Accident Analysis and Prevention 62 (2014) 102–109

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Internal validation of near-crashes in naturalistic driving studies: A continuous and multivariate approach

Jenny K. Jonasson, Holger Rootzén*

Department of Mathematical Statistics, Chalmers and Gothenburg University, SE-412 96 Gothenburg, Sweden

ARTICLE INFO

Article history:
Received 23 April 2013
Received in revised form 5 September 2013
Accepted 17 September 2013

Keywords:
Traffic safety
Rear-ending crash
Crash surrogate
Selection bias
Naturalistic driving study
Extreme value statistics

ABSTRACT

Large naturalistic driving studies give extremely detailed insight into how traffic accidents happen and what causes them. However, even in very large studies there are only relatively few crashes. Hence one additionally selects and studies crash surrogates, so called "near-crashes", i.e. situations when a crash almost happened. The selection procedures invariably entail severe risks of causing bias. In this paper we use extreme value statistics to develop two methods to study the extent and form of this bias. The methods are applied to a large naturalistic driving study, the 100-car study. Both methods identified a severe discrepancy between the rear-striking near-crashes and the rear-striking crashes. Perhaps surprisingly, one conclusion is that, for rear-striking and in this study, the crashes have little relevance for increasing traffic safety. We believe substantial efforts should be made to develop statistical methods for using near-crashes and crashes in future large naturalistic driving studies (such as the SHRP2 study).

1. Introduction

Quite considerable efforts and resources have already been spent on large naturalistic driving studies, and even much larger studies, such as the SHRP2 study (Strategic Highway Research Program 2 (2012)), are underway. These studies provide extremely detailed records of the crashes which occurred during the study, and unique insight into how accidents occur. Still, there are only relatively few crashes even in very large studies, and crash surrogates, "near-crashes", are used to augment the statistical basis for drawing conclusions about driver behavior and methods to decrease accident rates. The near-crashes are chosen to resemble the chains of events which lead up to real crashes as much as possible, and the assumption is that behavior and situations which cause near-crashes is similar to behavior and situations which cause traffic accidents. The definitions and selection of near-crashes vary between studies and are necessarily to some extent subjective.

As examples of research based on use of near-crashes, Dingus et al. (2006) investigated safety and fatigue in long-haul trucking; Guo and Fang (2012) studied how risk varies between individuals; Lee et al. (2010) assessed novice teenage crash experience; and McLaughlin et al. (2008) evaluated collision avoidance systems.

Near-crashes are selected in two steps (Wu and Jovanis, 2012, Section 3): first kinematic triggers are use for automated identification of interesting candidate events, then researchers review the recordings in a time window around the events, and, using carefully specified criteria, classify them as crashes, near-crashes, and others. Typically only a small percentage of the candidate events are classified as near-crashes, and even fewer are crashes. This is due to a large percentage of the candidate events being only kinematic (e.g. high deceleration) without any safety implications. Additionally, during the analysis phase some further near-crashes have to be excluded, e.g. because of absence of radar signals.

This procedure potentially can lead to a severe selection bias. As one example, in the Virginia Tech Transportation Institute 100-car study 34% of the crashes involved no reaction from the driver (Guo et al., 2010, Table 1). It seems likely that similar events often would not be caught by the kinematic triggers, and hence be under-represented among the near-crashes – and in fact there was no reaction from the driver in only 5% of the near-crashes. As another example, for rear-end striking the odds ratio for crash with max speed less than 25 km/h was 48. Thus, it appears to be 48 times more dangerous to drive slower than 25 km/h than at higher speeds. Is this due to selection bias? This is discussed in more detail in Sections 3.2 and 4.

The goal of this paper is to develop methods to understand the extent of this selection bias – and ultimately for drawing the correct conclusions from naturalistic driving studies. Our point of view is that a crash is an extreme event and that the most interesting factors in a crash are those which deviate from their values in normal driving – i.e. again those which are extreme. Hence we use extreme value statistics to attain this goal.

* Corresponding author. Tel.: +46 730794222.
E-mail addresses: jenny.jonasson@astrazeneca.com (J.K. Jonasson), hrootzen@chalmers.se (H. Rootzén).

0001-4575/$ – see front matter
http://dx.doi.org/10.1016/j.aap.2013.09.013

Table 1
Estimated dependence parameter α in fit of bivariate logistic GEV distribution to Max$\{-\text{TTC}\}$ and the indicated variable.

Variable	α
Max speed	1.00
Min distance left lane marking	1.00
Max time eyes off road in 3 s interval	1.00
Max variance of longitudinal acceleration	1.00
Min distance right lane marking	0.93
Max time eyes off road in 2 s interval	–
Longest glance of road last 15 s	–
Max variance of lateral acceleration	–
Max absolute value of yaw angle	–
Length of overlapping glance off road	*

–, A non-acceptable fit; *, the variable is not a maximum.

The methods require that an appropriate continuous crash proximity measure, such as Time To Collision (TTC), Time to Accident (TA), Time to Lane Crossing (TLC), or Post Encroachment Time (PET), . . ., can be computed for the near-crashes. Our aim is methods which (1) avoid the arbitrary discretization of continuous variables which is required for the commonly used odds ratio calculation and logistic regression methods, (2) makes possible quantitative and validated extrapolation from near-crash to crash frequencies and from behavior in lower risk events to behavior in higher risk ones, in a way which is not provided by logistic regression, (3) give new possibilities for understanding the sometimes complex and multidimensional chain of events which lead up to an accident, and (4) can make more efficient use of data.

The methods are tested on rear-striking crashes and near-crashes in the 100-car study (Wu and Jovanis, 2012; Dingus et al., 2005). Due to the limited size of the 100-car data set we here only make univariate and bivariate analyzes. An exiting future prospect is to use higher-dimensional methods to analyze data from the SHRP2 study, and from other future large studies.

Selection bias can be expected to be quite different for different types of crashes, and use of different crash proximity measures will also affect analysis (Hydén, 1987; van der Horst, 1990; Wu and Jovanis, 2012; Jovanis et al., 2011; Guo et al., 2010). Hence omnibus answers to the question "are near-crashes representative for crashes?" may be less useful. Instead careful separate analyzes for different types of situations are needed.

A useful distinction is between *internal validation*, i.e. to attempt to answer the question "are the near-crashes representative of the crashes in this driving study", for different types of crashes, and *external validation* which studies the question "are the crashes and/or near-crashes in this study representative of real crashes?". The latter question involves yet another round of risks of selection bias: the selection of the drivers in a study may be deliberately biased to include more risky drivers, drivers who agree to participate in a study may be different than those who do not want to participate, the population of cars in a study often is different than the general population of cars, etc., and accident data bases may also be subject to severe selection biases. Here we study internal validation. However, it is possible to adapt our methods also to external validation.

The literature on internal validation of crash surrogates in naturalistic driving studies is relatively recent. Wu and Jovanis (2012) give an authoritative review of the use of crash surrogates, with an emphasis on the crash to surrogate ratio, and make a logistic regression analysis of road departure events in the 100-car study. Jovanis et al. (2011) pinpointed a risk of substantial bias if environmental covariates are not included in the analysis. Guo et al. (2010) showed that using only crashes in the 100-car study led to higher odds ratios and much wider confidence intervals than if both near-crashes and crashes were used, and that the crash-to-near-crash ratio was highly scenario dependent.

Extreme value statistics in traffic research was introduced in a seminal paper by Campbell et al. (1996). Using short time fixed video recording of intersection traffic, Sogchitruksa and Tarko (2006) fitted the generalized extreme value distribution to Post Encroachment Times and were able to predict observed 3-year crash rates reasonably well. Tarko (2012) modeled extreme values of lane keeping measures in a driving simulator experiment. For an application of extreme value methods in a related area, aviation safety, see Panagiotakopoulos et al. (2009). Barnes et al. (2011) used so-called seemingly unrelated regression and extreme value techniques for external validation of road departure frequencies in a Michigan Field Operation Test. Results included that one of the surrogates, lateral deviation (LDEV), gave risk estimates which deviate from observed risks in real traffic, while for two others, Lane departure warning (LDW) and time to road edge crossing (TTEC), relative risks tended to agree with observed ones. Extreme value analysis of TTEC gave estimated crash frequencies which the authors deemed reasonable, but not in any way definitive.

Earlier influential studies, in particular Hydén (1987) and van der Horst (1990) used observation and recording of traffic at fixed locations, often intersections. Conclusions made in these studies include that TTC and TA are useful crash proximity measures, while TTCA (which is the same as TTC, except that accelerations instead of speed are assumed constant), may be less useful; that only events with minimum TTC smaller than a low limit (in particular the limits 1.5 s and 1 s were discussed) are useful as crash surrogates; that the crash proximity measure alone is not enough to predict crash severity but that conflict speed also is important; and that accident databases based on police reports are quite incomplete and hence also may include substantial selection biases.

Now a brief overview of the paper. The methods are introduced in Section 2. Section 3.1 gives a description of the data which is used for analysis, and the results of the analysis of the 100-car rear-striking data are presented in Sections 3.2 and 3.3. The results are discussed in Section 4. This section also contains a wider discussion of issues related to internal and external validation of near-crashes, and some perspectives for future research. Our conclusions are summarized in Section 5.

2. Methods

The aim is to develop and test general methods which aid internal validation and use of near crashes in future large naturalistic driving studies such as SHRP 2. The methods are based on statistical extreme value theory, and for completeness the first subsection gives a brief background on it. The next subsection considers validation through prediction of crash numbers. This is similar to the method used by Barnes et al. (2011) for external validation. The methods use the occurrence or not of a crash as the basis for validation, and severity of crashes are not modeled in this paper. However, severity modeling is an important future challenge, see Section 4. In the final subsection, multivariate extreme value statistics is introduced as a tool to obtain more detailed understanding of how near-crashes resemble real crashes and in which respects near-crashes and crashes differ.

2.1. Background and notation: extreme value statistics

Coles (2001) gives an accessible account of models; estimation methods; and model checking tools from extreme value statistics, and provides examples from hydrology, metrology, oceanography, materials science, finance, and sports. Gilleland et al. (2013) contains an up-to-date review of publicly available software.