水文確率分布の適合度と信頼性の評価指標に関する研究

A study on the index of adaptability and reliability of hydrological Probability distribution model

北海学園大学社会環境工学科　学生員　○島田暁仁 (Akihito Shimada)
北海学園大学社会環境工学科　学生員　飯束亮 (Ryo Iizuka)
北海学園大学社会環境工学科　正員　許士達広（Tatuhiro Kyoshi ）

１．はじめに

現在の河川流域における様々な水工計画は、所要の安全度のもとに降雨量や河川流量データの確率統計解析を行い、その統計量(水文量確率値)を定めたうえで策定されている。この水文量確率値を求める上で、多くの確率分布モデルや母数推定法などが用いられている。これらの検討した水文量確率値のうちどれが最適であるかが水工計画上の大きな課題であり、それには最適化のための手法および評価指標が重要となる。

評価指標としては、一般的に確率分布関数の原データへの分布の「適合度」を重視するものとして、SLSC（標準最小二乗規準）や相関係数（X-COR、P-COR)、AIC などが用いられている。また、実際の計画においては標本値および標本数 N の違いによる確率値の変動が小さいこと（変動性最小）が望まれる。これをここでは分布の「安定性」と呼ぶこととし、これには、Jackknife 法やブートストラップ法といったリサンプリング法の推定誤差分散の指標が用いられている。

どの確率分布モデルが最適かということに対して様々な研究がなされ、客観的・定量的な評価について近年いくつかの提案がされているが明確な答えはなく、実用的には分布の「適合度」に関する指標 SLSC を一定の条件におきつつ、観測データ毎に各種のモデル計算を行い、いくつかの手法の結果を比較して水文量確率値の決定を行っているのが実情である。

本研究では、これら分布の適合度と安定性の両方に関係する第 3 の指標として、「水文量確率値の信頼区間」に着目し、分布の「信頼性」に関する新たな指標を提案する。さらに、確率分布の「適合度」および「安定性」、「信頼性」に関わる評価指標を算出し、各指標の水文量確率分布に対する適用性および各確率分布の指標に対する特性について比較して考察を行う。

２．研究方法

本研究では評価指標の比較のため、既往の日雨量毎年最大値データに対し各確率分布モデルを用いて、各確率分布および評価指標の特性･類似性･関連性を検討する。評価指標は以下のものを対象とする。

①分布の適合度として、標準変量の理論値と実測値の相対誤差を表す「SLSC（標準最小二乗規準)」、相関係数「X-COR」および情報量基準 AIC を用いる。

② 水文量確率値の信頼区間について、水文量確率値と標準変量の直線回帰の考え方を用いて、「水文量確率値の信頼区間の指標 S_{CI}」を提案し、算出する。

③ 分布の安定性として、「Jackknife 法によるリサンプリング計算」を行い確率水文量の推定誤差分散を算出する。これは「水文量確率値の信頼区間」を表す指標でもあると考えられる。また上記の S の要素であるてこ比についても比較する。

確率分布については、河川計画に用いられる降雨量などの水文量確率値を求める上で、現在用いられている水文量確率分布モデルの大半は、対数正規分布、極値分布、ガンマ分布の三種類に大別される。本研究では、それぞれの代表的な方法として、2 母数対数正規分布、3 母数対数正規分布、グンベル分布、GEV 分布、および、対数ピアソンⅢ型分布を検討対象とする。また、これらの母数の算定には一般的には積率法、及び L 積率法が用いられるが、それらと対比しつつ主としてプロッティングポジションを利用した、最小二乗法、及び積率法を用いる。

検討に使用するデータについては、北海道内 8 ヶ所(札幌市、旭川市、函館市、帯広市、釧路市、網走市、根室市、寿都町）の気象台観測の日雨量毎年最大値データを用いることとする。

３．検討対象とする確率分布モデル

（１）確率分布

ａ）3 母数対数正規分布

水文諸量のヒストグラムは、多くが単峰な非対称分布を示す。このような場合に、対数変換を用いると近似的に正規分布にみなすことができ、正規分布に関する多くの確率分布特性が利用できる。このとき、対数正規分布の確率密度関数は以下のように表される。

$$f(x)=\frac{1}{(x-a)\sqrt{2\pi}\sigma_y}\exp\left[-\frac{1}{2}\left(\frac{\ln(x-a)-\mu_y}{\sigma_y}\right)^2\right] \quad 1)$$

ただし、水文量の対数　$y=\ln(x-a)$　$x > a$

この確率値 y_p は標準変量 Z_p の一次式　$y_p=\mu_y+\sigma_y\cdot Z_p$

x ：水文量、a ：分布下限値、Z_p：標準変量

μ_y ：yの平均 、 σ_y ：yの標準偏差

で表すことができる。ここで、標準変量 Z_pを求めるには以下の式を用いる。

$$Z_p = -\sqrt{\frac{t^2\left[(4t+100)t+205\right]}{\left[(2t+56)t+192\right]t+131}} \quad 2)$$

非超過確率 P≦0.5 のとき

係数 $t = -\ln(2p)$

非超過確率 P>0.5 のときは $t = -\ln\left[2(1-p)\right]$

となり、その時の標準変量Z_p は、式 2)を

$Z_p = -Z_{1-p}$ へ置き換えることにより求めることができる。

b）グンベル分布

最大値分布の第Ⅰ型形式として、最大洪水流量の分布関数に用いることの有用性が知られている、グンベル分布の確率密度関数は以下のように表される。

$$f(x) = \frac{1}{a}\exp\left[-\frac{x-a}{a} - \exp\left(\frac{x-c}{a}\right)\right] \quad 3)$$

$$(-\infty < x < \infty)$$

上式中の各項目として、x：水文量、a：尺度母数、c：位置母数　また、水文量xの確率値 x_pは標準変量 G_p の一次式で表すことができる。

$$x_p = c + a \cdot G_p \quad 4)$$

ここで、標準変量G_p は非超過確率 p を用いて、次式で表される。

$$G_p = -\ln\left[-\ln(p)\right] \quad 5)$$

c）対数ピアソンⅢ型分布

水文量の対数変換により水文量xの対数yがピアソンⅢ型分布に従うとき、分布は対数ピアソンⅢ型分布となり、そのときの確率密度関数は以下のように表される。

$$f(x) = \frac{1}{|a|\Gamma(b)x}\left(\frac{\ln x - c}{a}\right)^{b-1}\exp\left(-\frac{\ln x - c}{a}\right) \quad 6)$$

上式中の各項目として、x：水文量、a：尺度母数、

b：形状母数、c：位置母数、$\Gamma(\cdot)$：ガンマ関数

この確率値 y_p は、標準変量K_p　の一次式で表すことができる。

$$y_p = \mu_y + \sigma_y \cdot K_p \quad 7)$$

このとき、μ_y　σ_y がyの平均および標準偏差である。

標準変量K_pは次式で求められる。

$$K_p = \frac{2}{\gamma_y}\left(1 + \frac{\gamma_y Z_p}{6} - \frac{\gamma_y^2}{36}\right)^3 - \frac{2}{\gamma_y} \quad 8)$$

d）GEV 分布（一般化極値分布）

3 種類の極値分布（グンベル分布、対数極値分布 A 型および B 型）を１つの式形に統一したものが GEV 分布（一般化極値分布）と呼ばれており、累積分布関数は以下のように表される。

ここで、：γ_y ひずみ係数、Z_p：N(0,1) 従う標準正規変量

$$F(x) = \exp\left\{-\left[1 - \frac{k(x-c)}{a}\right]^{1/k}\right\} \quad 9)$$

$$(\ k \neq 0\ )$$

上式中の各項目として、x：水文量、a：尺度母数、c：位置母数、k：形状母数

また、水文量xの確率値 x_p は、標準変量 E_p の一次式で表すことができる。

$$x_p = c + \left(\frac{a}{k}\right)\cdot E_p \quad 10)$$

この標準変量E_pは、非超過確率 p を用いて次式より求められる。

$$E_p = 1 - \{-\ln(p)\}^k \quad 11)$$

このように、こられの検討対象確率分布における水文量xまたはその対数yの確率値は、一般的に標準変量

Z_p、G_p またはK_p、E_pの一次式で表され、直線回帰の問題として考えることができる。

４．適合度の評価指標

（１）確率分布の適合度を表す指標

a）SLSC（標準最小二乗規準 Standard Least-Squares Criterion）

SLSC は、確率分布モデルのデータへの適合度を表す評価基準として、次式によって定義される。

$$SLSC = \frac{\sqrt{\zeta^2}}{|S_{0.99} - S_{0.01}|} \quad 12)$$

$$\zeta^2 = \frac{1}{N}\sum_{i=1}^{N}(S_i - r_i)^2 \quad 13)$$

$S_{0.99}$、$S_{0.01}$：非超過確率を 0.99 及び 0.01 とした時の当該確率分布の標準変量

Si:順序統計量データを推定母数で変換した標準変量 S_i

r_i：プロッティングポジションに対応した理論確率水文量を推定母数で変換した標準変量

SLSC はその算出される値が小さいほどよく適合していると考えられ、近年の研究において河川流量において満足すべき適合度の基準として SLSC　0.04 以下とすることが適当であるとされている)。

b）相関係数(X-COR)

重相関係数は、同じプロッティングポジションの原データの順序統計量x_iと理論統計量y_iの相関関係を表すも

のである。その 2 乗は決定係数として回帰分析などに用いられる

$$X-COR = \frac{\sum_{i=1}^{N}(X_i-\overline{X})(Y_i-\overline{Y})}{\left[\left\{\sum_{i=1}^{N}(X_i-\overline{X})^2\right\}\left\{\sum_{i=1}^{N}(Y_i-\overline{Y})^2\right\}\right]^{1/2}} \quad 14)$$

ここに、N は資料の個数、(X_i, Y_i)は ｉ 番目における原データと、理論統計量の組の値。 $\overline{X}$ 、$\overline{Y}$ はそれぞれの変数の平均値である。

また重相関係数は変数がひとつの場合はデータと標準変量の相関係数に一致し

$$X-COR = \frac{\sum_{i=1}^{N}(X_i-\overline{X})(\varepsilon_i-\overline{\varepsilon})}{\left[\left\{\sum_{i=1}^{N}(X_i-\overline{X})^2\right\}\left\{\sum_{i=1}^{N}(\varepsilon_i-\overline{\varepsilon})^2\right\}\right]^{1/2}} \quad 15)$$

で表される。X-COR は値の絶対値が大きく 1 に近いものが良好と判断される。

（ｃ）情報量基準 AIC

適合度と母数の個数を加味したもので、次式で表される。

$$AIC = -2MLL + 2m \quad 16)$$

$$MLL = \sum_{i=1}^{N} \ln f(x_i, \hat{\theta}_j) \quad 17)$$

$f(x_i, \hat{\theta}_j)$：m個の母数θ_jを持つ当該分布の確率密度関数、$\hat{\theta}_j$：母数推定値　　ＭＬＬ：対数尤度の最大値

ＭＬＬの値が大きいほどモデルの適合度が良い。母数の数が増加すると、第 1 項目は小さくなるが第 2 項目が大きくなる。

2）確率分布の安定性を示す指標

ａ）Jackknife 法による誤差分散

水文資料の蓄積が進んでいくなかで、河川計画の面では、新たな標本が加わることによって水文量確率値が大きく変動しないような確率分布モデルと母数推定法を選定することが望まれている。誤差分散は新たなデータの付け加えによる確率値の変動性を表し、信頼性とともに安定性に関する指標としても考えられる。特に安定性を示す誤差分散算定の方法として、リサンプリング法があり、代表的なリサンプリング法として Jackknife 法があげられる。

Jackknife 法とは、まずＮ個のデータ X_iを用いて確率値$\hat{\theta}$をもとめる。次に ｉ 番目のデータを除いた（Ｎ－１）個のデータによる推定値を$\hat{\theta}_{(i)}$とすれば$\hat{\theta}_{(i)}$はＮ個得られる。$\hat{\theta}_{(i)}$の平均値$\hat{\theta}_{(\bullet)}$を求めると、確率値の推定誤差分散は

$$S^2 = \frac{1}{N}(N-1)\sum_{i=1}^{N}\left[\hat{\theta}_{(i)} - \hat{\theta}_{(\bullet)}\right]^2 \quad 18)$$

これを Jackknife 誤差分散と呼ぶ。

3）確率値の信頼性に関する指標

１ 章で示したように、主要な確率分布は標準変量の一次式で表される。本研究では、これを利用して回帰曲線の信頼区間の理論から水文量確率値の信頼性を推定する方法を提案する。適合度の指標はデータに対する分布の当てはまりを判定するが、実際に必要なのは分布ではなく確率値そのものの信頼性である。回帰分析の信頼区間は、その内容から見て、適合度と安定性を兼ね備えた指標と考えられ、水文確率分布の評価において重要な指標であると考えられる。

水文量確率値ｘ$_p$は、一般的に標準変量εの一次式

$$x_p = \overline{x} + \sigma \cdot \varepsilon \quad 19)$$

で表わされる。

いま、下図のように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。

水文量確率値の信頼区間$= t(n-2,\alpha) \times Se \times ho$　20)

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

誤差の標準偏差　$Se = \sqrt{\frac{\sum(\hat{x}_i - x_i)^2}{n-k}}$　21)

信頼係数　$ho = \sqrt{1 + \frac{1}{n} + \frac{(\varepsilon_p - \overline{\varepsilon})^2}{S_{\varepsilon\varepsilon}}}$　22)

残差平方和　$S_{\varepsilon\varepsilon} = \sum(\varepsilon_i - \overline{\varepsilon})^2$　23)

ε_i：標準変量、ε$_p$：求める確率値x_pにおける標準変量、$\overline{\varepsilon}$：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。

$$S_{CI} = Se \times h_0 \quad \textbf{24)}$$

この二乗のＳ$_{CI}{}^2$が確率値の誤差分散である。この信頼区間のこれは確率値の誤差の標準偏差であるから確率値そのものの信頼性を示し、指標の値が小さいものほど信頼性が高い。Se はデータの確率分布に対する回帰の誤差分散であり、ｋは分布の母数の数である。またｈ$_0$の中に、てこ比というデータの変動の伴う分布の回帰直線の変動を表す指標が含まれている。すなわち Se は適合度の指標であり、ｈ$_0$は安定性の指標と考えられるため、それらの積である Ｓ$_{CI}$ は確率値の総合的な信頼性を表す指標として､期待される。

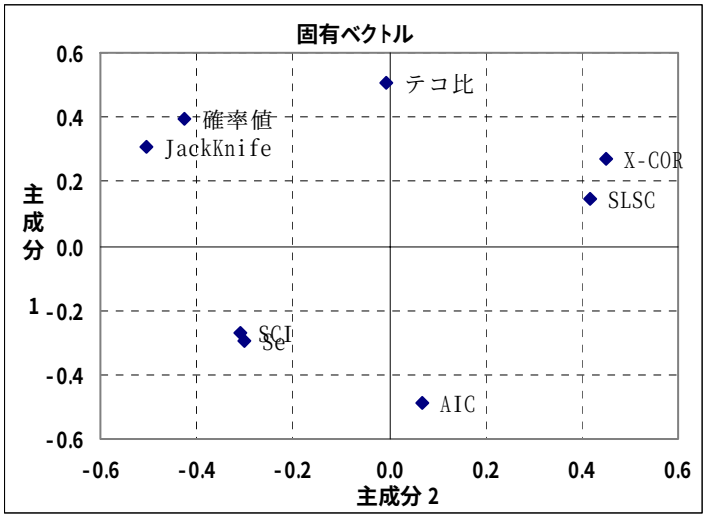
５．数値計算と考察。

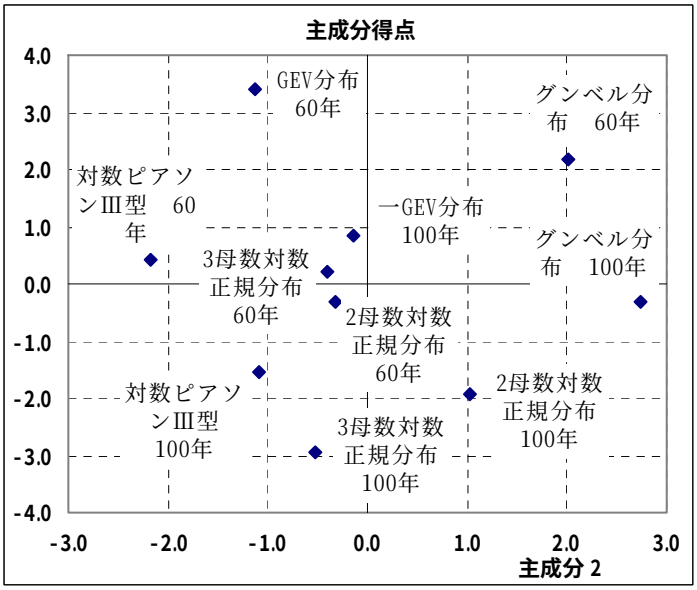
北海道内の主要都市の日雨量について、１/100 確率値及び各指標を求めた。表は旭川市のものであり、最近６０

年及び１００年のデータから求めている。分布関数５種　　に主成分分析を行なった。
類、指標を８種類選びそれぞれの関係と特性を見るため

表１－確率分布と指標の計算例

場所	分布	確率値	Se	SLSC	X-COR	AIC	テコ比	Sc I	JackKnife
札幌市	2 母数対数正規分布 60 年	182.85	10.291	0.0291	0.9862	454.0	0.1109	11.0328	184
	2 母数対数正規分布 100 年	166.29	8.537	0.0289	0.9864	716.6	0.0657	8.9030	180.0
	3 母数対数正規分布 60 年	183.39	9.720	0.0297	0.9851	449.1	0.1109	10.4203	191.6
	3 母数対数正規分布 100 年	167.87	9.762	0.0234	0.9914	745.5	0.0777	10.1804	176.5
	グンベル分布 60 年	181.94	5.414	0.0322	0.9814	374.9	0.1943	5.9167	181.9
	グンベル分布 100 年	168.79	4.395	0.0293	0.9848	581.9	0.1138	4.6389	168.8
	一般化極値分布 60 年	200.05	4.481	0.0253	0.9872	312.6	0.2450	5.0007	198
	一般化極値分布 100 年	186.21	3.732	0.0237	0.9890	551.2	0.1472	3.9976	187
	対数ピアソンIII型 60 年	196.58	9.995	0.0263	0.9888	452.5	0.1258	10.7870	196.1
	対数ピアソンIII型 100 年	182.65	7.360	0.0227	0.9917	689.0	0.0782	7.7201	183.5
	平均	181.66	7.369	0.0271	0.9872	532.7	0.1270	7.8598	184.74
	標準偏差	11.445	2.630	0.00321	0.003156	148.4	0.0559	2.7673	8.8884





主成分１から見ると、適合度を表す指標 AIC、Se とは反対の位置にある。同じく従来適合度を示す指標と思われている SLSC と X－COR は比較的テコ比のほうに近い位置にある。同じ適合度でも変動性のファクターを含んでいることが考えられる。S_{CI} は適合度と安定性の両方を見る総合的指標と期待されるが、式中に含む Se の影響が大きく、Se とあまり違わない性質になっている。主成分２から見ると SLSC と X-COR は性質が近く、変動性の指標 jackknife と相反する形になっている。

主成分得点から見ると、一般化極値分布はテコ比やジャックナイフなど安定性の良い分布の位置にあるが、確率値の影響を受けた可能性もある。3 母数対数正規分布は AIC など適合度が良いモデルといえる。グンベル分布は２母数であり、従来あまり適合度が良い指標とは考えられていないが、X-COR および SLSC は良い値になっている。

６．終わりに

水文統計において、どの確率モデルが良いかということは長く問題にされてきた。現在は適合度以外にも安定性や頑健性といった要素も考えられるようになっている。またモデルの良し悪しと同時に、何が真の確率値なのかという観点からの検討が重要である。今回は検討データが少ない段階であり明確な方向が見出せていないが、さらに検討を進めたい。

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)水文・水資源学会［編集］　水文・水資源ハンドブック　pp239-247