N 年最大統計量の適合性を考慮した分布関数の採択基準の提案

A NEW CRITERION FOR GOODNESS-OF-FIT USING MEAN VALUE OF N -YEAR MAXIMUM DATA ESTIMATED FROM RESAMPLING DATA

澁谷仁史¹・泉宮尊司²
Hitosi SHIBUYA, Takashi IZUMIYA

¹非会員 修(工) 山形県村山総合支庁 (〒991-8501 山形県寒河江市大字西根字石川西355)

²正会員 工博 新潟大学 工学部建設学科 (〒950-2181 新潟県新潟市西区五十嵐二の町8050)

The three candidate distributions which fits to the sample are selected through the goodness-of-fit criterion proposed by our previous study. In order to select the best fitted distribution, the mean of N -year maximum value are estimated using the resampling data of the sample. The number of resampling is 10,000 to estimate the mean value and its standard deviation of N -year maximum value. The mean values of N -year maximum value with the Monte Carlo simulation are compared with the results evaluated from the analytical expressions. The difference of mean values of the three candidate distributions is so distinct to select the best fitted distribution. The N -year maximum values extracted from the true distribution function show fairly good agreement with those computed from the analytical expression of the true distribution.

Key Words : N -year maximum value, goodness-of-fit, resampling method, criterion for best fitting

1. 緒言

ある限られた標本数の極値データから確率波高を精度よく推定するためには、精度の高い適合度評価基準および採択基準が必要である。これまで、相関係数を用いる方法(合田^{1),2)}や基準化最小2乗平均誤差適合基準(宝・高棹³⁾)などの適合度基準が提案されてきたが、それらの推定精度を比較検証した研究は数少なかった。泉宮・澁谷⁴⁾は、順序統計量の分散を考慮したIC基準を提案し、従来より用いられている相関係数を用いる方法およびSLSC基準よりも、Weibull分布の形状母数 k の値が1.4以上を除いて、精度が高いことを示した。しかしながら、候補分布関数を9種類以上とすると、分布によっては正解率が50%以下となる場合が存在し、正しい分布関数が得られないという欠点を有していた。

そこで本研究では、IC基準等の適合度基準を用いた評価で、上位3位までに入った分布関数の中に、真の分布関数が高い確率で含まれることを考慮して、それらの分布の N 年最大値の平均値をresampling手法により算定し、推定された母数から算定される理論値と比較することにより、分布関数を最終的に採択する手法を提案している。この手法によると、母分布関数を誤っても N 年最大統計量が小さくなる分布を避けることができるという利点を有しているのが特徴である。

2. 適合度評価法とその問題点

極値統計解析における適合度評価法として、相関係数を用いる方法、最大対数尤度を用いるAIC基準⁵⁾、および基準化最小2乗平均誤差適合基準(SLSC基準)³⁾を用いる方法などがある。最大対数尤度を用いるAIC基準は、母数推定法として最尤法を用いる場合には便利であるが、分布関数によっては解が存在しなかったり、標本数が少ない場合には負のバイアスが生じたりする欠点がある⁶⁾。相関係数を用いる方法は、最も計算が簡単で分かり易い方法ではあるが、分布関数によってはデータの大きい値の分散値が大きいために、相対的に相関係数が低下する傾向にあり、そのような分布関数は採用され難いという欠点を有している。

泉宮・澁谷⁴⁾は、AIC基準に準じて比較的簡単に計算できるように順序統計量の分散を考慮したIC基準を提案した。図-1および図-2に示すように、IC基準は相関係数を用いた方法およびSLSC基準よりも真の母分布関数の選択率が高い。しかしながら、分布関数によっては、正解率が50%以下となる場合があり、どの適合度評価基準を用いても誤り確率が高くなるという大変重大な問題が存在している。したがって、これまでのように適合度基準のみでは不十

分であり、さらに別の基準を用いて採択すべき分布関数を選別しなければならない。

そこで、IC基準を用いて適合度の順位を定め、上位3位までの分布関数の中に真の分布関数が存在するとすれば、その誤り確率は20%程度以下に収まることを考慮して、次節に示すような新たな採択基準を用いることを提案する。しかしながら、地球温暖化に伴う台風の強大化の影響は、非定常性を考慮した別の研究課題であるため、本論文では対象外としている。

図-1 適合度評価基準の比較 (FT-I 型分布 $N_D=100$)

図-2 適合度評価基準の比較 (Weibull 分布 $[k=0.75]$, $N_D=100$)

3. N 年最大値の平均値を用いた採択基準

(1) N 年最大値の平均値、最頻値および標準偏差

一般に、極値データ解析では、候補分布関数を何種類か用意し、それらの分布関数を当てはめて、それぞれの母数を推定する⁷⁾。それぞれの分布関数の母数が定まれば、 N 年最大値の分布関数も定まり、平均値および最頻値も理論的に算定することができ

る。母分布関数が極値I型および極値II型分布の場合、 N 年最大値の分布関数も同じ分布関数となり、 N 年最大値の平均値、分散および最頻値も容易に求められる(白石・上田⁸⁾、合田²⁾)。

一方、母分布関数がWeibull分布の場合には、このような性質がないので、泉宮⁹⁾による漸近理論による近似式によって算定することができる。以下に、FT-I型分布、GEV分布およびWeibull分布の N 年最大値の平均値、最頻値および分散を示す。

(1) 極値I型分布の N 年最大統計量

$$\left. \begin{aligned} \text{平均値: } \bar{x}_N &= B + A(\ln N + \gamma), \\ \text{最頻値: } (x_N)_{\text{mode}} &= B + A \ln N \\ \text{標準偏差: } \sigma(x_N) &= \pi A / \sqrt{6} \end{aligned} \right\} \quad (1)$$

(2) 一般化極値分布GEVの N 年最大統計量

$$\left. \begin{aligned} \text{平均値: } \bar{x}_N &= \frac{1}{\xi} A N^\xi [\Gamma(1-\xi) - 1] + \frac{1}{\xi} A (N^\xi - 1) B \\ \text{最頻値: } (x_N)_{\text{mode}} &= \frac{1}{\xi} A N^\xi \left[\left(\frac{1}{1+\xi} \right)^\xi - 1 \right] + \frac{1}{\xi} A (N^\xi - 1) B \\ \text{標準偏差: } \sigma(x_N) &= \frac{1}{\xi} A N^\xi [\Gamma(1-2\xi) - \Gamma^2(1-\xi)]^{1/2} \end{aligned} \right\} \quad (2)$$

(3) Weibull分布の N 年最大統計量

$$\left. \begin{aligned} \text{平均値: } \bar{x}_N &= A \left\{ (\ln N)^{1/k} + \frac{\gamma}{k (\ln N)^{1-1/k}} \right\} + B \\ \text{最頻値: } (x_N)_{\text{mode}} &= (\ln N)^{1/k} A + B \\ \text{標準偏差: } \sigma(x_N) &= \frac{\pi}{\sqrt{6}} \frac{A}{k (\ln N)^{1-1/k}} \end{aligned} \right\} \quad (3)$$

候補分布関数が真の分布関数に一致すれば、上式に示すような N 年最大値の平均値および最頻値も同じような変化を示すと予想される。

(2) 数値シミュレーションによる比較

ある分布関数に従う極値データを $N_D=100$ 個作成し、9種類の候補分布関数に当てはめて、重み付き最小二乗法で母数を推定した。その推定された母数を用いて、そのそれぞれの分布関数に従う極値データ数 N_D を10, 20, 30, 50, 70, および100個作成し、最大値を算定した。 N 年最大値の平均値および最頻値がどのくらいの標本数で安定した値が得られるかを検討した結果、10000ケース程度あればほぼ収束した結果が得られたので、その値に固定して数値シミュレーションを行った。

図-3は、真の分布関数がFT-I型分布の N 年最大値の平均値(黒丸印)の結果を示したものである。白丸印は、その標準偏差を示している。標準偏差実線で示すFT-I型分布の理論値と良く一致しているのが分かる。一方、IC基準で適合度順位が2位および3位で

図-3 推定母数を用いた N 年最大値の平均値の比較
(真の分布関数：FT-I 型分布)

図-6 推定母数を用いた N 年最大値の平均値の比較
(真の分布関数：Weibull 分布, $k=1.0$)

図-4 推定母数を用いた N 年最大値の平均値の比較
(真の分布関数：GEV 分布, $\xi=0.1$)

図-7 推定母数を用いた N 年最大値の平均値の比較
(真の分布関数：Weibull 分布, $k=1.4$)

図-5 推定母数を用いた N 年最大値の平均値の比較
(真の分布関数：GEV 分布, $\xi=0.3$)

図-8 推定母数を用いた N 年最大値の平均値の比較
(真の分布関数：Weibull 分布, $k=2.0$)

あったGEV分布($\xi=0.1$)およびWeibull分布($k=2.0$)は、標準偏差内に入っているものの、FT-I型分布とは異なる結果となっており、前者は N 年最大値の平均値が他の分布よりも大きく、逆に後者は小さくなっている。このように分布関数が似ているようでも、 N 年最大値の平均値等で比較することにより、より差異が明瞭に現れるので分布関数を採用する差異には有効な判断資料となり得る。

図-4は、真の分布関数がGEV分布($\xi=0.1$)の N 年最大値の平均値(黒丸印)とその標準偏差(白丸印)を示したものである。一点鎖線で示されたGEV分布($\xi=0.1$)の N 年最大値の平均値の理論値とほぼ完全に一致し、当然の結果ながら理論と数値シミュレーションの妥当性が確認された。またこの図より、GEV分布($\xi=0.1$)の極値データにGEV分布($\xi=0.2$)の分布を当てはめた場合には、 N 年最大値の平均値が大きめに評価されること、およびFT-I型分布を当てはめた場合には小さめに評価されることが分かる。

図-5は、同じく真の分布関数がGEV分布($\xi=0.3$)の N 年最大値の平均値(黒丸印)とその標準偏差(白丸印)を示したものである。この場合には、IC基準を用いた適合度基準により、FT-I型分布やWeibull分布は上位3位に入らなかったため、それらの分布は示されていない。一点鎖線で示されるGEV分布($\xi=0.3$)の理論値とシミュレーション結果との一致度は極めて高いことが分かる。

真の分布関数がWeibull分布の場合の結果を、図-6から図-8に示している。この分布の場合には、 N 年最大値の平均値の理論値は、厳密なものではなく漸近近似式であるが、黒丸印で示された数値シミュレーション結果と非常に良く一致している。ここには示されていないが、形状母数が $k=0.75$ のときも同様なことが確認されている。

図-6に示すWeibull分布($k=1.0$)のデータの場合には、GEV分布の $\xi=0.3$ および $\xi=0.4$ の分布に近いと判断されるが、 N 年最大値で比較すると大きな違いがあることが分かる。一方、Weibull分布($k=1.4$)のデータの場合には、GEV分布の $\xi=0.1$ および $\xi=0.2$ の分布に近いと判断され、しかも N 年最大値の変化の仕方もかなり類似している。このような場合、誤った分布であるGEV分布($\xi=0.1$)を採用しても、 N 年最大値の平均値が殆ど差異がないので、大きな問題とは思わないと思われる。

Weibull分布($k=2.0$)の場合には、FT-I型分布およびGEV分布($\xi=0.1$)が適合度判定で3位の中に入ってくるが、 N 年最大値の平均値にはやや有意な差異が見られるので、分布関数を選択する際には有効な判断指標となると考えられる。

(3) resampling dataを用いた N 年最大値の評価

実際に観測された波高などの極値データは、1つの標本であるため、その標本データから N 年最大値およびその平均値などを算定しなくてはならない。例えば50年間の極値の観測値があるとすれば、50年

最大値はその標本の最大値であるが、この値は確率変数そのものであるため、降順順位1位の確率分布に従って変動する。しかしながら、10年、20年、30年および40年最大値に関しては、resamplingを行えば、それぞれ ${}_{50}C_{10}$ 、 ${}_{50}C_{20}$ 、 ${}_{50}C_{30}$ および ${}_{50}C_{40}$ 通りの標本が得られるので、 N 年最大値も数多く得られる。因みに、 ${}_{50}C_{20}$ は1兆以上であり、 ${}_{50}C_{48}={}_{50}C_2$ でも1225通りにもなる。数値シミュレーションにおいては、ある極値分布のデータを100個作成し、その標本から10年、20年、30年、50年および70年分のデータを10,000回resamplingして、 N 年最大値の平均値を算定している。このような手法によることにより、観測された極値データだけを用いて、真の分布関数が不明であっても N 年最大値の平均値が評価できる利点がある。

図-9は、resampling手法を用いた方法により求めたもので、真の分布関数がFT-I型分布の N 年最大値の平均値(黒丸印)とその標準偏差(白丸印)を示したものである。この図の実線は、FT-I型分布を当てはめて推定した母数を用いて、式(1)に基づいて計算した結果である。同じく一点鎖線はGEV分布を、破線はWeibull分布を当てはめて計算したものである。この図に見られるように、FT-I型分布の理論値と最も一致度が高く、真の分布と一致していることが分かる。GEV分布の $\xi=0.1$ は、 N 年最大値の平均値はやや大きいですが、FT-I型分布の標準偏差内にはほぼ入っているため許容することも可能である。しかしながら、GEV分布の $\xi=0.2$ は N 年最大値の平均値が大きく外れているので、除外することができる。

図-10は、真の分布関数がGEV分布($\xi=0.1$)の結果を示したものである。この場合には、GEV分布の $\xi=0.1$ および $\xi=0.2$ の分布が標準偏差内に入っているものでいずれの分布を選択してもよいが、真の分布である $\xi=0.1$ を採用する際には、 N 年最大値の平均値がやや小さくなることを念頭に入れておく必要がある。FT-I型分布については、 N 年最大値の平均値が3つの分布の中で最も小さく、標準偏差からも外れかかっているため棄却してもよいと判断される。

真の分布関数がGEV分布($\xi=0.3$)の場合においても、図-11に示すようにほぼ同様なことが言え、 N 年最大値の平均値(黒丸印)に最も近い $\xi=0.3$ を採用することにより、外力を過小評価することなく分布関数を選択できる。また、この図に示されているように、GEV分布($\xi=0.3$)の場合には N 年最大値の平均値の標準偏差がかなり大きな値となることに注意しなければならない。このことは、信頼性設計を行う際に構造物の破壊確率等に直接関係してくるためである。

図-12から図-14は、真の分布関数がWeibull分布で、形状母数がそれぞれ $k=1.0$ 、 $k=1.4$ および $k=2.0$ の結果を示したものである。Weibull分布の場合には、 N 年最大値の平均値は漸近近似式であるにも拘らず、resampling結果との一致度はかなり高いと言える。

以上のように、適合度評価で上位3位までの分布

図-9 resampling data を用いた N 年最大値の平均値の比較 (真の分布関数 : FT-I 型分布)

図-12 resampling data を用いた N 年最大値の平均値の比較 (真の分布関数 : Weibull 分布, $k=1.0$)

図-10 resampling data を用いた N 年最大値の平均値の比較 (真の分布関数 : GEV 分布, $\xi=0.1$)

図-13 resampling data を用いた N 年最大値の平均値の比較 (真の分布関数 : Weibull 分布, $k=1.4$)

図-11 resampling data を用いた N 年最大値の平均値の比較 (真の分布関数 : GEV 分布, $\xi=0.3$)

図-14 resampling data を用いた N 年最大値の平均値の比較 (真の分布関数 : Weibull 分布, $k=2.0$)

関数を選択し、さらに実標本のresamplingにより N 年最大値の平均値を比較することで、真の分布関数をほぼ採択できる。もし分布関数を誤った場合においても、 N 年最大値を過小評価することなしに、外力を定量化できる利点があると言える。

4. 結論

本研究では、適合度基準であるIC基準を用いて、適合度の高い3つの分布関数を選択し、データのresampling手法により、 N 年最大値の平均値を算定し、それぞれの理論分布と比較することにより、より精度の高い分布関数を採択できる手法を開発した。その結果、以下の事柄が明らかとなった。

- (1) IC基準を含む適合度評価基準のみでは、真の分布関数を探し当てることは困難であり、適合度の上位3位までを選択することにより、誤り確率を20%程度以下に減少させることが可能である。
- (2) 極値I型分布、GEV分布およびWeibull分布に対して、それぞれ推定された母数を用いて、 N 年最大値の平均値を求め、その母数をもつ分布関数に従う極値データを発生させて算定した N 年最大値の平均値と比較したところ、極値I型分布およびGEV分布はもとより、近似式であるWeibull分布においても非常に良く一致していることが確認された。
- (3) 適合度が同程度であっても、 N 年最大値の平均値を比較すると、明確な違いが現れることが多く、分布関数の採択に当っては N 年最大値の平均値を指標として選定することにより、より精度よく採択することができ、外力を過小評価することもないという利点がある。
- (4) 1つの観測された極値資料より、resampling手法により、観測年数よりも数年以下の任意の N 年最大値の平均値が算定できることが示された。
- (5) N 年最大値の平均値の理論値が、resampling手法によるものよりもやや小さい分布関数は棄却し、それ以上の分布関数を採択することで、外力を過小評価する危険性を避ける利点がある。

参考文献

- 1) 合田良実 (1988) : 極値統計解析におけるプロットイング公式ならびに推定値の信頼区間に関する数値的検討, 港湾技術研究所報告, 第 27 巻, 第 1 号, pp. 31-92.
- 2) 合田良実(2008) : 耐波工学, 一港湾・海岸構造物の耐波設計一, 鹿島出版会, pp. 327 - 379.
- 3) 宝 馨, 高埴琢馬(1988) : 水文頻度解析における確率分布モデルの評価基準, 土木学会論文集, 第 393 号, II-9, pp. 151-160.
- 4) 泉宮尊司, 澁谷仁史(2009) : 順序統計量の分散を考慮した極値統計解析の適合度評価法に関する研究, 土木学会論文集B2 (海岸工学), 第 65 巻, pp.156-160.
- 5) Akaike, H.(1970): A new look at the statistical model identification, IEEE Transactions on Automatic Control, Vol. AC-19, No. 6, pp 716-723.
- 6) 泉宮尊司(2001) : 設計波の合理的な設定法, 2000 年度 (36 回) 水工学に関する夏期研修会講義集 B コース, pp.B-3-1-B-3-20.
- 7) 泉宮尊司, 齊藤雅弘 (1997) : 極値統計解析における順序統計量の分散を考慮した母数推定法, 海岸工学論文集, 第 44 巻, pp. 181-185.
- 8) 白石 悟, 上田 茂(1987): 港湾構造物及び海洋構造物の安全照査に関する検討, -作用荷重の変動係数と荷重係数に算定-, 港湾技術報告, 第 26 巻, 第 2 号, pp.493-576.
- 9) 泉宮尊司(1999): ワイブル分布に従う極値データの N 年最大統計量の定式化とその適用性, 海岸工学論文集, 第 46 巻, pp.236-240.