Accident Analysis and Prevention 51 (2013) 1–10

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks

Juan de Oña*, Griselda López, Randa Mujalli, Francisco J. Calvo

TRYSE Research Group, Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/Severo Ochoa s/n, 18071 Granada, Spain

ARTICLE INFO

Article history:
Received 13 April 2012
Received in revised form
11 September 2012
Accepted 26 October 2012

Keywords:
Cluster analysis
Latent Class Clustering
Bayesian Networks
Traffic accidents
Classification
Injury severity
Highways
Road safety

ABSTRACT

One of the principal objectives of traffic accident analyses is to identify key factors that affect the severity of an accident. However, with the presence of heterogeneity in the raw data used, the analysis of traffic accidents becomes difficult. In this paper, Latent Class Cluster (LCC) is used as a preliminary tool for segmentation of 3229 accidents on rural highways in Granada (Spain) between 2005 and 2008. Next, Bayesian Networks (BNs) are used to identify the main factors involved in accident severity for both, the entire database (EDB) and the clusters previously obtained by LCC. The results of these cluster-based analyses are compared with the results of a full-data analysis. The results show that the combined use of both techniques is very interesting as it reveals further information that would not have been obtained without prior segmentation of the data. BN inference is used to obtain the variables that best identify accidents with killed or seriously injured. Accident type and sight distance have been identify in all the cases analysed; other variables such as time, occupant involved or age are identified in EDB and only in one cluster; whereas variables vehicles involved, number of injuries, atmospheric factors, pavement markings and pavement width are identified only in one cluster.

1. Introduction

Traffic accidents are contingent events and analysing them requires awareness of the particularities that define them. In general, accidents are defined by a series of variables – generally discrete variables – that explain them. Once the nature of the variables is known, researchers select the method that is most appropriate for developing and implementing the best statistical models for analysing the data in each case (Lord and Mannering, 2010; Savolainen et al., 2011; Mujalli and De Oña, in press).

One of the main problems of accident data and their modelling process is their heterogeneity (Savolainen et al., 2011). If this is not taken into account during the analysis, certain relationships between the data may not be detected. Researchers often try to reduce heterogeneity by segmenting traffic accident data on the basis of expert domain knowledge, methodological decisions or the intention to study a specific problem. Although expert knowledge can lead to a workable segmentation, it does not guarantee that each segment consists of a homogenous group of traffic accidents (Depaire et al., 2008). That is why specific analysis techniques, such as cluster analysis (CA), are used as aids in traffic accident segmentation.

CA has been used in road safety analysis as a preliminary tool for attaining several aims. Karlaftis and Tarko (1998) used it to classify 92 areas of the state of Indiana into urban, sub-urban and rural areas. They applied Negative Binomial (NB) regression models to the results in order to analyse the influence of driver age on accidents. The results obtained with a model that used all the data and models based on clustered data showed statistically significant differences. Subsequently, Sohn (1999) used a Poisson regression model for previously clustered data (based on the latitude and longitude of each crash) to analyse accident frequency. Using CA, GIS (Geographic Information Systems) and NB models, Ng et al. (2002) developed an algorithm for estimating the number of accidents and evaluating their risk in a specific area. In a later study, Wong et al. (2004) proposed a method for evaluating the effect of a series of road safety strategies implemented in Hong Kong. They used CA as a preliminary step for grouping different road safety programmes and projects into smaller groups with significant road safety strategies. Ma and Kockelman (2006) used CA and a probit model to analyse the relationship between crash frequency and severity, road design, and the characteristics of use in the state of Washington.

Depaire et al. (2008) used Latent Class Cluster (LCC) and Multinomial Logit (MNL) models to study the severity of traffic accidents. In their study, they identified seven clusters that represent different types of traffic accidents. Subsequently, they applied an MNL model to the full set of data and on each of seven identified clusters. Their

* Corresponding author. Tel.: +34 958 24 99 79.
E-mail address: jdona@ugr.es (J. de Oña).

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2012.10.016

results showed that the clustered data provided information that would not have been obtained if only the full database had been used. Recently, LCC have also been used by Park and Lord (2009) and Park et al. (2010) in order to segment a database and analyses vehicle crash data. Finally, Pardillo-Mayora et al. (2010) used CA to analyse data from run off road accidents to calibrate a roadside hazardous index for two-lane roads in Spain. The four characteristics considered for the index were: roadside slope, non-traversable obstacles, safety barrier installation, and alignment. They used CA to group the 120 combinations of the four indicators into categories with homogeneous effects on severity.

Many previous studies have focused on compressing and identifying key factors that have an impact on the severity of the consequences of road accidents. Many different methodological approaches have been used to analyse severity (Savolainen et al., 2011): probit models (Bayesian ordered, binary, bivariate binary, bivariate ordered, heteroskedastic ordered, multivariate, ordered, random parameters ordered), logit models (Bayesian hierarchical binomial, binary, generalized ordered, heteroskedastic ordered, Markov switching multinomial, mixed generalized ordered, mixed joint binary, multinomial, nested, ordered, random parameters, random parameters ordered, sequential binary, sequential, simultaneous binary), log-linear model, partial proportional odds model, artificial neural networks, and classification and regression trees. Recently, Bayesian Networks (BNs) are being used to analyse traffic accident severity, with satisfactory results (Simoncic, 2004; De Oña et al., 2011; Mujalli and De Oña, 2011).

This paper presents an analysis of traffic accidents based on a combination of cluster analysis and Bayesian Networks. To the best of our knowledge, this is the first time that both approaches have been used together. The paper is structured as follows: Section 2 shows the methodology used to conduct the analysis, with a description of the Latent Class Clustering analysis and Bayesian Network techniques. Next, key characteristics of the data analysed are described. Section 4 shows the results and discussion, followed by the conclusions.

2. Methodology

2.1. Latent Class Clustering analysis

CA is an unsupervised learning technique within the field of Data Mining, where its principal objective is to group a finite subset of elements in a number of groups or clusters. CA is based on heuristics that try to maximize the similarity between in-cluster elements and the dissimilarity between inter-cluster elements (Fraley and Raftery, 2002). The similarity-based techniques include two main approaches: the hierarchical approach (e.g. Ward's method, a single linkage method) and the partitioning approach (e.g. K-means). Both approaches have been used in road safety (Sohn, 1999; Karlaftis and Tarko, 1998; Ng et al., 2002; Wong et al., 2004; Pardillo-Mayora et al., 2010), although the statistical properties of these methods are relatively unknown (Fraley and Raftery, 2002).

Another type of CA is Latent Class Clustering (LCC) (Moustaki and Papageorgiou, 2005; Vermunt and Magidson, 2002). In this type, the statistical properties of probability model-based clustering techniques are better understood (Bock, 1996; Fraley and Raftery, 2002). Although when using any kind of cluster analysis method it is inevitable to introduce some kind of subjective judgment, LCC have some important advantages over other types of cluster analysis methods (Hair et al., 1998; Magidson and Vermunt, 2002; Vermunt and Magidson, 2005), such as:

- Being able to use different types of variables (frequencies, categorical, metric variables or a combination of them), with no need for prior standardization that could have a bearing on the results.
- The method provides several statistical criteria that help to decide the most appropriate number of clusters.
- LCC allow probability classifications to be made by using subsequent membership probabilities estimated with maximum likelihood method.

Given a data sample of N cases (or accidents), measured with a set of observed variables, $Y_1, \ldots, Y_j$ which are considered indicators of a latent variable X; and where these variables form a Latent Class Model (LCM) with T classes. If each observed value contains a specific number of categories: Y_i contains I_i categories, with $i = 1, \ldots, j$; then the manifest variables make a multiple contingency table with $\prod_{i=1}^{j} I_i$ response patterns. If π denotes probability, $\pi(X_t)$ represents the probability that a randomly selected case belongs to the latent t class, with $t = 1, 2, \ldots, T$.

The regular expression of LCMs is given by:

$$\pi_{Y_i} = \sum_{t=1}^{T} \pi_{X_t} \pi_{Y_i|X_t} \quad (1)$$

With Y_i response-pattern vector of case i; $\pi(X_t)$ is the prior probability of membership in cluster t; $\pi_{Y_i|X_t}$ is the conditional probability that a randomly selected case has a response pattern $Y_i = (y_1, \ldots, y_j)$, given its membership in the t class of latent variable X. Local independence is the underlying assumption that needs to be verified, and therefore Eq. (1) is re-written:

$$\pi_{Y_i} = \sum_{t=1}^{T} \pi_{X_t} \prod_{i=1}^{j} \pi_{Y_{ij}|X(t)} \quad \text{with} \quad \sum_{i=1}^{j} \pi_{Y_{ij}|X(t)} = 1 \quad \text{and} \quad \sum_{t=1}^{T} \pi_{X_t} = 1 \quad (2)$$

For a detailed explanation of LCC analysis see Sepúlveda (2004).

The estimation of the model is based on the nature of the manifest variables, since it is assumed that the conditional probabilities may follow different formal functions (Vermunt and Magidson, 2005). The method of maximum likelihood is the most widely used method for estimating the model's parameters. Once the model has been estimated, the cases are classified into different classes by using the Bayes rule to calculate the a posteriori probability that each n subject comes from the t class (âre the model's estimated values):

$$\pi_{X_t|Y_i} = \frac{\pi_{X_t} \pi_{Y_i|X_t}}{\pi_{Y_i}} \quad (3)$$

In practice, the set of probabilities is calculated for each response pattern and the case is assigned to the latent case in which the probability is the highest. Thus, a specific accident may belong to different latent cases with a specific percentage of membership (with 100% being the sum total of membership probabilities).

2.2. Number of clusters selection

Given that the number of clusters is unknown at the start, the aim is to find the model that can explain or adapt the best to the data being used. In this paper we have used several information criterions for discovering the model that provides the most information on reality. The criterions are: Bayesian Information Criterion (BIC) (Raftery, 1986), Akaike Information Criterion (AIC) (Akaike, 1987) and Consistent Akaike Information Criterion (CAIC) (Fraley and Raftery, 1998).

In clustering contexts, the BIC criterion has shown better performance than other criteria (Biernacki and Govaert, 1999). In general, the lower the value of the indicators, the better the model is, because it is more parsimonious and adapts better to the data. Nonetheless, when analysing large samples, the BIC and other information criteria often do not reach a minimum value with increasing number of clusters (Bijmolt et al., 2004). In that case, the percentage of reduction in BIC between competing models must be analysed,