positive coefficient represent an increase in the probability of fatal and serious injuries compared to minor injuries and non-injuries. For the whole dataset, variables that significantly increase the probability of occurrence of fatal and serious injuries are as follows: workday, north Tamilnadu, hit from rear and side collision category, single lane roads, the presence of road separators, non-intersection spots, and uncontrolled junction spots. Additionally, the variables fault of the bicyclist, fault of the driver, and bicyclist behavior while diverging/merging and while going ahead/turning before the crash took place were found to increase the probability of fatal and serious crashes.

Two interesting observations can be viewed from the results of the injury severity analysis for the reported clusters and for the whole dataset. One is that some variables are significant for the whole dataset and for the individual clusters. The other is that variables are significant for the individual cluster but not for the whole dataset. The latter provides additional information by the clustering. For example, we find that weather conditions are not significant for the whole dataset, but significant for cluster 1 where the odds ratio is 0.178 for the rainy weather, which indicates that compared to baseline conditions (cloudy weather) bicyclist are less likely to receive fatal/serious injuries. This information was hidden in the whole data analysis but significant for cluster 1. Similarly, for cluster 2, the presence of traffic signage compared to the baseline condition of uncontrolled traffic is more likely to receive fatal / serious crashes by 110% (odds ratio = 2.157). In other words, bicyclists traveling on highways with median separators in the presence of traffic control are more likely to be involved in fatal/serious injuries. Speed limits between 30–40 kmph were not significant for whole data analysis. This factor is 0.2152 for cluster 3, which indicates that compared to low posted speed limits (<30kmph), bicyclists traveling on two-lane rural roads are 79.1% less likely to receive fatal/serious crashes. Also for cluster 3, crashes occurring in the daylight conditions are 14% more likely to be fatal/serious compared to crashes occurring in dark lighted conditions. Crashes occurring on a single lane in the presence of median separators when the bicyclist merges/diverges with the traffic flow (Cluster 4) is less likely to be fatal/serious by 31.5% (Odds ratio = 0.685) during the winter season compared to the summer season. All of these findings show that these variables were completely hidden in the full data model. Thus, cluster-based models reveal more information compared to the full data model.

From the results of the analysis performed, we also see that the effect of the significance of the variables is completely different for the whole dataset and for the individual clusters. For example, we find that the odds ratio for the crashes occurring during the workdays is 1.23 for whole data analysis, whereas it is 0.27 for cluster 1. Thus the whole data suggests that odds of crashes occurring on the workdays are 23% more likely to be fatal/serious than the baseline condition (crashes occurring during the weekends). However, the odds ratio estimated for the crashes occurring on the single-lane roads due to the fault of the driver (Cluster 1) is 73% less likely to be fatal/serious during the workdays compared to weekends.

Furthermore, earlier studies showed that the effect of some variables change direction between the cluster and for the whole data model and the reason for such behavior is unclear. These opposite effects cannot be understood clearly and it indicates that there is a need to validate some observations more closely. It also shows the interaction between network variables and bicyclist-involved crashes (Mohamed et al., 2013). Such a change of effect in variables is not found in the present study.

6. Discussions

The general conventional model, when used for analysis, would not have identified the hidden relationship between the injury severity analysis and influential factors due to the heterogeneous nature of the data. Therefore the LCC model was applied to separate the whole dataset into meaningful subgroups that helped in reducing heterogeneity. Five latent class clusters were identified for bicycle-vehicle collisions and severity models were developed for the identified clusters to determine the effect of influencing factors on the injury severity levels. The clusters were identified based on the variables number of lanes, road separation, road category, population setting, contributory factor for the crash, and bicyclist behavior just prior to the crash took place.

Few variables were found to have a direct dependence on injury severity both in the whole dataset and for the individual clusters. The variables include region, number of lanes, road separation, road category, and intersection. This implies that these variables are highly correlated with crash severity. It is found that the

Table 2
LCC results for the Bicycle- vehicle collisions (RADMS- 2009–2017).

		Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Work Day	Weekend	29.8	28.7	30.6	29.2	28.9
	Workday	70.2	71.3	69.4	70.8	71.1
Season	Monsoon	31.8	36.3	31.8	37.0	35.1
	Post Monsoon	15.6	16.7	14.9	15.1	16.2
	Summer	25.9	22.8	28.8	25.3	24.4
	Winter	26.7	24.2	24.5	22.7	24.3
Region	Central	23.6	20.3	21.7	26.4	2.6
	North	21.1	8.3	24.5	48.4	87.8
	South	17.0	57.2	28.8	19.5	0.1
	West	38.4	14.3	25.0	5.7	9.5
Collision Type	Head On	50.9	31.4	38.0	46.5	34.3
	Hit From Rear	32.0	50.8	43.3	33.4	55.3
	Hit From Side	9.3	11.4	11.6	11.5	9.3
	Others	6.3	4.5	5.6	6.8	1.0
	Overturning	0.6	0.9	0.6	1.3	0.0
	Ran Off Road	0.4	0.5	0.6	0.3	0.0
	Side Swipe	0.2	0.3	0.2	0.2	0.0
	Skidding	0.3	0.2	0.1	0.1	0.1
Number Of Lanes	Multiple	0.0	0.3	15.2	0.1	30.3
	Single	88.0	99.7	20.9	98.1	24.5
	Two	8.0	0.0	62.6	0.1	45.0
	Unknown	4.0	0.0	1.3	1.6	0.1
Road Separation	No	32.1	0.0	61.5	0.1	28.0
	Yes	67.9	100.0	38.5	99.9	72.0
Intersection	Intersection	9.8	4.9	25.9	3.8	2.3
	Non Intersection	85.2	48.6	72.3	10.8	96.9
	Unknown	5.1	46.4	1.9	85.5	0.9
Traffic Control	No Control	10.8	18.3	22.5	34.9	0.1
	Not At Junction	88.0	79.3	73.6	64.2	99.0
	Present	1.3	2.4	4.0	0.9	0.9
Road Category	Highways (National/State/Express)	53.8	98.9	81.5	95.4	93.0
	Major District Road	3.4	0.0	10.8	0.1	0.0
	Other Roads	24.3	0.1	6.7	0.9	2.2
	Unknown	1.3	0.2	0.4	3.1	1.0
	Village Roads	17.2	0.8	0.6	0.6	3.8
Light Conditions	Dark- Lighted	15.8	14.4	12.2	10.8	27.2
	Darkness - Unlighted	8.7	5.5	9.3	7.7	4.4
	Daylight	58.6	64.4	60.9	60.4	59.0
	Twilight	16.5	15.3	16.7	16.8	8.9
	Unknown	0.3	0.4	0.8	4.3	0.5
Weather Conditions	Cloudy	0.3	0.2	1.3	1.5	2.2
	Fine	99.2	99.2	98.0	98.0	96.6
	Mist/Fog/Smoke/Dust	0.0	0.1	0.1	0.1	0.8
	Others	0.3	0.2	0.4	0.3	0.4
	Rainy	0.1	0.3	0.2	0.0	0.0
Road Conditions	Good	97.9	100.0	99.2	96.2	99.8
	Others	0.8	0.0	0.7	0.0	0.0
	Poor	0.8	0.0	0.0	0.0	0.0
	Unknown	0.6	0.0	0.1	3.8	0.2
Traffic Movement	One-Way	4.8	2.3	4.7	3.7	8.9
	Two-Way	95.1	97.7	95.2	95.8	90.1
	Unknown	0.1	0.0	0.1	0.5	1.0
Posted Speed Limit	30–40	92.1	88.6	88.7	96.3	98.4
	40–50	3.4	2.5	4.4	2.6	0.7
	55+	3.4	8.4	6.2	0.8	0.3
	<30	1.1	0.5	0.6	0.3	0.7
Setting	Rural	82.6	80.3	84.9	83.5	65.6
	Urban	17.4	19.7	15.1	16.5	34.4
Contributing Factors	Aggressive/Impaired Driving	0.6	1.0	1.2	0.1	0.0
	Cause Not Known	2.0	1.8	9.7	40.0	1.6
	Defect In Road Condition	0.0	0.0	0.0	0.0	0.0
	Drunken Driver	0.1	0.0	0.1	0.0	0.0
	Fault Of Bicyclist	0.9	0.3	0.6	0.7	0.3
	Fault Of Driver	96.2	96.9	88.2	59.1	98.1
	Others	0.2	0.0	0.1	0.1	0.0
Colliding Vehicle	Bicycle	1.1	0.2	0.2	1.3	3.3
	Bus	8.1	11.7	15.1	11.1	8.8
	Heavy Goods Vehicle	12.4	11.6	22.2	11.1	10.6
	Lmv	25.5	28.2	31.2	26.3	27.9
	Motor Cycle	49.4	41.5	20.0	45.4	46.8
	Unknown Vehicle	3.4	6.7	11.3	4.9	2.6
Bicyclist Sex	Female	5.7	4.2	1.8	8.7	4.6
	Male	94.3	95.8	98.2	91.3	95.4
Bicyclist Age	18–25	9.1	8.3	5.5	10.5	11.0

(*continued on next page*)

Table 2 (*continued*)

		Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
	26–55	51.0	48.1	48.6	54.2	59.8
	55+	27.4	31.9	38.5	22.5	20.9
	<17	12.6	11.7	7.4	12.8	8.3
Bicyclist Behaviour	Stationary	7.6	2.7	3.2	0.3	0.1
	Turning	5.2	8.3	5.5	1.3	1.2
	Unknown	11.8	4.1	15.5	42.4	5.4
	While Crossing Traffic Stream	0.7	4.4	3.7	0.9	0.2
	While Diverging/Merging	28.1	15.5	24.2	46.0	23.8
	While Going Ahead /Overtaking	45.2	63.0	46.4	6.7	66.0
	While Starting/Stopping	1.3	2.0	1.5	2.5	3.3

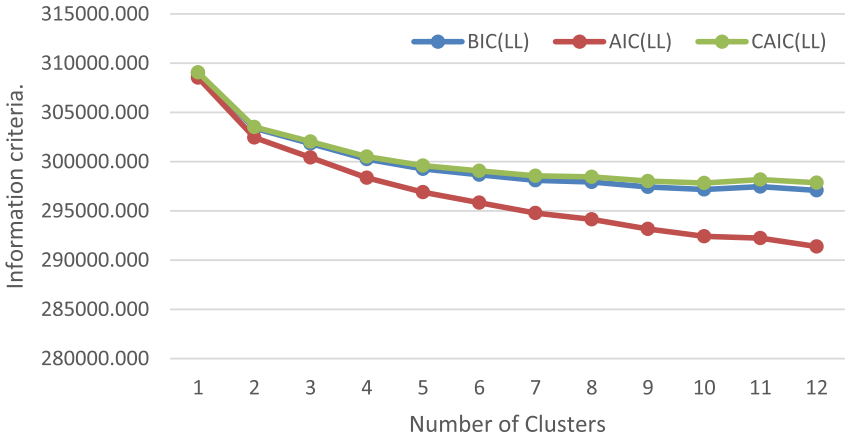

Fig. 1. Number of clusters identification.

Table 3
Cluster Summary.

Cluster	Characterization	Cases (%)
Cluster 1	crashes occurring on the single lane roads due to the fault of the driver	2882 (28.89)
Cluster 2	crashes occurring on the highways in presence of median separators	2980 (29.87)
Cluster 3	crashes occurring on rural two-lane roads	1531 (15.34)
Cluster 4	crashes occurring on a single lane in presence of median separators when the bicyclist merges/diverges with the traffic flow	1426 (14.29)
Cluster 5	crashes occurring on highways due to the fault of the driver when the bicyclist going ahead/overtaking the traffic at the time of the crash	1159 (11.61)

probability of fatal /serious crashes is high for single lane roads, (1.2–1.655), in the presence of median separators (0.69–1.064). Hence it is advised to take precautions against these crashes. Certain variables have a positive influence on the severity of fatal/serious crashes for the whole dataset and for clusters 1, cluster 2 and cluster 5. The variables include workday, region, collision type, traffic control, bicyclists' age, and bicyclist behavior just prior to the crash.

Results also show that combining latent cluster models along with binary logit models can identify the underlying hidden pattern and the logit models quantify the significance on the impact of certain contributory factors on injury severity outcomes of the bicyclists. Inferences from the results also show that certain variables that have not been identified as significant in the whole dataset are identified as significant for a specific cluster. For example, in cluster 2 presence of traffic control devices increases the probability of a fatal/serious crash by 110%, and for cluster 3 crashes occurring in the daylight are found to be critical.

The present study also confirms that segmenting the whole dataset into homogenous subgroups can help in identifying contributing factors that are hidden in the whole dataset. Such segmentation has helped in identifying targeted countermeasures that could be proved effective based on the specific reasons identified in those cluster of crashes. This also isolates the crash factors under certain conditions. Further, this study confirms that combined use of latent class models and binary logit models provide new information and insights on identifying contributing factors for crash severity. This is in line with the findings of previous studies that it is worthwhile to segment the data into homogenous subgroups before applying other analysis techniques (Sohn, 1999; Karlaftis & Tarko, 1998; Ng, Hung, & Wong, 2002; Wong, Leung, Loo, Hung, & Lo, 2004; Depaire et al., 2008). Further, the application of such latent class clustering can be extended to all types of crashes but not limited to bicycle-vehicle collisions for identifying better solutions for better safety. With respect to the limitations of the study, underreporting of minor injuries are common.

With the help of effective road safety strategic plan, developing nations like India must pay attention to bicycle-vehicle collisions to improve the safety of bicyclists. In order to reduce the severity of bicyclist crashes in Tamilnadu, targeted countermeasures at specific locations should be prioritized. Based on the scope of the study, we provide the following recommendations based on the results. Introduction of separate tracks for the bicyclists can reduce the severity of crashes and have proved to be beneficial in other developed nations based on previous research studies (Chataway, Kaplan, Nielsen, & Prato, 2014; De Rome et al., 2014; Kaplan et al., 2013, 2014). In India, there is no specific cycle track available exclusively for the bicyclist and hence they share up using the

Table 4
Binary Logit model estimation results for bicycle-vehicle collisions – whole data and clusters.

	Whole data			Cluster 1			Cluster 2			Cluster 3			Cluster 4			Cluster 5		
	Coeff	sig.	OR	Coeff	sig.	OR	Coeff	sig.	OR	Coeff	sig.	OR	Coeff	sig.	OR	Coeff	sig.	OR
Day of week (Ref: Weekend)																		
Workday	0.103	0.053	1.109	0.181	0.070	1.199												
Season(Ref:Monsoon)																		
Winter													−0.378	0.030	0.685			
Region (Ref: Central)																		
North	0.155	0.051	1.168	0.303	0.039	1.355												
South	−0.421	<0.001	0.656	−0.333	0.019	0.717	−0.258	0.069	0.772	−0.410	0.044	0.663	−0.556	0.008	0.573	−0.851	<0.001	0.427
West	−0.414	<0.001	0.661	−0.295	0.045	0.744	−2.772	0.005	0.663				−0.499	0.021	0.607	−0.720	0.003	0.486
Collision Type(Ref: Head On)																		
Hit From Rear	0.134	0.016	1.143													0.337	0.058	1.402
Hit From Side	0.186	0.032	1.204				0.467	0.004	1.596									
Ran Off Road	−0.724	0.050	0.485	−0.107	0.075	0.342												
Number Of Lanes (Ref: Multiple)																		
Single	1.268	<0.001	3.557	1.200	<0.001	3.319	1.313	<0.001	3.719	1.655	<0.001	5.238	1.063	<0.001	2.896	1.442	<0.001	4.228
Road Separation (Ref: N)																		
Y	0.720	<0.001	2.056	0.690	<0.001	1.994	0.624	<0.001	1.868	0.877	<0.001	2.404	0.775	<0.001	2.172	1.064	<0.001	2.899
Intersection(Ref: Intersection)																		
Non Intersection	0.519	<0.001	1.681	0.423	0.008	1.527	0.579	<0.001	1.785	0.931	<0.001	2.539	0.398	0.071	1.490			
Traffic Control (Ref : No Control)																		
Not At Junction	0.137	0.052	1.147				0.245	0.063	1.279									
Present							0.769	0.020	2.158									
Road Category(Ref: Highways(National/State/ Express))																		
Major District Road	−1.322	<0.001	0.267	−1.200	<0.001	0.301	−1.284	<0.001	0.277	−1.340	<0.001	0.260	−1.767	<0.001	0.171	−1.238	0.006	0.290
Other Roads	−0.521	<0.001	0.593	−0.416	0.008	0.660	−0.596	<0.001	0.551				−2.449	0.014	0.558	−0.940	0.002	0.391
Village Roads							0.542	0.017	1.721									
Light Conditions (Ref: Dark- Lighted)																		
Darkness - Unlighted	−0.289	0.008	0.748	−0.349	0.092	0.705							−0.554	0.067	0.574			
Daylight													−0.336	0.090	0.714			
Weather Conditions (Ref: Cloudy)																		
Others				−1.636	0.087	0.195												
Rainy				−1.724	0.065	0.178												
Posted Speed Limit (Ref :<30)																		
30–40										−1.535	0.003	0.215						
40–50	−0.714	0.004	0.490							−1.585	0.045	0.205						
55+	−0.874	0.010	0.417							−1.453	0.071	0.234				−2.310	0.056	0.099
Population Setting (Ref: R)																		
U	−0.211	<0.001	0.809	−0.427	<0.001													
Contributing Factors (Ref: Aggressive/Impaired Driving)																		
Drunken Driver																		
Fault Of Cyclist	1.249	0.005	3.490															
Fault Of Driver	0.790	0.006	2.205							0.970	0.095	2.637						
Colliding Vehicle (Ref: Bicycle)																		
Bus	−1.213	<0.001	0.297				−1.766	0.005	0.171							−2.396	0.026	0.091
Heavy Goods Vehicle	−1.498	<0.001	0.223	−1.271	0.049	0.280	−2.197	<0.001	0.111							−2.510	0.018	0.081
Lmv	−0.933	0.002	0.393				−1.423	0.024	0.241							−1.975	0.062	0.139
Cyclist Sex(Ref: F)																		
M	−0.229	0.077	0.795															
Cyclist Age (Ref: <17)																		
18–25	0.273	0.021	1.315	0.437	0.043	1.549												
55+	−0.343	<0.001	0.709				−0.435	0.012	0.647	−0.563	0.021	0.569	−0.440	0.079	0.644			

(*continued on next page*)

Table 4 (*continued*)

	Whole data			Cluster 1			Cluster 2			Cluster 3			Cluster 4			Cluster 5		
	Coeff	sig.	OR	Coeff	sig.	OR	Coeff	sig.	OR	Coeff	sig.	OR	Coeff	sig.	OR	Coeff	sig.	OR
Cyclist Behaviour (Ref: Stationary)																		
While Diverging/Merging	0.303	0.024	1.355				0.566	0.022	1.763									
While Going Ahead /Overtaking	0.292	0.025	1.339				0.446	0.060	1.563									
Intercept	0.256	0.694	1.292	1.096	0.379	2.992	0.073	0.954	1.076	−0.838	0.568	0.433	−1.308	0.522	0.270	3.693	0.106	40.168
Log likelihood at Zero	−12037			−3495.9			−3581.3			−1843.8			−1770.4			−1338		
Log likelihood at convergence	−10145			−2925.4			−2990.3			−1461.1			−1441.5			−1064.9		
Number of Observations	9978			2882			2980			1531			1426			1159		
ρ^2	0.157			0.163			0.165			0.207			0.185			0.204		
ROC	0.747			0.751			0.757			0.786			0.774			0.79		

roads with other road users. The results of the present study highlight the fact that 81% of the crashes occur on highways, which emphasize that exclusive cycle tracks are essential on the highways (including state, national, and express highways). The crashes are more common in highways because of higher speed differentials between the bicyclists and the motorists. Studies in the past also suggested having separate bicycle lanes in case of higher speed differentials (Kaplan et al., 2014; Kim et al., 2007; Klop & Khattak, 1999).

Secondly, most of the identified crashes occurred in the rural area (80%). The reason may be that the bicycle is still prevalent in rural households compared to urban areas. Fatality shares are higher on the rural roads compared to the average fatality share and it accounts for 23% of the total number of fatalities. Possible countermeasure to reduce the risk of collision in these areas would be to shift away the motorized vehicles from high cycling zones through a network level separation. A previous study also suggested that network level separation would reduce bicyclist exposure to motorized traffic (Schepers, Stipdonk, Methorst, & Olivier, 2017).

Thirdly, among the behavioral aspects of the bicyclists, the results of the current study show that when the bicyclist merges or diverges with the flow of traffic stream and during overtaking and moving ahead along the traffic alone account for 73% of the crashes.

Finally, for the benefit of bicyclists' safety, education campaigns and other educational programs should be conducted both for the bicyclists and other road users. Such programs should be effective, as a previous systematic review of these programs just increased the knowledge of bicyclist safety and has not translated into cycling safety (Richmond, Zhang, Stover, Howard, & Macarthur, 2014). Another common problem that increased the crash severity of the bicyclists, regardless of the nature of the crash pattern, is the neglect of the use of helmets. Providing proper bicyclists infrastructures, such as separate lanes with activated warnings, proper signage for motorists, and wide shoulders at the highways in the rural areas can improve bicyclists' safety.

7. Conclusions

This paper investigates the link between bicyclist injury severity outcomes of bicycle-vehicle collision and the contributing factors such as bicyclist demographics, bicyclist behavior, crash characteristics, and the built environment. Both LCC models and binary logit models were combined to identify the significant factors. Initially, the LCC model was used to segment the whole dataset into five clusters to increase the homogeneity, and then binary logit models were applied to reveal significant factors that influence severity outcomes in a bicycle-vehicle collision. Among the major findings are:

- Bicyclist behavior plays an important role in determining crash severity. Diverging and merging into traffic stream as well as going ahead/overtaking in regular traffic can increase the likelihood of fatal crashes.
- Variables that significantly increase the probability of occurrence of fatal and serious injuries are as follows: workday, north Tamilnadu, hit from rear and side collision category, single lane roads, the presence of road separators, non-intersection spots, and uncontrolled junction spots.
- Bicycle–vehicle crashes are more common in rural highways because of higher speed differentials between the bicyclists and the motorists.
- Bicyclists traveling on highways with median separators in the presence of traffic control are more likely to be involved in fatal/serious injuries.

- Some variables are influential in certain clusters indicating that those factors are significant in that cluster may be relevant only for safety in that context and may not pose as a safety concern in other conditions.

Acknowledgments

The authors wish to thank RADMS and TN police for providing the data. This research did not receive any specific research grants from the funding agencies in the public, commercial or not-for-profit sectors.

References

Akaike, H. (1987). Factor analysis and AIC. In: Selected Papers of Hirotugu Akaike (pp. 371–386). New York, NY: Springer.

Behnood, A., & Mannering, F. (2017). Determinants of bicyclist injury severities in bicycle-vehicle crashes: A random parameters approach with heterogeneity in means and variances. *Analytic Methods in Accident Research, 16*, 35–47.

Berry, M. J., & Linoff, G. (1997). *Data mining techniques: For marketing, sales, and customer support*. John Wiley & Sons Inc..

Biernacki, C., & Govaert, G. (1999). Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation, 64*(1), 49–71.

Bijmolt, T. H., Paas, L. J., & Vermunt, J. K. (2004). Country and consumer segmentation: Multi-level latent class analysis of financial product ownership. *International Journal of Research in Marketing, 21*(4), 323–340.

Chataway, E. S., Kaplan, S., Nielsen, T. A. S., & Prato, C. G. (2014). Safety perceptions and reported behavior related to cycling in mixed traffic: A comparison between Brisbane and Copenhagen. *Transportation Research Part F: Traffic Psychology and Behaviour, 23*, 32–43.

Chaurand, N., & Delhomme, P. (2013). Cyclists and drivers in road interactions: A comparison of perceived crash risk. *Accident Analysis & Prevention, 50*, 1176–1184.

Chong, S., Poulos, R., Olivier, J., Watson, W. L., & Grzebieta, R. (2010). Relative injury severity among vulnerable non-motorised road users: Comparative analysis of injury arising from bicycle–motor vehicle and bicycle–pedestrian collisions. *Accident Analysis & Prevention, 42*(1), 290–296.

de Ona, J., López, G., Mujalli, R., & Calvo, F. J. (2013). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis & Prevention, 51*, 1–10.

De Rome, L., Boufous, S., Georgeson, T., Senserrick, T., Richardson, D., & Ivers, R. (2014). Bicycle crashes in different riding environments in the Australian capital territory. *Traffic Injury Prevention, 15*(1), 81–88.

Depaire, B., Wets, G., & Vanhoof, K. (2008). Traffic accident segmentation by means of latent class clustering. *Accident Analysis & Prevention, 40*(4), 1257–1266.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal, 41*(8), 578–588.

Götschi, T., Garrard, J., & Giles-Corti, B. (2016). Cycling as a part of daily life: A review of health perspectives. *Transport Reviews, 36*(1), 45–71.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis*. New Jersey: Prentice-Hall International Inc..

Kaplan, S., & Prato, C. G. (2013). Cyclist–motorist crash patterns in Denmark: A latent class clustering approach. *Traffic Injury Prevention, 14*(7), 725–733.

Kaplan, S., Vavatsoulas, K., & Prato, C. G. (2013). Cyclist injury severity in a cycling nation: evidence from Denmark (No. 13-1547).

Kaplan, S., Vavatsoulas, K., & Prato, C. G. (2014). Aggravating and mitigating factors associated with cyclist injury severity in Denmark. *Journal of Safety Research, 50*, 75–82.

Karlaftis, M. G., & Tarko, A. P. (1998). Heterogeneity considerations in accident modeling. *Accident Analysis & Prevention, 30*(4), 425–433.

Kelly, P., Kahlmeier, S., Götschi, T., Orsini, N., Richards, J., Roberts, N., & Foster, C. (2014). Systematic review and meta-analysis of reduction in all-cause mortality from walking and cycling and shape of dose response relationship. *International Journal of Behavioral Nutrition and Physical Activity, 11*(1), 132.

Kim, J. K., Kim, S., Ulfarsson, G. F., & Porrello, L. A. (2007). Bicyclist injury severities in bicycle–motor vehicle accidents. *Accident Analysis & Prevention, 39*(2), 238–251.

Klassen, J., El-Basyouny, K., & Islam, M. T. (2014). Analyzing the severity of bicycle-motor vehicle collision using spatial mixed logit models: A City of Edmonton case study. *Safety science, 62*, 295–304.

Klop, J., & Khattak, A. (1999). Factors influencing bicycle crash severity on two-lane, undivided roadways in North Carolina. *Transportation Research Record: Journal of the Transportation Research Board, 1674*, 78–85.

Koglin, T., & Rye, T. (2014). The marginalisation of bicycling in Modernist urban transport planning. *Journal of Transport & Health, 1*(4), 214–222.

Macmillan, A., Connor, J., Witten, K., Kearns, R., Rees, D., & Woodward, A. (2014). The societal costs and benefits of commuter bicycling: Simulating the effects of specific policies using system dynamics modeling. *Environmental Health Perspectives, 122*(4), 335.

McLachlan, G., & Peel, D. (2000). Mixtures of factor analyzers. In: Proceedings of the Seventeenth International Conference on Machine Learning.

Mohamed, M. G., Saunier, N., Miranda-Moreno, L. F., & Ukkusuri, S. V. (2013). A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. *Safety Science, 54*, 27–37.

Moore, D. N., Schneider, W. H., IV, Savolainen, P. T., & Farzaneh, M. (2011). Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. *Accident Analysis & Prevention, 43*(3), 621–630.

Nabors, D., Goughnour, E., Thomas, L., DeSantis, W., Sawyer, M., & Moriarty, K. (2012). Bicycle road safety audit guidelines and prompt lists (No. FHWA-SA-12-018).

Ng, K. S., Hung, W. T., & Wong, W. G. (2002). An algorithm for assessing the risk of traffic accident. *Journal of Safety Research, 33*(3), 387–410.

Nicaj, L., Stayton, C., Mandel-Ricci, J., McCarthy, P., Grasso, K., Woloch, D., & Kerker, B. (2009). Bicyclist fatalities in New York City: 1996–2005. *Traffic Injury Prevention, 10*(2), 157–161.

Nilsson, P., Stigson, H., Ohlin, M., & Strandroth, J. (2017). Modelling the effect on injuries and fatalities when changing mode of transport from car to bicycle. *Accident Analysis & Prevention, 100*, 30–36.

Prati, G., Pietrantoni, L., & Fraboni, F. (2017). Using data mining techniques to predict the severity of bicycle crashes. *Accident Analysis & Prevention, 101*, 44–54.

Pucher, J., Buehler, R., & Seinen, M. (2011). Bicycling renaissance in North America? An update and re-appraisal of cycling trends and policies. *Transportation Research Part A: Policy and Practice, 45*(6), 451–475.

Raftery, A. E. (1986). A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society, Series B, 48*(2), 249–250.

Richmond, S. A., Zhang, Y. J., Stover, A., Howard, A., & Macarthur, C. (2014). Prevention of bicycle-related injuries in children and youth: A systematic review of bicycle skills training interventions. *Injury prevention, 20*(3), 191–195.

Road Accidents in India, 2017 Report, Ministry of Road Transport and Highways.

Rojas-Rueda, D., de Nazelle, A., Tainio, M., & Nieuwenhuijsen, M. J. (2011). The health risks and benefits of cycling in urban environments compared with car use: Health impact assessment study. *BMJ, 343*, d4521.

Rosenkranz, K. M., & Sheridan, R. L. (2003). Trauma to adult bicyclists: A growing problem in the urban environment. *Injury, 34*(11), 825–829.

Rowe, B. H., Rowe, A. M., & Bota, G. W. (1995). Bicyclist and environmental factors associated with fatal bicycle-related trauma in Ontario. *CMAJ" Canadian Medical Association Journal, 152*(1), 45.

Sasidharan, L., Wu, K. F., & Menendez, M. (2015). Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland. *Accident Analysis & Prevention, 85*, 219–228.

Schäfer, A., Heywood, J. B., Jacoby, H. D., & Waitz, I. A. (2009). *Transportation in a climate-constrained world*. MIT press.

Scheier, L. M., Abdallah, A. B., Inciardi, J. A., Copeland, J., & Cottler, L. B. (2008). Tri-city study of Ecstasy use problems: A latent class analysis. *Drug and Alcohol Dependence, 98*(3), 249–263.

Schepers, P., Agerholm, N., Amoros, E., Benington, R., Bjørnskau, T., Dhondt, S., & Niska, A. (2015). An international review of the frequency of single-bicycle crashes (SBCs) and their relation to bicycle modal share. *Injury Prevention, 21*(e1), e138–e143.

Schepers, P., Stipdonk, H., Methorst, R., & Olivier, J. (2017). Bicycle fatalities: Trends in crashes with and without motor vehicles in The Netherlands. *Transportation Research Part F: Traffic Psychology and Behaviour, 46*, 491–499.

Sohn, S. Y. (1999). Quality function deployment applied to local traffic accident reduction. *Accident Analysis & Prevention, 31*(6), 751–761.

Sun, M., Sun, X., & Shan, D. (2019). Pedestrian crash analysis with latent class clustering method. *Accident Analysis & Prevention, 124*, 50–57.

Sze, N. N., Tsui, K. L., Wong, S. C., & So, F. L. (2011). Bicycle-related crashes in Hong Kong: Is it possible to reduce mortality and severe injury in the metropolitan area? *Hong Kong Journal of Emergency Medicine, 18*(3), 136–143.

Tay, R., Rifaat, S. M., & Chin, H. C. (2008). A logistic model of the effects of roadway, environmental, vehicle, crash and driver characteristics on hit-and-run crashes. *Accident Analysis & Prevention, 40*(4), 1330–1336.

Vanparijs, J., Panis, L. I., Meeusen, R., & de Geus, B. (2015). Exposure measurement in bicycle safety analysis: A review of the literature. *Accident Analysis & Prevention, 84*, 9–19.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. *Applied Latent Class Analysis, 11*, 89–106.

Weiss, H. B., Kaplan, S., & Prato, C. G. (2016). Fatal and serious road crashes involving young New Zealand drivers: a latent class clustering approach. *International journal of injury control and safety promotion, 23*(4), 427–443.

Wong, S. C., Leung, B. S. Y., Loo, B. P., Hung, W. T., & Lo, H. K. (2004). A qualitative assessment methodology for road safety policy strategies. *Accident Analysis & Prevention, 36*(2), 281–293.

Xia, Ting, Zhang, Ying, Crabb, Shona, & Shah, Pushan (2013). Cobenefits of replacing car trips with alternative transportation: A review of evidence and methodological issues. *Journal of Environmental and Public Health, 2013*, 1–14. https://doi.org/10.1155/2013/797312.

Yan, X., Ma, M., Huang, H., Abdel-Aty, M., & Wu, C. (2011). Motor vehicle–bicycle crashes in Beijing: Irregular maneuvers, crash patterns, and injury severity. *Accident Analysis & Prevention, 43*(5), 1751–1758.

Sathish Kumar Sivasankaran received his master degree in mechanical systems design from IIITDM Kanchipuram in 2012 and is currently pursuing a Ph.D. at the Rehabilitation Bioengineering Group, Department of Engineering Design, IIT Madras. His current research focuses on road safety in India, transportation human factors, industrial and occupational ergonomics.

Venkatesh Balasubramanian is a Professor with the Department of Engineering Design, Indian Institute of Technology Madras, where he had founded the Rehabilitation Bioengineering Group (RBG), a team comprising physicians, health care professionals, engineers, and basic scientists. His research focus is on disruptive medical device development, human factors in product and process development, and innovation management. Dr. Balasubramanian mentors and advices for technology-based start-ups (both that have been incubated from or utilize technologies developed in RBG), SMEs, and automotive Tier-1 suppliers. He is also an Advisor with the Tamil Nadu Accident and Emergency Care Initiative in the state of Tamil Nadu, India.