平成20年度　土木学会北海道支部　論文報告集　第65号

非毎年確率の閾値決定法に関する研究

The Research on the Determination of Threshold of Partial Duration Series

北海学園大学社会環境工学科　学生員　○佐伯　健（Takesi　Saeki ）
北海学園大学社会環境工学科　学生員　菅野　秀則　（Hideniri Sugano）
北海学園大学社会環境工学科　正員　許士達広　（Tatsuhiro Kyoshi ）

１．はじめに

従来河川の計画では雨量や流量の年最大値を用いる毎年確率が用いられてきたが、毎年確率で使用されない年間第 2 位以下のデータの有効利用が従前から課題であった。近年河川整備計画検討において、降雨の毎年確率の適合度が低い場合の検討事項として、非毎年確率が位置付けられ非毎年確率の使用頻度が高まっている。また対象とする規模の現象が必ずしも毎年発生しない場合など、毎年確率値が当てはめにくい時には非毎年確率値で考える必要がある。

超過確率を扱う極値統計学は、水文学以外にも海岸の波浪や地震、金融や保険などで用いられる。それらの多くは、ある閾値より大きいデータを非毎年確率で扱う POT 解析と呼ばれるものである。閾値は非毎年確率の確率値算定に大きく影響するが、良い算定方法が無く、非毎年確率の欠点として各分野で問題となっている。

ここでは非毎年確率の課題である閾値の算定方法について回帰直線と、棄却検定の考え方を用いて考察する。

２．非毎年確率の式

非毎年確率として一般的に①指数分布または②一般化パレート分布をポアソン分布で変換したものが用いられている。①はグンベル分布、②は一般化極値分布（GEV 分布）の形となる。一般化極値分布はグンベル分布、対数極値分布 A 型、対数極値分布 B 型を一つの式形に統一したもので、グンベル分布は一般化極値分布の形状母数 k＝0 の場合である。

１）グンベル分布

ポアソン分布の式 $F(x)=\exp\{-\lambda[1-G(x)]\}$ において累積分布関数として指数分布

$$G(x)=1-\exp\left(-\frac{x-x_0}{a}\right)$$ を用いたときはグンベル分布となる。

$$\text{非超過確率 } p=F(x)=\exp\left\{-\lambda\exp\left[-\frac{x-x_0}{a}\right]\right\}$$

$$=\exp\left[-\exp\left(-\frac{x-c}{a}\right)\right] \quad \cdots 1)$$

$$c=x_0+a\ln\lambda$$

クオンタイル $x_p=c-a\ln[-\ln(p)]$ $\cdots 2)$

ここに $p\equiv F(x)$

２）一般化極値分布

ポアソン分布の式 $F(x)=\exp\{-\lambda[1-G(x)]\}$ の G(x)に一般化パレート分布を用いると一般化極値(GEV)分布になる。

$$G(x)=1-\left[1-\frac{k}{a}(x-x_0)\right]^{\frac{1}{k}}$$

$$p=F(x)=\exp\left\{-\lambda\left[1-\frac{k}{a_*}(x-x_0)\right]^{\frac{1}{k}}\right\}$$

$$=\exp\left\{-\left[1-\frac{k}{a_*}(x-c)\right]^{\frac{1}{k}}\right\} \quad \cdots 3)$$

$$a_*=a\lambda^{-k}$$

$$c=x_0+\frac{a}{k}\left(1-\lambda^{-k}\right)$$

クオンタイル $x_p=c+\left(\frac{a}{k}\right)\{1-[-\ln(p)]^k\}$ $\cdots 4)$

ここに　$p\equiv G(x)$　　$(k\neq 0)$

３．閾値の問題点

１）、３）式において閾値は分布の位置母数ｃである。非毎年確率モデルでは、ある閾値を超える値が指数分布あるいは一般化パレート分布をするという前提があり、閾値より小さい値は分布の対象からはずれる。このためｃは分布から求めるのではなく値を仮定して用いることになる。未知のパラメーターは尺度母数 a 一つになり、データから最適化する変数が一つ少なくなるため、非毎年確率の適合度および信頼性は低いものとなる。

閾値より大きなデータから位置母数 c を最適化して求めることは可能であるが、それは仮定した閾値に近い値になり、通常一致はしないが仮定した閾値に依存する。

また分布の平均値と分散が a で等しいという仮定になっているが、実際のデータではそうはならない。

この閾値の決定方法に合理的なものがないということが非毎年確率の最大の問題点であり、これが信頼性の上で決定的な欠点となっている。たとえば図‐１、図‐２は閾値を変えて雨量確率を計算したもので、ある閾値以上のデータ数が多くなるにつれてグンベル分布では確率値が小さくなり、一般化極値分布では大きくなる傾向がみられる。値しだいで値が大きく変わるのであれば確率値は定まらず、閾値の定め方が明確でなければ使い物に

ならないことになる。したがって閾値の決定が極値統計学において極めて重要な課題である。

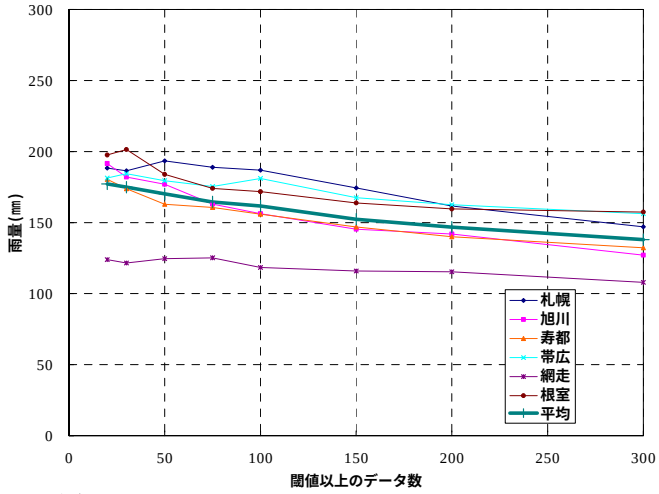


図-1　mbel分布における各観測所の閾値以上のデータ数と確率値（データ年数60年，確率1/100）

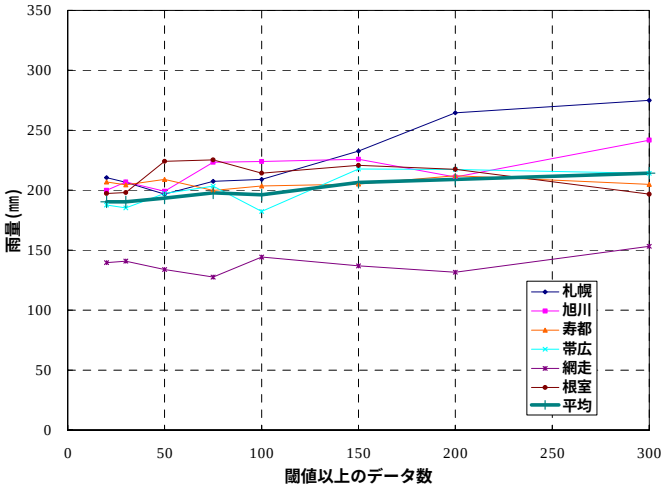


図-2　5 GEV分布における各観測所の閾値以上のデータ数と確率値（データ年数60年，確率1/100）

４．回帰曲線としての非毎年確率

この原因を考えるために、非毎年の水文データをｙ軸、分布の標準変量をｘ軸に取りプロットしてみる。

毎年に直した下からｉ番目の非超過確率ｐｉは

$$p=\exp\{-\lambda[1-G(x)]\}=\exp\left\{-\frac{M}{N}\left(1-\frac{i-\alpha}{M+1-2\alpha}\right)\right\} \quad 5)$$

Ｍ：閾値以上のデータ個数

Ｎ：データ年数　ｉ：プロッティングポジション

α：プロッティングポジションの定数

Ｇ(x)：非超過確率の分布関数、cdf

λ：ポアソン分布の閾値以上の値の年平均生起回数

指数分布によるグンベル分布の確率値は

$$x_p=c+a\{-\ln(-\ln(p))\}=c+a\left[-\ln\left\{\frac{M}{N}\left(1-\frac{i-\alpha}{M+1-2\alpha}\right)\right\}\right] \quad \cdots 6)$$

一般化パレート分布による一般化極値分布の確率値は

$$x_p=c+a\left[\frac{1}{k}\left\{1-\{-\ln(p)\}^k\right\}\right] \quad \cdots 7)$$

$$=c+a\left[\frac{1}{k}\left(1-\left\{\frac{M}{N}\left(1-\frac{i-\alpha}{M+1-2\alpha}\right)\right\}^k\right)\right] \quad \cdots 8)$$

いずれも［　］内がε i であり、

$$x_p=a*\varepsilon_i+c \quad \cdots 9)$$

の直線関係となる。

一般に非毎年確率の定数決定において、ｃは閾値として仮定され、ａとｋは積率法やＬ積率法で求められている。しかしここでは問題の解決を容易にするため、9）式のａおよびｃをデータｙと標準変量ε i から最小二乗法で定める方法をとる。一般化極値分布のｋはデータからＬ積率法などで、先に算出しておく。分布の確率値 xp はこれにより６)式及び７)式において、例えば１／100 確率値はｐ＝0.99、1／200 に対してはｐ＝0.995 に対応して求める。

最小二乗法では９)式において

$$a=\frac{S_{x\varepsilon}}{S_{\varepsilon\varepsilon}}=\frac{\sum(x_i-\bar{x})\sum(\varepsilon_i-\bar{\varepsilon})}{\sum(\varepsilon_i-\bar{\varepsilon})^2} \quad \cdots 10)$$

$$c=\bar{x}-a\bar{\varepsilon} \quad \cdots 8)$$

ただし $\bar{x}$ ：データの平均値　$\bar{\varepsilon}$ ：標準変量の平均値

である。また積率法でプロッティングポジションから係数を求めると

$$a^*=\sqrt{\frac{S_{xx}}{S_{\varepsilon\varepsilon}}}=\sqrt{\frac{\sum(x_i-\bar{x})^2}{\sum(\varepsilon_i-\bar{\varepsilon})^2}} \quad \cdots 11)$$

$$c^*=\bar{x}-a\bar{\varepsilon} \quad \cdots 12)$$

ａとａ*の間には　$a=a^*\times r_{w\varepsilon}$　・・・13）

ただし $r_{x\varepsilon}=\frac{S_{x\varepsilon}}{\sqrt{S_{xx}}\sqrt{S_{\varepsilon\varepsilon}}}$　(相関係数)

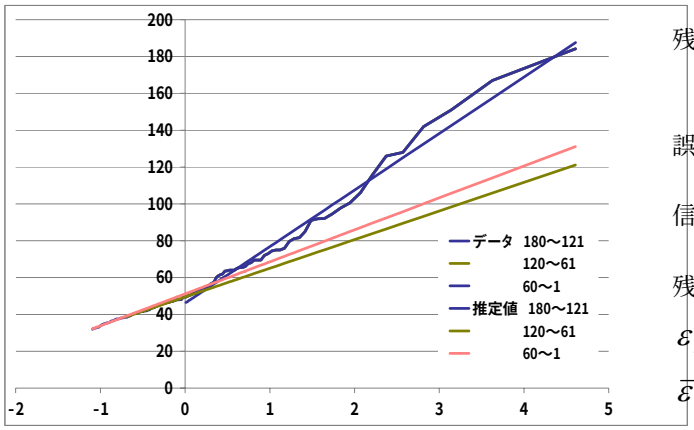
の関係がある。

最終的に必要なことは、確率値を求める関数である毎年確率に直した時のグンベル分布あるいは一般化極値分布において、どの値から上のデータを使うのが確率値推定に望ましいかを知ることである。したがってこの場合は先に閾値を仮定するのではなく、標準変量とデータの値の関係を見てその傾向から確率値を算出するのに適した閾値を見つけ出すことである。

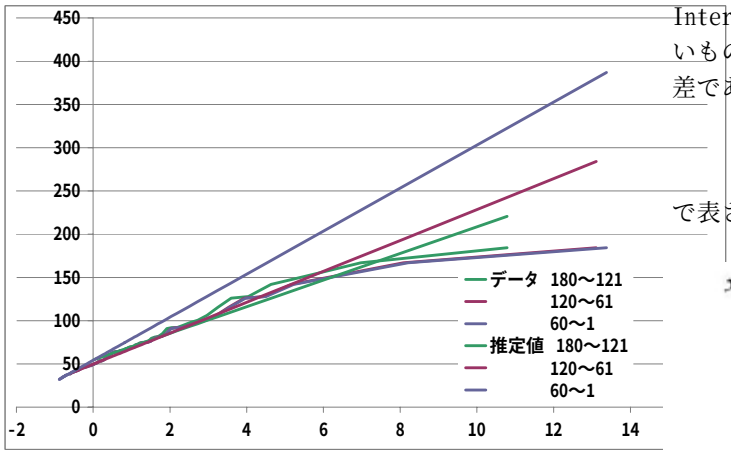
グンベル分布および一般化極値分布におけるｘとε i の関係を描いたのが図－3 及び図－４である。データ年数 60 年でデータ個数を大きいほうから 60,120,180 個取り、それぞれの標準変量にあわせてプロットしている。一般化極値分布の k 値はデータ数に合わせて、決めなおしている。

これで見るとグンベル分布は下に凸、一般化極値分布は上に凸であることが見て取れる。確率値はこのデータ

列の回帰直線上で標準変量から外挿される。図には最大の 60 個、（下から 121～180）、中間の 60 個（61～120）一番小さい 60 個（1～60）ごとに当てはめた回帰直線を記入している。回帰直線から分かるように、年に何回も起こるような非毎年データの小さい値を取り入れていくと、グンベル分布は確率値が小さく、一般化極値分布は確率値がやや大きくなっていく傾向が見られる。



図－3　グンベル分布のデータと標準変量のプロット



図‐４一般化極値分布のデータと標準変量のプロット

５．回帰.直線の信頼区間

非毎年確率がグンベル分布あるいは一般化極値分布をしているという仮定に立つと、データは横軸に標準変量、縦軸にデータ値をとったグラフにおいて直線で表される。

一般にデータのうちどの部分を選んで確率値などを定めるかを決めるのには、必要なデータとそれ以外のデータに分けて検定により判定する。この考え方を回帰直線に適応すれば直線回帰の傾向が同じものが確率値の推定において同じデータ群と考えられる。

直線回帰における誤差分散は以下のように示される。回帰直線における信頼区間の理論を水文量確率値に適用する。水文量確率値x_pは、一般的に標準変量εの一次式

$$x_p = \bar{x} + \sigma \cdot \varepsilon \qquad \cdots 14)$$

で表わされる。

いま、下図‐５のようにｙ軸を水文量x、x軸を標準変量εとする。このとき、水文量xの水文量確率値における信頼区間は、「t分布の自由度$n-2$および片側超過確率$\alpha/2$に対する臨界値ｔ(ｎ－２,α)」と「観測データごとに定まる分布の適合度を表す誤差の標準偏差 Se」及び「求める確率分布の標準変量に対する信頼係数h_0」の積によって表される。

$$\text{水文量確率値の信頼区間} = t(n-2,\alpha) \times Se \times ho \qquad \cdots 12)$$

誤差の標準偏差 Se　、信頼係数h_0およびそれに伴う残差平方和は次式より求めることとなる

$$\text{誤差の標準偏差} \qquad Se = \sqrt{\frac{\sum (\hat{x}_i - x_i)^2}{n-2}} \qquad 15)$$

$$\text{信頼係数} \qquad ho = \sqrt{1 + \frac{1}{n} + \frac{(\varepsilon_p - \bar{\varepsilon})^2}{S_{\varepsilon\varepsilon}}} \qquad 16)$$

$$\text{残差平方和} \qquad S_{\varepsilon\varepsilon} = \sum (\varepsilon_i - \bar{\varepsilon})^2 \qquad 17)$$

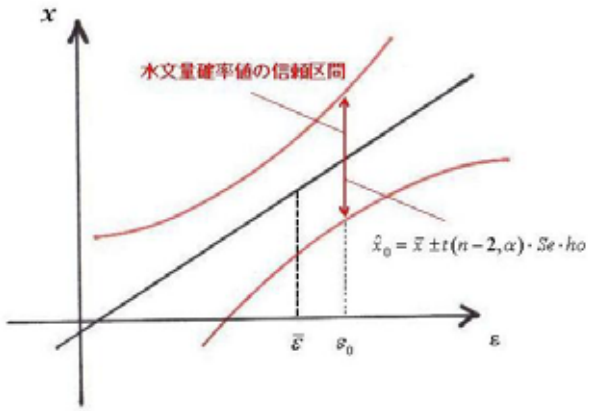
ε_i：標準変量、ε_p：求める確率値x_pにおける標準変量、

$\bar{\varepsilon}$：標準変量の平均値

ここで、t分布の臨界値ｔ(ｎ－２,α)は、水文確率分布の種類や水文データの値にも関係しない。本研究では臨界値を除く次式を信頼区間の指標S_{CI}（Confidence Interval）と定義する。この信頼区間の指標の値が小さいものほど信頼性が高い。これは確率値の誤差の標準偏差であるから

$$S_{CI} = Se \times h_0 \qquad \cdots \quad 18)$$

で表され、その二乗S_{CI}^2が確率値の誤差分散である。



図－5 直線の信頼区間

６．回帰直線による確率値の検定とデータの区分

非毎年確率でデータをプロットし、直線回帰で確率ｘｐを求める。このとき、前述したようにデータを取る範囲によって回帰直線が変化し、確率値も変わる。どこから上のデータで求めるのかが問題であるが、一つの方法として一定の信頼性をもって同じ確率値が得られるデータ集団で区切ることが考えられる。

ンをとり、データを２つに分け、データの値が大きいほうの個数 n1 個のデータと、値が小さいほうの個数 n2

個のデータの２つのデータ群のそれぞれを直線回帰させて、超過確率 p の標準変量に対応するデータ値すなわちそれぞれのデータによる確率値を求める。
上下二つのデータ群によるそれぞれの確率値ｘp1、xp2 の誤差分散は 13)~16）式よる S_{CI}^2 のようになる。

この２つの直線を分けたほうが良いか,分けないで１つにしたほうが良いかについては確率値の差の検定が有効である。２つのデータ群を比較する時に通常分散の比のＦ検定と平均値の差の検定が行なわれるが、確率値についても同じ考え方を適用する。

１）確率値の差の検定

確率値ｘp1、xp2 の差ｘp1－xp2 の分散はそれぞれの分散の和 $S_{CI1}^2+S_{CI2}^2$ で表される。それぞれの確率分布のデータ数を n_1 および n_2 とすると
これをｔ検定する場合

$$T=\frac{x_{p1}-x_{p2}}{\sqrt{S_{CI1}^2+S_{CI2}^2}} \qquad \cdots 17)$$

分散 S_{CI1}^2, S_{CI2}^2 が等しい時には自由度ｆは

$$f=n_1+n_2-4 \qquad \cdots 18)$$

分散が異なる時の自由度は

$$f=\frac{(S_{CI1}^2+S_{CI2}^2)^2}{\left(\frac{S_{CI1}^4}{n_1-2}+\frac{S_{CI2}^4}{n_2-2}\right)} \qquad \cdots 19)$$

有意水準αの両側検定である。１７)式で計算されるＴ値の絶対値が t(f,α/2)より大きい時確率値ｘp1、xp2 が等しいという仮説が棄却される。αは通常５%をとる。

分散が等しいかどうかはＦ検定により判断される。

$$F=S_{CI2}^2 \Big/ S_{CI1}^2 \qquad \cdots 20)$$

20)式で計算されるＦ値が
Ｆ＞Ｆ（n1－2，n2－2,α/２）あるいは
Ｆ＜F(n1－2，n2－2，１－α/２)の時に２つの分散が等しいという仮説は棄却される。T 検定の自由度は棄却されない時は 18）式、棄却さた時は 19）式をで定める。F 検定の有意水準αは５％あるいは 10%をとる。この場合は分母 S_{CI1}^2 が一番値が大きいデータ群の確率値の分散をとする。

２）確率値の信頼度の評価

確率値の誤差分散は確率値の信頼度を表し誤差分散が大きいものは信頼が低い。何十年、何年に１回といった確率値に近い大きい価のデータは使用されるが、年に何回も起きる規模のデータは、100 年に一度といった確率値を求めるどこまで取り入れるべきかが問題である。これらの値の小さいデータは標準変量の平均値 $\bar{\varepsilon}$ と確率値 ε_p との差が大きいため、それらによる確率値の分散が大きくなる傾向がある。したがって確率値の差の検定ばかりでなく、分散の比の検定で F の大きいものを棄却することも考えるべきであると。

7. 計算例

3 節に示したデータを用いて大小で分けたデータごとの確率値、および確率値の分散を求め、それらが同一とみなされるか検定したのが表ー1、表―2 である。

グンベル分布			
データ順	1～60	61～120	121～180
確率値	131.04	121.0	187.3
ｈｐ	6.189	3.412	1.121
SCI	1.598	0.938	4.162
F 検定	19.674	6.738	―
T 検定	12.620	15.537	―

Fα＝　0.05 のとき 1.546 以上は棄却
Fα=0.025 のとき 1.682 以上は棄却
ｔ　α=0.05 のとき 1.980 以上は棄却
α＝　0.025 のとき 2.270 以上は棄却

一般化極値分布			
データ順	1～60	61～120	121～180
確率値	386.0	283.6	220.0
ｈｐ	23.78	10.98	1.20
SCI	3.968	2.502	8.378
F 検定	1.436	5.464	―
T 検定	15.22	6.957	―

この結果、このデータでは大小 60 個ごとに区切られたデータごとの確率値は差が大きく、同一とみなされないことが分かった。

終わりに

非毎年確率値の閾値の決定は、各分野で問題になっているが、その決定法についてひとつの提案が出来たと思われる。今後検討する事項はいろいろありさらに研究を進めたい。

参考文献

1)星　清:開発土木研究月報、北海道開発土木研究所 pp.35-37、pp.38-40、pp.41-44、pp.46-48
2)(財)国土開発技術研究センター：高水計画の手引き検討の手引き(案) 平成 12 年 10 月　参考資料:水文統計解析の概要 pp5、pp7、pp10-17、pp24、pp20
3)竹内 啓 監修　芳賀 敏郎・野澤 昌弘・岸本 敦司 著 SAS による回帰分析 pp.17、21、197-198、200
4)ラリー・ゴミック、ウルコット・スミス 著　中村和幸訳確率統計が驚異的によくわかる pp187-206
10) G.W.Kite　FREQUENCYAND RISK ANALYSES IN HYDROGYbyWATER RESOURCES PUBLICATIONS pp.33-35、pp.69-77、pp.97、pp.105-113、pp.123-127