Contents lists available at GrowingScience

Journal of Future Sustainability

homepage: www.GrowingScience.com/jfs

Criticality trend analysis based on highway accident factors using improved data mining algorithms

Kumari Pritee[a*] and R. D. Garg[b]

[a]*Assistant Professor, Computer Science and Engg, Parul University, India*
[b]*Professor, Geomatics Engineering, IIT Roorkee, India*

C H R O N I C L E

Article history:
Received: October 2, 2022
Received in revised format: October 18, 2022
Accepted: November 13, 2022
Available online:
November 13, 2022

Keywords:
Data mining
Fp growth
Highway section
Accidents
Trend model

A B S T R A C T

Highway accident data analysis provides probability of occurrence of road accidents by associating different accident factors using data mining algorithms. Analysis can be improved by using advanced data mining algorithms that compute relationships with minimum processing time. As accident datasets are very heterogeneous in nature, it is difficult to identify the relationship between critical factors responsible for road accidents without data mining algorithms. In this study, K-modes for clustering and frequent pattern growth algorithms to extract relationships between critical accident factors have been used. The accomplished result concludes better relationships with better accuracy than earlier implemented data mining algorithms and has found meaningful hidden situations that would be beneficial for future work in decreasing the number of highway accidents.

1. Introduction

With the advanced technology in transportation, road accidents have decreased day by day but increasing vehicle speed makes highways dangerous for traveler health as well as economic conditions. Hence, comprehensive study of road accidents is an important task for safety analysts to identify the accident causative factors. On the basis of severity consequences, safety analysts can decide preventive actions to reduce the accident rate. Nowadays, the road accident analysis can be finally done by identifying the most critical factors affecting road accident frequency with their severity. The major problem with the dataset of road accidents is its heterogeneous nature. However, it is necessary to understand the pattern of road accidents clustered in critical road segments so that it can find critical factors responsible for black zones by frequently generated itemsets. In this case, frequently- occurred accident circumstances can be generated by data mining techniques only. Different statistical models with a number of hypotheses have been broadly used for accident analysis to extract relationships between crash factors and geometrical/environmental factors (Lee et al., 2002). However, in statistical analysis, main effects should be analyzed with interaction of road accident causative factors i.e., driver, vehicle, roadway and environmental factors should be analyzed to become more significant (Chen & Jovanis, 2000). Because of increasing complexity in large amounts of crash data with injury outcome, classical statistical tests are not enough for large dimensional analysis of crash datasets.

Data mining proposes various applications in the field of transportation (Barai, 2003). Road accident safety is one of the crucial areas of transportation in which nearby geographical features are actively involved. Significant research in the field of road accident safety has already been done using traditional statistical approaches (Joshua & Garber, 1990; Abdel- Aty & Radwan, 2000; Chen & Jovanis, 2000). Moreover, traditional statistical techniques have their own limitations on the basis of dependence of data attributes. Any mistake in the selection of these data attributes can make disastrous outcomes. Different previous studies with statistical approach have been done (Jones et al., 1991; Jones & Sheather 1991; Maher and Summersgill, 1996; Poch & Mannering, 1996; Karlaftis & Tarko, 1998; Savolainen et al., 2010) and data mining techniques (Chang & Chen, 2005; Kashani et al., 2011; Kumar & Toshniwal, 2015a, 2016b, 2016c; Prayag et al., 2017) have been carried out to estimate strong association between accident attributes and road accident criticality.

* Corresponding author.
E-mail address: priteegeo.kumari23@gmail.com (K. Pritee)

ISSN 2816-8151 (Online) - ISSN 2816-8143 (Print)

doi: 10.5267/j.jfs.2022.11.002

The objective of this study is divided into three sub-objectives:

The First sub-objective of this study is to preprocess the number of accident dataset of the period during January 2012 to January 2017 of National Highway sections of Karnataka state. The road project of the said sections of highways has been implemented by Project Implementation Unit i.e., PIU (Bangalore, Chitradurga, Dharwad, Gulbarga, Hospet and Mangalore). The accident dataset for the mentioned period collected from NHAI has been divided into homogeneous clusters using K-modes clustering. The second sub-objective is to reflect the relationship between the above-mentioned factors using the FP (Frequent Pattern)- growth association rule. The last sub-objective is to perform temporal trend analysis for each cluster on the basis of rules generated through Association Rule Mining and to predict the trend pattern for future road accidents.

2. Traditional Statistical approach for accident analysis

Statistical approach is totally different from data mining technique. This traditional statistical approach has been utilized for preliminary analysis for some time. Statistical approaches are carried out by the attribute information for generating predictive models. It also has an effective role in road accident safety research and analysis. The impact of rider's age on accident circumstances or patterns has been studied using negative binomial models and clustering techniques (Karlaftis & Tarko, 1998). The investigation has been done by clustering the data and afterward sorted that dataset of the accident into individual categories. Additionally, clustered output of the investigated dataset has been utilized to identify the reason of accident by Negative Binomial (NB) centering age of driver which may exhibit a few outcomes (Karlaftis & Tarko, 1998). A seven-year road accident dataset having 63 intersections in Bellevue (Washington) has been utilized using a negative binomial regression for the repetition of crashes at passage point. The assessment comes regarding estimation of crucial intersections along with traffic and geometric associated factors (Poch & Mannering, 1996). In that case, association of all these critical factors is still lacking. In this study, association of critical factors for critical highway sections has been shown using temporal trend analysis. This is where the data mining approach comes to remove the heterogeneous nature of accidents. Data mining can be termed as the nontrivial process to categorize meaningful patterns from a huge dataset (Fayyad et al., 1996).

3. Data Mining approaches for Accident Analysis

Data mining algorithms include classification, clustering, anomaly detection and association rule mining. Clustering algorithms such as k-means, k-modes clustering and hierarchical clustering are quite accepted algorithms in numerous domains. Heterogeneity of dataset is a main concern of data mining. A framework has been proposed for removal of heterogeneous nature of road accident dataset and concludes that clustering is more prior before analysis to deal with heterogeneous nature of traffic and accident dataset (Kumar & Toshniwal, 2015a). Latent class clustering (LCC) has also been used to remove heterogeneity from different types of data. LCC is suggested as an effective clustering technique for identifying optimal number of clusters using different cluster selection criteria. Moreover, a proportional research on road accident dataset from Haridwar (Uttarakhand) India has been performed (Kumar & Toshniwal, 2017). LCC and K-modes clustering techniques have been compared to cluster the data before performing the analysis. Additionally, they mined association rules using the FP-growth algorithm to mine the rules that explained accident prototypes in every cluster. They accomplished both techniques and found related effectiveness on cluster patterns and are proficient to eliminate the heterogeneous nature of the dataset. On the other hand, their conclusions were not appropriate to disclose the supremacy of one technique over another.

Heterogeneous dataset of road accidents is extremely unwanted and inescapable (Karlaftis & Tarko, 1998). The major shortcoming of heterogeneity of road accident data is that certain associations may possibly stay concealed such that accident factors related with exact vehicle type may not be considerable in complete data set. It is recommended that previous segmentation is extremely valuable in generating high-quality consequences for road accident data analysis. Earlier, data has been grouped into homogenous subgroups according to some skilled information, methodologies (Ulfarsson & Mannering, 2004; Islam & Mannering, 2006). However, division of factors into feasible groups can be done without assuring the homogeneity nature of the accident dataset. Therefore, data mining comes into play in which cluster analysis can be done to eradicate the heterogeneous nature of road accident data. Clustering technique has been enhanced in the various studies using K-modes to minimize the heterogeneity of road accident data (Depaire et al., 2008; Sasidharan et al., 2013). In this research, the authors claimed that K-modes clustering algorithms are extremely functional compared to other algorithms to minimize the heterogeneous nature of road accident data. The concluded outcomes can be used for providing precautionary rules to reduce road accident data.

In this study, critical highway sections (also termed as black highway sections) with their accident characteristics have been estimated using a data mining approach. An exploratory technique for frequent itemset has been used particularly to estimate promising accident patterns. Specifically, frequently occurred accident circumstances have been identified for critical highway sections. K-modes and hierarchical clustering have been applied to group accident data then classification approaches are used on clustered data for enhancing accuracy (Tiwari et al., 2007). K-mode clustering and association rules have been employed to mark the varied factors that are associated with the road accident frequency (Kumar et al., 2017).

This study represents a data mining approach using NHAI datasets (explained in previous chapter) to explore road accident factors by implementing K-modes clustering algorithm. After that, Frequent Pattern (FP) growth association rule mining has been applied on the clusters identified by K-modes to find out the relationships between various road accidents causative factors and critical highway sections. Frequent Pattern (FP) growth association rule mining of different clusters using K-modes clustering provides the hidden information by performing segmentation and association rules with minimum execution time. After that temporal trend analysis has been implemented using extracted rules on different time steps.

3.1 Data collection

Various meaningful circumstances have been revealed from the outcomes of these previous studies. The different factors used in the present study are: - Highway sections (HS), road features (ROF), road conditions (ROC), nature of accident (NOA), cause of accident (CAU) and vehicle responsible (VR). The analysis of such different factors is required to be done for a daily, fortnightly, semi-fortnightly and monthly basis for proper monitoring of road accidents. Therefore, the accuracy of such accident data analysis provides accurate critical information for formulating policies for prevention of road accidents.

4. Methodology

4.1 Data Preprocessing

After performing the three mentioned steps of preprocessing, the 6963 road accidents have been considered for further analysis.

4.2 Clustering

K-modes clustering

Data mining facilitates problem identification, data preprocessing and clustering and association pattern evaluation. Clustering is an unsupervised learning or segmentation in which groups are not predefined but are accomplished by determining the similarity among the data and similar items. Clustering algorithm has been applied to discover meaningful classes for unknown class labels in the form of clusters. The preliminary task for any cluster analysis is to compute the optimal number of clusters. Different approaches have already been applied to find out optimal number of clusters i.e., gap statistics or various information criteria e.g., AIC, BIC and CAIC (Raftery, 1986; Akaike, 1987; Fraley & Rafetery, 1998).

Methods for optimal number of clusters

- Elbow method
- Gap statistic method

Elbow method

This method can be defined on the basis of the minimum total within the cluster sum of squares. The total within-cluster sum of squares (WSS) measures the compactness of the clustering.

Algorithm

The optimal number of clusters can be defined as follow:

- Choose best clustering algorithm i.e. K-modes clustering for different values of k. For Example, k can be varying from 1 to 10 clusters.
- Calculation of the total within cluster sum of square (WSS) for each k.
- Design the curve of WSS on the basis of number of clusters k.
- The optimal number of clusters is considered on the basis of bend location in the plot.

Elbow method has been used for finding the optimal number of clusters by looking at the percentage of variance as a function of the number of clusters. The number of clusters is chosen by an angle in the graph via marginal gain drop. According to this method, 4 clusters have been suggested as shown in Fig. 1.

4.3 The gap statistic Method

It is a cluster classification technique that can be applied with any clustering algorithm (Tibshirani et al., 2001). Steps are as follows,

Step 1: The observed data has been clustered, varying the total number of clusters k = 1, 2, ..., K, giving within-dispersion

measures $W_{k,}$ k = 1, 2,. . . . , K.

Step 2: B reference data sets has been generated and clustered each one giving within-dispersion measures W^*_{kb}, b = 1, 2,. . . . , B, k = 1, 2,. . . . , K.

Compute the (estimated) gap statistics

$$\mathrm{Gap(k)} = (\frac{1}{\mathrm{B}}) \sum_{\mathrm{b}} \log\,(\,\mathrm{W}^*_{\mathrm{kb}}) - \log\,(\mathrm{W_k})$$

Step3: let $\mathrm{l} = \bar{l} = (\frac{1}{B}) \sum_b \log\,(\,\mathrm{W}^*_{\mathrm{kb}})$, compute the standard deviation

$$\mathrm{sd_k} = [(\frac{1}{\mathrm{B}}) \sum_{\mathrm{b}} \{\log(\mathrm{W}^*_{\mathrm{kb}}) - \bar{\mathrm{l}}\}^2\,]^{1/2}$$

And define $\mathrm{s_k} = \mathrm{sd_k}\sqrt{(1 + 1/\mathrm{B})}$. Finally choose the number of clusters via

$$\hat{\mathrm{k}} = \text{smallest k such that } \mathrm{Gap(k)} \geq \mathrm{Gap(k+1)} - \mathrm{s_{k+1}}$$

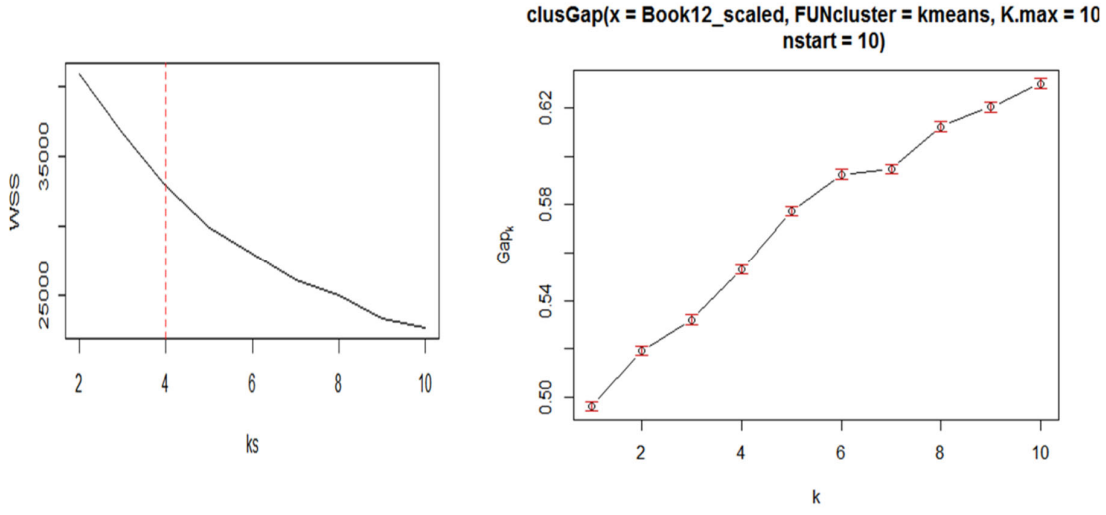

Fig. 1. Sum of WSS over Number of Clusters (a) Elbow Method (b) Gap Statistic method

According to previous studies (Depaire et al., 2008), Gap Statistic method is considered preferable over Elbow method to determine the optimal number of clusters. On the basis of this method, four clusters have been considered as an optimal number of clusters. Cluster selection results using the Elbow plot and Gap statistic method have been shown in the figure. The Figure demonstrated the Knee curve at model with cluster 4. The gap value at cluster model 4 has been indicated in the Figure that maximizes the value of $\mathrm{Gap_n}$ K. According to the above results of cluster selection criteria, four numbers of clusters have been chosen as an optimal number of clusters.

The K-means algorithm is a quite well-known clustering algorithm for huge numerical data analysis as shown in my previous studies. However, in this study, K-Modes clustering have been used to cluster the categorical data set. This approach has been proposed as an improved edition of conventional k-means algorithm along with many enhancements i.e. Iteration process, cluster center representation and distance measure (Chaturvedi et al., 2001). The K-modes algorithm used similarity measure criterion for clustering of categorical dataset.

Suppose X categorical attributes have been categorized using two qualitative data objects A and B. The similarity matching criterion between A and B has been computed using a quantity of similar attribute value of data objects. For more similarity among two objects, the number of matches must be more. In comparison to the K-means algorithm, K-modes algorithm employed modes irrespective of means for clustering the dataset. According to previous study, K-modes algorithm act as comparatively proficient in managing huge amounts of categorical dataset.

In this, the dataset is clustered into k clusters. There exists varied clustering algorithms but the variety of suitable clustering algorithms depends upon type and nature of data. The main purpose of this study is to discriminate the accident location based on their occurrence of frequency. Let's assume that X and Y is a matrix of m by n matrix of categorical data. The basic proximity coordinating computation between X and Y is the number of coordinating feature evaluation of the two values. The more significant number of matches is the comparability of two items. K-modes algorithm can be explained as:

$$d\,(Xi,Yi) = \Sigma\ \delta(Xi, Yi)\ mi = 1 \quad (1)$$

where $(Xi, Yi) = \{1, if\ Xi = Yi$
$0, if\ Xi \neq Yi$

In this study, K-modes clustering have been used to cluster the dataset into optimal number of clusters which is shown in results and discussion section.

4.4 Association algorithm

Association rules generate all sets of items having 'support' greater than the 'minimum support' and thereafter, generate the desired rules that have 'confidence' greater than the 'minimum confidence' using the large itemsets. This rule mining has already been discussed in the previous chapter. In this study, the FP-growth association algorithm has been used to obtain a descriptive analysis of critical highway sections. The Apriori algorithm has been already used in the previous chapter but the problem with these algorithms is that it uses candidate item set generation to test whether the item sets are frequent or not. Therefore, the Apriori algorithm acts as computationally expensive because it scans the database multiple times for candidate-sets generation. FP growth association rule mining technique has been proposed by Han et al., (2000). The major advantage of FP growth algorithm over Apriori is that its computation is faster than Apriori because of no requirement of candidate generation. FP growth algorithm works on a special data structure termed as FP tree, which conserves the relationship information of itemset. The support, confidence and lift interesting measures have been used to mine strong association rules from the data set. After construction of FP tree, the flowchart of FP growth algorithm is given as follows,

Algorithm: FP-Growth (Kumar et al., 2016)

Step1: A data set D, designated by FP-tree created and threshold assessment for support (Input); The group of frequent patterns(Output)
Step2: call FP-Growth(FP-tree, [ ]) // [ ] represents NULL Procedure FP-Growth(Tree, a)
Step3: If Tree consists a single prefix path then // Mining single prefix-path FP-tree
Step4: Let P be the single prefix-path part of Tree; Let Q be the multipath part with the top branching node replaced by a null root.
Step5: For each combination (denoted as A) of the nodes in the path P, generate pattern AU a with support = minimum support of nodes in A;
Step6: Let frequent-pattern-set(P) be the set of patterns so generated; otherwise Q be Tree;
Step7: For each item X_i in Q do // Mining multipath FP-tree, generate pattern A = X_iUX with support = X_i .support;
Step8: Construct A's conditional pattern-base and then A's conditional FP-tree Tree A; If Tree A ≠ Ø then
Step9: Call FP-Growth(Tree A,A);
Step10: Let frequent-pattern-set(Q) be the set of patterns so generated;
Step11: Return (frequent-pattern-set(P) ∪ frequent-pattern-set(Q) ∪ (frequent-pattern-set(P) × frequent-pattern-set (Q)))

4.5 Analysis for clustered based association rules using minimum confidence count

Cluster1: Rules for accidents in Hyderabad-Bangalore and Devihalli-Hassan highway sections

Table 1 shows the relationship between different accident factors i.e. CAU, ROC, ROF, NOA and HS. According to evaluated relationships in chosen factors, Hyderabad-Bangalore and Devihalli-Hassan are the most critical highway sections. The accident causative factors for critical highway sections are the main aspect for accident studies. According to above rules, the strongest accident causative relationship for Hyderabad-Bangalore section is due to overspeeding with right turn collision in four lane straight roads and for Devihalli-Hassan section due overspeeding with head on collision in four lane flat roads.

Table 1
Association rules for cluster 1 considering two factors

Rules	Measures
[ROF=FL_binarized=1, CAU=OVSD_binarized=1, HS=Hyderabad-Bangalore Section_binarized=1]: 492 ==> [ROC=SR_binarized=1, NOA=RTC_binarized=1]: 492	<conf:(1)> lift:(2.89) lev:(0.09) conv:(322)
[ROF=FL_binarized=1, CAU=OVSD_binarized=1, ROC=SR_binarized=1, HS=Hyderabad-Bangalore Section_binarized=1]: 492 ==> [NOA=RTC_binarized=1]: 492	<conf:(1)> lift:(2.5) lev:(0.08) conv:(294.96)
[ROF=FL_binarized=1, CAU=OVSD_binarized=1, NOA=RTC_binarized=1, HS=Hyderabad-Bangalore Section_binarized=1]: 492 ==> [ROC=SR_binarized=1]: 492	<conf:(1)> lift:(1.35) lev:(0.03) conv:(128.56)
[ROF=FL_binarized=1, NOA=HOC_binarized=1, HS=Devihalli-Hassan_binarized=1, ROC=FR_binarized=1]: 388 ==> [CAU=OVSD_binarized=1]: 370	<conf:(0.95)> lift:(1.22) lev:(0.02) conv:(4.47)
[NOA=HOC_binarized=1, HS=Devihalli-Hassan_binarized=1, ROC=FR_binarized=1]: 389 ==> [ROF=FL_binarized=1, CAU=OVSD_binarized=1]: 370	<conf:(0.95)> lift:(1.5) lev:(0.03) conv:(7.1)
[ROF=FL_binarized=1, CAU=OVSD_binarized=1, NOA=HOC_binarized=1, ROC=FR_binarized=1]: 390 ==> [HS=Devihalli-Hassan_binarized=1]: 370	<conf:(0.95)> lift:(6.56) lev:(0.08) conv:(189)
[CAU=OVSD_binarized=1, NOA=HOC_binarized=1, ROC=FR_binarized=1]: 391 ==> [ROF=FL_binarized=1, HS=Devihalli-Hassan_binarized=1]: 370	<conf:(0.95)> lift:(6.59) lev:(0.08) conv:(122)

Cluster2: Rules for accidents in Tumkur-Chitradurga highway section

Table 2 shows the relationship between different accident factors i.e. CAU, ROC, ROF, VR and HS According to evaluated relationships in choosen factors, Tumkur-Chitradurga is most critical highway section. According to above rules, the strongest accident causative relationship for Tumkur-Chitradurga section are due to heavy motor vehicle's overspeeding in four lane straight roads and overspeeding with right end collision in four lane straight roads.

Table 2
Association rules for cluster 2 considering two factors

Rules	Measures
[ROC=SR_binarized=1, CAU=OVSD_binarized=1, VR=HMV_binarized=1, HS=Tumkur-Chitradurga_binarized=1]: 161 ==> [ROF=FL_binarized=1]: 161	<conf:(1)> lift:(1.38) lev:(0.03) conv:(44.11)
[ROC=SR_binarized=1, CAU=OVSD_binarized=1, NOA=REC_binarized=1, HS=Tumkur-Chitradurga_binarized=1]: 150 ==> [ROF=FL_binarized=1]: 150	<conf:(1)> lift:(1.38) lev:(0.03) conv:(41.1)

Cluster3: Rules for accidents in Silk board to electronic city junction highway section

Table 3 shows the relationship between different accident factors i.e. CAU, ROC, ROF, DOA and HS. According to evaluated relationships in chosen factors, Silk board to electronic city junction is the most critical highway section. According to above rules, the strongest accident causative relationships for Silk board to electronic city junction are due to slightly inclined three lane roads and overspeeding in three lane roads at night.

Table 3
Association rules for cluster 3 considering two factors

Rules	Measures
[ROF=THL_binarized=1, ROC=SIN_binarized=1]: 116 ==> [HS=Silk board to electronic city junction_binarized=1]: 116	<conf:(1)> lift:(1.18) lev:(0.02) conv:(18.1)
[ROF=THL_binarized=1, DOA=Night_binarized=1, CAU=OVSD_binarized=1]: 134 ==> [HS=Silk board to electronic city junction_binarized=1]: 132	<conf:(0.99)> lift:(1.17) lev:(0.02) conv:(6.97)

Cluster4: Rules for accidents in Banglore – Neelamangala highway section

Table 4 shows the relationship between different accident factors i.e. ROC, ROF, VR and HS. According to evaluated relationships in chosen factors, Bangalore - Neelamangala are the most critical highway sections. According to above rules, the most strongest accident causative relationships for Bangalore - Nelamangala section are due to three lane straight roads and two-wheeler in straight road.

Table 4
Association rules for cluster 4 considering two factors

Rules	Measures
[ROF=THL_binarized=1, HS=Bangalore - Neelamangala_binarized=1]: 102 ==> [ROC=SR_binarized=1]: 102	<conf:(1)> lift:(1.24) lev:(0.02) conv:(19.98)
[VR=TW_binarized=1, HS=Bangalore - Neelamangala_binarized=1]: 112 ==> [ROC=SR_binarized=1]: 111	<conf:(0.99)> lift:(1.23) lev:(0.02) conv:(10.97)

5. Results and Discussion

The following diagrams show temporal trend analysis of the maximum persons affected by different types of accident over the years on the (a) daily, (b) semi-fortnightly, (c) fortnightly and (d) monthly basis due to different causative circumstances. These diagrams show the trend changing on different time steps.

Cluster1 (Hyderabad-Bangalore Section, RTC, OVSD, FL, SR): Most frequent accidents occurring due to these critical factors have increased from May 2015 as shown in Figure 2. In 2015, persons affected due to fatal accidents have increased in the months of April to June. Similarly, in 2016, persons affected due to fatal accidents have increased in the months of January and March. Mostly the trend has been continuously increasing compared to past trends on a semi-fortnight basis and varying in the first six months of year mainly in the month of March. It may slightly increase in future trends as compared to past trends as shown in Fig. 3.

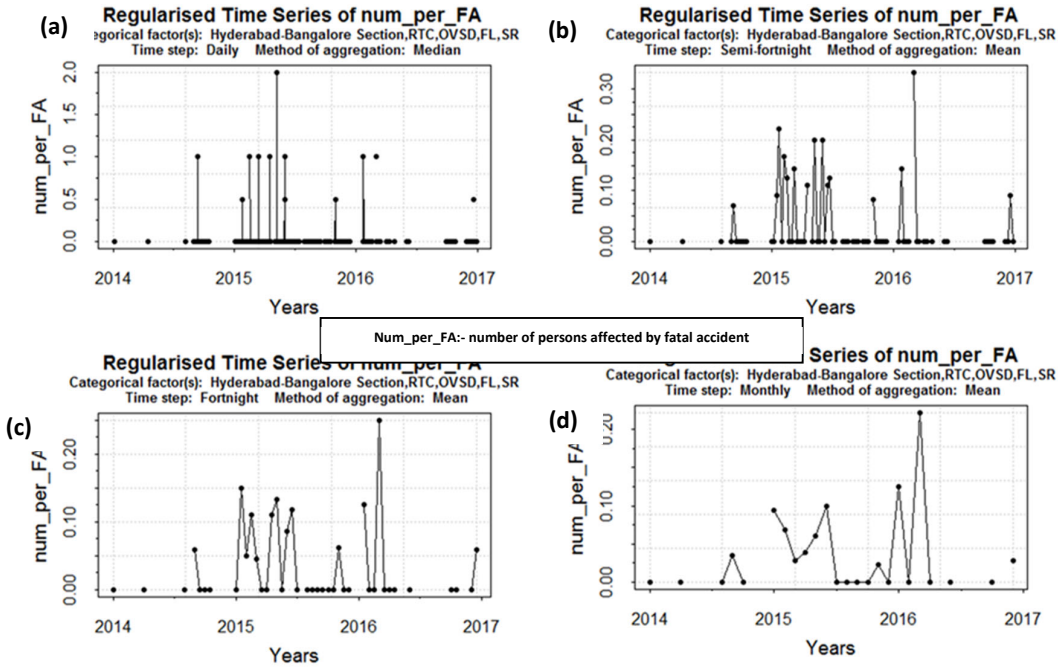

Fig. 2. Temporal trend analysis of number of persons affected by fatal accident due to critical categorical factor

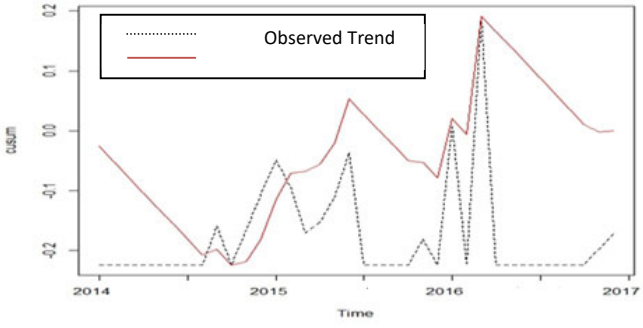

Fig. 3. Predicted monthly trend analysis of number of persons affected by fatal injury due to critical categorical factor

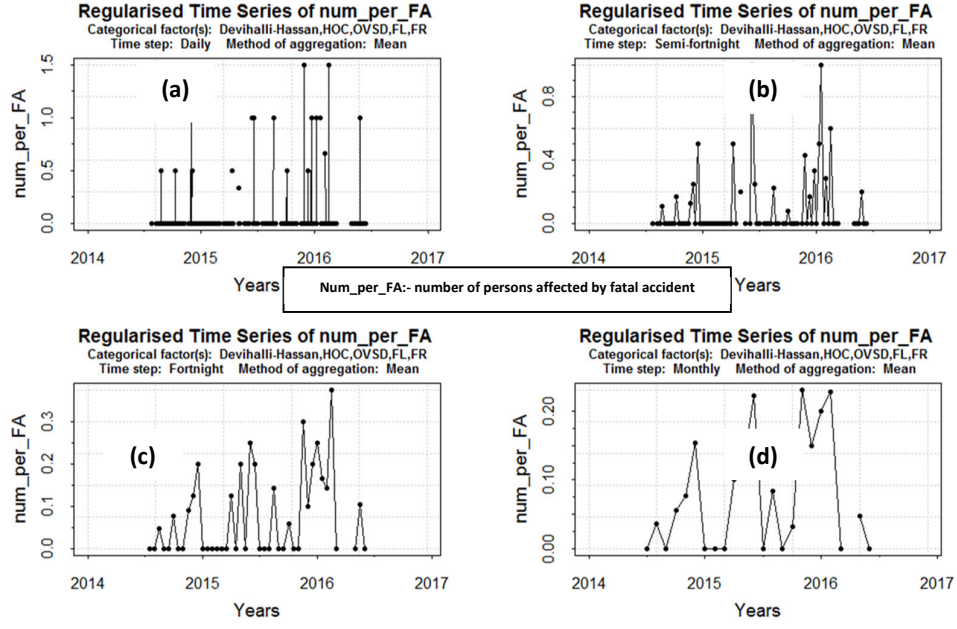

Fig. 4. Temporal trend analysis of number of persons affected by fatal accident due to critical categorical factor

Cluster1 (Devihalli-Hasan, HOC, OVSD, FL, FR): Most frequent accidents occurred due to previously explained critical factors have increased from July 2014 as shown in Fig. 4. In 2014, persons affected due to fatal accidents have increased in the months of September to December. Similarly, in 2015 and 2016, persons affected due to fatal accidents have increased in the months of March to June & October to November and in the month of February. Mostly the trend has been highly increasing compared to past trends and varying throughout the year mainly in the months of November and December. It may increase in future trends as compared to past trends as shown in the figure.

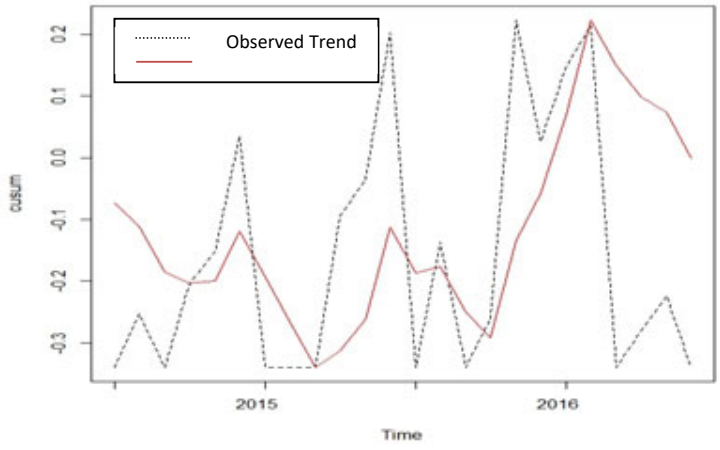

Fig. 5. Predicted monthly trend analysis of number of persons affected by fatal injury due to critical categorical factor

Cluster2 (Tumkur-Chitradurga, OVSD, FL, SR, HMV): Most frequent accidents occurring due to these critical factors have increased from July of 2014 as shown in Figure 6. In 2014, persons affected due to fatal accidents increased in the months of September to December. Similarly, in 2015 and 2016, persons affected due to fatal accidents have increased in the months of March to June & October to November and in the month of February. Mostly the trend has been highly increasing compared to past trends and varying throughout the year and mostly in the months of November and December.

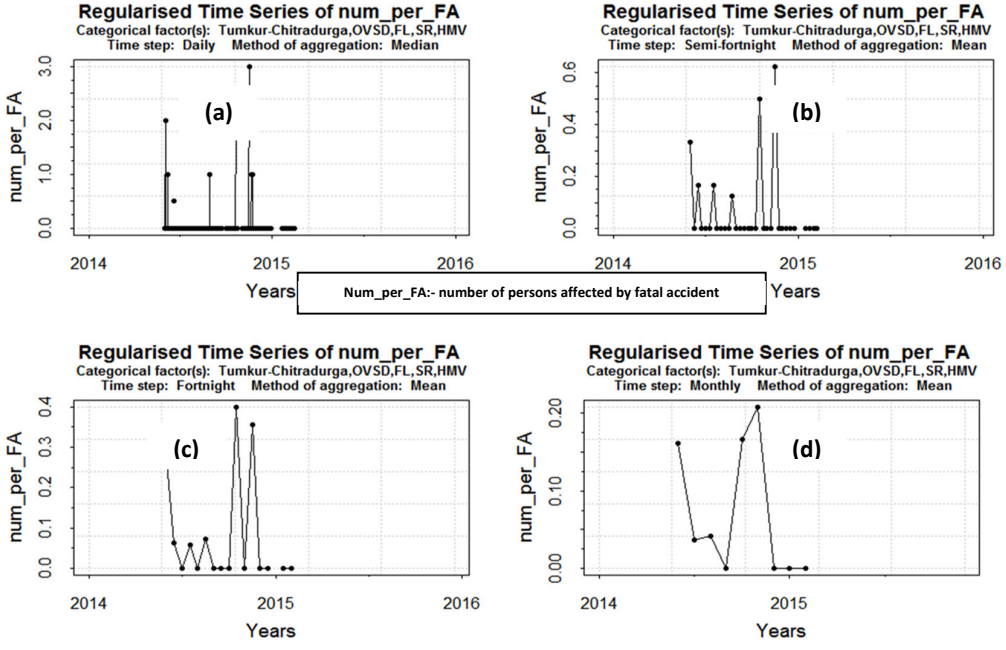

Fig. 6. Temporal trend analysis of number of persons affected by fatal accident due to critical categorical factor

Cluster2 (Tumkur-Chitradurga, OVSD, FL, SR, REC): Most frequent accidents occurring due to these critical factors have increased from July of 2014 as shown in Fig. 7. It varies the same as the above critical factors.

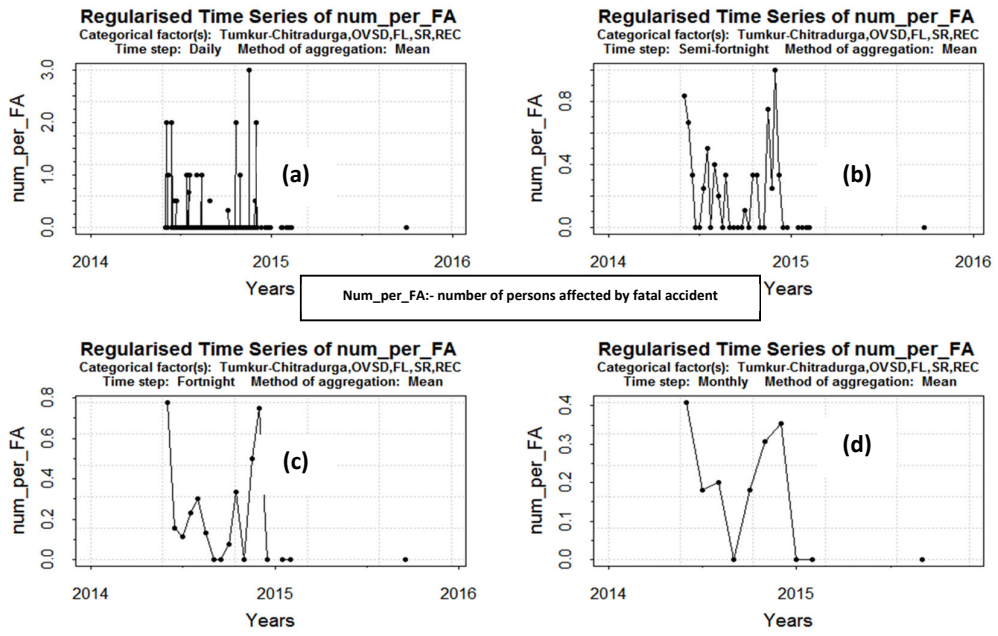

Fig. 7. Temporal trend analysis of number of persons affected by fatal accident due to critical categorical factor

Cluster3 (Silk Board to Electronic city junction, THL, SIN): Most frequent accidents occurred due to previously explained critical factor have increased from May 2015 as shown in Fig. 8. In 2015, persons affected due to fatal accidents have increased in the month of April to June. Similarly, in 2016, persons affected due to fatal accidents have increased in the month of March & August to September. Mostly the trend has been continuously increasing compared to past trends and varying in last six months of year mainly in month of March and September. It may slightly decrease in future trend as compared to past trend as shown in Fig. 9.

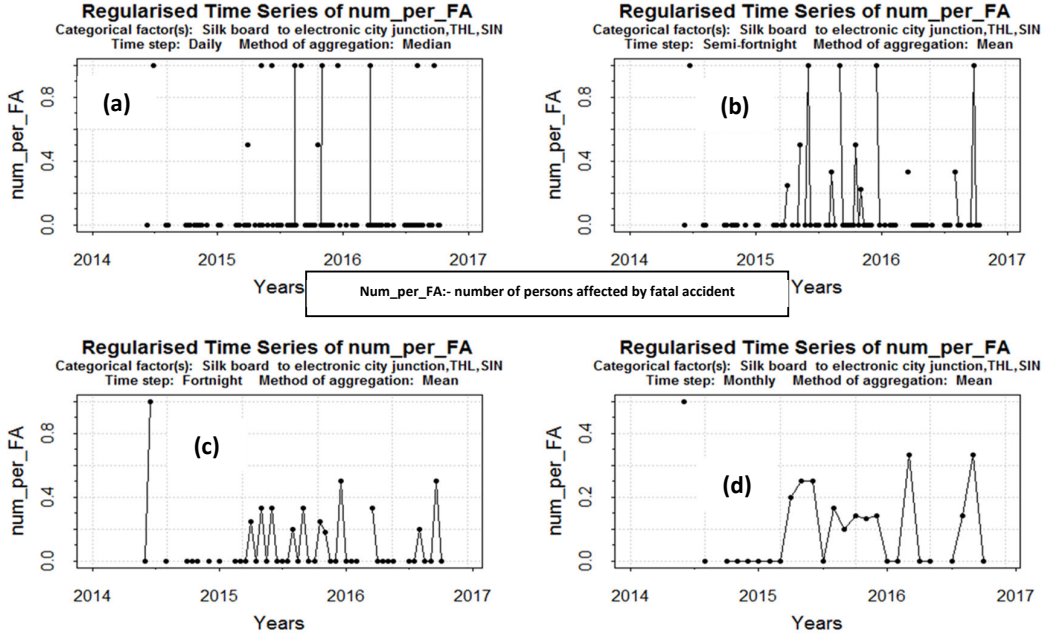

Fig. 8. Temporal trend analysis of number of persons affected by fatal accident due to critical categorical factor

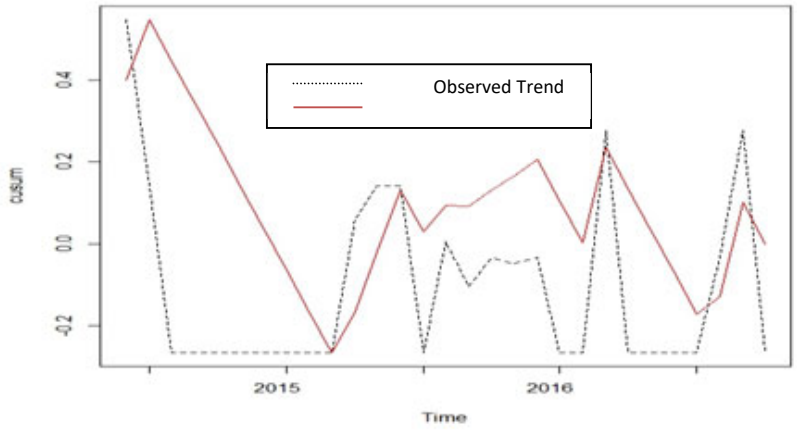

Fig. 9. Predicted monthly trend analysis of number of persons affected by fatal injury due to critical categorical factor

Cluster3 (Silk Board to Electronic city junction, OVSD, THL, NIGHT):- Most frequent accidents occured due to previously explained critical factors have increased from June 2014 as shown in Figure 10. In 2015, persons affected due to fatal accidents have increased in the month of July to September. Similarly, in 2016, persons affected due to fatal accidents have increased in the month of March. Mostly the trend has been highly increasing compared to past trends mainly in month of March and September. It may increase in future trend as compared to past trend as shown in Figure 11.

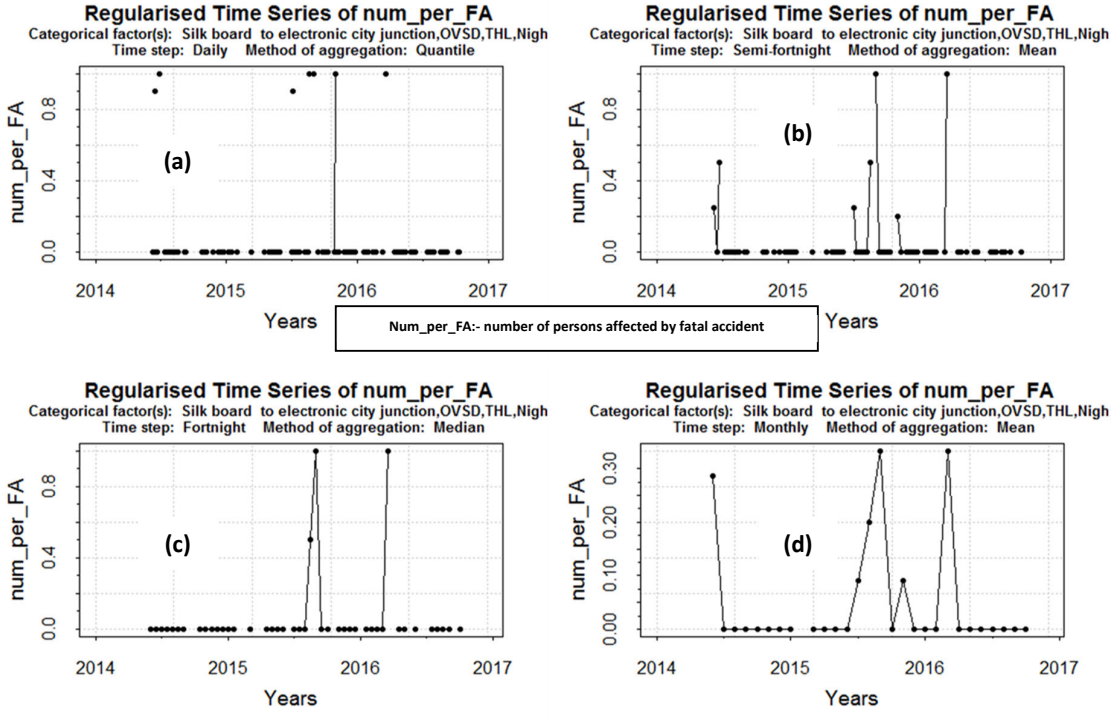

Fig. 10. Temporal trend analysis of number of persons affected by fatal accident due to critical categorical factor

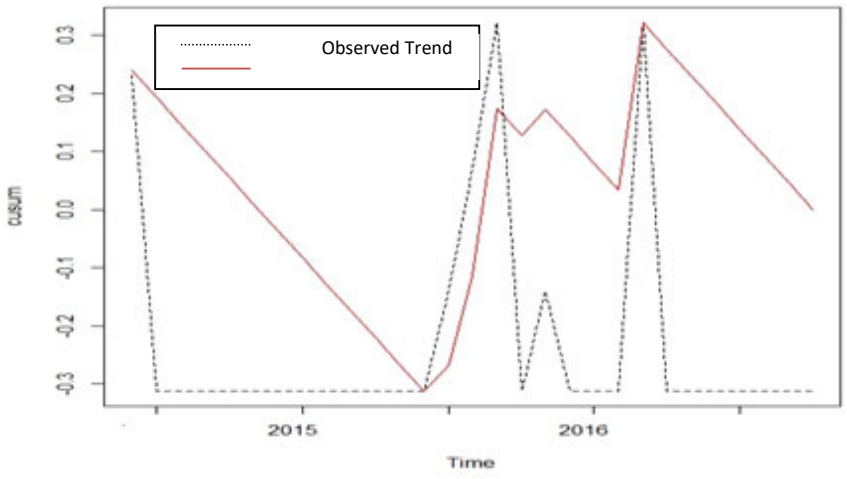

Fig. 11. Predicted monthly trend analysis of number of persons affected by fatal injury due to critical categorical factor

Cluster 4 (Bangalore-Neelmangla, THL, SR): Most frequent accidents occurred due to previously explained critical factor have increased from June 2014 as shown in Figure 12. In 2015, persons affected due to fatal accidents have increased in the month of July to September. Similarly, in 2016, persons affected due to fatal accidents have increased in the month of March. Mostly the trend has been highly increasing compared to past trends mainly in month of March and September. It may increase in future trend as compared to past trend.

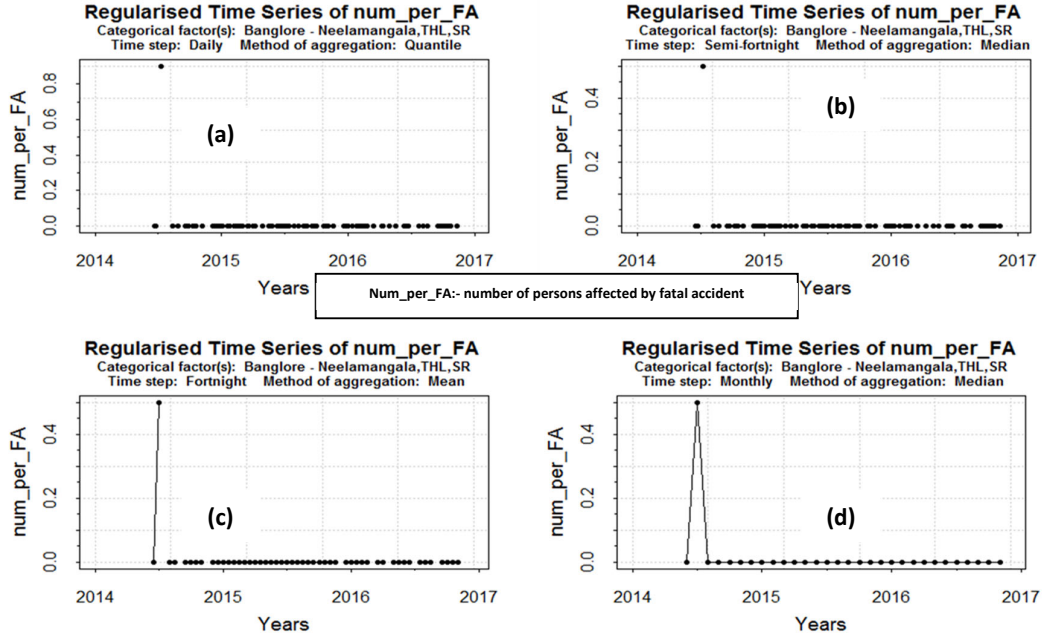

Fig. 12. Temporal trend analysis of number of persons affected by fatal accident due to critical categorical factor

6. Summary and Conclusions

Using advanced data mining techniques, the best correlation among the different factors responsible for accident occurrences have been extracted. This work introduces temporal trend analysis for different K-modes clusters with their strong association rules using FP-growth algorithm and proves effective for making policies for different types of accident causative factors on different time steps. Further, a temporal trend analysis has also been performed for each cluster. The analysis illustrates different trends in different clusters and predicts the future data sets on the basis of past and present accident data.

This method of analysis proves more effective in extracting precise substantial information which could be utilized in designing and executing a safe road transport system. The preciseness of the derived information depends upon the quality and quantity of the data being studied. In this study, though the collected data set is quite sufficient to extract a meaningful

result, an improvised data can lead towards a better result.

After implementing K-modes clustering and frequent pattern association rules, temporal trend analysis of a number of persons affected by fatal accident over the years on the (a) daily, (b) semi-fortnightly, (c) fortnightly and (d) monthly basis due to critical factors obtained by strong association rules has been performed. After analyzing all trends, maximum persons affected by fatal accidents have been taken into account and visualized on the basis of monthly trends. Most frequent critical accidents and maximum persons affected due to fatal & grievous accidents are due to below explained critical categorical factor as shown in Fig. 13.

***Critical term*: - Maximum persons affected due to fatal accident**

Critical Categorical factor:

- Hyderabad Bangalore Section, RTC, OVSD,FL, SR
- Devihalli-Hasan, HOC, OVSD, FL, FR
- Silk Board to Electronic city junction, THL, SIN
- Silk Board to Electronic city junction, OVSD, THL, NIGHT

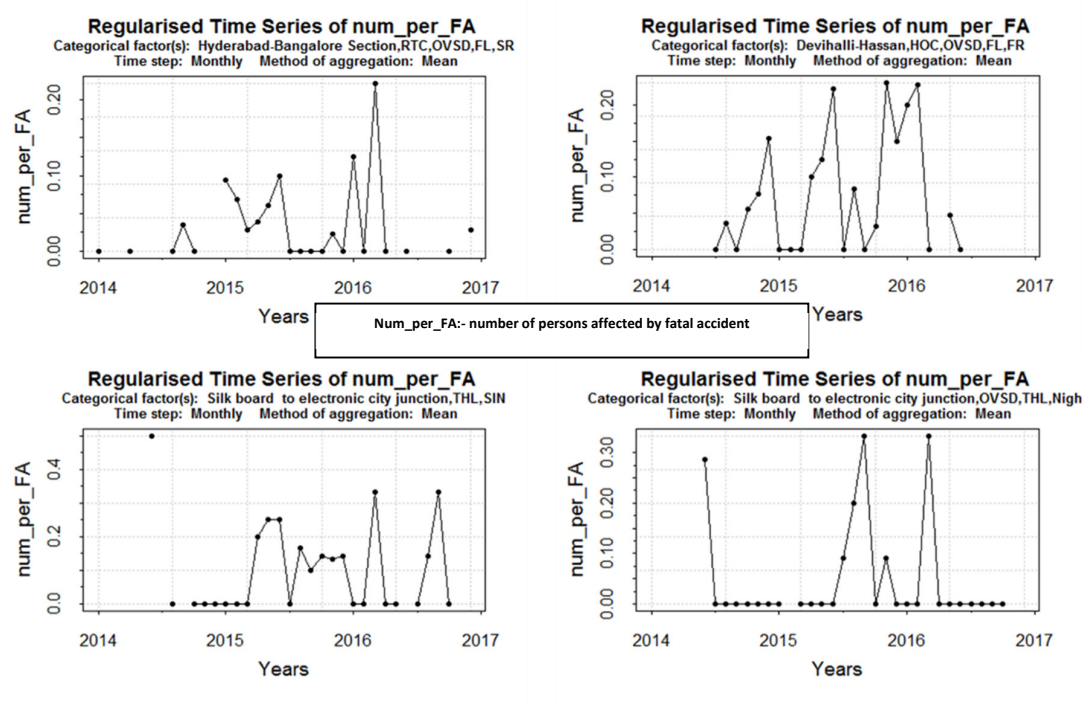

Fig. 13. Temporal monthly trend analysis of number of persons affected by fatal accident due to critical categorical factors

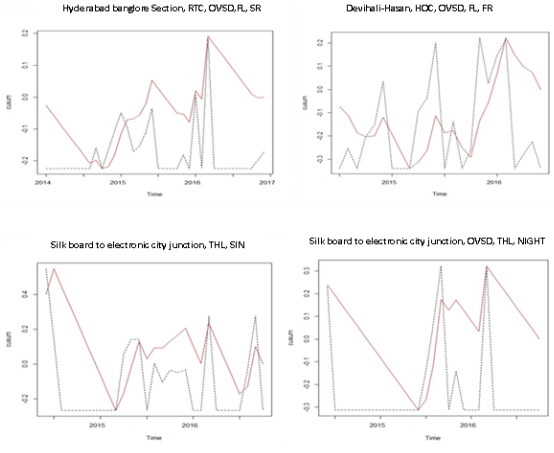

Fig. 14. Predicted monthly trend analysis of number of persons affected by fatal accident due to critical categorical factors

The trend analysis of persons affected due to fatal accidents has been shown in Fig. 14. It varies according to different critical factors as compared to past trends. It may increase or decrease in future trends compared to past trends depending upon different critical factors as explained above.

References

Abdel-Aty, M. A. & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention, 32*(5), 633-642.

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

Barai, S. K. (2003). Data mining applications in transportation engineering. *Transport, 18*(5), 216-223.

Chang LY, Chen WC (2005) Data mining of tree-based models to analyze freeway accident frequency. *Journal of safety research. 36*(4), 365-75.

Chaturvedi, A., Green, P. E., & Caroll, J. D. (2001). K-modes clustering. *Journal of classification, 18*(1), 35-55.

Chen, W. H., & Jovanis, P. P. (2000). Method for identifying factors contributing to driver-injury severity in traffic crashes. *Transportation Research Record, 1717*(1), 1-9.

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

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 1-37.

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

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *In ACM sigmod record 29*(2), 1-12. ACM.

Islam, S., & Mannering, F. (2006). Driver aging and its effect on male and female single-vehicle accident injuries: some additional evidence. *Journal of safety Research, 37*(3), 267-276.

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

Jones, M. C., & Sheather, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics & Probability Letters, 11*(6), 511-514.

Joshua, S. C., & Garber, N. J. (1990). Estimating truck accident rate and involvements using linear and Poisson regression models. *Transportation planning and Technology, 15*(1), 41-58.

Karlaftis M, Tarko A (1998) Heterogeneity considerations in accident modeling. *Accident Analysis Preview, 30*, 425–433

Kashani, A. T., & Mohaymany, A. S. (2011). Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. *Safety Science, 49*(10), 1314-1320.

Kumar, S., & Toshniwal, D. (2015). A data mining framework to analyze road accident data. *Journal of Big Data, 2*(1), 1-26.

Kumar, S., & Toshniwal, D. (2016). A data mining approach to characterize road accident locations. *Journal of Modern Transportation, 24*(1), 62-72.

Kumar, S., & Toshniwal, D. (2016). A novel framework to analyze road accident time series data. *Journal of Big Data, 3*(1), 1-8.

Kumar, S., & Toshniwal, D. (2017). Severity analysis of powered two wheeler traffic accidents in Uttarakhand, India. *European Transport Research Review, 9*(2), 1-24.

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.

Maher, M. J., & Summersgill, I. (1996). A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis & Prevention, 28*(3), 281-296.

Poch, M., & Mannering, F. (1996). Negative Binomial Analysis of Intersection-Accident Frequencies. *Journal of Transportation Engineering, 122*(2), 105-113.

Prayag, G., Hosany, S., Muskat, B., & Del Chiappa, G. (2017). Understanding the relationships between tourists' emotional experiences, perceived overall image, satisfaction, and intention to recommend. *Journal of Travel Research, 56*(1), 41-54.

Raftery, A. E. (1986). Choosing models for cross-classifications. *American sociological review, 51*(1), 145-146.

Sasidharan, R., Mustroph, A., Boonman, A., Akman, M., Ammerlaan, A. M., Breit, T. & van Tienderen, P. H. (2013). Root transcript profiling of two Rorippa species reveals gene clusters associated with extreme submergence tolerance. *Plant physiology, 163*(3), 1277-1292.

Savolainen, K., Alenius, H., Norppa, H., Pylkkänen, L., Tuomi, T., & Kasper, G. (2010). Risk assessment of engineered nanomaterials and nanotechnologies—a review. *Toxicology, 269*(2-3), 92-104.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411-423.

Tiwari, P., Madabhushi, A., & Rosen, M. (2007). A hierarchical unsupervised spectral clustering scheme for detection of prostate cancer from magnetic resonance spectroscopy (MRS). In International Conference on Medical Image Computing and Computer-Assisted Intervention, 278-286. Springer, Berlin, Heidelberg.

Ulfarsson, G. F., & Mannering, F. L. (2004). Differences in male and female injury severities in sport-utility vehicle,

minivan, pickup and passenger car accidents. *Accident Analysis & Prevention, 36*(2), 135-147.