Asian Journal of Information Technology 15 (12): 2005-2015, 2016
ISSN: 1682-3915

A Novel Approach to Analyze a Combination of IxJ Categorical Data for Estimating Road Accident Risk

[1]Geetha Ramani and [2]Shanthi Selvaraj
[1]Department of Information Science and Technology, Anna University, Guindy Campus, Chennai, Tamil Nadu, India
[2]Department of Computer Science and Engineering, Rathinam Technical Campus, Coimbatore, Tamil Nadu, India

Abstract: Road accident analysis is very challenging task and investigating the dependencies between the attributes become complex because of many environmental and road related factors. Use of feature selection algorithms such as feature ranking, Fisher's test, CFS etc which are using chi square test to find the statistical significance level of features is prevalent in data mining applications but it has been noted that such arbitrary usage may not be appropriate always. While analyzing the available data, sparse information may also be an issue wherein the study tends to use an option to probe the associations more rationally using Random Effect Model (REM). This approach could be much useful in understanding the rules of statistical significance that may be shadowed in other methods, especially chi square analysis. Results have shown the possible association together with amount of heterogeneity between the important variables involved in a data set related to road accidents that have occurred in Coimbatore, Tamil Nadu, India. Further, comprehensive analytical algorithms that are implemented in R have been provided to facilitate the approach for future replications.

Key words: Chi-square, classification, fatality, association, road accidents, random effects model

INTRODUCTION

Accidents are events wherein the number of occurrences is rare, although there are many opportunities for it to happen. Worldwide, >1.2 million people die annually in highway-related crashes and as many as 50 million more are injured and by 2030, highway-related crashes are projected to be the 5th leading cause of death in the world. National crime records bureau, India reported that the total number of deaths every year due to road accidents has now passed the 135,000 mark. About 60,000 lives are lost every year in road accidents. According to the latest Indian Road Transport Ministry report, among the states, Tamil Nadu stands second in the list with the highest number of road accidents. Road traffic accident is under the influence of many factors which makes it a complicated and uncertain system.

With an exponential growth of population, number of vehicles and the need for their use, understanding the multiple causes of road accident fatalities has become more significant especially in the advent of sophisticated technology (Singh and Suman, 2012). In recent years, with the growth of the volume and travel speed of road traffic, the number of traffic accidents, especially severe crashes has been increasing rapidly on a yearly basis. This is one of the primary factors responsible for road accidents. Identification of these factors can help improve the overall driving safety situation, not only by preventing accidents but also by reducing their severity.

The another concern is the way accidents are recorded; many reports and studies have expressed that the real numbers of fatalities could be much higher since many cases are not even reported. Out of the estimated 1.4 million serious road accidents/collisions occurring annually in India, hardly 0.4 million are recorded. Further, only a minimal percentage of these collisions are scientifically investigated, in the absence of which the real causes and consequences are never known. (http://www.irte.com/crash-investigation.html accessed on 10-10-2013). In most of the practical situations, enumerated data are collected for two or more variables and such data are tabulated for frequencies of observations of each of the categories, called contingency table. Table 1 shows the case of two variables with general notations followed in the analysis of such data.

Corresponding Author: Shanthi Selvaraj, Department of Computer Science and Engineering, Rathinam Technical Campus, Coimbatore, Tamil Nadu, India

Table 1: Notations for representing the levels of two categorical variables in I×J form

	Variable 2			
Variable 1	Level 1	Level 2	…	Row totals
Level 1	n_{11}	n_{12}	…	$R_1 = \sum n_{1j}$
Level 2	n_{21}	n_{22}	…	$R_2 = \sum n_{2j}$
……				……
Column totals	$C_1 = n_{i1}$	$C_1 = n_{i2}$	…	$N = \sum c_i = \sum R_i = \sum n_{ij}$

Researchers are often interested in finding an appropriate relationship between the categories of the classified variables. In statistical or data mining paradigm, such problems are attributed to finding associations between variables. Reliable information in terms of collected data and appropriate analysis tools are combined to establish these associations.

It can be observed that Chi-square test has been widely used in various feature selection algorithms of data mining. Many theoretical and application studies such as Mirkin (Mirkin, 2001) have attempted to portray the importance and elegance of Chi-square test of association. Many studies provide a wide comparison of association rules that include chi square method and it could be a prudent way to discover the correlation between the attributes (Fong *et al*., 2007; Dorn *et al*., 2008; Matsuoka *et al*., 2008). However, this widely accepted test for association has witnessed a concern for its applicability for categorical data which might lack few features that would deter in following chi square approximate procedures. Especially, small expected frequencies have drawn a special attention while applying Chi-square tests in data mining practices. Based on the above tabular display, expected frequency for a cell $E(I, J) = R_i \times C_j / N$ and large N will result with smaller E. All the cell counts may be equally spread or some of the cells contain low and/or zero counts whereas other cells have higher counts. Such polarized form, low or zero counts have received active attention in the data analysis environments. In such situations, chi square tests will not be appropriate and many studies have addressed alternatives such as Fisher exact test for 2×2 contingency tables (Agresti, 2002; Campbell, 2007; Howell, 2011; Kraus, 2012). Further, most of the recommendations have indicated that when association rules need to be compared between data sets of different sizes, the chi square test may not be preferred. More specifically, the chi-square test should only be used when all cells in the contingency table have expected values <1 and at least 80% of the cells have expected values <5(Brij *et al*., 2003). The researcher also listed some issues pertaining to the characteristic of a contingency table while focusing on objective measures of interestingness.

In view of this, the present study has identified the scope to analyze the categorical data using summary measure-based Random Effects Model (REM). In particular, REMs have been used to facilitate collapsing the given two-dimensional I×J data based on domain-based rationale and to understand the association in a more critical manner. Supporting data has been obtained from road accident information collected periodically by the Coimbatore City Traffic Head Quarters for the year 2013.

Application of random effects model has been widely appreciated in many experimental studies. Issues related to the choice of effect measure, handling small counts and/or zero counts are the major consideration of many research works. Literature is abundant in dealing with inference on summary measures especially for rare events (Sweeting *et al*., 2004) have compared the performance of different meta analysis methods for pooling odds ratios when applied to sparse event data with emphasis on the use of continuity correction. While analyzing the rare events (Bradburn *et al*., 2007) have evaluated the performance of 12 methods, eight for pooling odds ratio and four for pooling risk difference and the sparseness has been considered together with the usual recommendations on zero cells. However, these two extensive comparisons are mainly based on frequentist methods of estimation (Subbiah *et al*., 2008) have considered the methods that include the conditional approach and hierarchical Bayesian model on log-odds ratio for a study on sparse data. Similar other applications can be observed in many studies (Bollen and Brand, 2010; Chena and Chenb, 2010; Gebregziabher *et al*., 2012; Jackson *et al*., 2010).

In respect to accident related data analyses, most of the studies have applied random effects model not only with binomial distribution but also with Poisson and negative binomial distributions (Chin and Quddus, 2003) have examined the temporal and spatial effects in occurrence of the road traffic accidents at signalized intersections of Singapore using random effect negative binomial model. The results have showed the significant features on the safety at the intersection; yet the study has insisted improvements on analytical methods used both fixed and random-effects linear regression models to investigate the relationship between crash frequencies and roadway design. Also an ordered probit regression model has been used to examine the effects of speed limits as well as various geometric design features on crash severity. The speed limit appears to have no significant effect on crash severity.

Huang *et al*. (2010) examined Zero-Inflated Poisson regression with site-specific random effects (REZIP) with comparison to random effect poisson model and standard zero-inflated Poisson model using crash data in Singapore. Ulfarsson *et al*. (2002) used the negative multinomial model to form a predictive model of median crossover accident frequencies using a multi-year panel of cross-sectional roadway data of Washington. They found that the negative multinomial significantly

outperforms the negative binomial and the random-effects negative binomial in terms of fit but the negative multinomial model suffered from convergence problems.

Usman *et al.* (2011) illustrated the effects of data aggregation and correlation on disaggregated accident prediction models. The results showed that the effect of data aggregation had a significant effect on model results. A meta analysis has been conducted to enumerate the effects of cell phones on driving performance (Caird *et al.*, 2008). From the results they could derive that during driving the Reaction Time (RT) increases while using cell phones. The correlations between driver and vehicle involved in the same crashes in Singapore roads have been studied using bayesian hierarchical binomial logistic model (Huang *et al.*, 2008). The model could identify the significant factors affecting the severity level of driver injury and vehicle damage in traffic crashes at signalized intersections. Similar models could be found in many studies (Boucher and Guillen, 2009; Qi *et al.*, 2007; Quddus, 2008, 2013).

Cummings *et al.* (2006) and references thereon provide more details about case-control studies in helmet and head injury analysis. Wong and Chung (2008) have indicated the importance of heterogeneity in accident analysis and road safety models. A meta-analysis has been conducted to study the effect of road safety campaigns in reducing the number of road accidents and the model was evaluated by fixed and random effect meta-regression (Phillips *et al.*, 2011). Also, odds ratio has been the widely applicable summary measure when two variables of binary in nature have to be compared and most of the Meta analyses and case-control studies have exploited the advantages of odds ratio. Extensive literature is available that incorporate odds ratio as one of the summary measures to interpret the results and findings in distinct applications (Chen *et al.*, 2013; Lim *et al.*, 2010; Lv *et al.*, 2013; Park *et al.*, 2013; Petitti *et al.*, 2013) in particular recent studies (Donroe *et al.*, 2008; Elvik, 2011; Kuypers *et al.*, 2012) and references thereon provide applications of odds ratio specific to accident related data analyses. Further, (Eijkemans *et al.*, 2012; Fidler and Nagelkerke, 2013; Mavros *et al.*, 2013; Olivier and Bell, 2013) can be referred for detailed review about the usage of odds ratio in systematic combined studies.

However, it could be observed that the literature is quite abundant; still many methods depend on the available data structure. This study has identified similar situation in terms of limited number of attributes and sparse categorical data in I×2 contingency tables. Also the feature selection algorithms give only whether an attribute is significant or not but they do not mention the direction of significance.

Shanthi and Ramani (2012a, b) have analyzed accident related data sets using various data mining algorithms to understand the causes for road accidents; this includes the predicting the vehicle collision, pedestrian accidents, injury severity, accident patterns based on seating position. Further, these attempts are based on large data sets which are amenable for the application of data mining tools such as C4.5, C-RT, CS-MC4, Decision List, ID3, Naïve Bayes and Random Tree Feature Selection, CFS, FCBF etc.

Many studies have shown that driver-related attributes such as attitude, behaviour, knowledge and hazard perception are important determinants of the likelihood for collision and this was assessed using a Poisson regression model (Darby *et al.*, 2009). Literature on road accident related studies (Cummins *et al.*, 2011; Milburn *et al.*, 2006; Roudsari *et al.*, 2004) have indicated few frequently used factors that influence the severity of an accident. These include various environmental, behavioural and demographic variables and this paper has followed similar rationality in reshaping the original data.

However, all available data sets may not possess such computational flexibility, some of them require a more specific way to analyze to draw meaningful inferences. Similar to such situations where special data collection is not an option, analysis can be done by combining routine accident data from different sources or using any reasonable stratification methods. It could be observed that economic and time-related factors deter the collection of specific data on road accident units; yet to derive knowledge an available data set can be prepared in a consistent way with meaningful analysis plan. This in turn will be suitable to exploit the standard statistical procedures through sophisticated computing facilities to obtain better inferences and recommendations. Also, it has been suggested to consider the cause and effect of road accidents in Indian roads not only at a macro-level but also at a micro-level (Valli, 2005).

Hence, an REM approach has been attempted by collapsing the given I×J contingency table into multiple 2×2 tables based on reckoning appropriate stratifying variables that can be drawn from the studies. Further, in the advent of modern computing facilities, this study provides a comprehensive methodology to understand the nature of given data in terms of zero/low counts; recommends association tests and provides an alternative to traditional feature selection algorithms by applying REM based on proper collapsing conditions. It has been observed that this approach provides a better scope to understand the association between the variables when feature selection algorithms using Chi square analyses fail to identify the strength as well as the direction of underlying overall associations.

MATERIALS AND METHODS

Dataset description: The Coimbatore city traffic data records have details of accidents that have occurred in the city and the suburbs. This includes information on the number of fatal and non-fatal accidents each month based on classifications such as gender, time of accident, place of accident, type of roadways etc. Such a repository may be used to understand the number and characteristics of the units at risk and depending on the analysis, the corresponding data can be obtained. For this specific study, recent data sets that have been collected during January 2013-November 2013 are used and are available at www.shanthiselvarajresearch.info.

Micro-level categorization is done on the given data so as to form 2×2 contingency tables. As the public/private drivers tend to have different road behavior, the sample data are grouped based on vehicle types-public or private. Gender and road type have also been considered important variables to investigate; time slots of the accident occurrence are classified into two-day and night and peak hours and non-peak hours. Thus, the given samples are classified based on time: day/night, peak hours/non peak hours; type of vehicles: public/private; gender: male/female; road type: highways/non highways; month: vacation/non vacation. Table 2 provides the way in which the data in I×J tabular form has been collapsed to stratified 2×2 tables of the form as follows.

Based on these observations, the present study has reorganized the original data set so that REM can be applied to analyze the data more intrinsically. Table 3 provides the complete list of stratified variable 1 and 2. The two levels of variable II are the propensity of fatal or otherwise in the given road accident profile except for S. No 10 of Table 3 which has type of road (highways or otherwise) as two levels of variable 2. Further stratification is based on the eleven months during which the data has been collected.

Each of the 2×2 table cross-classifies the counts that fall within the respective levels of variables I and II that are labelled as quadruple (a, b, c, d). Hence, original data has a three dimensional table to depict month, categorizing variables and count of fatal and non-fatal related to the road accidents.

Table 4 exemplifies a typical data set and Table 5 presents this approach for the data from Table 4. This procedure of collapsing I×J table into 2×2 tables, subsequent analysis of tests for association and REM has been implemented in R. The source codes are available with researcher.

Table 2: Notation of stratified 2×2 table

	Variable 2	
Variable 1	Level 1	Level 2
Level 1	A	B
Level 2	C	D

Random effect model: REM is a statistical method to combine the results of individual studies. By statistically combining the results of similar studies it is possible to improve the precision of the estimates of study effect and assess whether study effects are similar enough to be combined. There are many approaches similar to pooling the counts in all individual studies as if they were part of one big study. However, this might be resulted with well known Simpson's paradox. In a similar way, it can be shown that simple average of study effects may not be a proper method to summarize the results as it is not uncommon for an analysis sources of variability, within-study and between study must be taken into account when making inferences about the population. Precisely, the population contrast is modelled by a two level process:

$$d_{ij} = \theta_i + e_{ij} \quad (1)$$

$$\theta_i = \mu + z_i \quad (2)$$

Table 3: List of variables that are used to categorize the given I×J tables in to 2×2 tables stratified by the months of original data collected

	Variable 1		Variable 2	
Study variables	Level 1	Level 2	Level 1	Level 2
Gender	Male	Female	Fatal	Non fatal
Hit run	Hit/run	Not H/R	Fatal	Non fatal
Age	Age <18	Age >18	Fatal	Non fatal
Person involved	Pedestrian	Non pedestrian	Fatal	Non fatal
Road Type	Highways	Non highways	Fatal	Non fatal
Time of accident	Day	Night	Fatal	Non fatal
Vacation	Vacation	Non vacation	Fatal	Non fatal
Vehicle age	<5 years	>5 years	Fatal	Non fatal
Vehicle type	Govt. vehicle	Private vehicle	Fatal	Fatalnon
Time of accident	Day	Night	Highways	Non highways

Table 4: A Sample data to illustrate the available accident related information in and around the place of study

Month	Time	Fatal	Nonfatal	Total
January	00.00-03.00	2	1	3
	03.00-06.00	3	3	6
	06.00-09.00	1	13	14
	09.00-12.00	2	10	12
	12.00-15.00	4	10	14
	15.00-18.00	4	12	16
	18.00-21.00	5	17	22
	21.00-24.00	4	8	12

Table 5: Collapsed 2×2 information from Table 2

Time	Fatal	Non-fatal
Day (6 AM-6 PM)	11	45
Night (6 PM-6 AM)	14	29

Where:
d_{ij} = The jth observed effect
θ_i = The mean effect for ith study

The first equation captures the within-subject variability and the between-subject variability is captured by the second equation. Further e_{ij} and z_i are Gaussian errors that have zero mean but variances σ_i^2 and τ^2 respectively. Also the summary statistic approach has been in practice; this is based on sample summary Y_i rather than on all the samples d_{ij}. Hence, Eq. 1 and 2 can be expressed as:

$$Y_i = \theta_i + e_i \quad (3)$$

$$\theta_i = \mu + z_i \quad (4)$$

Or equivalently:

$$Y_i \sim N(\theta_i, \sigma_i^2) \quad (5)$$

$$\theta_i \sim N(\mu, t^2) \quad (6)$$

where, μ and τ^2 are average effect size in the population and amount of heterogeneity in the effect size, respectively. In many applications, REM has been implemented using the robust summary statistic approach. The most widely used summary statistic in such analyses includes risk difference, risk ratio and odds ratio. Odds ratio is good for establishing causal relations and few of its advantages include:

- It is invariable across case control, follow-up and cross-sectional studies and thus it can be used to directly compare findings of different study designs
- It can be computed directly from the regression coefficients of logistic regression
- It is a good estimator of risk ratio if the disease is rare and the cases and controls are randomly selected from the population

Also, it could be observed that information regarding τ^2 might be of considerable scientific interest and multi-site studies closely resemble the statistical principles of meta-analysis. Several significance tests for the presence of heterogeneity have been discussed in the literature such as Q test, Wald, likelihood ratio and score tests for homogeneity. Classical methods for constructing confidence interval for τ^2 include Q statistic given by:

$$Q(t^2) = \sum_{i=1}^{k} \frac{(Yi - \hat{\mu})^2}{\tau^2 + \sigma_i^2} \quad (7)$$

Where:

$$\hat{\mu} = \frac{\sum w_i Y_i}{\sum w_i} \text{ and } w_i = \frac{1}{\tau^2 + \sigma_i^2}$$

The approaches using the log-likelihood function of μ and τ^2 under the REM include wald-type using Fisher information and REML have found considerable applications in the literature. Further, parametric and non-parametric bootstrap methods have been discussed in obtaining confidence intervals for τ^2 (Viechtbaue, 2007). The analysis will provide following summaries to understand the association between the variables in the individual and overall levels together with the amount of heterogeneity:

- Point estimate and confidence interval for the true θ_i of the ith study
- Point and interval estimates of μ to understand the presence or absence of an overall effect and its statistical significance
- Estimates of τ^2 indicating the variation between stations

Numerical summaries and graphical display of the results can directly be available from the appropriate computational tool such as R.

RESULTS AND DISCUSSION

Data analysis: From the collected data, study variables have been identified to understand the nature of associations based on tests for associations and summary measure techniques using REM. In each month from January to November, counts on fatal and non-fatal road accidents have been considered and categorized according to type and age of vehicle involved in the accident, time of accident, persons involved in the accident-classified by age as minor and major, possible hit-run instances, road type, gender, person type and time and road type. Two schemes of tests for the association are considered.

Based on the original I×J tables with Chi square tests and Fisher exact test as well as Chi-square tests for 2×2 tables. The underlying hypotheses are to test association of the first nine variables of Table 1 with the fatality occurrence in a given road accident. For example if the study of interest is to know whether gender involved in the accidents and fatality are statistically associated or not then the null and alternative hypotheses for this case will be:

- H_0: There is no statistically significant association between the gender involved in the accidents and fatality occurrence in an accident
- H_1: There is a statistically significant association between the gender involved in the accidents and fatality occurrence in an accident

Similarly, research and statistical hypotheses can be extended suitably to other variables for the two testing situations. A $p<0.05$ for such tests is commonly interpreted as statistical evidence rejecting the hypothesis of no difference (H_0); a $p\geq0.05$ suggests that observed difference could reasonably be attributed to chance alone.

The implementation of REM will result in the point and interval estimates of μ to understand the presence or absence of an overall effect and its statistical significance; estimates of τ^2 indicating the variation between months and confidence interval for the true θ_i of the ith month. The statistical significance of odds ratio and measure of heterogeneity are based on whether corresponding null values (Odds ratio = 1, $\tau^2 = 0$) lie within the respective confidence intervals.

However, when all cell counts of a row or column are zero, then the expected frequencies obtained using corresponding totals will also be zero which will deter to evaluate chi square statistics. For instance, in the case of I×J classification, fatal counts are zero in both arms of age classification for the months of Jan, Feb, Apr, Jul, Aug and Sep; for the same variable in the case of 2×2 tables, July and August have reflected similar feature. Hence, those tables are not included for the comparative study using chi square tests and Fisher test p-values. Table 6 presents p-values associated with the three test outcomes of variables that is amenable for applying chi square tests.

It can be observed from Table 6 that p values obtained from tests associated with 2×2 tables and that of chi square I×J tables are of same direction; there is no enough evidence to reject the hypotheses of independence in all the cases. However, there are few exemptions for this conclusion that could be observed in the case of time, fatality and road type classifications where an overall approach provides association whereas collapsed categories do not reveal such patterns. Only those cases where the original data has been considered have supported the hypothesis of association and similar conclusions can be made among time of accidents and fatality; 2×2 tables do reflect such associations in these cases. However, when the overall tables are investigated, most of the variables support no association though the accident-related literatures have indicated the significant associations among the similar variables under study.

Table 6: The p-values for the chi square test for independence based on original I×j tables and derived 2×2 tables together with fisher test results for 2×2 tables

		2×2		
Variables of interest	Months	Chi-square	Fisher	Chi-square
Time and road type	January	0.3134	0.3095	0.9096
	February	0.9040	1.0000	0.5820
	March	0.6987	0.6666	0.0392
	April	0.6189	0.5082	0.7236
	May	0.3320	0.3189	0.8307
	June	0.2252	0.1685	0.3466
	July	0.1125	0.0850	0.2947
	August	0.9660	0.8342	0.0021
	September	0.8786	1.0000	0.0664
	October	0.3353	0.2943	0.3053
	November	0.9577	1.0000	0.9434
Fatal and time	January	0.2177	0.1662	0.3199
	February	0.2305	0.1512	0.3959
	March	0.6077	0.6137	0.1141
	April	0.9639	1.0000	0.3693
	May	0.8384	1.0000	0.4863
	June	0.9824	1.0000	0.3368
	July	0.2695	0.1972	0.5747
	August	0.0862	0.0791	0.4609
	September	0.4900	0.4089	0.7434
	October	0.0209	0.0147	0.0411
	November	0.2639	0.2331	0.7624
Fatal and road type	January	0.8118	0.6749	0.1450
	February	0.7860	0.7871	0.1437
	March	0.3325	0.2789	0.7828
	April	0.7470	0.6346	0.6588
	May	0.5134	0.4984	0.2710
	June	0.0080	0.0035	0.0225
	July	0.9670	0.8123	0.9626
	August	0.6979	0.6438	0.0434
	September	0.4857	0.4516	0.5097
	October	0.5946	0.5208	0.2854
	November	0.8075	0.6700	0.8379

Further, the test results calculated using derived 2×2 have shown very similar p-values and these comparisons have been made among all the variables beyond those presented in Table 6. The most important inference is that at least one case in each of the classifications has shown an association between fatality pattern and the variables involved in the accidents. These observations could be thought of as a notion to investigate more on the associations beyond the tests for associations based on exact or approximate methods.

The results of REM have been presented in two panels of Table 7 that include point and interval estimates of odds ratio for each of the classifying variables corresponding to 11 months of study period; whereas, Table 8 summarizes estimates of overall odds ratio and associated measure of heterogeneity. In Table 7, Inf indicates extreme upper limits of individual odds ratio and the significance could be studied by observing the presence of null value (Odds ratio = 1) within a confidence interval.

Table 7: Point estimates of individual odds ratio for the association of fatality with the variables considered in the study together with lower and upper limits of 95% confidence interval

Variables	OR	LL	UL
Time and road type	0.6065	0.2725	1.1052
	0.9418	0.3606	2.4596
	1.2969	0.5543	3.0649
	1.3910	0.5712	3.3535
	1.6000	0.7261	3.5254
	0.4966	0.2231	1.2840
	0.4449	0.1845	1.0833
	0.8958	0.3867	2.0751
	1.0202	0.4724	2.1815
	0.6065	0.2645	1.4049
	1.0513	0.4966	2.2479
Fatal and vehicle type	1.0000	0.0993	10.0744
	0.6313	0.0707	5.6973
	1.3910	0.2516	7.6906
	10.8049	1.9155	61.5592
	3.0344	0.6250	14.7317
	1.6653	0.1604	16.4446
	2.4109	0.3753	15.4870
	1.7507	0.2982	10.2779
	0.4868	0.0257	9.2073
	4.3929	0.9802	19.8857
	1.6161	0.2923	8.9352
Fatal and time	0.5066	0.2019	1.2712
	0.4317	0.1409	1.3231
	1.5068	0.5273	4.2631
	0.8958	0.3198	2.4843
	0.9802	0.3791	2.5345
	1.1853	0.3570	3.9354
	0.5016	0.1827	1.3634
	0.3753	0.1381	1.0202
	0.5945	0.2039	1.7160
	0.2982	0.1153	0.7711
	0.5326	0.2101	1.3364
Fatal and vehicle age	1.1163	0.4493	2.7732
	0.6188	3.0042	1.9155
	0.9048	0.3396	2.3869
	1.0305	0.3791	2.8011
	1.0725	0.4190	2.7732
	0.5886	0.1791	1.9348
	0.7711	0.2837	2.0751
	0.7945	0.2952	2.1170
	1.6653	0.5712	4.8550
	0.9802	0.3135	3.0649
	1.0101	0.3946	2.5857
Fatal and person type	2.7732	1.0942	7.0287
	2.4109	0.7558	7.6141
	1.5841	0.6005	4.1787
	1.3499	0.4584	3.9354
	1.5841	0.6126	4.0960
	2.2705	0.7711	6.7531
	1.3364	0.4677	3.8190
	3.7434	1.3910	9.0250
	0.7634	0.2322	2.5093
	1.0618	0.4107	2.7183
	1.8589	0.7483	4.6646
Fatal and minor	11.1340	0.4360	Inf
	1.4918	0.0578	38.8613
	1.2092	0.0474	30.8766
	1.1052	0.0000	28.2191
	1.4918	0.0584	38.0918
	1.3910	0.0539	35.8735
	3.5254	0.0679	Inf
	3.6328	0.0693	Inf
	1.0202	0.0469	22.4210
	0.3329	0.0172	6.4883
	1.4918	0.0584	38.0918
Fatal and hit run	0.1670	0.0365	0.7558
	0.1423	0.0215	0.9418
	0.2671	0.0498	1.4191
	0.5827	0.0503	6.8210
	0.0863	0.0155	0.4819
	0.6907	0.1249	3.7810
	0.6440	0.1511	2.7456
	0.3606	0.0743	1.7507
	3.2544	0.1791	59.1455
	0.3570	0.1075	1.1972
	0.1075	0.0185	0.6313
Fatal and road type	1.2092	0.5220	2.8292
	1.3634	0.4493	4.0960
	0.5827	0.2369	1.4333
	1.3231	0.5016	3.4556
	1.4770	0.6188	3.5609
	7.1707	1.5841	32.4597
	0.8694	0.3396	2.2479
	1.3364	0.5273	3.4212
	1.6161	0.5945	4.3929
	1.3910	0.5886	3.2871
	1.2214	0.5169	2.9154
Fatal and gender	0.5769	0.2080	1.6000
	2.6379	0.3198	21.9771
	2.2034	0.6005	8.0849
	1.2712	0.3329	4.9037
	0.5220	0.1755	1.5683
	0.9139	0.2322	3.6328
	1.4049	0.4190	4.7588
	2.5857	0.5488	12.1825
	3.7062	0.8025	16.9455
	1.8404	0.4819	6.9588
	6.6859	0.8521	52.4573

Table 8: Point and interval estimates of overall Odds Ratio (OR) for the association of fatality together with the summaries of associated measure of heterogeneity

	Combined odds ratio			Measure of between variance		
Variables of interest	Estimate	LL	UL	Estimate	LL	UL
Time and road type	0.887	0.691	1.139	0.000	0.000	0.371
Fatal and vehicle type	2.226	1.271	3.896	0.000	0.000	1.245
Fatal and time	0.613	0.449	0.827	0.000	0.000	1.245
Fatal and vehicle age	0.942	0.691	1.271	0.000	0.000	0.000
Fatal and minor	1.553	0.571	4.179	0.000	0.000	0.000
Fatal and hit run	0.298	0.181	0.497	0.000	0.000	1.994
Fatal and road type	1.271	0.951	1.682	0.000	0.000	0.683
Fatal and gender	1.391	0.878	2.181	0.111	0.000	1.250
Fatal and person type	1.751	1.297	2.387	0.000	0.000	0.350

It can be observed from Tables 7 and 8 that odds for occurrence of an accident during night are uniformly more in the highways. S. No 1 and 3 of both tables exhibit this pattern with a significant overall association in this direction. Hence, duration of the accident and road type play a significant role in spotting the fatal cases. Further, odds for day time fatalities are also markedly more in the

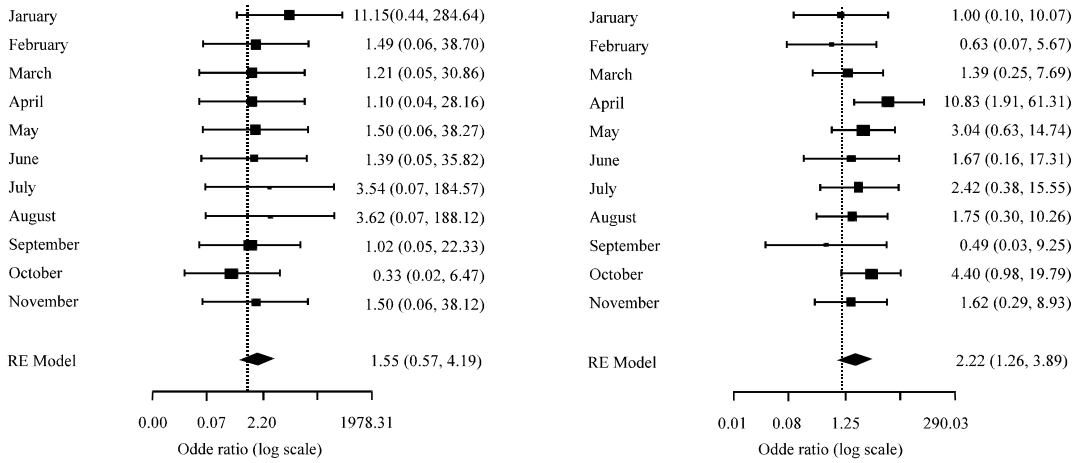

Fig. 1: Forest plot showing the results of individual and combined odds ratio for fatalities among the age group (left panel) and vehicle type (right panel) with corresponding 95% confidence intervals in the individual studies

highways during the months of March and June. This may be attributed to a heavy flow of traffic in the identified road network during day time of these months which record all type of road users. Similar behaviour could be observed from the comparison of fatality over the type of road. Except for two months, all estimates of odds ratio are considerably more for highways and overall estimate supports this findings though estimates lack statistical significance.

Second important association noted is the type of vehicle where odds for fatality is likely to be uniformly more over all the months among government vehicles compared to private users. With few individual significant associations, overall association for this relation is also of same direction and statistical significance can be observed for the type of vehicle. However, age of vehicle might not be so useful to identify a sharp difference among the odds where the individual odds are quite closer to no association value of OR = 1 except for three months. Overall odds also possess this feature and estimates obtained for this relation are statistically significant. Further, fatalities among minors are notably highly uniform over all the period.

Also, the information from Table 7 and 8 could conveniently be represented in pictorial method using forest plot; due to the paucity of space such presentation in Fig. 1 has been restricted to 2nd and 15th cases of Table 8 to illustrate the situation of no association and a possible association in overall measure. Though a lack of statistical significance has been observed for the estimates, odds for minors' fatality are of serious concern and a lower between-variance also emphasize the systematic pattern of minors' involvement in the accidents over the months of study and Fig. 1 exemplifies this observation. Similar feature can be noted when gender is considered and odds for male are high when compared to female, but this could be due to a general phenomenon that male outweigh in numbers in the usage of roads.

The comparison also reveals that odds for fatality in registered accidents are more and the computations yield statistically significant estimates in the case of overall measures and notably in few individual measures. Higher odds for fatality is visible among the pedestrians in all the months except one and overall odds ratio also tends to support this claim with a statically significant association department concerned has not involved in a systematic data collection or analyzing scientifically to understand the rationales and provide a prevention scheme or formulate viable alternatives to reduce the fatalities of human life. With an alarmingly high ratio of fatal-to-non-fatal patterns existing in Indian road accidents this study aims to provide a mechanism for a typical use of available data that helps to understand the significant attributes. Perceived knowledge from similar global scenario has also been useful to proceed with data mining tools to investigate the associations of variables. Hence, an attempt has been made to describe a method to study the nature of count data that are presented in a contingency table of arbitrary size.

First, presence of zeros and low counts >6 (Wong and Chung, 2008) are considered as measures for categorizing the underlying categorical data so as to supplement the data cleaning and understanding of any

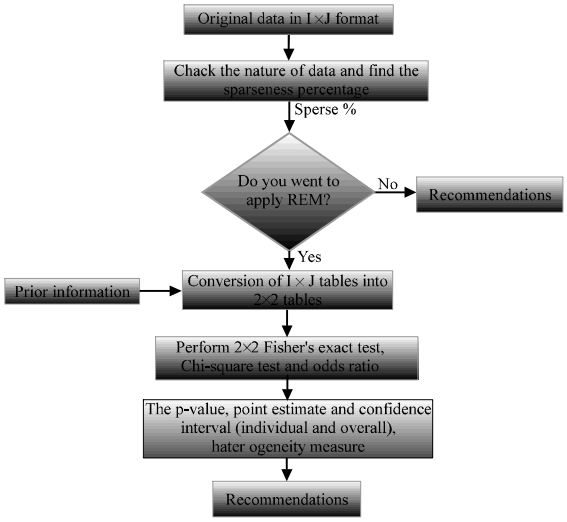

Fig. 2: Schematic representation of the proposed model

data mining process. Second, the limitations of highly used chi square of test of independence have been investigated and recommendations are provided with the usage of REM to probe more using a proper way of data aggregation; additionally Fisher exact test has also been used for a comparative purpose. Third, a comprehensive ready-to-use rcodes are provided for implementing the widely used techniques together with graphical output so as to build a complete data mining cycle from designing a problem to implementation of tools.

The entire representation from the initial data processing to setting analysis plan, preparation to alternative methodology and to conclusions is depicted in Fig. 2. The study has shown important highlights for two user communities.

One aspect for computational purpose, this study reinforces the proper usage of statistical tests and decisions based on p values with a viable alternative for modelling.

A proper understanding of the fatality pattern classified with vital features such as place, time and persons involved in the accidents. The higher rate of fatalities among pedestrians and minors are practically significant and pattern over different months also form reasonable observations together with road types and timings of accidents. This inference could be useful in assisting a typical traffic management with the available data and to set a reasonable experimental design for an extensive data management.

CONCLUSION

Although, descriptive data mining methods are clearly able to uncover reasonable information from the selected traffic accident data set, the results remain at a very general level such that they do not provide much previously unknown new knowledge for the traffic accident experts. Therefore, more detailed data is needed for finding novel facts from data. Data mining seems to produce very understandable and useful results. The lack of detailed accident-specific data hinders the analysis from the road network engineering point of view, because it is currently difficult to analyze the local defects in a particular road segment that might cause further accidents. For example, the data contain no information about seasonal speed limits, “no passing” zones, roundabouts, priority, median barriers, uphill/downhill degrees, curve radius, gravelling, salting, speeding, traffic rule violations (use of seat belts or helmet and aggressive/reckless/careless driving), type of vehicle (cross country vehicle, trailer, etc.), vehicle defects, protective devices (airbag), status/type of driving license, number of years with license, apparent suicide cases, sleepiness, etc. The literature review shows that many of these attributes have been available in other international case studies. Without all this information, it is difficult to evaluate the role of road building, deliberateness of accidents and so on.

It is well-known that chi square analysis has been followed in assessing statistical significance of association rule. Several alternatives as well as computations have been recommended in the place of this procedure especially for sparse yet large contingency tables. This paper makes an attempt to incorporate an REM approach especially for rare events and outcomes. The two-step algorithm provides easier computations of respective measures that facilitate the users to develop rational stratification rules. Such recommendations are deemed fit in analyzing a rare event that may face additional restrictions in terms of adopting special data collections like an epidemiological study.

Hence, the researchers are encouraged to combine the available data reasonably to extract better information for the study. Also, from the analysis point of view, more plausible collapsibility conditions could be designed in future and alternative paradigms such as Bayesian methods could be explored with suitable priors that are constructed from the existing expertise. Regression based on count models could also be attempted for aligning better predictive objectives.

REFERENCES

Bollen, K.A. and J.E. Brand, 2010. A general panel model with random and fixed effects: A structural equations approach. Soc. Forces, 89: 1-34.

Boucher, J.P. and M. Guillen, 2009. A survey on models for panel count data with applications to insurance. RACSAM. Rev. R. Acad. Cien. Serie A. Mat., 103: 277-294.

Bradburn, M.J., J.J. Deeks, J.A. Berlin and L.A. Russell, 2007. Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. Stat. Med., 26: 53-77.

Brijs, T., K. Vanhoof and G. Wets, 2003. Defining interestigness for association rules. Int. J. Inf. Theor. Appl., 10: 370-375.

Caird, J.K., C.R. Willness, P. Steel and C. Scialfa, 2008. A meta-analysis of the effects of cell phones on driver performance. Accid. Anal. Prev., 40: 1282-1293.

Campbell, I., 2007. Chi squared and Fisher-Irwin tests of two by two tables with small sample recommendations. Stat. Med., 26: 3661-3675.

Chen, L.F., H.C. Ho, Y.C. Su, M.S. Lee and S.K. Hung *et al*., 2013. Association between provider volume and healthcare expenditures of patients with oral cancer in Taiwan: A population-based study. PloS., Vol. 8,

Chen, T.H. and C.W. Chen, 2010. Application of data mining to the spatial heterogeneity of foreclosed mortgages. Expert Syst. Appl., 37: 993-997.

Chin, H.C. and M.A. Quddus, 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accid. Anal. Prev., 35: 253-259.

Cummings, P., F.P. Rivara, D.C. Thompson and R.S. Thompson, 2006. Misconceptions regarding case-control studies of bicycle helmets and head injury. Accid. Anal. Prev., 38: 636-643.

Cummins, J.S., K.J. Koval, R.V. Cantu and K.F. Spratt, 2011. Do seat belts and air bags reduce mortality and injury severity after car accidents?. Am. J. Orthopedics, 40: 26-29.

Darby, P., W. Murray and R. Raeside, 2009. Applying online fleet driver assessment to help identify, target and reduce occupational road safety risks. Saf. Sci., 47: 436-442.

Donroe, J., M. Tincopa, R.H. Gilman, D. Brugge and D.A. Moore, 2008. Pedestrian road traffic injuries in urban Peruvian children and adolescents: Case control analyses of personal and environmental risk factors. PloS., Vol. 3.

Dorn, M., W.C. Hou, D. Che and Z. Jiang, 2008. An empirical study of qualities of association rules from a statistical view point. J. Inf. Process. Syst., 4: 27-32.

Eijkemans, M., M. Mommers, M.T. Jos, C. Thijs and M.H. Prins, 2012. Physical activity and asthma: A systematic review and meta-analysis. PloS., Vol. 7.

Elvik, R., 2011. Publication bias and time-trend bias in meta-analysis of bicycle helmet efficacy: A re-analysis of Attewell, Glase and McFadden, 2001. Accid. Anal. Prev., 43: 1245-1251.

Fidler, V. and N. Nagelkerke, 2013. The mantel-haenszel procedure revisited: models and generalizations. PloS., Vol. 8.

Fong, J., S.M. Huang and H.Y. Hsueh, 2007. Online analytical mining association rules using Chi-square test. Int. J. Bus. Intell. Data Min., 2: 311-327.

Gebregziabher, M., L. Egede, G.E. Gilbert, K. Hunt and P.J. Nietert *et al*., 2012. Fitting parametric random effects models in very large data sets with application to VHA national data. BMC. Med. Res. Method., 12: 1-163.

Howell, D.C., 2011. Chi-Square Test: Analysis of Contingency Tables. In: International Encyclopedia of Statistical Science. Lovric, M. (Ed.). Springer Berlin Heidelberg, Heidelberg, Germany, ISBN: 978-3-642-04898-2, pp: 250.

Huang, H. and H.C. Chin, 2010. Modeling road traffic crashes with zero-inflation and site-specific random effects. Stat. Methods Appl., 19: 445-462.

Huang, H., H.C. Chin and M.M. Haque, 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: A bayesian hierarchical analysis. Accid. Anal. Prev., 40: 45-54.

Jackson, D., I.R. White and S.G. Thompson, 2010. Extending dersimonian and laird's methodology to perform multivariate random effects meta analyses. Stat. Med., 29: 1282-1297.

Krieg, A., T.A. Werner, P.E. Verde, N.H. Stoecklein and W.T. Knoefel, 2013. Prognostic and clinicopathological significance of survivin in colorectal cancer: A meta-analysis. PloS., Vol. 8.

Kuypers, K.P.C., S.A. Legrand, J.G. Ramaekers and A.G. Verstraete, 2012. A case-control study estimating accident risk for alcohol, medicines and illegal drugs. PloS., Vol. 7.

Lim, R.H., L. Kobzik and M. Dahl, 2010. Risk for asthma in offspring of asthmatic mothers versus fathers: A meta-analysis. PloS., Vol. 5.

Lv, X., S. Tang, Y. Xia, X. Wang and Y. Yuasn *et al*., 2013. Adverse reactions due to directly observed treatment strategy therapy in Chinese tuberculosis patients: A prospective study. PloS., Vol. 8.

Matsuoka, K., S. Yokoyama, K. Watanabe and S. Tsumoto, 2007. Data mining analysis of relationship between blood stream infection and clinical background in patients undergoing lactobacillus therapy. Proceedings of the IEEE/ICME International Conference on Complex Medical Engineering, May 23-27, 2007, IEEE, Beijing, China, ISBN: 978-1-4244-1078-1, pp: 1940-1945.

Mavros, M.N., V.G. Alexiou, K.Z. Vardakas, M.E. Falagas, 2013. Understanding of statistical terms routinely used in meta-analyses: An international survey among researchers. PloS., Vol. 8,

Mirkin, B., 2001. Eleven ways to look at the chi-squared coefficient for contingency tables. Am. Stat., 55: 111-120.

Olivier, J. and M.L. Bell, 2013. Effect sizes for 2×2 contingency tables. PloS., Vol. 8.

Park, K.H., J.H. Shin, S.Y. Lee, S.H. Kim and M.O. Jang, 2013. The clinical characteristics, carbapenem resistance, and outcome of acinetobacter bacteremia according to genospecies. PloS., Vol. 8.

Petitti, D.B., S.L. Harlan, G.C. Puente and D. Ruddell, 2013. Occupation and environmental heat-associated deaths in Maricopa County, Arizona: A case-control study. PloS., Vol. 8.

Phillips, R.O., P. Ulleberg and T. Vaa, 2011. Meta-analysis of the effect of road safety campaigns on accidents. Accid. Anal. Prev., 43: 1204-1218.

Qi, Y., B.L. Smith and J. Guo, 2007. Freeway accident likelihood prediction using a panel data analysis approach. J. Transp. Eng., 133: 149-156.

Quddus, M., 2013. Exploring the relationship between average speed, speed variation, and accident rates using spatial statistical models and GIS. J. Transp. Saf. Secur., 5: 27-45.

Quddus, M.A., 2008. Time series count data models: An empirical application to traffic accidents. Acci. Anal. Prev., 40: 1732-1741.

Ramani, R.G. and S. Shanthi, 2012. Classifier prediction evaluation in modeling road traffic accident data. Proceedings of the 2012 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), December 18-20, 2012, IEEE, Coimbatore, India, ISBN: 978-1-4673-1342-1, pp: 1-4.

Roudsari, B.S., C.N. Mock, R. Kaufman, D. Grossman and B.Y. Henary *et al*., 2004. Pedestrian crashes: Higher injury severity and mortality rate for light truck vehicles compared with passenger vehicles. Inj. Prev., 10: 154-158.

Shanthi, S. and R.G. Ramani, 2010. Classification of vehicle collision patterns in road accidents using data mining algorithms. Int. J. Comput. Appl., 35: 30-37.

Shanthi, S. and R.G. Ramani, 2012. A comparative evaluation of classification methods in the prediction of road traffic accident patterns. Proceedings of the International Conference on Future Communication and Computer Technology, May 19-20, 2012, International Research Association of Information and Computer Science, Beijing, China, pp: 978-988.

Shanthi, S. and R.G. Ramani, 2012. Classification of seating position specific patterns in road traffic accident data through data mining techniques. Proceedings of the Second International Conference on Computer Applications, January 27-31, 2012, ICCA, Pondicherry, India, pp: 98-104.

Shanthi, S. and R.G. Ramani, 2012. Feature relevance analysis and classification of road traffic accident data through data mining techniques. Proc. World Congress Eng. Comput. Sci., 1: 24-26.

Shanthi, S. and R.G. Ramani, 2012. Gender specific classification of road accident patterns through data mining techniques. Proceedings of the 2012 International Conference on Advances in Engineering, Science and Management (ICAESM), March 30-31, 2012, IEEE, Nagapattinam, Tamil Nadu, ISBN: 978-1-4673-0213-5, pp: 359-365.

Singh, R.K. and S.K. Suman, 2012. Accident analysis and prediction of model on national highways. Int. J. Adv. Technol. Civil Eng., 1: 25-30.

Subbiah, M., B.K. Kumar and M.R. Srinivasan, 2008. Bayesian approach to multicentre sparse data. Commun. Stat. Simul. Comput., 37: 687-696.

Sweeting, J.M., J.A. Sutton, and C.P. Lambert, 2004. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat. Med., 23: 1351-1375.

Usman, T., L. Fu and M.L. Miranda, 2011. Accident prediction models for winter road safety: Does temporal aggregation of data matter?. Transp. Res. Rec. J. Trans. Res. Board, 2237: 144-151.

Valli, P.P., 2005. Road accident models for large metropolitan cities of India. IATSS. Res., 29: 57-65.

Viechtbauer, W., 2007. Confidence intervals for the amount of heterogeneity in meta-analysis. Stat. Med., 26: 37-52.

Wong, J.T. and Y.S. Chung, 2008. Analyzing heterogeneous accident data from the perspective of accident occurrence. Accid. Anal. Prev., 40: 357-367.