時系列データの最大値・最小値のモデル化について

－フリーソフトRのextRemesを利用して－

On the modeling for the extreme values of a time series
－ by using the extRemes of R statistical software program －

ネットワーク情報学部　田中　稔

School of Network and Information　Minoru TANAKA

Keywords：extreme values, generalized extreme value distribution, R, extRemes

Abstract

This paper introduces the classical extreme value theorem and three examples are provided by way of illustration. The computations for the model estimations are carried out using the extRemes of R statistical software program.

1．はじめに

近年、特に注目されていることに、大洪水などの異常気象、異常発生による自然災害がある。一昨年のアメリカミネソタ州の橋の崩壊、地殻変動に伴う新潟沖大地震による原子力発電所の安全性なども記憶に新しい。これらはたいへん稀な出来事であるが、数理統計学の分野で考えると極値（最大値、最小値）理論に関係すると思われる。橋は建築後３０～４０年の疲労亀裂の大きさや金属疲労度（寿命）、原発は操業後２０年以来の地震の最大震度、大洪水は過去３０年内で最大降雨量に関連する問題である。建築物の耐震性の安全基準を設定するには、その建築地点における地震の最大震度を予測する必要があることは明らかである。

同様な問題として、古くから分析されているものに防波堤の設計がある。ある地点の波の高さは時間とともに変化する時系列と考えられるが、ここで問題とされることは一刻一刻の波の高さの分布ではなく、その波の高さの最大値（極値）であり、その分布である。ある期間に於ける最大値（例えば、２０年、１００年あたりの）の確率分布が分かれば、そのReturn level（帰還レベル）からある大きさを超えるReturn period（帰還期間）の予測に役立つ。

では最大値（Max）の分布はどうなっているのであろうか。時系列データの元々の分布と関係があることは想像できるが、例えばデータが正規分布に従っているとき最大値はどのような分布をしているのだろうか。以下の第２章では数理統計学から得られる最大値等の厳密な分布について述べる。

ところで我々がここで問題としている最大値（または最小値）は、一ヶ月、１年などの短いスパンではなく、２０年、５０年、１００年といった長いスパンでの極値である。そこで考えるべきものは極値の漸近的な振る舞い、すなわち極値の漸近的な確率分布である。第３章では古典的な極値理論でよく知られている事実である極値の漸近的な分布が３種類に限られること、さらにそれらはまた１つの一般化された極値分布で表現されることなどを紹介する。

そして第４章ではこれらの理論の応用として、実際のデータを使ってモデル（確率分布）を推定することを考える。そのために必要な、主に研究用に開発されている統計解析のフリーソフトRから極値モデルの解析用パッケージextRemesを紹介する。我々はこのソフトを使って、(A) シミュレーションデータ、(B) 公表されているFlood（米国の洪水による年間最大損害額）のデータ、さらに（C）気象庁の公表データから東京の年間最大降水量（1日辺り）の確率分布をそれぞれ推定する。特に東京の最大降水量データは101年間というサンプル数にも恵まれ、当てはめたモデルがデータに大変良く適合している。このモデルより一日に250mmから300mmを超える降水量は100年に一度くらいの比率で起こる現象であることなどが分かる。

2　極値の厳密分布について

いま確率変数Xは連続な確率密度関数f(x)を持ちその確率分布関数をF(x)とする。このときXからの大きさnの独立な標本確率ベクトルを$(X_1, X_2..., X_n)$、これを大きさの順に並び替えたものを$(X_{(1)}, X_{(2)}..., X_{(n)})$とする。すなわち$X_{(1)}< X_{(2)}<\cdots< X_{(n)}$；例えば$X_{(1)}=Min(X_1, \ldots, X_n)$は最小値であり、$X_{(n)}$は最大値である。これらは確率変数であり、その確率分布はもとの分布関数F(x)とその密度

関数 f(x) を使って表現できる。一般に r 次の順序統計量 X(*r*)，r ∈ {1, …, n}，の確率密度関数は

$$\frac{n!}{(r-1)!(n-r)!}F(x)^{r-1}(1-F(x))^{n-r}f(x)$$

となる。したがって最大値 の場合は

$$g(x)=nF(x)^{n-1}f(x)$$

となる ([1],[2])。

確率変数Xがよく知られている場合に、具体的な最大値 $X_{(n)}$の確率密度関数 $g(x)$およびそのグラフを求めてみる。

（1）X が一様分布 U[0,1] に従っている場合；

$$g(x)=nx^{n-1}$$

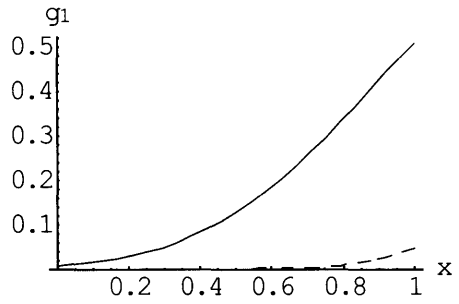


図 1.　n = 10, 30 の g(x) のグラフ

（2）標準正規分布 N(0, 1) に従っている場合；

$$g(x)=\frac{1}{\sqrt{p}}2^{1/2-n}e^{-x^2/2}n(1+Erf[x/\sqrt{2}])^{n-1}$$

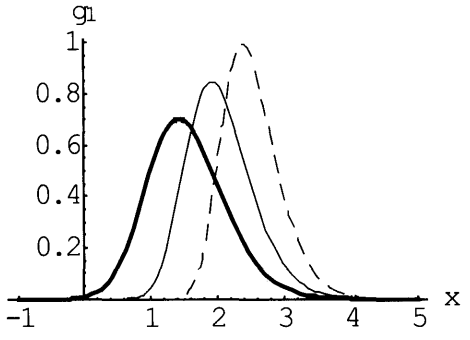


図 2.　n = 10, 30, 100 の g(x) のグラフ

（3）X がパレート分布に従っている場合；

$$g(x)=\frac{n}{2}x^{-2}(1-\frac{1}{2x})^{n-1}$$

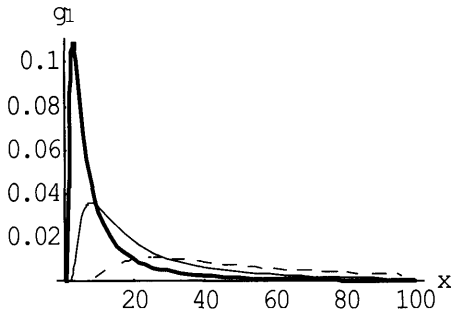


図 3.　n = 10, 30, 100 の g(x) のグラフ

3　極値の漸近分布について

第 2 章では、標本確率変数列の極値（最大値や最小値）の厳密な分布は元の確率分布に依存し多く存在することになる。しかし、標本数が大きくなると分布型がそれぞれ収束して、最終的な極値の漸近的確率分布関数は Type I. Gumbel 分布、TypeII. Frechet 分布、TypeIII. Weibull 分布の 3 つのタイプの族 (family) であることが古典的極値理論より知られている（参照 [3]）。それらの確率密度関数 f(x) およびグラフは以下のようになる。元の確率分布が正規分布や指数分布などの場合、パレート分布やコーシー分布の場合、一様分布などの場合はそれぞれその極値の漸近的確率分布は TYPE I,　TYPE II, TYPE III となる。

Type I.　Gumbel 分布

Variable: -∞ <x< ∞, Parameters: b(location),a > 0 (scale)

$$F(x)=Exp[-Exp[-\frac{x-b}{a}]]$$

確率分布の密度関数は右裾において指数関数的に急激に減少する。

$$f1=\frac{e^{-e^{\frac{b-x}{a}}+\frac{b-x}{a}}}{a};\quad \text{domain}[f1]=\{x,-\infty,\infty\}\ \&\&\ \{a>0,-\infty<b<\infty\};$$

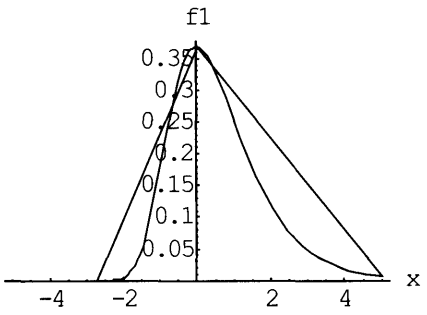


図 4.　Gumbel 分布

II. Frechet 分布 Variable: x>b、 Parameters: b (location), a>0 (scale), $\alpha > 0$ (shape)

$$F(x) = Exp[-(\frac{x-b}{a})^{-a}]$$

確率分布の密度関数は右裾において多項式関数的に比較的緩やかに減少する

$$f2 = \frac{e^{-\left(\frac{-b+x}{a}\right)^{-\alpha}}\left(\frac{-b+x}{a}\right)^{-1-\alpha}\alpha}{a};$$

domain[f2] = {x, b, ∞} && {α > 0, a > 0, -∞ < b < ∞};

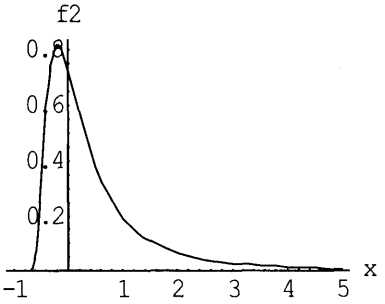


図 5. Frechet 分布

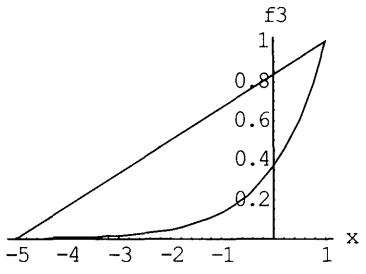
III. Weibull 分布 Variable: x < b, Parameters: b (location), a > 0 (scale), $\alpha > 0$ (shape)

$$F(x) = Exp[(\frac{x-b}{a})^{a}]$$

確率分布の密度関数は左裾において、TypeII の Frechet 分布と同様に多項式関数的に比較的緩やかに減少する

$$f3 = \frac{e^{\left(\frac{-b+x}{a}\right)^{\alpha}}\left(\frac{-b+x}{a}\right)^{-1+\alpha}\alpha}{a};$$

domain[f3] = {x, -∞, b} && {α > 0, a > 0, -∞ < b < ∞};



極値理論の初期の応用例ではこれら３つの分布族を候補として採用し、選ばれた分布族のパラメータを推定した。しかし３つのうちどれを選ぶかはかなりの熟練が必要であり大変難しい。一方、これら３つの極値の漸近的確率分布関数はさらに一つの分布族に一般化されることが知られている。一般化極値分布（the Generalized Extreme Value distribution, GEV）と呼ばれている（参照 [1], [2], [3]）。したがって実際に得られたデータの確率分布を推定する場合には、上の３つの分布族を候補とするよりも、一般化極値分布を仮定しそのパラメータを推定するほうが遥かに容易である。このＧＥＶの確率分布の密度関数は以下となる。shape パラメータξの符号によって上の３つの分布を得る。即ち、ξ>0 の場合は Frechet 分布、ξ<0 の場合 Weibull 分布、ξ→0 の場合が Gumbel 分布 となる。

Generalized Extreme Value distribution;
Variable: x > b, Parameters: μ (location), $\sigma > 0$ (scale), ξ > 0 (shape,)

$$F(x) = Exp[-(1+x(\frac{x-b}{a}))^{-1/x}]$$

(1) shape parameter ξ> 0 の場合；

$$f01 = \frac{e^{-\left(1+\frac{(x-\mu)\xi}{\sigma}\right)^{-1/\xi}}\left(1+\frac{(x-\mu)\xi}{\sigma}\right)^{-1-\frac{1}{\xi}}}{\sigma};$$

domain[f01] = {x, μ - σ/ξ, ∞} && {-∞ < ξ < ∞, σ > 0, -∞ < μ < ∞};

(2) shape parameter ξ < 0 の場合；

$$f02 = \frac{e^{-\left(1+\frac{(x-\mu)\xi}{\sigma}\right)^{-1/\xi}}\left(1+\frac{(x-\mu)\xi}{\sigma}\right)^{-1-\frac{1}{\xi}}}{\sigma};$$

domain[f02] = {x, -∞, μ - σ/ξ} && {-∞ < ξ < ∞, σ > 0, -∞ < μ < ∞};

(3) shape parameter ξ → 0 の場合；

$$f1 = \frac{e^{-e^{\frac{b-x}{a}}+\frac{b-x}{a}}}{a};$$ domain[f1] = {x, -∞, ∞} && {a > 0, -∞ < b < ∞};

4 Ｒのパッケージソフト extremes による極値のモデル化の応用例

実際にデータ解析する場合に比較的扱いやすいコンピュータソフトとして、フリーウエアーソフトＲのパッケージ extremes が知られている。極値データの分析は一般的なテーマではなく特殊なためそれに関する分析ソフトは市販のソフトの中ではあまり見かけない。しかし最近注目され始めているフリーソフトのＲには多くの研究者が各自の専門分野の統計解析用パッケージを公開している。あくまでもフリーであるため信頼性には絶対的な保障はないが、多くのユーザーによる検証により日々ソフトの更新が行われている。我々は最新のものをインストールすることでソフトウエアーの信頼性を確保できるものと思われる。このＲの中の極値モデルの解析用パッケージ extRemes は G.Young や E.Gilleland 等によって開発され、また現在も更新されている。[4],[5].

以下ではこのソフトを使って実際にデータ解析をしてみることにする。ここで解析に使うデータは次の３つのデータである。

（A）擬似乱数で一般化極値分布からシミュレーションしたデータ

（B）パッケージ extremes に入っている米国の Flood データ（洪水による損害額の年間最大値）

（C）東京周辺の一日辺りの降水量の年間最大値

パッケージ extremes によるモデル化の方法はデータの入力と散布図、一般化極値分布 GEV の最尤法によるパラメータ推定、４つの検証用グラフ作図（Probability Plot, Quantile Plot, return level curve, Density Plot with Hisotgram）[1] 参照。最後にこれらの結果からモデルの当てはまりの良さを考察する。

４．１　データ解析の実例

（A）図は GEV[μ , σ , ξ] = GEV[0, 1, 0.2] (location parameter μ =0, scale parameter σ =1.0, shape parameter ξ=0.2) からシミュレーションで発生させた大きさ５０の標本の散布図である。このヒストグラムは下図にある。

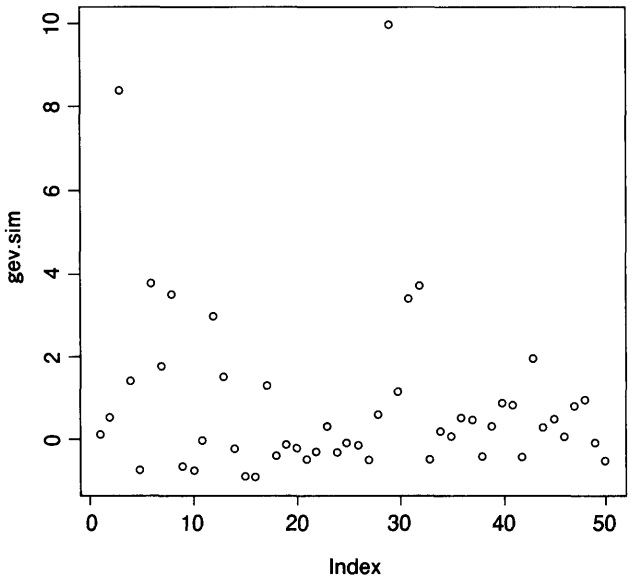


図７．GEV[0, 1, 0.2] からの大きさ５０の標本の散布図

このデータに一般化極値分布 GEV[μ , σ , ξ] を当てはめ、最尤法によるパラメータ推定を行った。下の出力結果より、パラメータ推定値は

μ = -0.075

σ = 0.741

ξ = 0.464

すなわち、推定されたモデルは GEV [μ , σ ,ξ] = GEV [-0.075, 0.741, 0.464] となる。またパラメータの共分散行列より、それぞれのパラメータの 95％信頼区間が漸近的に、μ は [-0.31, 0.16], σ は [0.52, 0.96] そして ξ は [0.188, 0.739] となる。μ の値はゼロの可能性が高いが、他のものは有意と考えて良いであろう。

```
[1] "Maximum Likelihood Estimates:"

MLE Stand.Err.MU:(identity) -0.07467      0.12024
SIGMA:(identity)             0.74121      0.11259
Xi:(identity)                0.46437      0.14056

[1] "Negative log-likelihood: 77.2372558627211"

Parameter covariance:[,1][,2][,3]
[1,] 0.014457046  0.009767753  -0.004453080
[2,] 0.009767753  0.012675911   0.001537216
[3,] -0.00445308  0.001537216   0.019756846
```

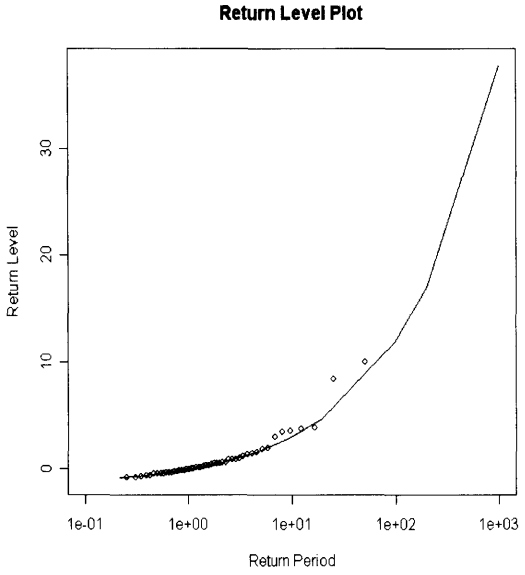


図８．Return Level Plot

検証用グラフ（図９）の結果より

Probability Plot：直線上に点が乗っていてモデルの当てはまりは良い。

Quantile Plot：X座標の値が大きいとき以外は直線に沿っています。

Return Level Plot：９５％信頼区間の中に納まっている。

Density Plot(with Histogram)：分布の特徴が捉えられている。

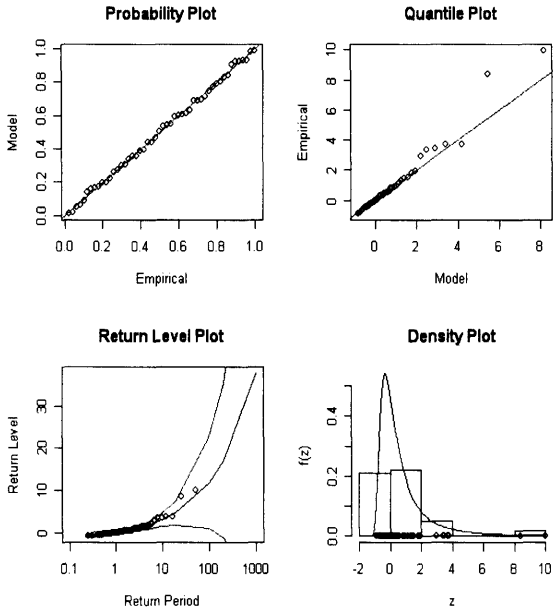


図９．検証用の４つのグラフ

（B）パッケージ extremes に入っている既存の Flood データ（年間の洪水による損害額の最大値、大きさ n = ５０）。下図は Flood データの散布図である。この図より（A）で分析したシミュレーションのデータに分布の様子が酷似していることが分かる。従って、このデータに当てはめる確率モデルの候補は一般化極値分布ＧＥＶ(μ , σ , ξ) と考えても良いだろう。

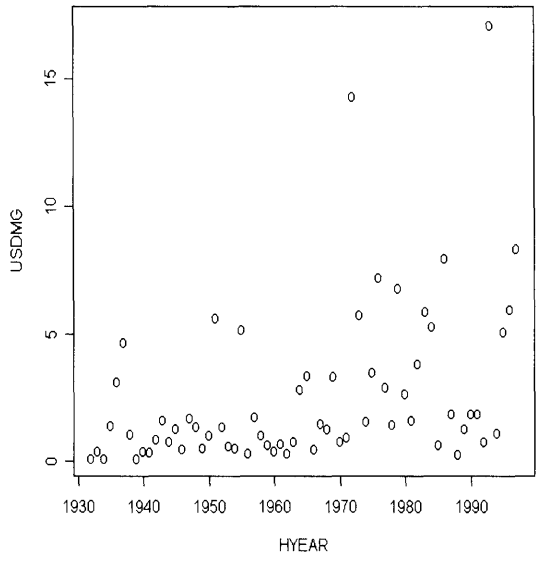


図１０．Flood データの散布図

一般化極値分布ＧＥＶ(μ , σ , ξ) をデータに当てはめ、最尤法によるパラメータ推定を行った。プログラムの出力結果（下記に推定値と各とその分散協分散行列がある）より、パラメータの推定結果は

μ = 0.99
σ = 0.96
ξ = 0.70

すなわち、推定されたモデルは GEV[μ , σ , ξ] = GEV[0.99, 0.96, 0.70] となる。従って、shape parameter ξ = 0.70 (> 0、標準誤差 0.15226) よりこのデータの確率分布は Frechet 分布 に近いと考えられる。

"Maximum Likelihood Estimates:"

MLE	Stand.Err.	
MU:(identity)	0.99150	0.13924
SIGMA:(identity)	0.96229	0.15199
Xi:(identity)	0.70376	0.15226

[1] "Negative log-likelihood: 127.429708565582"

Parameter covariance:[,1][,2][,3]

[1,]	0.019389158	0.017124769	-0.005656303
[2,]	0.017124769	0.023099825	0.002793059
[3,]	-0.0056563	0.002793059	0.023182714

モデル検証用グラフの解析：

Probability Plot：各点が直線上に近く分布していて、よく当てはまっている。

Quantile Plot：シミュレーションの結果と同じ傾向である。

Return Level Plot：９５％信頼区間の中に納まっている。

Density Plot(with Histogram)：モデルの密度関数がデータのヒストグラムに近い。

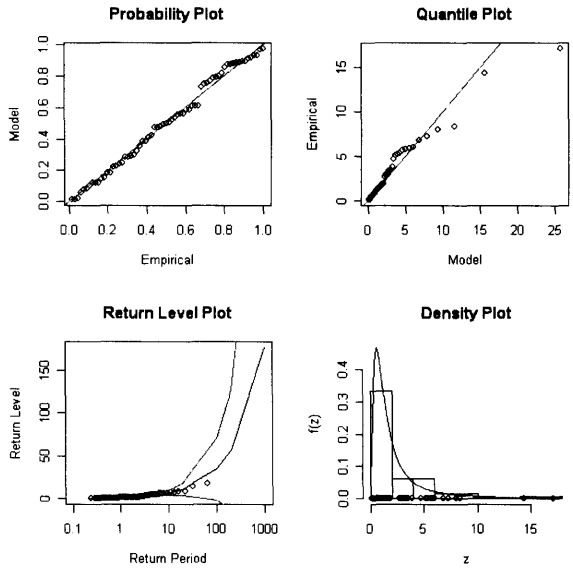


図１１．検証用の４つのグラフ

当てはまりの良さについての考察：モデルが適切でないという証拠は特にないといえる。特に Density Plot ではデータと推定されたモデルの確率密度度関数は良く当てはまっているといえる。ただし、Quantile Plot においては、各点が直線とかけ離れており良い当てはまりとはいえない。別のモデルを考慮する余地は十分にあるといえる。
もし一般化極値分布ＧＥＶ以外の別のモデルを当てはめた場合には、尤度比検定を行いどちらがより有効であるかを調べることが出来る。

（Ｃ） 最後に東京周辺の一日辺りの降水量の年間最大値（block maxima、単位は mm）を 1901 年から 2007 年までのデータ（気象庁のホームページを参照）を解析してみる。1958 年には東京でこの 101 年間最大の 371.9 ｍｍを記録している。散布図から上で解析した（Ａ）や（Ｂ）の分布に似ていることが分かるであろう。

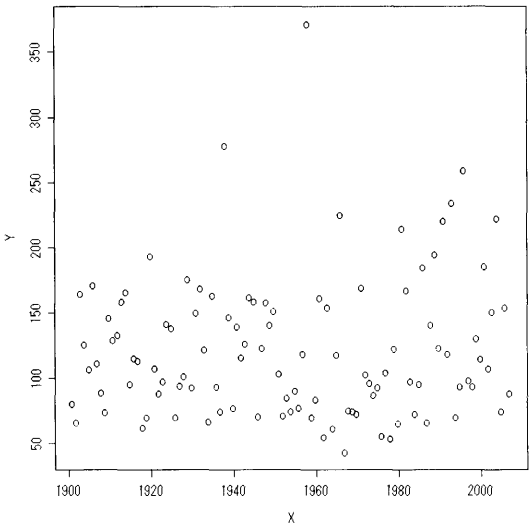


図１２．東京周辺の一日辺りの降水量の年間最大値の散布図

そこで一般化極値分布ＧＥＶ(μ , σ ,ξ) をこのデータに当てはめ、最尤法によるパラメータ推定を行った。プログラムの出力結果（括弧内は標準偏差）より、パラメータの推定結果は

μ = 96.43427 (3.99951)
σ = 35.70065 (3.12339)
ξ = 0.13926 (0.08816)

すなわち、推定されたモデルは GEV[μ , σ ,ξ] = GEV [96.43, 35.70, -0.139] となる。従って、shape parameter はξ = 0.139 であるが、標準誤差が 0.08816 であり、９５％信頼区間にはゼロが含まれるためこのデータの確率分布は ξ = 0.0 である Type I の Gumbel 分布 に近いと考えても良いであろう。

```
[1] "Maximum Likelihood Estimates:"
 MLE Stand.Err
.MU:(identity)      96.43427     3.99951
 SIGMA:(identity)  35.70065     3.12339
 Xi:(identity)       0.13926     0.08816

[1] "Negative log-likelihood: 560.03270499351"

 Parameter covariance:[,1][,2][,3]
 [1,] 15.9960614  6.79910970    -0.133392894
 [2,] 6.7991097   9.75554493    -0.065424682
 [3,]-0.133392    -0.06542468    0.007772035
```

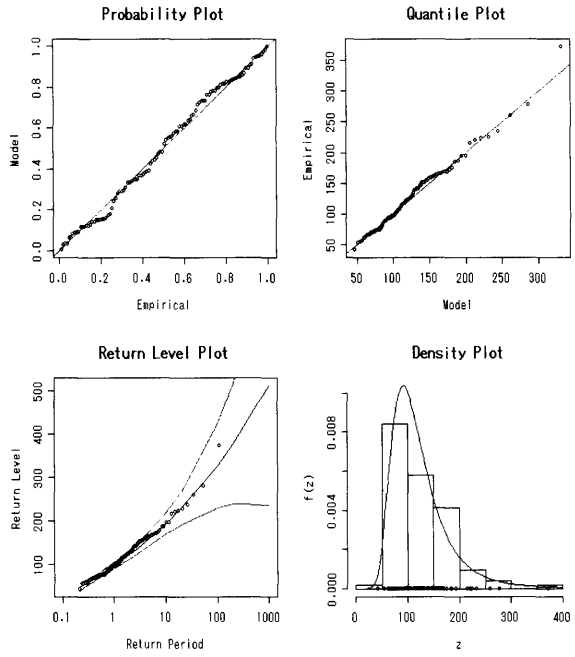


図１３．検証用の４つのグラフ

当てはまりの良さについての考察：モデルが適切でないという証拠は特にないといえる。特に Density Plot ではデータと推定されたモデルの確率密度度関数は良く当てはまっているといえる。また Probability Plot においては多少直線と離れた点も見うけられるが、全体的には問題ないと思われる。最後に Return Level Plot から分かることとして、もしこの当てはめた分布が正しければ日最大が 300 ｍｍを超えるのはおよそ 90 ～ 100 年に一度と考えれる。ところがこの現象が頻繁に起こるのであれば、自然界の構造の変化、すなわち環境破壊が進んでいることを暗示することになるであろう。

５　おわりに

本稿では極値（特に最大値，Block Maximum）の確率分布に関する話題の中でよく知られている漸近的な極値分布を紹介した。そして実際の Block Maximum のデータ

である東京周辺の一日辺りの降水量の年間最大値に一般化極値分布を当てはめ、そのパラメータを推定しモデル化を行った。その結果Type IのGumbel分布によく当てはまっていることがわかった。しかしこのモデルはデータが定常で独立な時系列であることを仮定しており、平均値は一定であると仮定している。このモデルでは近年話題となっている温暖化現象などの効果は考慮されていない。極値理論では非定常な時系列のモデル化として、閾(threshold)モデル、Point Process Characterizationなどいろいろモデルが提案されているので、これらを使えば温暖化などのより詳しい議論ができると思われる。いろいろな地域の降水量のデータや気温データなど範囲を広げて分析することが今後の課題である。

参考文献

[1] S.G.Coles : An Introduction to Statistical Modeling of Extreme Values, Springer Verlag, New York (2001).

[2] P. Embrechts, G. Kluppelberg and T. Mikosch : Modelling Extremal Events for Insurance and Finance, Springer Verlag, New York (1998).

[3] M.R Leadbetter, C. Lindgren and H. Rootzen : Extremes and Related Properties of Random Sequences and Series, Springer Verlag, New York (1983).

[4] E. Gilleland and R. W. Katz : The Extremes Toolkit (extremes) Weather and Climate Applications of Extreme Value Statistics, http://www.assessment.ucar.edu/toolkit.

[5] Statistics of Weather and Climate Extremes, http://www.isse.ucar.edu/extremevalues / extreme. html.