Neurocomputing 181 (2016) 53–63

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Investigating macro-level hotzone identification and variable importance using big data: A random forest models approach

Ximiao Jiang [a], Mohamed Abdel-Aty [b], Jia Hu [a,*], Jaeyoung Lee [b]

[a] Office of Operation R&D, Federal Highway Administration, McLean, VA 22101, USA
[b] Departmen of Civil, Environmental & Construction Engineering, University of Central Florida, Orlando, FL 32816, USA

ARTICLE INFO

Article history:
Received 15 March 2015
Received in revised form
3 June 2015
Accepted 27 August 2015
Communicated by T. Heskes.
Available online 28 November 2015

Keywords:
Hotzone identification
Big data
Connected Vehicle
Variable importance
Random forest
Wilcoxon test

ABSTRACT

As Connected Vehicle technologies begin to be deployed along roadway networks, they will be providing massive amount of data. This big data can be useful in identifying safety hazardous zones, which can be complicated and unreliable today. Without sufficient data, past studies had to focus mostly on the micro-level networks. Research on macro-level hotzone identification is limited, and until this point, the contribution of various macroscopic features on the macro-level crash risks is still in dispute. This paper, with the help of massive amount of data, investigates the feasibility of using random forest for hotzone identification at macro-level – the Traffic Analysis Zone (TAZ) level. At the same time, the most influential macro-level crash risk determinants were identified by applying a series of random forest models in combination with the cross validation methods. The differences of all features between hotzones and normal TAZs were also recognized through Wilcoxon tests. Crash data of three counties in Florida during 2008 and 2009 were employed. Crash risks by different injury levels and collision types were investigated separately. Finally, the significance of various macroscopic variables was determined by different types of crash risks using variable importance analysis.

The research results suggest that the distribution of road network and socio-economics are the two most important factors when proactively alleviating traffic safety issues. For developed urban areas, it is desirable to formulate specific traffic safety management strategies that accounts for zone-level socio-economics and development of road infrastructure. For zones with a higher percentage of school enrollment, pedestrian and bicycle friendly roadway system design are most beneficial. It is also desirable to take efficient countermeasures such as law enforcement and driving school training to regulate young drivers' behavior in school zones. For areas with high minority residence, there might be a need to use awareness campaigns in multiple languages to relieve pedestrian safety issues. Finally, additional attention should be paid to improve intersection design and management during the planning and operation processes.

Published by Elsevier B.V.

1. Introduction

As the roll-out of the Connected Vehicle (CV) technology, vehicles become part of the road network monitoring system. Diagnostic sensors installed will enable vehicles to collect much more information than the conventional data collection equipment. A great number of measurements that are previously unavailable would become known, which include but not be limited to: vehicle speeds, positions, arrival rates, rates of acceleration and deceleration, queue lengths, stopped time and so on. These information, if mined properly, could significantly help us understand traffic. In the past, many studies have researched on this area from a traffic operation point of view, for instance, real time traffic state estimation [41,51], but few has looked at it from a safety perspective. One potential way of making use of big data for safety is safety hazardous areas identification. By doing this, together with GPS system installed on CV enable vehicles, warnings could be sent to drivers through Dedicated Short Range Communication (DSRC) system to keep drivers on alert in order to reduce collisions.

While many studies have contributed to the identification and improvement of traffic safety issues on individual traffic sites such as road segments and intersections, a growing body of literature shows significant influence of macro-level features on the occurrence of crashes. Due to the sustained high number of traffic crashes in the United States, there is an urgent need to conduct efficient transportation planning from the traffic safety perspective. Accordingly, the

* Corresponding author.
E-mail addresses: xjiang10@vols.utk.edu (X. Jiang), M.Aty@ucf.edu (M. Abdel-Aty), jh8dn@virginia.edu (J. Hu), jaeyoung@knights.ucf.edu (J. Lee).

http://dx.doi.org/10.1016/j.neucom.2015.08.097
0925-2312/Published by Elsevier B.V.

Transportation Equity Act for the 21st Century (TEA-21) and the Safe, Accountable, Flexible, Efficient, Transportation Equity Act: A Legacy for Users (SAFETEA-LU) proposed the requirement to incorporate safety into the transportation planning process [42,43].

In recent decades, quite a few scholars have investigated traffic safety propensity on various macroscopic levels, such as block groups [29], traffic analysis zones or TAZs [3,33,37,38,47], census tracts [13,30,45,48], counties [23,5,6], and other levels [26]. Among these spatial units, TAZ has been widely adopted for traffic safety analysis. A TAZ is a statistical entity delineated by officials at states' Department of Transportation (DOTs) and/or local Metropolitan Planning Organizations (MPOs) for tabulating traffic-related census data such as, journey-to-work and place-of-work statistics [44]. Therefore, from a transportation planning perspective, TAZs seem to be preferred spatial entities as compared to other spatial units.

An increasing number of researchers have attempted to estimate the effects of various macroscopic features on the occurrence of crashes. Different types of crashes in terms of injury severity levels, collision types, non-motorized crashes and many others were investigated. Variables considered in previous studies can be categorized into classes. The major predictors in these classes include:

- Road network features: percentage of roads by functional classes [14,21,34], percentage of roads by speed limits [22,3,34], road network structures [14,49].
- Demographic characteristics: population density [22,29], percentage of age and gender groups [22,5,6], percentage of minorities [23,39].
- Land use attributes: household units [7,22,20], total shopping/commercial, educational, recreational and other land use per unit area [3,32].
- Socio-economical factors: Vehicles per household [28,34]; income level [13,37], employment rate [20,34].

From a methodological perspective, a wide spectrum of modeling approaches has been adopted to estimate the effects of macroscopic features on the occurrence of crashes. The primary methods include Negative Binomial (NB) models [14,20,21,27,37,6,5], ordinary least square regression models [48], log-linear models [50], geographically weighted regression (GWR) models [20], spatial lag models [29], Bayesian hierarchical models [34] and Bayesian models accounting for spatial autocorrelation [23,49].

The above mentioned studies have contributed substantially to macro-level crash risk research. However, the significance of various features is still in dispute, and the ranking of these factors in contributing to macro-level traffic safety is still unknown. In addition, regression models have the capacity to estimate the influence of independent variables on crash frequency, but they have difficulty in detecting and interpreting complex or high-order interactions among independent variables [31,40]. Moreover, most regression models have specific assumptions and pre-defined underlying relationships between dependent and independent variables. If these assumptions are violated, this could lead to biased parameter estimates. On the contrary, machine learning techniques can identify associations between dependent and independent variables without requiring assumptions. Especially, these techniques are superior to traditional regression procedures in that they are robust to outliers and are capable of detecting complex interactions [17].

The major machine learning techniques applied to traffic safety studies include but are not limited to decision tree based models [10,12], neural networks [35,36], and artificial neural networks [1,15,2]. Among these models, classification and regression trees are known for their transparency [46]. However, decision trees are susceptible to small perturbations in the learning set. Due to this limitation, the random forest (RF) technique was introduced by Breiman [8], which incorporates Breiman's bagging idea into Ho's "random subspace method". The random subspace method was created to construct a collection of decision trees with controlled variations [25]. In recent years, the RF method has been widely employed for classification and prediction purposes in various applications due to its superior performance and its relative simplicity in design. For example, Abdel-Aty et al. [4] employed the RF method to evaluate traffic safety on Dutch freeways. There, the authors argued that the RF method is a more robust variable selection tool as it exhausts a collection of multiple tree classifiers as compared to one single decision tree. Harb et al. [24] used RF models to recognize important factors that are associated with traffic crash avoidance maneuvers. Results suggest that the RF method is capable of identifying important determinants of crash avoidance maneuvers. Particularly, Siddiqui et al. [38] employed decision tree and RF models to identify and examine important variables that are associated with total crashes and severe crashes per TAZ in four counties in Florida. The study provided a rank of variable importance. However, their study focused solely on trip production and attraction related variables, and the efficiency of using RF models for macro-level hotzone identification was not investigated.

The current research defines the TAZs with top 25% crash risk measures as hotzones, and the remaining as normal zones. The objective of this paper is to investigate the feasibility of using RF models on big data to conduct TAZ-level hotzone identification, and to recognize the most important variables that could predominately determine the TAZ-level crash risks. These factors would help transportation planners to incorporate safety into the transportation planning process. Crashes by injury levels (total, fatal and injury (FI)), non-motorized crashes (pedestrian and bicycle) and collision types (rear-end, angle and sideswipe) were investigated separately. At the same time, the appropriateness of alternative crash risk measures (crashes per square mile, crashes per mile and crashes per MVMT) was examined, and the one that achieved the best performance was selected for the current study. In order to quantitatively understand the effects of all variables, wilcoxon tests were conducted to compare the differences of each variable between hotzones and normal zones. The complete organization of this paper is sketched in Fig. 1.

2. Methodology

2.1. Data preparation

Three counties in Central Florida were selected for this study; Orange, Seminole and Osceola. These counties are divided into 1116 TAZs. The sizes of these TAZs range from 0.02 mi^2 to 170.66 mi^2, with a mean of 2.54 mi^2. Crashes that occurred in 2008 and 2009 were obtained for the analysis. Two forms of crash report are used in the State of Florida, short form and long form crash reports. A long form is used when the following criteria are met: (1) death or personal injury; (2) leaving the scene involving damage to attended vehicles or property, and (3) driving while under the influence. Also the police officer can complete a long form for a PDO crash at his discretion. Whereas a short form is used to report other types of property damage only (PDO) traffic crashes. Both long form and short form Crash data were collected from Florida Department of Transportation (FDOT) and MetroPlan Orlando (the Orlando Metro-Planning Organization-MPO), respectively. As a result, a total of 86,828 crashes were collected and among them, 22,575 are injury or fatality involved crashes,