Data Analysis in Extreme Value Theory

Xin Liu
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill

April 30, 2009

1 Introduction

Extreme value theory has emerged as one of the most important statistical disciplines for the applied sciences. Extreme value techniques are also becoming widely used in many other disciplines. For example, for portfolio adjustment in the insurance industry; for risk assessment on financial markets; and for traffic prediction in telecommunications. Here we do some basic data analysis in extreme value theory.

In Section 1, we introduce the extreme value model which represents the basis of extreme value theory. We analyze an example base on daily maximum winter temperature (degrees centigrade) from 1945 through 1995 for Sept-Iles, Quebec. Then we talked about threshold model with data analysis of hourly precipitation for Denver, Colorado in the month of July from 1949 to 1990 in Section 2. Finally, we append all the figure in Section 3.

2 Extreme value distribution

2.1 Asymptotic models

The extreme value model focuses on the statistical behavior of

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

where $X_1, X_2, \ldots, X_n$ are independent random variables with identical distribution F. Analogous to central limit theorem, it can be shown that there exist constants $a_n > 0, b_n$

such that

$$Pr\left\{\frac{M_n - b_n}{a_n} < x\right\} = (F(a_n x + b_n))^n$$

converges to some $H(x)$ as $n \to \infty$.

The Extremal Types Theorem (Fisher-Tippett) asserts that if nondegenerate H exists, it belongs to one of the following families:

$$\begin{aligned} I : H(x) &= \exp\{-e^{-x}\}, \ -\infty < x < \infty \ \text{(Gumbel)} \\ II : H(x) &= \exp\{-x^{-\alpha}\}, \ x > 0; \ [H(x) = 0, \ x \le 0] \ \text{(Frechet)} \\ II : H(x) &= \exp\{-|x|^{\alpha}\}, \ x < 0; \ [H(x) = 0, \ x \ge 0] \ \text{(Weibull)}. \end{aligned}$$

Here, $\alpha > 0$ in Frechet and Weibull distributions.

The three types can be combined into a single family, known as the Generalized Extreme Value (GEV) Distribution

$$H(x) = \exp\left\{-\left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right\}, \tag{2.1}$$

where $\mu, \sigma > 0$ and ξ are location, scale and shape parameters respectively. It is straightforward to check that the type II and type III classes of extreme value distribution correspond respectively to the class $\xi > 0$ and $\xi < 0$. The subset of the GEV family with $\xi = 0$ is interpreted as the limit of (2.1) as $\xi \to 0$, leading to the Gumbel family.

2.2 Example: Daily Maximum Winter Temperature at Sept-Iles

This analysis is based on the series of daily maximum winter temperature at Sept-Iles, Quebec over the period 1945-1995, as described in Figure 1. From Figure 1 it seems reasonable to assume the data is stationary over the observation period, so we model the data as independent observations from the GEV distribution.

Using maximum likelihood estimation, we get the estimates of parameters (standard errors in parentheses):

$$\hat{\mu} = 18.20(0.50) \quad \hat{\sigma} = 3.13(0.36) \quad \hat{\xi} = -0.14(0.12).$$

Figure 2 shows the profile log-likelihood for ξ, from which a 95% confidence interval for ξ is obtained as $(-0.1396, -0.1395)$ which doesn't cover 0. Therefore, the shape parameter ξ is different from 0 at the 5% significance level.

Estimates and confidence intervals for return levels are obtained by Figure 3. The confidence intervals for 10-year, 50-year, 100-year, 150-year return levels are respectively

$$(22.88, 26.50), \quad (25.58, 33.38), \quad (26.37, 36.73), \quad (28.00, 50.17).$$

The various diagnostic plots for assessing the accuracy of the GEV model are shown in Figure 4. Both the probability plot and the quantile plot show the reasonability of the GEV fit. The return level curve asymptotes to a finite level as a consequence of the negative estimate of ξ. Finally, the correspond density estimate seems consistent with the histogram of the data. Consequently, all four diagnostics plots support the fitted GEV model.

3 Exceedances over thresholds

3.1 Threshold models

Let $X_1, X_2, \ldots, X_n$ be independent random variables with identical distribution F. It is natural to regard as extreme events those of the X_i that exceed some high threshold. Denoting an arbitrary element in the X_i sequence by X. Consider the distribution of X conditionally on exceeding some high threshold u by the conditional distribution

$$Pr\{X < u + y | X > u\} = \frac{F(u+y) - F(u)}{1 - F(u)}, \ y > 0.$$

Since, in practical application, F is unknown, we need to approximate the conditional distribution with high threshold.

The main result is the following theorem.

Theorem 3.1. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with identical distribution F. Define*

$$M_n = \max\{X_1, X_2, \ldots, X_n\}.$$

Suppose, for large n,

$$P(M_n < x) \approx H(x),$$

where H is the generalized extreme value distribution defined by (2.1). Then,
for large enough u, the distribution function of $X - u$, conditional on $X > u$, is approximately

$$G(y, \sigma, \xi) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)_+^{-1/\xi}.$$

$G(y, \sigma, \xi)$ is called the generalized Pareto Distribution(GPD).

3.2 Example: Hourly precipitation for Denver

The example is about hourly precipitation (mm) for Denver, Colorado in the month of July from 1949 to 1990. Figure 5, 6 and 7 show the scatter plots of precipitation against hour, day, and year respectively.

It is important to choose the appropriate thresholds. If the threshold is chosen too high, then there are not enough exceedances over the threshold to obtain good estimators of the extreme value parameters, and consequently, the variances of the estimators are high. Conversely, if the threshold is too low, the GPD may not be good fit to the excesses over the threshold and there will be bias in the estimations. Figure 8 (mean excess plot) and Figure 9 (parameter stability plot) suggest a threshold $u = 0.395$. Maximum likelihood estimates (standard errors in parentheses) in this case are

$$\hat{\sigma} = 0.29(0.064) \quad \hat{\xi} = -0.08(0.157).$$

The profile log-likelihood for ξ is shown in Figure 10. Figure 11 gives the diagnostics for the GDP fit.

Often, threshold excesses are not independent. For example, a hot day is likely to be followed by another hot day. In such situation, we employ a declustering scheme to filter out a set of approximately independent threshold excesses. The threshold excesses in Figure 5 (scatter plots of precipitation against hour) appears dependent. We calculate the daily maximum precipitation for each day in each year. Consider a GPD fit to the aggregated daily maximum precipitation data, and obtain the fitted parameter values (standard errors in parentheses):

$$\hat{\sigma} = 0.32(0.070) \quad \hat{\xi} = -0.12(0.152).$$

The profile log-likelihood for ξ is shown in Figure 12. Figure 13 gives the diagnostics for the GDP fit.

Decluster the original precipitation with 'declustering parameter' $r = 1$ and refit the newly declustered data to the GPD. The fitted parameters values (standard errors in parentheses) are:

$$\hat{\sigma} = 0.31(0.068) \quad \hat{\xi} = -0.10(0.154).$$

Figure 14 gives the diagnostics for the GDP fit.

There's no big difference between the three fitted models, which may imply that the exceedances above the threshold $u = 0.395$ can be regarded independent. The diagnostics show that GDP fit is reasonable here.

4 Appendix

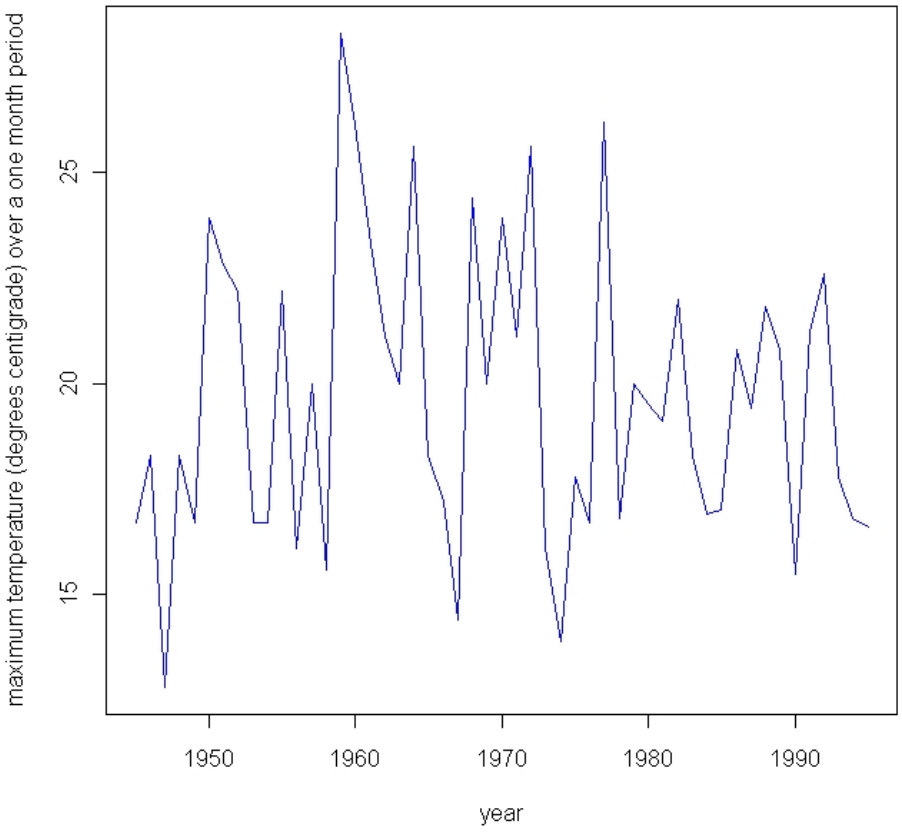

Figure 1: Daily maximum winter temperature at Sept-Iles, Quebec over the period 1945-1995.

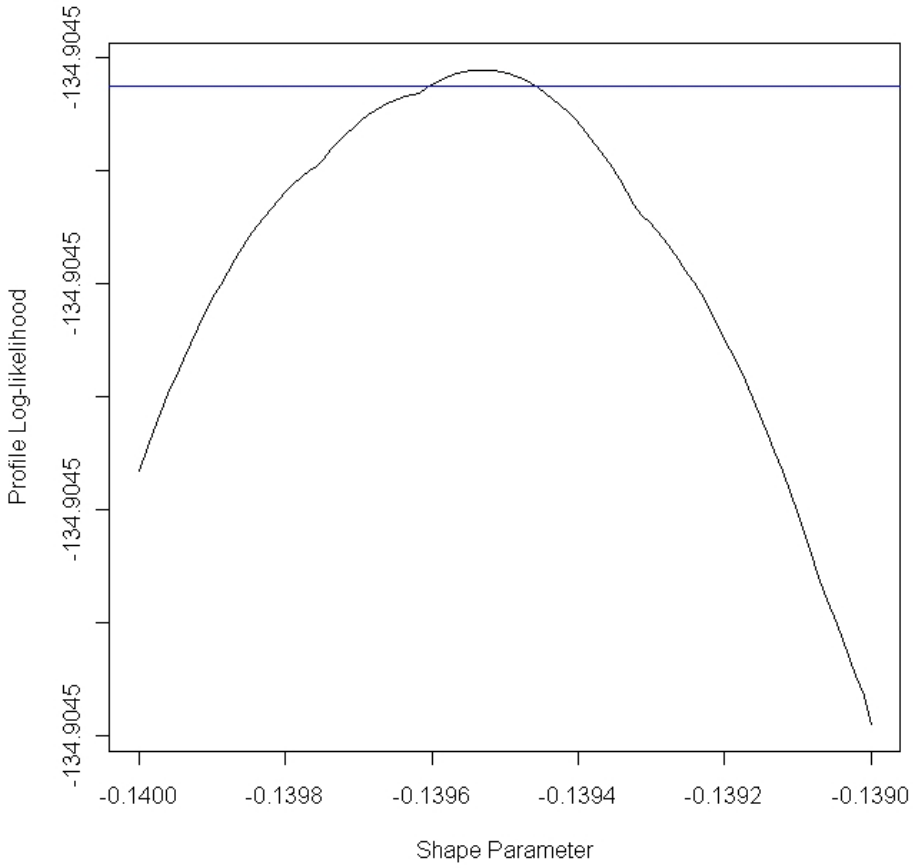

Figure 2: Profile likelihood for ξ in the Sept-Iles example.

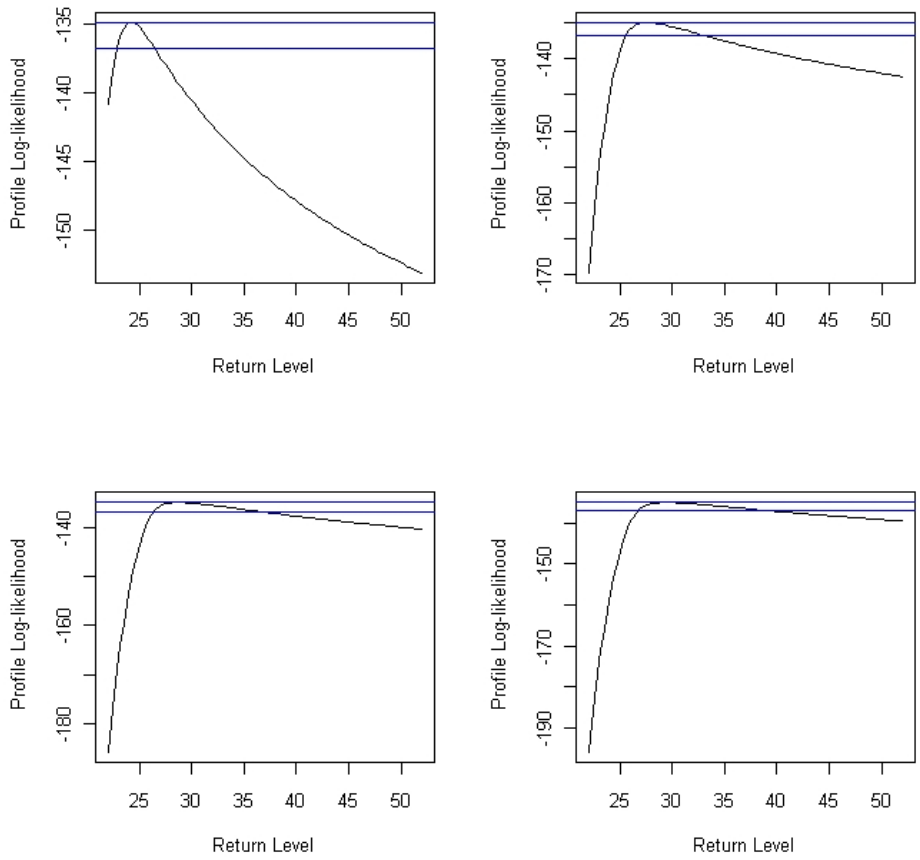

Figure 3: Profile likelihood for 10-year, 50-year, 100-year, 150-year return level in the Sept-Iles example.

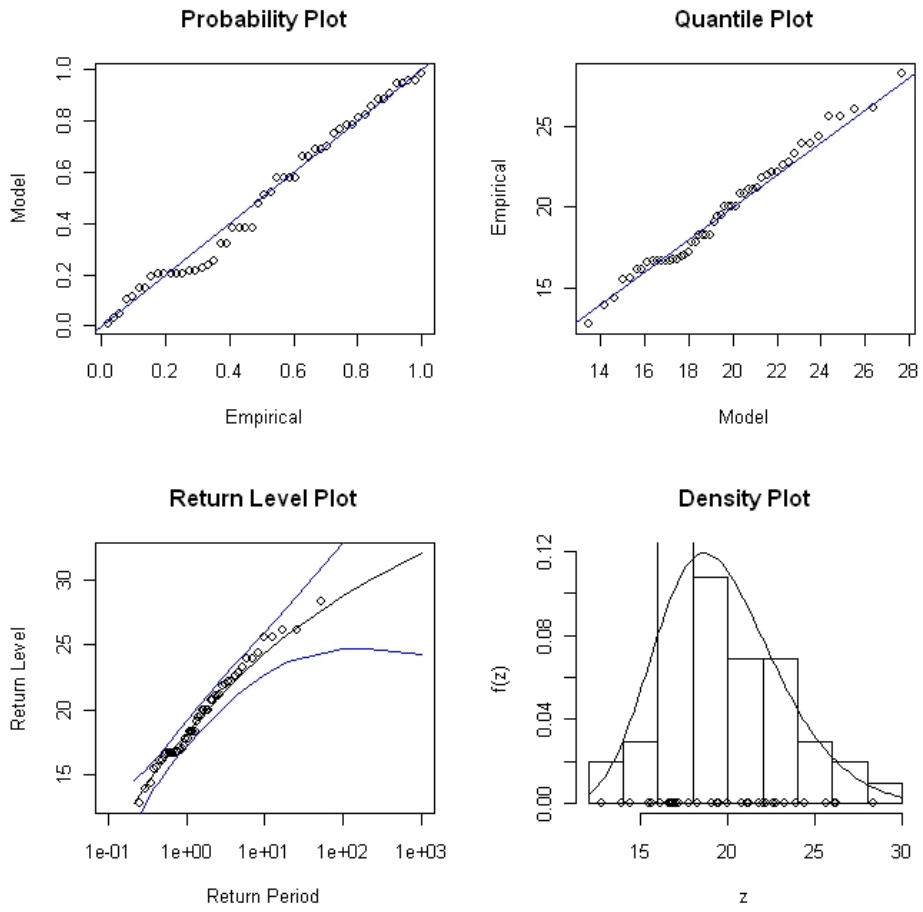

Figure 4: Diagnostic plots for GEV fit to the Sept-Iles example.

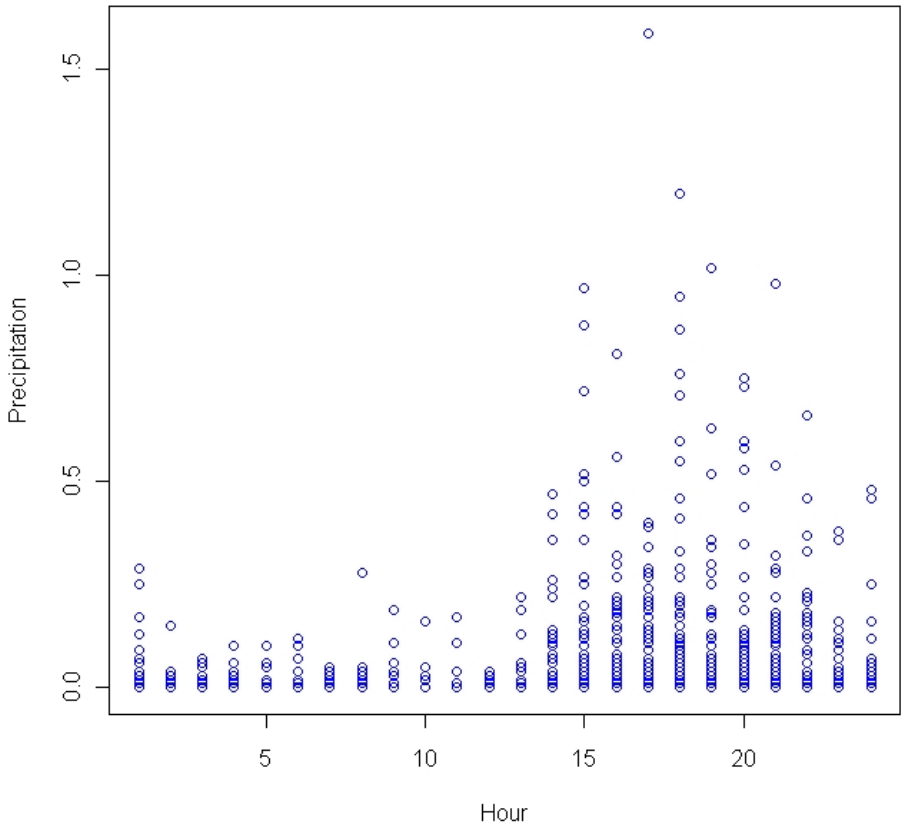

Figure 5: Scatterplot of precipitation agaist hour.

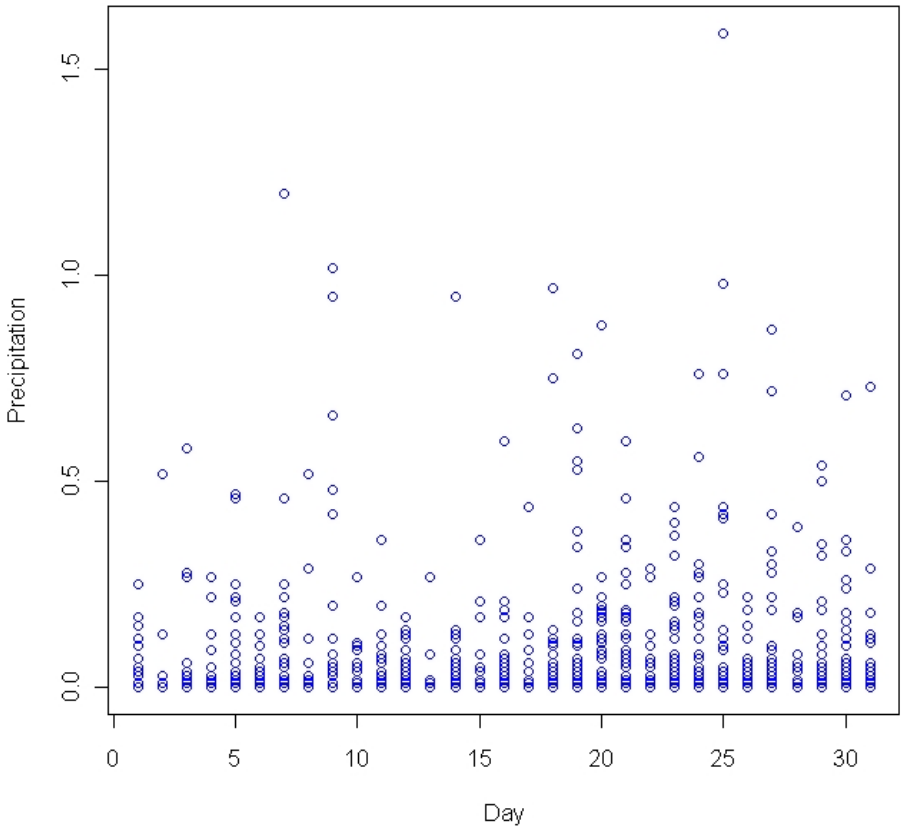

Figure 6: Scatterplot of precipitation agaist day.

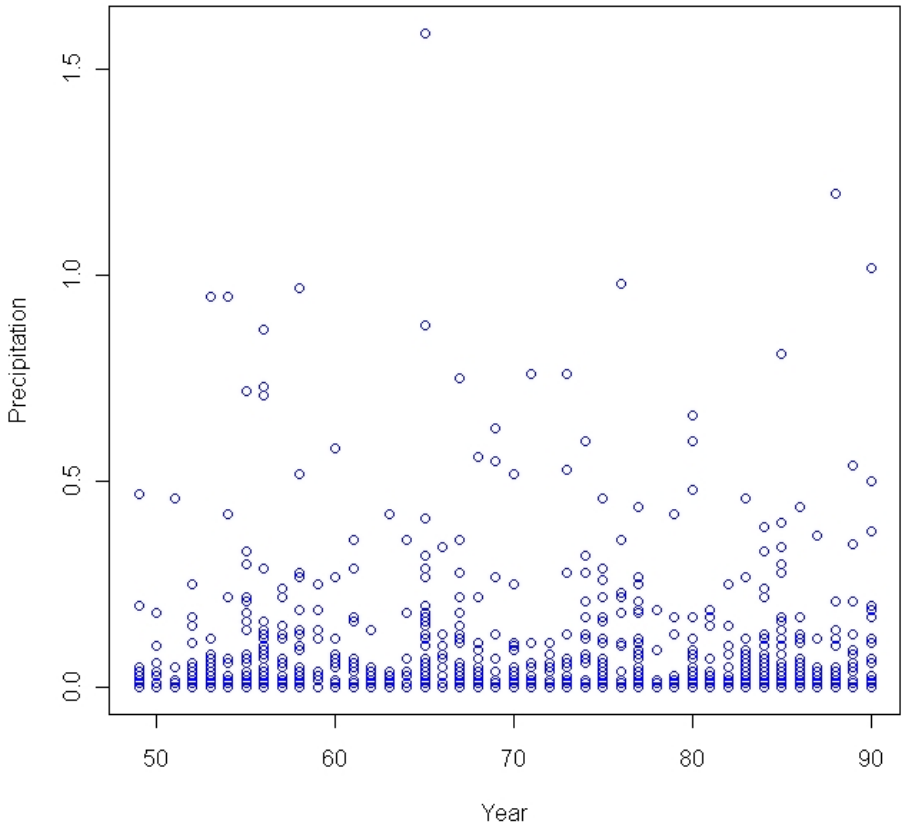

Figure 7: Scatterplot of precipitation agaist year.

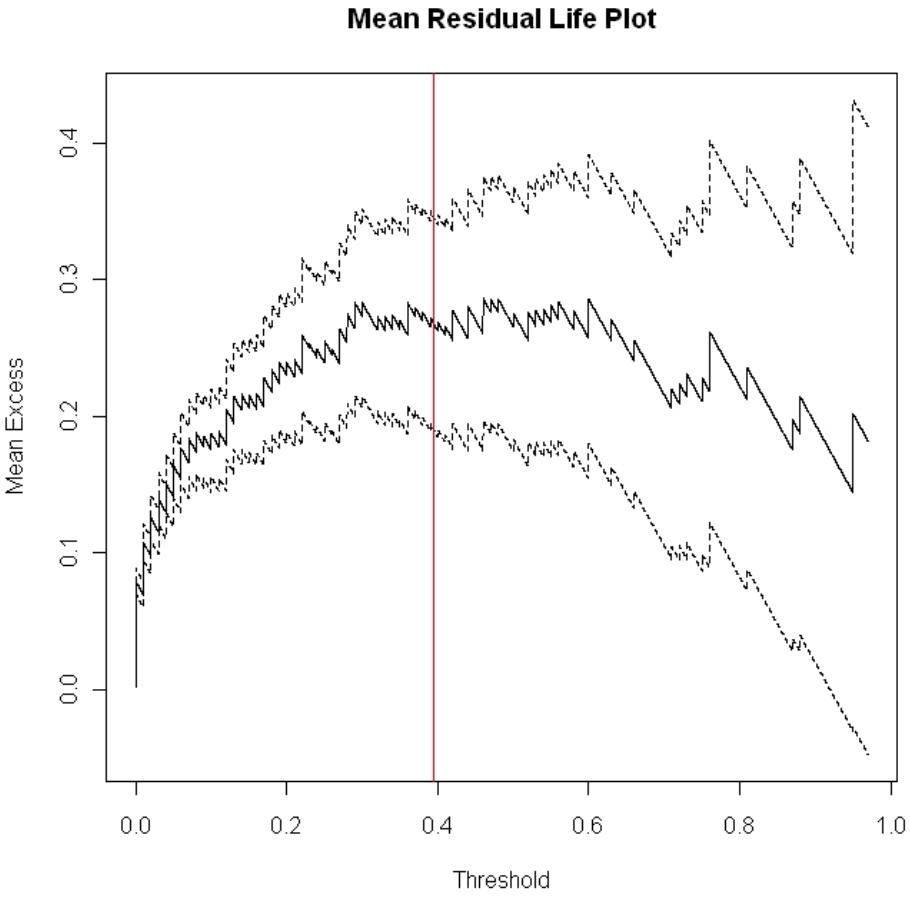

Figure 8: Mean excess plot for Denver example.

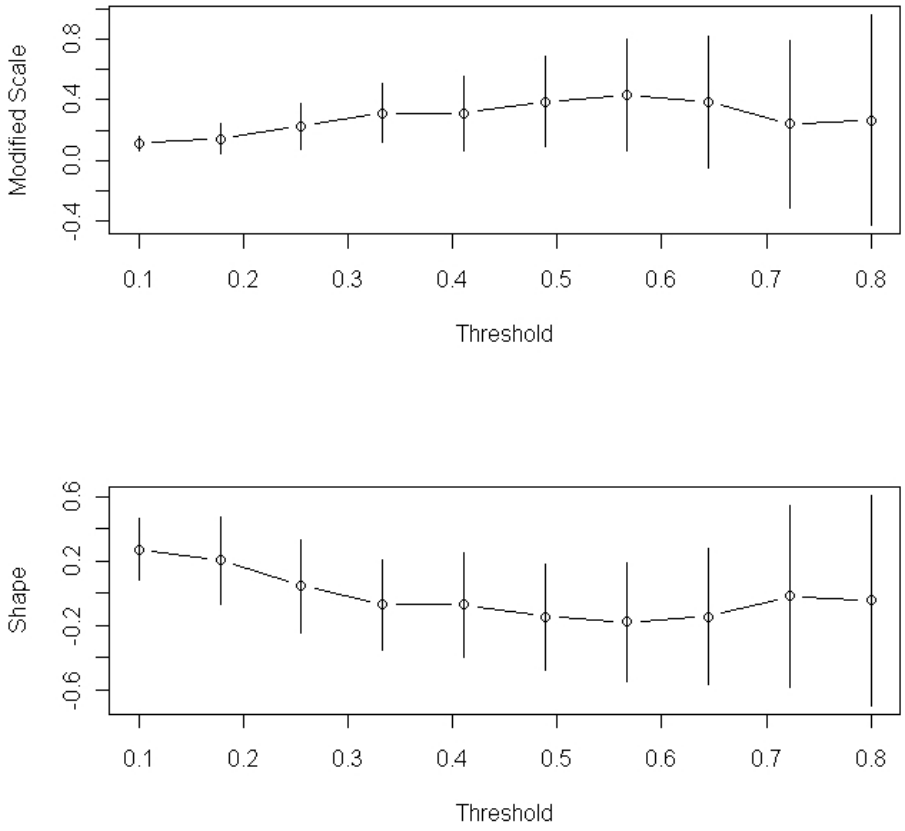

Figure 9: Parameter stability plot for Denver example.

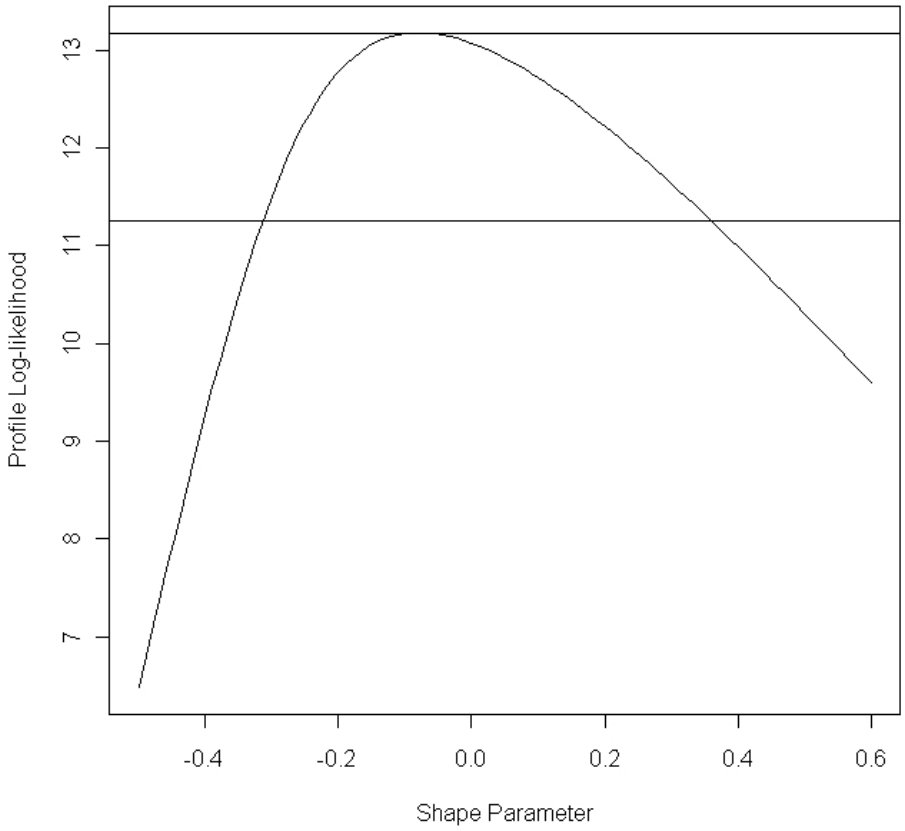

Figure 10: Profile likelihood for ξ in the Denver example.

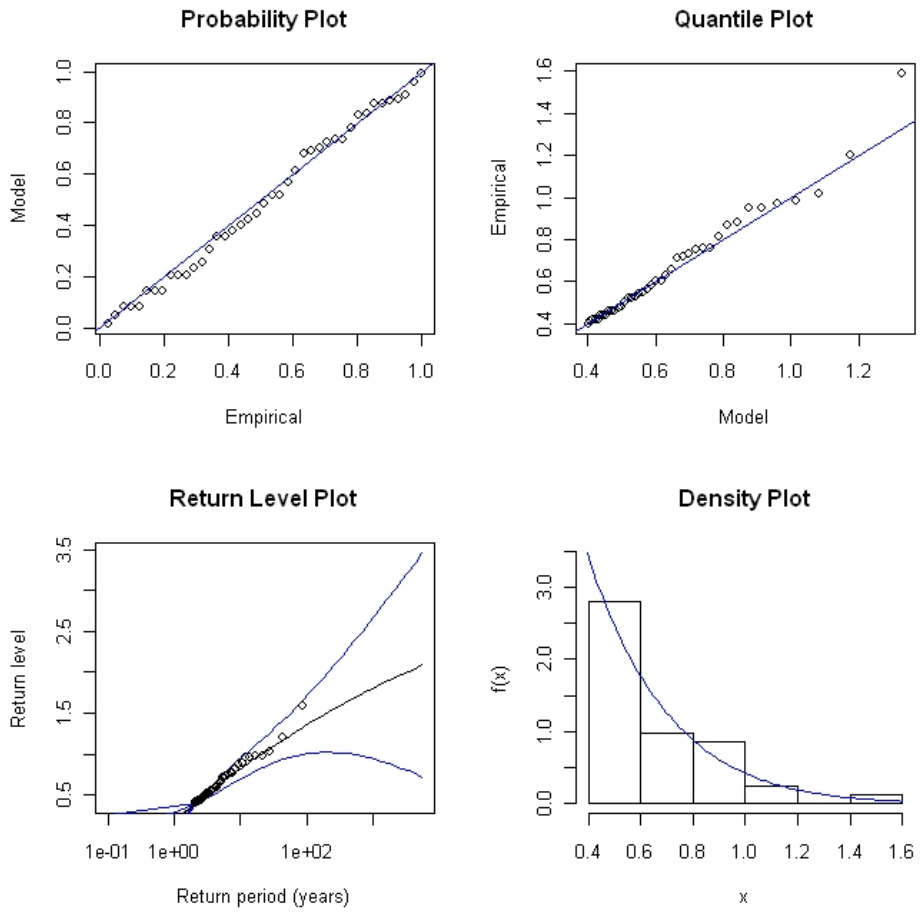

Figure 11: Diagnostic plots for GEV fit to the Denver example.

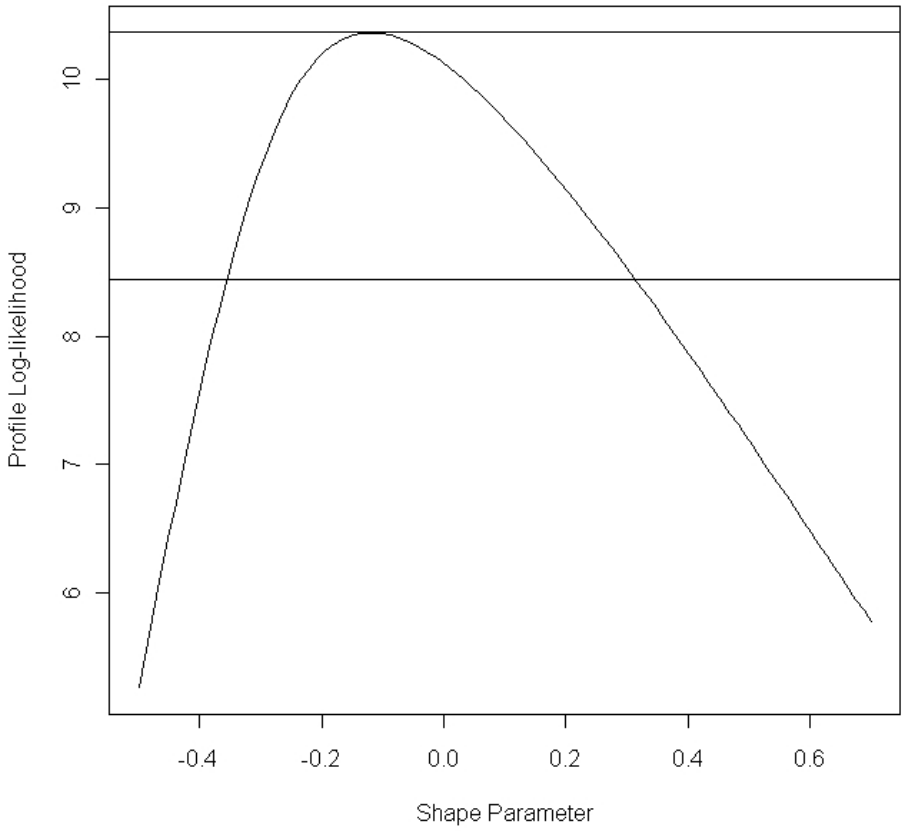

Figure 12: Profile likelihood for ξ for the aggregated Denver data.

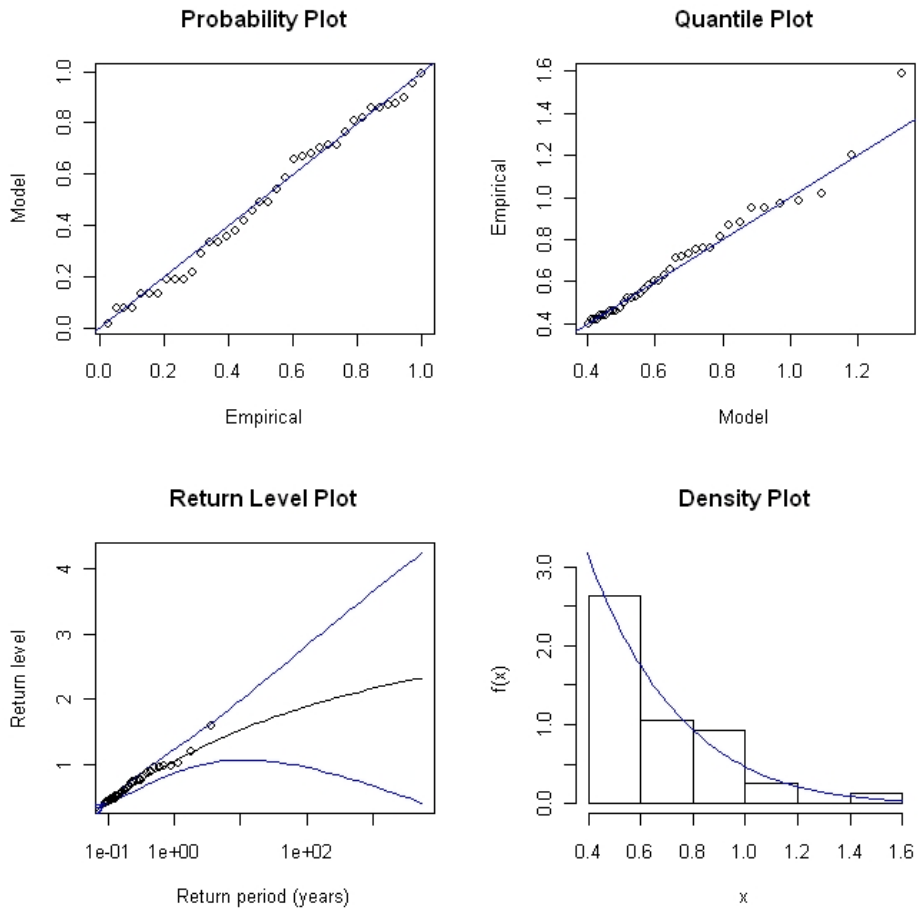

Figure 13: Diagnostic plots for GEV fit to the aggregated Denver example.

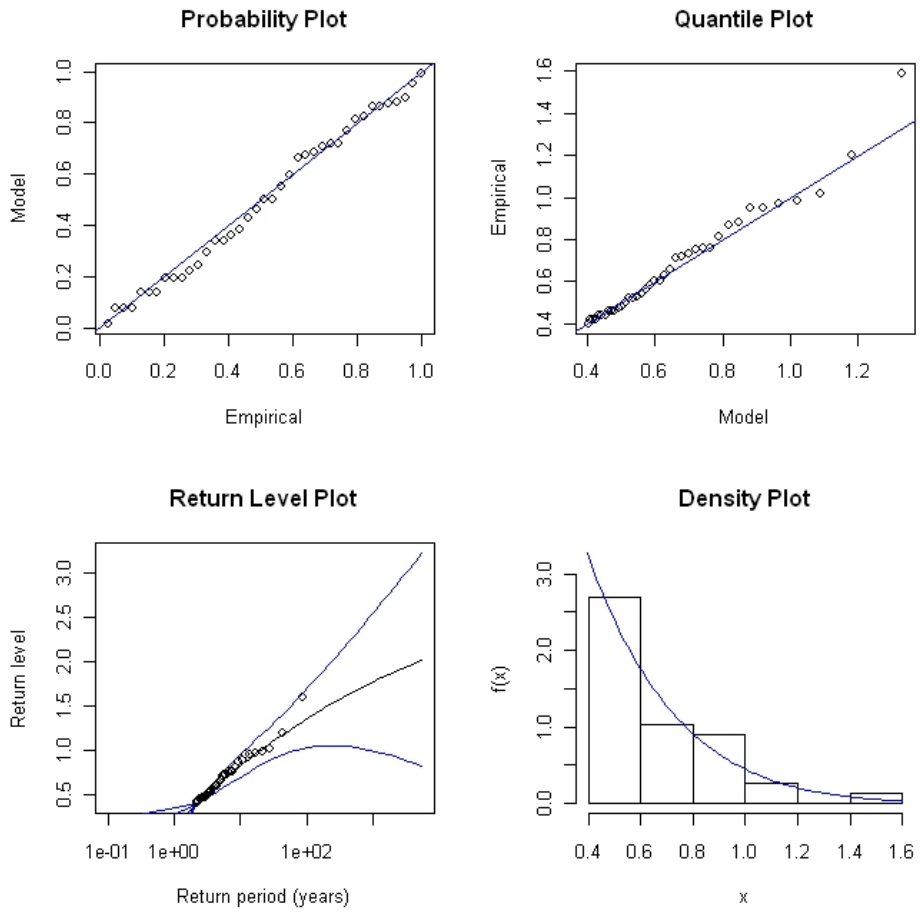

Figure 14: Diagnostic plots for GEV fit to the declustered Denver example.