Investigation of the Contributing Factors to Road Traffic Accidents In Thailand by Using Latent Class Analysis

Tarn Laochareonsuk[1], Sahassawat Runganothai[2] and Mongkut Piantanakulchai[3,*]

[1,2,3] School of Civil Engineering and Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, THAILAND

**Corresponding author; E-mail address: mongkut@siit.tu.ac.th*

Abstract

Traffic accident is a leading cause of death in Thailand. Understanding the patterns of accident helps determine the measures. Hence, Latent class clustering (LCC) analysis was conducted on 41,489 traffic road accident cases from the Transport Accident Management Systems (TRAMS) provided by the Ministry of Transport between 2021 and 2022 to show the patterns of traffic accidents. The analysis, which included six variables: road type, collision type, vehicle group, weather, time, and presumed cause, revealed four and five clusters of road accident patterns in the 2021 and 2022 datasets, respectively. Four comparable pairs of clusters are matched in those years. In addition, the vehicle group analysis was performed by LCC, giving similar results for six vehicle groups in both years. Vehicle group analysis showed the effect of vehicle variables on the characteristics of accident patterns as cluster to population ratio ***(C/P).*** The results suggest road safety policies should focus on the cluster of run-off-roadway accidents on straight sections caused by speeding. The LCC analysis also provides an advantage in the further application and policy evaluation, as progress can be tracked by the change in clustering behavior of the future accidents dataset.

Keywords: data mining, latent class clustering, road accident

1. Introduction

Road injury is one of the top causes of death in Thailand. According to the Ministry of Public Health, road injury ranked among the top 5 of the cause of death with 26.3 casualties per 100,000 people in 2020 [1]. The loss of lives and resources through road accidents impacts the overall economy and the quality of life of the people in the country. As many accidents occurred, the information on accidents was collected and used to create the accident database. Utilizing the information improves understanding of the accident and helps determine the measures. Recent research on the analysis of road accident topics from other countries uses road accident data which have many different aspects of road traffic characteristics such as traffic law, driver's behavior, and characteristics of road features. With many new advanced techniques, these studies have been proven effective in explaining contribution factors to road accidents in these local areas of study. However, directly employing the models from other research may not be suitable for Thailand due to differences in road traffic characteristics and available data. Some methods in previous studies might not be applicable to Thailand's road accident database because of lacking certain information.

For adapting any methods for Thailand's accident cases, compatibility between available data and method need to be reviewed. Recent researchers conducted analysis on the data requested from public organizations such as highway police departments and accident directorate offices, while few studies found alternative open data sources on social media platforms, online news, and onboard measuring equipment [2,3]. The accident information was categorized by the factors regarding each accident recorded at the crash sites. The factors included but not limited to time, location, accident type, weather condition, lighting condition, road characteristics, demographic information, and casualty report [2,3,4,5]. To obtain the road traffic accident contributing factors, various studies both used a single and two or more methods combined. Kaplan & Prato reported that the best results they obtained are from combining

two or more analytical methods, so the analysis of the outcomes is strengthened. In addition, more information that helps state road and traffic conditions should be included to increase the accuracy of the analysis [6]. De Oña et al. used a combination of Latent Class Clustering (LCC) and Bayesian Network (BN) and the analyzing result from the inferences of the Bayesian Network are the contributing factors of high severity accidents and recommendations for road safety; collisions with pedestrians occurred on rural highways, and sight distance is restricted by topography such as traffic signs and buildings [2]. Depaire et al. analyzed road traffic accident data that occurred in Belgium [7]. The heterogeneous data are processed by Latent Class Clustering, which revealed seven clusters that represent various types of traffic accidents. Clusters with high probabilities showed a link between accidents at crossroads without traffic lights, no priority roads, and with adult pedestrians included. Afterward, the clusters were used for accident severity analysis by Multinomial Logit models.

Upon reviewing the data collected by the Ministry of Transport (MOT), it was found that the majority of the recorded attributes are categorical. Hence, LCC is selected for analysis in this study. Clustering is a technique for dividing and organizing items into groups (clusters) searching for them to be distinct from other elements from other clusters while they have common attributes [3]. Clustering algorithms can be classified into many subtypes based on their similarity function. Latent Class Clustering is a clustering technique with distribution probability-based functions. It works with data with qualitative elements or discrete elements. The effectiveness of this method can be evaluated by many factors.

According to the Bureau of Highway Safety annual report, the current utilization of road traffic accident information by Thailand government agencies is only studying accident data on a single attribute [8]. The research on road accident in Thailand usually used logistic regression method to identify the factors that affect to probability and severity of the accidents [9,10,11,12]. Even though logistic regression methods can identify multiple significant factors, it cannot show which of these factors really occur together in nature. In fact, road traffic accidents contain numerous characteristics, thus assessing the accidents with a single characteristic is inappropriate. LCC enables utilizing all available data to the fullest as it can analyze several variables with different types at the same time. In addition, scope of Thailand traffic accident studies in the past are focused on a specific area [10,11,12] due to performance of the method which does not support large feature datasets, so their application stays local. The major benefits of LCC over other methods are reducing data discrimination, working well with large data sets, and providing graphically interpretable results [3]. LCC also reveals unobservable characteristics (latent variables) and the correlation among the variables. With results from LCC analysis, the patterns and contributing factors can be identified more specifically. Hence, instead of using many general measures, a particular solution for a specific pattern of major accidents can be obtained, enhancing the efficiency of budget consumption in road safety management. The results are also expected to be an indicator for evaluating road safety policies in the past as the change in road accident patterns reflects the effect of those solutions on road traffic accidents.

Therefore, this research was decided to use a different statistical technique, Latent Class Clustering on a countrywide scale accident database. The study is expected to achieve two objectives. The first objective is to uncover any hidden correlations among the fields of collected accident data attributes. To accomplish this task, we will employ LCC analysis, which is expected to reveal the contributing factors to traffic accidents. The second objective is to apply the outcome of the analysis for road safety management. Based on statistics of clustering results, recommendations will be provided to focus on important accident clusters. The prominent relationships found in each cluster direct the policies by eliminating these significant contributing factors and encourage detailed study on important specific accident patterns. Lastly the statistics of each cluster are expected to be used as indicators for evaluating the effectiveness of the policies.

2. Data

2.1 Data description

Thailand's Ministry of Transport manages the countrywide accident data in the Transport Accident Management System (TRAMS). The accident information is then published on the open government data site of Thailand [13]. This database includes all cases that occurred under the Department of Highways' responsibility and a few thousand cases under the Department of Rural Roads and Expressway Authority of Thailand's responsibility. In this study, sets of data in the range

of the latest two years (2021-2022) were used because of the changes in available data fields. The number of cases in the year 2021 and 2022 are 20,457 and 21,032 accordingly which are comparable amounts of sampling size varies with the different studies ranging from 3000 to 17,000 selected cases [2,4,14]. Using multiple data sets may indicate any change in the clustering of the data with different years. The information displayed several attributes that present the characteristic of each accident. The attributes chosen to be in the analysis are called variables. For controlling the degree of freedom of data, some variables were recategorized to reduce the complexity.

2.2 Variables

In the LCC analysis of accident population, six variables were included which are road type, collision type, time, weather, presumed cause, and vehicle group. Filtering the missing or corrupt data is the general procedure to ensure the quality and consistency of the result and improve workability in the analysis algorithm [3]. Hence, the variables are classified into fewer categories according to the Department of Highways' categorization [8]. Road type, collision type, and presumed cause contained a total of 23, 11, and 49 types of response and were recategorized to contain 10, 10, and 29 types respectively. Weather is the only variable that was used in the analysis without any modification. Time originally was a continuous variable and was categorized into 4 categories: morning peak (6.31-9.30 a.m.), mid-day (9.31 a.m. to 4.30 p.m.), evening peak (4.31-6.30 p.m.), and night (6.31 p.m. to 6.30 a.m.) according to the traffic volume recorded by Department of Highways [15]. The vehicle group was derived by LCC analysis of the 9 binary variables of the involved vehicles. Vehicle group was used instead of 1st vehicle variable for the reason that it reveals all the vehicles involved in the accident. Although there are additional variables in the data, like the location and highway numbers, these data contain wide ranges of information that may render the investigation too complex.

2.3 Other attributes

Killed or seriously injured (KSI) is another important attribute that is used to express the severity of the cluster in further analysis after LCC is done. A case is considered as "KSI" if it has at least one casualty or serious injury.

Table 1 Descriptive statistics of variables

Variables	Value	Total & (%)	
		2021	2022
Road type	Straight section	69.96	68.99
	Horizontal curve without elevation	11.50	10.76
	Horizontal curve with elevation	3.87	4.18
	Others	9.89	11.18
Collision type	Run-off-road on straight section	40.19	41.40
	Rear-end collision	29.05	30.05
	Run-off-road on curve	12.56	11.34
	Head-on collision	4.67	4.07
	Obstacle collision	3.43	3.56
	Others	7.71	7.14
Weather	Clear	85.26	84.98
	Rainy	13.54	13.74
	Fog/smoke/dust	0.64	0.59
	Overcast	0.36	0.29
	Natural disaster	0.06	0.04
	Others	0.12	0.36
Presumed cause	Violating Speed limit	74.82	72.83
	Sudden cutting off	7.84	7.32
	Drowsiness	4.76	6.04
	Vehicle defects	3.81	3.46
	Others	3.13	3.60
Time	Night	42.75	40.52
	Mid-day	32.66	33.78
	Morning peak	12.51	13.07
	Evening peak	12.08	12.63
Vehicle group	Pickup	36.97	37.05
	Car	24.34	26.72
	Motorcycle	13.83	12.32
	Trailer-Truck	10.44	7.46
	Others	8.37	8.52
	Truck	6.04	7.94

3. Methodology

3.1 Latent class clustering

Clustering is a machine learning technique that aims to divide a set of finite data into clusters by maximizing similarity among data in the same clusters [16]. There are mainly 2 types of clustering methods by similarity approach: distance base and probability base. Distance-based methods work only on continuous data. This type of method includes K-means clustering and K-nearest neighbor clustering, while probability-based methods can be used on many different types of data at once. Latent class clustering (LCC) is in this type of method.

Latent class clustering is a probabilistic approach technique that calculates the posterior probability of each data case from

their observed variables to assign them to a cluster. Interpreting result from LCC analysis inevitably forms subjective judgment like in other clustering methods [2]. However, the LCC's significant advantage over other techniques is that this method provides many statistical criteria that help in result interpretation. As this research hypothesis is that the determination of accident cluster characteristics will improve understanding of the accident patterns. LCC can extract the related factors in each cluster which may help decide the road safety policies from the statistical results.

Given data of N cases, if the j variables are included in the analysis, an arbitrary case x_i can be expressed as a set of variables $(V_1, V_2 \quad _j)$. Let each variable contains a specific number of possible responses: variable V_i contains T_i types of response, with . Therefore, the number of possible patterns of the parameter set equals to $\prod_{i=1}^{j} T_i$. Then each case will be assigned to a cluster of C-class models by its posterior probability. The posterior probability of a randomly selected case x_i belonging to cluster Z_j, $P(Z_j \mid x_i)$ is shown in Eq. (1),

$$P(Z_j \mid x_i) = \frac{P(Z_j)P(x_i \mid Z_j)}{P(x_i)} \quad (1)$$

for and ; where $P(Z_j)$ is a probability of an arbitrary case belonging to cluster Z_j. $P(x_i \mid Z_j)$ is a probability of finding case with the same response as x_i in cluster Z_j, and $P(x_i)$ is a probability of an arbitrary case give the same response as x_i.

snowLatent version 2.3.3 package in Jamovi (www.jamovi.org) organizes each case to the cluster with highest probability, like most analysis software. Therefore, a specific case may be assigned to a different cluster with the different percentage of members. The maximum log-likelihood is introduced to compare each partitioning for software to select the model of partition.

3.2 Number of cluster selection

The number of clusters from LCC is unknown from the beginning. Although increasing the number of clusters in an LCC model results in finer data partitioning, it is not always useful if the model causes more complexity. As a result, certain clusters involving only a few percentages of the dataset's population may be generated. Selecting the number of clusters aims to optimize the complexity of the model for a good interpretation and control over the result while the model is still proof of having a satisfy level of partitioning. The number of clusters was selected by 3 criteria.

Value of AIC, BIC, CAIC: Akaike information criterion (AIC) [17], Bayesian information criterion (BIC) [18], and Consistent Akaike Information Criterion (CAIC) [19] are common statistical parameters in the field of data modeling calculated from maximum log-likelihood. The lower value of them indicates the better partitioning of the model. Increasing the number of clusters is stopped, if the improvement of AIC, BIC, and CAIC is less than 1% [20]. Improvement of AIC, BIC, and CAIC value of the n-class model ($\%VI_n$) can be calculated by Eq. (2)

$$\%VI_n = \frac{V_{n-1} - V_n}{V_{n-1}} \times 100 \quad (2)$$

where V_n is statistical parameter (AIC, BIC, or CAIC) of model with n clusters.

Entropy: entropy expresses the distribution of data with values varied from 0 to 1. The closer value to one suggests better cluster separation. The model must have an entropy of at least 0.7 to be selected. The entropy of the C-class model with N cases of data is calculated by Eq. (3) [21].

$$I(j) = 1 - \frac{\sum_{i=1}^{N}\sum_{j=1}^{C} P(Z_j \mid x_i)\ln P(Z_j \mid x_i)}{N \ln(1/C)} \quad (3)$$

where $P(Z_j \mid x_i)$ is the posterior probability of a randomly selected case with pattern of x_i belonging to cluster Z_j.

Population of each cluster: Clusters with small numbers of members are not worth consideration as a domain-usefulness principle. All clusters' populations should be larger than 5% of the total population [22].

3.3 Cluster characteristic determination

Results of latent class clustering analysis from Jamovi are shown as posterior probabilities of each accident case belonging to each cluster. The software assigns the case to the cluster that gives the highest probability. Considering all cases in each cluster and their attributes that are used as variables in LCC analysis, the characteristics of each cluster can be obtained with the following criteria:

Maximum type: the type which has a percentage of more than 60% will be considered as a characteristic of the variable of the cluster.

Percentage ratio of in-cluster data to total data: a percentage ratio value that is close to 1.0 in most variable types indicates an insignificant change in each variable's types from clustering. If the ratios of most variable types are in the range of 0.9 to 1.1, the variable is excluded from being selected as characteristic of the cluster. The percentage ratio (C/P) of k type in variable v can be calculated by Eq. (4)

$$C/P = \frac{C_{k,v}}{P_{k,v}} \quad (4)$$

where $C_{k,v}$ is percentage of k-type in variable v in a cluster, and $P_{k,v}$ is percentage of k-type in variable v of total population.

4. Results

4.1 Vehicle group clustering

Jamovi gave very similar results of vehicle group analysis in both the 2021 and 2022 datasets. The improvement of AIC, BIC, and CAIC became less than 1% when the number of clusters increased to seven. The model of 6 clusters was selected to use in both datasets and satisfied the rest criteria. The clusters are named by their vehicle type with maximum probability which are pickup (VH1), passenger car (VH2), motorcycle (VH3), truck (VH4), trailer truck (VH5), and other (VH6). Five clusters contain their maximum vehicle type of 100% except for VH5 which is about 90%. By considering C/P, the effect of the vehicle group on the probability of accident characteristics compared to the total population can be inspected. The higher C/P value indicates the more significant effect of the vehicle group on other characteristics. However, C/P can be affected by the number of members in the cluster. Hence the characteristics with more than 5% of the total population are considered. Each vehicle group's characteristics and their effects on accident characteristics are shown in Table 2.

4.2 Cluster analysis

LCC analysis returned the value of AIC, BIC, and CAIC improvement of 0.95%, 0.78%, and 0.76% respectively once the model with five clusters was tested. The entropy criterion was also satisfied. Therefore, the four-cluster model was selected for explaining the 2021 accident dataset.

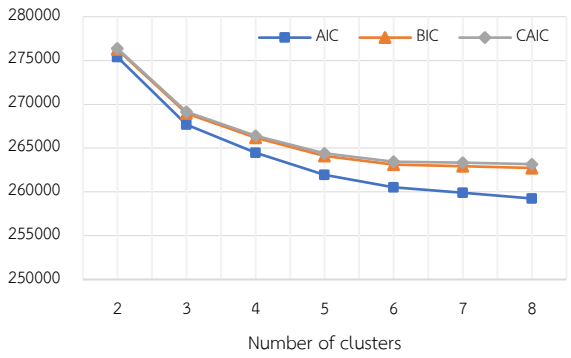

Fig. 1 Evolution of AIC, BIC, CAIC to the number of clusters in 2021

The description of the four clusters representing 2021 accidents is provided in the subsequent sections:

Cluster 1 (C1/21): This cluster consists of 91.05% of the cases on the straight section, which happened at mid-day (30.06%) and night (49.23%). Run-off-road is popular as a collision factor in the cluster (72.83%). Violating the speed limit is also commonly detected (84.76%). Furthermore, VH1 and VH2 are the main contributors to the accident in the cluster (67.42%). C1/21 could be labeled as "*Run-off accidents from speeding on straight section outside rush time by pickup and passenger car*".

Cluster 2 (C2/21): 64.91% of the accidents in this cluster occurred at the morning peak and early afternoon with clear weather. The most common collision type is the rear-end collision (83.28%), taking part by the VH1, VH2, and VH3 (89.75%). The straight section is featured in 68.89% of all cases. Note that this cluster is also reported with the highest suddenly cutting-off behavior (32.41%). The cluster could be called "*Rear-end collision on straight section in daytime.*"

Cluster 3 (C3/21): Most of the accidents in this cluster occurred on the curves (90.28%), together with the highest violating speed limit detected in all clusters (87.69%). VH1 and VH2 were included in 69.70% of the cases. Most vehicles were run-off-roadway with 83.28% included in the cluster. The weather condition variable was not distinct enough for identifying the relationship, but it can be noted that rain contribution is the highest of all clusters. "*Run-off on curved section by pickup and passenger car*" could be used to identify this cluster.

Cluster 4 (C4/21): The cluster included 61.12% of the accident that happened on the straight section. Most of the collision types were unspecified (57.72%) and obstacle collision. It is worth mentioning that the C/P of intersection collision is drastically increased (12.18) which contributes to 13.04% of all cases. The top vehicles involved are VH6 (40.91%) with VH3

Table 2 Member probability and characteristic of each vehicle group

Cluster	Member probability (%) [2021/2022]	Cluster characteristic (%) [2021/2022]	Effect on accident characteristic (C/P) [2021/2022]
VH1: Pickup	Pickup: [100/100] Motorcycle: [15.21/15.21]	Membership: [36.97/37.05] KSI rate: [17.82/16.44]	Collision type Run-off road on curve [1.30/1.31] Run-off road on straight section [1.22/1.24]
VH2: Passenger car	Passenger car: [100/100] Pickup: [21.40/24.51] Trailer truck: [5.08/5.14]	Membership: [24.34/26.72] KSI rate: [11.93/10.93]	Collision type Rear-end collision [1.25/1.30]
VH3: Motorcycle	Motorcycle: [100/100] Passenger car: [39.39/40.16]	Membership: [13.83/12.32] KSI rate: [41.46/40.41]	Collision type Head-on collision [1.25/1.23] Rear-end collision [1.35/1.32] Other collision [2.92/3.09] Time Evening peak [1.41/1.39] Presumed cause Suddenly cutting off [2.29/2.41]
VH4: Truck	Truck: [100/100] Pickup: [0.00/23.00] Motorcycle: [18.01/18.01]	Membership: [6.04/7.94] KSI rate: [20.55/18.93]	Collision type Run-off road on curve [1.32/1.16]
VH5: Trailer truck	Trailer truck: [90.62/88.98] Motorcycle: [10.21/10.21] Van: [5.81/10.21] Truck: [10.5/0.00]	Membership: [10.44/7.46] KSI rate: [19.48/15.88]	Collision type Obstacle collision [1.36/1.42] Run-off road on curve [1.28/1.64]
VH6: Other	Other: [100/100] Pickup: [17.70/20.81] Motorcycle: [9.52/9.52]	Membership: [8.37/8.52] KSI rate: [17.28/13.62]	Collision type Obstacle collision [2.77/2.68] Other collision [1.50/1.20] Time Morning Peak [1.45/1.44]

(32.12%). The collision type is also unidentified. In addition, the C/P of substance usage is raised significantly. This cluster could be labeled as "*Residual and unidentified cases*".

For 2022 dataset, LCC analysis gave value of AIC, BIC, and CAIC improvement of 0.53, 0.35, and 0.33% respectively when the number of clusters became six. Hence, model of five cluster was inspected by other criteria. Entropy of the model is 0.875 satisfying the criteria, while there is a small cluster containing only 3.34% of population which is not meet the criteria. Although the four-cluster model was selected, the small cluster with exactly the same members was still valid. This piece of evidence suggests the significant distinctiveness of the small cluster. Therefore, the five-cluster model was selected.

The description of the four clusters representing 2022 accidents is provided in the subsequent sections:

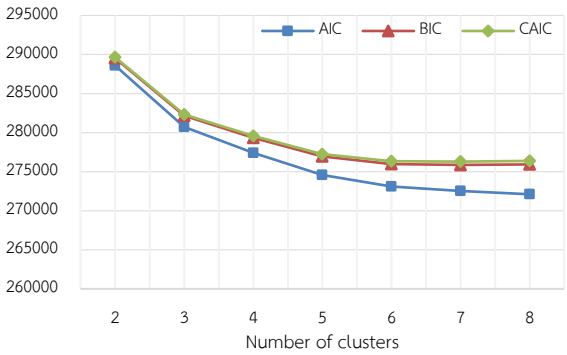

Fig. 2 Evolution of AIC, BIC, CAIC to the number of clusters in 2022

Cluster 1 (C1/22): it includes 86.79% of run-off-roadway. The accidents occur on the straight section without elevation (83.13%), and 80.21% of accidents occur in non-congestive traffic time (mid-day and night). The major vehicle groups in this cluster are VH1 (46.23%) and VH2 (23.93%). Violating the speed limit is the leading presumed cause of this cluster with 79.92% of the cases. C/P indicates an insignificant effect of weather on

this cluster. This cluster could be labeled as "*Run-off accidents from speeding on straight section outside rush time by pickup and passenger car*".

Cluster 2 (C2/22): it contains 75.47% of accidents occur on straight section without elevation, and 84.53% of its members are considered rear-end collision. Most of the cases occur in clear weather condition (95%) and daytime (27.77% probability of nighttime). The major vehicle groups of this cluster are VH2 (38.97%), VH1 (24.88%), and VH3 (19.99%). The major presumed causes are violating speed limit (73.83%) and suddenly cutting off (20.75%). This cluster could be called: "*Rear-end collision on straight section in daytime*".

Cluster 3 (C3/22): it contains 91.56% of accidents that occur on the curve with/without elevation which 81.85% of the cases are run-off-roadway. Weather conditions of this cluster are 69.61% clear and 29.33% rainy with C/P of 0.82 and 2.14 respectively. VH1 and VH2 are the major vehicle groups of this cluster with members of 47.86% and 22.57%. VH4 and VH5 group is considered 9.95% and 12.08% of the cases but they have C/P of 1.25 and 1.62 which are the highest among 5 clusters. The presumed cause of 83.53% of the cases is expected to be violating the speed limit. Time is not a feature of this cluster due to insignificant changes in variable types. "*Run-off on curved section by pickup and passenger car*" could be used to identify this cluster.

Cluster 4 (C4/22): it includes 59.66% of accidents occurred on straight section with/without elevation, and the cases with 82.67% of the population are unspecified or other collision types. Most members in this cluster come from vehicle groups VH3 and VH1 with the percentage of 43.73% and 19.95% respectively. There is not any dominant presumed cause, and C/P implies weather and time are not feature characteristics of this cluster. This cluster could be named: "*Residual and unidentified cases*."

Cluster 5 (C5/22): it consists of 66.10% of accidents that occur on straight section without elevation. The principal collision types of this cluster are rear-end collision (46.23%), and obstacle collision (19.49%). Half of the cases (50.42%) occur at night. Vehicle groups involved with this cluster are greatly dominated by VH6 (89.69%). While presumed causes of 97.32% are also other/unspecified reasons. C/P indicates an insignificant effect of weather on this cluster. This cluster could be explained as: "Other/*special vehicle cases.*"

Note that cluster naming represents only some of the most dominant factors, looking at the more specific details in each clustering assessment is recommended to uncover any hidden correlation. The summary of attribute probabilities used for describing distinctive clusters can be found in Table 3. Tables 4 and 5 illustrate the population, percentage of KSI, and cases with more than one vehicle involved.

Table 3 Variables with probabilities of each member in clusters

Variable - value	C1 (%)		C2 (%)		C3 (%)		C4 (%)		C5 (%)
	2021	2022	2021	2022	2021	2022	2021	2022	2022
Vehicle group: Pick-up	44	46	22	25	47	48	10	20	7
Vehicle group: Passenger car	23	24	34	39	23	23	8	17	0
Vehicle group: Motorcycle	5	6	33	20	6	4	32	43	0
Vehicle group: Others	7	6	2	6	3	3	41	7	90
Collision type: Run-off-roadway	73	87	0	0	83	85	5	0	8
Collision type: Rear-end collision	22	8	69	85	4	0	15	2	48
Collision type: Others	1	0	9	0	5	5	58	83	8
Road type: Straight section	91	89	69	76	0	1	61	52	66
Road type: Horizontal curve	0	0	3	1	90	88	14	10	13
Presumed cause: Speeding	85	80	61	74	88	83	23	37	0
Presumed cause: Unidentified	0	0	0	0	1	1	38	1	97
Presumed cause: Cutting off	1	0	32	21	1	1	3	13	0
Time: Mid-day	30	27	43	41	36	36	29	32	28
Time: Nighttime	50	53	20	28	35	37	48	43	50
Time: Morning-evening peak	20	20	37	31	28	27	23	25	21
Weather: Clear	84	82	91	95	72	70	87	89	89
Weather: Rainy	16	17	2	3	27	29	6	5	10

Table 4 Percentage of population, KSI and cases with more than one vehicle involved of each cluster in 2021

Cluster	Population (%)	KSI (%)	Cases with more than one vehicle involved (%)
1	54.57	16.28	25.00
2	22.17	27.45	93.43
3	15.09	16.01	12.15
4	8.17	31.40	45.69
Total	100.00	19.95	39.92

Table 5 Percentage of population, KSI and cases with more than one vehicle involved of each cluster in 2022

Cluster	Population (%)	KSI (%)	Cases with more than one vehicle involved (%)
1	46.72	13.43	12.42
2	28.61	23.25	97.91
3	13.86	12.83	9.81
4	7.44	41.11	48.02
5	3.37	2.12	67.66
Total	100.00	17.83	41.03

There are several aspects of the clustering characteristic that can be noticed from the analysis result. First, the clustering showed the four pairs of clusters that have comparable characteristics. Second, the severity of the accident is also distributed among different clusters despite the fact that this attribute was not the one of variables used for LCC. This information can be used for linking the factors of the accidents to how severe they are. For example, rear-end collision cases tended to have more severity. Moreover, it was found that the group with a higher KSI rate is the motorcycle. However, it is difficult to reveal any contributing factors in the cluster with unidentified components. Third, the cluster with the greatest population is the "Rollover/runoff accidents on the straight section outside rush time".

5. Conclusion and recommendation

This study aims to analyze the contributing factors to road traffic accidents by using latent class clustering analysis and to provide recommendations supporting road safety policies. The road accident datasets between 2021 and 2022 provided by Thailand's Ministry of Transport were used in this analysis. LCC shows the result of analyzing road accident cases with 6 variables included in four and five clusters of accidents in the 2021 and 2022 datasets respectively. Each cluster consists of the probability of factors that share a certain relationship, which then reveals the contributing factor in the individual group. The clusters in both years have identical attributes and can be matched to four equivalent pairs. The result suggests an insignificant change in accident behaviors in two years. Most clusters' characteristics are collision type, road type, and time which explicitly help understand major accident patterns. However, the contribution of vehicles to accidents is not clearly shown in this analysis. Hence, the vehicle group analysis was performed to inspect the effect of this variable. LCC gave a very similar result to the six-cluster model for vehicle group analysis in both years. Vehicle group analysis also shows the tendency of accident patterns in each group of vehicles.

Using results obtained from the analysis, the authors can apply them in the recommendation insight for traffic accident prevention policies. However, this is a very challenging task to focus on every accident type. The most dominant cluster was selected for the detailed breakdown. The cluster with the biggest size (54.57% and 46.72%) is related to run-off accidents caused by speeding. The inspection should be conducted at the location of the accident site to retrieve any road information regarding the safety issues. In this case, speed-control features should be especially focused. Then, try to install additional features to reduce the speeding behavior. After that, the accident pattern can be reevaluated after a period of time with the LCC analysis. In this manner, the effectiveness of the countermeasures can be determined by the change in population of the cluster with the same characteristics. The results from the LCC can also be combined with the other analytic approaches for expanding insight into the accident data. Implementation of the probabilistic model, such as the Bayesian network, and multinomial logit analysis are some of the methods for discovering detailed relationships in accident severity [2,14]. We hope for the possibility of applying these recommendations to the next study and traffic accident practices in the future.

There are several considerations and suggestions regarding the study. First, Jamovi currently cannot perform LCC on the continuous and discrete variables simultaneously. The users can only run categories at a time, which does not suit the dataset that may contain both types of variables. Second, the analysis outcome might be improved with more detailed data entry.

Research from other countries shows some of the factors that could be used for evaluating the contributing factors. Additional evidence such as shoulder style, crash barriers, driver info, and missing safety protocol is valuable data for further investigation [7,10]. The issue is much more concerned with the presumed cause factor. More than 75 percent of accident presumed causes come from driver's behavior, and there is only one attribute in the dataset that recorded this information. An annual report from the Bureau of Highway Safety showed the accident report form, which contains completed information regarding each factor in detail [8]. However, the TRAM system published only partial data compared with the data from full accident report form. The analysis performance could be improved by dividing the presumed cause into more subcategories or in checklists format such as any traffic law violation. Third, according to the analysis result, a cluster of residual or unidentified cases in each year contains about eight percent of the total accident. These clusters come from incomplete-data cases and poor categorization of the dataset for grouping unidentified and many specific categories altogether. In the data collection process, information indicating incomplete cases should be provided. This problem can be solved by using another analysis method such as the Bayesian network, and Random Forest classification, which are capable of predicting incomplete variables.

Acknowledgment

The authors are grateful to Dr. Setthaluth Pangkreung for continuously giving constructive advice and suggestions. Gratitude is extended to the School of Civil Engineering and Technology at SIIT for their indispensable support in facilitating this research. The first author would like to express his gratitude to NSTDA and SIIT for the scholarship provided.

References

[1] Ministry of Public Health (2020). *Public Health Statistics A.D. 2020*. Retrieved September 18, 2022, from: https://bps.moph.go.th/new_bps/sites/default/files/2563_0.pdf

[2] de Oña, J., López, G., Mujalli, R. and Calvo, F. J. (2013). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis and Prevention*, 51, pp.1–10. https://doi.org/10.1016/j.aap.2012.10.016

[3] Gutierrez-Osorio, C. and Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents: A review. *Journal of Traffic and Transportation Engineering (English Edition)*, *7*(4), pp.432–446. https://doi.org/10.1016/j.jtte.2020.05.002

[4] Abdullah, T. and Nyalugwe, S. (2019). A Data Mining Approach for Analysing Road Traffic Accidents. *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, pp.1–6. https://doi.org/10.1109/CAIS.2019.8769587

[5] Champahom, T., Jomnonkwao, S., Watthanaklang, D., Karoonsoontawong, A., Chatpattananan, V., and Ratanavaraha, V. (2020). Applying hierarchical logistic models to compare urban and rural roadway modeling of severity of rear-end vehicular crashes. *Accident Analysis & Prevention*, *141*, 105537. https://doi.org/10.1016/j.aap.2020.105537

[6] Kaplan, S. and Prato, C. G. (2012). Risk factors associated with bus accident severity in the United States: A generalized ordered logit model. *Journal of Safety Research*, *43*(3), pp.171–180. https://doi.org/10.1016/j.jsr.2012.05.003

[7] Depaire, B., Wets, G. and Vanhoof, K. (2008). Traffic accident segmentation by means of latent class clustering. *Accident Analysis and Prevention*, *40*(4), pp.1257–1266. https://doi.org/10.1016/j.aap.2008.01.007

[8] Bureau of Highway Safety (2022). *Traffic Accident on National Highways in 2021*. Retrieved September 20, 2022, from: http://bhs.doh.go.th/files/accident/64/report_accident_2564.pdf

[9] Lakman, T. and Jaiwongya, R. (2018). Factors Affect to Violence of Traffic Accidents in Responsible Areas of Highway 1st Offices. *National Undergraduate Conference on Statistics (NUCS2018),* Chiang Mai, Thailand, 12-13 March 2018, pp.168-179.

[10] Lonluai, P. and Charerntanyarak, L. (2022). Prevalence and Associated Factors of Motorcycle Accident Among Senior High School Students in Phukieo District, Chaiyaphum Province. *Community Public Health*, 8(3), pp. 119-130.

[11] Lamaigase, J., Siripaiboon, C. and Komonmalai, W. (2018). Risk Factors for Death of Road Traffic Accident Patients at the Emergency Department in the State Hospitals.

Boromarajonani College of Nursing, Suphanburi, 1(2), pp.66-78.

[12] Tanthong, S. and Nathapindhu, G. (2019). Prevalence and Associated Factors of Motorcycle Accident Among Senior High School in Namsom District Udonthani Province. *Disease Prevention and Control 9th Nakhon Ratchasima,* 25(2), pp.67-77.

[13] Ministry of Transport (2023). *TRansport Accident Management Systems: TRAMS*. Retrieved January 28, 2023, from: https://datagov.mot.go.th/dataset/roadaccident

[14] Çelik, A. K. and Oktay, E. (2014). A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey. *Accident Analysis & Prevention*, *72*, pp.66–77. https://doi.org/10.1016/j.aap.2014.06.010

[15] Thammasat University Research and Consultancy Institute (2020). *Road Safety Management Information System Development Project for the Department of Highways (Phase 1) (TIMS)*. Retrieved January 24, 2023, from: http://bhs.doh.go.th/files//Project/Tipam/finalreport_tipam.pdf

[16] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), pp.611–631. https://doi.org/10.1198/016214502760047131

[17] Akaike, H. (1987). Factor analysis and AIC. *Psychometrika* 52, pp.317–332.

[18] 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, pp.249–250

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

[20] Scheier, L. M., Abdallah, A. B., Inciardi, J. A., Copeland, J. and Cottler, L. B. (2008). Tri-city study of Ecstasy use problems: A latent class analysis. *Drug and Alcohol Dependence*, *98*(3), pp.249–263. https://doi.org/10.1016/j.drugalcdep.2008.06.008

[21] McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

[22] Nasserinejad, K., Rosmalen, J. V., de Kort, W. and Lesaffre, E. (2017). Comparison of criteria for choosing the number of classes in Bayesian finite mixture models. *PLoS ONE*, *12*(1). https://doi.org/10.1371/journal.pone.0168838