https://doi.org/10.65646/3rc20dmd1v1192

MODELING ROAD TRAFFIC ACCIDENTS IN COLOMBO USING BINARY LOGISTIC REGRESSION MODEL AND SPATIAL ANALYSIS

V. Dayapoorani* and R. Punchi-Manage

Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Sri Lanka

Road traffic accidents have become a serious concern, causing millions of deaths and severe injuries annually. In the analysis of road accidents in Sri Lanka, Colombo contributes significantly. This study aims to identify and quantify the factors influencing road accidents and accident severity in the Colombo Municipal Council area using a binary logistic regression model, spatial statistics and circular statistics. The accident records were obtained from the City Traffic Police Station, Fort. The fitted model assesses the probability of a binary level of accident severity (minor, severe) based on predictors such as "Type of Day", "Area Type", "Day of the Week", "Weather Condition", "Location Type", "Number of Vehicles Involved", "Light Conditions", "Weekday or Weekend", "Time Factor". The model was optimized using stepwise AIC selection and assessed via odds ratios and p-values. The number of vehicles involved in a collision, "Day or Night" and "Weekend or Weekday" variables are showing a positive association with the accident severity. Odds ratios suggest that accidents that occurred at night are 52.2% more likely to be serious than those during the day, and accidents on weekends are 48.1% more likely to be serious compared to weekdays. Hence, we can conclude that accident severity tends to be higher during nighttime, weekends, and weekday mornings compared to other times. Circular statistics indicated that the frequency of the accidents is high between 7:00-8:00 a.m., 1:00-2:00 p.m., and 4:00-5:00 p.m. in a day, which may be due to the high volume of traffic before and after school hours and the end of the typical workday. Moran's I statistic showed a positive moderate spatial autocorrelation of 0.4383, which indicates that accident counts are clustered and not randomly spread, suggesting that there is a high accident count near other high counts or a low accident count near other low counts. Spatial maps show that the highest number of accidents occurred in the Cinnamon Garden Grama Niladhari Division. These findings aim to facilitate decision-making regarding infrastructure and safety interventions that can be implemented to improve traffic safety and ultimately contribute to mitigating accidents and enhancing overall road safety in the region.

Keywords: binary logistic regression model, road accident severity, spatial autocorrelation, Global Moran's I, circular statistics

**Corresponding Author: s19585@sci.pdn.ac.lk*

MODELING ROAD TRAFFIC ACCIDENTS IN COLOMBO USING BINARY LOGISTIC REGRESSION MODEL AND SPATIAL ANALYSIS

V. Dayapoorani* and R. Punchi-Manage

Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Sri Lanka.

INTRODUCTION

Worldwide, nearly 1.35 million people lose their lives in traffic accidents each year, and between 20 and 50 million more suffer major injuries, according to the World Health Organization (WHO). Currently, the eighth most common cause of mortality worldwide, road accidents could rise to the seventh position by 2030 if present trends continue (Danthanarayana & Mallikahewa, 2021). In order to develop effective strategies to reduce the frequency and severity of traffic accidents, this study primarily focuses on identifying the factors that contribute to these incidents. The appropriate authorities can then carry out focused interventions to successfully address these factors. Spotting high-risk areas is also central to many safety analyses, as knowledge of these locations helps maximize the efficiency of investments in traffic safety improvements within a country. This study attempts to identify and quantify the factors influencing road accident severity in Colombo Municipal Council area (CMC) using a binary logistic regression model. Depending on predictors like weather conditions, day type (workday/holiday), time of day, number of vehicles, and other contextual factors, the model assesses the likelihood of various accident severity levels (e.g., minor, severe). It analyses which variables are most significantly linked to increased or decreased severity using statistical techniques (p-values, odds ratios, and stepwise AIC selection).

METHODOLOGY

Study area: The dataset was collected form the City Traffic Police Station, Fort. It contains 10,412 records of accidents that occurred from 2019 to 2023 in the Colombo Municipal Council (CMC) Area. Roads in the Colombo Municipal Council area are extracted from a shape file (https://data.humdata.org/dataset/hotosm_lka_roads; Fig. 1).

Binary Logistic Regression: We used a binary logistic regression model to predict accident severity (Y_i), where the outcome is binary (severe = 1, minor = 0). The model has the following form: $ln\left[\frac{\Pr(Y_i=1|X)}{\Pr(Y_i=0|X)}\right] = \beta^T X$ ---(1). Where β is the vector of coefficients and X represent the set of predictor variables. $Pr(Y_i = 1|X)$ is the probability of observing a severe accident given predictors X is defined as: $Pr(Y_i = 1|X) = \frac{\exp(\beta^T X)}{1+\exp(\beta^T X)}$ ---(2). Here, $Pr(Y_i = 0|X)$ represents the probability of observing minor accidents (reference category) is expressed as: $Pr(Y_i = 0|X) = \frac{1}{1+\exp(\beta^T X)}$ ---(3). The odds ratio (OR) calculated as: $OR = \exp(\beta_j)$---(4). It represents the change in the odds of $Y_i = 1$ for a unit increase in X_j, holding the other variables constant. To assess the statistical significance of individual coefficients, Wald test statistic is used: $Z = \frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)} \sim N(0, 1)$. ---(5). This study Y_i has two severity levels (0 and 1). The value 0 represents the minor severity accidents and serves as the baseline category. In this model we used several predictors X: "*Type of Day*", "*Area Type*", "*Day of the Week*", "*Weather Condition*", "*Location type*", "*Number of vehicles involved*", "*Light Condition*", "*Weekday or Weekend*" and "*Time factor*". Each predictor has multiple levels, with one level designated as the baseline. The analysis was performed using the glm() function in R (R Core Team, 2024) and stepwise AIC (Akaike Information Criterion) based selection. This selection method systematically adds or removes predictors to identify the most parsimonious model with optimal explanatory power. The stepAIC() function from the MASS package in R was used to select the best fitted model (Venables & Ripley, 2002). The model diagnostics were conducted by assessing coefficients, standard error, z-values and p-values, supported by the computation of odds ratio to quantify the effect of each factor on accident severity. The confusion matrix was used to evaluate its predictive accuracy.

Spatial autocorrelation- Global Moran's I: In this research, Global Moran's I was used to determine the spatial autocorrelation, which is used to identify whether the spread of accidents is clustered, dispersed or random over the study area. $I = \frac{N\,\Sigma_i \Sigma_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\Sigma_i (x_i - \bar{x})^2\, \Sigma_i \Sigma_j w_{ij}}$ ---(6). *N*-total number of spatial units, x_i and x_j represent the observed accident counts at locations i and j, respectively, $\bar{x}$ is the mean of the accident count, and w_{ij} is the spatial weight between locations i and j. Moran's I >0 indicates a positive spatial autocorrelation, which means accident counts are clustered together (e.g., high accidents count near other high counts, or low near low). Moran's $I \approx 0$ means there is no spatial autocorrelation. Therefore, the accident counts are random, and the spatial arrangement does not influence the distribution of values. Moran's I < 0 indicates a negative spatial autocorrelation. That are dissimilar values are adjacent (i.e., high values surrounded by low values, and vice versa). This suggests a dispersed pattern or checkerboard like spatial pattern. Spatial analyses were performed using "spatstat", and "tmap" packages in R (Baddeley, Rubak, Turner 2015; Tennekes M 2018). Moran's I test was performed using the "spdep" package in R (Bivand, Pebesma, Gómez-Rubio, 2013).

Circular statistics: Circular statistic is used to visualize the distribution of road accidents across different times of the day using a rose diagram (circular bar plot), which makes it easier to detect time-of-day patterns.

RESULTS AND DISCUSSION

Binary Logistic Regression: We obtained the following results (Table 1). The model uses "minor" accident severity as the baseline category. Positive coefficients indicate higher log-odds of a serious accident relative to a minor one, while negative coefficients suggest a lower likelihood. Several "*Location Type*" categories are significantly associated with reduced odds of serious accidents. *Location types* 1,2,3,5,6,7, and 8 show statistically significant negative coefficients, indicating substantially lower odds of serious accidents. For example, accidents at "*Location Type 7"* are approximately 76.3% less likely to be serious, while those at "*Location Type* 8", which is approximately 87.9% less likely. With a coefficient of 0.324 ($p < 0.001$), the number of vehicles involved in an accident is positively associated with accident severity, meaning that the log-odds of a serious accident increase with each additional vehicle (Table 1). A significant positive coefficient of 0.420 ($p < 0.001$) for the "*Day or Night*" variable indicates that accidents at night are 52.2% more likely to be serious than those during the day, with an odds ratio of roughly 1.522 (Table 1). Additionally, there is a positive correlation between severity and the "*weekday or weekend*" variable (0.393, $p < 0.001$), meaning that accidents that occur on weekends are 48.1% more likely to be serious compared to weekdays (Table 1). The "*Time Factor*" variable also contributes to increased severity (0.244, $p < 0.01$), suggesting that accidents occurring on weekdays between 6 and 8 AM are 27.6% more likely to be serious than during the reference time (Table 1). The model intercept is -0.556 and is not statistically significant, indicating no strong baseline log-odds of serious accidents after accounting for all predictors (Table 1). The VIF values are very low; therefore, no multicollinearity is present. The Hosmer–Lemeshow Goodness of Fit test does not reject the null hypothesis (H_0: the model fits well). The confusion matrix shows 93.8% sensitivity, indicating that the model almost always predicts class 1 correctly. However, it has low specificity.

Table 1: Result of Binary Logistic Regression

Independent variables	Coefficient	Std. Error	Odds ratio
Location Type 1[Stretch of road]	-0.702*	(0.334)	0.496
Location Type 2[4-leg junction]	-0.669*	(0.339)	0.512
Location Type 3[T-junction]	-0.654	(0.337)	0.519
Location Type 4[Y-junction]	-0.562	(0.371)	0.569
Location Type 5[Roundabout]	-0.912**	(0.344)	0.402
Location Type 6[Multiple-road junction]	-1.016**	(0.372)	0.362
Location Type7 [Entrance by road]	-1.440**	(0.452)	0.237
Location Type 8[Railroad crossing]	-2.114***	(0.584)	0.120
Location Type9 [Others]	-0.588	(0.365)	0.555
Number of Vehicles	0.324***	(0.050)	1.382
Day or Night [Night]	0.420***	(0.050)	1.522
Weekday or Weekend [Weekend]	0.393***	(0.053)	1.481
Time Factor [Morning Weekday]	0.244**	(0.091)	1.276

Time Factor [Noon Weekday]	-0.214**	(0.079)	0.807
Constant	-0.556	(0.349)	0.574
Observations 8,921	Log Likelihood -5,866.177		
Akaike Inf. Crit. 11,762.350	Model Deviance: 11732.35		
Model AIC: 11762.35			
Note: *p<0.05; ** p <0.01; ***p <0.001. Only significant results were shown.			

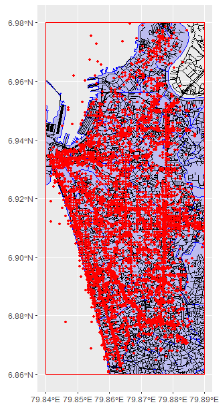

Figure 1: Spatial Pattern of Road accidents in CMC area

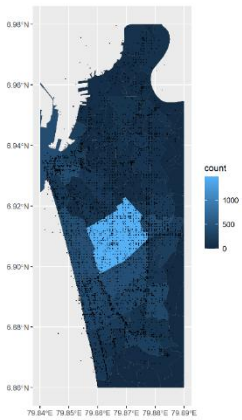

Figure 2: Road accidents count in each GN divisions

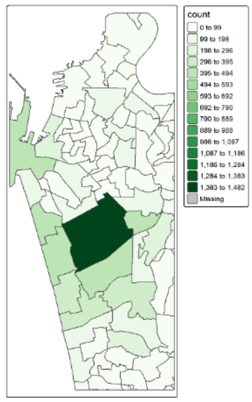

Figure 3: Road accident count areal data.

Spatial Maps & Moran's I Statistics: Figure 1 indicates the spatial distribution of the accidents (in red) and road networks (black). In Figure 2 & Figure 3, the intensity of accident counts in the Grama Niladari divisions (GND) is shown. Cinnamon Garden GN division indicates the highest intensity for accidents in Colombo Municipal council area. The results of Moran's I statistic show a moderate positive spatial autocorrelation of 0.4383, which can be concluded that the accident counts are clustered and not randomly spread. It has a high accident count near other high counts, or a low accident count near low counts.

Circular Statistics: The rose diagram depicts, most of the accidents occur between 7:00 and 8:00 a.m., 1:00 and 2:00 p.m., and 4:00 and 5:00 p.m., which means that the time of day may significantly affect the frequency of accidents. These peaks are probably caused by heavy traffic at the end of the regular workday and during the opening and closing hours of schools.

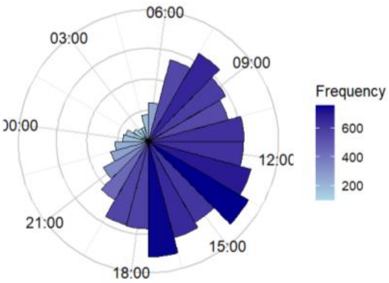

Figure 4: Rose Diagram of 24-hour accident distribution

CONCLUSIONS/RECOMMENDATIONS

The fitted binary logistic regression model evaluates the impact of several independent factors on accident severity, with plans to visualize these effects as percentage changes in an upcoming graph. The results indicate that accidents tend to be more severe at night, on weekends, and on the mornings of weekdays compared to other times. Furthermore, there is a clear association between the number of vehicles involved and the severity of the accident. The additional variables, like driver demographics, vehicle types, road conditions, traffic volume, and enforcement levels, could improve the model's accuracy and provide more thorough insights into the factors influencing accident severity. Targeted interventions can be put in place to improve traffic safety by identifying these factors, which will ultimately help to reduce accident rates and improve overall road safety in the region. The accident-prone areas of the Colombo Municipal Council were identified using spatial data and advanced spatial modelling techniques. Given that high accident clusters frequently point to underlying causes like traffic congestion or road geometry, this strategy may make data-driven decisions about infrastructure and urban safety easier.

REFERENCES

Baddeley A, Rubak E, Turner R (2015). Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press, London.

Bivand R, Pebesma E, Gómez-Rubio V (2013). Applied spatial data analysis with R, Second edition. Springer, NY.

Danthanarayana, C. T., & Mallikahewa, S. N. (2021). An analysis of the enduring factors of road traffic accidents in Sri Lanka. *Sri Lanka Journal of Economic Research, 8*(2), 1–22.

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Tennekes M (2018). tmap: Thematic Maps in R. Journal of Statistical Software, 84(6), 1-39.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

ACKNOWLEDGMENT
The data used in this study is provided by City Traffic Police Station, Fort. We gratefully acknowledge their support.