International Co-operation on Theories and Concepts in Traffic Safety

Investigating the optimal sample size for traffic conflict observation using extreme value theory approach

[1]Lai Zheng*, [2]Jiayi Li, [2]Shuanghu Ma
*lead presenter
[1]zhenglai@hit.edu.cn, Harbin Institute of Technology, China
[2] Harbin Institute of Technology, China

Introduction

Traffic conflict techniques have gained increasing popularity as a surrogate measure of crash for road safety analysis. However, what is the optimal sample size for traffic conflict observation is still an open question. The selection of the sample size is a critical challenge in traffic conflict related studies. On the one hand, excessive samples can lead to additional cause of resources, making the analysis process increasingly complex and time consuming. On the other hand, a low sample size would do harm to the representativeness of the results, which can lead to erroneous conclusions. Currently, the selection of sample sizes for various types of traffic conflicts is generally based on empirical judgment rather than a well-established theoretical framework. To bridge this gap, this studies introduces a novel as well as practical approach to determine the optimal sample sizes based on the extreme value theory. To be specific, generalized Pareto distribution models are developed based on traffic conflicts of different sample sizes, and the optimal sample size is determined based on the stability of estimated crash return level.

Methodology

The peak over threshold (POT) approach that takes observations over a predetermined threshold as extremes is used, and it provides a class of models to enable extrapolation from the frequent events (e.g., traffic conflicts) to infrequent events (e.g., crashes). Mathematically, let X_1, X_2, ..., X_n are independently and identically distributed random variables with unknown distribution function $F(x)=\Pr(X_i \leq x)$. For a sufficiently high threshold u, the conditional distribution function $F_u(x) = \Pr(X - u \leq x | x > u)$ could be approximated by a generalized Pareto distribution (GPD), and the form is as follows:

$$G(x; u, \sigma, \xi) = 1 - \left(1 + \frac{\xi}{\sigma}(x - u)\right)^{-\frac{1}{\xi}} \quad (1)$$

where u is the predetermined threshold; $\sigma > 0$ is the scale parameter; $-\infty < \xi < \infty$ is the shape parameter.

Base on the GPD, the safety indicator crash return level can be obtained. The crash return level z_p associated with the return period $T=1/p$ gives

$$z_p = \Pr(z \leq z_p) = 1 - p \quad (2)$$

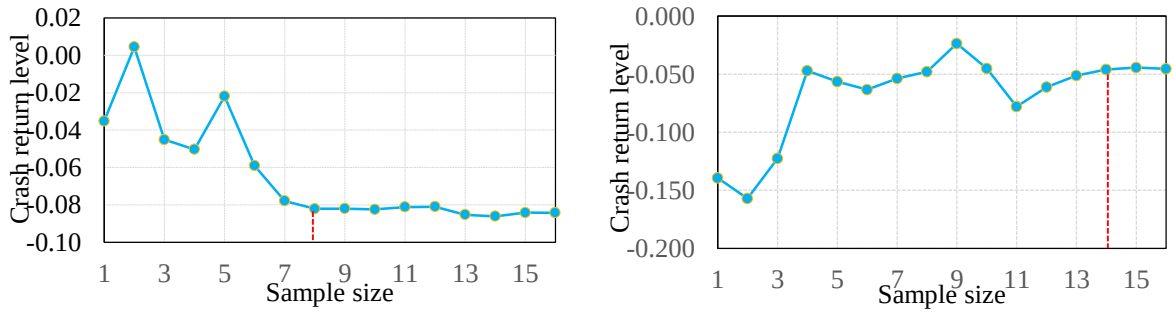
where p is a specified probability under the tail area. The crash return level can be interpreted as the average waiting time until next occurrence of a crash is T years, or the average number of crashes occurring in a T-year period is one. In case of negated PETs, the one-year return level $z_p \geqslant 0$ implies that a positive crash frequency is expected during a year, and the $z_p \leqslant 0$ implies that zero crash frequency is expected. In general, the lower the z_p value is, the lower the crash risk. For an investigated entity, its expected crash return level should be constant with the allowance of sampling errors. This argument suggests that, a sample size that is large enough

should be the one from which the estimated crash return levels remain near-constant with the increase of the sample size.

Results

The proposed method was applied to the data collected from a signalized intersection in Harbin, China with a total duration of six days, 8 hours per day. The data were collected using the LiDAR, and algorithms were developed to extract trajectories of different road users (i.e., vehicles, bicycles, and pedestrians) and calculate the PET between left-turn vehicles and other vehicles. The events with PET⩽4s were considered as traffic conflicts.

With a 3-hour interval, the collected 48-hours data were divided into 16 groups, and GPD models were estimated based on sample sizes of 3 hours, 6 hours, …, and 48 hours, considering different types of traffic conflicts. A total of 32 models were developed at last, and the corresponding estimated crash return levels were calculated. It is noted that the thresholds were determined using the threshold stability plot and mean residual plot. For both the conflicts between motorized and motorized vehicles and conflicts between motorized and non-motorized vehicles, the thresholds were 1.5s. The estimated crash return levels corresponding to different sample sizes is shown in Figure 1. It can be found that the estimated crash return levels become stable when the sample size is 8*3=24 hours for traffic conflicts between motorized vehicles, while the sample size for traffic conflicts between motorized and non-motorized vehicles is 14*3=42 hours. It was also found that the average hourly conflict rate for motorized-motorized vehicles was 40.7, while that for motorized-non-motorized vehicles was 11.5. It indicates the sample size needed for conflicts that occur more frequent is smaller.



(a) motorized-motorized vehicles (b) motorized and non-motorized vehicles

Figure 1 Crash return levels corresponding to different sample sizes

Conclusion

This study proposed an approach to determine the optimal sample size of traffic conflict observation based on the extreme value theory. The generalized Pareto distributions was employed to fit the conflict extremes of different sample size, and the crash risk return levels were calculated as a safety indicator. The basic principle is that, the expected value of the safety indicator should be constant, and thus the minimum sample size needed for traffic conflict observation should be the one from which the estimated crash return level become stable with the increase of the sample size. The application of the approach to left-turn related conflicts shows that the optimal sample sizes for conflicts of motorized-motorized vehicles and conflicts of motorized and non-motorized vehicles are 24 hours and 42 hours, respectively.