Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 37 (2019) 195–202

Transportation Research Procedia

www.elsevier.com/locate/procedia

21st EURO Working Group on Transportation Meeting, EWGT 2018, 17th – 19th September 2018, Braunschweig, Germany

Freeway rear-end collision risk estimation with extreme value theory approach. A case study

Gregorio Gecchele[a], Federico Orsini[b], Massimiliano Gastaldi[b,*], Riccardo Rossi[b]

[a] *Atraki s.r.l., Via A. Diaz 4 - 37015 S. Ambrogio di Valpolicella (Verona), Italy*
[b] *University of Padova - Department of Civil, Environmental and Architectural Engineering, via F. Marzolo 9 - 35131 Padova, Italy*

Abstract

The current practice in crash-based safety analysis is hindered by some weaknesses: rarity of crashes, lack of timeliness, mistakes in crash reporting. Researchers are testing alternative approaches to safety estimation without the need of crash data. This paper presents an application of Extreme Value Theory in road safety analysis, using Time-To-Collision as a surrogate safety measure to estimate the risk to be involved in a freeway rear-end collision. The method was tested using data from an Italian toll-road with good results.

Keywords: Safety Analysis; Extreme Value Theory; Freeways; Surrogate Safety Measures.

1. Introduction

Road safety studies have usually adopted crash-based approach although it is hindered by several weaknesses, namely randomness and rarity of crash events, lack of timeliness, mistakes in crash reporting (Tarko et al. 2009; Tarko 2012). Crash-based safety analysis is a reactive approach, since a large amount of car accidents must happen (over years) before some action can be undertaken (Sayed and Zein, 1999). Given these issues, road safety analysis can benefit from alternative approaches which use surrogate safety measures. Effective surrogate measures should be based on observable non-crash events, predictably and reliably related to crashes, and provide a practical procedure for converting the non-crash events into crash frequencies data (Tarko et al., 2009). The Extreme Value Theory (EVT) is effectively used in several research areas to study extreme event distributions (Coles, 2001; Rocco, 2014;

* Corresponding author. Tel.: +39-049-827-5574;
E-mail address: massimiliano.gastaldi@unipd.it

10.1016/j.trpro.2018.12.183

Torrielli et al., 2013), but its application to road safety analysis is still limited. Songchitruksa and Tarko (2006) proposed a proactive method to estimate the risk of right-angle collisions at signalized intersections from measures of Post-Encroachment Time (PET). Tarko (2012), later on, proposed a general formulation of the application of EVT to road safety analysis, then applied by other researchers to different crash types (Jonasson and Rootzén, 2014). Zheng et al. (2014a) first tested two alternative EVT approaches, Block Maxima (BM) and Peak-Over-Threshold (POT), to relate PET and crashes in lane change maneuvers on freeways, and then proposed a shifted Gamma-Generalized Pareto Distribution to estimate the same type of crashes (Zheng et al., 2014b). Farah and Azevedo (2017) tested the Generalized Extreme Value distribution in the BM approach and the Generalized Pareto Distribution in the POT approach, to estimate head-on collisions in passing maneuvers on two-lane rural highways; following a previous work (Azevedo and Farah, 2015), they used the minimum Time-To-Collision (TTC) with the vehicle travelling in the opposite direction as surrogate safety measure.

In this study we applied EVT to estimate the probability of being involved in a freeway rear-end collision adopting TTC as a surrogate safety measure. Vehicles' trajectory information was derived from real-world traffic data from an Italian toll-road. The preliminary results show that the approach produces acceptable collision predictions compared to the actual observed values. This work contributes to the study of EVT application in road safety analysis, which appears a promising but still unexplored tool in this research field.

2. Methodology

2.1. Modelling details

Two main approaches in Extreme Value Theory can be identified: Block Maxima (BM) and Peak-Over-Threshold (POT). The first one examines maxima (or minima) over blocks of time (or space) and uses the Generalized Extreme Value (GEV) distribution, the latter, in which all values above a certain level are selected, adopts the Generalized Pareto (GP) distribution.

In the present study the Block Maxima is used. In BM approach observations are binned into time (or space) intervals (blocks). The maximum value from each block is considered an extreme event and is sampled into the estimation dataset, which is used to fit the model.

Considering a set of independently and identically distributed random observations $\{X_1, X_2, \dots, X_n\}$ with unknown distribution function $F(x) = \Pr(X_i \leq x)$, the maximum $M_x = \max(X_1, X_2, \dots, X_n)$ will converge to a GEV distribution when $n \to \infty$ (1):

$$G(x) = exp\left\{-[1 + \xi((x - \mu)/\sigma)]^{-\frac{1}{\xi}}\right\} \qquad (1)$$

where $-\infty < \mu < \infty$ is the location parameter, $\sigma > 0$ the scale parameter, and $-\infty < \xi < \infty$ the shape parameter.

Determining the block interval size is crucial: both, too small and too large intervals may lead to a poor fit.

When there is a limited number of blocks, but these blocks contain a sufficient number of observations, r-largest order statistic is usually adopted, because it allows to sample more extreme events to fit the GEV distribution. In an opposite situation, in which a large amount of block maxima is available, the fitting can sometimes be improved by subsampling the maxima, i.e. setting a limit value and then estimating the model only on the maxima above that value (Farah and Azevedo, 2017; Jonasson and Rootzén, 2014; Orsini et al., 2018). This is somehow similar to the POT approach, with the difference that for POT the threshold is applied to sample values from the base dataset, whereas in this case it is used to subsample from a sample (i.e., the block maxima) of the base dataset.

The BM method is also suitable for studying minima within the blocks, by considering the maxima of the negated values.

2.2. Data collection

In this study traffic and crash data were collected for a 5km-length, 3-lanes segment of the A4 freeway in Veneto Region, Italy.

Traffic data were collected for four weeks in 2013, as a result, a total of 1,044,444 vehicles were included in the analysis. Table 1 shows the Average Daily Traffic (ADT) for weekdays and the total week volume observed.

Table 1. Traffic data summary.

Week	Period	ADT Weekdays	Total week volume
1	14-20/01/2013	35,140	222,031
2	11-17/03/2013	38,190	242,032
3	15-21/07/2013	46,890	316,680
4	14-20/10/2013	40,660	263,701

Detailed vehicle-by-vehicle traffic information was collected using microwave Doppler radar: vehicle class, vehicle speed, time headway, and vehicle lane. Based on these data, Time-To-Collision (TTC) was calculated for each vehicle (the follower), analyzing the interaction with the lead vehicle in the same lane:

$$TTC = \frac{R}{RR} \qquad (2)$$

where R (range) is the separation between the leading and the following vehicle, and RR (range rate) is the speed difference between the following and the leading vehicle.

Only TTC values in which RR>0 were considered in the analysis, since they represent potentially dangerous conditions, in which the follower is approaching the leader at a higher speed. As a result, 544,362 positive TTC values were available for the analysis in the four weeks. For the model fitting activity, original TTC were converted to negated-TTC, in order to find extreme maxima values. This is a common practice in EVT theory applied to road safety estimation (Farah and Azevedo, 2017; Orsini et al., 2018; Songchitruksa and Tarko, 2006; Zheng et al., 2014a).

Crash data were collected by the administrative department of the A4 freeway for 6 years, from January 2011 to December 2016. Within the observed segment 73 crashes occurred and 21 of them were rear-end collisions. In order to better relate crash data to surrogate safety measure, in this study only working days (Monday to Friday) crashes were analyzed. As a result, a total of 16 crashes remained.

3. Results and analysis

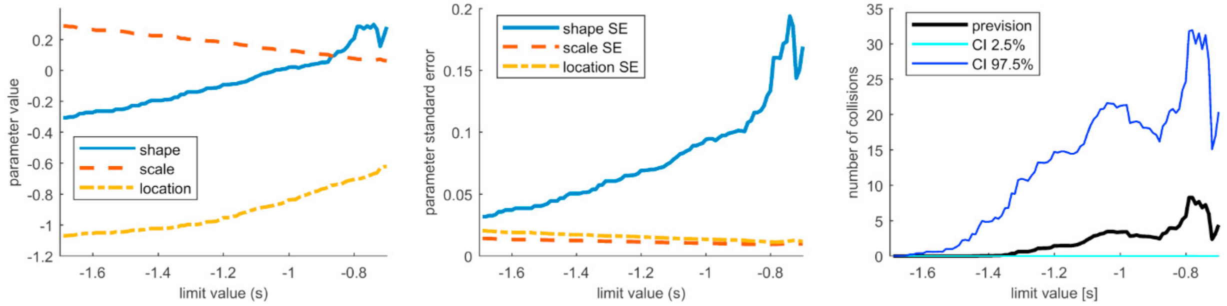
Traffic data were analyzed with the GEV model, using the Block Maxima method. Two different datasets were used: the first one contained hourly maxima (BM-hour), the second one r-largest daily maxima (BM-day). In the latter approach both stationary and non-stationary conditions were analyzed. GEV distributions were fitted using maximum likelihood (ML) estimation method.

3.1. Block Maxima with hourly values (BM-hour)

In this analysis the time block was one-hour long. Only daytime (07:00-19:00) working days (Monday to Friday) blocks were considered. This limitation was introduced to reduce inhomogeneity between blocks (Ferreira and De Haan, 2015). A total of 240 blocks was defined and the maximum negated-TTC value in each block was considered an extreme and included in the estimation dataset.

Following Farah and Azevedo's (2017) approach, these maxima were first pre-processed, in order to eliminate measurements that were not real extremes. A sensitivity analysis was carried out to assess the effect, on parameters and rear-end collision estimations, of the limit value that discriminates between non-extremes and extremes. The GEV was fitted multiple times against the measurements above the limit, the latter varying between -1.7 s and -0.7

s. After analyzing the stability plot of GEV parameters and parameters standard errors (Fig 1a, Fig 1b), all limits above -0.9 s were excluded, due to the shape parameter erratic trend. Among the remaining limit values, the one which maximized the number of estimated collisions was chosen for safety reasons (Fig 1c); this limit was -1.04 s.



a b c

Fig. 1. BM-hour graphs: (a) parameter stability plot; (b) parameter standard error stability plot; (c) estimated collisions changing the limit value.

This limit value does not only have a statistical meaning, but also a physical one. Past studies concluded that, in the case of rear-end collision, TTCs smaller than a low limit (typically 1 to 1.5 s) were useful as crash surrogates (Hydén, 1987; Jonasson and Rootzén, 2014).

After the original 240-maxima dataset had been filtered, 138 records had a negated-TTC above -1.04 s. The GEV model was fitted to these measurements and resulted in the parameters and respective standard errors shown in Table 2. Fig 3a presents the kernel probability density function of the empirical and modelled negated-TTC, and Fig 3b presents the simulated quantile-quantile plot (QQ plot), obtained by plotting the model data quantiles against the empirical quantiles, in order to compare the two probability distributions. These graphs allow to conclude that the modelled distribution has adequate fitting results.

Collision frequency was defined as the cumulative GEV probability of negated-TTC being higher than or equal to 0. This represents the probability of a collision to occur inside a one-hour block. By multiplying this probability by the number of blocks inside the observation time (in this case 138), the four-weeks collision probability is obtained. Note that this means that blocks excluded from the original 240-blocks dataset are assigned collision probability zero; this choice is in line with what made by Farah and Azevedo (2017). Further, multiplying by 52/4, the annual predicted number of rear-end collisions resulted in 3.48 predicted collisions, with a 95% confidence interval of [0.00;21.59].

The confidence interval was computed with the same simulation-based inference method presented by Songchitruska and Tarko (2006). Collision frequency is a scalar function of the model parameters, which, under regularity conditions, can reasonably be assumed to follow the multivariate normal distribution. The parameter values were generated 10^6 times and the collision frequency was computed each time, giving its empirical distribution; the 95% confidence interval was then calculated from the empirical distribution.

Table 2. Estimation results for the best fitted BM-hour model.

	Parameter					
	Location	Scale	Shape	Samples	Limit value	NLL
Value	-0.864	0.1361	0.0053	138	-1.04	-56.2071
Standard error	0.0139	0.0106	0.0892	n/a	n/a	n/a

3.2. Block Maxima with daily values (BM-day)

In this second approach the block was defined as whole day (00:00-23:59); only working days were considered, resulting in 5 blocks per week and 20 in the four-weeks observation time.

Following the approach used by Songchitruska and Tarko (2006) we defined as extremes and included in the estimation dataset, not just the maximum value from each block, but r-largest negated-TTC values.

The main difficulty in this approach is to determine the *r-value*. Different methods to choose the *r-value* were adopted in non-transport related applications. In a study on wave heights, Guedes Soares and Scotto (2004) fitted their GEV model several times, increasing the *r-value* by one each time, and applying the likelihood ratio test to compare two subsequently fitted models. The chosen *r-value* was the smallest *r* for which a model fitted on a dataset of (r+1)-largest values did not result significantly different to the one fitted on r-largest values.

R-largest order statistic was used also by Said et al. (2011) for predicting deteriorations in internet network traffic. To choose the *r-value* they also fitted the model several times, then they visually analyzed the trend of parameters' standard errors, plotted against *r-value*.

In the present study, similarly to the above-mentioned works, the GEV model was fitted multiple times, increasing the *r-value* from 1 to 50. Looking at parameters trends (Fig 2b), they showed a much more regular trend for *r-values* higher than 20. Likelihood ratio test was also conducted (alpha=0.05). Plotting the negative-log-likelihood against the *r-value* (Fig 2a), it was observed that it decreased for *r-values* lower than 21 and then rapidly increased. In addition to this, the highest collision estimation value was obtained with *r*=21 (Fig 2c).

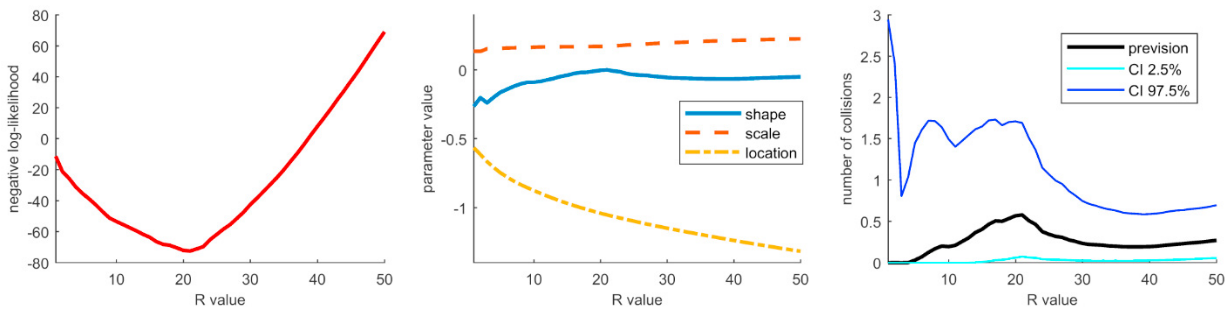


Fig. 2. BM-day graphs: (a) negative log-likelihood plot; (b) parameter stability plot; (c) estimated collisions changing the R-largest value.

Considering all these indicators, it was finally chosen an *r-value* equal to 21, and the model was fitted on the resulting dataset (Table 3). Fig 3c and Fig 3d show the probability density function of the empirical and modelled negated-TTC and the simulated QQ plot for the best fitted model. Both figures indicate a good fit between the modelled GEV distribution and the empirical data.

The annual collision estimation and its 95% confidence interval were obtained with the same procedure presented in section 3.1. The estimation was significantly lower (0.58) compared with that of the BM-hour and the confidence interval much smaller ([0.07;1.69]).

Table 3. Estimation results for the best fitted BM-day stationary model.

	Parameter					
	Location	Scale	Shape	Samples	R value	NLL
Value	-1.0527	0.1727	0.0016	420	21	-72.4123
Standard error	0.0097	0.0072	0.0424	n/a	n/a	n/a

It was then analyzed the effect of a covariate (daily traffic volume) in the model. The parameter estimations and standard errors of the non-stationary model are presented in Table 4; probability and QQ plots indicate a good fit (Fig 3e and Fig 3f).

The non-stationary model was tested against the stationary model using the likelihood ratio test, resulting in a p-value of 0.0083, significantly smaller than alpha=0.05: it confirms that the inclusion of the covariate in the model produced an improvement in the goodness-of-fit of the model.

The predicted number of annual collision was 0.71, with confidence interval [0.08;2.18].

Table 4. Estimation results for the best fitted BM-day non-stationary model.

	Parameter						
	Location (μ0)	Location (μ1)	Scale	Shape	Samples	R value	NLL
Value	-1.1943	3.4959*10-6	0.1696	0.0188	420	21	-75.8922
Standard error	0.0098	2.2290*10-8	0.0072	0.0484	n/a	n/a	n/a

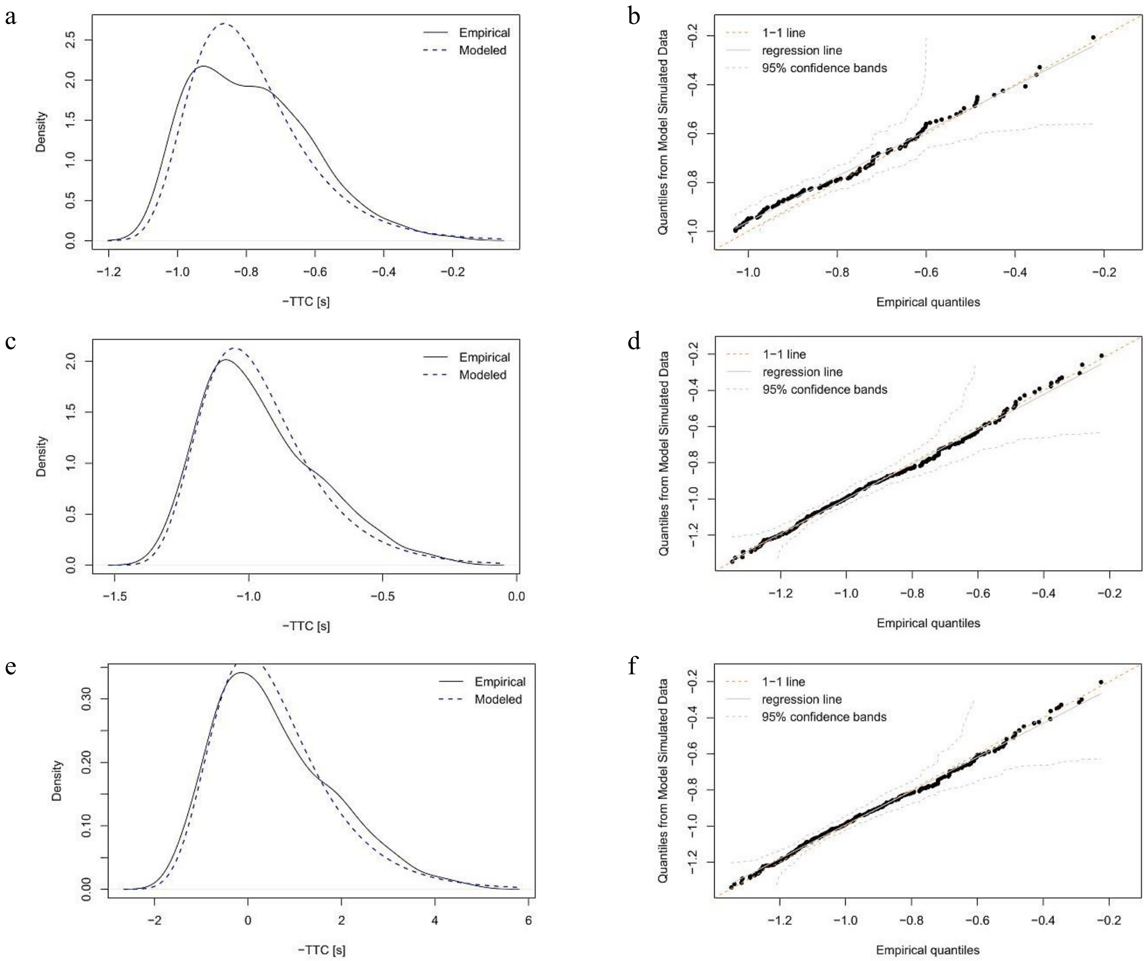


Fig. 3. Model fitting: BM-hour – (a) Kernel probability density plot, (b) Simulated QQ plot; Stationary BM-day – (c) Kernel probability density plot, (d) Simulated QQ plot; Non-stationary BM-day – (e) Kernel probability density plot, (f) Simulated QQ plot.

3.3. Comparing estimations with actual crash data

In transport-related GEV applications it is sometimes possible to compare modelling results with the number of crashes actually occurred during the observation time. This is usually the case of driving simulator experiments, like in the works of Tarko (2012), Farah and Azevedo (2017) and Orsini et al (2018).

In real-world applications this rarely happens, and estimations are compared with historical crash data. Crashes are usually rare events, so these data are aggregated over one or more years.

The physical location where accidents happen is usually the same location in which surrogate measure is recorded.

For example, Songchitruska and Tarko (2006) used as surrogate measure PET to estimate the number of right-angle collisions at signalized intersections. They collected these data videotaping traffic at 12 intersections in Lafayette area, Indiana. For each of these intersections they compared the estimated number of crashes with the historical number of crashes (collected over a 4-year period). In this case, collection and accident point were virtually the same exact place, i.e., the conflict point.

In a freeway case study, Zheng et al. (2014) were interested in predicting lane-change-maneuver collisions using as surrogate measure PET. They videotaped traffic on 29 road segments in Jinzhu, Yuegan and Kaiyang freeways in China. They compared the crash estimation in each segment with the historical crashes that happened within the observed segments. In their case collection and accident point were not the same place, but both points were located inside a certain road segment, identified by the camera angle.

In the present work, the surrogate measure (TTC) is collected in a road cross-section, but the aim is to predict collisions that happen inside a road segment. The issue is identifying the road segment size, as the number of actual crashes depends on its length: the longer the segment, the more accident happened and vice-versa.

Table 5 shows how the observed average annual number of collision changes with the length of the road segment (the analyzed section represents the middle section of the road segment).

The BM-day model¸ especially the non-stationary one, is able to predict with good accuracy accidents happening within 1 km road segment; this size may be reasonable and comparable to segments length in Zheng et al. (2014).

The BM-hour model is very sensitive and the presence of even a single very low value of TTC can greatly overestimate the number of collisions.

In the BM-day model the effect of a single extreme value still has an impact on the crash estimation, but confidence intervals are, in general, much smaller than BM-hour model. This is of course also related to the fact that the BM-day datasets contain more records than the BM-hour dataset.

In the view of a future practical application, looking at this comparison, the BM-day model seems more promising.

Further tests will involve applying this approach to other sections and verify if it is possible to identify a road distance within which the collision estimations are reliable.

Table 5. Mean, max and min annual number of accidents changing the road segment length (analyzed section is the middle section of the road segment).

Segment Length (m)	Nr. of accidents/year			Total accidents (6-year period)
	Min	Mean	Max	
0-1,000	0	0.67	2	4
0-2,000	0	1.50	3	9
0-3,000	0	2.00	3	12
0-4,000	1	2.50	4	15
0-5,000	1	2.67	4	16

4. Conclusions

This paper proposes an Extreme Value Theory approach (EVT), using Time-To-Collision as a surrogate safety measure, to estimate freeway rear-end collision risk for an Italian freeway. Based on the estimations from our data, for the proposed Block Maxima (BM) approaches, we can conclude that: the BM-hour approach (with hourly values) provided less satisfactory results, overestimating road accidents; the BM-day approach (with daily values) performed better than the BM-hour one: it produced the closest collision prediction to the actual observed value (especially with 1 km road segment length as reference), showing the narrowest confidence interval; the BM-day approach, especially the non-stationary model, appears to be promising in view of practical application.

This work can be seen as a precious contribution in the study of EVT in road safety analysis, which appears to be a promising but still unexplored approach for safety evaluation. There are several directions in which this work could be extended in the future: the introduction of new covariates for the BM-day model; the application of the Peak Over Threshold (POT) approach; the application of the models to other road sections and/or other freeways; an analysis of the transferability of the results to other countries; the application of the models to other types of accidents.

Acknowledgements

The work was supported by the University of Padova [grant code BIRD177377/17], and partially developed within the European Social Fund project "Traffic control center to integrated mobility management platform".

References

Azevedo, C.L., Farah, H., 2015. Using Extreme Value Theory for the Prediction of Head-On Collisions During Passing Maneuvres. 2015 IEEE 18th Int. Conf. Intell. Transp. Syst. 268–273. https://doi.org/10.1109/ITSC.2015.53

Coles, S., 2001. An introduction to statistical modeling of extreme values, Springer Series in Statistics. London. https://doi.org/10.1007/978-1-4471-3675-0

Farah, H., Azevedo, C.L., 2017. Safety analysis of passing maneuvers using extreme value theory. IATSS Res. 41, 12–21. https://doi.org/10.1016/j.iatssr.2016.07.001

Ferreira, B.Y.A., De Haan, L., 2015. On the block maxima method in extreme value theory: PWM estimators. Ann. Stat. https://doi.org/10.1214/14-AOS1280

Guedes Soares, C., Scotto, M.G., 2004. Application of the r largest-order statistics for long-term predictions of significant wave height. Coast. Eng. 51, 387–394. https://doi.org///doi.org/10.1016/j.coastaleng.2004.04.003

Hydén, C., 1987. The development of a method for traffic safety evaluation:the Swedish traffic conflicts technique. Bull. Lund Inst. od Technol. 70.

Jonasson, J.K., Rootzén, H., 2014. Internal validation of near-crashes in naturalistic driving studies: A continuous and multivariate approach. Accid. Anal. Prev. 62, 102–109. https://doi.org/10.1016/j.aap.2013.09.013

Orsini, F., Gecchele, G., Gastaldi, M., Rossi, R., 2018. Collision prediction in roundabouts: a comparative study of extreme value theory approaches. Transp. A Transp. Sci. https://doi.org/10.1080/23249935.2018.1515271

Rocco, M., 2014. Extreme value theory in finance: A survey. J. Econ. Surv. 28, 82–108. https://doi.org/10.1111/j.1467-6419.2012.00744.x

Said, A.M., Hasbullah, H., Dahab, A.Y., 2011. R-largest order statistics for the prediction of bursts and serious deteriorations in network traffic. 2011 IEEE 3rd Int. Conf. Commun. Softw. Networks 579–583. https://doi.org/10.1109/ICCSN.2011.6014960

Sayed, T., Zein, S., 1999. Traffic conflict standards for intersections. Transp. Plan. Technol. 22:4, 309–323. https://doi.org/10.1080/03081069908717634

Songchitruksa, P., Tarko, A.P., 2006. The extreme value theory approach to safety estimation. Accid. Anal. Prev. 38, 811–822. https://doi.org/10.1016/j.aap.2006.02.003

Tarko, A.P., 2012. Use of crash surrogates and exceedance statistics to estimate road safety. Accid. Anal. Prev. 45, 230–240. https://doi.org/10.1016/j.aap.2011.07.008

Tarko, A.P., Davis, G., Saunier, N., Sayed, T., Washington, S., 2009. Surrogate Measures of Safety - White Paper, FHWA. Washington D.C., USA.

Torrielli, A., Repetto, M.P., Solari, G., 2013. Extreme wind speeds from long-term synthetic records. J. Wind Eng. Ind. Aerodyn. 115, 22–38. https://doi.org/10.1016/j.jweia.2012.12.008

Zheng, L., Ismail, K., Meng, X., 2014a. Freeway safety estimation using extreme value theory approaches: A comparative study. Accid. Anal. Prev. 62, 32–41. https://doi.org/10.1016/j.aap.2013.09.006

Zheng, L., Ismail, K., Meng, X., 2014b. Shifted Gamma-Generalized Pareto Distribution model to map the safety continuum and estimate crashes. Saf. Sci. 64, 155–162. https://doi.org/10.1016/j.ssci.2013.12.003