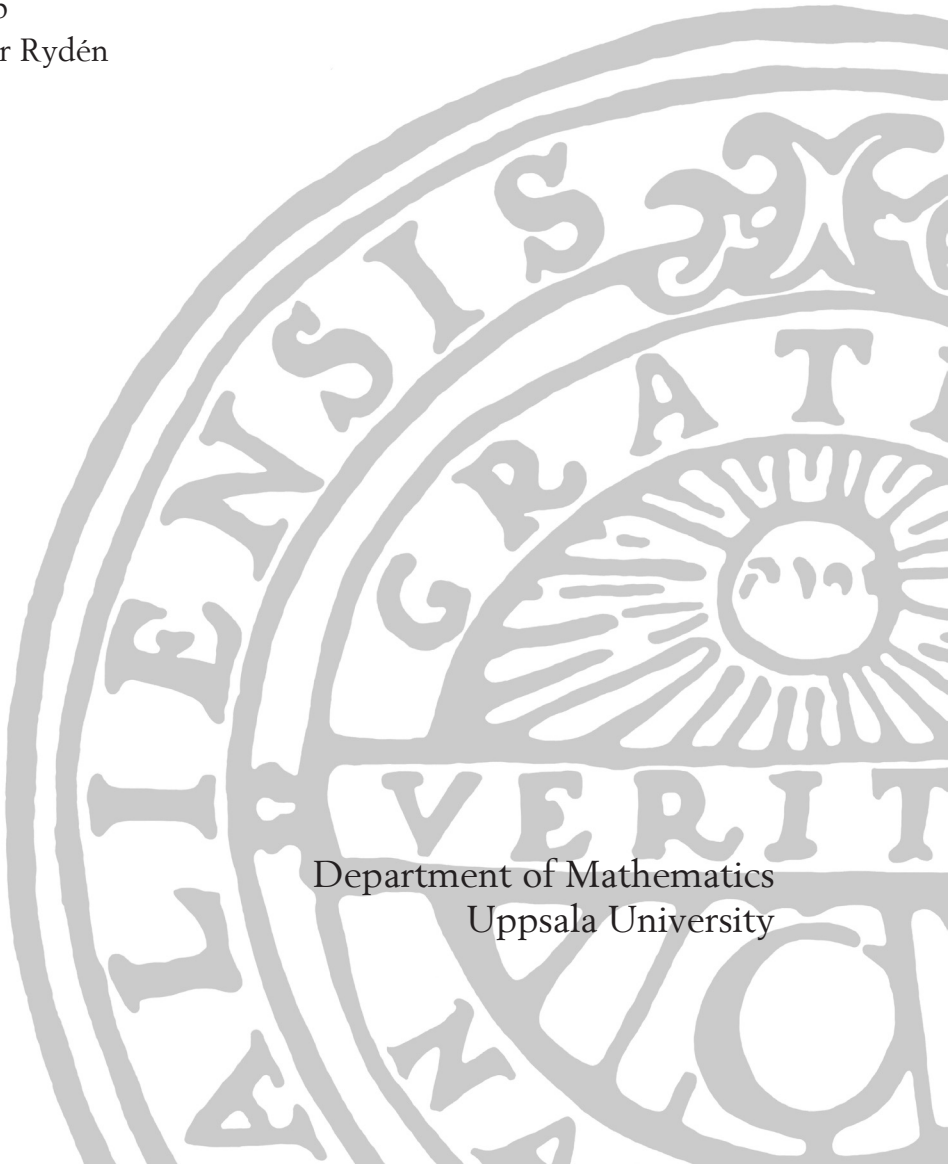
UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2014:8

Statistical studies of the Beta Gumbel distribution

Fredrik Jonsson

Examensarbete i matematik, 15 hp
Handledare och examinator: Jesper Rydén
Februari 2014



Department of Mathematics
Uppsala University

Statistical studies of the Beta Gumbel distribution

Fredrik Jonsson

February 28, 2014

Abstract

Nadarajah and Kotz (2004) introduced a generalization of the Gumbel distribution, referred to as the Beta Gumbel (BG) distribution. It was demonstrated that the BG provided more flexible tail behaviour compared to the Gumbel distribution. We will study the BG and compare it to the Gumbel distribution. We will also model extreme rainfall for series of length 51 years with the BG, GEV and Gumbel distribution, respectively and assess whether BG can be superior to the standard distributions in modelling tail behaviour of data. Furthermore, we derive estimates of the return periods of length 100 up to 10 000 years for the aforementioned distributions. We use the maximum likelihood to find estimates of the model parameters and use the delta method as well as bootstrapping with resampling to find approximate confidence intervals of the return periods.

1 Introduction

Generalizing distributions have been discussed frequently in statistics in problems of trying to fit and model observed data in various areas. A generalized class of the Beta distribution was first given by Eugene et al. (2002) in [4], where the Beta Normal distribution was introduced as a generalization of the Normal distribution. In comparison to the classical normal distribution this generalization had greater flexibility of the shape of the distribution.

Generalized distributions (often) have potential to yield significantly better fits than the classical distributions and in this thesis we will investigate a generalization of the Gumbel distribution, referred to as the Beta Gumbel (BG) distribution. The standard Gumbel distribution has been widely applied in engineering and other fields, especially in extreme-value analysis. In [6] (2004), Nadarajah and Kotz introduced the Beta Gumbel distribution and showed that it had greater flexibility in explaining the variability of the tails than the Gumbel distribution. It was demonstrated that the extra parameters gave greater variability of the skewness and kurtosis.

Many examples in the research literature show that weather variables such as precipitation (heavy rain) display heavy-tail behaviour. To study

the applicability of modelling data with tail-behaviour we will study the BG distribution in comparison to the Gumbel distribution. We decide to compare BG to the Gumbel distribution as well as to the generalized extreme distribution (GEV) and study the goodness of fit.

In many fields of engineering geographical variables such as hurricanes, flooding, precipitation or earthquakes have to be modelled to estimate the probability of rare (extreme) events. For instance, a dam must be dimensioned to withstand rare extreme flooding (due to extreme rainfall). To estimate the probability and the magnitude of events (extreme rainfall) one may use extreme value distributions as model based on historical records. In risk analysis one often talks about the return period which is an estimate of rare events which has to be estimated on the basis of historical data.

A return period of 100 years (or 100-year flood) is said to have 1 % risk of being exceeded on the average in any given year. Usually, one is interested in estimation of return periods of 100 or 1000 years and model selection has to be done with care.

The paper is organized as follows, first we present the basics of extreme value theory. In Section 3 we give the definition of the Beta Gumbel distribution, and in the following Section 4 we provide the means of calculating the quantiles, confidence intervals and discuss return periods and how to fit the BG to a given sample. In Section 5 we perform some simulation studies to observe the behaviour of the BG and Gumbel distribution. Next, in Section 6, we fit annual maxima of daily rainfall data to the Gumbel and BG distribution. We will also fit the rainfall data to the generalized extreme value distribution (GEV). Furthermore, we compare the three different distributions in terms of modelling diagnostics and how well they fit real data. We also provide the point estimates and corresponding confidence intervals in Section 6.4 for the 100-, 500-, 1000-year return periods for each distribution. Finally, in Section 6.5, we investigate the behaviour of longer return periods of up to 100,000 years. Conclusions are given in Section 7.

2 Extreme-value modelling

In extreme value modelling one is interested in finding the distribution of a series containing the maxima (or minima) over regular timed measurements. We begin by formulating the basics of extreme value theory from (Coles) [2]. Study the statistical behaviour of

$$M_n = \mathbf{max}\{X_1, \ldots, X_n\}$$

where $X_1, \ldots, X_n$ is a sequence of independent and identically distributed variables. Each observation X_i is measured at equidistant intervals for example hourly, daily, or weekly - the selection of the interval distance will not

be discussed thoroughly. In most cases the observations come from a common distribution F, where F is an unknown distribution and consequently the exact behaviour of $\{X_i\}_{i=1}^n$ is usually hard to find.

This is known as the classical block maxima model for extremes, where we group a series into m blocks each of period n. For example, if we choose a block size of $n = 365$ days, M_n corresponds to the annual maximum. We thus study a series containing maxima i.e. $M_{n,1}, \ldots, M_{n,m}$. In the example above where the period was chosen as 365 days the series $M_{365,1}, M_{365,2}, \ldots M_{365,p}$ correspond to the annual maxima over a length of p years.

It can be proved that under suitable conditions, the distribution of M_n can be approximated for large values of n. This is a result of the extremal types theorem, which states that the distribution of M_n belongs to a single family of distributions, regardless of the unknown distribution F.

The extremal types theorem says that if there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\mathbf{P}\left\{(M_n - b_n)/a_n \leq x\right\} \to G(x) \quad \text{as } n \to \infty,$$

where G is a non-degenerate distribution function, then the distribution G belongs to one of the following three families of distributions: Gumbel, Fréchet and Weibull, respectively. We also call them the extreme value distribution of type I, II and III, respectively. These families of distributions can be combined into one single family of distributions called the Generalized Extreme Value (GEV) distribution. The distribution of GEV has the following form

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

where x is defined for $1 + \xi(x-\mu)/\sigma > 0$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$, where μ is a location parameter, σ a scale parameter, and ξ a shape parameter. For $\xi = 0$ Eq. (1) is understood as a limit with distribution function

$$G(x) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, \tag{2}$$

with two parameters. This distribution Eq. (2) corresponds to the Gumbel family distribution or the extreme value distribution of type I.

The extremal types theorem may be used under certain regularity conditions – for example the selection of the block size must be chosen wisely. A too short block size may violate the conditions under which the limit exists. On the other hand having larger blocks instead generates fewer blocks and we incur larger variance when we make estimates. A standard block size is

yearly blocks since this mitigates any seasonality effect. We usually assume that time series of annual length are time-homogeneous, where we record the observations at intervals of days. In this paper, we will study extreme annual daily rainfall for two different locations, and investigate the annual maxima with different alternatives to the extreme value distribution.

3 Generalization of the Gumbel distribution

3.1 Gumbel Distribution

Recall that a special case of the GEV distribution in Eq. (1) is $\xi = 0$ and is called the Gumbel distribution. Its distribution function is given by

$$G(x) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, \quad -\infty < x < \infty, \tag{3}$$

where $-\infty < \mu < \infty$ and $\sigma > 0$. Nadarajah and Kotz (2004) [6] introduced a generalization of the Gumbel distribution as the Beta Gumbel distribution (BG) in hope it would attract greater applicability in engineering, demonstrating that it was more flexible than the Gumbel distribution. By adding two parameters a, b which mainly control the skewness and kurtosis it is possible the BG can explain the tail behaviour far more superior to the Gumbel distribution.

3.2 Beta Gumbel

Following [6], let G be the cumulative distribution function, then a generalized class of Beta distribution functions can be defined by

$$F(x) = I_{G(x)}(a, b) \tag{4}$$

where $I_{G(x)}(a, b)$ is the incomplete beta ratio function. In this paper we study the Beta Gumbel distribution in which $G(x)$ belongs to the Gumbel distribution. The generalization in Eq.(4) can be rewritten as

$$I_{G(x)}(a, b) = \frac{B_{G(x)}(a, b)}{B(a, b)}, \quad a > 0,\ b > 0 \tag{5}$$

where $B_{G(x)}(a, b)$ is the incomplete beta function given by

$$B_{G(x)}(a, b) = \int_0^{G(x)} t^{a-1}(1-t)^{b-1}\,\mathrm{d}t, \quad a > 0,\ b > 0. \tag{6}$$

If we let G correspond to the Gumbel distribution in Eq.(5) with parameters a and b it gives us a generalization of the original (parental) distribution G which we call the Beta Gumbel distribution, $\mathrm{BG}(\mu, \sigma, a, b)$. For the special

case where $a = 1$ and $b = 1$ the distribution coincides with the Gumbel distribution. We can now define the probability density function as

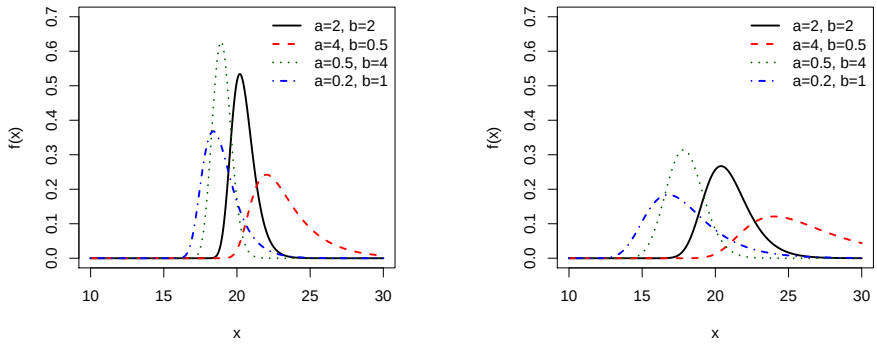
$$f(x) := F'(x) = \frac{d}{dx}\frac{1}{B(a,b)}\int_0^{G(x)} t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$$
$$= \frac{g(x)}{B(a,b)}G(x)^{a-1}\left[1-G(x)\right]^{b-1} \tag{7}$$

where $g(x)$ is the density function of the parental distribution. From Eq.(7) it follows that the density function of the Beta Gumbel distribution is given by

$$f(x) = \frac{1}{\sigma B(a,b)}ue^{-au}\left[1-e^{-u}\right]^{b-1} \quad -\infty < x < \infty, \tag{8}$$

for $-\infty < \mu < \infty$, $\sigma > 0$, $a > 0$, and $b > 0$, where $u = \exp\{-(x-\mu)/\sigma\}$.

In Figure 1, the density function of the (BG) is plotted for different values of the parameters. In the left figure $\mu = 20$ and $\sigma = 1$ and in the right figure $\mu = 20$ and $\sigma = 2$ where we vary the parameters a and b. It can be noted that the parameter b affects the skewness highly whenever b is chosen below 1, close to 0. In simple terms one could say that the b parameter is sensitive in terms of skewness of the density curve. This was also noted in the original paper [6] where a low value of b rapidly amplifies the skewness and kurtosis of the density function.



(a) $\mu = 20$, $\sigma = 1$. (b) $\mu = 20$, $\sigma = 2$.

Figure 1: The density function of the BG-distribution for different values of a, b and σ with μ held fixed at 20.

4 Estimation techniques

Consider a random sample $x_1, \ldots, x_n$ that are i.i.d observations from Eq.(8). We use the method of maximum likelihood to find the best estimates of the parameter vector $\Theta = (\mu, \sigma, a, b)$. The log-likelihood function following [6] is

$$\begin{aligned}\log \mathcal{L}(\mu, \sigma, a, b \mid x) = &- n \log \sigma + (b-1) \sum_{i=1}^{n} \log \left[1 - \exp \left\{ - \exp \left(- \frac{x_i - \mu}{\sigma} \right) \right\} \right] \\ &- \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} - a \sum_{i=1}^{n} \exp \left(- \frac{x_i - \mu}{\sigma} \right) - n \log B(a, b).\end{aligned} \tag{9}$$

Taking the partial derivatives of Eq.(9) for each parameter we obtain [1],

$$\begin{aligned}\frac{\partial \log \mathcal{L}}{\partial \mu} = &\frac{n}{\sigma} - \frac{a}{\sigma} \sum_{i=1}^{n} \exp \left(- \frac{x_i - \mu}{\sigma} \right) \\ &+ \frac{b-1}{\sigma} \sum_{i=1}^{n} \frac{\exp(-(x_i - \mu)/\sigma) \exp\{-\exp(-(x_i - \mu)/\sigma)\}}{1 - \exp\{-\exp(-(x_i - \mu)/\sigma)\}},\end{aligned}$$

$$\begin{aligned}\frac{\partial \log \mathcal{L}}{\partial \sigma} = &- \frac{n}{\sigma} + \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} \left\{ 1 - a \exp \left(- \frac{x_i - \mu}{\sigma} \right) \right\} \\ &+ \frac{b-1}{\sigma^2} \sum_{i=1}^{n} \frac{(x_i - \mu) \exp(-(x_i - \mu)/\sigma) \exp\{-\exp(-(x_i - \mu)/\sigma)}{1 - \exp\{-\exp(-(x_i - \mu)/\sigma)\}},\end{aligned}$$

$$\frac{\partial \log \mathcal{L}}{\partial a} = n\psi(a+b) - n\psi(a) - \sum_{i=1}^{n} \exp \left(- \frac{x_i - \mu}{\sigma} \right),$$

$$\frac{\partial \log \mathcal{L}}{\partial b} = n\psi(a+b) - n\psi(b) + \sum_{i=1}^{n} \log \left[1 - \exp \left\{ - \exp \left(- \frac{x_i - \mu}{\sigma} \right) \right\} \right].$$

The best estimates of μ, σ, a and b are found by setting the partial derivatives to zero and solving the subsequent simultaneous equations.

[1] Note that $\frac{\partial \log \mathcal{L}}{\partial \mu}$ is slightly different from the one in [6] which is likely due to a misprint.

4.1 T-year return periods

In extreme analysis one is interested in estimation of the T-year return period, which is defined to be the value x_T that will on average be exceeded once over a period of T years (time units). The T-year return period can be found by solving the equation,

$$F(x_T) = 1 - 1/T \tag{10}$$

where F is the cdf. Solving the equation for x_T by inverting the cumulative distribution function can sometimes be difficult or impossible if no closed formula exists. For the case of continuous distributions the inverse of a cdf is usually a well-defined function on (0,1) and an analytical function may sometimes be found.

4.2 Confidence intervals

Denote the maximum likelihood estimate of Θ as $\tilde{\Theta}$. One can show that under suitable regularity conditions as n is large, $\tilde{\Theta}$ is asymptotically normal distributed. In most cases we are interested in estimation of functions of Θ. If the regularity conditions are satisfied a result with use of Taylor's formula enables us to find estimates of functions of the maximum likelihood estimates (MLE) . The result says that an estimate of a function say $g = g(\Theta)$ is simply found by $g(\tilde{\Theta})$. In particular the return period can be viewed as a function giving us a tool to construct approximate confidence intervals. This method is commonly referred to as the delta method, which we will present for the particular case where F belongs to the classical Gumbel distribution. The inverse of the Gumbel cumulative distribution function for x_T is,

$$x_T = \mu - \sigma \ln\left(-\ln\left(1 - 1/T\right)\right), \quad T > 1 \tag{11}$$

and a confidence interval for the T-year estimate with approximately $1 - \alpha$ confidence is given by

$$x_T = [\tilde{x}_T \pm \lambda_{\alpha/2}\tilde{\sigma}], \tag{12}$$

with variance

$$V(x_T) = \nabla x_T(\mu, \sigma)^T V \nabla x_T(\mu, \sigma), \tag{13}$$

where V is the covariance-variance matrix evaluated at $(\tilde{\mu}, \tilde{\sigma})$, and

$$\nabla x_T = [\frac{\partial x_T}{\partial \mu}, \frac{\partial x_T}{\partial \sigma}]. \tag{14}$$

V can be approximated by $\hat{\Sigma} = [-\ddot{l}(\hat{\mu}, \hat{\sigma})^{-1}]$.

Since we have no analytical formula for the quantile function of the Beta Gumbel distribution we cannot apply the delta method. Instead we use

bootstrapping with resampling to estimate the standard errors to find approximate confidence intervals. The delta method and bootstrapping technique will be used thoroughly in this paper.

5 Simulation studies

In this section we will we make simulations to get a grasp of the Beta Gumbel distribution. To compare with the Gumbel distribution, we will investigate the behaviour of the parameters a and b. Recall that the special case of the Beta Gumbel distribution happens when a and b equal to one.

To study this, we generated a sample of random numbers from the classical Gumbel distribution and used maximum likelihood estimates to fit a Beta Gumbel distribution as well as the Gumbel distribution (the original) to this sample. The computations were done using R and to find the maximum likelihood estimates we used two optimizing methods; BFGS a quasi-Newton method (1970) and Nelder-Mead a simplex algorithm that can be applied to non-differentiable functions as well. It proved to be difficult to find maximum when we applied the log-likelihood function of the BG due to the calculations of the gradient. For instance providing the exact gradients of the log-likelihood function did not give us maximum but local extreme points (saddle points). Instead we let the in-built methods carry out the calculations of the gradient with a finite-difference approximation. This alternative way works reasonably well. We used both the BFGS and Nelder-Mead algorithm interchangeably.

5.1 Quantile function

The BG distribution can be written as a composition of functions. To see this, BG can be expressed as a composed function where $F(X) = F_{Beta}(G(X))$, where G is the cdf of the parental distribution function. To find a random number X using the uniform distribution U it suffices to solve $X = F^{-1}(U)$.

$$\begin{aligned} F_{Beta}(G(X)) &= U \\ F^{-1}_{Beta}(G(X)) &= F^{-1}_{Beta}(U) = B \\ X &= G^{-1}(B) \end{aligned} \tag{15}$$

where the inverse of the Gumbel distribution is $G^{-1}(x) = \mu - \sigma \log[-\log(x)]$. With the same argument and using that the distribution function of the Beta Gumbel is right-continuous and strictly increasing on $p \in (0, 1)$ for $F^{-1}(p)$ the quantile function of BG is

$$Q_{BG} = \mu - \sigma \log\{-\log[Q_{Beta(a,b)}(p \mid a, b)]\} \tag{16}$$

for $p \in (0,1)$, where $Q_{Beta(a,b)}(p \mid a,b)$ is the quantile function of the Beta distribution, with $p = 1-q, q = 1/T$. Note that the quantile function of the Beta distribution must be calculated numerically.

5.2 Parameter estimates

In the first case we simulated 1000 samples of 1000 random numbers from the Gumbel distribution with parameters $\mu = 10$, and $\sigma = 2$. To compare with the Beta Gumbel distribution we then find the maximum likelihood estimates of these samples. We did not supply the gradient but let it be approximated. In Figure 2 we provide the histograms of each estimated parameter of the BG distribution.

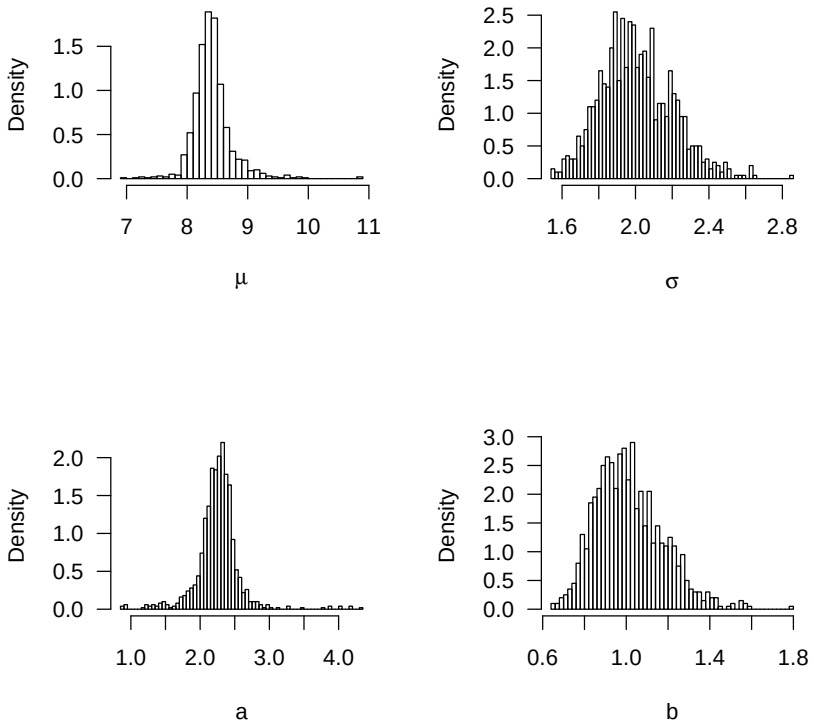


Figure 2: Histograms of parameter estimates with original $\mu = 10$, $\sigma = 2$.

We see from the Figure 2 that μ is somewhat close to the original μ of the Gumbel distribution, furthermore σ is almost identical. We note that the extra parameters a has slightly larger variance than the parameter of

b. See also Table 1. Indeed, the parameters a and b are not simultaneously equal to one, meaning that the BG does show some flexibility in modelling.

	μ	σ	a	b
Estimate	8.41	2.01	2.25	1.01
Variance	0.11	0.038	0.099	0.026

Table 1: Parameter estimates as $\mu = 10$ and $\sigma = 2$.

We repeated the same computations for the parameters $\mu = 20$ and $\sigma = 5$, simulating 1000 samples each with 1000 random numbers from the Gumbel distribution, then fitting the BG distribution to each sample.

	μ	σ	a	b
Estimate	19.76	5.00	1.13	1.01
Variance	2.65	0.17	0.30	0.021

Table 2: Parameter estimates as $\mu = 20$ and $\sigma = 5$.

In Table 2 the estimated parameters a and b are close to 1.0, in which case the BG does not provide flexibility, since the fitted BG is that of the original Gumbel distribution. This is also seen in Figure 3.

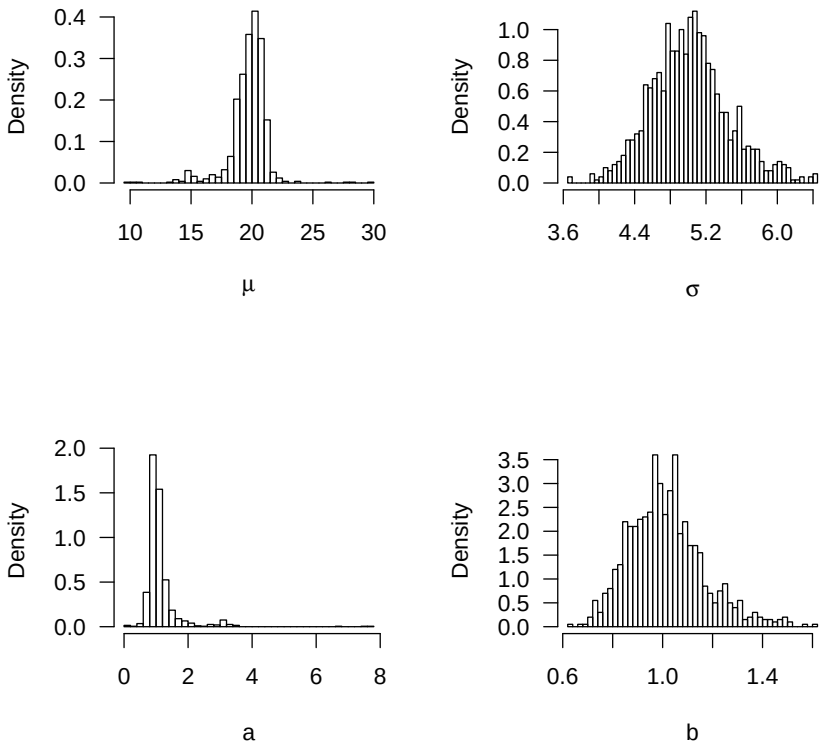


Figure 3: Histograms of parameter estimates with original $\mu = 20$, $\sigma = 5$.

As can be noticed in Figure 3 the estimated parameter a has some outliers for some samples. A provided boxplot in Figure 4 confirms this, only a few of the parameter estimates of a are above 4.

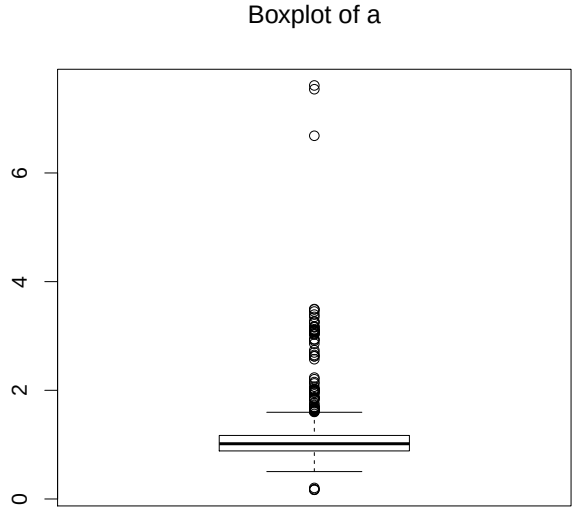


Figure 4: Boxplot of the estimated parameter a from 5000 samples with Gumbel($\mu = 20$, $\sigma = 5$).

5.3 Quantile estimation

In this section we discuss the behaviour of the quantiles of the BG and the Gumbel distribution. For calculations of the quantiles of the Beta Gumbel distribution we use the formula in Eq.(16), but this gives us by no means a methodology to find standard errors (the formula is not closed). For that we had to use bootstrapping techniques to find approximate confidence intervals. Estimates of the quantiles and the corresponding standard errors of the Gumbel distribution was calculated using the delta method. The simulation was carried out as follows, in principle we simulated one *single* sample of 1000 numbers from the Gumbel distribution. For the bootstrapping procedure we bootstrapped this sample 5000 times with resampling. In the Table 3 the mean and variance of these 5000 estimates are given. We also provide the histograms of these estimates in Figure 5.

	μ	σ	a	b
Estimate	18.38	4.28	1.45	0.86
Variance	3.86	2.80	1.18	0.83

Table 3: Parameter estimates after 5000 bootstraps, $\mu = 20$ and $\sigma = 5$.

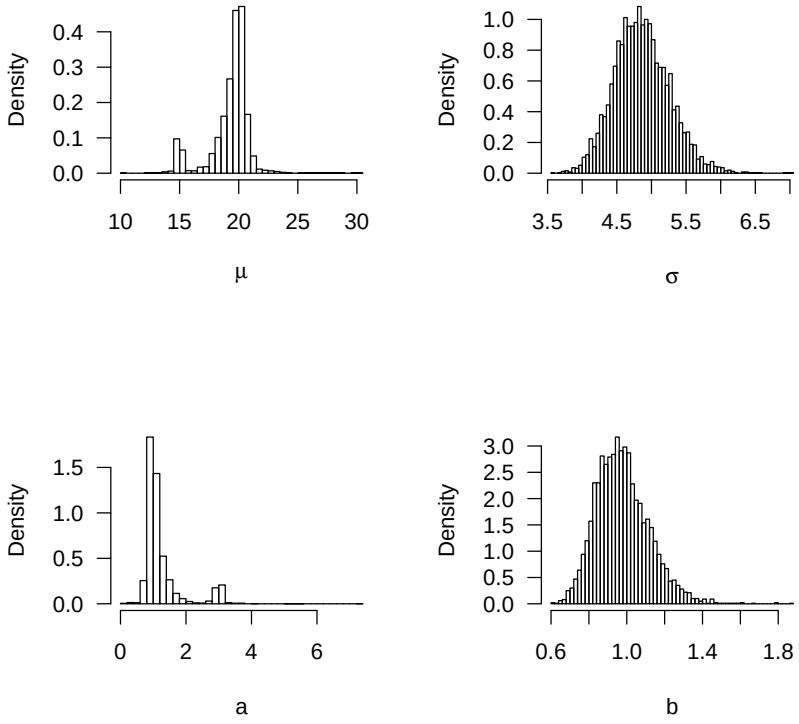


Figure 5: Histogram of the parameters after 5000 bootstraps.

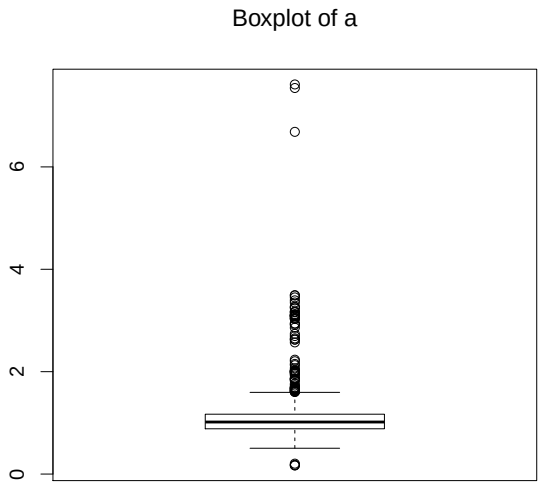
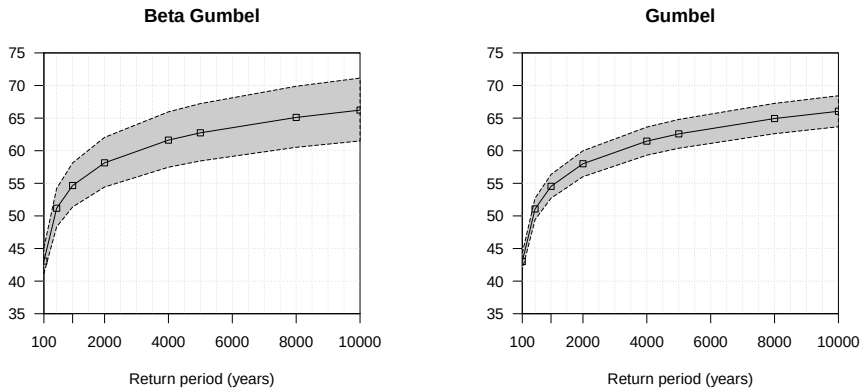


Figure 6: Boxplot for the single estimated parameter a from 5000 bootstraps.

In Figure 6 we see there are some outliers for the estimates of the parameter a.

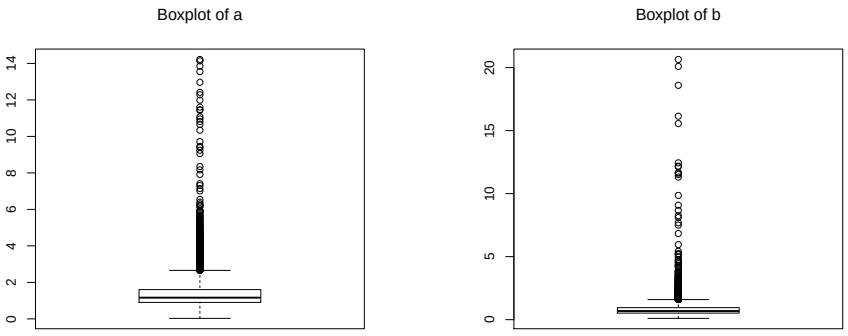
To study the behaviour of the return periods of both BG and Gumbel of longer return periods, we calculate the return periods of 100 up to 10,000 years and the corresponding confidence intervals. In Figure 7 we notice that the Beta Gumbel has wider confidence intervals. From a reliability point of view, this may be of interest when we want to estimate an upper bound used as a threshold value.



(a) Return periods for Beta Gumbel. (b) Return periods for Gumbel.

Figure 7: Return periods and confidence intervals from 100 of up to 10,000 years.

It should be mentioned that these exact computations for samples of 100 random numbers instead of 1000 yielded indefinite values, which is likely due to numerical issues. The problem originates from the calculation of the inverse of the Beta Gumbel for different sets of parameter estimates. In this case, 5000 bootstrapping samples gives sets of parameter estimates that gives difficulty with computations of the inverse of the Beta Gumbel distribution function. A provided boxplot in Figure 8 gives reasonable doubt of the estimates a and b for some of the parameter sets after bootstrapping.



(a) Estimated parameter a. (b) Estimated parameter b.

Figure 8: Boxplots for some of the estimated parameters after bootstrapping 5000 times.

6 Real data

Investigating the applicability of modelling real data we will study annual maximum daily rainfall in Sweden for two different locations; Stockholm and Härnösand. The series of annual maximum daily rainfall extends from 1961 to 2011 and was retrieved from `http://www.hurvarvadret.se` which handles weather data with courtesy of SMHI, Swedish Meteorological and Hydrological Institute. The locations, Stockholm and Härnösand lie in areas of Sweden where some of the most extreme rainfall events (defined as at least 90 mm precipitation during 24 hours) have occurred, especially the latter one [11]. Most extreme rainfall events in Sweden happen in the region of Svealand and in the southern coast of Norrland; regions in which the investigated locations are situated.

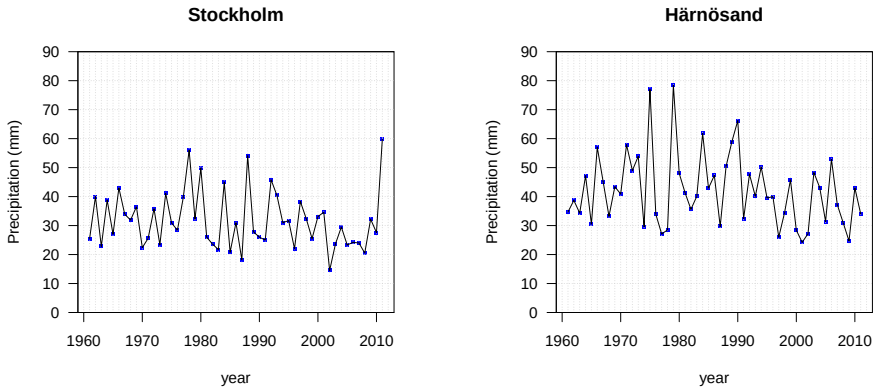
6.1 Tools of measurements and error sources

Precipitation can be measured in two main ways, either by gathering and measuring the fallen precipitation in a fixed location, say one weather station, or by using numerous weather stations scattered around a large area of at least 1000 km^2 and combining data from all stations in the location and finding the most extreme one.

Measuring the amount of rainfall is done by rain gauges which gathers and measures the accumulated amount of liquid over a specific period of time. Due to limitations the amount of precipitation cannot be measured accurately. During hurricanes or windy weather it is difficult to gather the rainfall which leads to underestimation of the precipitation. Moreover, any evaporation will reduce the amount of measured precipitation. In numbers the total underestimation is on average of 5-10 %. In winter any snow gathered by the instrument will be melted and the melted water is measured.

For exact definitions how precipitation is measured see [9]. Extreme precipitation is defined by SMHI if the accumulated precipitation over the last 24 hours exceeds 90 mm. [10].

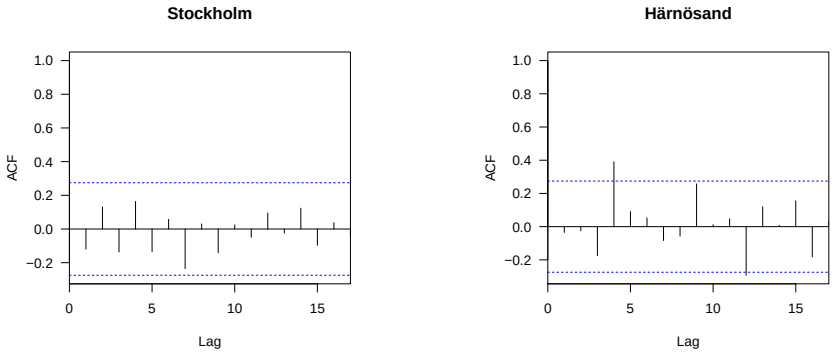
6.2 Introductory analysis



(a) Stockholm. (b) Härnösand.

Figure 9: Maximum daily rainfall records in Stockholm (left) and Härnösand (right).

In Figure 9, the time series of annual maxima is plotted for Stockholm and Härnösand, respectively. We use the same scale, and, evidently Härnösand has on average higher maximum daily rainfall. Furthermore, we notice that there is no apparent trend. We also provide the ACF plots up to lag 15 in Figure 10 to see whether there is dependence between the observations. By visual inspection we see that most autocorrelation values are close to 0. On the other hand the ACF of Härnösand shows a cut-off at lag 4, but the dependence is weakly so there is no apparent reason to justify non-stationarity of this series.

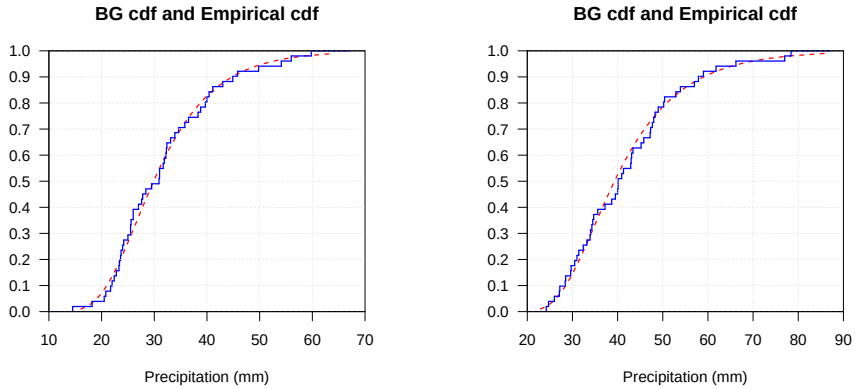


(a) Stockholm. (b) Härnösand.

Figure 10: ACF for both sample sets.

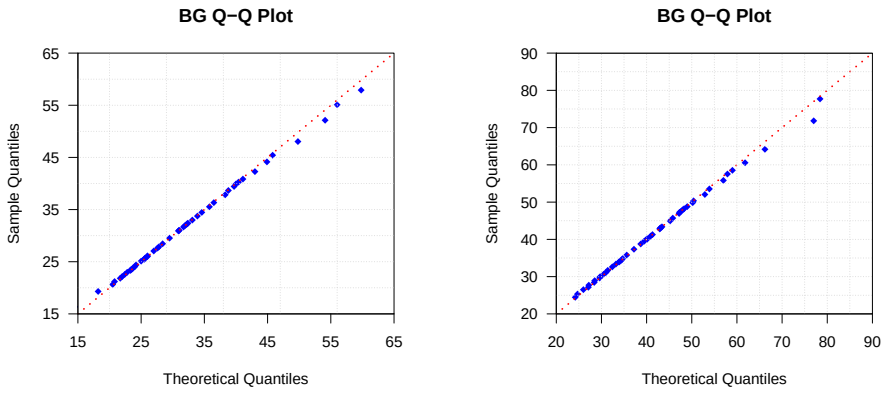
6.3 Model Diagnostics

A main problem is to choose from a variety of models and assess which model is the most appropriate one. Therefore, it is of interest to check that the chosen model fits well to the particular data. We use two graphical methods to assess the goodness-of fit. In Figure 11 the empirical distribution and the fitted BG model is plotted. Our model agrees reasonably well with the empirical cdf. We also provide the QQ-plot in Figure 12, and notice no apparent departures from the unit diagonal except at a few points for dataset 2 (Härnösand). From these plots we can not reject that the Beta Gumbel distribution is a good candidate model.



(a) Stockholm. (b) Härnösand.

Figure 11: Empirical distribution versus CDF.



(a) Stockholm. (b) Härnösand.

Figure 12: QQ-plot.

To study how well BG competes with other candidate models such as the Gumbel distribution and the GEV, we will perform likelihood-ratio tests and use the Akaike information criterion (AIC).

	$\log \mathcal{L}(\hat{\Theta})$
Dataset 1	
BG	-184.0948
GEV	-184.1427
Gumbel	-184.1654
Dataset 2	
BG	-195.9717
GEV	-196.1058
Gumbel	-196.1798

Table 4: The maximum log-likelihood values for each distribution.

The maximum log-likelihood values for each distribution are given in Table 4. For these data sets we had to use the Nelder-Mead algorithm for optimizing the log-likelihood function of Beta Gumbel. We notice that the highest log-likelihood value is obtained by fitting the BG. However, the differences in the log-likelihood values between the distributions are very small. We can use the log-likelihood ratio test to check whether one higher order parameter model describes the variability significantly better. A log-likelihood statistic is $D = 2[\log(M_1) - \log(M_0)]$ where M_0 is a reduction of the model M_1. The statistic D is chi-square distributed with $p-k$ degrees of freedom, where p and k is the dimension of the parameter space of M_1 and M_0, respectively. The null hypothesis is rejected if $D > \chi^2_{p-k}$, favouring the M_1 model which describes the variability of the data significantly better. Comparing the Gumbel distribution to the GEV distribution is equal to testing

$$H_0 : \xi = 0 \qquad \text{against} \quad H_1 : \xi \neq 0,$$

and note that the test statistic is very small $2(-184.1427 - (-184.1654)) = 0.0454$ with p-value of almost 1. Similarly for dataset 2 the statistic is very small. We can also test the Gumbel against the BG distribution since it is a reduction of the BG distribution of the parameter space $a \times b$

$$H_0 : a = 1,\ b = 1 \qquad \text{against} \quad H_1 : a \neq 1,\ b \neq 1$$

with $4 - 2 = 2$ d.f. For dataset 2 the test statistic is $2(-195.9717 - (-196.1798)) = 0.4162$ with p-value of about 0.81. This suggests that the

Gumbel distribution is adequate for modelling the data equally well as the BG distribution. The likelihood ratio test does not give us any findings whether any distribution models the data significantly better.

In statistics it is not desirable to use complicated models, and most frequently the simplest model is most likely to be correct, and one can test whether the more complicated model explains the variability significantly better. To test whether one model with higher number of parameters models the data significantly better than another candidate model with lower parameters, we can use the Akaike information criterion that is a test statistic that penalizes over-fitting. The test statistic is given by $\mathrm{AIC} = -2\log\mathcal{L}(\hat{\Theta}) + 2p$, where p is the number of model parameters.

		Parameter estimate				
	AIC	μ	σ	a	b	ξ
Dataset 1						
BG	372.1896	16.53	6.44	3.89	0.77	
GEV	371.2854	27.19	7.51			0.023
Gumbel	370.3308	27.28	7.57			
Dataset 2						
BG	395.9434	19.17	6.50	6.43	0.56	
GEV	395.2116	36.16	9.36			0.048
Gumbel	394.3596	36.41	9.55			

Table 5: AIC values and parameter estimates of the different distributions.

For both datasets the Gumbel distribution has the lowest AIC whereas the AIC of the BG and the GEV are almost equal, meaning that the models are indistinguishable in terms of modelling. Note that in Table 5 the scale parameter ξ for GEV, is almost zero. To test whether ξ is significantly nonzero we use that the confidence interval for a maximum likelihood estimate of a single parameter is $\tilde{\xi} \pm \lambda_{\alpha/2}\sigma_{\xi}$. We find that the standard error of ξ is 0.13 and we can reject the null hypothesis that ξ is nonzero with 1 % significance. So 0 lies inside a 99 % confidence interval for ξ. Therefore, one may argue that the GEV model should be rejected and that the Gumbel distribution models the data equally well. A reduction would be preferable here, but for the sake of comparison we will keep the GEV. It should be mentioned that the estimated parameters a and b of the BG for both sets are not simultaneously equal to one, meaning that BG does indeed show some flexibility in modelling data with tail-behaviour.

6.4 Results

In this section we provide the estimated return periods for each distribution and the corresponding confidence intervals. The confidence intervals of the GEV and Gumbel distribution was derived as usual with the delta method. To find approximate confidence intervals of the BG distribution we used bootstrapping with resampling. Both samples were resampled 2000 times. As usual this same procedure was done in Section 5.

	Return levels with 95% C.I.		
	$T = 100$ (x_{100})	$T = 500$ (x_{500})	$T = 1000$ (x_{1000})
Dataset 1			
BG	64.2 (51.2, 78.0)	77.6 (58.5, 97.6)	83.4 (61.7, 106.0)
GEV	63.6 (46.2, 81.0)	77.3 (45.4, 109.2)	83.3 (43.6, 123.1)
Gumbel	62.0 (53.6, 70.6)	74.3 (63.2, 85.4)	79.5 (67.4, 91.7)
Dataset 2			
BG	85.8 (66.5, 105.9)	104.5 (74.6, 133.6)	112.5 (77.9, 145.7)
GEV	84.4 (57.6, 111.1)	104.0 (52.1, 155.8)	112.9 (47.0, 178.8)
Gumbel	80.3 (69.5, 91.1)	95.7 (81.6, 109.8)	102.3 (86.8, 117.9)

Table 6: Return level estimates and corresponding confidence intervals.

From Table 6, we notice that the both GEV and BG distribution give higher estimates of the return periods; especially for the 1000-year return period. Moreover, the estimates do not differ largely from a practical point of view. The Gumbel gives lower values of the return period, thus we see that GEV and BG are more conservative in estimating the return period. This is valid for both datasets.

6.5 Longer return periods

Next we investigate the behaviour of longer return periods for the different distributions. Estimating longer return periods is only meaningful if the assumption of stationarity is valid but it is nonetheless useful to talk about return periods of up 10,000 years. For instance in the construction of a dam, it is not unusual that it must be designed to withstand extreme rainfall events with return periods of 2,000–20,000 years as means of flood protection.

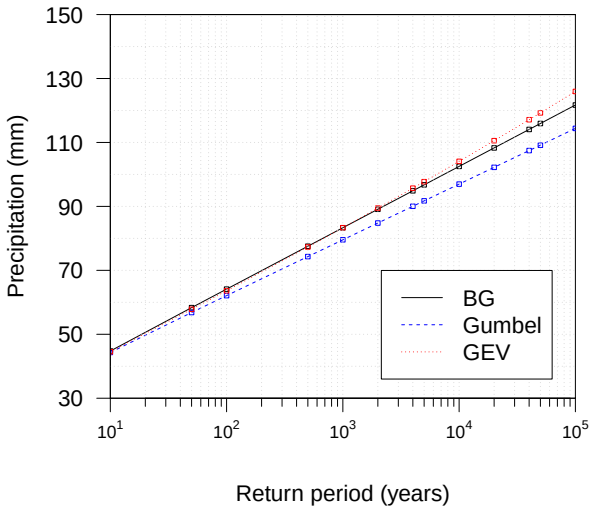


Figure 13: Return levels for Stockholm.

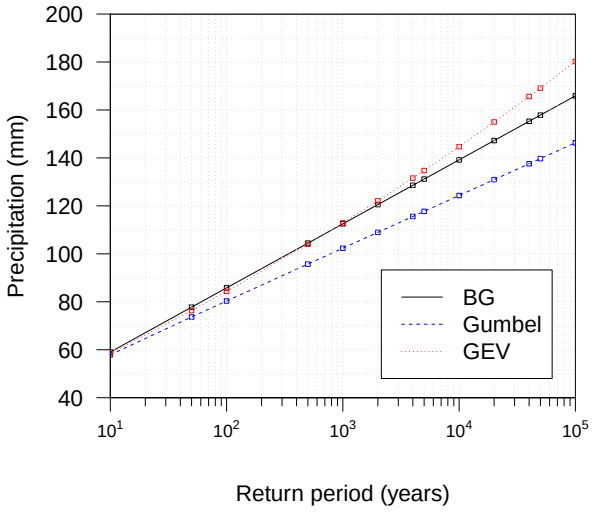


Figure 14: Return levels for Härnösand.

As noted before we see in Figure 13 and 14 that the Gumbel distribution gives lower estimates and that the GEV gives higher estimates compared to the BG distribution for longer return periods.

7 Conclusions

We have studied one generalization of the Gumbel distribution, the Beta Gumbel distribution with two additional parameters which allows skewness and varying of the tail weights. Applications of real data show that the BG does indeed provide more flexibility than the Gumbel distribution in modelling data with heavy-tail behaviour.

However, we encountered problems finding confidence intervals using bootstrapping for some parameter sets. When applying the BG to real data we had problem finding standard errors based on bootstrapping. This was related to the calculation of the inverse of the Beta Gumbel distribution function (quantile function). The formula for the quantiles given in Eq.(16) involves the inverse of the Beta distribution function which has to be calculated numerically. After resampling the real data 2000 times, some parameter sets made it tricky to find quantiles q, $F_{Beta}(q) = p$, where p was a number near 1 since it was highly sensitive to precision errors giving us indefinite quantiles. As noted the inverse is injective only on $(0, 1)$ and for some parameter sets the inverse of the Beta distribution yielded us a value of 1.0 whenever we tried to find quantiles for p close to 1.0. This is of course not well-defined and gives us quantiles that are indefinite (infinite). All numerical computations were done in R which uses finite precision arithmetic which basically means that at most 15-20 digits are correct (or accurate) in the computations. We therefore had to rely on an alternative software that uses arbitrary-precision arithmetic to do the computations, meaning that any number of precision of digits can be used. We used Mathematica (a Computer algebra system) that uses arbitrary-precision arithmetic to calculate the inverse and find that precision of at least 30 up to 39 digits had to be used to perform the computations of the quantiles. Although a difference as negligible as 10^{-39} is almost $(1 - 10^{-39} \doteq 1)$ the logarithmic function in the formula in Eq.(16) is not defined at 1. The quantile function also involves computation of a composition of functions (logarithmic function within a logarithmic function). This gives large differences in evaluating the formula (16) for values $(0, 1)$ close to 1.

Finding maximum likelihood estimates to the log-likelihood function was also tricky since for arbitrary values of a, b maximum estimates could not be found (especially with b close to 0). If this is related to the curvature of the 4-dimensional function or a matter of numerical issue is unclear. The generalized Beta Gumbel distribution involves the incomplete beta function and also makes it more difficult to work with.

We compared BG to the Gumbel distributions as well as the more standard extreme value distribution, GEV in modelling real data. We performed likelihood ratio tests and discussed if BG could serve as a good candidate model. It should be noted that BG which is a 4-parameter model compared to the 3-parameter model of GEV makes the numerical work more problematic. Optimizing over a 4-dimensions is highly more difficult than over a 3-dimensional space. We found that the BG is useful for modelling data with tail-behaviour. However, we could not conclude whether BG is a better candidate model to use. Further work has to be done to study this distribution and its applicability.

References

[1] Alexander C., Cordeiro G., Ortega E., Sarabia J. (2011). *Generalized Beta-Generated Distributions*

[2] Coles S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London: Springer-Verlag

[3] Cordeiro G., Castro M. (2009). *A new family of generalized distributions*

[4] Eugene N., Lee C., Famoye F. (2002). Beta-Normal Distribution and its applications, *Communications in Statistics - Theory and Methods*, **31:4**, 497-512, DOI: 10.1081/STA-120003130

[5] Morais A. (2009). *A Class of Generalized Beta Distributions, Pareto Power Series and Weibull Power Series.*

[6] Nadarajah S., Kotz S. (2004). The Beta Gumbel Distribution, *Mathematical Problems in Engineering*, **2004(4)**, 323-332, DOI: 10.1155/S1024123X04403068

[7] Reiss R., Thomas M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*, Birkhäuser Verlag

[8] Rychelik I., Rydén J. (2006). *Probability and Risk Analysis: An Introduction for Engineers*. Berlin: Springer-Verlag

[9] Wern L., (2012). Extrem nederbörd i Sverige under 1 till 30 dygn, 1900 - 2011. *SMHI Meteorologi*, **143**, 5-22

[10] Extrem nederbörd, (2012, 10th of July). Retrieved July 16, 2013, from http://www.smhi.se/kunskapsbanken/meteorologi/extrem-nederbord-1.23060

[11] Extrem punktnederbörd, (2013, 4th of February). Retrieved July 16, 2013, from http://www.smhi.se/kunskapsbanken/meteorologi/extrem-punktnederbord-1.23041

[12] Svenska nederbördsrekord, (2013, 13th of February). Retrieved July 16, 2013, from http://www.smhi.se/kunskapsbanken/meteorologi/svenska-nederbordsrekord-1.6660

A Source code

A source code listing to implement in R is given for some of the basic functions related to the Beta Gumbel distribution.

```
#probability density function
f <- function(x,mu,sigma,a,b) {
        return(exp(-(x-mu)/sigma)*1/(sigma*beta(a,b))
            *exp(-a*exp(-(x-mu)/sigma))
            *(1-exp(-exp(-(x-mu)/sigma)))^(b-1))
}


#quantiles beta gumbel
qbg <- function(q, mu, sigma, a, b) {


        k<-qbeta(1-q,a,b)
        x<-mu-sigma*log(-log(k))

}


#loglikelihood beta gumbel
logl <- function (par, x) {
        mu <- par[1]
        sigma <- par[2]
        a <- par[3]
        b <- par[4]
        n <- length(x)

    return(-n*log(beta(a,b))-n*log(sigma)
            +(b-1)*sum(log(1-exp(-exp(-(x-mu)/sigma))))
            -sum((x-mu)/sigma)
            -a*sum(exp(-(x-mu)/sigma)))

    }
```

B Datasets

25.5	40.0	22.8	38.8	27.0	43.0	33.9	31.9
36.5	22.4	25.6	35.8	23.4	41.1	30.9	28.4
39.7	56.0	32.3	49.8	26.0	23.6	21.7	44.9
20.8	31.0	18.2	54.1	27.8	26.0	25.0	45.8
40.4	31.0	31.7	22.0	38.3	32.4	25.5	33.1
34.6	14.5	23.7	29.5	23.3	24.2	24.0	20.5
32.2	27.6	59.8					

Table 7: Stockholm data set.

34.7	38.7	34.3	47.2	30.5	57.0	45.2	33.2
43.4	40.9	57.9	49.0	53.9	29.6	77.0	33.9
27.1	28.5	78.4	48.0	41.3	35.6	40.1	61.8
42.9	47.3	29.7	50.4	59.0	66.2	32.4	47.7
40.1	50.2	39.5	40.0	26.0	34.5	45.8	28.4
24.2	27.2	48.3	43.1	31.4	52.9	37.2	31.0
24.7	43.0	34.0					

Table 8: Härnösand data set.