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 © 2013 Elsevier Ltd. All rights reserved.
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{ − 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.

Extreme value statistics has two main methods, the Peaks-over-Thresholds method, where all values above some high level are used, and the Block Maxima method in which maxima over blocks of time (or space) are considered. Here we only use the Block Maxima method, and the blocks of time are the annotated intervals which contain the near-crashes.

In the one-dimensional Block Maxima method the observed maxima in different blocks of time, e.g. the longest off-road glances in the different near-crashes, are assumed to be statistically independent of one another and to have the generalized extreme value (GEV) cumulative distribution function (cdf) G which has the following form,

$$G(x) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right\}, \quad \text{for } x \text{ such that } 1+\xi\frac{x-\mu}{\sigma} > 0. \tag{1}$$

Here $\mu \in (-\infty, \infty)$ is a location parameter, $\sigma > 0$ is a scale parameter and $\xi \in (-\infty, \infty)$ is a shape parameter. For $\xi > 0$ this is the Fréchet cdf which has the finite lower endpoint $\mu - \sigma/\xi$, if $\xi < 0$ it is the (reversed) Weibull cdf with finite upper endpoint $\mu + \sigma/|\xi|$, and if $\xi = 0$ it reduces to the Gumbel cdf. In applications the model parameters are estimated from observations.

The reason for using the GEV cdf as a model for Block Maxima is similar to the central limit theory argument for using the normal distribution as a model for "random errors". The argument applies quite generally both to independent and dependent observations, and to observations made continuously in time or at discrete points of time. The present situation includes a further twist: the selection of near-crashes can be thought of as "subsampling of maxima". The use of the GEV cdf also in this situation is supported by the good fit to the observed values.

The Block Maxima method can also be used to study minima: one just considers maxima of the negated values instead of minima of the original values, since minus the maximum of the negated variables equals the minimum, (i.e. $\min\{x_1, \ldots x_n\} = -\max\{-x_1, \ldots -x_n\}$). This is how TTC is handled in the sequel.

In the multivariate Block Maxima method the observations are two or more maxima (e.g. the maximum of −TTC, the maximum speed, the maximum yaw rate, etc.) which occur in the same block (e.g. an annotated interval around a near-crash). For simplicity we only discuss the bivariate case where one considers two maxima from each block. As before we assume that maxima from different blocks (i.e. maxima coming from different near-crashes) are statistically independent.

Similar arguments as for the one-dimensional Block Maxima method provide a bivariate GEV as model for the vectors of Block Maxima. This bivariate GEV is built from two parts: the marginal distributions of the individual components of the Block Maxima vectors, and a dependence function which specifies the dependence between the components. The marginal cdf-s $G_1(x)$ and $G_2(x)$ of the first and second components are one-dimensional GEV-s of the form (1), usually with different parameters for the first component and for the second one. The bivariate GEV cdf has the form

$$G(x_1, x_2) = \exp\{-\ell(-\log G_1(x_1), -\log G_2(x_2))\}, \tag{2}$$

where $l(y_1, y_2)$ is the so-called stable dependence function. There is a large flexibility in how one can choose the stable dependence function, and different choices give different statistical models for the vectors of maxima. One simple and useful choice is the so-called bivariate logistic model in which the stable dependence function is determined by a single parameter $\alpha \in (0, 1]$, and has the form

$$l(v_1, v_2) = (v_1^{1/\alpha} + v_2^{1/\alpha})^{\alpha}.$$

Here the parameter α is a measure of the strength of dependence between the two components: $\alpha = 1$ means that the component Block Maxima are independent, while letting $\alpha \to 0$ leads to a model with complete dependence between the components.

2.2. Internal validation through prediction of crash frequencies

This validation method consists of using the near-crashes to predict the crash frequency in a naturalistic driving study. If the prediction is reasonably accurate, then it is an indication that the near-crashes are similar to crashes. On the other hand, if the predicted crash frequency is substantially different from the observed ones, this may be because near-crashes do not catch the same type of events as those which result in crashes.

The method consists of the following steps: (1) fit a GEV distribution to the observed maxima of a crash proximity measure, here −TTC, in all the near-crashes, (2) say that a crash happens when the crash proximity measure TTC crosses zero, or equivalently, that there is a crash if $\max\{-\text{TTC}\} > 0$, so that hence the probability of a crash is $\Pr(\max\{-\text{TTC}\} > 0)$, (3) compute an estimate, say $\hat{p}$ of this probability using the fitted GEV, and (4) compare $\hat{p}$ with the observed crash frequency, i.e. with f= (number of crashes)/(total number of near-crashes and crashes).

Step (1) contains a non-standard ingredient: since $-\text{TTC} > 0$ means a there has been a crash, this does not occur in any near-crash. Thus we should not fit an ordinary GEV distribution to the observed $\max\{-\text{TTC}\}$, but instead a GEV distribution conditional on the values being <0. In formulas, setting $M = \max\{\text{TTC}\}$, this conditional GEV distribution $G_0(z)$, for $z \le 0$ is

$$G_0(z) = \Pr(M \le z | M \le 0) = \frac{\Pr(M \le z)}{\Pr(M \le 0)} = \frac{G(z)}{G(0)}, \tag{3}$$

with $G(x)$ given by (1). An appendix describes how the parameters in (3) and their standard errors can be estimated.

2.3. Internal validation through multivariate near-crash modeling

The second validation method consists of (1) finding continuous "explanatory" variables which potentially could contribute to causing crashes and hence could be different/more extreme in near-crashes than in ordinary driving, and the more so the closer the near-crash is to a crash (for examples of possible such variables see Table 1), (2) fitting a multivariate GEV to $\max\{-\text{TTC}\}$ and the maxima (or minima, as appropriate) of the explanatory variables, as before conditional on $0 \ge \max\{-\text{TTC}\}$, (3) computing the distribution of the explanatory variables conditional on $\max\{-\text{TTC}\} > 0$, and (4) comparing this distribution with the distribution of the same variables in the crashes. If these two distributions are similar it is an indication that near-crashes can be representative for crashes, and if not it means that near-crashes and crashes differ in ways which have to be understood and handled.

3. Rear-striking in the 100-car study

In this section the validation methods from the previous sections are applied to rear-striking near-crashes in the 100-car study. We first describe the data set, then apply the relative frequency based validation method, and finally try out the multivariate approach. Our analysis was crucially aided by the availability of the free software NatWare (Dozza, 2012).

3.1. Data

Data acquisition techniques for the 100-car data base are extensively reported in Dingus et al. (2005), and a careful summary is given in Wu and Jovanis (2012). The data base contains 761 near-crashes and 69 crashes. Out of those we used the 384 near-crashes and 14 crashes which were classified as "rear-striking". For the near-crashes minimum TTC (=distance between the instrumented car and the lead car, divided by the difference in speed of the two

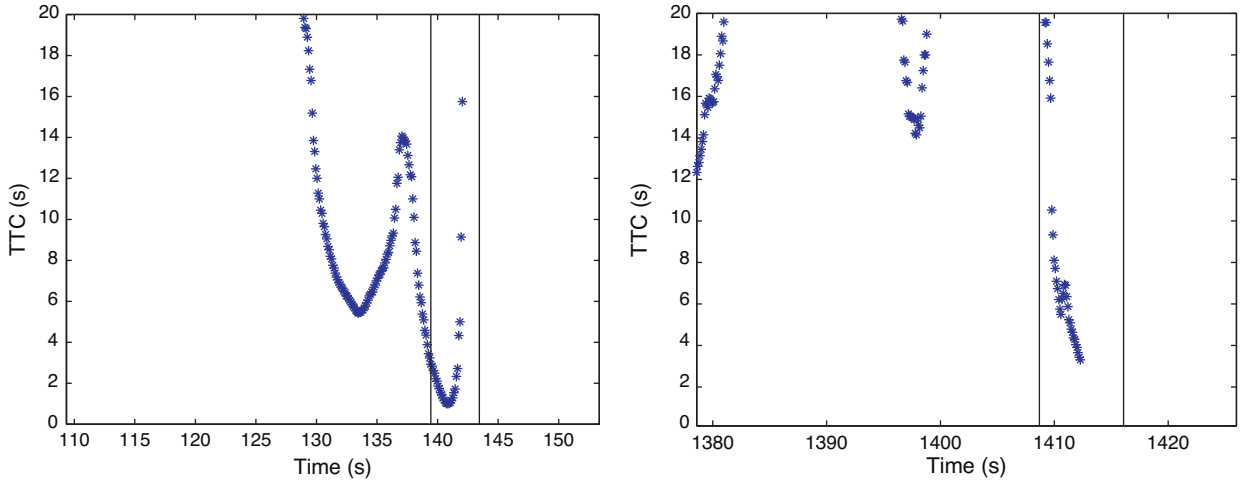

Fig. 1. *Left*: a TTC curve where the minimum TTC was used in the analysis. *Right*: a TTC curve where the minimum TTC was deemed not appropriate for analysis. Vertical lines mark the precipitating event, and the end of the near-crash, respectively.

cars) was computed manually from radar signals. We only kept the near-crashes were a clear minimum TTC could be observed in the interval between beginning and the end of the near-crash record. Due to absent or poor quality radar signals it was only possible to compute TTC for 47 of the near-crashes. An example of a computed TTC curve which was deemed to provide a useable minimum TTC and one who wasn't is shown in Fig. 1. Moreover, events with minimum TTC larger than 1.5 s were also discarded, see the discussion at the end of the introduction. (We used the limit 1.5 s rather than 1 s in order not to make the sample size too small.)

3.2. Results: crash frequencies

Fitting a GEV distribution (see Appendix), to the 29 rear-striking near-crashes with usable TTC-values gave the parameter estimates $\hat{\mu} = -1.21$, $\hat{\sigma} = 0.208$ and $\hat{\xi} = -0.0958$. The quantile plot in Fig. 2 indicated good model fit. Using the fitted GEV distribution, the probability that $\max\{-\mathrm{TTC}\} < 0$ was estimated to be 0.00020 with a 95% confidence interval (0.00017, 0.00022).

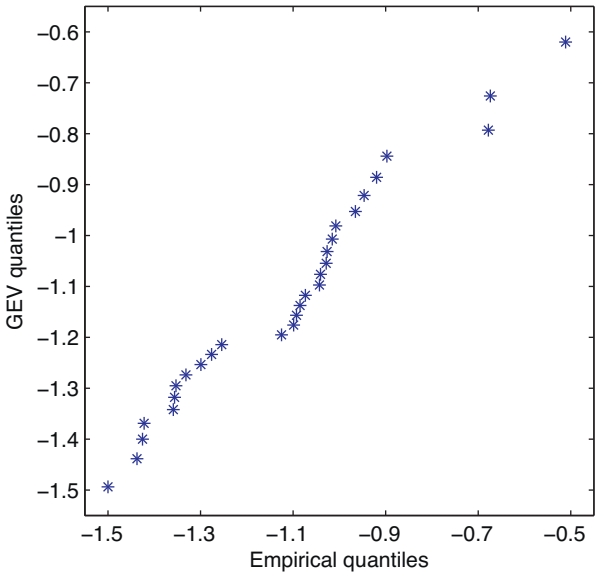

Fig. 2. qq-Plot of fit of GEV distribution for $-\min\{\mathrm{TTC}\}$ ($=\max\{-\mathrm{TTC}\}$). If model fit is good, the points fall close to a 45° line.

There were 14 rear-striking crashes and 384 rear-striking near-crashes in the 100-car data. Thus, the frequency of crashes is estimated to be 14/398 = 0.035, with a 95% binomial confidence interval (0.017, 0.053). This is 175 times larger than the estimate obtained from the fitted GEV distribution.

Individual study of the records showed that 12 of the 14 crashes were in slowmoving/queuing stop-and-go traffic, while the 29 near-crashes with usable TTC were in relatively free-flowing traffic. If these 12 crashes are excluded from the calculations, the relative frequency of crashes is changed to 2/386 = 0.0052, with a 95% confidence interval (0, 0.012). Thus, this estimated frequency is much smaller, but still is 26 times larger than the GEV estimate.

3.3. Results: multivariate distributions

Separate bivariate logistic GEV distributions were fitted to the joint distribution of $\max\{-\mathrm{TTC}\}$ and maxima or minima of 9 other variables, as shown in Table 1. We further studied one variable which was not a maximum, the "length of overlapping glance off road", with "overlapping" taken to mean that the glance off road either overlapped the precipitating event, or (to take the uncertainty in the annotator's timing of the precipitating event into account) ended within one second before the precipitating event.

For 2 of the 9 variables it was not possible to fit a bivariate GEV because of poor or missing signals. Additionally, 2 of the eye glance related variables were too discontinuous to fit to a GEV model. For the remaining 5 variables, the α-values indicated that there was no dependence with $\max\{-\mathrm{TTC}\}$. Also for the 4 variables which gave bad fit, scatterplots indicated little or no dependence with $\max\{-\mathrm{TTC}\}$. The left panel in Fig. 3 illustrates the lack of dependence by showing a plot of minimum TTC versus maximum speed. In 4 of the near-crashes video signals were not recorded, and in 13 there was no overlapping glance. The right panel in Fig. 3 shows that in the remaining 12 near-crashes the overlapping glance tended to be much longer when $\max\{-\mathrm{TTC}\}$ was closer to zero.

The left panel in Fig. 4 shows that density function obtained by using the fitted bivariate GEV to compute an estimate of the conditional density of maximum speed given that $\max\{-\mathrm{TTC}\} > 0$ is quite different from the histogram of the maximum speeds in the crashes.

As background, the distribution of maximum speed in the near-crashes with usable TTC was similar to the maximum speed distribution among all the near-crashes (Fig. 4, right panel). As

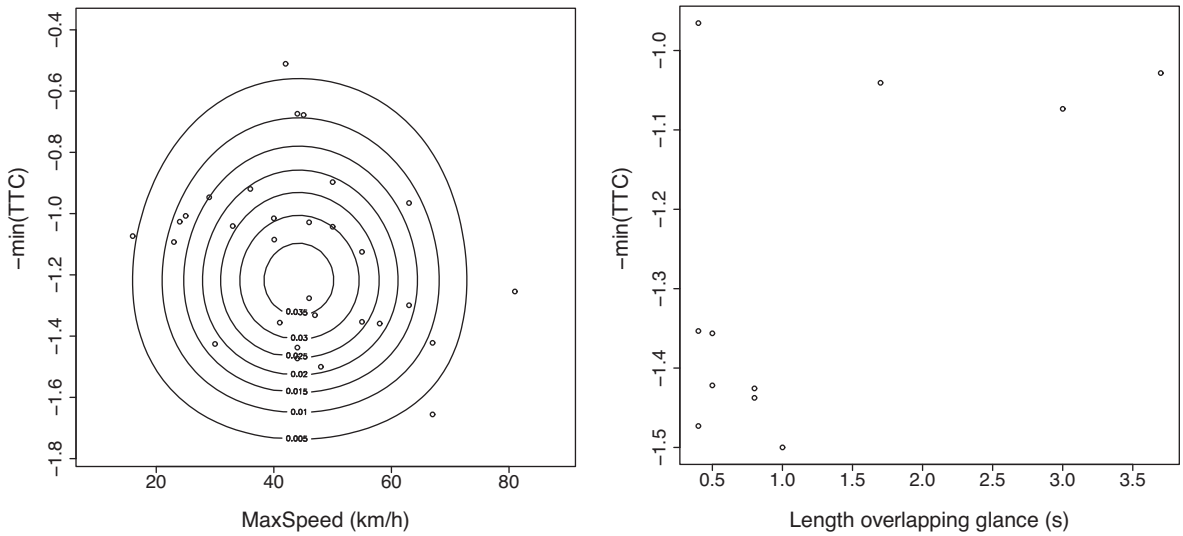

Fig. 3. *Left*: Scatterplot of −min{TTC} (=max{ − TTC}) vs. maximum speed, with superimposed level curves for fitted bivariate density. *Right:* Scatterplot of −min{TTC} vs. length of overlapping glance off road.

further background, using all the rear-end striking crashes and near-crashes, the odds ratio for crash with max speed less than 25 km/h is 48, cf. the introduction. If instead the odds ratio is computed from the near-crashes with usable TTC it is 38. Taking statistical uncertainty into account, these estimates are similar.

4. Discussion

In this section we discuss selection bias in the 100-car study, comment on the use of odds ratios to study selection bias, and point to some possibilities for future research.

4.1. Selection bias in the 100-car study

Section 3.2 showed that for rear-striking the predicted crash frequency obtained from the GEV distribution was very different from the frequency of crashes amongst the total number of crashes plus near-crashes. In Section 3.3 the distribution of the maximum event speed predicted from the near-crashes was found to be quite different from the actual distribution of the maximum speed in the crashes, see the right panel of Fig. 4. Thus both methods pointed to a considerable selection bias.

Informal investigation of the events revealed one source of this bias: 12 of the 14 crashes were at low speed/stop-and-go traffic, while the near-crashes with usable radar signals were in relatively free flowing traffic. Thus the near-crashes were not representative of these crashes. After removal of the crashes which occurred in slowmoving/stop-and-go traffic only two crashes remained, and the difference between predicted crash frequency and observed crash frequency was much smaller.

Nevertheless, two crashes are more than what is predicted by the univariate analysis. We have not been able to explain this remaining discrepancy – possible explanations could be (1) insufficient data quality, in particular for TTC, (2) bias in the selection of near-crashes with usable TTC, (3) bias in the selection of near-crashes in general; or (4) inaccuracies in the GEV predictions. It seems likely that (1) at least is a contributing factor. As for (2), the checks we have made did not

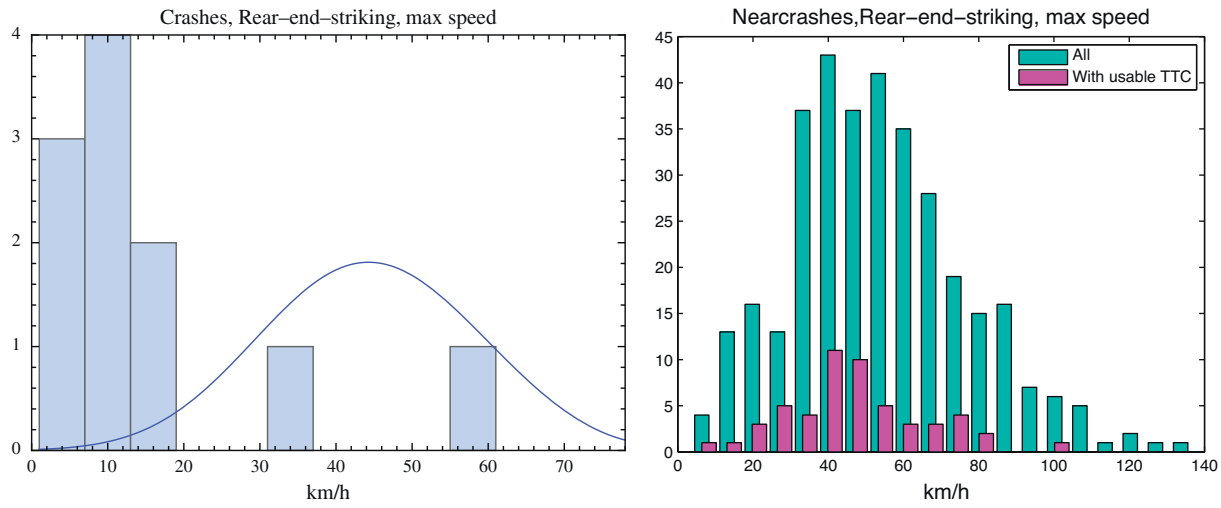

Fig. 4. *Left*: Histogram of maximum speed for rear-end crashes. Solid line is density function of maximum speed as predicted by the bivariate GEV analysis. *Right*: Histograms of maximum speeds for all rear-end striking near-crashes, and for the 29 rear-end striking near-crashes with usable TTC.

reveal any clear cause of bias: in particular, the right panel in Fig. 4 indicates that the distribution of maximum speed in the selected near-crashes and the distribution of maximum speed in all near-crashes are similar. The agreement in odds ratios computed from all near-crashes and from the near-crashes with usable TTC supports this impression. Cause (3) is always present, since there always are crashes which would not have been caught by pre-crash kinematic triggers, cf. Wu and Jovanis (2012). Finally, the good fit of the GEV distribution speaks against cause (4).

The rear-striking crashes in the 100-car study were of a kind which typically only leads to minor damage and hence is of less interest for traffic safety, and which would seldom be included in accident databases. Thus we believe that the crashes, for rear-end striking and in this study, have little relevance for increasing traffic safety. Instead the near-crashes in the study may potentially be more useful as tools for improving traffic safety. However, to understand whether this in fact is the case requires further validation. Specifically, external validation using accident statistics, in-depth accident investigations, and, potentially, event data recorders could help with this.

The results in Section 3.3 indicate that none of the 9 first variables in Table 1 are good predictors of crash risk, while "length of overlapping glance" seemed to be substantially bigger for more severe incidents. The latter is in agreement with recent findings by Lee et al. (2013). That min{TTC} does not seem to depend on speed is discussed in van der Horst (1990).

4.2. Relation to odds ratio methods

In a more general perspective, measured variables such as driver characteristics, road properties, traffic situation and other traffics states, and driver actions, may be of two kinds: those which are unrelated to accident risk, and those which contribute to accidents occurring. The first kind, by definition, has the same distribution in normal driving and in crashes, and if there is no selection bias they should also have the same distribution in near-crashes. In particular, risk and odds ratios for such variables should be close to one.

On the other hand, variables which influence crash risk should be present to a greater degree in near-crashes, than in normal driving, and to an even greater degree in crashes. Hence, firstly, risk and odds ratios computed using near-crashes and crashes in a naturalistic driving study will be biased downwards for such variables (Guo et al. (2010)).

Secondly, that factors which influence crash risk are present to a greater degree for more dangerous situations, makes standard odds ratio calculations for continuous variables such as visual inattention times problematic: these require a subjective choice of a cut-off which separates low and high values of the variable, and the more extreme the cut-off, the higher the odds ratio. This in particular carries a risk that researchers, consciously or unconsciously, select cut-offs which give the most interesting results, or the results which agree with present prejudices. (In fact, the cut-off speed for the odds ratio calculations above was chosen to underline the point we wanted to make.)

It is unclear to us how odds ratio calculations can contribute to internal validation of near-crashes, beyond discovery of clearly unintuitive properties. Of course, on the other hand, odds ratios potentially could be suited for external validation using variables which are available in both real crashes and in the naturalistic driving study.

These problems are circumvented by the methods we propose in this paper. The methods can also be used for external validation, to the extent that continuous variables are reliably available in reports of real crashes.

4.3. Future possibilities: crash severity estimation, parametric covariate models

In principle relative risk estimation using the methods from Section 2.2 is straightforward. E.g. if the risk of a rear-striking crash is significantly different when it is light and when it is dark it is inappropriate to apply the method in Section 3.2 to the data set consisting of all near-crashes. One could then instead divide the near-crashes into two sets, one consisting of the near-crashes which occurred when it was light and one with the crashes which happened when it was dark. The crash probabilities would then be estimated separately for each of the two data set using the methods from Section 2.2, and relative risk could be estimated by dividing these two probabilities. But, each of these two data sets would then be smaller and statistical uncertainty would increase and perhaps take over. The problem of increasing statistical uncertainty would be compounded further if one is interested in not just one covariate (light/dark), but several (e.g. light/dark and age of driver).

However, instead of dividing up into subgroups one can use covariate models where some the parameters of the GEV distribution, say the location parameter, depend on the covariates, but others are the same for all values of the covariates, see Coles (2001), Chapter 5. This potentially can reduce statistical uncertainty.

The second possibility is to use extreme value statistics to model the joint distribution of the minimum of the crash proximity measure and speed at the time when this minimum is attained. (This is different from the present paper: here we study the maximum of the speed during the entire event.) This joint distribution then provides an estimate of the distribution of speed at the time of impact in a crash: it is the conditional distribution of speed given that the minimum of the crash proximity measure is negative. In a second step, this speed distribution can be combined with well established knowledge about the relation between speed at impact and crash severity, to estimate a crash severity distribution. This is related to the approach in Hydén (1987), which used the speed at the start of the evasive maneuver as a part of crash severity prediction.

5. Conclusion

This paper develops frequency-based and multivariate methods for internal validation of near-crashes. Both methods were able to identify a substantial internal selection bias for rear-striking events in the 100-car naturalistic study. Further investigation revealed a major cause of this bias: all except two of the crashes were in slowmoving/stop-and-go traffic, while the near-crashes were in relatively free-flowing traffic. As argued above we believe that this indicates that the rear-striking crashes in the 100-car study are not useful as a means to improve traffic safety. On the other hand, the near-crashes potentially may be more similar to real crashes and crashes in accident data bases, and hence more relevant for improving traffic safety. To decide whether this is the case requires further study.

Our methods require the availability of an appropriate crash proximity measure. Such measures necessarily have to be different for different types of crashes, and some proximity measures may be more useful than others. Selection bias will be different for different types of crashes, and presumably also for separate naturalistic driving studies, cf. Wu and Jovanis (2012) and van der Horst (1990). We believe that successful use of naturalistic driving studies to improve traffic safety will require careful case-by-case internal and external validation and understanding of the nature of near-crashes and crashes.

Estimation of risk ratios can be improved if one has access to suitable baseline data on normal driving, and, of course, baseline data is also essential for the general understanding of near-crashes and crashes. Further, odds ratio calculations may be less useful for internal validation.

Naturalistic driving studies have large potential for adding to the understanding of detailed causes of traffic accidents, and, in a second step, to improve safety. However, the statistical methodology which is one of the ingredients needed for realizing this potential are still in their infancy, and in our opinion major research efforts should be directed at developing appropriate statistical approaches to analysis of the studies.

Acknowledgments

We thank D. Zholud (SAFER and Mathematical Sciences, Chalmers) for help with figures and comments; M. Dozza (SAFER, Chalmers) for comments and help with NatWare; J. Bärgman and T. Victor (SAFER, Chalmers); O. Nerman (Mathematical Sciences, Chalmers) for comments. Research supported by the VINNOVA SeMiFOT2 project; the US Strategic Highway Research Program 2; the SSF Gothenburg Mathematical Modelling Centre; and the Knut and Alice Wallenberg foundation. Data made available by Virginia Tech Transportation Institute.

Appendix A. Maximum likelihood estimation

The methods which were used in Section 3 to estimate parameters and standard errors are briefly described in this appendix.

Eq. (3) is the cdf of max{ − TTC} conditional on it being less than zero. Differentiating with respect to z and taking logarithms gives that the log-likelihood function for a sample $z_1, \ldots, z_n$ of max{ − TTC} from n near-crashes is

$$l(\mu,\sigma,\xi;z_1,\ldots,z_n) = -n\log\sigma - \left(\frac{1}{\xi}+1\right)\sum_{i=1}^{n}\log\left\{1+\frac{\xi}{\sigma}(z_i-\mu)\right\} - \sum_{i=1}^{n}\left(1+\frac{\xi}{\sigma}(z_i-\mu)\right)^{-1/\xi} + \sum_{i=1}^{n}\left(1-\frac{\xi}{\sigma}\mu\right)^{-1/\xi}, \quad \text{for } z_1,\ldots z_n \le 0. \tag{4}$$

This is the usual log-likelihood for fitting a GEV distribution except for the additional last "correction" term. Estimates $\hat{\mu}, \hat{\sigma}$ and $\hat{\xi}$ of the parameters are obtained by maximizing the log likelihood function with respect to μ, σ and ξ. An estimate of the probability of a crash, i.e. the probability $p = 1 - G_0(0)$ that max{ − TTC} is positive, is then obtained as

$$\hat{p} = 1 - \exp\left\{-\left(1-\frac{\hat{\xi}}{\hat{\sigma}}\hat{\mu}\right)^{-1/\hat{\xi}}\right\}. \tag{5}$$

By maximum likelihood theory (Coles, 2001, Section 2.6.4) the inverse of the observed information matrix provides an estimate $\hat{V}$ of the covariance matrix of the estimators. Using the delta method the standard error of $\hat{p}$ is estimated by $\nabla\hat{g}^T\hat{V}\nabla\hat{g}$ where $\nabla\hat{g}$ is obtained by replacing μ, σ, ξ by $\hat{\mu}, \hat{\sigma}, \hat{\xi}$ in

$$\nabla g^T = \left(\frac{\partial g}{\partial\mu}\ \frac{\partial g}{\partial\sigma}\ \frac{\partial g}{\partial\xi}\right),$$

where T denotes "transpose" and

$$\frac{\partial g}{\partial\mu} = \left(1+\xi\frac{z-\mu}{\sigma}\right)^{-1/\xi-1}\exp\left\{-\left[1+\xi\frac{z-\mu}{\sigma}\right]^{-1/\xi}\right\}$$

$$\frac{\partial g}{\partial\sigma} = \frac{z-\mu}{\sigma^2}\left(1+\xi\frac{z-\mu}{\sigma}\right)^{-1/\xi-1}\exp\left\{-\left[1+\xi\frac{z-\mu}{\sigma}\right]^{-1/\xi}\right\}$$

$$\frac{\partial g}{\partial\xi} = \left(\xi^{-2}\log\left\{1+\xi\frac{z-\mu}{\sigma}\right\}+\xi^{-1}\left(\frac{\sigma}{z-\mu}\right)\right)\left(1+\xi\frac{z-\mu}{\sigma}\right)^{-1/\xi} \times\exp\left\{-\left[1+\xi\frac{z-\mu}{\sigma}\right]^{-1/\xi}\right\}.$$

The correction term in (4) arising from conditioning on max{ − TTC} being less than zero, and the estimated probability of a crash, were small for the data studied in this paper. Hence omitting it would only change estimates slightly. However, for other data sets it could have a more important influence. Only crashes with min{TTC} < 1.5 were selected. We did not include conditioning on this.

For the bivariate analysis we again used maximum likelihood estimation with the likelihood function being the one obtained from the logistic model, conditional on max{ − TTC} being less than zero. Since this is completely parallel to the univariate analysis details are omitted.

References

Barnes, M., Blankespoor, A., Blower, A., Gordon, T., Green, P.E., Kostyniuk, L., LeBlanc, D., Bogard, S., Cannon, B.R., McLaughlin, S.B., 2011. Development of analysis methods using recent data: a multivariate analysis of crash and naturalistic event data in relation to highway factors using the gis framework, Technical report. University of Michigan Transportation Research Institute.

Campbell, K., Joksh, H.C., Green, P.E., 1996. A bridging analysis for estimating the benefits of active safety technologies. Technical Report UMTRI-96-18. University of Michigan, Transportation Research Institute.

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer-Verlag, New York.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A.D., Lee, S.E., Sudweeks, J., Perez, M.A., Hankey, M.A., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z.R., Jermeland, J., Knippling, R.R., 2005. The 100-car naturalistic driving study, phase ii – results of the 100-car field experiment. Technical Report. National Traffic Safety Admin (contract no. DTNH22-00-C-07007).

Dingus, T.A., Neale, V.L., Klauer, S.G., Petersen, A.D., Carroll, R.J., 2006. The development of a naturalistic data collection system to perform critical incident analysis: an investigation of safety and fatigue issues in long-haul trucking. Accident Analysis and Prevention 38, 1127–1136.

Dozza, M., 2012. What factors influence drivers' response time for evasive maneuvers in real traffic? Accident Analysis and Prevention (in press).

Gilleland, E., Ribatet, M., Stephenson, A.G., 2013. A software review for extreme value analysis. Extremes 16, 103–119, http://dx.doi.org/10.1007/s10687-012-0155-0.

Guo, F., Fang, Y., 2012. Individual driver risk assessment using naturalistic driving data. Accident Analysis and Prevention (in press).

Guo, F., Klauer, S.G., Hankey, J.M., Dingus, T.A., 2010. Near crashes as crash surrogate for naturalistic driving studies. Transportation Research Record 2147, 66–74.

Hydén, C., 1987. The development of a method for traffic safety evaluation: The Swedish Traffic Conflicts Technique. In: Bulletin 70. Lund Institute of Technology, Department of Traffic Planning and Engineering.

Jovanis, P.P., Aguero-Valverde, J., Wu, K.-F., Shankar, V., 2011. Analysis of naturalistic driving event data: omitted variable bias and multilevel modeling approaches. Transportation Research Record 2236, 49–57.

Lee, S.E., Simons-Morton, B.G., Klauer, S.E., Ouimet, M.C., Dingus, T.A., 2010. Naturalistic assessment of novice teenage crash experience. Transportation Research Record 2147, 66–74.

Lee, J.D., Victor, T.W., Dozza, M., 2013. Timing matters: distraction, glances, and crash risk (in preparation).

McLaughlin, S.B., Hankey, J.M., Dingus, T.A., 2008. A method for evaluating collision avoidance systems using naturalistic driving data. Transportation Research Record 40, 8–16.

Panagiotakopoulos, D., Majumdar, A., Ochieng, W., 2009. Characterizing the distribution of safety occurrences in aviation an approach using extreme value theory. Transportation Research Record 2106, 129–140.
Sogchitruksa, P., Tarko, A.P., 2006. The extreme value approach to safety estimation. Accident Analysis and Prevention 38, 811–822.
Tarko, A.P., 2012. Use of crash surrogates and exceedance statistics to estimate road safety. Accident Analysis and Prevention 45, 230–240.
van der Horst, A.R.A., 1990. A Time-based Analysis of Road User Behaviour in Normal and Critical Encounters. TNO Institute of Perception, Soesterberg, The Netherlands.
Wu, K.-F., Jovanis, P.P., 2012. Crashes and crash-surrogate events: exploratory modeling with naturalistic driving data. Accident Analysis and Prevention 45, 507–516.