An Extreme Value Approach to Road Safety Analysis

Johanna Lägnert

Supervisor: Nader Tajvidi

June 28, 2019

Abstract

In this thesis we study the feasibility of applying extreme value theory to data regarding road safety. In particular, we propose a model for assessing the risk of collision and near collision using extreme value theory. The thesis is relevant for road safety analysis in order to both understand whether extreme value theory is useful for modelling the collected data and to check if there is a need for collecting more data in future. Collecting this kind of information is very time consuming and expensive so efficient use of data is essential in this type of applications.

The data consists of Time to Collision (TTC) and Post-Encroachment Time (PET) for right turning vehicles against bicycles in a four-way intersection in Barcelona. The dataset is the result of a 24-hour film sequence. The modeling is done with Generalized Extreme Value distribution and Generalized Pareto distribution with block maxima and peaks over threshold method. In addition, a homogeneous Poisson process model is suggested to make predictions on the number of collisions/near collisions in a longer time frame than the observed period.

Keywords: *Time To Collision minimum, Post-Encroachment Time, Road Safety, Extreme Value Theory, block maxima, peaks over threshold, Generalized Extreme Value distribution, Generalized Pareto distribution, Poisson distribution*

Acknowledgment

I would first like to give a special thank you to my supervisor Nader Tajvidi, for all his help and commitment through this project. He has been a great support throughout the whole process.

I would also like to thank Carl Johnson for the help and support with all my questions about road safety.

Outline

In the first part, the background of the subject is presented. It includes a brief introduction to road safety in general and the concept of the risk measurements Time To Collision and Post-Encroachment Time.

The second part concerns the univariate extreme value theory. There are two distributions; Generalized Extreme Value distribution and Generalized Pareto distribution that are presented, both in general and in an adopted fashion for the intended data. Methods that are used for the two distributions are also presented, which are block maxima and peaks over threshold. The section finishes with some statistical theory, which has been used in the analysis.

Further details of the data and the applications of extreme value theory are presented in the third part. It is followed by conclusions in the fourth part and finishes with proposals for further research.

Contents

1. Introduction

Extreme Value Theory (EVT) applied to transportation engineering is a relatively new approach. One of the first applications of EVT was by *Hyde and Wright, (1986)* [12] where they estimated road traffic capacity from varied traffic flow during normal conditions. Another early application was made by *Sharma et al. (1999)* [2] where they predicted the violation of air quality standards at an urban intersection. The branch "Extreme Value Theory applied to Road Safety" has above all developed over recent years. The methods vary depending on the problem under study [8].

In studies by *Tarko (2012)* [11] and by *Zheng (2013)* [8] focus lies on the lane change maneuvers. In work by *Tarko (2012)* the intention is to fit a distribution to estimate crashes while in work by *Zheng (2013)* a comparison between the block maxima approach and peaks over threshold approach is made. When using block maxima and peaks over threshold, the predetermined sequence is negated before applying the methods.

One of the advantages when using EVT applied to road safety is that it is not necessary to use data of actual crashes which is quite rare. The ambition is to use the information of near collisions which is more common and from this information be able to estimate the risk of an actual collision.

1.1 Road Safety

Classifying traffic interaction by severity and frequency is done with a concept first introduced by *Hydén (1987)* [6]. It is a pyramid of traffic events, see Figure 1.1. The pyramid is divided into different layers, each layer representing events with similar severity and the volume of each layer representing the frequency of each severity. For example, a collision is placed at the top of the pyramid since it is both the most severe traffic event and has the lowest frequency. Two measurements that are used in road safety analysis are Time To Collision and Post-Encroachment Time. The two measurements are independent of each other and are used to model the risk of an accident.

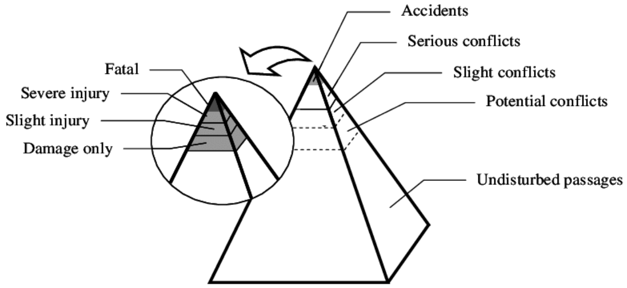


Figure 1.1: Conceptual Safety Pyramid
Source: How to analyze accident causation? A handbook with a focus on vulnerable road users [9].

The general idea with extreme value theory applied to road safety is to be able to predict the risk of an accident by analyzing the near accidents situations. It is based on the assumption that the conceptual safety pyramid is an accurate description of reality (Figure 1.1).

Post-Encroachment Time (PET) is defined as:

Definition 1. "PET is calculated as the time between the moment that the first road user leaves the path of the second and the moment that the second reaches the path of the first; in other words, PET indicates the extent to which they have missed each other" [9].

This is illustrated in Figure 1.2.

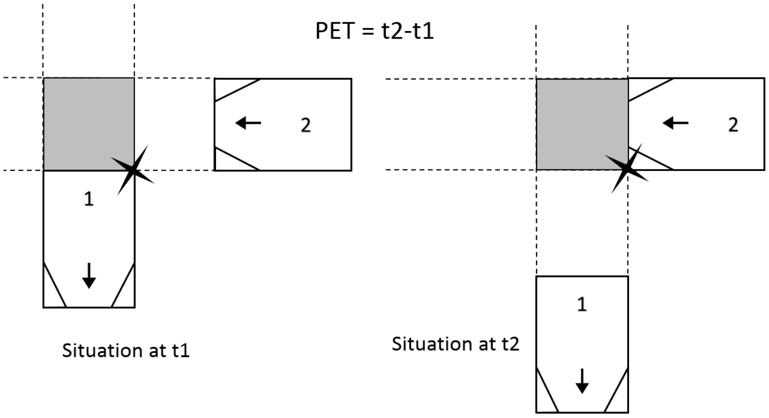


Figure 1.2: Illustration of PET

Time To Collision (TTC) is defined as:

Definition 2. "TTC is the time until a collision would occur between road users if each continued on their present course at their present rate" [9].

See Figure 1.3 for illustration of TTC. The lowest value of TTC for each event is called TTC minimum, denoted by TTCmin.

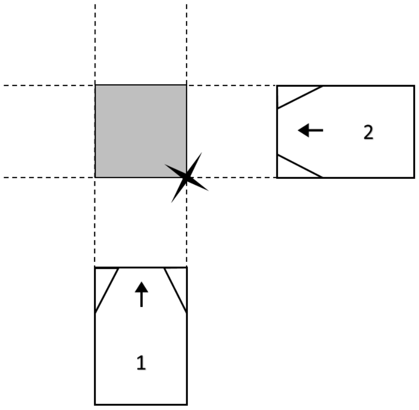


Figure 1.3: Illustration of TTC

The two measurements have a lower limit of zero seconds which is equivalent that a collision has appeared. There is no specific upper limit though larger values are of no interest. The data is generated from a film sequence from the critical area. Hence, the upper limit depends on how the cameras in the intersection are set. Are they far from the scene, higher values of PET and TTCmin are expected [7]. Values that are commonly used as critical values when classifying severity are 1.5, 2 and 3 seconds for TTCmin and values between 1 and 5 seconds for PET. These values are not scientifically determined but it is common to use them. PET and TTCmin was chosen to be the measurements used in the thesis because they are the most common ones regarding road safety [13].

1.2 Problem Formulation

At the Department of Road Safety at LTH (V-sektionen) research including the risk of accidents and near accidents is done. One way to do this is to analyze TTCmin and PET which is collected by analyzing film sequences of the critical area. This technique is costly and time consuming. During the recent years, data regarding PET and TTCmin have been gathered from an intersection in Barcelona (as part of the European research project InDev, Horizon 2020). Various parts of this data have been structured but yet been analyzed with an extreme value approach. Due to previous applications with extreme value theory on road safety, there is a desire to apply it to this dataset as model. It is however not clear if the structured data is adequate for an extreme value approach. The main question in this essay that is to be studied is:

- What is the risk of collision?

This leads to the following questions to be answered:

- Is there enough of the intended data for an application of extreme value theory?
- If so, are the fitted models, done with Generalized Extreme Value distribution, Generalized Pareto distribution and combined with Poisson distribution, a valid description of the data?
- If the models are valid for the data, do they provide a valid description of reality?

1.3 Purpose

The purpose of this thesis is to model the risk of collision and near collision with Generalized Extreme Value distribution and Generalized Pareto distribution.

This is done with block maxima and peaks over threshold on the intended data. It will also, in combination with the obtained result, model risk of accident and near accident with the Poisson process. Finally, it will evaluate if the data is sufficient for the application of the extreme value theory.

2. Theory

2.1 Extreme Value Theory

2.1.1 Generalized Extreme Value Distribution (GEV)

The model for GEV focuses on the statistical behavior of

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

where $X_1, ..., X_n$ are independent identically distributed random variables. The sequence usually consists of values from a process measured on regular time-scale. The maximum of the process over n time units of observations is denoted by M_n. Theorem 2.1 is called Extremal Types Theorem which leads to Generalized Extreme Value distribution, *Coles, (2004)* [3].

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

$$P\{(M_n - b_n)/a_n \leq z\} \rightarrow G(z), \quad \textit{as} \quad n \rightarrow \infty,$$

for a non-degenerate distribution function G, then G is a member of the GEV family

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

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

2.1.2 GEV fitted to negative data

Since the idea is to fit a distribution that describes the risk of collision, the model should focus on the statistical behavior of

$$m_n = \min\{X_1, .., X_n\}$$

where $X_1, ...X_n$ is a sequence of independent random variables having a common distribution function F. There are two different approaches when fitting a GEV distribution for minima. The first uses Theorem 2 below which is from *Coles, (2004)* [3].

Theorem 2. *If there exists sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$P\{(m_n - b_n)/a_n \leq z\} \to \tilde{G}(z), \quad as \quad n \to \infty,$$

where $\tilde{G}$ is an non-degenerate distribution function, then $\tilde{G}$ belongs to the GEV family of distributions for minima:

$$\tilde{G}(z) = 1 - exp\Big\{ - \Big[1 - \xi\Big(\frac{z - \tilde{\mu}}{\sigma}\Big)\Big]^{-1/\xi}\Big\}, \tag{2.1}$$

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

The second approach uses the relation $Y_i = -X_i$ for $i = 1, .., n$. Small values of Y_i corresponds to large values of X_i. Let

$$m_n = \min\{X_1, .., X_n\}$$
$$M_n = \max\{Y_1, .., Y_n\}$$

then it holds that:

$$\min\{X_1, .., X_n\} = -\max\{Y_1, .., Y_n\}. \tag{2.2}$$

Fitting a GEV distribution to the negative sequence generates parameters $(\hat{\mu}, \sigma, \xi)$. The scale σ and shape parameter ξ are consistent for the two methods, while a sign correction is done for location parameter due to the relation $\tilde{\mu} = -\hat{\mu}$, see [3] for further details.

2.1.3 Block maxima and Minima

Blocking the data into blocks of equal length is essential. The blocks should neither be too small or too large. If they are chosen too small the approximation by Theorem 2.1 is likely to be poor and the estimation and extrapolation will then be biased. If the blocks are chosen too large, there will be few blocks, leading to large estimation variance. Data can be blocked yearly/monthly/daily/hourly etc or by a chosen block size n resulting in a new sequence of size *m=(length of data series)/n.* The sequence

$$M_{(n,1)}, .., M_{(n,m)},$$

is fitted to a GEV distribution when block maxima is used. For Block Minima the sequence

$$m_{(n,1)}, .., m_{(n,m)}$$

is fitted to the GEV distribution. Based on Equation 2.2 it is equivalent to use

$$-M_{(n,1)}, .., -M_{(n,m)}$$

as a sequence to fit a GEV distribution. In that case it is also assumed that the relation $Y_i = -X_i$ holds and small values of Y_i corresponds to large values of X [3].

2.1.4 Generalized Pareto Distribution (GPD)

The second distribution that is used in the thesis is the Generalized Pareto distribution. Theorem 3 is the Generalized Pareto distribution Theorem from *Coles, (2004)* that focus on the statistical behaviour of $(X - u|X > u)$, [3].

Theorem 3. *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 satisfy Theorem 2.1, 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 $\{y : y > 0$ and $1 + \xi y/\tilde{\sigma}) > 0\}$, where

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

2.1.5 Peaks over Threshold (POT)

The method that is used when fitting a GPD to data is called Peaks Over Threshold (POT). Extreme events are identified by defining a high threshold u for which the exceedances are $\{x_i : x_i > u\}$. With exceedances $x_{(1)}, ..., x_{(k)}$ the threshold excesses are $y_j = x_{(j)} - u$ for $j = 1, ..., k$. To choose a suitable threshold is essential. A threshold that is too high leads to high variance due to few excesses with which the model can be estimated. A threshold that is too low is likely to violate the asymptotic bias of the model, leading to bias.

Below follow two methods regarding how to choose an appropriate threshold; mean residual plot and plotting parameter estimates against different thresholds.

Mean residual plot

The mean residual plot is computed from the mean of GPD. Let $\xi < 1$ and assume that GPD is a valid model for the exceedances over threshold u_0 derived from following stochastic sequence $X_1, ..X_n$. The mean of GPD is

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

Since GPD is assumed to be a valid model for the exceedances over u_0, it should be equally valid for thresholds $u > u_0$. If $\sigma_u = \sigma_{u_0} + \xi u$, it then follows that

$$\begin{aligned} E(X-u|X>u) &= \frac{\sigma_u}{1-\xi} && (2.3)\\ &= \frac{\sigma_{u_0}+\xi u}{1-\xi}. && (2.4) \end{aligned}$$

Consequently, the conditional mean is linear for $u > u_0$. Additionally, the result from Equation 2.3 imply that the sample mean of the threshold excesses of u provides an empirical estimate. Plotting a locus of points:

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

should thereby be approximately linear in u for exceedances $x_{(i)}, ..., x_{(n_u)}$. Number of exceeded observations is denoted by n_u. Confidence interval is included in the plot which is based on the approximate normality of sample means [3].

Parameter estimates against different thresholds

This approach involves fitting the GPD to a range of thresholds and looking for some stability of the parameter estimates. The smallest threshold u should be selected in which linearity is constant to higher values and within the confidence interval. It is because if GPD is a proper model to describe the exceedances over threshold u_0 it should be as good for threshold u, $u > u_0$. The shape parameter is ξ independent of the threshold while the scale parameter σ_u is dependent on u and denoted by:

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

After reparameterization, the scale parameter is written as:

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

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.

The plots consist of the estimated parameters $\hat{\xi}$ and $\hat{\sigma}^*$ with respectively confidence interval. The confidence interval is derived from the variance-covariance matrix and the delta method [3].

2.1.6 Peaks Over Threshold For Negative Data

When POT is used for GPD adapted to m_n (or $-M_n$) the fitting of the distribution is done on the excess loss. For clarity Figure 2.1 illustrates a very simplified version of POT used on a sequence of positive random variables $\{Z_1, Z_2, Z_3, Z_4\}$. The relation $\tilde{Z}_i = -Z_i$ holds and threshold is chosen to u respectively $-u$. As familiar, POT method is used to model the distribution of the values large enough to exceed a chosen threshold. The same principle applies when the data and threshold is negative. Due to the relation $\tilde{Z} = -Z$, a large excess from $\tilde{Z}$ also represents a small value from the sequence Z. The exceedances that are modeled from $\tilde{Z}$ are called excess loss or shortfalls.

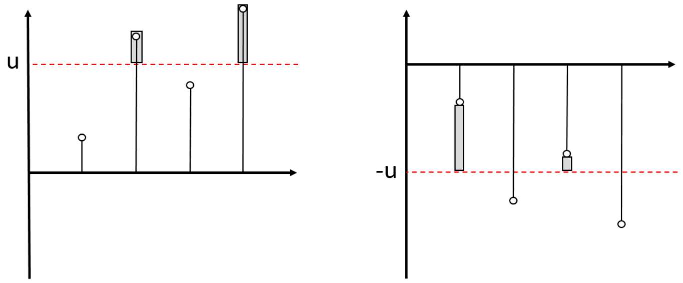


Figure 2.1: Simplified version of POT for positive and negative data

2.1.7 Conditional Probability fitted to negative data

Let $\tilde{X}_i = -X_i$ for $i = 1, ..., n$ be a sequence of independent stochastic variables and choose threshold $\tilde{u} = -u$, $\tilde{u} < 0$. Fit GPD to $\tilde{X} - \tilde{u}$ so

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

which also can be written as:

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

where F is the unknown distribution function. It follows that

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

where

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

which is the same as number of exceedances from sequence $\tilde{X}$ over threshold $\tilde{u}$. It follows from $P(\tilde{X} > x)$ that

$$P(-X > x) = \eta_{\tilde{u}} \left[1 + \xi \Big(\frac{x - (-u)}{\sigma} \Big) \right]^{-1/\xi} \quad 0 < x < u \tag{2.5}$$

$$P(X < -x) = \eta_{\tilde{u}} \left[1 + \xi \Big(\frac{u - (-x)}{\sigma} \Big) \right]^{-1/\xi} \quad -x < u \tag{2.6}$$

$$P(X < y) = \eta_{\tilde{u}} \left[1 + \xi \Big(\frac{u - y}{\sigma} \Big) \right]^{-1/\xi} \quad y < u \tag{2.7}$$

for $0 < y < u$ [3].

2.1.8 Lower endpoint

For GPD it holds that distribution of exceedances over the threshold u has an upper bound, $\sigma/|\xi|$, if the shape parameter $\xi < 0$. This means that the distribution of the original random variable has an upper endpoint of $u + \frac{\sigma}{|\xi|}$. If $\xi > 0$, it has no upper endpoint [3]. The same principle applies for the distribution of excess loss. A lower endpoint exists if $\xi < 0$ but not for $\xi > 0$. From Equation 2.7 for $\xi < 0$ it holds that

$$P(X < u - y) = P(X < u)\big(1 - |\xi|\frac{y}{\sigma}\big)^{1/|\xi|} \tag{2.8}$$

for $0 < y < \sigma/|\xi|$. By reforming $u - y = z$ to $y = u - z$ Equation 2.8 is written as

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

for $0 < u - z < \sigma/|\xi|$. The lower endpoint will then be

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

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

The same principle applies to GEV distributions. If $\xi < 0$ an upper endpoint exists and is given by

$$\tilde{\mu} + \frac{\sigma}{|\xi|}.$$

Thus, the following equation:

$$-\Big(\tilde{\mu} + \frac{\sigma}{|\xi|}\Big).$$

is the lower endpoint for an equal distribution that is a 180° rotation around the y-axis [3].

2.1.9 Continuous distribution with point mass

If a distribution has a lower endpoint that is negative, the region below zero represents "the probability that negative time after a collision appears". This is irrational and a method to cope with this kind of problem is required. Boundaries are introduced so that the distribution function is valid between zero and the chosen threshold u. When introducing restrictions two point masses are also introduced. The Equation 2.7 imply that point masses m_1 and m_2 are:

$$\begin{aligned} m_1 &= \eta_{\tilde{u}}\Big(1+\xi\frac{u}{\sigma}\Big)^{-1/\xi} \\ m_2 &= 1-\eta_{\tilde{u}} \end{aligned}$$

see Appendix A for further calculations. In Figure 2.2 a graphical description of the two point masses and the density function $f(x)$ are illustrated. The density $f(x)$ is derived from Equation 2.7 and results in:

$$f(x) = \frac{\eta_{\tilde{u}}}{\xi}\Big(1+\xi\frac{u-x}{\sigma}\Big)^{-1/\xi-1}$$

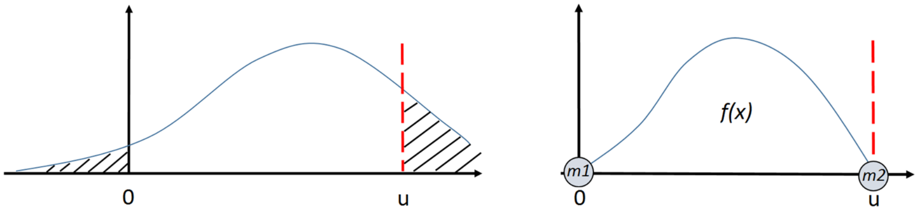


Figure 2.2: The left figure depicts a density function $f(x)$ without restrictions. The right figure depicts the same function $f(x)$ but with restrictions and two point masses m_1 and m_2

Finally, the probability is denoted by:

$$P(X < x) = \begin{cases} \eta_{\tilde{u}} \left(1 + \xi \frac{u}{\sigma}\right)^{-1/\xi}, & x \leq 0 \\ \eta_{\tilde{u}} \left(1 + \xi \frac{u-x}{\sigma}\right)^{-1/\xi}, & 0 < x < u \\ 1, & x \geq u \end{cases} \tag{2.9}$$

which is called censored GPD and can be thought as a mixture of discrete and continuous distributions.

2.2 General Statistical Theory

2.2.1 Poisson Process

In *Gut, (2009)* [4] the homogeneous Poisson process is defined by:

Definition 3. A Poisson process is a stochastic process $\{X(t), t \geq 0\}$ with independent, stationary, Poisson distributed increments. Also, $X(0) = 0$. The increment $X(t+h) - X(t)$ is Poisson distributed with expected value λh for every t and $h \geq 0$. In other words,

- the increments $\{X(t_k) - X(t_{k-1}), 1 \leq k \leq n\}$ are independent random variables for all $0 \leq t_0 \leq t_1 \leq t_2 \leq ... \leq t_{n-1} \leq t_n$ and all n;
- $X(0) = 0$ and there exists $\lambda > 0$ such that

  $$X(t) - X(s) \in \mathrm{Po}(\lambda(t-s)), \quad 0 \leq s < t.$$

  The constant λ is called the intensity of the process.

The intensity is calculated by

$$\int_0^t \lambda(u)\, du = \lambda t$$

for $t \geq 0$ and the probability mass function with respect to counting measure is:

$$p_{X(t)}(x) = \mathrm{P}(X(t) = k) = e^{(-\lambda t)} \frac{(\lambda t)^k}{k!}, \quad k = 0, 1...,$$

A nonhomogeneous Poisson process is a Poisson process with time-dependent intensity. For the increment $X(t_2) - X(t_1)$ the intensity is:

$$X(t_2) - X(t_1) \in \mathrm{Po}(\int_{t_1}^{t_2} \lambda(u)\, du).$$

The variance is calculated by

$$Var(X) = \sqrt{\lambda}$$

for X [10].

2.2.2 Binomial distribution

The probability mass function for a binomial distributed stochastic variable X is

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, ..., n.$$

Parameter n is number of independent experiments, k is number of successes in the experiment and $0 \leq p \leq 1$ is the probability for success. The variance is:

$$Var(X) = \sqrt{np(1-p)}$$

for X [5].

2.2.3 Maximum Likelihood

The method of maximum likelihood is based on the likelihood function:

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

where f is the density function, x_i the values from the stochastic independent variables $X_1, ..., X_n$ and θ is the unknown parameter in parameter space Θ, $\theta \in \Theta$. The idea is to use different values of $\theta \in \Theta$ and select the θ that generates the largest value of $L(\theta)$. This is because finding the highest likelihood is the same as finding the model with the highest probability. The θ that results in the highest value of L is chosen as the estimation and is called $\tilde{\theta}$. When maximizing the likelihood function the log-likelihood function:

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

is often used instead. This is because it is generally easier to maximize a sum $l(\theta)$ compared to a product $L(\theta)$. The two functions takes it's maximum at the same point since the logarithm is a monotone transformation [5].

2.2.4 Akaike Information Criterion (AIC)

When working with models that are not nested the Akaike information criterion (AIC) is useful when comparing models. It is defined as:

$$AIC = 2k - 2\ln(\hat{L})$$

where k is the number of parameters and $\hat{L}$ is the maximum value of the likelihood function for the model. A small value of AIC is better than a large value [1].

2.2.5 Delta Method

Computing the variance of variable $\sigma^*(\sigma_u, \xi)$ is done in following manner:

$$Var(\sigma^*) \approx (\nabla\sigma^{*T})V(\nabla\sigma^*),$$

where

$$\nabla\sigma^{*T} = \Big[\frac{\partial\sigma^*}{\partial\sigma_u}, \frac{\partial\sigma^*}{\partial\xi}\Big]$$

and V is the variance-covariance matrix for σ_u and ξ. In general it can be shown that $\sqrt{n}(\sigma^* - \sigma) \rightarrow N(0, var(\sigma^*))$ [3].

2.2.6 Confidence Interval

A 95% confidence interval is calculated by

$$CI = mean \pm 1.96 \cdot \sqrt{Var}$$

where $mean$ is the estimated parameter and Var the variance for the intended parameter [5].

2.2.7 Model Diagnostics

Two graphical techniques that are used to evaluate how well a model fits to observed data are **probability plot** and **quantile plot**. The ordered sample of independent observations

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

has unknown distribution function F and an estimated distribution function $\hat{F}$. If $\hat{F}$ is a reasonable estimation, the points should lie close to the unit diagonal. For the probability plot the points consist of

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

and following points

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

forms the quantile plot [3].

3. Data Analysis

The first part of Data Analysis is an overview of how the data is collected. An overview of the structures of PET and TTCmin is also presented in this part. In the second part, several analyses on the negated data are performed. The estimated parameters for the used distributions are presented in following order: GEV distribution, GPD and Poisson process. Finally, the probability for collision and near collision using the estimated parameters from the different distributions are presented in the result.

3.1 Data

The analysis is based on data filmed in a four-way intersection in Barcelona consisting of one-way streets, see Figure 3.1. The data is from the European research project InDev which is within the framework Horizon 2020 by the European Commission. The film sequence is from $2017 - 06 - 20$.

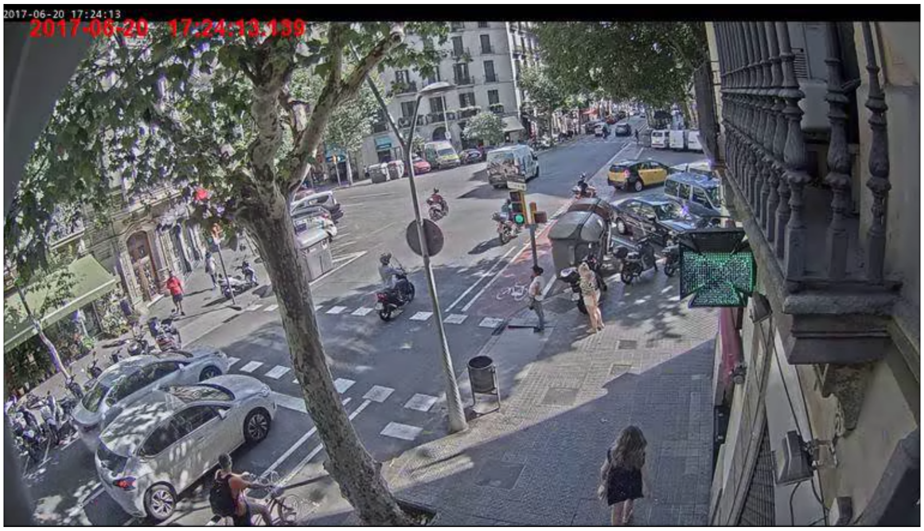


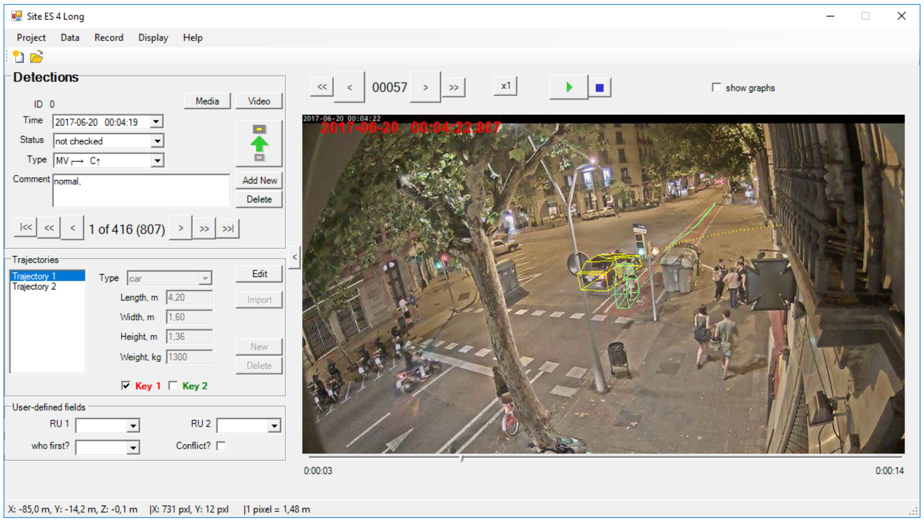
Figure 3.1: The four-way intersection in Barcelona that was filmed

The data has been generated in the following way. Two students have analyzed 24 hours of film by first marking every trajectory between right turning motor vehicles and cyclists. This was done based on the following criteria:

- Both vehicles are in the intersections at the same time. However, if the cyclist/motor vehicle has crossed the cyclist path when the other vehicle enters, it is considered to be no risk of collision and is therefore not marked as a trajectory
- The two vehicles need to be in motion at some point in the intersection
- If there are more cyclists/motor vehicles in the intersection, the trajectory should be done on the most relevant (closest) vehicles

which resulted in 415 encounters. For every marked encounter, the students fit "boxes" over the vehicles in every fourth frame (the film sequences are run in 15 frames per second) and the moving average is calculated for the other frames. Lastly, PET and TTCmin are computed in the program according to their definitions, see Figure 3.2 for a print screen from the program. Note the boxes over the car and bike together with the future path of the vehicles.

Some of the values in PET gets the same value. It does not necessarily mean that the actual situation was exactly alike it is just a consequent when computing PET in the way it is. It can be observed in the plots in Figure 3.11.



ID 0

Time 2017-06-20 00:04:19

Status not checked

Type MV ┌→ C↑

Comment normal,

1 of 416 (807)

Trajectory 1
Trajectory 2

Type car

Length, m 4,20

Width, m 1,60

Height, m 1,36

Weight, kg 1300

☑ Key 1 ☐ Key 2

User-defined fields

RU 1 RU 2

who first? Conflict? ☐

00057

X: -85,0 m, Y: -14,2 m, Z: -0,1 m |X: 731 pxl, Y: 12 pxl |1 pixel = 1,48 m

Figure 3.2: Print from the program where TTCmin and PET are computed

In Table 3.1 the amount and types of motor vehicles are presented.

Table 3.1: Table of encounters during 24 hours

	Bus	Car	Minivan	Truck
Bicycle	2	373	29	11

The two sequences are treated as independent random variables. In Figure 3.3 histogram and plots are presented for the sequences TTCmin and PET. TTCmin was calculated for 168 of the encounters and PET 413 of the encounters.

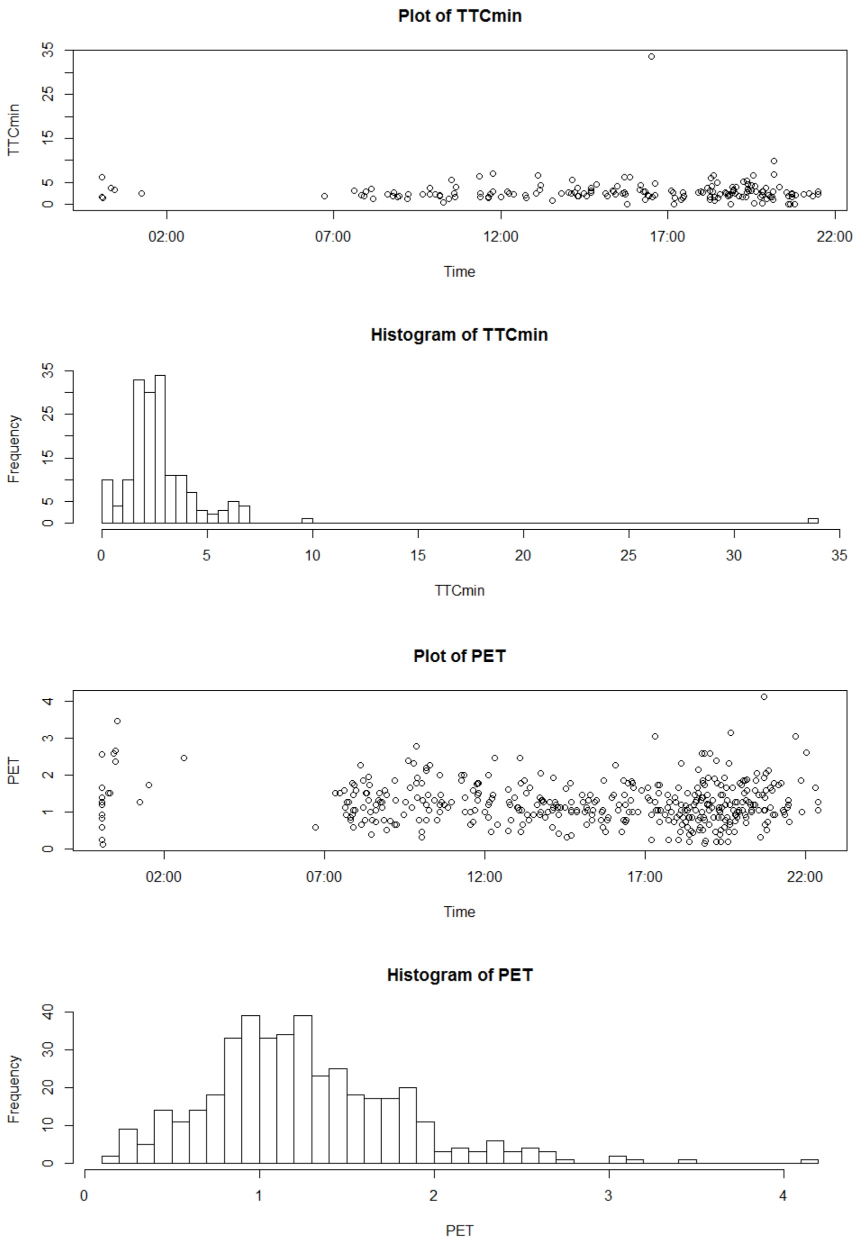


Figure 3.3: Plot and Histogram of PET and TTCmin

The lowest values the sequences takes are 0.01 seconds in TTCmin and 0.13 seconds in PET.

3.2 Modelling of the road safety data

3.2.1 Generalized Extreme Value Distribution: GEV

Block-size selection

The blocking is done on -TTCmin and -PET. The first approach is to select the largest values of every hour and generate a sequence of a maximum of every hour. Yet, as one can see in Table 3.2 are the range of trajectories during the day very wide. There should be enough data in every block when selecting a maxima so hours with less than ~ 10 observations are not included.

Table 3.2: Number of observations for every hours of -PET and -TTCmin

Hour	-PET	-TTCmin
00	11	2
01	2	1
02	1	0
03	0	0
04	0	0
05	0	0
06	1	1
07	17	4
08	25	7
09	16	5
10	21	9
11	16	8
12	16	6
13	25	6
14	22	13
15	22	12
16	20	11
17	25	10
18	59	25
19	52	24
20	39	17
21	15	5
22	4	0
23	0	0

The second approach is to choose a block size n and generate a sequence with the largest values from every block. Different values for n is used and for -PET the block size is selected to $n_{-PET} = 12$, based on the diagnostics plots, see Figure 3.4. The quantiles in the probability and quantile plot follow the diagonal indicating that the GEV distribution can be considered as an appropriate fit to depict -PET.

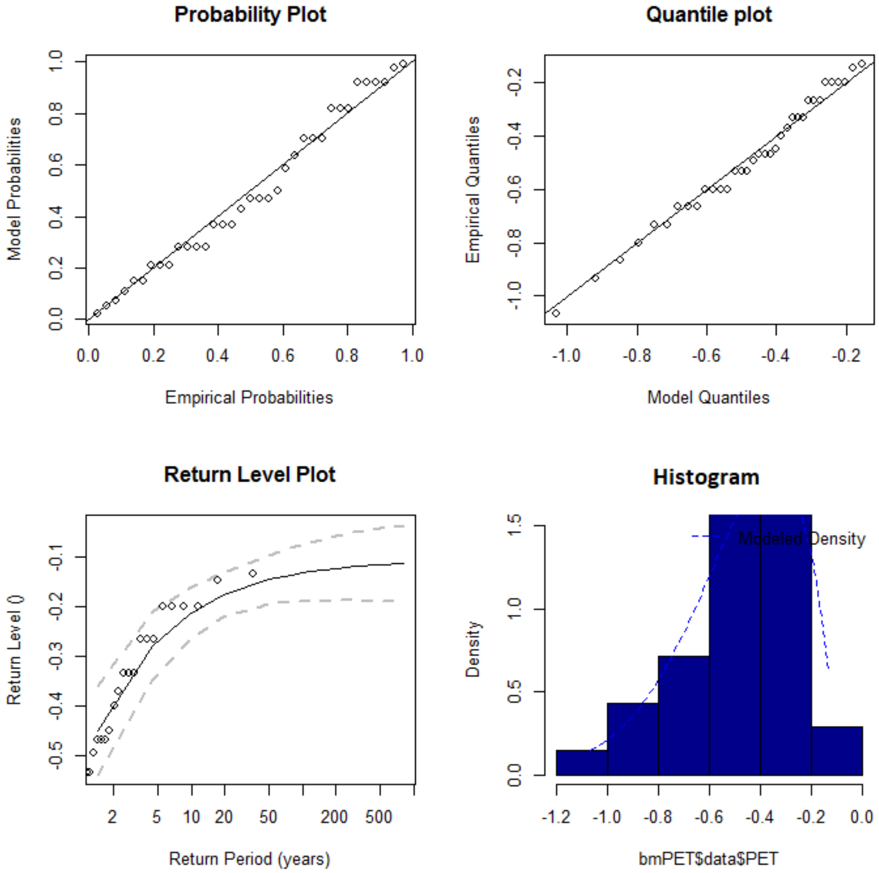


Figure 3.4: GEV distribution fitted to -PET with block size $n = 12$

A GEV distribution is fitted to -TTCmin with various block sizes (not less than 10). The distributions that are obtained are poor images of the underlying data -TTCmin. A likely explanation is that there is not enough observations to fit a GEV distribution.

PET: Parameter estimation

The estimation of the parameters with 95% confidence interval for the fit for -PET are presented in Table 3.3. Note: $\tilde{\mu}$ is the estimated location parameter for -PET while μ is for the original data PET. This is due to the relation $\tilde{\mu} = -\mu$.

Table 3.3: Estimated parameters with confidence interval for -PET

Confidence interval:	95% Lower bound	MLE	95% Upper bound
$\hat{\tilde{\mu}}$	-0.628	-0.533	-0.439
$\hat{\sigma}$	0.178	0.258	0.339
$\hat{\xi}$	-0.896	-0.604	-0.312

The lower endpoint for PET is:

$$-\Big(\hat{\tilde{\mu}} + \frac{\hat{\sigma}}{|\hat{\xi}|}\Big) = 0.106,$$

so the generated distribution covers all observations in PET.

3.2.2 Generalized Pareto Distribution: GPD

Threshold selection

By studying mean residual plots for -TTCmin and -PET (Figures 3.5 and 3.6) and threshold range plots (Figures 3.7 and 3.8) appropriate thresholds are selected. It is also important to choose a threshold that generates a distribution with a lower endpoint close to zero and not too far from the lowest value in TTCmin/PET. The diagnostics plots: probability plot, quantile plot, histogram and return level plot are also considered in the decision of threshold.

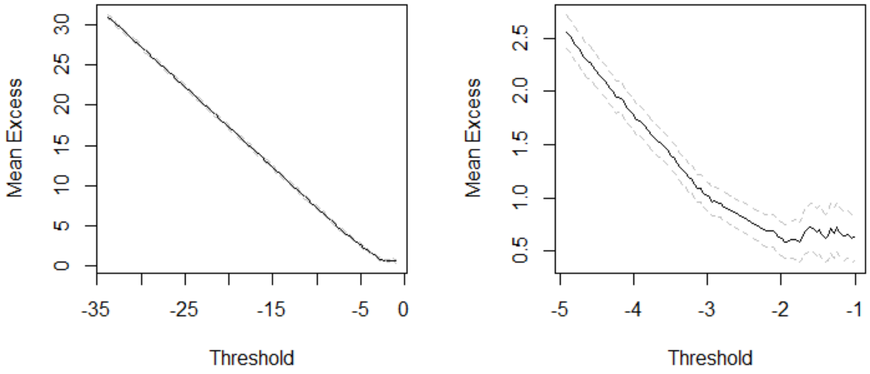


Figure 3.5: Mean Residual Plot for -TTCmin. The left picture is original with all the data and the right figure is zoomed in, note the axes

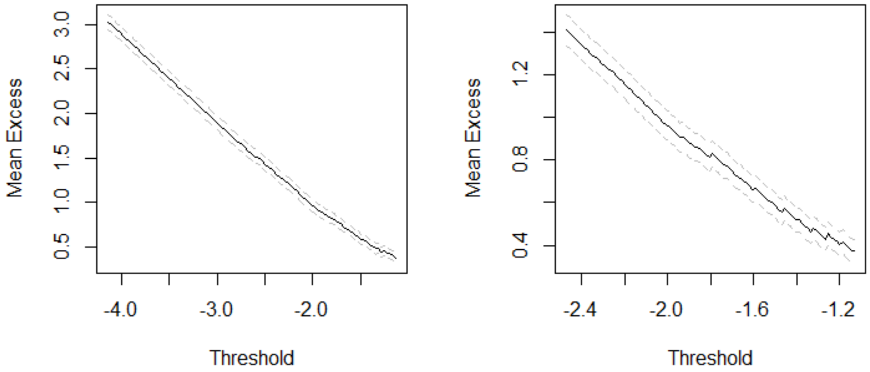


Figure 3.6: Mean Residual Plot for -PET. The left picture includes all the data and the right figure is zoomed in, note the axes

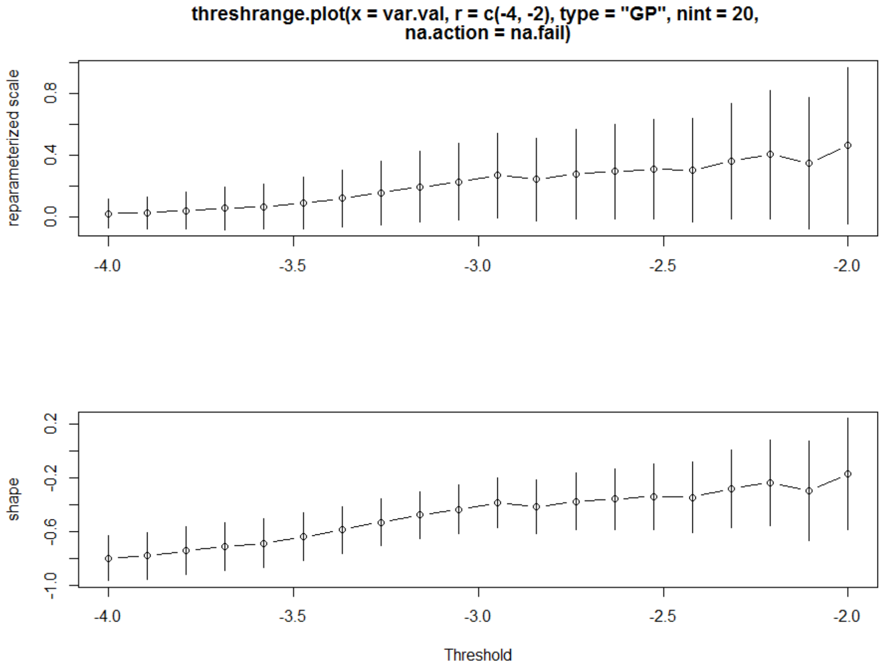


Figure 3.7: Threshold range plot for -TTCmin

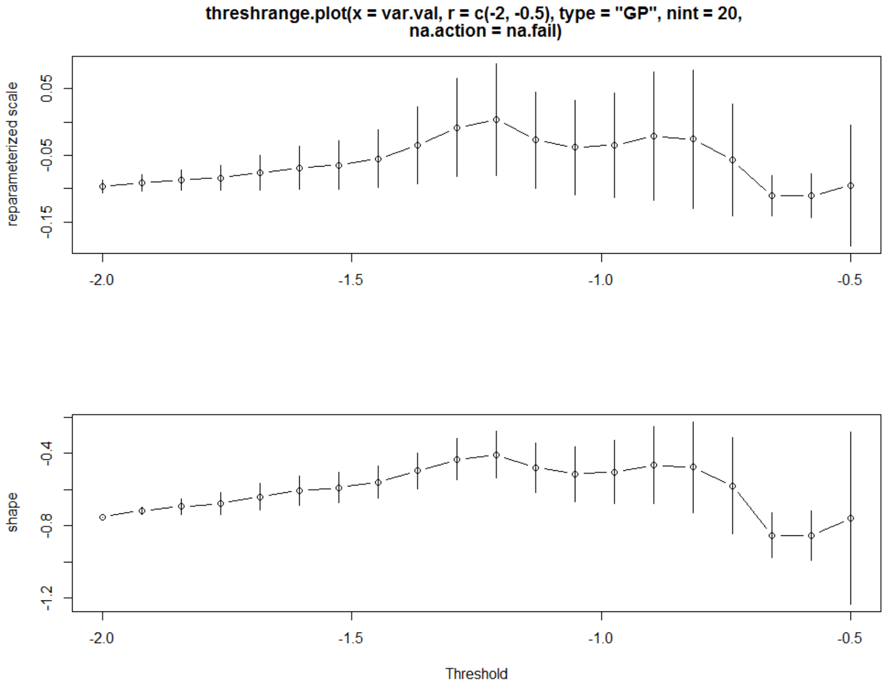


Figure 3.8: Threshold range plot for -PET

Due to linearity in the figures the threshold for -TTCmin is selected to -3 and a suitable threshold for -PET is chosen to -1.5.

TTCmin

The analysis on TTCmin is done with two different approaches. The first approach leads to estimates of σ and ξ such that

$$u \approx \frac{\hat{\sigma}}{|\hat{\xi}|}.$$

The generated distribution will then be valid between zero and u which is desired. The second approach is based on the threshold selection $\tilde{u} = -3$ from previous section.

Method I: **TTCmin**

The modelling is done on -TTCmin for the encounters between cars and cyclists, denoted by -TTCmin$_{cars}$. It is due to a comparison between AIC of the fit of GPD on -TTCmin ($AIC = 31.75$) and -TTCmin$_{cars}$ ($AIC = 29.65$). Since AIC is lower for -TTCmin$_{car}$ than for -TTCmin the model will be estimated for the encounters between cars and cyclists. The threshold is selected to $\tilde{u} = -1.7$ and the goodness of fit is presented in Figure 3.9.

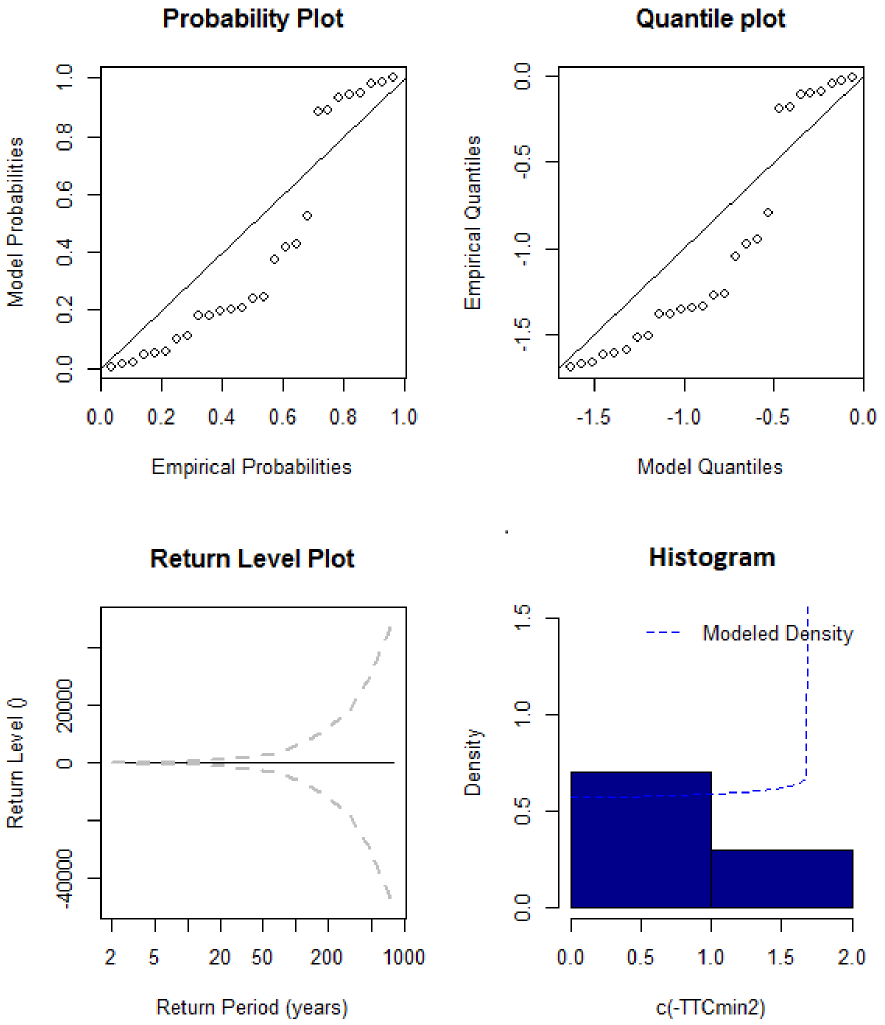


Figure 3.9: Method I: Probability-plot, Quantile-plot, Return Level Plot and histogram for -TTCmin$_{cars}$

***Method I*: Parameter estimation**

The generated parameters are presented in Table 3.4.

Table 3.4: Estimated parameters for TTCmin$_{cars}$

Confidence interval:	95% Lower bound	MLE	95% Upper bound
$\hat{\sigma}$	1.760	1.760	1.760
$\hat{\xi}$	-1.045	-1.041	-1.038

Since the shape parameter is negative, a lower endpoint exists which is calculated to

$$u - \frac{\hat{\sigma}}{|\hat{\xi}|} = 0.009.$$

Since the lower endpoint is rounded up to 0.01 it is clear that the generated distribution covers all values in TTCmin.

***Method II*: TTCmin**

As in method I, the fitting of GPD is done on -TTCmin$_{cars}$. It is due to a lower value of AIC only involving cars (AIC= 206.1621) compared to (AIC= 238.7365) when all motor vehicles are included.

In Figure 3.10 diagnostics plots are presented for -TTCmin$_{cars}$.

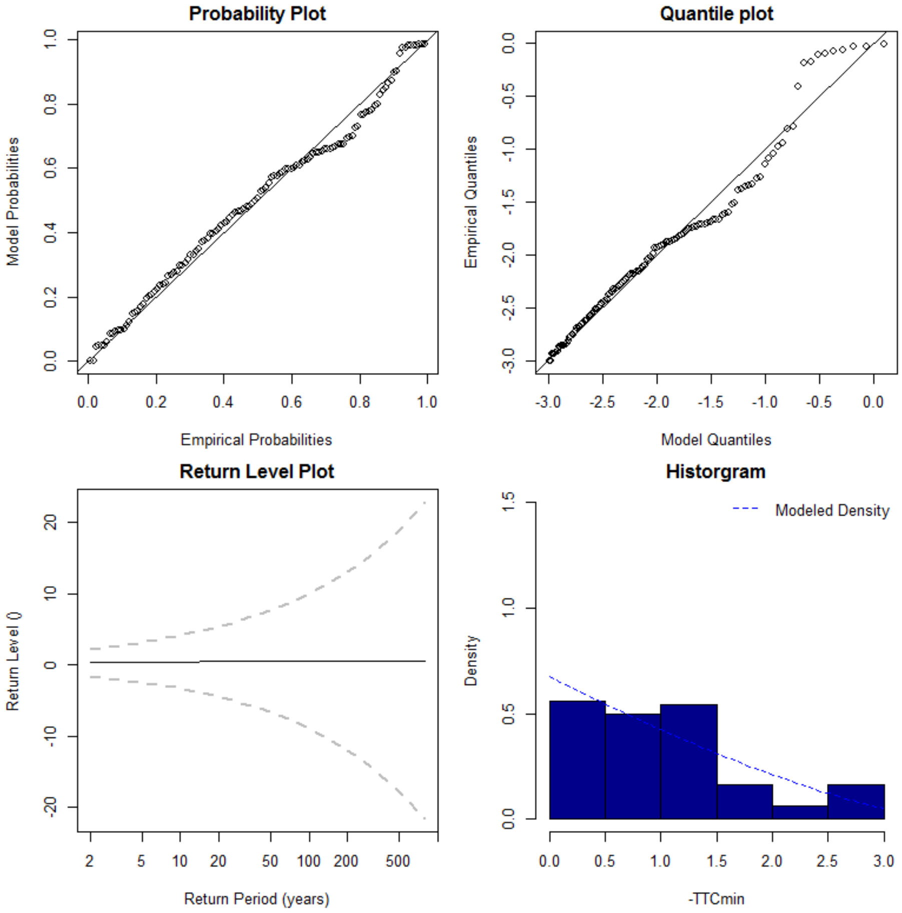


Figure 3.10: Method II: Probability-plot, Quantile-plot, Return level plot and histogram for -TTCmin$_{cars}$

Method II: **Parameter estimation**

The parameters are presented in Table 3.5.

Table 3.5: Estimated parameters for TTCmin$_{cars}$

Confidence interval:	95% Lower bound	MLE	95% Upper bound
$\hat{\sigma}$	1.057	1.421	1.786
$\hat{\xi}$	-0.575	-0.389	-0.203

One can see in the diagnostics plot that the fit is accurate for lower values in the quantile and probability plot. Since the shape parameter is negative, a lower endpoint exists which is calculated to

$$u - \frac{\hat{\sigma}}{|\hat{\xi}|} = -0.655$$

PET

In the same manner as for -TTCmin. The first approach is done by fitting GPD to -PET, meaning all right turning motor vehicles are included. The second approach is to select data with exclusively right turning cars. The same threshold $\tilde{u} = -1.5$ for the two approaches. The result is similar to each other, yet the AIC for the second approach is lower (AIC= 113) than for the first approach (AIC=123). The approach with only cars and bikes are therefore selected. The parameters are found in Section 3.2.2 and the probability plot, quantile plot, return level plot and histogram for -PET$_{cars}$ are illustrated in Figure 3.11.

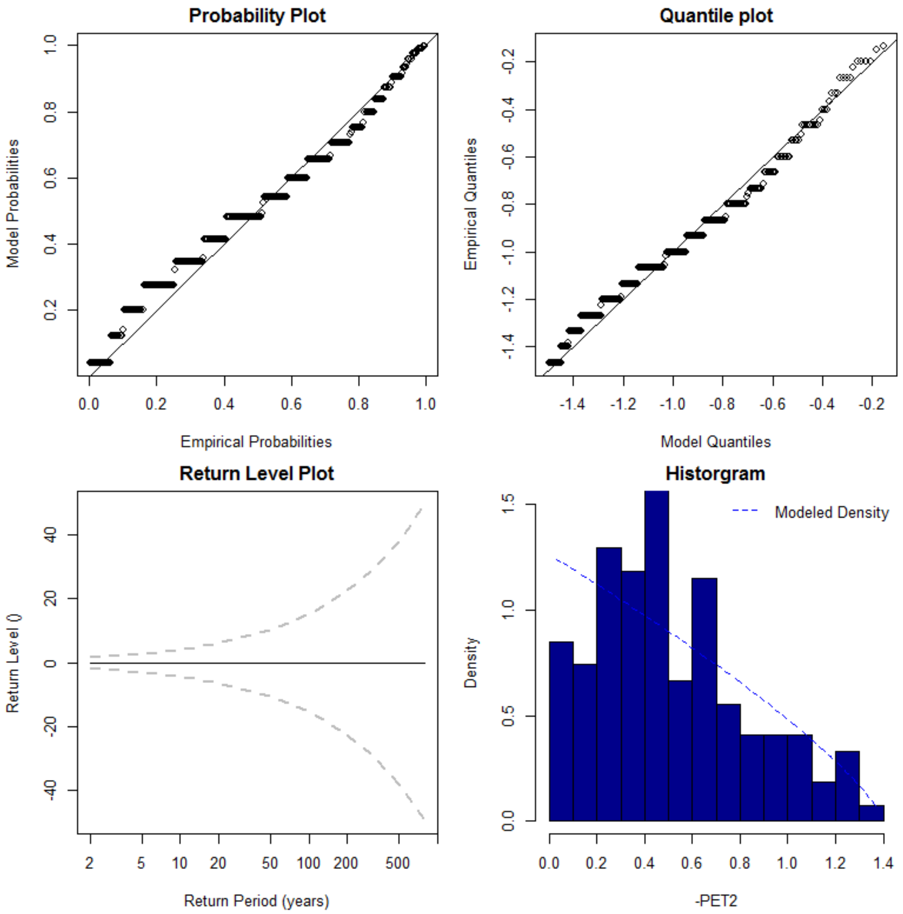


Figure 3.11: Probability-plot, Quantile-plot, Return level plot and histogram for -PET$_{cars}$

PET: Parameter estimation

The estimated parameters, with confidence interval, are presented in Table 3.6.

Table 3.6: Estimated parameters for PET

Confidence interval:	95% Lower bound	MLE	95% Upper bound
$\hat{\sigma}$	0.684	0.792	0.899
$\hat{\xi}$	-0.658	-0.564	-0.471

As $\xi < 0$ a lower endpoint exists. It is is calculated to:

$$u - \frac{\hat{\sigma}}{|\hat{\xi}|} = 0.097$$

and the generated distribution covers all observations in PET.

3.2.3 Poisson Process

The process with *all* registered encounters is assumed to be described as two independent Poisson processes $Y(t) = X(t) + Z(t)$. Process $\{Y(t),\, t \geq 0\}$ is the process for all encounters, process $\{X(t),\, t \geq 0\}$ is for encounters between cars and cyclists and $\{Z(t),\, t \geq 0\}$ is the process for cyclists and the remaining motor vehicles. In Figure 3.12 process X and Y are depicted. The intensity for process Y is computed to $\lambda_{motor} = 415/day$ and for process X is it $\lambda_{car} = 373/day$. The time lap is 24 hours.

Figure 3.12: Poisson process Y(t) and X(t)

The idea with using Poisson process in combination with GPD/GEV distribution is to estimate the risk of collision/near collision per year (365 days). Let p_1 be the risk of collision/near collision for one encounter between car and cyclist and t number of days that are of interest. The intensity for the Poisson process is then:

$$t\, \lambda_{car}\, p_1 \tag{3.1}$$

Specifically, the intensity for risk of collision during one year using GPD is:

$$365\, \lambda_{car}\, \eta_{\tilde{u}} \Big(1 + \xi \frac{u}{\sigma}\Big)^{-1/\xi}.$$

The same approach applies when determine the risk of near accidents. This concept is called thinning of Poisson process.

3.3 Results

There are no recorded accidents in the data that are comparable to the estimated probability. The probability for a near accident can on the other hand both be estimated and empirically calculated. For PET/TTCmin the probability is computed to $P(0 < X < a) = x/N$, where x is number of encounters under a seconds and N is the total number of encounters. The outcome is a binomial random variable and the confidence interval is calculated by:

$$I_{x/N} = \frac{x}{N} \pm 1.96\sqrt{\frac{\frac{x}{N}\left(1 - \frac{x}{N}\right)}{N}}.$$

For encounters between cars and bikes $N = 373$ for both PET and TTCmin. When all encounters are included is $N = 415$. For different values of a are Table 3.7 constructed.

Table 3.7: Empirical probability

		95% Lower Bound	Point Estimate	95% Upper Bound
PET:	$P(0 < X < 0.05)$	0	0	0
	$P(0 < X < 0.2)$	0	0.005	0.012
	$P(0 < X < 1)$	0.259	0.303	0.348
PET_{car}:	$P(0 < X < 0.05)$	0	0	0
	$P(0 < X < 0.2)$	0	0.005	0.013
	$P(0 < X < 1)$	0.261	0.308	0.355
TTCmin_{car}:	$P(0 < X < 0.05)$	0	0.008	0.017
	$P(0 < X < 0.2)$	0.007	0.022	0.036
	$P(0 < X < 1)$	0.012	0.030	0.047

The number of near accidents under a seconds in the data are presented in Table 3.8.

Table 3.8: Numbers of near accidents in the data

$0 < X < a$	TTCmin_{car}	PET_{car}	PET
$0 < X < 0.05$	3	0	0
$0 < X < 0.2$	8	2	2
$0 < X < 1$	11	113	126

Estimation of accident/near accident using GPD

The probability for accidents and near accidents is based on the estimated parameters using GPD. A more detailed explanation of *how* the probability is computed in R is found in Appendix B. The calculations are straight forward, however the thinking is ”reversed” due to the use of POT on negated data.

Table 3.9: Probability of accident/near accident using GPD

	I: TTCmin$_{car}$	II: TTCmin$_{car}$	PET$_{car}$
$P(X = 0)$	0	0.003	0
$P(0 < X < 0.05)$	0.002	0.004	0
$P(0 < X < 0.2)$	0.009	0.007	0.007
$P(0 < X < 1)$	0.043	0.037	0.332

Estimation of accident/near accident using GEV

The estimated distribution $\tilde{G}(z)$ using block maxima with n number of blocks is:

$$\tilde{G}(z) = 1 - \big(1 - F(z)\big)^n. \tag{3.2}$$

The calculations are based on Equation 2.1 and 2.2 and is rewritten to:

$$\hat{F}(z) \approx 1 - \big(1 - \hat{\tilde{G}}(z)\big)^{1/n}. \tag{3.3}$$

where the block size is $n = 12$. The probability is computed by the use of parameters in Table 3.10 and Equation 2.1 and 3.3. The result of the calculations are presented in 3.10.

Table 3.10: Probability of accident/near accident using GEV distribution

	PET
$P(X = 0)$	0
$P(0 < X < 0.05)$	0
$P(0 < X < 0.2)$	0.007
$P(0 < X < 1)$	0.246

Estimation of accident/near accident using Poisson process

Number of expected near accidents to occur during one day are presented in Table 3.11. The calculations are done by $p_1\, \lambda_{car}\, days$ where p_1 is the probability of near accident from Table 3.9. The probability that theses encounters occur are done with Poisson process and presented in Table 3.12

Table 3.11: Expected number of near accidents

p_1	days	I: TTCmin$_{car}$	PET$_{car}$
$P(0 < X < 0.05)$	1	1	$\sim$
$P(0 < X < 0.2)$	1	4	3
$P(0 < X < 1)$	1	17	124

Table 3.12: Probability of near accidents using Poisson process, $P(X = k)$

	k	I: TTCmin$_{car}$	PET$_{car}$
$0 < X < 0.05$	1	0.353	$\sim$
$0 < X < 0.2$	3	0.221	0.219
	4	0.183	0.144

Additionally, the expected numbers of accidents for a year using the result for "II: TTCmin$_{car}$" is 119 accidents with probability 0.544. The expected value of near accidents under 0.2 seconds using the result from 3.10 is approximately 3 ($\approx 415 * 0.007$) with probability 0.223. The expected value of near accidents under one second is approximately 103 ($\approx 0.246 * 415$) with probability 0.039.

Remarks: Fitting of GPD using POT on -PET$_{cars}$ with one collision in the data is done. The estimated parameters are:

$\hat{\sigma}$	$\hat{\xi}$
0.749	-0.488

and the lower endpoint is -0.035. The probability of collision is computed to 0.0003 and for a year it would be $\sim$ 44 collisions. It is based on 374 encounters per day for 365 days ($0.0003 * 374 * 365$). If two collisions are added in the data the parameters are

$\hat{\sigma}$	$\hat{\xi}$
0.755	-0.484

and the risk of collision would be 0.0009 with 124 collisions per year. Even though this data is fabricated and generates a high probability of collision it gives an idea of how much more data that are required. A sequence that is triple the size of the current one could be a good start.

4. Conclusions

This thesis aims to propose a model for analyzing the road safety data. The modeling has been done with Generalized Extreme Value distribution and Generalized Pareto distribution. Moreover, a model using the probabilities from GPD along with Poisson distribution has been constructed.

The modeled probabilities using POT with GPD on the negated data are more precise than using block maxima with GEV distribution. Furthermore, the focus is on the encounters between cars and cyclists when GPD is fitted. The proposed model for PET seems to fit the data well. It is proper for near collision estimations and the modeled probability is within the confidence interval for the empirical probability. Regarding TTCmin are two models constructed with GPD and the outcomes are reasonable for near accidents. The risk of collision according to one of the methods is $\sim 0.3\%$ and the question that arises is: is it likely that the risk of collision is 0.3%?

The obtained probabilities for TTCmin and PET using GPD in combination with Poisson distribution resulted in various outcomes. The estimated probability seems reasonable. Although, as stated in Section 3.3 under Remark an interesting approach that probably generates a better result is to focus on the rush hours and use a nonhomogeneous intensity[1]. Also, more data is required for both TTCmin and PET. Specifically, small values so estimation of collision can be done. More things can be made to continue the project of applying extreme value theory on road safety. For example:

- Use a nonhomogeneous Poisson process
- Investigate other approaches of the point mass in the origin
- Use data from other locations to see differences in the fitted models.

[1] The majority of the small values (< 0.3s) in PET are between 17.00-20.00 o'clock.

5. Bibliography

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716 – 723, 1974.

[2] Prateek N. Sharma Mukesh Khare Sila P. Chakrabarti. Application of extreme value theory for predicting violations of air quality standards for an urban road intersection. *Transportation Research Part D: Transport and Environment*, 4(3):201–216, 1999.

[3] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag, London, 2001.

[4] Allan Gut. *An Intermediate Course in Probability.* Springer, New York, 2009.

[5] Gunnar Blom Jan Enger Gunnar Englund Jan Grandell Lars Holst. *Sannolikhetsteori och statistikteori med tillämpningar.* Studentlitteratur, Lund, 2013.

[6] Christer Hydén. The development of a method for traffic safety evaluation: the swedish traffic conflict technique.

[7] Carl Johnsson. *PhD Student, Departement of Road Safety, LTH.* 2019.

[8] Xianghai Meng Lai Zheng, Karim Ismail. Freeway safety estimation using extreme value theory approach: A comparative study. *Accident Analysis Prevention*, 62.

[9] T. Polders, E. Brijs. *How to analyse accident causation? A handbook with focus on vulnerable road users.* Deliverable 6.3. Horizon 2020 EC Project, InDeV, Hasselt, Belgium: Hasselt University, 2018.

[10] George Lindgren Holger Rootzén Maria Sandsten. *Stationary Stochastic Processes For Scientists And Engineers.* CRC Press, Taylow Francis Group, Boca Raton, 2014.

[11] Andrew P. Tarko. Use of crash surrogates and exceedance statistics to estimate road safety. *Accident Analysis Prevention*, 45.

[12] T. Hyde Christopher C. Wright. Extreme value methods for estimating road traffic capacity. *Transportation Research Part B: Methodological*, 20(2):125–138, 1986.

[13] Aliaksei Laureshyn Carl Jonsson Tim De Ceunynck Åse Svensson Maartje de Goede Nicolas Saunier Paweł Włodarek Richard van der Horst Stijn Daniels. Review of current study methods for VRU safety Appendix 6 – Scoping review: surrogate measures of safety in site-based road traffic observations. *InDeV: In-Depth understanding of accident causation for Vulnerable road users, HORIZON 2020 - the Framework Programme for Research and Innovation.*

A. Calculations of Point masses

Point mass m_1 is computed by

$$\begin{aligned} m_1 &= \int_{-\infty}^{0} \eta_{\tilde{u}} \Big(1 + \xi \frac{u-x}{\sigma}\Big)^{-(1+\frac{1}{\xi})} \\ &= \eta_{\tilde{u}} \Big[\big(1 + \xi \frac{u-x}{\sigma}\big)^{-1/\xi}\Big]_{-\infty}^{0} \\ &= \eta_{\tilde{u}} \big(1 + \xi \frac{u}{\sigma}\big)^{-1/\xi} - \lim_{x \to -\infty} \eta_{\tilde{u}} \big(1 + \xi \frac{u-x}{\sigma}\big)^{-1/\xi} \\ &= \eta_{\tilde{u}} \big(1 + \xi \frac{u}{\sigma}\big)^{-1/\xi} \end{aligned}$$

Point mass m_2 is computed by:

$$\begin{aligned} m_2 &= 1 - \eta_{\tilde{u}} \big(1 + \xi \frac{u}{\sigma}\big)^{-1/\xi} - \eta_{\tilde{u}} \Big[\big(1 + \xi \frac{u-x}{\sigma}\big)^{-1/\xi}\Big]_{0}^{u} \\ &= 1 - \eta_{\tilde{u}} \big(1 + \xi \frac{u}{\sigma}\big)^{-1/\xi} - \Big(\eta_{\tilde{u}} 1^{-1/\xi} + \eta_{\tilde{u}} \big(1 + \xi \frac{u}{\sigma}\big)^{-1/\xi}\Big) \\ &= 1 - \eta_{\tilde{u}} \end{aligned}$$

B. Calculations of Probability in R

As said before in Section 3.3 it is straight forward to compute the probability of near accident using GPD in R. However, the thinking is "reversed" due to the use of POT on negated data, with threshold $\tilde{u}$. Let the fitted parameters be σ, ξ and the threshold $u\,(= -\tilde{u})$. By the use of function *pgpd* in package *evd* (and *fExtremes*) the probability of an encounter under a seconds is:

$$P(0 < X < a) = \eta_{\tilde{u}}(1 - pgpd(u - a, \sigma, \xi)) \tag{B.1}$$

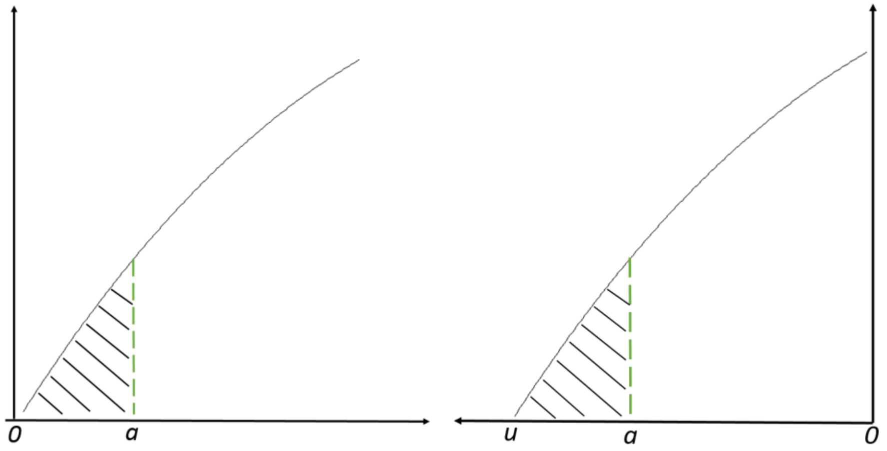


Figure B.1: The left Figure depicts the probability function for parameters and the right picture is a way to think when computing the probability for a distribution that was fitted for negated data