Accident Analysis and Prevention 60 (2013) 402–411

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Multilevel analysis in road safety research

Emmanuelle Dupont[a], Eleonora Papadimitriou[b,*], Heike Martensen[a], George Yannis[b]

[a] IBSR, Belgian Road Safety Institute, Belgium
[b] National Technical University of Athens, Greece

ARTICLE INFO

Article history:
Received 7 July 2011
Received in revised form 1 March 2013
Accepted 29 April 2013

Keywords:
Road safety
Multilevel models
Hierarchical structures
Geographical dependences
Road accident process dependences

ABSTRACT

Hierarchical structures in road safety data are receiving increasing attention in the literature and multilevel (ML) models are proposed for appropriately handling the resulting dependences among the observations. However, so far no empirical synthesis exists of the actual added value of ML modelling techniques as compared to other modelling approaches. This paper summarizes the statistical and conceptual background and motivations for multilevel analyses in road safety research. It then provides a review of several ML analyses applied to aggregate and disaggregate (accident) data. In each case, the relevance of ML modelling techniques is assessed by examining whether ML model formulations (i) allow improving the fit of the model to the data, (ii) allow identifying and explaining random variation at specific levels of the hierarchy considered, and (iii) yield different (more correct) conclusions than single-level model formulations with respect to the significance of the parameter estimates. The evidence reviewed offers different conclusions depending on whether the analysis concerns aggregate data or disaggregate data. In the first case, the application of ML analysis techniques appears straightforward and relevant. The studies based on disaggregate accident data, on the other hand, offer mixed findings: computational problems can be encountered, and ML applications are not systematically necessary. The general recommendation concerning disaggregate accident data is to proceed to a preliminary investigation of the necessity of ML analyses and of the additional information to be expected from their application.

1. Introduction

Most of the data of interest for road safety research happen to be hierarchically organized, i.e., to belong to structures with several hierarchically ordered levels. This implies that the observations can be unambiguously attributed to one and only one unit at higher level(s).[1] For a part, these hierarchical structures result from the *spatial (and temporal) spread* of the data: Observations belong to larger geographical areas or units (road sites, segments, or intersections, counties, regions, etc.), or are made on a recurrent basis over a given time period. For another part, this hierarchical organization of observations results from the very nature of *accidents*, as each road-user, driver, or vehicle observation "pertains" to one and only one accident.

One of the main problems associated with hierarchical data organization is the dependence that it generates among the observations (Hox, 2002). Observations that are sampled from the same geographical units have in common a series of unobserved characteristics that are proper to these larger geographical areas (Langford et al., 1999). One can think of risk studies that are based on crash-frequency data aggregated over a sample of road intersections or segments, which may themselves exhibit different road geometrics, traffic, or other unobserved environmental characteristics that are all likely to affect accident frequency. In a similar vein, observations that are made at time points that are close from each other will also tend to be more similar than observations that are made at two remote time points. One can doubt of the possibility to *exhaustively* account for these heterogeneities by measuring and including them as covariates in a model. One can also doubt that all of these heterogeneities will be measurable at all (Huang and Abdel-Aty, 2010). Similarly, observations made on individuals occupying the same vehicles and involved in the same accident are likely to resemble each other more than observations made on individuals involved in different vehicles or accidents. This is so because these observations will be commonly influenced by vehicle and accident characteristics that are often left unobserved in a given analysis.

The estimations obtained from most standard analysis techniques rest on the assumption that the observations are sampled from a single homogeneous population, and that the residuals are independent. However, the hierarchical organization of data fundamentally challenges these assumptions. Hence, applying traditional

* Corresponding author. Tel.: +30 2107721380.
E-mail address: nopapadi@central.ntua.gr (E. Papadimitriou).

[1] There are also cases where observations can simultaneously be attributed to different higher-level units. These cases are discussed later on in this article (Section 5.1).

0001-4575/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2013.04.035

statistical techniques (linear or generalized linear models) to hierarchically organized data is likely to result in underestimated standard errors and exaggeratedly narrow confidence intervals (Kreft and De Leeuw, 1999). The risk is consequently that incorrect conclusions be derived about the significance of the parameters whose effects are investigated.

Statistical models have been developed that allow accounting for hierarchical data structures, and taking into account the dependence they introduce among the data. Because the hierarchical structure is specified in the model, predictors that characterize the different levels considered can also be correctly defined (no need for aggregation or disaggregation). These models are labelled multilevel models, hierarchical models, mixed-effect models, random coefficients or random parameter models. In the remainder of this article the terms multilevel (ML) or hierarchical (HL) models will be used indifferently.

Although there are good statistical and conceptual arguments for the application of ML models in road safety research, so far no review based on road safety analyses has been conducted to assess the actual added value they can offer compared to "traditional" modelling techniques in this field of research. This article starts with a description of the hierarchical structures most commonly encountered in road safety studies. HL models are then defined and their statistical and conceptual interest is discussed. The second part of this article provides a review of several ML analyses conducted on the basis of three types of road safety data: (1) aggregated accident data, (2) disaggregated accident data, and (3) behavioural indicators. In each case, the review focuses on the questions of knowing whether ML model formulations (i) allow identifying significant random variation of the observations at the various levels of the hierarchy considered, (ii) allow improving the fit of the model to the data, and (iii) yield different conclusions than single-level model formulations with respect to the significance of the estimates of the effects of explanatory variables. The necessity and feasibility of applying ML models is finally discussed distinguishing the three types of data.

2. Prevailing hierarchies in road safety research: spatial distributions of data and the nature of the accident process

One can distinguish two prevailing hierarchies in road safety data, namely: geographical and accident hierarchies.

As illustrated in Fig. 1, road safety data are organized in geographical units that are nested into each other (for example: road-sites nested into counties that are themselves nested into regions and countries). Similarly, the observations made on individual road users involved in accidents are nested into vehicles, which are themselves nested into different accidents.

The two hierarchies are actually complementary and have been incorporated into a single framework to represent prevailing data structures in road safety (Huang and Abdel-Aty, 2010). An adapted version of this general hierarchical framework is presented in Fig. 2.

Because road sites can be considered to belong to both types of hierarchies, they constitute the link between geographical and accident hierarchies, the macro- and microscopic ML structures.

Repeated measurements in particular can be included as a horizontal 'time' dimension in this framework (Huang and Abdel-Aty, 2010; Aguero-Valverde and Jovanis, 2006). The multilevel structure can also be a multiple membership structure, as indicated by the double arrow inside the pyramid, or a cross-classification structure, as indicated by the crossed arrows inside the pyramid. These complex structures are detailed in Section 5.1.

Depending on the research question, driver characteristics can be associated to the "vehicle" (e.g., all information about driver behaviour or manoeuvres) or to the "road users" level (e.g., the characteristics that are likely to affect the severity of accident outcomes such as age or gender).

The "measurements/responses" level has been included in Fig. 2 to specify the capacity of multilevel models to handle complex types of response variables as being nested within individuals (i.e., multivariate responses, e.g. Duncan et al. (1999) multinomial responses, or repeated measurements).

Intuitively, geographical hierarchies call for macroscopic analysis, while accident hierarchies, with individual road users or drivers as unit of analysis are the ideal basis for microscopic analysis (e.g., "What are the accident, vehicle, or driver characteristics that help predicting the occurrence of accidents and/or their outcomes?").

3. "Hierarchical/multilevel models" – definition and general model formulation

ML/HL models are regressions (linear or generalized linear models) in which the parameters (intercept and/or estimates of covariates effects) are assigned a probability model. As a consequence, this "higher-level (probability) model has parameters of its own (mean, variance). These are termed the "hyperparameters" of the model–which are also estimated from the data" (Huang and Abdel-Aty, 2010: p. 1560).

In this sense, hierarchical models are grounded in the Bayesian paradigm: The model parameters are assigned a probability distribution that summarizes the knowledge the researcher has about each parameter, prior to any data observation. These "prior distributions" may be either informative (when, for example, existing knowledge allows reasonable assumptions to be made about the mean value of the parameter and its variance), or vague. In the latter case, "typical" distributions with relatively large variances are assigned to the parameters, so as to account for the lack of knowledge prior to observation. In the Bayesian approach, inference about the parameters is based on the posterior distribution, which combines the prior information (defined by the prior distribution) with information derived from the observations. Carriquiry and Pavlovich (2004), as well as Miaou and Lord (2003) provide a thorough discussion of hierarchical model formulation in relation to the distinction between Empirical and Full Bayes estimation.

Following Lord and Mannering (2010), it is important to distinguish between models allowing random variation of the parameters and "truly" hierarchical models. In the first case, the intercept and covariate parameters are allowed to vary *across the observations*, and are thus assigned a probability distribution. HL models, on the other hand, specify the observations units (the lowest level of observation, for example, crash counts aggregated at various road intersections) as being clustered into higher-level units (for example, the "corridors" to which the various road segments belong to). In the latter case, the higher-level units are themselves considered *a sample from a larger population* (a sample from the "corridor population"). In such cases, the hyperparameters of the model define the random variation of the model's parameters *across the units at the higher level(s)* (the corridors). The total variation in the observations can consequently be partitioned, or structured, along the different levels included in the model.

As we will see, although the first type of model takes account of the unobserved extra variations, it does not account for the hierarchical structure in itself, and does not offer any information about the proportion of variation in the observation that is