Aggregate nonparametric safety analysis of traffic zones

Chowdhury Siddiqui, Mohamed Abdel-Aty, Helai Huang

A R T I C L E   I N F O
Article history:
Received 15 April 2011
Received in revised form 15 June 2011
Accepted 26 July 2011

Keywords:
Macro-level safety analysis
Total crash
Severe crash
Data mining
Random forest
Safety planning

A B S T R A C T
Exploring the significant variables related to specific types of crashes is vitally important in the planning stage of a transportation network. This paper aims to identify and examine important variables associated with total crashes and severe crashes per traffic analysis zone (TAZ) in four counties of the state of Florida by applying nonparametric statistical techniques such as data mining and random forest. The intention of investigating these factors in such aggregate level analysis is to incorporate proactive safety measures in transportation planning. Total and severe crashes per TAZ were modeled to provide predictive decision trees. The variables which carried higher weight of importance for total crashes per TAZ were — total number of intersections per TAZ, airport trip productions, light truck productions, and total roadway segment length with 35 mph posted speed limit. The other significant variables identified for total crashes were total roadway length with 15 mph posted speed limit, total roadway length with 65 mph posted speed limit, and non-home based work productions. For severe crashes, total number of intersections per TAZ, light truck productions, total roadway length with 35 mph posted speed limit, and total roadway length with 65 mph posted speed limit were among the significant variables. These variables were further verified and supported by the random forest results.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The recent emphasis on transportation safety planning (TSP) issues has shed light on the importance in dealing with macro-level crash data. Incorporating safety in the transportation planning stage has been challenging from different perspectives. For example, aggregated crash data is criticized of having problems with heterogeneity. Also, planning for an unused land area or even for an existing area requires considerations from socio-economic, demographic and different traffic-related factors. The interplay between these factors becomes a significant problem for prediction during the planning stage.

For the last decade researchers have been trying to capture significant factors associated with crashes at different aggregation levels. In this study we consider traffic analysis zones (TAZs) as the base spatial unit of analysis. That means, crashes were aggregated for a TAZ. TAZs are thought to be homogenous in demographic and land use patterns (You et al., 1997). Also they are widely used building blocks of transportation planning networks and frequently used by metropolitan planning organizations (MPOs) in their travel demand models (TDMs) and various traffic related analysis.

This paper aims to identify and examine important variables associated with total and severe crashes per TAZ. In doing so, different trip related variables and road-traffic related factors were investigated. Very few studies have addressed the effect of trip related variables onto crashes. Trips usually depend on demographic and land use patterns of an area. Thus, trip variables were thought to be advantageous in accounting for different correlated effects between demographic and land use factors. Investigating these factors in such aggregate level analysis will be helpful in incorporating proactive safety measures for long range transportation planning. Data mining techniques such as decision trees and random forests were applied to find important factors and develop predictive crash models. These techniques are nonparametric in nature and do not depend on any functional form. Also, they account for missing values and outliers and can handle a dataset with large number of predictors without being affected by the multicollinearity problem between them. The superiority and robustness of tree based algorithms are extensively discussed in Karlafis and Golas (2002), Chang and Wang (2006), Harb et al. (2009), and Yan et al. (2010).

2. Review of literature

Macroscopic safety modeling has been addressed so far at different spatial aggregation starting from census block to state level. Noland (2003) analyzed how various road infrastructure