Development of hierarchical safety performance functions for urban mid-blocks

Meera J. Krishnan\textsuperscript{a}, Anjana S.\textsuperscript{b}, Anjaneyulu M.V.L.R.\textsuperscript{c}

\textsuperscript{a}M.Tech student, Department of Civil Engineering, National Institute of Technology-Calicut, 673601, Kerala, India.
\textsuperscript{b}Research scholar, Department of Civil Engineering, National Institute of Technology-Calicut, 673601, Kerala, India.
\textsuperscript{c}Professor, Department of Civil Engineering, National Institute of Technology-Calicut, 673601, Kerala, India.

Abstract

Crash frequency and severity are influenced by a variety of variables that represent regional, site, crash and driver-vehicle unit characteristics. In the traditional methods of crash prediction, all the variables are considered at a single level and the multilevel structure inherent in the crash data is ignored. Hierarchical modelling is a statistical technique that allows multilevel data structure to be properly specified and estimated. In the present study, a hierarchical modelling approach was used to estimate the crash frequency and severity of single and dual carriageway roads. Since the crash patterns of single and dual carriageways were found to be different, separate models were developed for these facilities. A two level design was adopted for crash frequency prediction and four level design for crash severity prediction. The two levels in the crash frequency prediction are geographic region level and traffic site level. The additional levels in severity prediction are crash level and driver-vehicle unit level. The study indicates that hierarchical models performed better for crash frequency and severity prediction. Hierarchical models are strongly advocated for crash data that has correlated observations within groups.

Keywords: Mid-block; crash frequency prediction; crash severity prediction; hierarchical models

1. Introduction

Even though there was a marginal decline in road crashes during the year 2011, problem of road safety still remains acute in India. During the year 2011 there were around 4,98,000 road crashes, in which 1,42,485 people were killed and more than 5,00,000 persons were injured. According to the International Road Federation India suffers a loss of Rs. 1 lakh crore (15 billion USD) every year due to traffic crashes. Due to tremendous life and
property loss, more and more attention is placed in various ways to improve the road safety. One promising way is road safety management.

The road safety authorities in India are handicapped without tools and techniques for identifying the correct countermeasures for many road safety problems. Obtaining unbiased and relatively accurate estimation and prediction of traffic system safety is vital for road safety management. By merely seeing, one cannot assess the safety level of a road section. If the road safety can be quantified in terms of some influencing variables, it is possible to detect whether a road section is safe or not. It is also necessary to identify the influence of variables at different levels like regional, traffic site, crash and driver-vehicle unit level on crashes, to have a complete understanding of the road safety problem.

Although the methods to estimate the system safety vary extensively, most studies on road safety rely on traffic crash statistics to address a range of safety related concerns. Hauer (1992) defines system safety as the expected number of crashes in each severity class, which is a characteristic property of a system during a period of time. Crash frequency and severity are the two major safety indicators in measuring road safety (Hauer, 2006). It is the usual practice in safety research to establish a statistical relationship between various risk factors in crash causation and the crash occurrences (Abbas, 2004; Abdel-Aty, 2000; Harwood et al., 2007). This statistical model is called a crash prediction model or safety performance function (Dixon and Zheng, 2013). Safety performance functions provide better understanding of the road safety problems. They aid traffic and road safety engineers to formulate more targeted countermeasures to address road safety issues.

A significant number of studies were conducted to investigate the suitability of various crash prediction models ranging from multiple linear regression models to generalised linear models (Nambuusi et al., 2008). However, these models suffer from an underlying limitation that all samples in the dataset are assumed to be independent of one another. That is, each observation in the estimation procedure corresponds to an individual situation. The residuals from such models exhibit independence. Generally, this independence assumption is not true since multilevel data structures exist extensively in the crash data, either intrinsically or extrinsically from data clustering (Huang et al. 2008). Disregarding the possible heterogeneity between groups produces models with unreliable parameter estimates and statistical inferences (Jones and Jorgensen, 2003). A method to distinctly address the multilevel data structure is to use hierarchical models (also called as multilevel model or random effects model) (Xie et al. 2003). Hierarchical modelling is a statistical technique that allows multilevel data structures to be properly specified and estimated.

This study aims at developing safety performance functions for urban mid-blocks by taking into consideration the hierarchical structure of crash data. The study is conducted to fulfil the following objectives.

1) To identify the variables that influence the crash frequency and severity of urban mid-blocks.
2) To quantify the influence of those variables on the crash frequency and severity.
3) To develop safety performance functions to predict the crash frequency and severity.

In the present study, crash frequency and severity prediction models are developed for single and dual carriageway roads in urban areas using hierarchical modelling approach. The work is limited to the development of safety performance functions of urban mid-blocks only. Data for the study are collected from four cities in Kerala, namely Thiruvananthapuram, Thrissur, Ernakulum and Kozhikode.

2. Literature review

Several research works focus on the relationship between crash occurrences and roadway, traffic, and operational factors. The effect of lane width and number of lanes on safety is inconsistent across several research
works (Bauer, 2004; Hadi et al., 1995; Elvik and Vaa, 2004). Increasing the shoulder width is beneficial to road safety (Hauer, 2000; Harwood et al., 2000). Among the studies that explored the effect of median, Elvik and Vaa (2004) found a decrease in injury crashes and increase in property damage crashes due to presence of median in urban areas. Hauer (2000) found that with increase in median width, cross median crashes decrease and median related crashes increase. Zegeer et al. (1988) observed that removing roadside obstacles from clear zone, decreasing utility pole density and increasing pole offsets reduce crashes. Abdel et al. (2004) observed that longer the length of the roadway section, the more likely the occurrence of crashes. Jiamming (2008) found that injury crash rate increases with an increase in traffic volume. Baruya and Finch (1994) and Garber et al. (2000) found that crash levels increase with increase in speed.

In predicting the crash frequency, Poisson distribution was traditionally employed to model the count data (Mayock and Hall, 1984; Joshua and Garber, 1990; Jones et al., 1991; Miaou and Lum, 1993). In some studies (Miaou, 1994; Vogt and Bared, 1998), the crash data is found to be significantly over-dispersed, that is the variance is much higher than the mean. Under such circumstances the standard Poisson regression model is invalid. Researchers like Miaou (1994), Poch and Mannering (1996) and Abdel-Aty and Radvan (2000) employed negative binomial models and found that negative binomial regression model is better than Poisson model in fitting the over-dispersed crash data.

The distribution of annual crash frequencies with extra zeros may be quantitatively different from the simple Poisson and over dispersed Poisson distributions. To handle this situation, the zero-inflated Poisson regression model is used (Miaou, 1994; Kumara and Chin, 2003).

When crash severity is concerned, discrete outcome distributions are generally used, such as those in nominal models or ordered discrete models. A number of researchers (Mannering and Grodsky, 1995; Shankar and Mannering, 1996; Mercier et al., 1997) used binomial or multinomial logit models to explore the significance of risk factors by taking crash severity as a nominal variable. Other researchers (O’Donnell and Connor, 1996; Rifaat and Chin, 2005; Abdel-Aty and Keller, 2005) employed ordered logit or probit models to account for the ordered nature of severity levels.

The first application of hierarchical models in traffic safety was by Shankar et al. (1998), who showed that introducing site specific random effects and time indicators into the negative binomial regression model can significantly improve the explanatory power of crash models. Chin and Quddus (2003) and Kim et al. (2007) employed hierarchical models for predicting crash frequency while Jones and Jorgensen (2003), Lenguerrand and Laumon (2006) and Huang et al. (2008) developed hierarchical models to identify factors affecting crash severity.

A hierarchical model is a regression model in which the parameters or the regression coefficients are given a probability function. The higher level model has parameters of its own which are estimated from the data. The hierarchical generalised linear model works on the link function that adds disturbance terms to the model corresponding to different sources of variation in the multilevel data.

The main advantages of hierarchical modelling approach are

1) Hierarchical modelling provides a coherent model that simultaneously incorporates both individual-level and group-level effects. In classical models, it is possible to include covariates from all levels, but it is not possible to include the group indicator to account for the omitted or unobservable cross-group heterogeneity.

2) In case of multilevel data, modelling with a complete pooling across all groups would give the average estimate, ignoring variation among groups. Complete pooling ignores variation between groups and the separate models for each cluster overstates it. The modelling paradigm of hierarchical analysis represents a preferred partial pooling, a compromise between these two extremes.
Since hierarchical modelling combines information from both individual level variation and group-level effect, it is possible to model for groups with small sample size. This is impossible in a traditional model where only the local information is used.

Generally, a four level hierarchy represents the general framework of multilevel data structure in traffic safety. The levels are regional, site, crash and driver-vehicle unit level. Inter-regional studies generally include traffic data collected from the regions of interest. This level is normally associated with a number of contextual factors potentially affecting the traffic safety situation such as driving regulations, road density, spatial features, population and other socioeconomic features. Nested under geographic region level is traffic site level, which is of greatest interest in many traffic safety studies. It constitutes the basic road network, namely road segments (link) and road junctions. It is intuitively reasonable that characteristics of crashes occurring at same site should be correlated due to the same context in terms of geometry, traffic and regulatory control factors. Measures like collision type, season and time are used to characterize the traffic crashes. Driver-vehicle unit level including drivers and vehicles involved in a crash is the most concerned entity in traffic safety because it directly relates to the life and property loss. The factors considered at this level are driver age, gender, severity, etc.

3. Data collection and analysis

The various data collected for the study can be classified as regional level, site level, crash level and driver-vehicle unit level data. The regional level data include road length, population density, literacy rate, vehicle ownership, new vehicle registration, annual rainfall, minimum temperature and maximum temperature of the city. The site level data include geometric data, traffic volume data and crash frequency for each accident site. Geometric and traffic data were collected manually. Three hour (8.00 A.M. to 11.00 A.M) classified volume count of the study sites were collected as Average Daily Traffic (ADT) volumes were not available. For the study, 24-hour volume count was taken at two representative locations in each of the four cities. From the 24 hour volume counts, adjustment factors were found out for converting short duration volume count to average daily traffic volume. Exact ADT values of the study locations will greatly improve the precision of the prediction.

Crash data were collected from archived data sources maintained by the police department. One issue with the crash data collected from police records is that not all accidents are reported and not all reported accidents are correctly recorded. Many crashes that do not involve injury are unreported. This leads to overrepresentation of fatal and injury accidents in the crash database. It was seldom easy to trace the exact location of individual crashes from the database. This can affect the accuracy of crash counts of road sections, especially in the case of short sections. The driver-vehicle unit level variables include driver age, gender, vehicle type and severity for each crash. The crash database also lacked details of the victims and only the details of the accused were available for analysis. Reliable estimates of safety can be produced only with the help of comprehensive and accurate crash, geometric, traffic and operational data. In India there is an immediate need for reasonably reliable data to be easily and readily available for analysis and future research in road safety.

Preliminary analysis of the data indicated the variables that are likely to influence the crash frequency and severity of single and dual carriageways. The site level variables that influence crash frequency and severity are traffic volume, access road density, shoulder width, carriageway width, median height and presence of road markings. The other significant variables that affect crash severity are driver’s age and gender, time of crash occurrence and vehicle type. Total vehicle ownership and road length are the variables considered at regional level for severity prediction.

4. Crash frequency prediction

For crash frequency prediction, a two entity level design, with higher level as regional level and lower level as site level was used. The hierarchical Poisson model was used for model calibration. The dependent variable is the
crash frequency expressed as crashes/year (CF). The variables considered at site level are lane width in metres (LW), volume per lane in PCU/day/lane (VPL), number of lanes (NL), number of roadside obstructions per km (RO), number of signs per km (NS/km). The independent variable at regional level is road length in kilometers (RL). The crash frequency prediction models for single and dual carriageway roads are given in equations 1 and 2.

**Hierarchical Poisson model for single carriageway roads**

**Level 1**
\[
\log(CF) = \beta_0 + 0.215752 \times LW + 0.669509 \times NL + 0.005797 \times RO + 0.032913 \times VPL
\]

**Level 2**
\[
\beta_0 = -1.865649 + 0.000553 \times RL + 0.00007
\] (1)

**Hierarchical Poisson model for dual carriageway roads**

**Level 1**
\[
\log(CF) = \beta_0 - 0.015536 \times NS - 0.013202 \times RO + 0.001157 \times VPL - 0.223790 \times LW
\]

**Level 2**
\[
\beta_0 = 4.795284 - 0.000703 \times RL + 0.18068
\] (2)

The models show that increase in traffic volume increases crash frequency. Increase in lane width and number of lanes increases crash frequency of single carriageways. This is due to the increase in operating speed of vehicles. Presence of roadside obstructions also increases crash frequency of single carriageways. For dual carriageways, increase in lane width decreases crash frequency. This is due to the presence of enough space for error correction from the part of the driver. Presence of adequate road signs reduces crash frequency.

The significance of hierarchical model can be found out from four estimates namely reliability estimate, ICC (Intra class Correlation Coefficient), variance component and deviance. Reliability estimate gives how far the regional level variables influence the crash frequency. ICC is an indicator of the within crash correlation. The random effect associated with level 2 is given by the variance component. Deviance is a measure of lack of fit between the model and the data. The various statistics for the single and dual carriageway models are given in Table 1.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Single carriageway</th>
<th>Dual carriageway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability estimate</td>
<td>0.004</td>
<td>0.974</td>
</tr>
<tr>
<td>ICC</td>
<td>10.00</td>
<td>12.64</td>
</tr>
<tr>
<td>Deviance (Current model)</td>
<td>10.74</td>
<td>10.77</td>
</tr>
<tr>
<td>Deviance (Null model)</td>
<td>10.46</td>
<td>11.19</td>
</tr>
</tbody>
</table>

The reliability estimate for single and dual carriageway crash frequency models are 0.004 and 0.974. This means that the variables that influence total crash frequency vary 0.4% and 97.4% between the regions in single and dual carriageways respectively. ICC value of 10 and 12.64 means 10% and 12.64% of the unexplained variation in total crash frequency is due to regional heterogeneity in single and dual carriageway roads. Deviance is less for the current model as compared to the null model, which means that hierarchical models give a better fit to the data.
5. Crash severity prediction

For crash severity prediction, a four level design; regional level, site level, crash level and driver-vehicle unit level is used. The dependent variable at lowermost level is crash severity (SEV). A hierarchical binomial logit model with two outcomes is used for model development. The two possible outcomes are high severity (Y=1) and low severity (Y=0). The fatal and grievous injury crashes are considered as high severity and minor and property damage only crashes are considered as low severity crashes. The region level variables considered are road length in km (RL), vehicle ownership (VO) and new vehicle registrations per year (NVR). Site level variables considered for single carriageway are carriageway width (CW) in metres, shoulder width (SW) in metres, volume per lane (VPL) expressed as PCU/day/lane, presence of road markings (RM), number of lanes (NL) and access road density (AD) expressed as access roads/km. For dual carriageways median width (MW) and median height (MH) in metres are also considered at site level. The variables considered at the crash level are season (SEASON), time (TIME) and collision type (CT). The seasons considered are summer, monsoon and winter. The time of the crash is either day or night. Six types of collisions are considered such as fixed object, inside vehicle, sideswipe, rear end, head on and pedestrian crash. The driver – vehicle unit level variables are driver age (AGE), gender (GEN) and vehicle type (VT). The different types of vehicle considered are cycle, two-wheeler, auto-rickshaw, car, bus and truck. The crash severity prediction models for single and dual carriageway are given in equations 3 and 4 and the various statistics are given in Table 2.

Hierarchical binomial logit model for single carriageway

Level 1
\[
\log \left( \frac{\mu}{1-\mu} \right) = \beta_0 + 0.001596 \cdot AGE
\]

Level 2
\[
\beta_0 = \beta_1 + 0.049327 \cdot CT + 0.02750
\]

Level 3
\[
\beta_1 = \beta_2 - 0.184898 \cdot NL + 0.037822 \cdot LW - 0.067958 \cdot SW - 0.001272 \cdot VPL + 0.00863
\]

Level 4
\[
\beta_2 = 0.616460 - 0.000114 \cdot RL + 0.02269
\]

Here \( \mu \) is the probability that the crash is of high severity that is, \( \mu = Pr (Y=1) \).

Hierarchical binomial logit model for dual carriageway

Level 1
\[
\log \left( \frac{\mu}{1-\mu} \right) = \beta_0 + 0.007178 \cdot VT
\]

Level 2
\[
\beta_0 = \beta_1 + 0.004135 \cdot SEASON + 0.00007
\]

Level 3
\[
\beta_1 = \beta_2 - 0.902949 \cdot RM - 0.012810 \cdot SW - 0.000005 \cdot VPL + 0.00011
\]

Level 4
\[
\beta_2 = 0.792159 + 0.000068 \cdot RL + 0.00001
\]
Table 2. Summary statistics of single and dual carriageway crash severity model

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Single carriageway</th>
<th>Dual carriageway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability estimate of level 2 coefficient</td>
<td>0.101</td>
<td>0.069</td>
</tr>
<tr>
<td>Reliability estimate of level 3 coefficient</td>
<td>0.108</td>
<td>0.074</td>
</tr>
<tr>
<td>Reliability estimate of level 4 coefficient</td>
<td>0.811</td>
<td>0.025</td>
</tr>
<tr>
<td>ICC of level 2 variance</td>
<td>82.90%</td>
<td>55.0%</td>
</tr>
<tr>
<td>ICC of level 3 variance</td>
<td>26.17%</td>
<td>15.80%</td>
</tr>
<tr>
<td>ICC of level 4 variance</td>
<td>68.50%</td>
<td>0.426%</td>
</tr>
<tr>
<td>Deviance of null model</td>
<td>10.097</td>
<td>10.10</td>
</tr>
<tr>
<td>Deviance of current model</td>
<td>9.897</td>
<td>9.897</td>
</tr>
</tbody>
</table>

With increase in lane width, speed increases and hence severity increases. With increase in shoulder width, severity decreases because of the use of wider shoulders as emergency lane. Traffic volume has negative influence on crash severity. From fixed object collision to pedestrian crashes, severity increases. With increase in age, severity increases. With increase in the number of lanes, severity reduces. Reliability estimate and ICC values are calculated for levels 2, 3 and 4. From the reliability estimate, it is clear that the variables influencing crash severity vary 10.1% between crashes, 10.8% between sites and 81.1% between regions. 82.9% of the unexplained variation in crash severity results due to between crash heterogeneity, 26.17% results due to between site heterogeneity and 68.5% results due to regional heterogeneity. Here also deviance of the current model is less than that of the null model. So hierarchical models give a better fit to the crash severity data.

In case of dual carriageway roads, presence of road markings reduces crash severity. With an increase in shoulder width and traffic volume, severity reduces. From summer to monsoon season severity increases. When accused vehicle changes from two-wheeler to truck, severity increases. From the reliability estimate, it is clear that the variables influencing crash severity vary 6.9% between crashes, 7.4% between sites and 2.5% between regions. 55% of the unexplained variation in crash severity is due to between crash heterogeneity, 15.8% is due to between site heterogeneity and 0.426% is due to regional heterogeneity. Here also deviance of current model is less than that of the null model.

RMSE (Root Mean Square Error) values of the hierarchical models were compared with that of single level Poisson and negative binomial models. The RMSE values for single carriageway and dual carriageway crash frequency models are shown in Table 3.

Table 3. RMSE values for various crash frequency prediction models

<table>
<thead>
<tr>
<th>Model</th>
<th>Single carriageway</th>
<th>Dual carriageway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson model</td>
<td>8.04</td>
<td>18.26</td>
</tr>
<tr>
<td>Negative binomial model</td>
<td>6.81</td>
<td>22.64</td>
</tr>
<tr>
<td>Hierarchical model</td>
<td>5.80</td>
<td>8.57</td>
</tr>
</tbody>
</table>
From the Table 3, it is clear that the RMSE value of two level hierarchical models is less compared to single level Poisson and negative binomial model for single and dual carriageways. This shows that hierarchical models best fit the data.

Crash severity models are validated using 141 data of single carriageways and 115 data of dual carriageways. The validation is done based on the predictability of models. A comparison of predictability of hierarchical and single level binomial logit model for single and dual carriageway is given in Table 4.

Table 4. Predictability of hierarchical and single level binomial logit model

<table>
<thead>
<tr>
<th>Model</th>
<th>Single carriageway</th>
<th>Dual carriageway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical binomial logit</td>
<td>73%</td>
<td>70%</td>
</tr>
<tr>
<td>Binomial logit</td>
<td>31%</td>
<td>38%</td>
</tr>
</tbody>
</table>

From the Table 3, it is understood that the predictability of hierarchical model is better compared to the single level model.

6. Conclusion

This study identified the risk factors that influence crashes at urban mid-blocks, considering the hierarchical structure of crash data. The study showed that crash frequency is influenced by regional level and site level variables whereas crash severity is influenced by regional level, site level, crash level and driver-vehicle unit level variables. Road length, lane width and road side obstructions have a positive influence on single carriageway crash frequency whereas they have a negative influence on dual carriageway crash frequency. Traffic volume has a positive influence on crash frequency of both single and dual carriageway roads. The crash severity prediction models show that larger lane width causes more severe crashes at single carriageway roads. With a wider shoulder width and increase in traffic volume, the severity of crashes at single and dual carriageways decreases. The presence of road marking decreases severe crashes in dual carriageway. Increase in victim’s age and change in collision type from fixed object to pedestrian collision results in more severe crashes at single carriageways. When the accused vehicle changes from two-wheeler to truck, the severity of crashes at dual carriageways increases. For crash prediction, hierarchical approach has more predictability compared to traditional approaches.

Acknowledgements

We express sincere gratitude to the National Institute of Technology Calicut, Ministry of Urban Development and Centre of Excellence in Urban Transport, IIT Madras for funds and support rendered throughout the research work.

References


