

適合性と変動性を考えた水文確率評価の研究

A Study on Evaluation Criterion of Hydrological Probability Value from the Viewpoint of Fitness and Variability

北海学園大学大学院工学研究科 ○学生員 岡田拓巳 (Takumi Okada)
 北海学園大学工学部社会環境工学科 フェロー 許士達広 (Tatsuhiro Kyoshi)

1. はじめに

河川計画に用いられる水文確率値の算定については、一般には既に一定の理論的整理がなされていると考えられている。しかし現実にはどのような確率モデルが良いか、その算定方法はどの方式が良いか、それを何で評価するかについては、依然として明確な答えは得られていない。本研究では水文確率の最適値の算定に重要な根本的問題である「何をもって最適とするか」の判断基準について考察する。

2. 適合性と変動性

河川などの計画において目標値を定める場合、既往のデータに合わせてモデルと呼ばれる確率密度関数の定数を定めて、100年に1回といった確率値を算出する。この時一般的に用いられるのは図-1に示すデータの点と推定する線（関数）のとの誤差を小さくする手法で、これはデータと関数の「適合性」の面から最適化を図っている。しかし観測されるデータが変わったり追加されたりすると、推定する式が変化し推定に不確実性が生ずる。これを「変動性」と呼び、図-2に示すもう一つの誤差を生じる要素である。この両者はデータへの適合性が高いモデルはデータの変化に追従して変動しやすいため、

一般的に両立しない。適合性のみで判断されることもあるが、両方合わせて最適化する方法は知られておらず、現状では各モデルについて各種の指標を算出し、確率値と合わせて人間が総合的判断をしている。

3. 適合性の指標

1) 決定係数

回帰モデルの代表的指標に決定係数があるが、 y の式のパラメータ数 k （例えば2母数は2,3母数ならば3）の扱いによって以下のように異なる。標本のみ考える場合は決定係数 R^2 、母集団を考える場合は自由度調整済み決定係数 R^{2*} 、予測を考える場合には二重自由度調整済み決定係数 R^{2**} が対応する。

$$R^2 = 1 - \frac{s_e^2}{s_y^2} \quad \dots \quad 1)$$

$$s_y^2 = \frac{1}{n} \sum (y - \bar{y})^2 \quad s_e^2 = \frac{1}{n} \sum (y - \hat{y})^2$$

$$R^{2*} = 1 - \frac{s_e^{2*}}{s_y^{2*}} \quad \dots \quad 2)$$

$$s_y^{2*} = \frac{n}{n-1} s_y^2 \quad s_e^{2*} = \frac{n}{n-k} s_e^2$$

$$R^{2**} = 1 - \frac{s_e^{2**}}{s_y^{2**}} \quad \dots \quad 3)$$

$$s_y^{2**} = \frac{n+1}{n-1} s_y^2 \quad s_e^{2**} = \frac{n+k}{n-k} s_e^2$$

この違いは母数推定にデータが使われれば母数自体の変動が加わることを表現しており、 R^{2*} や R^{2**} を用いれば2母数より3母数の方が変動性が大きくなる。一般的に母数が多ければ適合度による誤差は小さくなるが変動性の誤差は大きくなる。回帰分析においては無関係なパラメータによる精度向上を防ぐために R^2 より R^{2*} が推奨されている。 R^{2**} が用いられることは少ないが s_e^{2**} は最終予測誤差と呼ばれるものであり、これから以下のようにAICに等しいものが導出され、相応の意味を持つものである。

$$L = n \log s_e^{2**} = n \log s_e^2 + n \log \left(\frac{n+k}{n-k} \right)$$

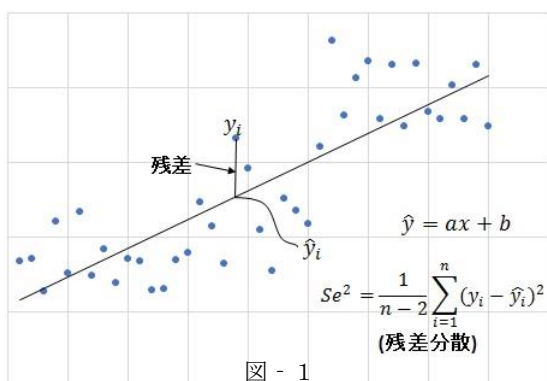


図-1

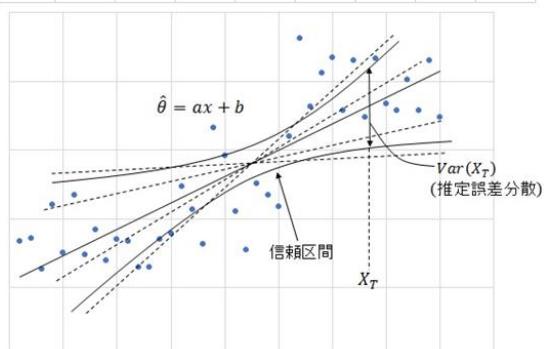


図-2

$$\begin{aligned} \log\left(\frac{n+k}{n-k}\right) &= \log\left(1 + \frac{2k}{n+k}\right) \\ &\approx n \frac{2k}{n+k} \approx 2k \end{aligned}$$

$$L \approx \log s_e^2 + 2k = \text{AIC}$$

2) SLSC

近年 SLSC がモデル評価に多く用いられている。

$$\text{SLSC} = \frac{\sqrt{\varepsilon^2}}{|s_{0.99} - s_{0.01}|} \quad \varepsilon^2 = \frac{1}{N} \sum (s_i - r_i)^2 \quad (5)$$

$S_{0.99}$ 、 $S_{0.01}$:それぞれの非超過確率 0.99 と 0.01 に対する当該確率分布の標準偏差

si:順序統計量データを推定母数で変換した標準変量

ri:プロット位置に対応した理論クオンタイルを推定母数で変換した標準変量

SLSC は小さければ小さいほど適合度が良い。前出の決定係数と比較すると、決定係数はデータ方向の残差分散 s_e^2 をデータ方向の分散 s_y^2 で除している。これに対して SLSC は標準変量方向の残差分散の平方根を標準変量の範囲で割っている。残差をとる方向が違うが本質的にやっていることは同じであり、二つの指標はほとんど同じ傾向を示す。しかし SLSC は以下のような問題点を有する。

①SLSC が 0.04 以下であることが目安とされるがデータに影響され、標本数が多いほど達成しやすくなる。これは分散は大数の法則等によりデータ数が多ければ分母の絶対値はデータ数により変わらないためである。

②自由度調整済み決定係数のような、パラメータによる変動の影響が考慮されていない。2母数でも3母数でもデータとの残差のみで評価されている。

③通常の水文確率計算ではほとんど影響がないが、同じデータでも $y = ax + b$ と $x = cy + d$ は相関係数が小さくなるにしたがってかなり違ったものとなる。また勾配が小さければ s_e^2 が小さくても ε^2 かなり大きいといったことも生ずる。

4. 変動の指標

1) 理論式

変動を表す評価指標は各モデルにおける理論式によるものとジャックナイフ法やブートストラップ法の理論式によるものがある積率法の誤差分散の一般理論式は以下のように与えられる。

$$S_T^2 = \frac{\mu_2}{n} \left\{ 1 + K\gamma_1 + \frac{K^2}{4} [\gamma_2 - 1] + \frac{\delta K}{\delta \gamma_1} [2\gamma_2 - \right.$$

$$\left. 3\gamma_1^2 - 6 + K(\gamma_3 - 6\gamma_1\gamma_2/4)] + \left(\frac{\delta K}{\delta \gamma_1}\right)^2 [\gamma_4 - 3\gamma_3\gamma_1 - 6\gamma_2 - \frac{9\gamma_1^2\gamma_2}{4} + \frac{35\gamma_1^2}{4} + 9] \right\} \quad \dots (6)$$

$$\gamma_1 = \mu_3/\mu_2^{3/2} \quad \gamma_2 = \mu_4/\mu_2^2$$

$$\gamma_3 = \mu_5/\mu_2^{5/2} \quad \gamma_4 = \mu_6/\mu_2^3$$

$\mu_2 \sim \mu_6$: 2~6 次モーメント

ε_p : 確率値の標準変量 $\bar{\varepsilon}$: 標準変量の平均

σ_ε : 標準変量の標準偏差

モーメントはデータから求め一般化極値分布の場合

$$\varepsilon_p = \frac{1}{k} \{1 - (\ln p)^k\} \quad \bar{\varepsilon} = \frac{1}{k} \{1 - \Gamma(1+k)\} \quad k: \text{形状母数}$$

$$K_p = \frac{\varepsilon_p - \bar{\varepsilon}}{\sigma_\varepsilon} = \frac{\frac{1}{k} \{1 - (\ln p)^k\} - \frac{1}{k} \{1 - \Gamma(1+k)\}}{\frac{1}{|k|} \sqrt{(1+2k) - \Gamma^2(1+k)}}$$

7)

$$\frac{\partial K_p}{\partial \gamma_1} = \frac{K_p^+ - K_p^-}{2\Delta \gamma}$$

K_p^+ : ひずみ係数が $\Delta \gamma$ 増加した時の K_p

K_p^- : ひずみ係数が $\Delta \gamma$ 減少した時の K_p

グンベル分布の場合 $\frac{\partial K_p}{\partial \gamma_1} = 0$ であり、モーメントはデータから求めるほか、モーメントの理論値を用いた場合は以下の式になる。

$$S_T^2 = \frac{s_e^2}{n} \left\{ 1 + 1.1396(\varepsilon_p - 0.5772) \frac{\sqrt{6}}{\pi} + 1.1(\varepsilon_p - 0.5772)^2 \frac{6}{\pi^2} \right\}$$

s_e^2 : 変量の残差分散 n : データ数 ε_p : 標準変量 (8)

2) リサンプリング法の内ジャックナイフ誤差分散は以下の式である。

$$s^2 = (N-1) \sum [\hat{\theta}(i) - \hat{\theta}(\cdot)]^2 / N \quad (9)$$

s^2 : 推定誤差分散 $\hat{\theta}$: 推定値 $\hat{\theta}(\cdot)$: $\hat{\theta}(i)$ の平均値

$\hat{\theta}(i)$: (N-1) 個のデータによる推定値

3) 理論式とジャックナイフ法の誤差分散の比較

予測誤差の算出にジャックナイフ誤差分散を使う場合、理論値でどの程度違いが生ずるか確認する。図3はグンベル分布について赤線が 8) 式の理論分散とジャックナイフ誤差分散の比、青線がデータから求めた理論分散とジャックナイフ誤差分散の比である。図-4 は一般化極値分布について両者の比をとったものである。

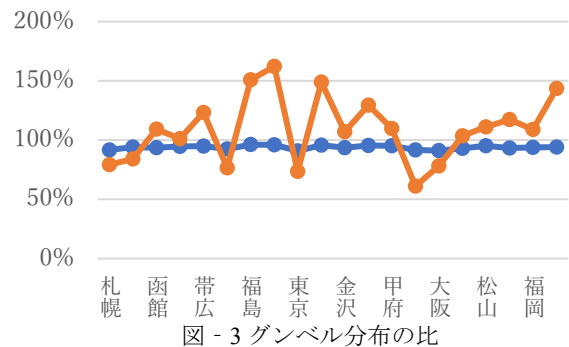


図-3 グンベル分布の比

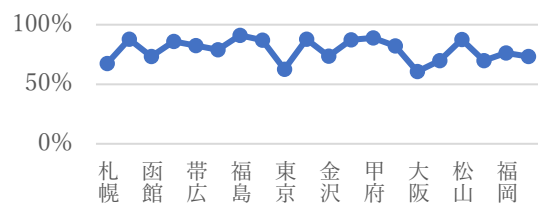


図-4 一般化極値分布の比