Using data mining techniques to predict the severity of bicycle crashes

Gabriele Prati*, Luca Pietrantoni, Federico Fraboni

Dipartimento di Psicologia, Università di Bologna, Viale Europa 115, 47521 Cesena, FC, Italy

A B S T R A C T

To investigate the factors predicting severity of bicycle crashes in Italy, we used an observational study of official statistics. We applied two of the most widely used data mining techniques, CHAID decision tree technique and Bayesian network analysis. We used data provided by the Italian National Institute of Statistics on road crashes that occurred on the Italian road network during the period ranging from 2011 to 2013. In the present study, the dataset contains information about road crashes occurred on the Italian road network during the period ranging from 2011 to 2013. We extracted 49,621 road accidents where at least one cyclist was injured or killed from the original database that comprised a total of 575,093 road accidents. CHAID decision tree technique was employed to establish the relationship between severity of bicycle crashes and factors related to crash characteristics (type of collision and opponent vehicle), infrastructure characteristics (type of carriageway, road type, road signage, pavement type, and type of road segment), cyclists (gender and age), and environmental factors (time of the day, day of the week, month, pavement condition, and weather). CHAID analysis revealed that the most important predictors were, in decreasing order of importance, road type (0.30), crash type (0.24), age of cyclist (0.19), road signage (0.08), gender of cyclist (0.07), type of opponent vehicle (0.05), month (0.04), and type of road segment (0.02). These eight most important predictors of the severity of bicycle crashes were included as predictors of the target (i.e., severity of bicycle crashes) in Bayesian network analysis. Bayesian network analysis identified crash type (0.31), road type (0.19), and type of opponent vehicle (0.18) as the most important predictors of severity of bicycle crashes.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

It is recognized that the use of bicycle as a mode of transport is associated with environmental and societal benefits (De Nazelle et al., 2011; Xia et al., 2013; Macmillan et al., 2014) as well as health benefits (Kelly et al., 2014; Götschi et al., 2016). However, there are also societal costs of bicycle use, especially in terms of consequences of bicycle crashes.

In Europe, 8% of people choose bicycles as the most common mode of daily transport (European Commission, 2014). Nevertheless, cyclists still represent one of the road user categories with the highest risk of injuries and fatalities. From 2004–2013, cyclists' fatalities decreased by 32%, but from 2010 this tendency has stagnated, with less than a 1% year-to-year reduction. Furthermore, 31% of the fatalities happen at junctions (European Commission, 2015). Risks for non-fatal accidents are higher for cyclists than for car drivers (De Hartog et al., 2010).

Similar to European data, in Italy, 6% of the population indicates the bicycle as the most common mode of transport (European Commission 2014). In 2014, there were 18.055 bicycle accidents and 273 cyclists' fatalities recorded in Italy, leading to a 9% increase in comparison to 2013. In Italy, the mortality index (deaths every 100 accidents) for cyclists is 1.42, which is more than double compared to car users (Istat, 2015).

Various contributing factors to bicycle crashes have been identified in literature. Accident analysis revealed that violation of traffic rules plays a key role in fatal crashes involving cyclists. Red-light violation is one typical violation behaviour among cyclists (Wu et al., 2012; Pai and Jou 2014). Other violations commonly associated with collision were riding against traffic, in a wrong-way, or coming from an unexpected side of the road (Atkinson and Hurst 1983; Ashbaugh et al., 1995; Kim and Li 1996; Wachtel and Lewiston 1996; Wessels 1996; Räsänen et al., 1998; Vandenbulcke et al., 2014; Hamann et al., 2015).

Although fall and collisions with non-motorized users may happen more frequently, collisions involving motor vehicles account for majority of the reported bicyclists' fatalities and serious injuries (Rosenkranz and Sheridan, 2016; Rowe et al., 1995; Nicaj et al., 2009; Chong et al., 2010; Sze et al., 2011). Exposure to traffic

increases the risk of collision (Hagel et al., 2014; Chen 2015). Another possible reason might be blind spot conflicts (Wachtel and Lewiston 1996).

Different approaches have been employed to investigate these factors. One of these approaches is based on investigating factors that increase the severity of bicycle crashes. Usually, collision data are gathered from official sources (Klassen et al., 2014). Using this approach, factors contributing to the severity of bicycle crashes have been studied at various levels: crash characteristics (e.g., type of collision and opponent vehicle), infrastructure characteristics (e.g., type of carriageway, road type, road signage, and type of road segment), cyclists (e.g., gender and age), and environmental factors (e.g., time of the day, darkness, day of the week, and weather). In terms of infrastructure characteristics, crashes in straight sections have been found to be the most fatal (Klop and Khattak 1999; Bíl et al., 2010). Bicycle crashes occurring at signalized intersections were less severe than those elsewhere (Eluru et al., 2008).

Among the crash characteristics, the involvement of trucks and heavy vehicles was found to increase the severity of bicycle crashes compared to other types of vehicles (Kim et al., 2007; Moore et al., 2011; Yan et al., 2011). Moreover, head-on and angle collisions were found to increase the level of bicyclist injury severity (Kim et al., 2007; Bíl et al., 2010; Yan et al., 2011).

In terms of environmental characteristics, the consequences of bicycle crashes tend to be less severe if they occur at day-time under good visibility, whereas crashes occurring in night-time traffic in places without streetlights have the worst consequences for cyclists (Klop and Khattak 1999; Kim et al., 2007; Eluru et al., 2008; Bíl et al., 2010; Yan et al., 2011). Cycling in the night (e.g., from midnight to 6 a.m.) has been found to increase the likelihood of fatal injury (Stone and Broughton 2003; Eluru et al., 2008). Furthermore, inclement weather (Kim et al., 2007) and foggy weather (Klop and Khattak 1999) were associated with more severe consequences of bicycle crashes. Concerning factors related to cyclists, there is evidence that male cyclists are more likely to suffer a bicycle fatal injury than female cyclists and that older cyclists (e.g., above 55 or 65 years old) are the most vulnerable age group (Kim et al., 2007; Eluru et al., 2008; Bíl et al., 2010).

The analysis of the predictors of the severity of bicycle crashes has been conducted using different types of analysis such as the generalized linear model of logistic regression, binary logit model, multinomial logit model, and mixed logit model (Klassen et al., 2014). However, because of the mass of complicated data on road accidents, it is difficult to use regression models to investigate the predictors of the severity of bicycle crashes. Firstly, regression models rely on different and strong statistical assumptions such as no outliers, linearity in modelling the relationship (Harrell 2001; Cohen et al., 2003; Tabachnick and Fidell 2012), which are hardly to be valid for accident data (Chang and Wang 2006; Yan et al., 2010; De Oña et al., 2011). Secondly, interaction may occur in complex forms and its detection using cross-product terms may be a daunting task (Yan et al., 2010). Thirdly, regression models may not satisfactorily handle many discrete variables or variables with a high number of categories (Harrell 2001; Cohen et al., 2003; Tabachnick and Fidell 2012).

Data mining techniques refer to an analytic process aimed at exploring large amounts of data (also known as 'big data' in the popular press) in search of structures, commonalities, hidden patterns (or rules) among data (Hand et al., 2001; Pujari, 2001; Han et al., 2012). Data mining techniques such as CHAID decision tree technique and Bayesian network analysis have the following advantages: (1) no problem with outliers, (2) no assumption on variable distributions is made and a priori probabilistic knowledge about the severity of bicycle crashes is not needed, (3) many discrete variables or variables with a high number of categories are more properly handled compared to regression models, and (4)

it is possible to extract information from large amounts of data (Breiman et al., 1984; Friedman et al., 1997; Sutton 2005; Strobl et al., 2009). CHAID decision tree technique and Bayesian network analysis have been successfully applied to investigate the predictors of head injury for pedestrians and cyclists (Badea-Romero and Lenard 2013), train–vehicle crashes at passive highway-rail grade crossings (Yan et al., 2010), traffic injury severity (Chang and Wang 2006; Mujalli et al., 2016), traffic accident injury severity on rural highways (De Oña et al., 2011; De Oña et al., 2013), and driver injury severity in rear-end crashes (Chen et al., 2015). However, to our knowledge, no research has used both CHAID decision tree technique and Bayesian network analysis in the study of the severity of bicycle crashes.

1.1. Study objectives

The main aim of the present study was to identify factors and rules crucial to the occurrence of fatal bicycle crashes. Crash characteristics (type of collision and opponent vehicle), infrastructure characteristics (type of carriageway, road type, road signage, pavement type, and type of road segment), cyclists (gender and age), and environmental factors (time of the day, day of the week, month, pavement condition, and weather) were considered as predictors of bicycle injury severity.

2. Method

2.1. Road transport in Italy

Based on data from the National Institute of Statistics, the total population of Italy in 2016 is 60,665,552. According to Eurostat, in Italy in 2013, the total length of motorways was 6751 kilometres (based on the last available data from 2013), whereas the total length of other roads was 249,288 kilometres. The total length of state, provincial and communal roads was 19,920, 154,948, and 74,420 kilometres, respectively. The motor vehicles movement on national territory was 51,293 million vehicles-kilometres. The estimated passenger road transport on national territory was 770,590 million of passenger-kilometres.

2.2. Data

The data used in this study were provided by the Italian National Institute of Statistics (ISTAT). The ISTAT gathers data about all road crashes collected by public institution. The data are collected through a broad collaboration among different public institutions: ISTAT, Italian Automobile Club, the Italian Ministry of Transport and Infrastructure, different National Police organizations, and local Municipalities.

In the present study, the dataset contains information about road crashes occurred on the Italian road network during the period ranging from 2011 to 2013. At the time of the study, 2013 was the most recent available ISTAT data. In 2010 (Law L. 29/7/2010 n. 120) a new national traffic law was approved, with minor changes involving also bicycle use. Therefore, to have a trade-off between the need to have a large sample size and the need to control for change in road regulation, we chose a three-year period ranging from 2011 to 2013.

The ISTAT database does not include a distinction between different levels of injuries, thus making a distinction only between road crashes resulting in injuries or fatalities (within 30 days). As shown in Table 1, the database was rearranged and 15 categorical variables were selected: (1) month of the year, (2) day of the week, (3) time of the day (4) cyclist age, (5) cyclist gender, (6) road type, (7) accident location, (8) road pavement type, (9) road pavement condition, (10) type of junction, (11) road signage, (12)