Data Analysis in Extreme Value Theory
: Non-stationary Case

Soyoung Jeon
April 28, 2009

1. Background

We are already familiar with the background on extreme value theory. $Z_1, Z_2 \cdots$ are independent random variables with the same probability distribution and let $M_n = max\ (Z_1, Z_2, \cdots, Z_n)$. Then there exist normalizing constants $\mathrm{a_n} > 0$ and $\mathrm{b_n}$ such that

$$P\left(\frac{M_n - a_n}{b_n} > z\right) \rightarrow G(z) \text{ as } \mathrm{n} \rightarrow \infty$$

where G is a non-degenerate distribution function and has one of the three Extreme Value Distribution. The three types of EVD are written as a single family distribution and it is called as Generalized Extreme Value (GEV) distribution.

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

where $y_+ = max\{y, 0\}, -\infty < \mu,\ \xi < \infty$ and $\sigma > 0$.

Using this notation of GEV distribution, we apply it to the data of non-stationarity with trends in Section 2. And it can be applied to the non-stationary case with cyclic variation in Section 3.

2. Non-stationarity with Trends

The climate change in the environmental data has trends of extreme weather events through time. For example, maximum temperature or minimum temperature shows apparent trends over the time period. In this case non-stationary processes with changes through time are assumed. Thus using the notation of GEV(μ, σ, ξ) it follows that the model for Z_t in time $t = 1, 2, \cdots$ has GEV distribution,

$$Z_t \sim \mathrm{GEV}(\mu(t), \sigma(t), \xi(t))$$

$$\begin{aligned} \text{where } \mu(t) &= \mu_0 + \mu_1 \cdot t \\ \sigma(t) &= exp\ (\sigma_0 + \sigma_1 \cdot t) \\ \xi(t) &= \xi. \end{aligned}$$

Non-stationarity can be expressed in terms of the location and scale parameter with trends and shape parameter.

2.1 Data: Maximum Sea Levels at Fremantle

The annual maximum sea level data at Fremantle is discussed as one example for non-stationary case with trends. From 1897 to 1989 the annual maximum sea level is observed at Fremantle, Western Australia. There is an increase in annual maximum sea levels through time as shown in Figure 1 (a).

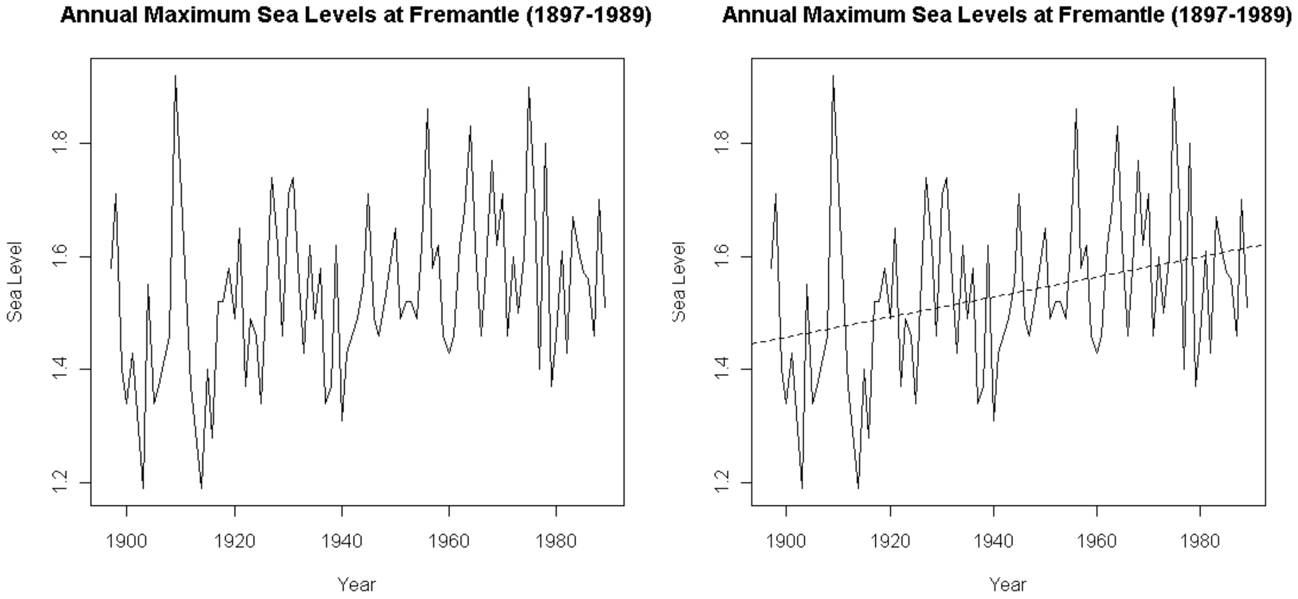

Figure 1 (a) Annual maximum sea levels at Fremantle, 1897-1989, (b) Linear regression fitting of year against maximum sea levels

To determine whether there exists linear trend, a linear regression of year against maximum sea levels is fitted. The lm function is used in R. The estimate of slope in year was 0.002 and the test for 'slope=0' was significant (p-value: 0.003). But the intercept estimate of linear regression is -1.899 and the test for 'intercept=0' didn't show a significant result (p-value: 0.091). In Figure 1(b), the dashed line indicates the linear line adjusted to the annual maximum sea levels. This result implies that it is reasonable to assume a linear trend for this data.

First we fit the GEV without any trend to compare with models including trends. Fitting the model without a trend in time we got the three parameters: $\hat{\mu} = 1.482,\ \hat{\sigma} = 0.141$ and $\hat{\xi} = -0.217$. The second model is GEV fit with linear trend in only location parameter and third one is with linear trend only in scale parameter. The last model is fitted with a trend in both location and scale parameter. The detail result of each fit is shown below:

Model (GEV)	$\hat{\mu}_0$	$\hat{\mu}_1$	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\xi}$	NLLH
without trend	1.482	0	0.141	0	-0.217	-43.567
lin. in location	1.387	0.002	0.125	0	-0.129	-49.790
lin in scale	1.493	0	-1.751	-0.006	-0.149	-44.824
lin in loc & scale	1.397	0.002	-1.929	-0.004	-0.138	-50.570

Table 1 Parameter Estimates and Negative Log Likelihood (NLLH) of GEV Fits

We get the result of likelihood ratio test from NLLH values in Table 1. For the test of $\mu_1 = 0$, we got the significant result (p-value: 0.0007) but the test for $\sigma_1 = 0$ was not significant (p-value: 0.2115). Thus the fitted model is

$$Z_t \sim \text{GEV}(\hat{\mu}(t), \hat{\sigma}(t), \hat{\xi}(t))$$

$$\text{where} \quad \begin{aligned} \hat{\mu}(t) &\approx 1.387 + 0.002 \cdot t \\ log\,\hat{\sigma}(t) &\approx 0.125 \\ \hat{\xi}(t) &\approx -0.129. \end{aligned}$$

The diagnostic plots, Figure 2(b) shows better performance than the model without a trend (See Figure 2(a)). Thus we have a significant result suggesting that the model with linear trend in location explains more about the Fremantle data than the stationary model.

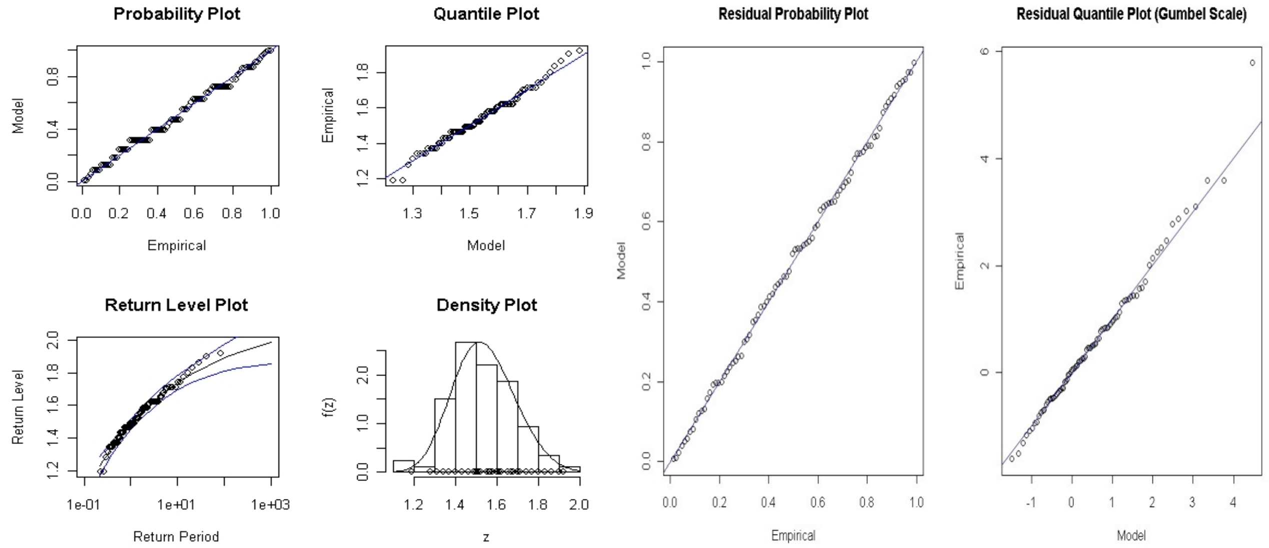

Figure 2 (a) Diagnostic plots for model without a trend, (b) Diagnostic plots for model with linear trend in location parameter

3. Non-stationarity with Cyclic Variation

In meteorological data lots of variables have annual, seasonal or diurnal cycles. In the modeling cyclic changes in threshold exceedance, it is useful to specify a model with different parameters in each cycle.

There are two approaches to analyze this sort of data. One thing is the orthogonal approach. First we fit an annual cycle to Poisson rate parameter. In detail, we make a binary vector where 1 indicates an excess over threshold and 0 indicates an observation under threshold. Then we make two vectors containing the cyclic trend over time. Thus the model to obtain Poisson rate parameter is

$$log\, \lambda(t) = \lambda_0 + \lambda_1\, sin\left(\frac{2\pi t}{T}\right) + \lambda_2\, cos\left(\frac{2\pi t}{T}\right).$$

Then the Generalized Pareto distribution (GPD) is fitted with cycle in scale parameter. The cyclic model can be expressed by

$$log\, \sigma^*(t) = \sigma_0^* + \sigma_1^*\, sin\left(\frac{2\pi t}{T}\right) + \sigma_2^*\, cos\left(\frac{2\pi t}{T}\right).$$

Another approach is using the point process model. We fit the point process model with the cyclic trend in location and scale and those location and scale parameter are obtained from GEV re-parameterization. Three parameters can be expressed as

$$\mu(t) = \mu_0 + \mu_1\, sin\left(\frac{2\pi t}{T}\right) + \mu_2\, cos\left(\frac{2\pi t}{T}\right)$$
$$log\, \sigma(t) = \sigma_0 \ + \sigma_1 \ sin\left(\frac{2\pi t}{T}\right) + \sigma_2 \ cos\left(\frac{2\pi t}{T}\right)$$
$$\xi(t) = \xi.$$

3.1 Data: Denver Hourly Precipitation

Hourly precipitation (mm) is observed for Denver, Colorado in the month of July from 1949 to 1990. In Figure 3 each plot shows hourly precipitation according to year (1949-1990) and hour (1-24). The red line from those plots indicates a threshold, 0.395 which was reasonable as the result of previous study.

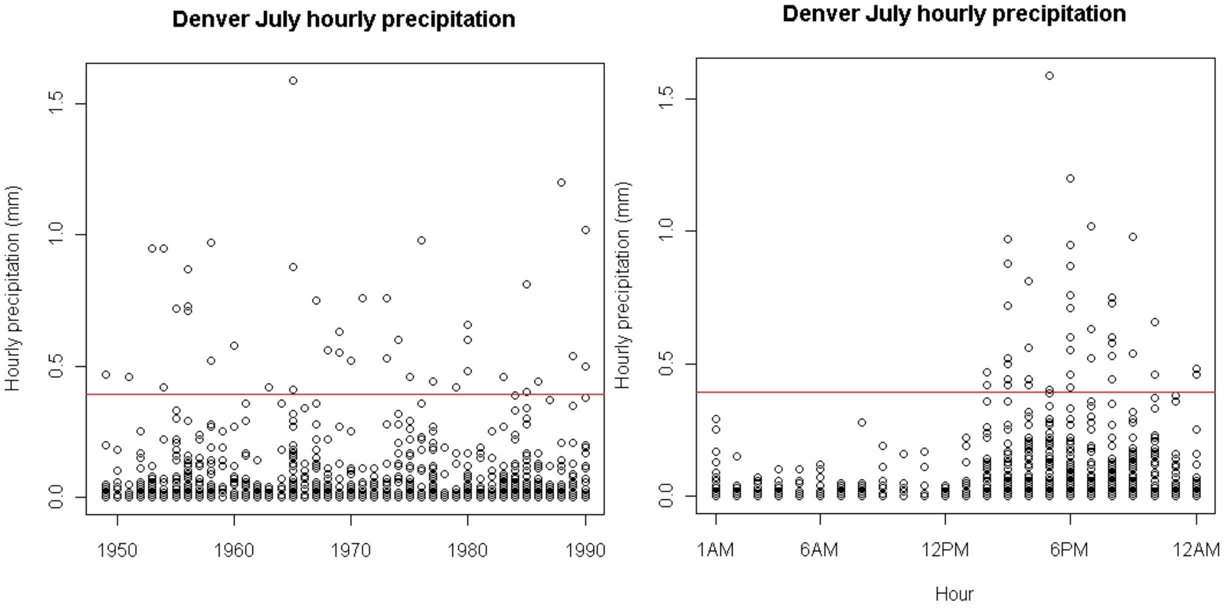

Figure 3 Denver Hourly Precipitation (a) through year, (b) over hour

To find the cycle explaining hourly precipitation we fit the generalized linear model for each annual cycle and hourly cycle. Fitting annual cycle in Poisson rate parameter we got the parameter,

$$log\,\hat{\lambda}(t) = -6.682 - 0.433\,sin\left(\frac{2\pi t}{T}\right) + 0.004\,cos\left(\frac{2\pi t}{T}\right)\;,\quad T = 744\text{ (31 days × 24 hours)}$$

but the test for coefficients showed that those parameter estimates are not significant. Thus now we fit the hourly cycle instead. Using the glm with Poisson family we get

$$log\,\hat{\lambda}(t) = -8.131 - 2.888\,sin\left(\frac{2\pi t}{T}\right) - 0.024\,cos\left(\frac{2\pi t}{T}\right)\;,\quad T = 24\text{ (hours)}$$

and the significant result was reported for the test of $\lambda_1 = 0$ (p-value< 0.001). It makes sense that only the vector with *sin* cycle has effects on hourly precipitation since we already showed high precipitation during 12PM-12AM than 1AM-12PM from Figure 3(b). From the GPD fitting with hourly cycle in scale parameter we got the following estimates,

$$log\,\widehat{\sigma}^{*}(t) = -2.660 - 1.877\,sin\left(\frac{2\pi t}{T}\right) + 0.022\,cos\left(\frac{2\pi t}{T}\right).$$

For the test of parameter estimates, only σ_1 was significant (p-value=0.028).

Using the GEV re-parameterization approach, we got GEV model with hourly cycle in following parameters,

$$\hat{\mu}(t) \approx -1.581 - 0.876 \cdot sin\left(\frac{2\pi t}{T}\right)$$

$$log\,\hat{\sigma}(t) \approx -0.381 + 0.031 \cdot sin\left(\frac{2\pi t}{T}\right)$$
$$\hat{\xi}(t) \approx -0.243.$$

In the likelihood ratio test for $\mu_1 = 0$, we got the significant result (p-value< 10^{-6}) but the test for $\sigma_1 = 0$ was not significant (p-value > 0.10). Thus we finally fit the model with hourly cycle in location parameter. Diagnostic plots for model with cycle are below:

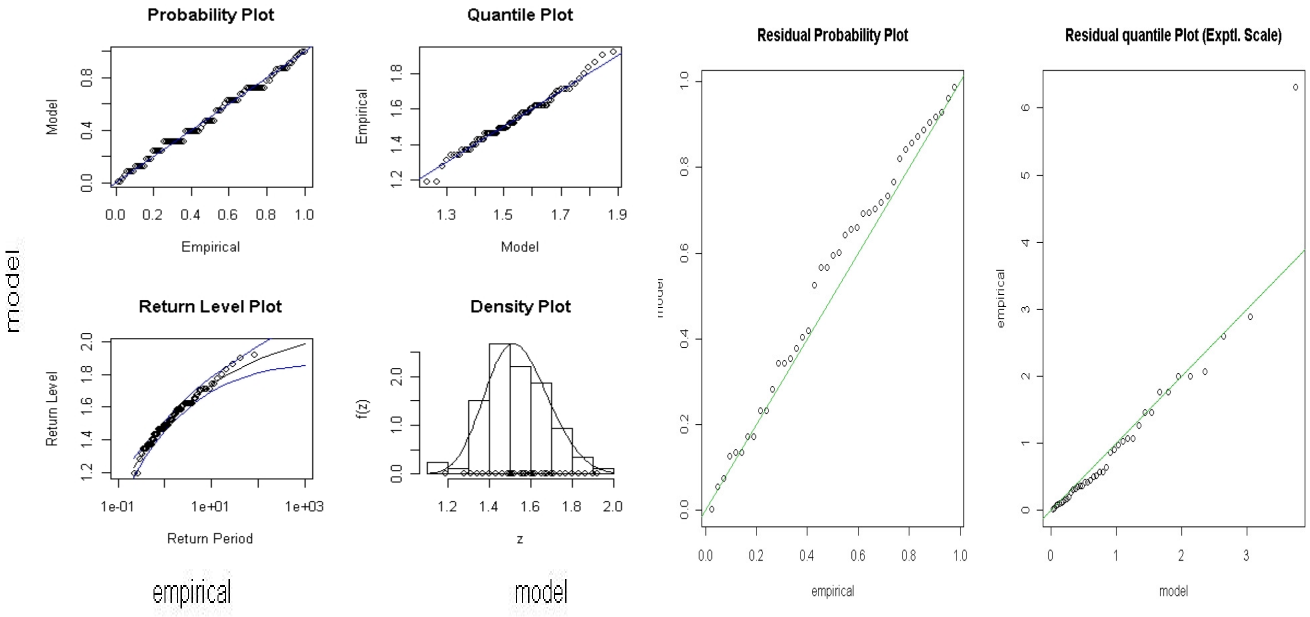

Figure 4 (a) Diagnostic plots for point process model without cycles, (b) Diagnostic plots for point process model with hourly cycle in location parameter

Diagnostic plots do not show better fitting than diagnostic plot for the model without cycles. It might be obtained better results from other possible cycle, for example day time cycle vs. night time cycle.

4. Conclusion

As it is shown with two data analysis, those methods are very specialized for non-stationary extreme data. Also the general theory can not be extended for non-stationary series but the advantage is that we can interpret the trend with covariates. Thus it is useful to adopt to the pragmatic approach.

References

Coles, S. G. (2001) An Introduction to Statistical Modeling of Extreme Values. London: Springer.

Stephenson A and Eric Gilleland. (2006) Software for the analysis of extreme events: The current state and future directions. Extremes, 8:87.109.

Eric Gilleland, Rick Katz and Greg Young. (2009) extRemes: Extreme value toolkit, R package

Stuart Coles and Alec Stephenson. (2009) ismev: An Introduction to Statistical Modeling of Extreme Values, R package

Interdisciplinary Workshop. (2009) Effects of climate change: coastal systems, policy implications and the role of statistics.