Indian Journal of Data Mining (IJDM)
ISSN: 2582-9246 (Online), Volume-2 Issue-1, May 2022

Criticality Trend Analysis Based on Different Types of Accidents using Data Mining Approach

Kumari Pritee, R. D. Garg

Abstract: ***Safety on roads and prevention of accidents are the prime concern of any highway system. Data mining is a source of retrieval of information for knowledge discovery approach. Many data mining methodologies have been applied to accident data in the recent past years. There is need to analyze the relationship between different factors related to accidents i.e. number of persons affected by fatal, minor, grievous, non-injury, road feature (ROF), road condition (ROC), cause of accident (CAU) and vehicle responsible (VR) according to daily, fortnightly, semi-fortnightly and monthly basis. The objective of this study is divided into three sub-objectives. The First sub-objective of this study is to divide number of accident dataset of National Highway sections of Karnataka state implemented by Project Implementation Unit i.e. PIU (Bangalore, Chitradurga, Dharwad, Gulbarga, Hospet and Mangalore) during January 2012 to January 2017 collected from NHAI (National Highway Authority of India) in homogeneous clusters using K-means clustering. The second sub-objective is to reflect the relationship between different factors i.e. a number of persons affected by fatal, minor, grievous, non-injury, CAU, ROC, ROF and VR using Apriori association rule. The last sub-objective is to perform temporal trend analysis for each cluster on the basis of rules generated by Association Rule Mining.***

Keywords: ***Road accidents, Data mining, K-means clustering Algorithm, Association Rule Mining and Temporal Trend Analysis.***

I. INTRODUCTION

Road accident is responsible for unnatural deaths, property damage, and disability. India is classified amongst larger accident rate countries in the world due to approx. 0.4 million accidents per year as stated by the MORTH report, 2014 1. According to this report, there is a minor decrease in accident rate from the year 2012 to 2013 but it is not confirmed whether this decreasing trend will continue in the future years or not. Due to increasing capacity of modern computer system to store more data, the size of the dataset has also increased day by day. However, there is need of fruitful research in order to analyze circumstances of accident occurrence.

Manuscript received on 28 March 2022 | Revised Manuscript received on 25 April 2022 | Manuscript Accepted on 15 May 2022 | Manuscript published on 30 May 2022.
*Correspondence Author
Kumari Pritee*, Faculty, Department of Computer Science and Engineering, IIIT Ranchi. (Jharkhand) India. Email: priteegeo.kumari23@gmail.com
R.D.Garg, Professor, Department of Geomatics Engineering, IIT Roorkee (Uttarakhand) India. Email:- garg_fce@iitr.ac.in

Data mining 2 is a set of techniques to extract implicitly unknown information and hidden patterns from a large amount of data i.e. searching for new and interesting hypothesis than confirming the present ones. Police stations only report the limited number of accident data which have occurred in their territories or regions covering a limited portion of highways. Raj 3 discussed that in India, the basic and non-analytical reports prepared at accident sites and the method used for collecting, compiling and recording the accident data requires major modifications because it is not updated as compared to other countries i.e. US and European countries. Data mining provides various functions related to transportation systems such as data analysis for road accident, road pavement, road roughness etc. Various studies in India 45 shows that road accident data were mostly collected manual for analysis and hence this data is limited and not long lasting. Because of this, data mining techniques proves not much fruitful to extract hidden information from this data. The studies of different countries have applied data mining for road accident data analysis and the outcomes have meaningful results 678910. NHAI is another source of road accident data in India which serves and keeps track of every accident record and cover information of entire State Highways (SH) road accidents. Basically, the number of victims involved in accidents is a basic measure of accident severity. It concludes accident severity type and major circumstances that influence the injury severity of accidents. Accident data collected from NHAI are very much heterogeneous in nature making the data tough to analyze. Data segmentation can be done to minimize the heterogeneity of data. In the present study, the large database having different records of road crashes have been analyzed for the highways of Karnataka state using data mining techniques so that policymakers can get ideas to improve highway safety. Various data mining techniques such as clustering, classification, and association rule mining have been applied to analyze road accident data, predict factor responsible for high-frequency accidents on the highway, and predict future trend. These are utilized for finding yet unrecognized and unsuspected facts to reduce highway accident 11. Data Mining enables to handle a huge amount of accident data stored in databases. It provides the integration of different technologies i.e. database technology etc. and different principles i.e. Associations, Sequential Patterns, Classifications, Predictions, and Clustering in different areas 1213. Cluster results have been implemented using Negative Binomial (NB) to correlate driver age's impact on road accidents 14.

Clustering has been applied for grouping of accident data into different clusters and additionally, probit model 15 has also been generated for the identification of relationship by correlating different accident characteristics. Moreover, Mohamed et al. 16 predicted the increase in the fatal crashes risk in Montreal Canada, based on the clustered dataset analysis with bad visibility due to bad weather as a major factor. The Euclidean k-medians have been used for Clustering based on k-means 17. Valent et al. 18 found that Sundays and holidays act as significant risk factors in clustered data analysis. Anderson 19 considered environmental characteristics for classification of accident hotspots into relatively homogeneous types. In addition, extraction of information using cluster-based analysis of road accident proved better rather analyzing data without clustering 20 21. Therefore, use of segmentation of road accidents is very necessary. Therefore, Cluster analysis has been introduced as a data mining technique for homogeneous groups of road accidents.

Kannov and Janson 22 reported an association between accident frequency and other factors i.e. road geometry, road features, vehicle responsible and traffic information for the effective solution of accident prevention. Lee et al. 23 illustrate statistical models for estimating the correlation between the accident and different traffic and geometric factors. The association-rule mining is basically used to estimate all strong association rules fulfilling user specification for minimum support and minimum confidence constraints 24. Geurts et al. 25 implemented an association rule data mining technique to mine different accident circumstances occurred on high-frequency accident locations in Belgium. Even though, regression models generally provide an association between dependent and independent variables with their own model-specific assumptions. Violation of assumptions may provide erroneous results 26. Certain relationships between independent and dependent variables with model specific assumptions have been created which when not followed may lead to errors and provide erroneous outcomes 27.

This paper represents data mining approach to explore road accident factors provided by NHAI by implementing K-means algorithm. After that association rule mining has been used for the identification of different situations using different factors associated with accident occurrence in different identified clusters. Association rule mining of different clusters using K-means clustering provides the hidden information by performing segmentation and association rules. The association rules are used to establish a relationship between a number of persons affected by fatal, grievous, minor and non-injury considering different categorical factors. Furthermore, a temporal trend analysis has also been implemented for four clusters i.e. a maximum number of persons affected by fatal, grievous, minor and non-injury which concludes different trends in different clusters on daily, semi-fortnightly, fortnightly & monthly basis using the maximum method of aggregation. The results can be utilized to put some accident prevention efforts in the areas identified for different categories of accidents to reduce the number of accidents.

II. DATA COLLECTION

Accident dataset collected by NHAI implemented by PIU (Bangalore, Chitradurga, Dharwad, Gulbarga, Hospet, and Mangalore), contain the accident details of different highway sections for Karnataka from January 2012 to January 2017 i.e. a total number of 6963 accident cases. The data collected was in an Excel file format with 16 attributes to individually describe each record as shown in Table 1. Attributes have been taken which are meaningful for the analysis and are categorized them into 4 clusters: maximum number of persons affected by fatal, grievous, minor and non-injury. Different highway sections are also listed in Table 1. The datasets analyzed here comprised several data sources obtained from NHAI for Karnataka implemented by PIU (Bangalore, Chitradurga, Dharwad, Gulbarga, Hospet, and Mangalore) during January 2012 to January 2017. Maximum attributes of the accident data are categorical attributes as given in Table 1.

Table 1 List of accident data attributes for different highway section in Karnataka

Factors	Sub-factors	Code	Factors	Sub-factors	Code
Month of Accident MOA	January	JAN	Nature of Accident NOA	Overturning	OVT
	February	FEB		Head on collision	HOC
	March	MAR		Rear end collision	REC
	April	APR		Collision brush/Side Wipe	CB/SW
	May	MAY		Right turn collision	RTC
	June	JUN		Skidding	SKD
	July	JUL		Others	Others
	August	AUG	Classification of Accident COA	Fatal	F
	September	SEP		Grievous Injury	GI
	October	OCT		Minor Injured	MI

	November	NOV		Non Injury	NI
	December	DEC	Cause of Accident	Drunken	DRUNK
Season of Accident	Summer	SUM	CAU	Over-speeding	OVSD
SOA	Rainy	RNY		Vehicle out of control	VOC
	Autumn	ATN		Fault of driver	FOD
	Winter	WNT		A Defect in mechanical condition of motor vehicle /road.	DIMC
Year of Accident	2012	-		No Property Damage	NPD
YOA	2013	-	Road Feature	Single lane	SL
	2014	-	ROF	Two lane	TL
	2015	-		Three lanes or more without central divider (median)	THL
	2016	-		Four lanes or more with central divider	FL
	2017	-		Straight road	SR
Day of Accident	Morning	-		Slight Curve	SC
DOA	Afternoon	-		Sharp Curve	SHC
	Evening	-	Road Condition	Flat Road	FR
	Night	-	ROC	Gentle incline	GIN
Highway Section	Hoskote-Dobbaspet	-		Steep incline	SIN
HS	Karnataka border	-		Hump	H
	Mulbagal-Karnataka	-		Dip.	D
	Neelamangala to Devihalli	-	Weather Condition	Fine	FN
	Hyderabad-Bangalore Section	-	WC	Mist/Fog	M/F
	Silk board to electronic city junction	-		Cloud	C
	Banglore - Neelamangala	-		Light rain	LR
	Banglore-Hoskote-Mudbagal Section	-		Heavy rain	HR
	Tumkur-Chitradurga	-		Hail/sleet	H/S
	Belgaum &Khanpur-Knt/Goa border	-		Snow	S
	Belgaum-Dharwad	-		Strong Wind	SW
	Maharastra Border-Belgaum	-		Dust Storm	DS
	Bijapur - Hungund Section	-	Vehicle Responsible	Light Motor Vehicles	LMVs
	MH/KNT Border Sangareddy	-	VR	Heavy Motor Vehicle	HMV
	Hungund-Hospet	-		Pedestrian	PD
	Kundapur-Surathkal	-		Medium Air Vehicle	MAV
	Devihalli-Hassan	-		Water Vehicle	WV
Number of persons affected by	Fatal Accident	Num_per_FA	Help Provided by Ambulance	Private Vehicle	PV
Num_per	Grievous Injury	Num_per_GA	HPA	Ambulance	AMB
	Minor Injury	Num_per_MA		Ambulance/Private Vehicle	AMB/PV
	Non-Injury	Num_per_NA			
Number of animals killed	Number of animals killed	NAK	Accident Severity	Critical	CR
NAK			AS	Non-Critical	NC

Retrieval Number:100.1/ijdm.C1618051322
DOI:10.54105/ijdm.C1618.051322
Journal Website: www.ijdm.latticescipub.com

Published By:
Lattice Science Publication (LSP)

III. METHODOLOGY

Data mining extracts individual patterns & trends among datasets which are not visible in a large amount of data 12. A data mining task provides the solution for different problems i.e. problem identification, data preprocessing, clustering and association pattern evaluation. The proposed implementation procedure for the data mining techniques consists of four steps: Pre-processing, Clustering, Association rule mining and Temporal trend analysis as shown in Figure 1.

A. Data Preprocessing

Data preprocessing 28 is the primary task to be done prior to analysis to get the data ready for analysis. As good data can only provide good results, data preprocessing becomes necessary prior to analysis. First of all, the data has been prepared for data mining before knowledge discovery. Several operations are performed in preprocessing such as the noise or outliers' removal, handling discrepancy in consistency properly i.e. missing values to improve the dataset quality, and transformation data into data structures to perform data mining on the Road Asset Management data. To perform this operation, three steps have been followed in this study. Data cleaning has been performed to diagnose the errors related to data volume, data type, and error rate and is removed by editing the data i.e. meaningless and contradictory data removal by removing missing values, noise and duplicate data, frequent misspells, space between two words or lack of necessary spaces, having hyphens. Additionally, data integration has been performed to combine data from heterogeneous sources into trusted data i.e. meaningful and valuable information. The accident data set includes a variety of data but required categorical and numerical attributes have been taken into account. Furthermore, data selection has been performed for the restoration of effective data for analysis. Finally, data conversion has been performed to convert the data into relevant forms for useful information extraction. After all these steps of preprocessing of data set, 6963 road accidents have been considered for further analysis .

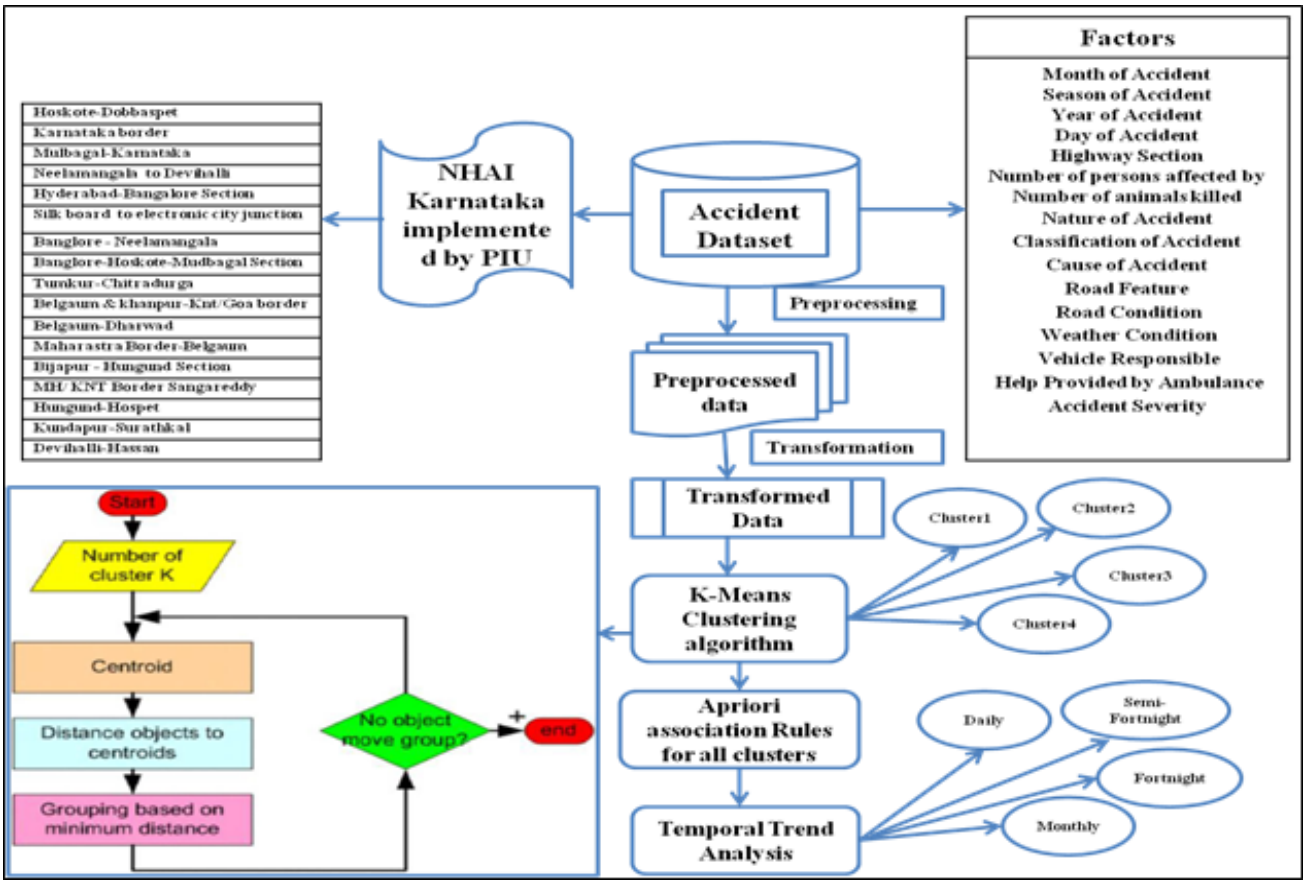

Figure 1 Architecture of Temporal Trend Analysis using Data mining approach

B. Clustering

Clustering is unsupervised learning or segmentation in which groups are not predefined but accomplished by determining the similarity among the data and similar items fall into same groups. Clustering algorithm has been applied to discover meaningful classes for unknown class labels. Here K-means clustering algorithm has been applied to discriminate the number of persons affected by accidents based on their different type of injuries. K-means is one of the simplest unsupervised learning data mining algorithms for clustering problem by forming *k* clusters of *n* objects. It depends upon high intra-cluster similarity but low inter-cluster similarity. Unsupervised learning technique means to describe unknown result clusters before the execution of clustering algorithm.

K-means algorithm steps 228:

- Randomly choose k objects from a data set S of n objects as the initial cluster means or centers.
- When all objects have been assigned, recalculate the positions of the k centroids and reassign each object which is distributed to a cluster on the basis of most similar or the nearer cluster center.
- Repeat above steps or relocating objects from one group to another group until the centroids update the cluster means by calculating the mean value of the object for each cluster or mean distance between the objects by using equation 1 which is basically used in minimizing WCSS (Within cluster sum of squares) i.e. sum of distance functions of each object to k centroids in a cluster.

$$\underset{S}{\arg\min} \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2 \quad \ldots\ldots(1)$$

where μ_i is the mean of points in S_i.
$x_1, x_2, \ldots, x_n$ as a set of observations.
S_i as a number of data points in i^{th} cluster $S = \{S_1, S_2 \ldots S_k\}$
k as cluster centers

- Output: A dataset of k clusters by calculating minimum distance between objects. K-means algorithm is a basis for all other clustering algorithms to calculate the mean values.

Table 2 Different Clusters According To Number Of Persons Affected By Different Injuries

Count	Number of Persons affected by Cluster_case1				Classified term
	FA	GA	MA	NA	
71 (VLFAC)	37	47	200	1895	Maximum persons affected by Non-injury
513 (LFAC)	150	1299	226	58	Maximum persons affected by Grievous injury
4291 (HFAC)	910	1404	1346	1240	Maximum persons affected by Fatal injury
2088 (MFAC)	133	272	3281	2173	Maximum persons affected by Minor injury

As shown in Table 2, the first cluster VLFAC (Very Low-Frequency Accident Count) indicates maximum persons affected by non-injury. The second cluster LFAC (Low-Frequency Accident Count) indicates maximum persons affected by grievous injury. The third cluster HFAC (High-Frequency Accident Count) indicates maximum persons affected by fatal injury. The fourth cluster MFAC (Moderate Frequency Accident Count) indicates maximum persons affected by minor injury.

C. Association Rules

Association rule mining 24 is a descriptive analytics technique that extracts interesting correlations among a different set of attributes in the data. It also discovers set of significant rules and underlying patterns showing variable category conditions that occur frequently together in a large dataset. By considering transactional databases of dataset D of n transactions for each transaction $T \epsilon D$, various relations occurred i.e. $X \subseteq T$, $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$, where $I = \{I_1, I_2 \ldots I_n\}$ be a set of items. An association rule is basically expressed in the terms of X and Y i.e. $X \Rightarrow Y$ (X satisfies Y). Various interesting measures such as support, confidence and lift, are there which identifies the strong rules with their quality. As shown in Equation 2, these interesting measures for a rule $X \Rightarrow Y$ can be defined as follows: The support count or frequency constraint indicates the frequency of occurrence of X and Y together in the data set or ratio of transactions that satisfy both X and Y to the number of transactions in databases. Support count is a very effective constraint to generate frequent itemsets by satisfying certain support threshold and those frequent item sets generate association rules by considering other measures. On the other hand, the confidence of a rule $X \Rightarrow Y$ defines the reliability of a rule i.e. ratio of the frequency of occurrence of X and Y to the frequency of occurrence of X. Higher the confidence values of above rule, higher the chances of frequent occurrence of Y with the frequent occurrence of X. The lift of above rule is the ratio of the observed support to expected support if X and Y were independent or a ratio of the confidence and the expected confidence of a rule. Normally it ranges from 0 to ∞. Lift values higher than 1 concludes highly potentially rule for prediction of the consequent in future data sets. Association rules generate all sets of items having support greater than the minimum support and after that generate the desired rules that have confidence greater than the minimum confidence using the large itemsets.

$$\text{Rule: } X \Rightarrow Y \rightarrow \begin{cases} Support = \dfrac{frq(X,Y)}{N} \\[2mm] Confidence = \dfrac{frq(X,Y)}{frq(X)} \\[2mm] Lift = \dfrac{Support}{Supp(X) \times Supp(Y)} \end{cases} \quad \ldots\ldots(2)$$

Here Apriori algorithm has been used for computing association rules because it is one of the most extensively used and popular techniques for outcomes of association rules 29.

Apriori Algorithm 1230

- Generate k frequent itemsets having length 1.
- Repeat till no new frequent itemsets are determined.
- Generate candidate itemsets of length (k+1) from frequent itemsets of length k or using large item set of the previous pass that is combined with itself to generate all itemsets having size higher than 1.
- Remove infrequent candidate itemsets having subsets of length k, leaving only those that are frequent.
- Calculate the number of k frequent candidate itemsets contained in a (k+1)-itemset.

Retrieval Number:100.1/ijdm.C1618051322
DOI:10.54105/ijdm.C1618.051322
Journal Website: www.ijdm.latticescipub.com

Published By:
Lattice Science Publication (LSP)

- Calculate the support count of each candidate by examining the Database.

Initially, the accident data has been divided into four categories namely HFAC, MFAC, LFAC, and VLFAC. Apriori algorithm has been used in R-Studio in order to generate strong association rules. Different minimum support for each group has been considered for generating association rules. The below tables indicate the association rules generated for each category considering some interesting measures. Here the factors i.e. CAU, ROC, ROF and VR have been considered as the main factor. The rules for HFAC, MFAC, LFAC and VLFAC considering two and three attributes are discussed as follows:

Cluster1: Rules for maximum persons affected by Non-Injury (VLFAC)

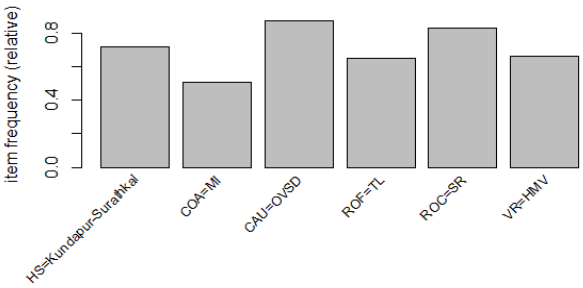

Figure 2 Histogram of Different Attributes of Accident Datasets

Table 3 Association Rules for Cluster1 Considering Two Factors

Rules	Support	Confidence	Lift
{ROC=SR} => {CAU=OVSD}	0.7042254	0.8474576	0.9704757
{ROF=TL} => {HS=Kundapur-Surathkal}	0.6338028	0.9782609	1.3618926
{VR=HMV} => {CAU=OVSD}	0.6056338	0.9148936	1.0477008
{HS=Kundapur-Surathkal} => {CAU=OVSD}	0.6056338	0.8431373	0.9655281
{HS=Kundapur-Surathkal} => {ROC=SR}	0.5915493	0.8235294	0.9910269
{ROF=TL} => {CAU=OVSD}	0.5633803	0.8695652	0.9957924

Table 4 Association Rules For Cluster1 Considering Three Factors

Rules	Support	Confidence	Lift
{HS=Kundapur-Surathkal, ROF=TL} => {ROC=SR}	0.5492958	0.8666667	1.0429379
{ROF=TL,ROC=SR} => {HS=Kundapur-Surathkal}	0.5492958	0.975	1.3573529
{HS=Kundapur-Surathkal, ROC=SR}=> {ROF=TL}	0.5492958	0.9285714	1.4332298
{HS=Kundapur-Surathkal, ROF=TL} => {CAU=OVSD}	0.5492958	0.8666667	0.9924731
{CAU=OVSD,ROF=TL}=> {HS=Kundapur-Surathkal}	0.5492958	0.975	1.3573529
{HS=Kundapur-Surathkal, CAU=OVSD} => {ROF=TL}	0.5492958	0.9069767	1.3998989

Figure 2 shows the relationship between different accident factors i.e. CAU, ROC, ROF, VR and HS through histogram. Here over-speeding is the main factor which is most responsible for the accident and for increasing the number of persons affected by Non-Injury. There is a strong correlation between CAU and ROC, the Moderate correlation between four factors i.e. CAU, ROC, ROF, VR and low correlation between five factors i.e. CAU, ROC, ROF, VR, and HS. Here, two and three categorical attributes of cluster 1 have been chosen to generate strong association rules using different interesting measures i.e support, confidence and lift as shown in table 3 and 4. The analysis of strong association rules shown in Table 3 and 4 for cluster 1 rule indicate 46.5 % multi-vehicle accidents caused due to over-speeding in two lanes straight roads in Kundapur Surathkal highway section in which a maximum number of persons are affected by non-injury. Kundapur Surathkal highway section is more vulnerable highway section for accidents and over-speeding in straight roads is the key factor for road accidents. The optional case is 53.5% Heavy motor vehicle (passenger, goods, and RPO & canter vehicle) accidents occur due to over speeding, drinking or vehicle out of control in one or more lane straight road. Although CAU, ROC, ROF, and VR are chosen as main factors, cluster1 rules show that over-speeding in two lanes straight roads is responsible for maximum persons affected by non-Injury by considering two and three categorical attributes in above rules. Therefore, over-speeding in two lanes straight road has been considered as a main categorical variable for temporal trend analysis.

Cluster2: Rules for maximum persons affected by Grievous Injury

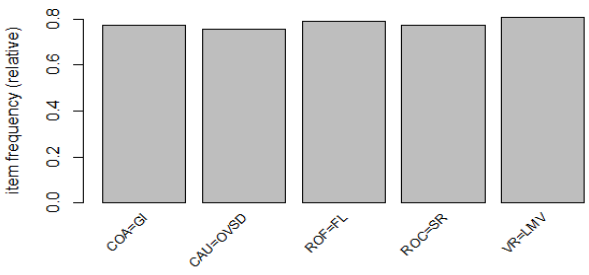

Figure 3 Histogram of different attributes of accident datasets

Table 5 Association rules for cluster2 considering two factors

Rules	Support	Confidence	Lift
{ROF=FL}=> {VR=LMV}	0.6549708	0.8296296	1.0280193
{CAU=OVSD}=> {ROF=FL}	0.6510721	0.8586118	1.0875750
{COA=GI}=> {VR=LMV}	0.6354776	0.8211587	1.0175227
{CAU=OVSD}=> {VR=LMV}	0.6179337	0.8149100	1.0097798
{ROC=SR}=> {ROF=FL}	0.6120858	0.7929293	1.0043771
{ROC=SR}=> {VR=LMV}	0.6081871	0.7878788	0.9762846

Table 6 Association rules for cluster2 considering three factors

Rules	Support	Confidence	Lift
{CAU=OVSD,ROF=FL}=> {VR=LMV}	0.5438596	0.8353293	1.035082
{ROF=FL,VR=LMV}=> {CAU=OVSD}	0.5438596	0.8303571	1.0950468
COA=GI,ROF=FL}=> {VR=LMV}	0.5107212	0.8424437	1.0438977
{COA=GI,VR=LMV}=> {ROF=FL}	0.5107212	0.803681	1.0179959
{ROF=FL,VR=LMV}=> {COA=GI}	0.5107212	0.7797619	1.0076017
{COA=GI,CAU=OVSD}=> {ROF=FL}	0.5009747	0.8538206	1.0815061

Figure 3 shows the relationship between different accident factors i.e. CAU, ROC, ROF, and VR through histogram. Here LMVs (Light motor vehicles) in four lanes is the main factor responsible for accidents and for increasing the number of persons affected by Grievous Injury. There is a strong correlation between ROF and VR, moderate correlation between four factors i.e. CAU, ROC, ROF, and VR. There is no low correlation. Here, two and three categorical attributes of cluster 2 have been chosen to generate strong association rules using different interesting measures i.e support, confidence and lift as shown in table 5 and 6. The analysis of strong association rules as shown in Table 5 and 6 for the cluster 2 rules indicates 40.35% LMVs (passenger, goods, commercial, ambulance, two wheeler) accidents caused due to over-speeding in four lanes straight roads in which maximum number of persons are affected by grievous injury and involves most of Neelamangala-Devihalli highway section. The optional case is 57.9% LMVs (passenger, goods, commercial, ambulance and two-wheeler) accidents occur due to over speeding, drinking or vehicle out of control in two or more lane straight road. Although CAU, ROC, ROF, and VR are chosen as the main factor, cluster 2 rules show that over-speeding of LMVs in four lanes straight road is responsible for maximum persons affected by grievous injury by considering two and three categorical attributes in above rules. Therefore, over-speeding of LMVs in four lanes straight road has been considered as a main categorical variable for temporal trend analysis.

Cluster3: Rules for maximum persons affected by Fatal Accidents

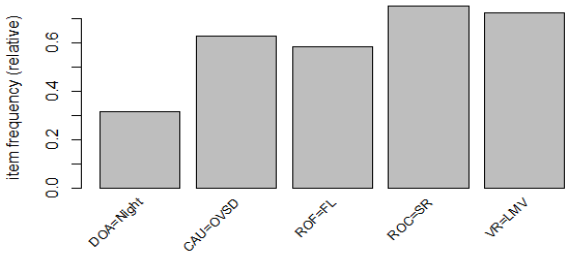

Figure 4 Histogram of different attributes of accident datasets

Table 7 Association rules for cluster3 considering two factors

Rules	Support	Confidence	Lift
{VR=LMV}=> {ROC=SR}	0.5273829	0.7281210	0.9712052
{CAU=OVSD}=> {VR=LMV}	0.4581683	0.7297699	1.0075426
{ROF=FL}=> {CAU=OVSD}	0.4551387	0.7802637	1.2428031
CAU=OVSD}=> {ROC=SR}	0.4551387	0.7249443	0.9669680
{ROF=FL}=> {ROC=SR}	0.4327663	0.7419097	0.9895973
{ROF=FL}=> {VR=LMV}	0.4264740	0.7311227	1.0094103

Table 8 Association Rules for Cluster3 Considering Three Factors

Rules	Support	Confidence	Lift
{CAU=OVSD,ROF=FL}=> {VR=LMV}	0.3309252	0.7270865	1.0038379
{ROF=FL,VR=LMV}=> {CAU=OVSD}	0.3309252	0.7759563	1.2359422
{CAU=OVSD,VR=LMV}=> {ROF=FL}	0.3309252	0.7222787	1.2382333
{CAU=OVSD,ROF=FL}=> {ROC=SR}	0.3295269	0.7240143	0.9657275
{ROF=FL,ROC=SR}=> {CAU=OVSD}	0.3295269	0.7614432	1.2128258
{CAU=OVSD,ROC=SR}=> {ROF=FL}	0.3295269	0.7240143	1.2412088

Figure 4 shows the relationship between different accident factors i.e. CAU, ROC, ROF, VR and DOA through histogram. Here LMVs in the straight road is the main factor responsible for maximum accidents and for increasing the number of persons affected by fatal Accidents. There is a strong correlation between VR and ROC, moderate correlation between four factors i.e. CAU, ROC, ROF and VR and low correlation between five factors i.e. CAU, ROC, ROF, VR, and DOA. Here, two and three categorical attributes of cluster 3 have been chosen to generate strong association rules using different interesting measures i.e support, confidence and lift as shown in table 7 and 8.

The analysis of strong association rules as shown in Table 7 and 8 for the cluster 3 rules indicate 52.7% LMVs (passenger, goods, commercial, ambulance, two wheeler) accidents caused on straight roads due to mostly over-speeding in four lanes roads of maximum Banglore-Hoskote-Mudbagal Highway Section in which maximum number of persons is affected by fatal accident. The optional case is 48.9% LMVs (passenger, goods, commercial, ambulance, and two-wheeler) accidents occur due to over speeding, drinking or vehicle out of control in two or more lane on straight roads.

Although CAU, ROC, ROF, and VR are chosen as main factors, cluster 3 rules show that Over-speeding of LMVs in mostly four lane straight road is responsible for maximum persons affected by fatal accidents by considering two and three categorical attributes in above rules. Therefore, Over-speeding of LMVs in mostly four lane straight road has been considered as a main categorical variable for temporal trend analysis.

Cluster4: Rules for maximum persons affected by Minor injury

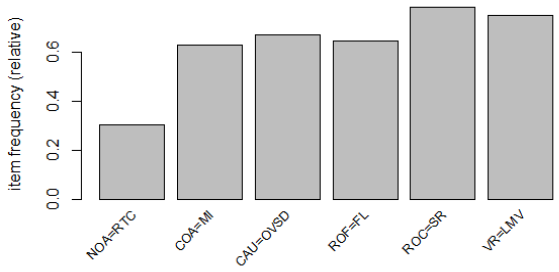

Figure 5 Histogram of different attributes of accident datasets

Table 9 Association rules for cluster4 considering two factors

Rules	Support	Confidence	Lift
{VR=LMV}=> {ROC=SR}	0.5790230	0.7725240	0.9871665
{CAU=OVSD}=> {ROC=SR}	0.5086207	0.7591137	0.9700302
{ROF=FL}=> {CAU=OVSD}	0.4966475	0.7664449	1.1439150
{ROF=FL}=> {ROC=SR}	0.4947318	0.7634885	0.9756206
{CAU=OVSD}=> {VR=LMV}	0.4861111	0.7255182	0.9679758
{COA=MI}=> {ROC=SR}	0.4846743	0.7719298	0.9864073

Table 10 Association Rules for Cluster4 Considering Three Factors

Rules	Support	Confidence	Lift
{CAU=OVSD,ROF=FL}=> {ROC=SR}	0.3697318	0.7444552	0.9512989
{ROF=FL,ROC=SR}=> {CAU=OVSD}	0.3697318	0.7473379	1.1153977
{CAU=OVSD,ROC=SR}=> {ROF=FL}	0.3697318	0.7269303	1.1218259
{COA=MI,VR=LMV}=> {ROC=SR}	0.3649425	0.7552032	0.9650332
{COA=MI,ROC=SR}=> {VR=LMV}	0.3649425	0.7529644	1.0045941
{ROC=SR,VR=LMV}=> {COA=MI}	0.3649425	0.6302730	1.0038215

Figure 5 shows the relationship between different accident factors i.e. CAU, ROC, ROF, VR and NOA through histogram. Here LMVs in the straight road is the main factor responsible for maximum accidents and for increasing the number of persons affected by minor injury. There is a strong correlation between VR and ROC, moderate correlation between four factors i.e CAU, ROC, ROF and VR and low correlation between five factors i.e. CAU, ROC, ROF, VR, and NOA. Here, two and three categorical attributes of cluster 4 have been chosen to generate strong association rules using different interesting measures i.e support, confidence and lift as shown in table 9 and 10. The analysis of strong association rules as shown in Table 9 and 10 for the cluster 4 rules indicate 74.95% LMVs (passenger, goods, commercial, ambulance, two wheeler) accidents caused mostly due to over speeding in four-lane road of maximum Hyderabad-Bangalore Section in which a maximum number of persons is affected by minor injury. The optional case is 53.3 % light motor vehicle accidents (passenger, goods, commercial, ambulance, and two-wheeler) occurs due to over speeding, drinking or vehicle out of control in two or more lane straight roads. Although CAU, ROC, ROF, and VR are chosen as the main factor, cluster 4 rules show that Over-speeding of LMVs in mostly four lane straight road is responsible for the maximum persons affected by Minor injury by considering two and three categorical attributes in above rules. So, Over-speeding of LMVs in mostly four lane straight road has been considered as a main categorical variable for temporal trend analysis.

D. Temporal Trend Analysis

Trend analysis or Time series analysis has been defined as an effective technique for information collection for both partly or completely hidden by noise, for extracting an underlying pattern of accidents and to perform time series trend by obtaining evenly spaced time points 3132. Even though trend analysis is basically applied to predict future events but it has been proved good for determining uncertain and unpredictable past events on the basis of data such as predict accident count between two dates.

The trend can be of different types e.g. linear trend analysis, nonlinear trend analysis etc. Linear trend analysis can be summarized in terms of regression analysis, as explained in trend estimation 33. Other than linear trend analysis, different shape trend testing can be done by non-parametric techniques e.g. Mann-Kendall test etc. Smoothing can also be used for testing and visualization of nonlinear trends. Aggregation methods are types of calculations used to group attribute values into a metric for each dimension value. Here a regularized time series trend has been built in which aggregation of raw data has been interfaced using an appropriate method considering max value for the same period (day, week, month and year).

This method is implemented by the "pastecs" package 34. For final time series, data frequency has been selected that allows the selection of the time step using 'maximum method of aggregation'. Different options for classic frequencies are available: daily, semi-fortnightly, fortnightly, monthly and yearly. Using this, all raw data has been aggregated by daily, semi-fortnightly, fortnightly, monthly. Time step has been chosen on the basis of the theoretical sampling frequency of raw data for getting maximum information without any loss by creating too many missing values. Time step has been chosen by computing the minimum and maximum period with the meantime. In this regularized time series trend using maximum method of aggregation algorithm as shown in table 11, the interface recommended the daily time step if mean time between two observations is less than 5 days; semi-fortnightly time step for mean time between 5 to10 days; fortnightly time step between 10 to 23 days; monthly time step between 23 to 60 days and over 60 days Yearly (annual) time step is recommended.

Table 11 Maximum method of aggregation algorithm

Maximum	Selects the maximum value for the metric.
	Used for numbers, dates, times, and durations.
	Used for multi-value attributes.

After that maximum method of aggregation has been selected for raw data that enables trend analysis aggregated data at the previously chosen time step. Further, regularized time series trend has been built through a plot or a table that will be automatically saved.

IV. RESULTS AND DISCUSSIONS

Cluster1: Figure 6 shows temporal trend analysis of a number of persons affected by non-injury over the years on the (a) daily, (b) semi-fortnightly, (c) fortnightly and (d) monthly basis due to over-speeding in two-lane straight roads. It is observed that most frequent non-injury accident and maximum persons affected due to non-injury correspond to first six months in 2014 on the daily and semi-fortnightly basis as shown in figure 6(a) & 6(b). However, on the fortnightly and monthly basis, as shown in figure 6(c) & 6(d), it is observed that most frequent non-injury accident and the maximum persons affected due to non-injury also correspond to the first six months in 2014. There is a distributed number of persons affected by non-injury, not on regular basis but there are some gaps between the number of persons affected by non-injury over the years on the different time series basis. In 2013 and 2014, mostly persons are affected by non-injury accident due to due to over-speeding in two lanes straight roads. After 2014, the trend has been decreased in 2015 and 2016. By considering past trend in 2015 & 2016, a number of persons affected due to non-injury on the daily basis may be decreased in future due to due to over-speeding in two lanes straight roads.

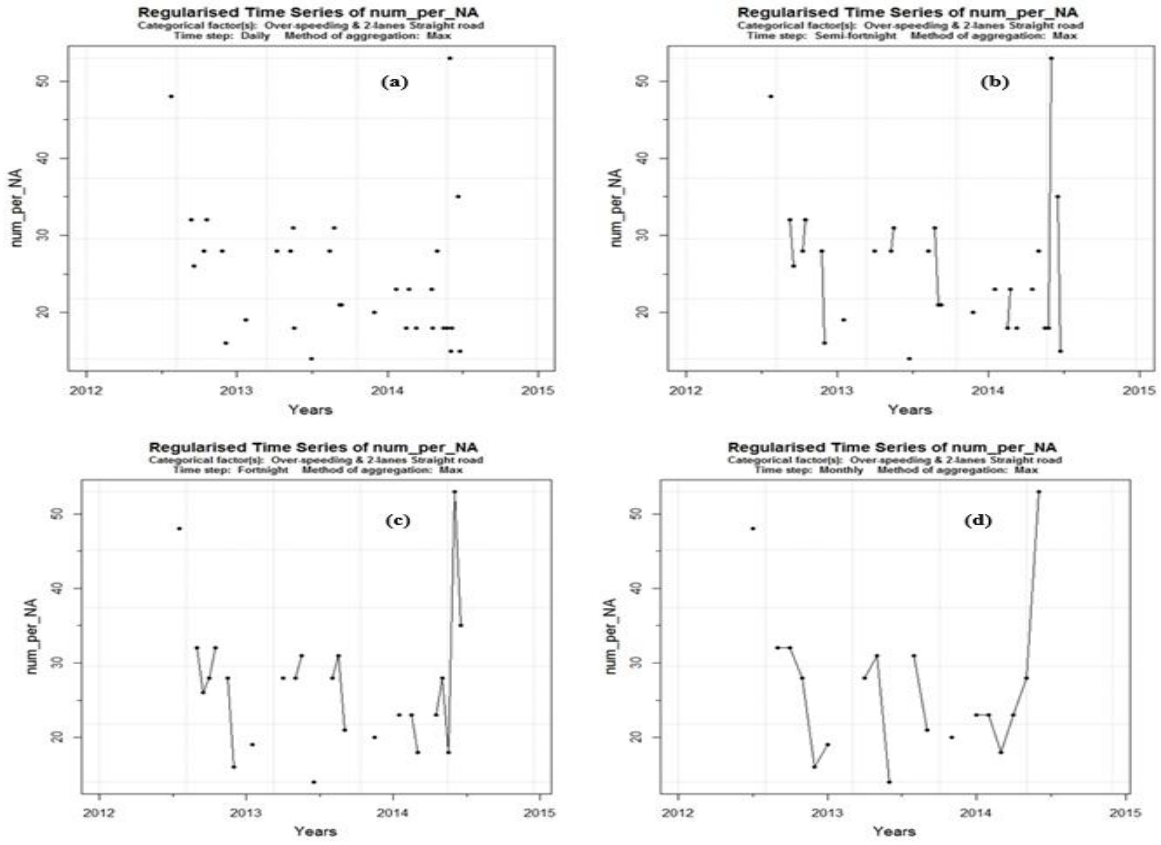

Figure 6 Temporal trend analysis of number of persons affected by non-injury due to over-speeding in two lanes straight road on (a) Daily (b) Semi-fortnightly (c) Fortnightly (d) Monthly basis

Cluster2: Figure 7 shows temporal trend analysis of a number of persons affected by grievous injury over the years on the (a) daily, (b) semi-fortnightly, (c) fortnightly and (d) monthly basis due to over-speeding of LMVs in four-lane straight roads. As shown in figure 7(a), most frequent grievous accidents and maximum persons affected due to grievous injury correspond to the year 2015 on the daily basis, but as per figures 7(b) & 7(c), it is observed that most frequent grievous accidents occur and maximum persons affected due to grievous injury correspond to the year 2016 on the semi-fortnightly and fortnightly basis. However figure 7(d) shows that on the monthly basis, most frequent grievous accident started in 2015 and continues up to 2017 and maximum persons due to grievous injury were affected in 2016. There are distributed a number of persons affected by grievous injury not on regular basis but there are gaps between the number of persons affected by grievous injury over the years on the different time series basis. Most persons are affected by Grievous injury due to over-speeding of LMVs in four lanes straight roads in 2015 and 2016. After 2014, the trend has been increased in 2015 and 2016. By considering past trend in 2015 & 2016, a number of persons affected due to non-injury on the on the daily, semi-fortnightly, fortnightly and monthly basis may be increased in future due to over-speeding of LMVs in four lanes straight roads.

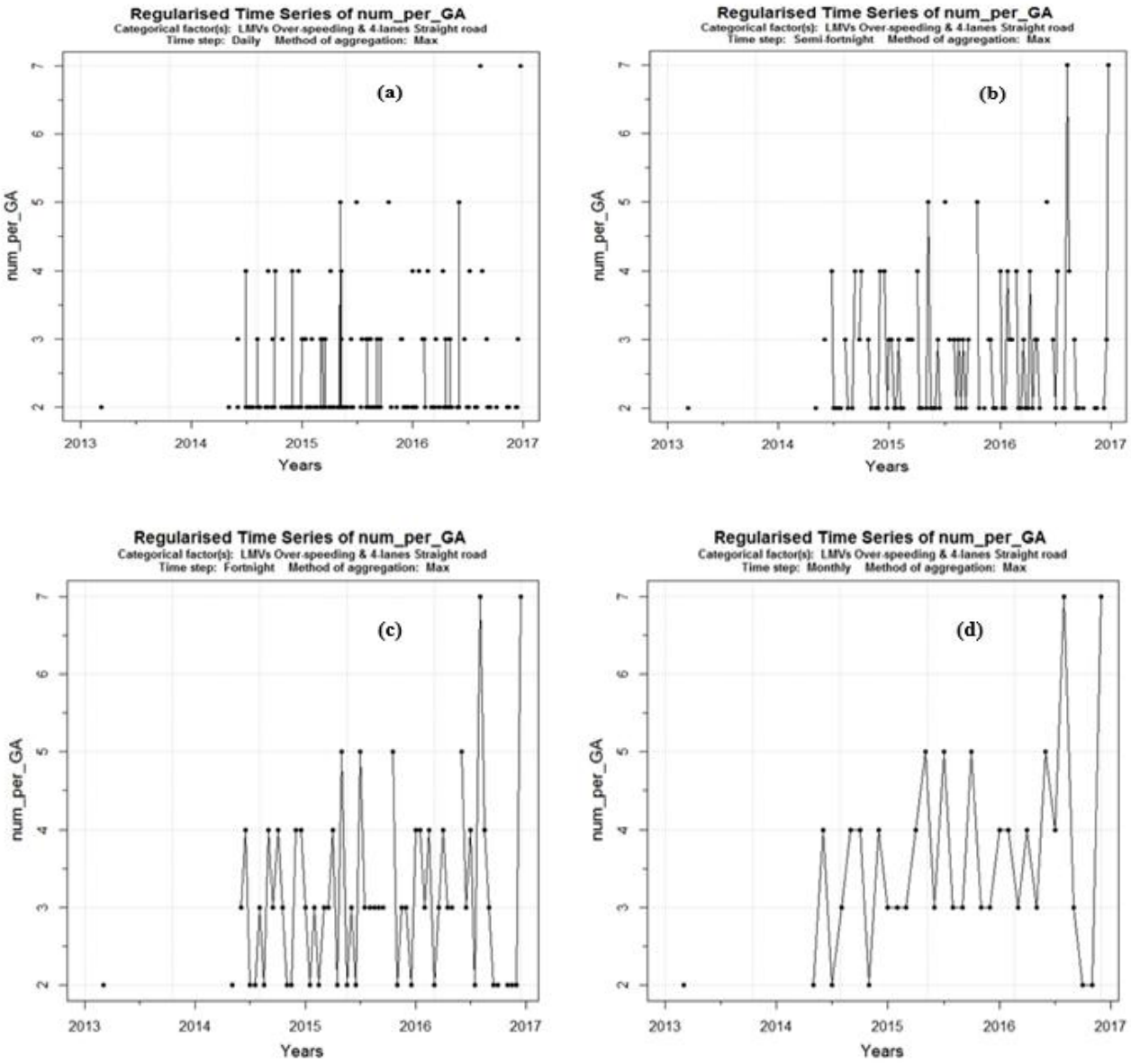

Figure 7 Temporal trend analysis of number of persons affected by grievous injury due to over-speeding of LMVs in four lanes straight roads on (a) Daily (b) Semi-fortnightly (c) Fortnightly (d) Monthly basis

Cluster3: Figure 8 shows temporal trend analysis of a number of persons affected by fatal injury over the years on the on the (a) daily, (b) semi-fortnightly, (c) fortnightly and (d) monthly basis due to over-speeding of LMVs in mostly four lanes straight roads. As shown in Figures 8(a), 8(b), 8(c) & 8(d), most frequent fatal accident was started from March 2014 and continue up to 2016 and maximum persons due to fatal injury were affected in 2016 on the daily, semi-fortnightly, fortnightly and monthly basis.

There are distributed number of persons affected by Fatal injury not on regular basis but there are gaps between the number of persons affected by Fatal injury over the years on the different time-series basis. Most persons are affected by fatal injury due to over-speeding of LMVs in four lanes straight roads in 2015 and 2016. After 2014, the trend has been increased in 2015 and 2016. By considering past trend in 2015 & 2016, a number of persons affected due to fatal injury on the daily, semi-fortnightly, fortnightly and monthly basis may be increased in future due to over-speeding of LMVs in mostly four lanes straight roads.

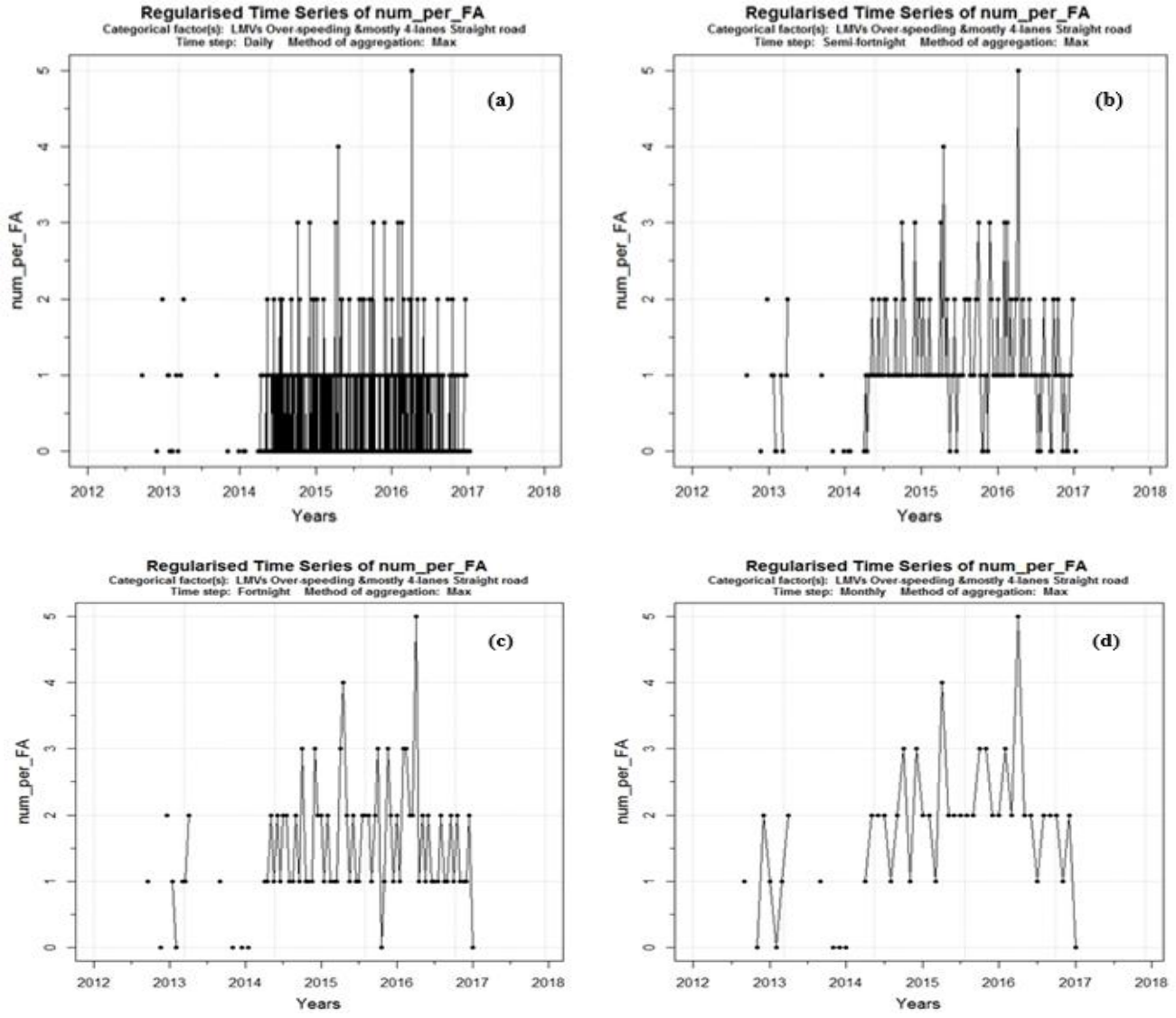

Figure 8 Temporal trend analysis of number of persons affected by fatal injury due to over-speeding of LMVs in mostly four lanes straight roads on (a) Daily (b) Semi-fortnightly (c) Fortnightly (d) Monthly basis

Cluster 4: Figure 9 shows temporal trend analysis of a number of persons affected by minor injury over the years on the (a) daily, (b) semi-fortnightly, (c) fortnightly and (d) monthly basis due to over-speeding of LMVs in mostly four lanes straight roads. As shown in figure 9(a), most frequent minor injury accidents started from May 2014 and continued up to 2016 and maximum persons due to minor injury were affected in 2015 on the daily basis. However on the semi-fortnightly & fortnightly basis, most frequent minor injury accident started from September 2013 and continued up to 2016 and maximum persons due to minor injury were affected in 2015 as shown in Figures 9(b), 9(c). It is observed that on the monthly basis, most frequent minor injury accident starts from March 2014 and continues up to 2016 and maximum persons due to minor injury were affected in 2015 as shown in figure 9(d). There are a distributed number of persons affected by minor injury not on regular basis but there are gaps between the number of persons affected by minor injury over the years on the on the different time series basis. Most persons are affected by fatal accidents due to over-speeding of LMVs in four lanes straight roads in 2015 and 2016. After 2014, the trend has been increased in 2015 and 2016. By considering past trend in 2015 & 2016, a number of persons affected due to a minor injury on the daily, semi-fortnightly, fortnightly and monthly basis may be increased in future due to LMVs over-speeding in mostly four lane straight road.

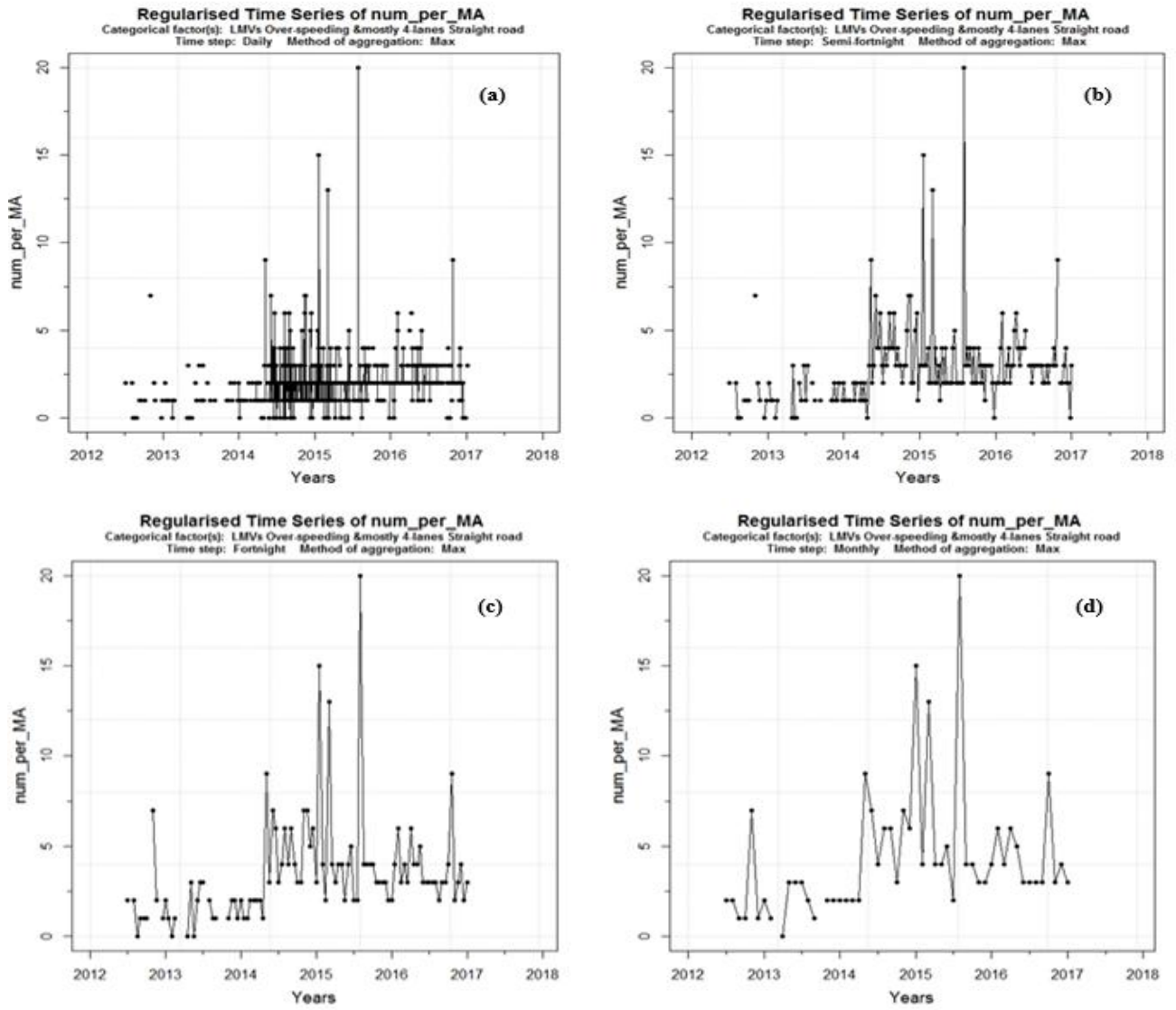

Figure 9 Temporal Trend Analysis of Number of Persons Affected by Minor Injury Due to Over-Speeding of Lmvs in Mostly Four Lanes Straight Roads On (A) Daily (B) Semi-Fortnightly (C) Fortnightly (D) Monthly Basis

V. SUMMARY AND CONCLUSIONS

The best correlation among different factors responsible for accident occurrences has been extracted using data mining techniques from road accident data for highways of Karnataka state implemented by PIU (Bangalore, Chitradurga, Dharwad, Gulbarga, Hospet, and Mangalore) during January 2012 to January 2017 collected through NHAI. Because of a large amount of data having 16 attributes, data mining has been applied for highway road accident data analysis. K-Means algorithm is used to identify clusters i.e. VLFAC (maximum number of persons affected by non-injury), LFAC (maximum number of persons affected by grievous injury), MFAC (maximum number of persons affected by minor injury) and HFAC (maximum number of persons affected by fatal accident) based on accident count using some technical measures. Association rule mining using Apriori algorithm has been applied to identify the relationship among different sets of attributes or various frequently occurring circumstances that are associated with an accident occurrence for the clusters identified by K-means clustering algorithm. The first cluster rule indicates 46.5% multi-vehicle accidents caused due to over-speeding in two-lane straight roads in Kundapur Surathkal highway section in which a maximum number of persons are affected by non-injury. The second cluster rule indicates 40.35% LMVs (passenger, goods, commercial, ambulance, and two-wheeler) accidents are caused due to over-speeding in four lanes straight roads in which a maximum number of persons is affected by grievous injury and involves most of Neelamangala-Devihalli highway section. The third cluster rule indicates 52.7% LMVs (passenger, goods, commercial, ambulance, and two-wheeler) accidents are caused on straight roads due to mostly over-speeding in four-lane roads of Banglore-Hoskote-Mudbagal Highway Section in which a maximum number of persons affected is by a fatal accident. The fourth cluster rule indicates 74.95% LMVs accidents (passenger, goods, commercial, ambulance, and two-wheeler) are caused mostly due to over speeding in four lanes roads for mostly Hyderabad-Bangalore Section in which a maximum number of persons is affected by minor injury.

The rules generated for every cluster revealed the above-considered factors associated with a number of persons affected by road accidents. The analysis using K-means clustering and Apriori association rule mining algorithm of road accident dataset predicts that this approach can prove more effective and provide some more meaningful results with more accident data attributes by correlating various other factors and circumstances associated with highway road accidents. The collected data set is quite sufficient to find out reasonable information but more data can improve the information. Further, a temporal trend analysis has also been performed for each cluster which illustrates different trends in a different cluster and predicted future data sets on the basis of past and present accident data. Finally, temporal trend analysis is applied to study the rules from the preprocessed data on the daily, semi-fortnightly, fortnightly and monthly basis for identifying patterns from the derived data set. It concludes that the number of persons affected by different injuries might increase on the semi-fortnightly, fortnightly and monthly basis but it would decrease on the daily basis.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Human Resource Development [grant number MHR-02- 23-200-429].

REFERENCES

1. MORTH (2014) Road Accidents in India 2013. New Delhi: Ministry of Road Transport and Highways Transport Research Wing, Government of India.
2. Tan PN, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson Addison-Wesley.
3. Ponnaluri RV (2012) Road traffic crashes and risk groups in India: analysis, interpretations, and prevention strategies. IATSS Research 35(2): 104-110. [CrossRef]
4. Parida M, Jain SS, Kumar CN (2012) Road traffic crash prediction on national highways. Indian Highways 40(6).
5. Kumar CN, Parida M, Jain SS (2013) Poisson family regression techniques for prediction of crash counts using Bayesian inference. Procedia-Social and Behavioral Sciences, 104: 982-991. [CrossRef]
6. Bandyopadhyaya R, Mitra S (2013) Modelling Severity Level in Multi-vehicle Collision on Indian Highways. Procedia – Social and Behavioral Sciences, 104: 1011-1019. [CrossRef]
7. Jones B, Janssen L, Mannering F (1991) Analysis of the Frequency and Duration of Freeway Accidents in Seattle. Accident Analysis and Prevention 23(4): 239-255[CrossRef]
8. Miaou SP, Lum H (1993) Modeling Vehicle Accidents and Highway Geometric Design Relationships. Accident Analysis and Prevention 25(6): 689-709. [CrossRef]
9. Miaou SP (1994) The Relationship between Truck Accidents and Geometric Design of Road Sections–Poisson versus Negative Binomial Regressions. Accident Analysis and Prevention 26(4): 471-482. [CrossRef]
10. Poch M, Mannering F (1996) Negative Binomial Analysis of Intersection-Accident Frequencies. Journal of Transportation Engineering, 122(2): 105-113. [CrossRef]
11. Mathuriya N, Bansal A (2012) Comparison of k-means and back propagation data mining algorithms. International Journal of Computer Technology and Electronics Engineering (IJCTEE), ISSN 2249-6343, Volume 2, Issue 2.
12. Han J, Kamber M (2006) Data Mining: Concepts and Techniques. Second Edition. Morgan Kaufmann. ISBN 978-92-4-156437-3, San Fransisco.
13. Piatetsky-Shapiro G (1996) Advances in knowledge discovery and data mining. Fayyad UM, Smyth P, Uthurusamy R, editors. Menlo Park: AAAI Press.
14. Karlaftis MG, Tarko AP (1998) Heterogeneity considerations in accident modeling. Accident Analysis & Prevention. 30(4): 425–33. [CrossRef]
15. Ma J, Kockelman K (2006) Crash frequency and severity modeling using clustered data from Washington state. IEEE Intelligent Transportation Systems Conference. Toronto Canada. ITSC'06. IEEE, pp. 1621-1626.
16. Mohamed MG, Saunier N, Miranda-Moreno LF, Ukkusuri SV (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. [CrossRef]
17. Arora S, Raghavan P, Rao S (1998) Approximation Schemes for Euclidean k-median and Related Problems. Proceedings of the thirtieth annual ACM symposium on Theory of computing. ACM, pp. 106-113. [CrossRef]
18. Valent F, Schiava F, Savonitto C, Gallo T, Brusaferro S, Barbone F (2002) Risk factors for fatal road traffic accidents in Udine, Italy. Accident Analysis & Prevention, 34(1): 71-84. [CrossRef]
19. Anderson TK (2009) Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis & Prevention 41(3): 359-364. [CrossRef]
20. Savolainen PT, Mannering FL, Lord D, Quddus MA (2011) The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis & Prevention. 43(5): 1666–76. [CrossRef]
21. Depaire B, Wets G, Vanhoof K (2008) Traffic accident segmentation by means of latent class clustering. Accident analysis and prevention. 40(4):1257-66. [CrossRef]
22. Kononov J, Janson B (2002) Diagnostic methodology for the detection of safety problems at intersections. Transportation Research Record: Journal of the Transportation Research Board. (1784): 51-56. [CrossRef]
23. Lee C, Saccomanno F, Hellinga B (2002) Analysis of crash precursors on instrumented freeways. Transportation Research Record: Journal of the Transportation Research Board. (1784): 1-8. [CrossRef]
24. Kumar S, Toshniwal D (2016) A data mining approach to characterize road accident locations. Journal of Modern Transportation. 24(1):62-72. [CrossRef]
25. Geurts K, Wets G, Brijs T, Vanhoof K (2003) Profiling of high-frequency accident locations by use of association rules. Transportation Research Record: Journal of the Transportation Research Board. (1840): 123-30. [CrossRef]
26. Chang LY, Chen WC (2005) Data mining of tree-based models to analyze freeway accident frequency. Journal of safety research. 36(4): 365-75. [CrossRef]
27. Bandyopadhyaya R, Mitra S (2013) Modelling Severity Level in Multi-vehicle Collision on Indian Highways. Procedia – Social and Behavioral Sciences. 104: 1011-1019. [CrossRef]
28. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM computing surveys (CSUR) 31(3): 264-323. [CrossRef]
29. Kotsiantis S, Kanellopoulos D (2006) Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering. 32(1): 71-82.
30. Kwon OH, Rhee W, Yoon Y (2015) Application of classification algorithms for analysis of road safety risk factor dependencies. Accident Analysis & Prevention. 75:1–15. doi:10.1016/j.aap.2014.11.005. [CrossRef]
31. Dongre J, Prajapati GL, Tokekar SV (2014) The role of Apriori algorithm for finding the association rules in Data mining. Issues and Challenges in Intelligent Computing Techniques (ICICT). 2014 International Conference on. IEEE. pp. 657-660. [CrossRef]
32. Henderson R, Jaffe AB, Trajtenberg M (1998) Universities as a source of commercial technology: a detailed analysis of university patenting, 1965–1988. Review of Economics and statistics. 80(1): 119-27. [CrossRef]
33. Immerwahr J (2004) Public Attitudes on Higher Education: A Trend Analysis, 1993 to 2003. National Center Report Number 04-2. Public Agenda.
34. Grosjean P, Ibanez F (2004) Package for Analysis of Space-Time Ecological Series. PASTECS version 1.2-0 for R v. 2.0. 0 & version 1.0-1 for S+ 2000 rel: 3.

AUTHOR PROFILE

Kumari Pritee, I have done B.Tech in Computer Science and Engineering from BIT SINDRI, M.Tech in informatics from IIT DHANBAD and completed Ph.D. in "Trend Model for Road Accident Analysis: Spatial Data Mining Approach" from IIT ROORKEE. I have worked on cloud based GIS applications for Roorkee City and spatial data mining approach for highway sections of Karnataka. These applications provide a solution to researchers, policymakers to improve road conditions by making decisions on the basis of time-series spatial analysis. This application plays a very important role in reducing accident severity.

I am **R. D. Garg,** working as a Professor, Geomatics Engg., Civil Engineering Department Indian Institute of Technology (IIT), Roorkee – 247667. I have done B.E in Civil Engineering from University of Roorkee, M.Tech in Remote Sensing from University of Roorkee and completed Ph.D. in GIS and remote sensing from IIT ROORKEE. My Current Areas of Research are Geomatics Remote Sensing, DIP, GIS, WebGIS, Land Surveying, GPS, Thermal, Microwave and Hyperspectral remote sensing, Ground Penetrating Radar (GPR).

Retrieval Number:100.1/ijdm.C1618051322
DOI:10.54105/ijdm.C1618.051322
Journal Website: www.ijdm.latticescipub.com

Published By:
Lattice Science Publication (LSP)