Chapter 3

Methodology

3.1 Preface

This chapter describes the statistical modeling approaches utilized in this thesis for modeling conflict data . First, this chapter introduces the Extreme value theory (EVT), used to model conflict and predict the probability of crashes. Section 3.1 outlines the basic theory related to extreme value models, modeling approaches, and threshold selection techniques. Application of EVT in previous conflict studies is reviewed in section 3.2. The Bayesian modeling approach is described in Section 3.3.

3.2 Extreme Value Theory

Extreme value theory (EVT) enables the researchers to model and predict the behaviour of extreme events which are low frequency and high severity events. Since EVT enables to predict the unobserved level of extreme event, conflict-based crash risk assessment is gaining popularity in road safety literature. Since traffic conflicts are extreme events, Extreme value theory (EVT) has been used extensively in segregation and modeling conflict. EVT enables to predict crash probability and crash frequency without using crash

data. Songchitruksa and Tarko [181] introduced EVT based model for conflict-based crash risk estimation at signalized intersections. Since then, many studies have shown that EVT is a promising tool for modeling conflict [182–184].

Selection of proper sampling approach for EVT-based models is important for selecting extreme observations. Also, researchers have found that using bivariate EVT models lead to more accurate and precise estimates of crashes than univariate models [185–187]. While using a bivariate EVT model, Zheng et al. [187]) found that the combination of two independent conflict indicators leads to better crash estimation. Borsos [188] used speed related surrogate measures to model conflict severity and found that temporal indicator alone is not good enough to make inferences about severity of conflict events. The author suggested that speed related surrogate measures may be used along with temporal indicators for safety assessment. Cavadas et al. [189] used two dependent SSMs to model crash probability in overtaking manoeuvres using detailed trajectory data from a driving simulator. They suggested that extreme value copula may be used to model SSMs when the dependence structure is not consistent. The authors further suggested that driver and roadway geometry should be used for defining conflicts or safety events. Literature suggests that accuracy of extreme value models can be improved by using joint site multivariate EVT models by combining data from similar sites [42, 190].

Extreme events can be modeled using two approaches, namely (1) block maxima and (2) peak over threshold approach. In block maxima approach, the dataset is divides into blocks, and a single extreme observation is selected from each block. Previous conflict studies identified few issues with the BM approach. First, selecting block intervals is subjective, with no guidelines for selecting appropriate block sizes. In a recent study, Kar, Venthuruthiyil, and Chunchu [177] pointed out problems with selection in block size, where the same block size may not contain extreme events for all vehicle types in heterogeneous traffic. Second, selecting a single observation from each block may

discard other extreme observations within the same block. Researchers suggest peak over threshold approach for modeling conflicts as it selects extreme observation more efficiently [182, 183, 191]. The detail about peak over threshold approach used for modeling conflict is presented in next section.

3.2.1 Univariate Peak-Over Threshold Model

Pickands [192] proposed the peak over threshold approach to model all the extreme observations exceeding a specific high threshold. Let $X_1, X_2, X_3, \ldots$ be a sequence of independent random variables with the common distribution function F, and let

$$M_n = \max\{X_1, X_2, X_3, \ldots, X_n\} \tag{3.1}$$

And for large n,

$$\Pr\{M_n \leq z\} \approx G(z), \tag{3.2}$$

where

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{3.3}$$

for some $\mu, \sigma > 0$ and ξ. Then for large enough u, the distribution function of $(X-u)$, conditional on $X > u$, is approximated by a univariate generalized Pareto distribution (GPD):

$$H(y) = 1 - \left(1+\frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{3.4}$$

defined on $\{y : y > 0 \text{ and } \left(1+\frac{\xi y}{\tilde{\sigma}}\right) > 0\}$, where

$$\tilde{\sigma} = \sigma + \xi(u-\mu) \tag{3.5}$$

3.2.2 Bivariate Peak-Over Threshold Model

Bivariate peak over threshold model is analogous to univariate peak over threshold model in which threshold excesses for both variables are modeled as joint probability distribution. Let $\{(X_1,Y_1),(X_2,Y_2),(X_3,Y_3),\ldots,(X_n,Y_n)\}$ be a sequence of independent random variables drawn from $F(X,Y)$. For large enough thresholds, u_x and u_y, threshold excesses, $(X > u_x$ and $Y > u_y)$ will follow Generalized Pareto Type distribution (GPD) with parameters (σ_x,ξ_x) and (σ_y,ξ_y) for X and Y respectively as stated in Eqn.3.6 [172].

$$G(z) = 1 - \left\{1 + \xi\left(\frac{z-u_z}{\sigma_z}\right)\right\}^{-1/\xi_x} \quad \text{for } z > u_z \tag{3.6}$$

where, σ and ξ are scale and shape parameters respectively, $\varsigma_x = \Pr(X > u_x)$ and $\varsigma_y = \Pr(Y > u_y)$. Consider $\tilde{X}$ and $\tilde{Y}$ as the transformations (defined in Eqn.3.7 and Eqn.3.8) of X and Y given that $X > u_x$ and $Y > u_y$. $\tilde{X}$ and $\tilde{Y}$ are approximately standard Fréchet distributed (F).

$$\tilde{X} = -\left(\log\left\{1 - \varsigma_x\left[1 + \left(\frac{\xi_x(X-u_x)}{\sigma_x}\right)^{-\frac{1}{\xi_x}}\right]\right\}\right)^{-1} \tag{3.7}$$

$$\tilde{Y} = -\left(\log\left\{1 - \varsigma_y\left[1 + \left(\frac{\xi_y(Y-u_y)}{\sigma_y}\right)^{-\frac{1}{\xi_y}}\right]\right\}\right)^{-1} \tag{3.8}$$

The joint distribution, $F(x,y)$, of threshold excesses $X > u_x$ and $Y > u_y$ for such random variables can be written as (Coles, 2001):

$$F(x,y) \approx G(x,y) = \exp\left\{-V(\tilde{X},\tilde{Y})\right\}, \quad \text{for } x > u_x \text{ and } y > u_y \tag{3.9}$$

Where, $(\tilde{x},\tilde{y})$ are the transformations of (x,y) and $V(\tilde{x},\tilde{y})$ is the dependence function of the marginal Fréchet distribution $F^{\sim}$ and is defined as:

$$V(\tilde{x},\tilde{y}) = 2\int_0^1 \max\left(\frac{w\tilde{y}}{\tilde{x}+\tilde{y}}, \frac{(1-w)\tilde{x}}{\tilde{x}+\tilde{y}}\right) dH(w) \tag{3.10}$$

Where, $H(w)$ is called the spectral measure, defined as a distribution function on [0,1], satisfying the following criteria:

$$\int_0^1 w dH(w) = \frac{1}{2} \tag{3.11}$$

$$V(\tilde{x},\tilde{y}) = 2\left(\frac{1}{\tilde{x}}+\frac{1}{\tilde{y}}\right)\int_0^1 \max\left(\frac{w\tilde{y}}{\tilde{x}+\tilde{y}}, \frac{(1-w)\tilde{x}}{\tilde{x}+\tilde{y}}\right) dH(w) \tag{3.12}$$

Let

$$A(t) = \int_0^1 \max\{w(1-w)t, (1-w)t\} dH(w) \tag{3.13}$$

Then, $F(x,y)$ can be written as

$$F(x,y) \approx G(x,y) = \exp\left\{-\left(\frac{1}{x}+\frac{1}{y}\right)A\left(\frac{x}{x+y}\right)\right\} \tag{3.14}$$

Where, $A(t)$ is called the Pickands dependence function, which is a convex function [193] and satisfies $(1-t)\vee t \leq A(t) \leq 1$, where $t \in [0,1]$.

In parametric estimation, various distributions can be assumed for spectral measure function, H. The logistic distribution function is most commonly used in this case [185, 187]. The other functions include negative logistic, asymmetric logistic, asymmetric negative logistic, bilogistic, negative bilogistic and Husler-Reiss. More details about these functions can be obtained from Beirlant et al. [193]. The details of parametric family of distribution used in present study are given in Table 3.1.

Table 3.1 Parametric bivariate extreme value distributions used in the present study

Name of distribution	Functional form	Characteristics
Logistic	$G(x,y) = \exp\left\{-\left(x^{1/r}+y^{1/r}\right)^r\right\}$, where $0 < r < 1$	Completely dependent: $r \to 0$ Independent: $r \to 1$
Negative Logistic	$G(x,y) = \exp\left\{-x-y+\left(x^{-r}+y^{-r}\right)^{-1/r}\right\}$, where $r > 0$	Completely dependent: $r \to \infty$ Independent: $r \to 0$
Husler-Reiss	$G(x,y) = \exp\left\{-x\Phi\left[r^{-1}+\frac{1}{2r}\log\left(\frac{x}{y}\right)\right] -y\Phi\left[r^{-1}+\frac{1}{2r}\log\left(\frac{y}{x}\right)\right]\right\}$, where $r > 0$, Φ is the standard normal distribution function	Completely dependent: $r \to \infty$ Independent: $r \to 0$

3.2.3 Threshold Selection Procedure

Threshold excess model enables efficient use of data in modeling of extreme events. Estimation of proper threshold is an important step in fitting peak over threshold models. Selecting a low threshold leads to entry of non-extreme data into the model, which causes bias. Also, a very high threshold leads to reduces the number of data points causing high variance. EVT offers several techniques (Mean residual life plot, threshold stability plot, spectral measure plot) for selection of appropriate thresholds for modeling extreme events.

1. Mean residual life plot

Mean residual life plot is based on the mean of GPD. If $X = \{X_1, X_2, \ldots, X_n\}$ is a random variable whose threshold excess ($X > u_0$) follows GPD, then the mean of excesses can be derived as [172]:

$$\mathbb{E}[X - u_0 \mid X > u_0] = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0} + \xi u}{1-\xi} \tag{3.15}$$

For $u > u_0$, $E((X-u) \mid (X > u))$ is a linear function in u, given that GPD is a valid model for u (Eqn.3.15). This result leads to the mean residual life (MRL) plot, which is the

locus of (P,Q) given by $\{(P = u, Q = \text{mean of excesses: } u < x_{\max})\}$. A suitable threshold u can be chosen on the MRL curve wherever the curve is approximately linear.

2. Threshold stability plot

The second method of threshold selection is based on the threshold stability plot (TSP). For the random variable $X = \{X_1, X_2, \ldots, X_n\}$, if the generalized Pareto distribution (GPD) is a valid model for threshold excesses $(X > u_0)$, then for any threshold $u > u_0$, the excesses would also follow a GPD with the same shape parameter ξ and scale parameter σ_u given by the following equation [172]:

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0) \tag{3.16}$$

Equation (14) implies that $\sigma_u - \xi u = \sigma_{u_0} - \xi u_0 = \sigma^*$ (say), which is independent of u. The parameters σ^* and ξ remain constant when exceedances $(X > u > u_0)$ follow GPD. In the threshold stability plot, u is plotted against σ^* and ξ, and the lowest u_0 for which the estimate remains nearly constant is selected as the threshold.

3. Spectral measure plot

Beirlant et al. [193] proposed a spectral measure plot (H versus k) for selecting thresholds for joint bivariate distribution. Both the variables (with n observations) are transformed into standard Fréchet distributions as follows

$$x^*_{ij} = -\frac{1}{\log u_{ij}}, \quad i = 1,2,3,\ldots,n \quad \text{and} \quad j = 1,2 \tag{3.17}$$

The radial coordinate r_i is defined as the sum norm of transformed variables (x^*_{i1}, x^*_{i2}) as stated below:

$$r_i = x^*_{i1} + x^*_{i2} \tag{3.18}$$

Radial coordinates r_i are sorted (descending order) to obtain r_0 with k as the index. Then, k corresponding to empirical spectral measure $H = r_0 \times k/n = 2$ (which is the theoretical value of H for a bivariate case) is selected as the position of the threshold. The highest $(n - k_0)$ observations of the sorted data would be used for a bivariate GPD model. Further details about spectral measure plot can be found in Beirlant et al. [193].

3.3 Bayesian Modeling Approach

3.3.1 Basic theory

Bayesian inference, based on Bayes' theorem, is a method for incorporating uncertainty in measurements. Bayes' theorem offers a way to update the probability of an event or parameter using the observed data and prior knowledge. The updated probability is derived by conditioning the prior probability on the observed data. In the Bayesian framework, a model is first constructed similar to classical statistics. Next, a probability distribution known as the prior distribution is assumed for the unknown parameters in the model. This is termed "prior" and is used because it is not based on data but rather on subjective reasoning before incorporating observed data. It could be defined based on experience with observations and data. The parameters' probability distribution is then updated based on the observed data and Bayes' theorem. The resulting probability distribution is referred to as the posterior distribution. This posterior distribution integrates both the data and the prior and depicts a model for the parameters. Using the Bayesian method for estimating model parameters, the posterior distribution for the given observation y can be written as

The posterior distribution $p(\theta|y)$ is given by

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{\int p(y|\theta)\, p(\theta)\, d\theta} \tag{3.19}$$

where $p(y|\theta)$ is the likelihood, $p(\theta)$ is the prior distribution, and $\int p(y|\theta)\, p(\theta)\, d\theta$ is the normalizing term.

3.3.2 Bayesian hierarchical inference

Bayesian hierarchical models are an extension of Bayesian modeling where parameters are organized into a hierarchy, reflecting the structure of the data. A hierarchical framework involves a population with a multilevel or hierarchical structure. Population in these cases is comprised of many sub-groups. Bayesian hierarchical models are used when each population sub-group contributes information about other sub-groups. The hierarchy allows for partial pooling of information across groups, providing more stable and reliable estimates, especially when dealing with sparse data in some groups. This approach is highly beneficial in traffic modeling, where data might be collected from different locations, times, or vehicle types. By using Bayesian hierarchical inference, one can model complex dependencies and variations within the data, leading to more accurate and robust predictions. Complete pooled models fit the same monolithic parameters, assuming that the population is homogeneous and a universal model is appropriate for all subgroups in the population. Fig. 3.1. depicts a general framework for a pooled and hierarchical model.

In a pooled model, the population is assumed to be homogeneous, and each observation $x_1, x_2, x_3, \ldots, x_n$ is exchangeable.. If the traffic stream is composed of multiple vehicle types, assuming the population to be homogeneous and fitting a pooled model is not appropriate. In a hierarchical model, the population is composed of several heterogeneous groups, and observations $x_1, x_2, x_3, \ldots, x_n$ are exchangeable only within the groups [194]. The Bayesian hierarchical modeling approach enables data integration from groups with similar characteristics and allows the model parameter to vary across the group. This framework can overcome the problems of data scarcity in case of extreme value models, by combining information from different groups [172]. The Bayesian analysis outputs the

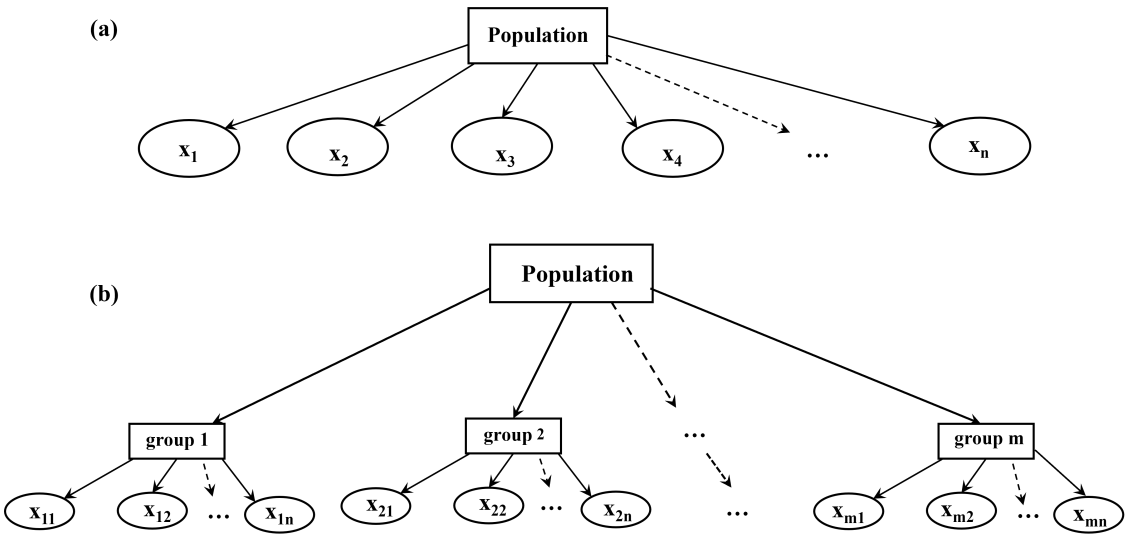

Fig. 3.1 Framework for (a) pooled and (b) hierarchical models

distribution of parameters called as posterior distribution instead of point estimates, as in the case of maximum likelihood estimation. Therefore, the uncertainty in measurements and estimation of parameters is available for a more complete inference. The Bayesian model enables us to use the prior data from similar studies, making model inference more accurate.