Q1 # Investigating the role of socio-economic factors in comprehension of traffic signs using decision tree algorithm

Q3 Q2 ## Madhar Taamneh

*Department of Civil Engineering, Hijjawi Faculty for Engineering Technology, Yarmouk University, P.O. Box 566, Irbid 21163, Jordan*

## ARTICLE INFO

## ABSTRACT

*Introduction:* Drivers' ability to comprehend the meaning of traffic signs is essential to safe driving. Drivers' personal characteristics are believed to play a crucial role in determining drivers' comprehension of traffic signs. *Method:* This study investigates the role of age, gender, marital status, license category, educational level, driving experience, monthly income, and number of traffic violation during the last five years in drivers' comprehension of 39 posted traffic signs in the city of Irbid, Jordan. These signs include 15 regulatory signs, 17 warning signs, and 7 guidance signs. A total of 400 paper-based surveys were completed by drivers with different socio-economic characteristics. Subsequently, a decision tree was created for each category of traffic signs to identify the most influential factors affecting drivers' comprehension. Each tree was created twice; once using the whole data set for building and validating the tree, and a second time only using 80% of the data for building and 20% for validating. *Results:* The accuracy of the generated trees in predicting drivers' comprehension of regulatory, guidance, and warning traffic signs was 70%, 71%, and 66.5%, respectively, when using the whole data for building and validating the tree, and was 65%, 62.5%, and 61.3%, respectively, when using only 80% of the data for building and the remaining for validating. *Conclusions:* The generated decision trees showed that driving experience, marital status, age, and education background are the most influential factors in determining drivers' comprehension of traffic signs as they were primary splitters in such trees. *Practical application:* The rules obtained from the decision tree can be utilized by transportation agencies to determine the drivers who need help with understanding the road traffic signs.

© 2018 National Safety Council and Elsevier Ltd. All rights reserved.

## 1. Introduction

Road safety is always a major concern for highway and traffic engineers. According to the World Health Organization (WHO), Jordan is ranked 32nd out of 186 countries in terms of number of road fatalities (WHO, 2015). Traffic accident statistics reported by the Jordan Traffic Institute (JTI, 2015) indicated that 102,441 traffic accidents occurred in 2014, resulting in a total of 688 deaths and 14,790 injuries. Traffic signs are essential for drivers' safety on roads. Failure to comprehend the meaning of traffic signs leads to a number of traffic accidents; according to the Jordanian Traffic Institute (JTI, 2015), 2.2% of traffic accidents in Jordan were attributed to failing to comply with regulatory traffic signs. This paper aims to understand to what extent the personal characteristics of drivers influence their understanding of traffic signs in Jordan. Several studies were conducted to evaluate drivers' comprehension of traffic signs (Kirmizioglu & Tuydes-Yaman, 2012; Shinar, Dewar, Summala, & Zakowska, 2003; Shinar & Vogelzang, 2013). These studies aimed at evaluating drivers' ability to comprehend traffic signs to decide if the signs need to be improved. It was found that a considerable number of traffic signs were not recognized by drivers. Also, adding text to the sign was found to improve the comprehension. In these studies, however, no connection was made between the personal characteristics of drivers and their comprehension of traffic signs. Other researchers studied the influence of the personal characteristics of drivers on the severity of traffic accidents (Abdel-Aty, Chen, & Schott, 1998; Boufous & Williamson, 2009; Hammoudi, Karani, & Littlewood, 2014; Issa, 2016; Massie, Campbell, & Williams, 1995; McGwin Jr. & Brown, 1999; Oltedal & Rundmo, 2006). They found that age and gender are the most influential factors in determining the severity of traffic accidents. Another set of studies explored the impact of the personal characteristics of drivers on their comprehension of traffic signs. Age, gender, marital status, educational background, accident rate, monthly income, driving experience, and license type were the most commonly studied personal characteristics. In a study conducted by Al-Madani (2000), the relationship between drivers' understanding of posted traffic signs and their safety-related characteristics was investigated. The safety-related characteristics studied were: driving experience, accident involvement, experience per accident, citations received in the last three years on speed limit violations, and seat belt usage. The results showed that years of driving experience has a positive correlation with traffic sign comprehension. Seat belt usage also showed positive correlation with understanding of posted signs. Al-Madani and Al-Janahi (2002a, 2002b) investigated the relationship between drivers' personal characteristics and their comprehension of 28 posted traffic

*E-mail address:* mtaamneh@yu.edu.jo.

signs. Questionnaires were distributed and collected from five countries including: Bahrain, Kuwait, Oman, Qatar, and United Arab Emirates. The results showed that only 56% of the posted signs were known to drivers. Educational level, monthly income, and nationality characteristics were found to be crucial in traffic sign comprehension. Western drivers were found to comprehend traffic signs better than other nationalities. The results also indicated that male drivers comprehended traffic signs better than female drivers. Furthermore, the results revealed that age, marital status, experience, and accident rates had no effect on drivers' comprehension of signs.

Ng and Chan (2008) also investigated the impact of driver's characteristics on the comprehensibility of traffic signs. The results showed that driving experience and education background played a major role in traffic sign comprehensibility. On the other hand, driver age had no effect on traffic sign comprehension. Furthermore, a clear correlation was found between traffic sign comprehensibility and familiarity. Makinde and Oluwasegunfunmi (2014) investigated the relationship between the understanding of traffic control devices and drivers' personal characteristics such as age, gender, educational level, and marital status. The results showed that education level has an impact on understanding traffic control devices, while age and gender have no effect.

All aforementioned studies use ordinary statistical analysis methods to find the relationship between the personal characteristics of drivers and their comprehension of traffic signs. Such methods face significant challenges in analyzing large, highly dimensional data sets. As a result, many researchers recently started using data mining techniques to uncover hidden patterns in large, highly-dimensional data sets. Data mining can help determine the interactions between variables that would be very difficult to establish directly using ordinary statistical modeling techniques (Baluni & Raiwani, 2014; Shankar et al., 1996).

Although there is an abundance of studies using data mining techniques to predict the injury severity of traffic accidents (Abellan, López, & De Oña, 2013; Alkheder, Taamneh, & Taamneh, 2017; Chang & Chien, 2013; Chang & Wang, 2006; De Oña, López, & Abellán, 2013; Delen, Sharda, & Bessonov, 2006; Kashani & Mohaymany, 2011; Kashani, Rabieyan, & Besharati, 2014; Kwon, Rhee, & Yoon, 2015; Taamneh, Alkheder, & Taamneh, 2017; Taamneh, Taamneh, & Alkheder, 2016), there is, to the best of our knowledge, only one study using such techniques to investigate the relationship between the personal characteristics of drivers and their comprehension of traffic signs (Rahimi, Kazemi, & Pakgohar, 2012). In that study, a well-known data mining method, called Classification And Regression Tree (CART) was used to investigate the impact of five driver characteristics on their attention to pedestrian crossing signs. These driver characteristics are: age, gender, job, marital status, and educational level. The results showed that the job and education level were more important in determining how drivers pay attention to pedestrian crossing signs than the other three factors.

In this study, a decision tree algorithm called Iterative Dichotomiser 3 (ID3) was employed using a suite of data mining software developed by Frank et al. (2016) to highlight the main contributing personal characteristics factors that affect the comprehension of traffic signs. The reasons for choosing this method are that it is able to handle both numerical and categorical data, simple to understand and interpret, and efficient for large data sets. In this paper, eight personal characteristics were investigated. These characteristics are believed to be the most influential factors in determining the drivers' comprehension of traffic signs (Al-Madani, 2000; Al-Madani & Al-Janahi, 2002a, 2002b; Ma, Shao, Song, & Chen, 2014; Rahimi et al., 2012).

## 2. Research methodology

In this paper, an effort was made to understand the influence of gender, age, marital status, educational background, license category, driving experience, monthly income, and number of traffic violations on drivers' comprehension of three categories of traffic signs: regulatory, guidance, and warning signs. The participant's comprehension of traffic signs is quantified based on his/her answers to questions designed specifically for that purpose. The 50th percentile is then used to split the drivers into two categories, Good (more knowledgeable) drivers and Poor (less knowledgeable) drivers. The 50th percentile was used to create a balanced data set where the classes are represented equally (i.e., a data set with roughly equal numbers of good and poor drivers). The problem with the imbalanced datasets is that classifiers have good accuracy on the majority class but very poor accuracy on the minority class.

### 2.1. Survey administration and participants

Jordan is divided into 12 governorates. In this study, three governorates with different socio-economic characteristics were selected. These governorates are Amman (capital of Jordan), Irbid, and Ma'an. A total of 400 respondents were recruited by convenience sampling technique from shopping centers, universities, bus stations, workplaces and gas stations. The survey was carefully designed to measure drivers' comprehension of 15 regulatory signs, 17 warning signs, and 7 guidance signs in the city of Irbid, Jordan. These signs are presented in Table 1. In this study, drivers 18 years and older with a valid drivers license were targeted. Participants agreed to participate in this study voluntarily with no incentives given, and were assured that their personal information and responses will remain anonymous. Table 2 shows the distribution of participants based on their characteristics.

### 2.2. Survey design

The paper-based questionnaire consists of two parts. In the first part, descriptive information such as age, gender, marital status, license category, educational background, driving experience, monthly income, and the number of traffic violations during the last five years are gathered. The second part consists of 39 questions that are used to measure drivers' comprehension of 39 traffic signs. The driver's comprehension of a specific sign is evaluated by answering the following question: "what is the meaning of this traffic sign." The correct answer is coded by one and the wrong answer by zero.

### 2.3. Decision tree algorithm

Decision tree learning is a widely used technique in data mining. It aims to build classification models that predict the value of a target attribute based on the input attributes. The decision tree constructs classification models in the form of trees. Each interior node in these trees represents one of the input variables, and it has a number of branches equal to the number of possible values of that input variable. Each leaf node holds a value of the target attribute. The leaf node represents the decision made based on the values of the input variables from the root to the leaf. In the decision trees, primary splitters (i.e., input variables that appear in the upper levels of the tree) that are directly connected with leaf nodes are considered the most influential factors in determining the value of the target attribute. In this work, the Iterative Dichotomiser 3(ID3) algorithm is applied on gathered data in order to identify the factors affecting drivers' comprehension of regulatory, warning, and guidance traffic signs. ID3 is a simple decision tree learning algorithm that constructs the decision tree by employing a top-down, greedy search through the given sets to teach each attribute at every tree node.

For each category of traffic signs, two decision trees (i.e., classifier models) were built. The first tree was built and validated using the whole data set. The second tree was built using percentage split of 80%, where 80% of data are used for building the model, and the remaining 20% are used for validating. The reason for building the second model is that using the whole dataset for building the classifier may cause over-fitting of the model. Over-fitting happens when the model works fine for the training sets and fail to predict future observation reliably. In order to check if using the whole dataset for both building and testing the model can affect the ability of the generated classifier in