TRAFFIC SAFETY ANALYSIS BY SURROGATE MEASURES

AN EXTREME VALUE APPROACH

HEIDI MACH

Master's thesis
2022:E12

LUND UNIVERSITY

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

Master's Theses in Mathematical Sciences 2022:E12
ISSN 1404-6342

LUNFMS-3107-2022

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/

Populärvetenskapligt sammanfattning

Trafiksäkerhetsanalys är nödvändigt för förebyggandet av dödsolyckor i trafiken. I Europa, är ambitionen "Nollvisionen", det vill säga, ingen ska skadas eller dö i en trafikolycka. I rapporten används olika distans mellan två fordon i en interaktion för att uppskatta risken för krock eller nära till att krockas. Däremot, är distansen ett mått på hur nära bilarna är varandra, därför går det inte att avgöra om situationen är allvarlig eller inte.

Data som användes är samlat av en forskningsgrupp hos avdelningen Trafik och väg på Institutionen för Teknik och samhälle vid Lunds tekniska högskola. Lunds universitet. Sammanlagt innehåller datan 1512 observationer av en fyrvägskorsning. De situationer som analyserades var de med två bilar involverade på en vägbana; en bil som kör rakt och den andra svänger vänster. Två olika distanser registrerades vid olika moment. Den ena momentet är den minsta distansen mellan bilarna i korsningen. Den andra är första momentet då Post Enchroachment Time beräknas (PET). det vill säga, momentet då första bilen lämnar den andra bilens bana.

Syftet med rapporten är att uppskatta risken att krockas med hjälp av Extreme Värdes Teori (EVT) applicerad på trafikdata. De modeller som används för att modellera data är Generalized Extreme Value (GEV) och Generlized Pareto Distribution (GPD). Metoderna de användes tillsammans med är blockmaxima och Peak Over Threshold (POT), respektive. Det bästa resultatet gavs av GPD tillsammans med POT där sannolikheten för att det blir en krock är 0.0173%.

Abstract

Road safety analyses are required for the prevention of road accident fatalities. In Europe, the ambition is "Vision Zero". Data that was used is collected by the research group Transport and Roads which is part of Department of Technology and Society at LTH, Lund University. The dataset of video-recorded traffic situations used in the study was limited to encounters in which one motor vehicle turns left at an intersection and a straight-passing vehicle approaches. Distance between the cars were registered and used as surrogate measure for the risk of collision, specifically, the Minimum Distance (MD) between the involved motor vehicles during an interaction and Post Encroachment Distance (PED). The PED is the distance computed at the moment when the first road-user leaves the lane of the second road-user. The nearness to collision is of interest, thus, the probability that distances are less than 0 need to be computed.

Modelling was done with Generalized Extreme Value Distribution (GEV) and Generalized Pareto Distribution (GPD) together with block maxima and Peak Over Threshold (POT), respectively. The model GPD yielded the best results with probability of collision being .0173%.

Acknowledgements

Writing the thesis has been challenging, but I have learnt a lot. I want to thank Nader Tajvidi for being supportive during my journey. I would also like to thank Oksana Yastremska-Kravchenko, Aliaksei Laureshyn and Carmelo D'agostino for their insights on road safety.

Contents

Chapter 1

Introduction

1.1 Importance of Traffic Safety

Every year more than 25 000 people die on the roads and more than 135 000 people are seriously injured in the countries of the European Union. The road crashes are now the most common cause of death for children and young people between 5 and 29 worldwide. The European Commission has now set together the Road Safety Policy Framework 2021-2030. Its target is to halve the number of fatalities caused by road crashes, as well as, halving the number of serious injuries by 2030. The target is part of the long-term goal of European Union, which is to move the number of deaths caused by road crashes close to zero ("Vision Zero") [1].

The results from Vision Zero have been very promising in several countries. In Sweden, the number of road accident fatality rate has decreased in the recent years and are in compliance with Sweden's Vision Zero. During 2018 the number of accident fatality was 324. Comparing it to 2020 which has in total decreased with 120 persons, corresponding to approximately 37 percent [2].

Road safety analyses are required for the prevention of road accident fatalities. Previous research has relied on crash data. However, crash data has been known to have a lot of limitations such as access to small samples of data due to the data being reactive, i.e., more crash accidents are needed which is contradicting to wanting to prevent accidents. Instead, it is more

beneficial to utilize observable non-crash events as a surrogate or a complement to crashes [3]. Thus, surrogate measure of safety is instead used, which is measure of how close the road-users are to collision. The practicality is the additional information; type of driver, type of vehicles, evasive maneuvering, type of intersection at which near-to-collision happened. It is more advantageous since the dataset available is much larger [4].

Extreme value theory (EVT) has been widely in applied science for over 50 years. The main objective of extreme value analysis is attempting to quantify the behavior of stochastic processes at extreme values, i.e., unusually large- or small values. The analyses usually require estimation of the probability of the more extreme values that haven't previously been observed [5]. The extreme value theory provides a framework for extrapolation. which is suitable for application in Traffic Road Safety. Viewing the crash accident as extreme values, appropriate measures can be used for quantifying the frequency of the crash data.

1.2 Surrogate measures

In this report different surrogate measures of road crashes are used. The different indicators that will be introduced are minimum distance, post encroachment distance and Delta-V. The former two are surrogate measures of the nearness of collision, while Delta-V are often used as an indicator of severity of collision. The indicators are defined as:

- **Minimum distance (MD)**. is the smallest distance between two vehicles in an interaction. For example, Situation 1 and 2 are moments during an interaction and the smallest distance of these two is the minimum distance (see in Figure 1.1).

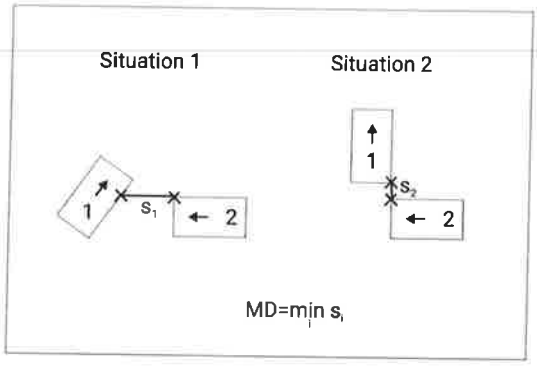

Figure 1.1: Minimum distance between two vehicles.

- **Post encroachment distance (PED)** is distance between the involved vehicles at the moment when the first vehicle (who reached the conflict area first) is leaving the area of potential collision, when PET (Post Encroachment Time) is computed. Given there is two road-users in an interaction. The PET indicator is the time between two moments; the first moment is when the first car leaves the path of the second car, and the other moment is when the second car arrives at the path of the first car (Figure 1.2) [6].

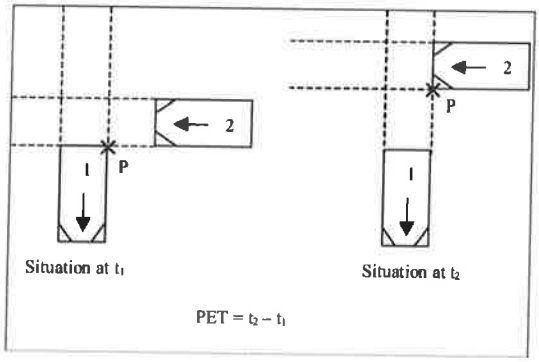

Figure 1.2: Illustration of PET [6].

- **Delta-V** (Δv), the change of velocity during a crash. Delta-V will

be defined below. Assume there are two road-users; user 1 and user 2. Let m_1 and m_2 denote their corresponding masses with v_1 and v_2 as speeds. In an elastic collision, as shown in Figure 1.3, Delta-V's can be computed for both road-users as,

$$\Delta v_1 = \frac{m_2}{m_1 + m_2} \cdot \sqrt{v_1^2 + v_2^2 - 2v_1 v_2 \cos\alpha}$$
$$\Delta v_2 = \frac{m_1}{m_1 + m_2} \cdot \sqrt{v_1^2 + v_2^2 - 2v_1 v_2 \cos\alpha}$$

where Δv_1 and Δv_2 are Delta-V's of road-user 1 and road-user 2, respectively [3].

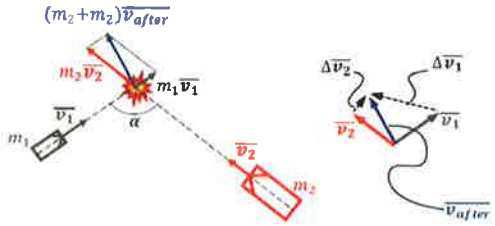

Figure 1.3: Illustration of Delta-V [3].

1.3 Problem formulation

In the recent years, EVT has been found to be useful for predicting the risk of accident in road safety analysis. Temporal surrogate measures are mostly used in road safety, but has been found to result in some contrary effects. Another surrogate measure, which does not have this problem is distance. Additionally, the distance can be measured free from the need of a collision course [7]. In this thesis, the indicators MD and PED and the corresponding Delta-V's will be analyzed with EVT to better describe road safety. The related data are collected by a research group at Transport & Roads which is apart of the Department of Technology and Society at LTH, Lund University. The purpose is to assess the risk of collision using GEV and GPD. The method used in combination with the models are block maxima and peak over threshold. The following questions that will be attempted to be answered are:

- Is there a significant correlation between MD and PED with its Delta-V's?
- Are the EVT models vaild for modelling the given safety data?

Chapter 2

Theoretical background

2.1 Extreme Value Theory (EVT)

Extreme value theory deals with extreme values in data. The two EVT families that will be introduced are GEV and GPD. In GEV, the aim is to model maxima (or minima), while GPD models excess (or excess loss). Methods that work well with these models are block maxima and POT, respectively.

2.1.1 Generalized Extreme Value Distribution (GEV)

In extreme value theory it is often of interest to analyze the statistical behaviour of

$$M_n = \max\{X_1, ..., X_n\}$$

where $n \in \mathbb{N}$ and $X_1, ..., X_n$, is a sequence of independent random variables having common distribution F. The distribution of M_n can be derived as,

$$\mathbb{P}(M_n \leq z) = \{F(z)\}^n \tag{2.1}$$

In practice, the distribution F is usually unknown. The distribution function can be estimated from observed data. The problem with this method is that small discrepancies in the estimate of F can lead to large discrepancies in F^n. Instead, if F is assumed to be unknown, a possible way to to approximate the distribution of M_n is to use approximate families of models for F^n. By usual practice, observe the behaviour of F^n for $n \to \infty$. Due to the asymptotic

distribution being degenerate, a linear renormalization of M_n is performed, given by,

$$M_n^* = \frac{M_n - b_n}{a_n}$$

where $\{a_n\}$ and $\{b_n\}$ are sequences of constants. They are chosen, such that it stabilizes the location and scale of M_n^* as n increases.

Theorem 1 *If there exist sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) \quad as \quad n \to \infty \tag{2.2}$$

for a non-degenerate distribution function G, then G is a member of one of the following families:

1. *Gumbel distribution*

$$G(z) = \exp\left\{-\exp\left\{-\left(\frac{z-b}{a}\right)\right\}\right\}, -\infty < z < \infty$$

2. *Fréchet distribution*

$$G(z) = \begin{cases} 0, & z \leq b \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b; \end{cases}$$

3. *Weibull distribution*

$$G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b; \\ 1, & z \geq b \end{cases}$$

for parameters $a > 0$, b *and, in the case of families 2. and 3.,* $\alpha > 0$.

The the families can be unified into a single family of extreme value distribution called the Generlized Extreme Value (GEV) family, given by,

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{2.3}$$

defined on $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$, *where* $-\infty < \mu < \infty, \sigma > 0$ *and* $-\infty < \xi < \infty$. *Different values of* ξ *corresponds to any of the three models mentioned above; if* $\xi > 0$ *the model corresponds to a Frechet distribution and if* $\xi < 0$ *Weibull distribution and if* $\xi = 0$, *the model is the Gumbel distribution. In the case of the Gumbel distribution, it can be simplified to,*

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}, \quad -\infty < z < \infty. \tag{2.4}$$

The GEV is commonly used in extreme value theory [5].

2.1.2 GEV fitted to negative data

Instead of being interested in the behavior of the maxima, the minima is sometimes a more interesting case. Let $\tilde{M}_n = \min\{X_1, ..., X_n\}$, where X_i denotes the independent and identically distributed random variables. Let $Y_i = -X_i$ for $i = 1, ..., n$, the small values of X_i now corresponds to the large values of Y_i. If $\tilde{M}_n = \min\{X_1, ...X_n\}$ and $M_n = \max\{Y_1, ..., Y_n\}$, then $\tilde{M}_n = -M_n$. The proof is given by the following,

$$\begin{aligned}
\Pr\{\tilde{M}_n \leq z\} &= \Pr\{-M_n \leq z\} \\
&= \Pr\{M_n \geq -z\} \\
&= 1 - \Pr\{M_n \leq -z\} \\
&\approx 1 - \exp\left\{-\left[1 + \xi\left(\frac{-z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \\
&= 1 - \exp\left\{-\left[1 - \xi\left(\frac{z-\tilde{\mu}}{\sigma}\right)\right]^{-1/\xi}\right\}
\end{aligned}$$

on $\{z : 1 - \xi(z - \tilde{\mu})/\sigma > 0\}$, where $\tilde{\mu} = -\mu$. The derived distribution is called the GEV distibution for minima. It can be restated formally as the following theorem [5].

Theorem 2 *If there exist sequencese of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$\Pr\{(\tilde{M}_n - b_n)/a_n \leq z\} \to \tilde{G}(z), \text{ as } n \to \infty$$

for a non-degenerate distribution function $\tilde{G}$, then $\tilde{G}$ is a member of the GEV family of distribution of minima:

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

defined on $\{z : 1 - \xi(z - \tilde{\mu})/\sigma > 0\}$, where $-\infty < \mu < \infty$, $\tilde{\sigma} > 0$ and $-\infty < \xi < \infty$ [5].

2.1.3 Block Maxima Approch

The method that are usually used together with GEV are block maxima or block minima. In the case block maxima, let n be the size of the dataset. The data can be blocked into k blocks with block size $m = \frac{n}{k}$.
In each block the maximum will be taken, resulting in the sequence

$$M_{(k,1)}, ..., M_{(k,k)}.$$

The sequence will then be fitted to a suitable GEV distribution. The procedure for block minima is similar, but the minimum will be taken in each block given as,

$$m_{(k,1)}, ..., m_{(k,k)}.$$

which is equivalent to,

$$-M_{(k,1)}, ..., -M_{(k,k)}[5].$$

2.1.4 Generalized Pareto Distribution

Another distribution that is not as wasteful as GEV is the Generlized Pareto distribution (GPD). The distribution is considers the random variable X over a chosen threshold u, i.e., $(X - u|X > u)$.

Theorem 3 (Generalized Pareto Distribution) *Let $X_1, ..., X_2$ be a sequence of independent random variables with common distribution function F, and let*

$$M_n = \max\{X_1, ..., X_n\}.$$

Denote an arbitrary term in the X_i sequence by X, and suppose that F satisfies the equation 1, so that for large n,

$$P\{M_n \leq z\} \approx G(z)$$

where

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

for some $\mu, \sigma > 0$ and ξ. Then, for large enough u, the distribution function of $(X - u)$, conditional on $X > u$, is approximately

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

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\tilde{\sigma} > 0)\}$, where $\tilde{\sigma} = \sigma + \xi(\mu - \sigma)$ [5].

2.1.5 Peak over threshold (POT)

A framework that are usually used together with GPD is the Peak Over Threshold method. In this method the extremes are chosen with regards to a given high threshold. Given independently and identically distributed measurments $x_1, ..., x_n$. The method considers the extremes events to be chosen according to a high threshold u, such that we the exceedences are $\{x_i : x_i > u\}$. Labelling the exceedences by $x_1, ..., x_k$, where k is the number of exceedences, the threshold excesses can be defined by $y_j = x_j - u$, for $j = 1, ... k$. In this method choosing the threshold is crucial. Too low a threshold will lead to the violation of the asymptotic assumption, thus, resulting in high bias, while a too high threshold will result in the model using too few data and causing high variance instead. In this report, the two methods used for choosing a suitable threshold are the mean residual plot and plotting the parameter estimates against different thresholds.

Mean residual plot

The mean residual plot is created by plotting the mean of the GPD. Assume that GPD is a valid model for the excess over a chosen threshold u_0 derived from the following stochastic sequence $X_1, ..., X_n$. The mean of the GPD is given by,

$$E(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1 - \xi}$$

The assumption that GPD is a valid model for the exceedences over u_0, indicate that it should also be valid for the thresholds $U > U_0$. If $\Sigma_u = \sigma_{u_0} + \xi u$, then

$$E(X - u | X > u) = \frac{\sigma_u}{1 - \xi} \tag{2.5}$$

$$= \frac{\sigma_{u_0} + \xi u}{1 - \xi} \tag{2.6}$$

Therefore, for $u > u_0$ the conditional probability $E(X - u | X > u)$ is a linear function of u and also the mean. Plotting the locus of points:

$$\left\{ (u, \frac{1}{n_u} \sum_{i=1}^{n_u} x_{(i)} - u); u < x_{max} \right\}$$

This should make a linear function.

Parameter estimates against different thresholds

The shape parameter ξ is independent, while the scale parameter σ_u is dependent on u and denoted by:

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

After the reparameterization, the scale parameter is written as:

$$\sigma^* = \sigma - \xi u$$

and is constant with respect to u. Accordingly, the estimates ξ and σ^* should be constant above u_0 if u_0 is a suitable threshold [5].

2.1.6 Conditional probability to fitted negative data

Instead of excess, For $i = 1, ..., n$, let $\tilde{X}_i = -X_i$ be a sequence of independent stochastic variables. Given a threshold $\tilde{u} = -u$, for $\tilde{u} < 0$. Fitting GPD to $\tilde{X} - \tilde{u}$, we have,

$$P(\tilde{X} > \tilde{u} + x | \tilde{X} > \tilde{u}) \sim GPD(\sigma, \xi), x > \tilde{u},$$

which can be rewritten:

$$P(\tilde{X} > x) = \frac{1 - F(\tilde{u} + x)}{1 - F(\tilde{u})}$$

where F is the unknown distribution function. Knowing the distribution of $\tilde{X}$, we have that,

$$P(\tilde{X} > x) = \eta_{\tilde{u}} \left[1 + \xi \left(\frac{x - \tilde{u}}{\sigma}\right)\right]^{-1/\xi} \quad x > \tilde{u}$$

where

$$\eta_{\tilde{u}} = P(\tilde{X} > \tilde{u}) = P(-X > -u) = P(X < u)$$

which can be seen is the same number of exceedences over the threshold $\tilde{u}$ for $\tilde{X}$. It follows from $P(\tilde{X} > x)$ that

$$P(-X > x) = \eta_{\tilde{u}} \left[1 + \xi \left(\frac{x - (-u)}{\sigma}\right)\right]^{-1/\xi}, \qquad 0 < x < u \qquad (2.7)$$

$$P(X < -x) = \eta_{\tilde{u}} \left[1 + \xi \left(\frac{u - (-x)}{\sigma}\right)\right]^{-1/\xi}, \qquad -x < u \qquad (2.8)$$

$$P(X < y) = \eta_{\tilde{u}} \left[1 + \xi \left(\frac{u - y}{\sigma}\right)\right]^{-1/\xi} \qquad y < u \qquad (2.9)$$

for $0 < y < u$ [5].

2.1.7 Lower- and upper endpoint

In extreme value theory, the endpoint is also of interest and can tell you if there is a threshold for the data. For the GEV distribution, the endpoints are estimated by $\mu - \frac{\sigma}{|\xi|}$. the following points decides if it is a lower- or upper endpoint:

- upper endpoint if $\xi < 0$.
- lower endpoint if $\xi > 0$.

This leads to violation of the standard asymptotic behavior and leads to the following results:

- maximum likelihood has usual asymptotic properties for $\xi > -0.5$
- for $-1 < \xi < -0.5$ maximum likelihood estimators are possible to estimate, however does not have the standard asymptotic behavior
- for $\xi < -1$ the estimators are most probably not possible to estimate

On the other hand, GPD has an upper bound, $\sigma/|\xi|$ over the threshold $u > 0$. Indicating that the original variable has an upper endpoint $\mu + \sigma/|\xi|$. However, there is no endpoint for $\xi > 0$. By Equation (2.9), if $\xi < 0$, then it holds that

$$P(X < u - y) = P(X < u)\left[1 - |\xi|\frac{y}{\sigma}\right]^{1/|\xi|} \tag{2.10}$$

for $0 < y < \sigma/|\xi|$. Reforming $u - y = z$ to $y = u - z$, Equation (2.10) can be rewritten as

$$P(X < z) = P(X < z)\left[1 - |\xi|\frac{u - z}{\sigma}\right]^{1/|\xi|}$$

for $0 < u - z < \sigma/|\xi|$. The lower endpoint is given by

$$u - \frac{\sigma}{|\xi|}$$

since it holds that $u - \frac{\sigma}{|\xi|} < z < u$ [5].

2.2 Basic Statistics

This section will introduce necessary basic statistics. Kendall's tau is mentioned as a measure of linearity between two variables. The estimation of model parameters are done by using maximum likelihood method. In order to make inference, the profile likelihood and delta method was used to compute the confidence interval. Lastly, model diagnostics is introduced as a method for quantifying models.

2.2.1 Kendall's tau

In order to see if there are any linearity involved between variables, the Kendall's tau was chosen. In particular, the measure is more robust to extreme values and to non-linear data. Assume X and Y are random variables sampled from a bivariate distribution, Kendall's tau is defined as

$$\tau_{XY} = \mathrm{E}\left\{\mathrm{sign}(X_1 - X_2)\mathrm{sign}(Y_1 - Y_2)\right\}$$

where (X_1, Y_1) and (X_2, Y_2) pairs of variables sampled independently from the same population [8].

2.2.2 Maximum Likelihood Estimation

A general estimation of an unknown parameter θ is the maximum likelihood. Let $x_1, ..., x_n$ be independent observations of a random variable X with probability density function $f(x;\theta)$. The maximum likelihood method is based on the likelihood function and defined as,

$$L(\theta) = \prod_{i=1}^{n} f(x;\theta).$$

The maximum likelihood estimator of θ is the one that result in maximizing the likelihood function, i.e.,

$$\tilde{\theta} = \max_{\theta \in \Theta} L(\theta)$$

where Θ is the parameter space. By convenience the log-likelihood function is usually used to compute the estimator. It is computed as,

$$l(\theta) = \sum_{i=1}^{n} \log f(x;\theta).$$

It can already be seen that the log-likelihood is a sum, which is much easier to compute than a product [5].

2.2.3 Profile Likelihood

In inference, another method is the profile likelihood. Defining the profile log-likelihood for θ_i, we have,

$$\ell_p(\theta_i) = \max_{\theta_i} \ell(\theta_i, \theta_{-i}).$$

where $\ell(\theta_i, \theta_{-i})$ is the log-likelihood of θ_{-i}, which is the set including all θ excluding θ_i [5].

2.2.4 Delta Method

In some situations, a parameter θ_0 might not be of interest, and instead some function $\theta_0 = g(\theta_0)$ that also has a different dimension. Let θ_0 be a d-dimensional parameter with approximate variance-covariance matrix V_0. Given an estimator and the large-sample maximum likelihood, which will be denoted $\hat{\theta}_0$. Assuming $\phi = g(\theta)$ is a scalar function, then the maximum likelihood estimator of $\phi_0 = g(\theta_0)$ satisfies

$$\hat{\phi}_0 \sim \mathrm{N}(\phi_0, V_\phi)$$

where

$$V_\phi = \nabla\phi^T V_\theta \nabla\phi$$

with

$$\nabla\phi = \left[\frac{\partial\phi}{\partial\theta_1}, ..., \frac{\partial\phi}{\partial\theta_d}\right]^T$$

evaluated at $\hat{\theta}_0$ [5].

2.2.5 Confidence interval

In inference theory, a common practise is to compute the confidence interval. Let θ be a parameter belonging to the parameter space Θ. Denote the confidence interval, $I_\theta(X)$, with a confidence level $1-\alpha$. Let $x_1, ..., x_n$ be the observation of a random variable $X \sim N(\mu, \sigma^2)$, with parameters μ and σ. Computing $\hat{\sigma}$ as an estimator of σ, the confidence interval is given computed as,

$$I_\mu = \left(\hat{x} \pm \lambda_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right).$$

where $\hat{x}$ is the estimated mean, λ some normal quantile and α the sigificant level [9].

2.2.6 Model Diagonostics

There are different graphical techniques for evaluating the model fit. In this thesis the two that will mention are the probability plot and the quantile plot. Let the ordered sample of the independent observations $x_1, x_2, ..., x_n$ be,

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

with an unknown distribution function F. The estimated distribution function will be denoted $\tilde{F}$. Plotting the probability plot for $\tilde{F}$, we have the plot:

$$\left\{ \left(\tilde{F}(x_{(i)}), \left(\frac{i}{n+1} \right) \right) : i = 1, ..., n \right\}$$

and the quantile plot are formed by

$$\left\{ \left(\tilde{F}^{-1} \left(\frac{i}{n+1} \right), x_{(i)} \right) : i = 1, ..., n \right\}.$$

Assuming that $\tilde{F}$ is a reasonable estimation, the quantile plot should approximately be a unit diagonal [5].

Chapter 3

Data Analysis

3.1 Data Description

The dataset comes from the research group Transport and Roads which is part of Department of Technology and Society at LTH, Lund University. Records from year 2010 of a regular signalized intersection with two-phases in Minsk for two days (between 6AM and 9PM) was used to produce the dataset. Traffic interactions of two vehicle on one lane were registered. Specifically, the situations where there are two road users driving in the opposite direction and one wants to turn left and the other is driving straight. Each road user corresponds to a data point in the dataset, and in total the number of observations is 1512. Note that no accidents occurred during the recording. Computation of MD and PED was conducted by Transport and Roads using the program Pyramid. Minimum distance was computed at a moment when distance (pink line) between two vehicles is the smallest during the interaction of two vehicles (Figure 3.1). The PED is computed at the first moment of PET. In this case it is the red car leaving the dotted yellow trajection of the yellow truck (Figure 3.2). At these two different moments the Delta-V's are computed and denoted Delta-V_{MD} and Delta-V_{PED}.

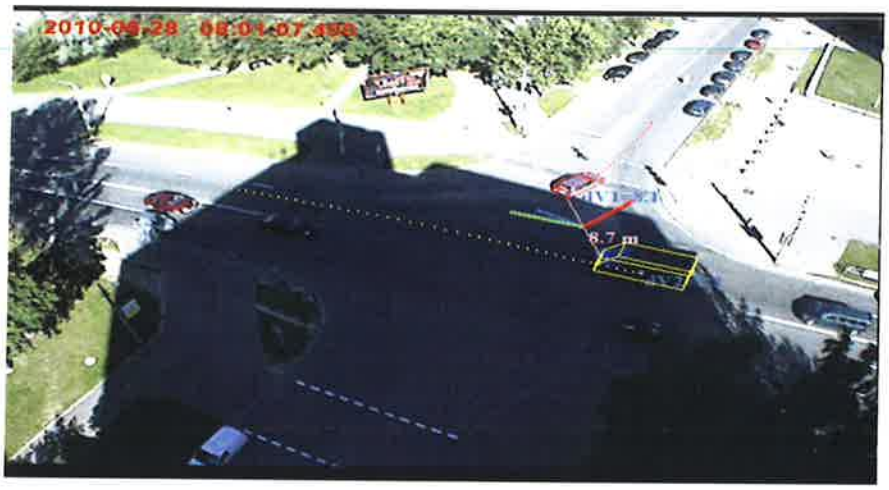

Figure 3.1: Minimum distance (MD).

Figure 3.2: Post encroachment distance (PED).

Scatter plots and histograms of -MD and -PED are shown in Figure 3.3. The descriptive statistics of Delta-V_{MD}, MD, Delta-V_{PED} and PED are shown in Table 3.1. According to the table, MD has a minimum which is equal to zero (and also the only one in the dataset). This is caused by some numerical approximation by the program. However, it is also important to be reminded that the minimum distance is an indicator of the nearness to collision and does not imply anything about the severity of a collision. Thus, for minimum distances equal to zero, an accident might not have happened since the severity might be low. It can be noted that the mean of PED is higher than the mean of MD, which could indicate that the values are generally higher in PED.

Table 3.1: Descriptive Statistic

Indicator	MD	Delta-V_{MD}	PED	Delta-V_{PED}
Min	0	0.56	1.76	0.54
Max	40.72	16.19	95.14	20.93
Mean	12.65	7.57	37.33	8.75
Median	10.99	7.84	34.4	7.3
Stdev	7.66	2.8	18.63	4.56
Skewness	0.97	-0.13	0.42	0.45
Kurtosis	0.74	0.02	-0.61	-1.08

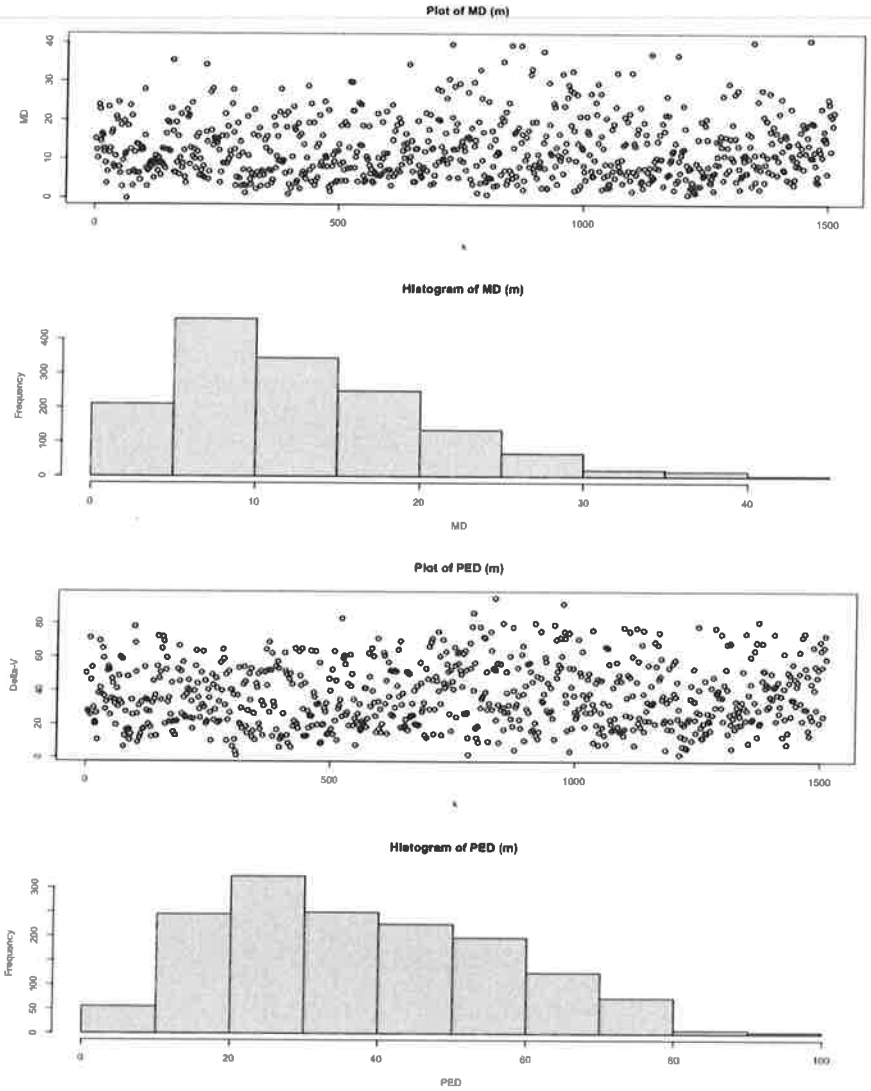

Figure 3.3: Scatter plots and histograms of MD and PED.

Intuitively, the indicators minimum distance and Delta-V_{MD} should be correlated. This means that a binary model might be appropriate. The closer the vehicles are to one another, the lower the velocities since the drivers tend to brake to avoid a potential car crash. In other words, lower minimum distance should result in lower Delta-V_{MD}. Moreover, drivers were separated into drivers going straight and left turning drivers (Straight and Left drivers) to see if there was a difference. Observing the Delta-V_{MD} plotted against -MD for all the different groups in Figure 3.4, it can be noticed that most of the points are concentrated between the distances -20m and 0m, and between the Delta-V_{MD}'s 0m/s and 10m/s. There does seem to be some kind of pattern in data, and there are some outliers at around -MD equal to -40m. The Kendall's correlation deemed appropriate to compute. which is shown in Table 3.2. All the p-values indicate that there is a significant negative correlation for all drivers, and the groups Straight- and Left drivers. Left drivers had higher absolute value of the correlation with its 0.269 as compared to 0.0786 for Straight drivers, which could be due to Left drivers breaking more when turning at the intersection. This implies that a binary model indeed seems to be a suitable model for this dataset.

Table 3.2: Kendall's rank correlation tau for -MD and Delta-V_{MD}

Type driver	τ	p-value
Straight & Left	-0.172	$< 2.2 \cdot 10^{-16}$
Left	-0.269	$< 2.2 \cdot 10^{-16}$
Straight	-0.080	0.00121

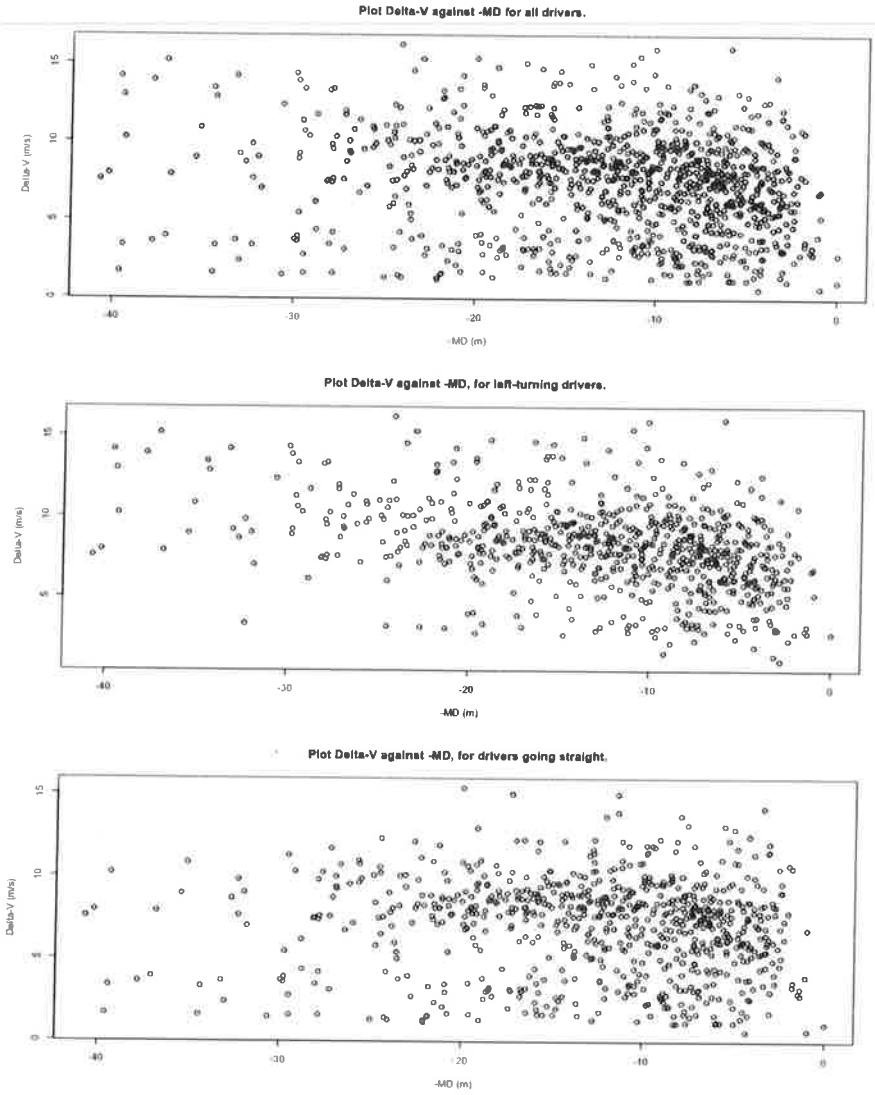

Figure 3.4: Delta-V_{MD} plotted against -MD.

For the -PED and Delta-V_{PED}, the Straight drivers a linear pattern could be noticed (see Figure 3.5) and the correlation is significant with negative value equal to -0.324 according to Table 3.3. An explanation to left drivers not having a significant correlation could be due to the fact that most of the velocities are being registered when the drivers are leaving the collision course at a low speed. Thus, binary model for Straight drivers is most suitable. However, it could also be of interest to model collision for Straight & Left drivers, since there was still a correlation and Left drivers are also of interest. Even though the binary model can better assess the frequency of accidents, for this thesis it was decided that only the nearness of collision will be assessed (i.e. modelling of distance). The binary problem will be left for further research.

Table 3.3: Kendall's rank correlation tau for -PED and Delta-V_{PED}

Type driver	τ	p-value
Straight & Left	-0.135	$3.56 \cdot 10^{-15}$
Left	-0.013	0.5996
Straight	-0.324	$< 2.2 \cdot 10^{-16}$

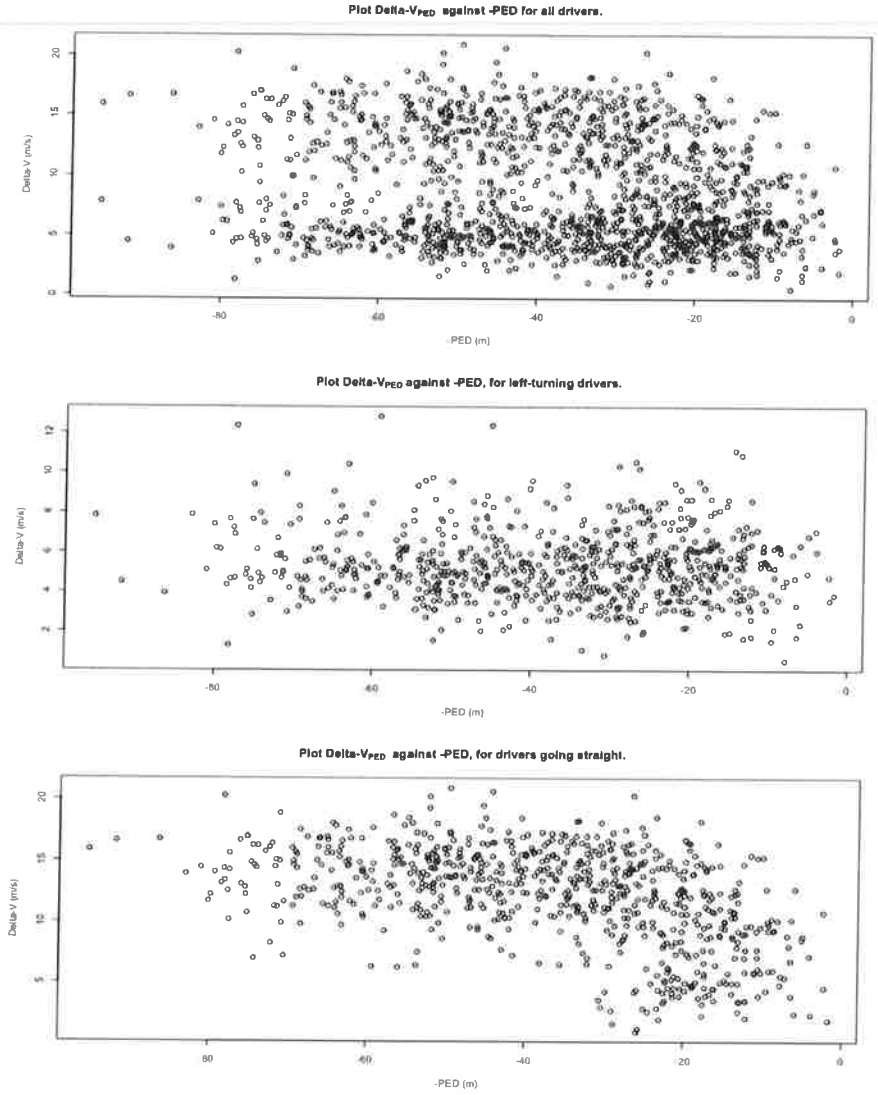

Figure 3.5: Delta-V_{PED} plotted against -PED.

3.1.1 Modelling of the road safety data

Modelling with Generalized Extreme Value Theory

Maximum block method is conducted over the negated data of the minimum distance. The number of blocks to use for the method is chosen by iterating over different block sizes and choose the one that performed well according to the diagnostic plots. Resulting in the number of blocks to be 45 for -MD and 80 for -PED. Estimation of the model parameters with a 95% confidence interval are shown in Table 3.4. In the table, the local variable μ is estimated for negated data. Observing that $\xi < 0$, there exist an upper bound. meaning that there is a lower endpoint of the original data computed by.

$$\tilde{\mu} - \frac{\sigma}{|\xi|}$$

where $\tilde{\mu} = -\mu$. Using the values from Table 3.4, the lower point of MD and PED are computed as -1.278 and -0.168. respectively, which means that the distribution covers all the observations.

Table 3.4: Parameters of GEV.

Indicator		LB	MLE	UB	End point	#Blocks
-MD	μ	-3.150	-3.563	-3.975	1.278	45
	σ	0.987	1.279	1.571		
	ξ	-0.453	-0.264	-0.075		
-PED	μ	-16.571	-15.074	-13.576	0.168	80
	σ	5.181	6.266	7.350		
	ξ	-0.547	-0.411	-0.275		

Diagnostic plots are shown in Figure 3.6. As can be seen the quantile plots and the probability plots for both indicators are well aligned on the diagonal line, indicating that the GEV distribution might be appropriate for this dataset.

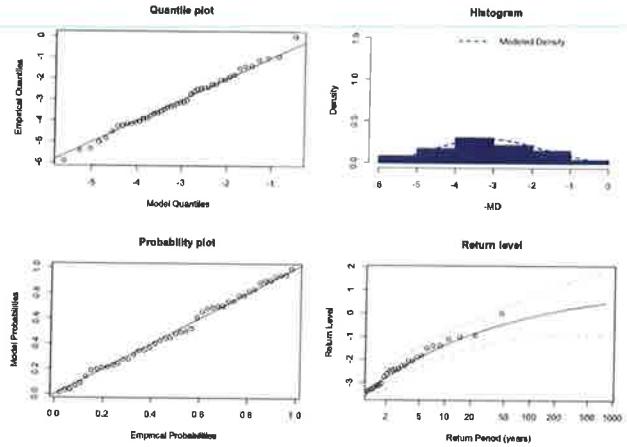

Figure 3.6: Diagnostics for GEV fitted to -MD.

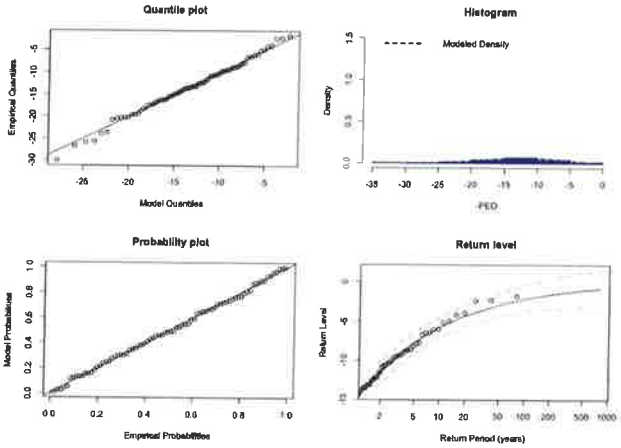

Figure 3.7: Diagnostics for GEV fitted to blocked -PED.

3.1.2 Modelling with Generalized Pareto Distribution

An important part of the GPD is the threshold selection. The mean residual plot (MRL) of -MD and -PED (Figure 3.8 and Figure 3.9) shows that the graph is curved until approximately -5 and -15, and that there might be some linearity for larger values of the negated data. The parameter threshold plots (Figure 3.10 and Figure 3.11) further convinces us that a threshold of -5 and -15 are appropriate choices. It is also important that the lower endpoint is close to the smallest datapoint in the dataset.

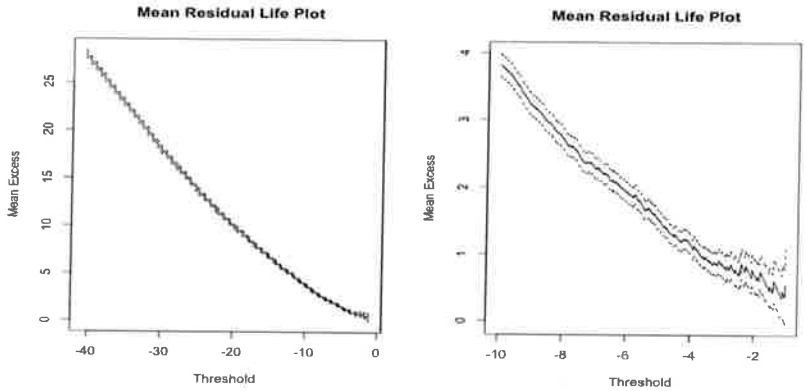

Figure 3.8: Mean residual plot for -MD. The left plot is over all threshold and the right plot is a zoomed in version over a smaller interval.

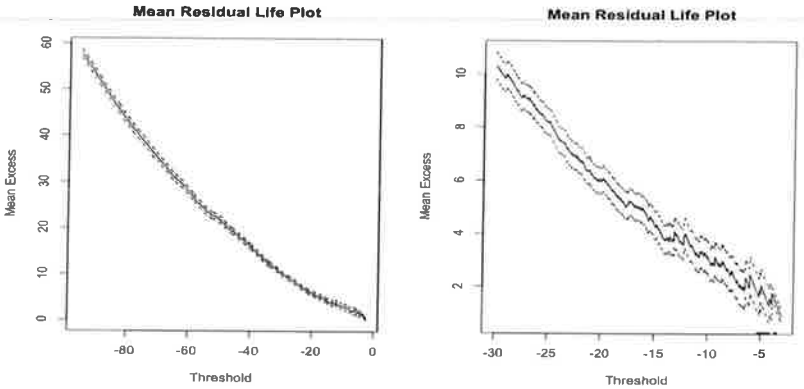

Figure 3.9: Mean residual plot -PED. The left plot is over all threshold and the right plot is a zoomed in version over a smaller interval.

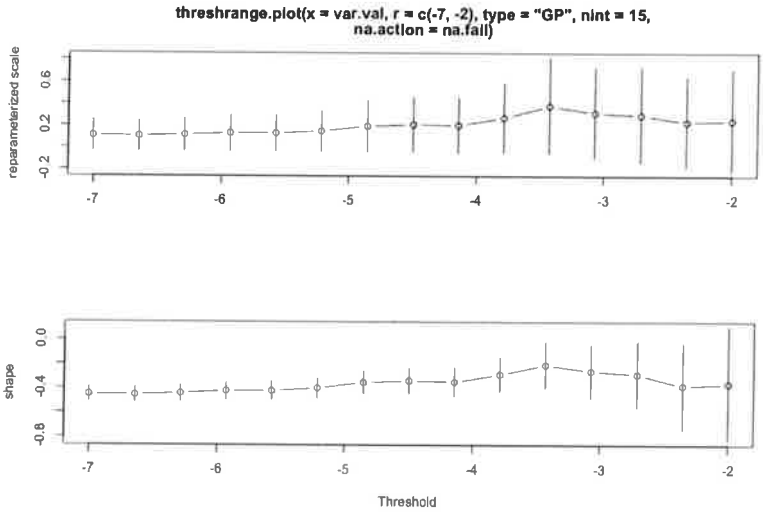

Figure 3.10: Parameter thresholds for -MD.

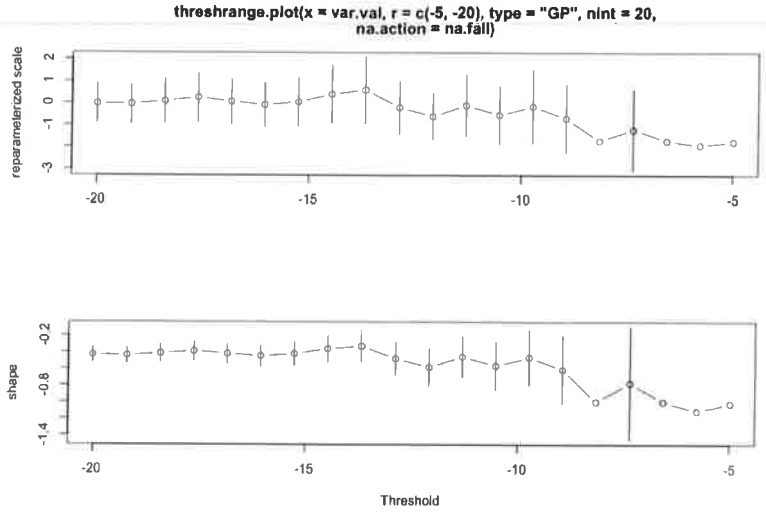

Figure 3.11: Parameter thresholds for -PED.

The diagnostic plots of the chosen model is shown in Figure 3.12. The quantile and probability plot seem to be aligned well on the diagonal line, which indicate that the model is appropriate. Estimating the parameters of the GPD model, the values are shown in Table 3.5. Furthermore, computing the lower endpoint from the table results in the value -0.420 for MD and -0.478 for PED, which covers all the values of the observations.

Table 3.5: Parameters of GPD.

Indicator		LB	MLE	UB	End point	Threshold
-MD	σ	1.757	2.069	2.381	0.420	-5
	ξ	-0.467	-0.381	-0.296		
-PET	σ	4.804	6.044	7.283	0.478	-15
	ξ	-0.538	-0.390	-0.243		

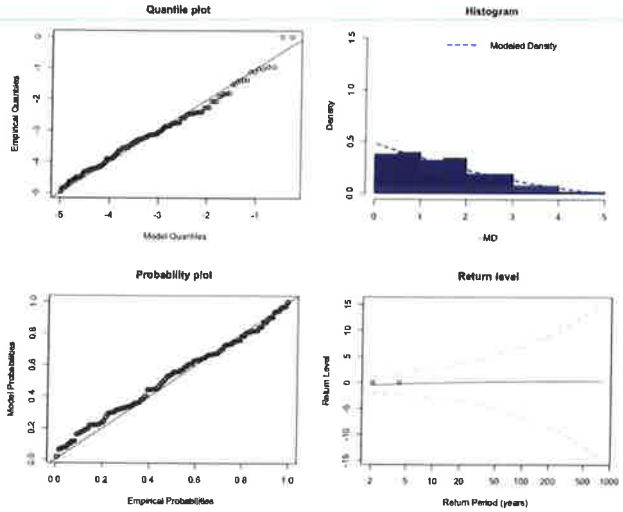

Figure 3.12: Diagnostics for GPD fitted to -MD.

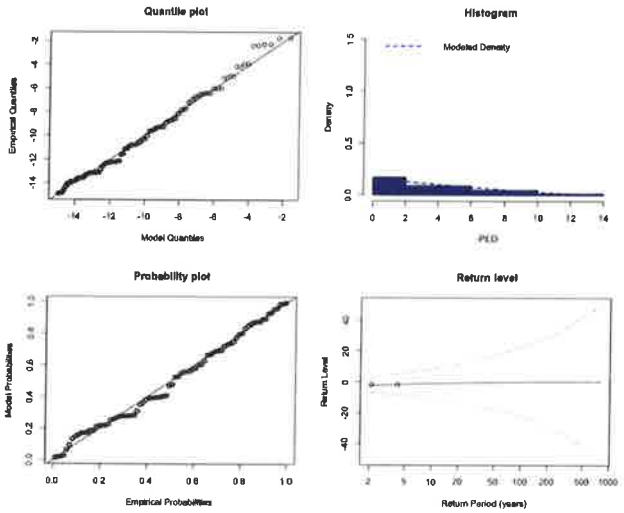

Figure 3.13: Diagnostics for GPD fitted to -PED.

3.1.3 Results

The nearness of collision could be described by the probability $P(Z \leq 0)$, where Z is a random variable for distance. In this section, the results are generated using the different models. Computations of a 95% confidence using delta method and profile likelihood are presented. The delta method is quite straight forward, on the other hand, the profile likelihood is a bit more complicated and also requires some optimization algorithms.

Confidence interval using delta method for GEV

For $\xi \neq 0$ the GEV distribution for excess loss is computed as

$$P(Z \leq z) = \exp\left\{-\left[1-\xi\left(\frac{z-\tilde{\mu}}{\sigma}\right)\right]^{-1/\xi}\right\}$$

on $\{z : 1-\xi(z-\mu)/\sigma > 0\}$. Let $p(\theta) = \exp\{-g(\theta)\}$, where $g(\theta) = \left[1-\xi\left(\frac{z-\tilde{\mu}}{\sigma}\right)\right]^{-1/\xi}$ and $\theta = (\mu, \sigma, \xi)$.

$$\begin{aligned}\Delta p &= \left[\frac{\delta p}{\delta \mu}, \frac{\delta p}{\delta \sigma}, \frac{\delta p}{\delta \xi}\right] \\ &= \left[\frac{\delta p}{\delta g}\frac{\delta g}{\delta \mu}, \frac{\delta p}{\delta g}\frac{\delta g}{\delta \sigma}, \frac{\delta p}{\delta g}\frac{\delta g}{\delta \xi}\right]\end{aligned}$$

By some computation the derivatives are given by.

$$\begin{aligned}\frac{\delta p}{\delta g} &= \exp\left(g(\theta)\right) \\ \frac{\delta g}{\delta \tilde{\mu}} &= \left(-\frac{1}{\sigma}\right)\left(g(\theta)\right)^{\xi+1} \\ \frac{\delta g}{\delta \sigma} &= \left(-\frac{z-\tilde{\mu}}{\sigma^2}\right)\left(g(\theta)\right)^{\xi+1} \\ \frac{\delta g}{\delta \xi} &= \left[\frac{1}{\xi^2}\ln\left[1-\xi\left(\frac{z-\tilde{\mu}}{\sigma}\right)\right] + \frac{\sigma}{\xi(z-\tilde{\mu})}\right] g(\theta)\end{aligned}$$

Using the delta method we have that.

$$V(p) = \Delta p^T V_\theta \Delta p$$

where V is the covariance matrix for the parameters. A 95% confidence interval was computed for the point estimate, $P(\tilde{M}_n \leq 0)$, for the GEV models shown in Table 3.6. The delta method produced negative lower bound, which could mean that the log-likelihood function is skewed and certainly not suitable for this case.

Table 3.6: Confidence intervals using delta method are in scale 10^{-2}.

Indicator	LB	$P(\tilde{M}_n < 0)$	UB
MD	-2.514	0.644	3.803
PED	-0.081	0.00174	0.0848

Confidence interval using profile-likelihood method for GEV

For $\xi \neq 0$, GEV fitted to a minima is computed by,

$$P(\tilde{M}_n \leq z) = 1 - \exp\left\{-\left[1 - \xi\left(\frac{z-\tilde{\mu}}{\sigma}\right)\right]^{-1/\xi}\right\}$$

on $\{z : 1-\xi(z-\tilde{\mu})/\sigma > 0\}$. Differentiating the distribution function we have the following density function,

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

Computing the log-likelihood function,

$$\begin{aligned}\ell(z;\mu,\sigma,\xi) &= \sum_{i=1}^{m} \log f(z_i) \\ &= -m\log\sigma - (1+1/\xi)\sum_{i=1}^{m}\log\left[1+\xi\left(\frac{z_i-\tilde{\mu}}{\sigma}\right)\right] \\ &\quad -\sum_{i=1}^{m}\left[1+\xi\left(\frac{z_i-\tilde{\mu}}{\sigma}\right)\right]^{-1/\xi}\end{aligned} \tag{3.1}$$

provided that

$$1+\xi\left(\frac{z_i-\tilde{\mu}}{\sigma}\right) > 0, \text{for } i = 1,..,m$$

Let $p = P(\tilde{M}_n \leq z)$ and solving for μ, we have

$$\tilde{\mu} = \frac{\sigma}{\xi}\left(1 - [-\log(1-p_0)]^{-\xi}\right). \tag{3.2}$$

Replacing $\tilde{\mu}$ in (3.1) with (3.2), the new log-likelihood is now dependent on the parameters p, σ and ξ and can be expressed as $\ell(z; p, \sigma, \xi)$. The parameters are optimized using existing optimizing algorithms in R. Using this method a confidence interval was computed for $p_0 = P(\tilde{M}_n \leq 0)$. The profile log-likelihood for the model is shown in Figure 3.14. The two red intersections correspond to a confidence interval for p. The function used to plot this graph is in Appendix A.1.1.

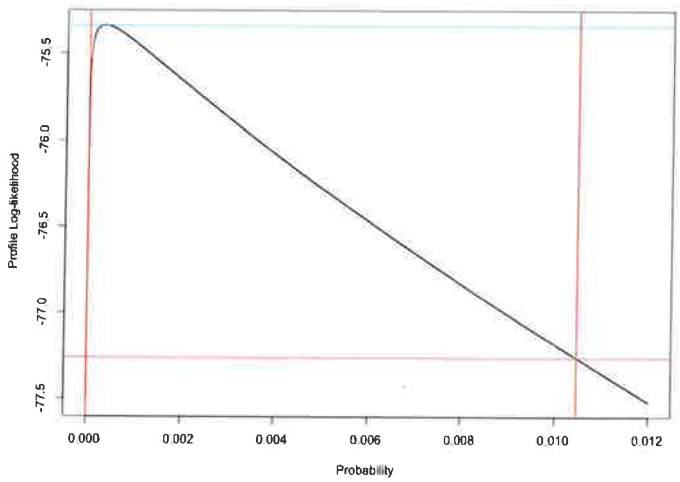

Figure 3.14: Profile likelihood of GEV. The blue horizontal line is the maximum of the graph and the red horizontal line is χ^2_1. The two vertical line intersecting χ^2_1 correspond to a 95% confidence interval.

Observing the profile likelihood plot, the distribution is skewed and the MLE is closer to the lower bound. The resulting confidence interval is shown in Table 3.7. Comparing to the delta method the lower bound is positive, which means the MLE is significant. Furthermore, it is non-zero, which

might provide more information for future predictions. The table shows the confidence interval for the point estimate using PED. For this case it is a bit problematic since the lower bound is 0. It is unlikely that future predictions are zero. An explanation is that GEV models might not be suitable for this indicator.

Table 3.7: Confidence intervals using profile likelihood method are in scale 10^{-2}.

Indicator	LB	$P(\tilde{M}_n < 0)$	UB
MD	0.00646	0.644	5.091
PED	0	0.00174	0.0150

Confidence interval using delta method for GPD

Let $\theta = (\sigma, \xi)$ and

$$g(y;\theta) = \eta_u \left[1 + \xi\left(\frac{u-y}{\sigma}\right)\right]^{-1/\xi}$$

Differentiating g with respect to the parameters we have

$$\frac{\delta g}{\delta \sigma} = \frac{\eta_u(u-y)}{\sigma^2}\left[1+\xi\left(\frac{u-y}{\sigma}\right)\right]^{-1/\xi-1}$$
$$\frac{\delta g}{\delta \xi} = \left[\frac{1}{\xi^2}\ln\left(\frac{u-y}{\sigma}\right) - \frac{\sigma}{\xi(u-y)}\right]\eta_u g(\sigma,\xi)$$

Similar to before, using the delta method we have.

$$\Delta g = \left[\frac{\delta g}{\delta \sigma}; \frac{\delta g}{\delta \xi}\right]$$

the variance computed as,

$$V(z) = \Delta g^T V_\theta \Delta g$$

where V is the covariance matrix for the parameters. The resulting confidence interval for the GPD models are shown in Table 3.9.

Table 3.8: Confidence intervals using delta method are in scale 10^{-2}.

Indicator	LB	$P(\tilde{M}_n < 0)$	UB
-MD	-0.186	0.0173	0.221
-PED	-0.0578	0.00143	0.0607

It can be noticed that the confidence interval, just like GEV, also has negative lower bound. This further confirms that the delta method is not suitable to compute the confidence interval for the point estimate.

GPD: Confidence interval using profile likelihood

The minima of the GPD distribution is given by,

$$P(Z \leq z) = \eta_{\tilde{u}} \left[1 + \xi \left(\frac{u-y}{\sigma}\right)\right]^{-1/\xi}, y < u$$

let $p = P(Z \leq z)$, and solving for σ,

$$\sigma = (u-z)\xi \left(\frac{p^{-\xi}}{\eta} - 1\right)^{-1}. \tag{3.3}$$

Computing the profile log-likelihood,

$$\ell(z;\sigma,\xi) = m\left(\log z - \log \sigma\right) - (1/\xi + 1)\sum \log\left[1 + \xi\left(\frac{u-y}{\sigma}\right)\right]. \tag{3.4}$$

for $\{z : 1 + \xi\left(\frac{u-y}{\sigma}\right) > 0\}$. Replacing σ in (3.4) with (3.3), the profile log-likelihood will now be expressed as $\ell(z;\xi,p)$, i.e, it is now instead dependent on the new variable p. By usual optimization practice, the log-likelihood function will be optimized with respect to the new variables. The profile log-likelihood plot of GPD fitted to -MD is shown in Figure 3.15. Analogous to GEV, the two red intersections correspond to a 95% confidence interval. The algorithm used to plot the profile log-likelihood and compute the confidence interval is shown in Appendix A.1.2.

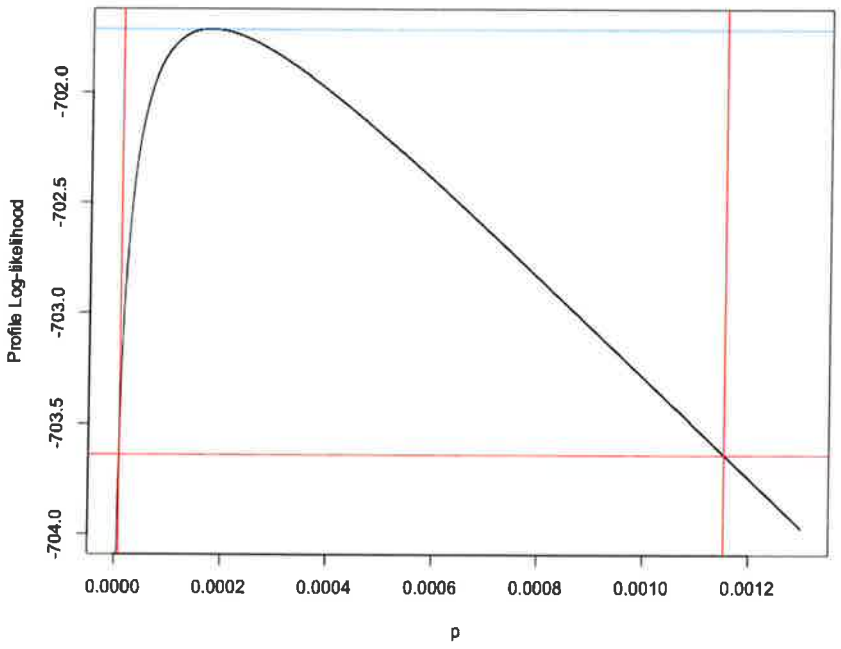

Figure 3.15: Profile likelihood of GPD. The blue horizontal line is the maximum of the graph and the red horizontal line is χ^2_1. The two vertical line intersecting χ^2_1 correspond to a 95% confidence interval.

The values of the confidence interval can be seen in Table 3.9. The lower bound of the point estimate for -MD is nonnegative and might provide information about future prediction. On the other hand the point estimate of -PED has is zero and might not be as useful.

Table 3.9: Confidence intervals using profile likelihood method are in scale 10^{-2}.

Indicator	LB	$P(\tilde{M}_n < 0)$	UB
MD	0.000698	0.0173	0.115
PED	0	0.00144	0.109

Chapter 4

Conclusions

In this thesis, the purpose was to model risk of collision in traffic using traffic safety data. The situations of interests were of encounters between two drivers in a four-way intersection. Extreme value theory was applied to the data, specifically the families Generalized extreme value distribution (GEV) and Generalized Pareto distribution (GPD). For the models different methods were used to model the data. Block maxima was used together with GEV and peak over threshold (POT) was used together with GPD.

All modelling was done over negated MD and PED. Analyzing the diagnostics plots, the GEV and GPD model provided appropriate models for analyzing the nearness of collision. Based on these models the probability of the minimas of the MD and PED less than zero were computed. Using GEV models the probabilities for MD and PED were computed as $\sim$0.644% and $\sim$0.00174%, respectively. The point estimate using PED were much lower, which seems reasonable since the series included more higher values than MD. Computing the point estimate using GPD models, the values were $\sim$0.0173% for MD and $\sim$0.00144% for PED. The point estimate for MD differed quite a lot, which means the GEV model might not be suitable for this indicator. Both GEV and GPD for the points estimate using PED were quite similar. Overall POT with GPD provided better models.

Additionally, different methods to compute the confidence interval of the point estimate were tried. The delta method resulted in negative lower bound and might not be accurate. Alternatively, the profile likelihood method was implemented. The computed confidence interval was nonnegative. In particular, the point estimate of GPD model for MD had a non-zero lower bound, which might provide information about future prediction. However, with

GPD on PED, the lower bound of the point estimate resulted in zero, implying that the point estimate might not be significantly different from zero. Intuitively, it is unlikely and to improve the results would require that more and smaller values are in the data. It can be concluded that the nearness of collision is best approximated by GPD on the MD with the probability being ~0.0173%. Whether the percentage is reasonable or not, remains to be verified with another dataset.

In this report, merely the nearness of collision was measured. It does not account for severity, which is an important aspect of road safety. For example, if two vehicles have a velocities close to zero, there might not be any implication of danger. analyzing the correlation, there were a significant correlation for both MD and PED to Delta-V. Thus. for further research, a binary model including Delta-V as a surrogate measure for severity would provide more interesting result for traffic safety.

Appendix A

R Codes

A.1 Confidence interval

The two different methods for computing the confidence interval for GEV and GPD are the delta method and profile likelihood method. All the listed codes are documented and accessible from https://github.com/machstat/evt.git.

A.1.1 Confidence interval for GEV

Delta method function

The function *gev.delta* computes a 95% confidence interval of a point estimate of GPD using the delta method.

```
gev.delta <- function(p,z,theta, V) {
  # parameters
  mu <- theta[1]
  sgm <- theta[2]
  xi <- theta[3]

  # g.theta
  g.theta <- (1-xi*((z-mu)/sgm))^(-1/xi)
  dp.g <- exp(-g.theta) # derivative of p wrt. g

  # differentiate g.theta wrt. parameters
  dg.m <- (-1/sgm)*(g.theta^(1+xi))
  dg.sgm <- (-(z-mu)/sgm^2)*(g.theta^(1+xi))
  dg.xi <- ((1/xi^2)*log(1-xi*((z-mu)/sgm))+sgm/(xi*(z-mu)))*g.theta

  # p differentiated wrt. different parameters
  dp.mu <- dg.m*dp.g
  dp.sgm <- dg.sgm*dp.g
  dp.xi <- dg.xi*dp.g

  # use delta method
  p.delta <- c(dp.mu,dp.sgm,dp.xi)
  V.p <- p.delta %*% matrix(V, nrow=3,ncol=3) %*%p.delta

  # confidence interval
  LB <- p-1.96*(sqrt(V.p))
  UB <- p+1.96*(sqrt(V.p))

  CI <- data.frame("LB"=LB,"UB"=UB, "MLE"=p)
  return(CI)
}
```

Profile method

In order to produce a confidence interval for GEV, two functions were created; *gev.pplikMinima* and *gev.pprofMinima*. The *gev.pprofMinima* uses *gev.pplikMinima* to compute the profile log-likelihood and generates profile likelihoods over a given interval.

```
gev.pplikMinima <- function(x, theta, z0, p0) {
  # initial guesses
  sgm <- theta[1]
  xi <- theta[2]
  muTilde <- z0 + sgm/xi * (( - log(1 - p0))^( - xi) - 1)

  # normalizing
  y <- (x - muTilde)/sgm
  y <- 1 - xi * y

  # negative log-likelihood
  if(is.infinite(muTilde) || sgm <= 0 || any(y <= 0))
    l <- 10^6
  else l <- length(y) * log(sgm) + sum(y^(-1/xi)) + sum(log(
    y)) * (1/xi + 1)
  return(l)
}

gev.pprofMinima <- function(x, theta0, z0, p.low, p.up, nint = 1000, conf = 0.95, pplot = 0) {
  l <- numeric(nint)
  p <- seq(p.low, p.up, length = nint)
  theta0 <- theta0[2:3]

  # optimizing algorithm for the parameters
  for(i in 1:nint) {
    p0 <- p[i]
    opt <- optim(par = theta0, fn = gev.pplikMinima, x = x, p0 = p0, z0 = z0)
    theta0 <- opt$par

    # negative log-likelihood for p
    l[i] <- opt$value
  }

  # MLE of probability
  mle.p <- p[which.max(-l)]

  # maximum value of log-likelihood function
  plik.max <- max(-l)

  # quantile of chi-square
  q <- plik.max - 0.5 * qchisq(conf, 1)

  # find the index of -q in l for lower bound (LB) and upper bound (UB)
  plik.lb <- l[l <= -q ][1] ## bigger values
  ind.lb <- match(plik.lb, l) ## 21

  plik.ub <- l[l>=-q & p>mle.p][1]
  ind.ub <- match(plik.ub, l)

  # confidence interval
  ci.lb <- p[ind.lb]
  ci.ub <- p[ind.ub]

  # plotting of the values
  if (pplot){
    plot(p, - l, type="l", xlab = "Probability", ylab =
           " Profile Log-likelihood")
    abline(h = plik.max - 0.5 * qchisq(conf, 1), col = "red")
    abline(h = plik.max, col = 4)
    abline(v=ci.lb, col="red")
    abline(v=ci.ub, col="red")
  }

  return(data.frame("LB" = ci.lb,"UB" = ci.ub, "MLE" = mle.p))
}
```

A.1.2 Confidence interval for GPD

Delta method

The function *gpd.delta* shown in the code snippet down below computes a 95% confidence interval for a point estimate of GPD using the delta method.

```
gpd.delta<- function(p, x, u, etau, theta,V) {
  # parameters
  sgm <- theta[1]
  xi <- theta[2]

  # delta method
  dg.sgm <- -((etau * (u-x))/ (xi*sgm^2)) + (1 + xi*((u-x)/sgm))^(-1/xi-1)
  dg.xi <- (1/xi^2 * log((u - x)/sgm) - (sgm / (xi * (u - x)))) * etau * (1+xi*((u-x)/sgm))^(-1/xi)
  p.delta <- c(dg.sgm, dg.xi)
  V.p <- p.delta%*%V%*%p.delta

  # confidence interval
  ci.lb <- p - 1.96*(sqrt(V.p))
  ci.ub <- p + 1.96*(sqrt(V.p))
  CI <- data.frame("LB"=ci.lb,"UB"=ci.ub)

  return(CI)
}
```

Profile likelihood

Analogous to GEV, the functions *gpd.pplikMinima* and *gpd.pprofMinima* are created to compute confidence interval for GPD.

```
# This file contains the following functions:
# gpd.pplikMinima, gpd.pprof

gpd.plikMinima <- function(x, theta, u, eta, p0, z0) {
  # initial guesses
  xi <- theta
  sgm <- ((u - z0) * theta)*((p0 / eta)^(-theta) - 1)^(-1)

  # normalizing
  y <- (u - x) / sgm
  y <- 1 + theta * y

  # negative log-likelihood
  if(any(y <= 0))
    l <- 10^6
  else l <- -(length(x) * (log(eta) - log(sgm)) - (1/theta + 1) * sum(log(y)))
  l
}

gpd.pprofMinima <- function(x, theta0, u, z0, p.low, p.up, nint = 1000, conf = 0.95, pplot = 0){
  l <- numeric(nint)
  p <- seq(p.low, p.up, length = nint)

  theta <- theta0[2]
  eta <- sum(x < u)/length(x)
  x <- x[x < u]

  # optimizing algorithm for the parameters
  for (i in 1:nint) {
    p0 <- p[i]
    opt <- optim(theta, fn=gpd.plikMinima, method = "BFGS", x = x, u = u, eta = eta, p0 = p0, z0 = z0
        )
    theta <- opt$par

    # negative log -likelihood  for p
    l[i] <- opt$value
  }

  # MLE of  probability
  mle.p <- p[which.max(-l)]

  # maximum  value  of log -likelihood  function
  plik.max <- max(-l)

  # quantile  of chi -square
  q <- plik.max- 0.5 * qchisq(conf, 1)

  # find  the  index  of -q in l for  lower  bound (LB) and  upper  bound (UB)
  plik.lb <- l[l <= -q ][1]
  ind.lb <- match(plik.lb,l)

  plik.ub <- l[l >= -q & p > mle.p][1]
  ind.ub <- match(plik.ub ,l)

  # confidence  interval
  ci.lb <- p[ind.lb]
  ci.ub <- p[ind.ub]

  # plotting  of the  values
  if(pplot) {
    plot(p,  -l, type = "l", xlab = "p", ylab =
           "Profile Log-likelihood", ylim = c(-704, plik.max))
    abline(h = q, col = "red")
    abline(h = plik.max, col = 4)
    abline(v = ci.lb, col = "red")
    abline(v = ci.ub, col = "red")
  }
  data.frame("LB" = ci.lb, "UB" = ci.ub,"MLE" = mle.p)
}
```

Bibliography

[1] Valean Adina-loana. Next steps towards 'vision zero'. *European Commission. Brussels*, 2020.

[2] Jonathan Hedlund Khabat Amin. Analys av trafiksäkerhetsutvecklingen 2020. *Trafikverket*, 2021.

[3] Aliaksei Laureshyn, Tim De Ceunynck, Christoffer Karlsson, Åse Svensson, and Stijn Daniels. In search of the severity dimension of traffic events: Extended delta-v as a traffic conflict indicator. *Accident Analysis & Prevention*, 98:46–56, 2017.

[4] Emilsson Sara and Filip Vestin. Statistical inference for traffic safety analysis using the generalized pareto distribution. *LUNFMS-3096-2020*, 2020.

[5] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*. volume 208. Springer, 2001.

[6] A Richard A Van der Horst, Maartje de Goede. Stefanie de Hair-Buijssen, and Rob Methorst. Traffic conflicts on bicycle paths: A systematic observation of behaviour from video. *Accident Analysis & Prevention*, 62:358–368, 2014.

[7] Fay Patterson. The adverse effects of paradigm and pragmatism on road safety with case studies in traffic conflicts technique and cyclist safety at roundabouts. *University Research Support Grant Scheme.*, 2020.

[8] Roger Newson. Parameters behind "nonparametric" statistics: Kendall's tau, somers' d and median differences. *The Stata Journal*, 2(1):45–64, 2002.

[9] Dragi Anevski. *A Concise Introduction to Mathematical Statistics*. Studentlitteratur., 2017.