

流出負荷量の区間推定における bootstrap 法の適用について An application of bootstrap method for the interval estimation of diffuse loads

○ 多田 明夫, 田中丸治哉
○ Akio TADA and Haruya TANAKAMARU

1. はじめに

山林や農地などの面源・ディフューズ・ソースからの、ある期間における総流出負荷量の把握は、統合的な流域管理において大切である。特に山林からの流出負荷量は、流域内の生活排水や農業排水を希釈する効果があると考えられており、バックグラウンド負荷量として重要である。上記の観点から、著者らは山林からの溶存イオンの流出負荷量の区間推定法の確立に向けて研究を行っている。

流出負荷量の良好な区間推定を実現するためには、(1)適切なサンプリングが実施されること、(2)適切な負荷量の計算法が選択されること、(3)適切な区間推定の構成法が採用されること、の3点が必要である。著者らはこれまで(1)・(2)について主に検討を加えてきたが、本報告では、この(3)について、種々の bootstrap 法による区間推定について比較を行い、どのような方法が負荷量の区間推定として望ましいか検討した。また良好な区間推定法を確立するための課題の所在についても明らかにしたい。

2. 使用データと負荷量の計算方法

本報告では奈良県の山林流域(12.82ha)で2009年10月21日～2010年12月9日に観測された10分間隔の K^+ , Cl^- , Na^+ 濃度データと流量データを利用した。本報告では期間中の10分間単位の観測負荷量の合計値を負荷量の真値として扱う。サンプリング方法としては4日間隔の定期サンプリングを仮定した。負荷量計算法として、USGSの7パラメータモデル・(1)式をレーティングカーブとして利用し、これにバイアス修正法である smearing 法・(2)式を組み合わせた方法を採用した。

$$\ln l_i = a_0 + a_1 \ln q_i^* + a_2 (\ln q_i^*)^2 + a_3 T_i^* + a_4 (T_i^*)^2 + a_5 \sin 2\pi T_i^* + a_6 \cos 2\pi T_i^* + \varepsilon_i \quad (1)$$

ここで、 l_i は時刻 T_i の負荷量、 q_i は時刻 T_i での流量、 $a_0 \sim a_6$ は偏回帰係数である。*は共線性を除去す

るための変換を施していることを意味する。

smearing バイアス修正項(BCF)は、(1)式から求めた時々刻々の瞬間負荷量の積算値に乗数として用いられるバイアス修正量(係数)であり、次式で表される。

$$BCF = \frac{1}{n} \sum_{i=1}^n \exp(\varepsilon_i) \quad (2)$$

ここで、BCFは smearing バイアス修正係数、 ε_i は(1)式の残差、 n は観測されたデータ数である。この計算方法は容易でかつ応用が利き、結果も比較的良好である。同様の特性を有するノンパラメトリックなバイアス修正法には Composite 法があるが、今回は(2)式をバイアス修正項として採用した。

3. 信頼区間の構成方法～Bootstrap 法

負荷量の区間推定法としては、誤差分布に特定の確率分布を仮定しないノンパラメトリックな方法が望ましい。また複雑な推定量に対しても、適用が容易であることが望ましい。この条件を満たす信頼区間の構成方法にノンパラメトリックな bootstrap 法がある。

現在提案されている代表的なノンパラメトリックな bootstrap 法は、①基本的ブートストラップ法(推定量 θ の $\alpha\%$ 推定量が $2\hat{\theta} - \hat{\theta}_{B(1-\alpha)}$ 、 $\hat{\theta}$ は標本集団の推定値、*はブートストラップ複製の値、 B はブートストラップ複製標本数を表す)、②パーセンタイル法($\hat{\theta}_{Ba}^*$)、③ブートストラップ t 法($t = (\hat{\theta}^* - \hat{\theta})/\hat{\sigma}$ により区間推定)、④分散安定化ブートストラップ t 法(VS boot- t 法と略記、分散が1となるように変換を施すと、推定量は $h^{-1}(h(\hat{\theta}^*)) - (dh(\theta)/d\theta) \cdot \hat{\sigma} \cdot z_{1-\alpha}$ 、 h は分散安定化変換、 $z_{1-\alpha}$ は標準正規分布の $1-\alpha$ 分位点)、⑤BCa法(偏りの修正係数と偏りあたりの分散の変化率を考慮した方法)、⑥ABC法(解析的にBCa法と同様な区間推定を行う方法)、⑦マルチスケールブートストラップ法(サンプルサイズの増減と推定される確率の分散の関係を利用した不偏な確率推定法に基づくもの)等がある。これら手法の詳細は文献1)～3)を参照されたい。①・②は1次の精度($O(n^{-1/2})$ 、 n はサ

(所属) 神戸大学大学院農学研究科, Graduate School of Agricultural Science, Kobe University
(キーワード) 流出負荷量, 区間推定, ブートストラップ法, 精度, 溶存成分

ンプルサイズ), ③・④・⑤・⑥は2次の精度 ($O(n^{-1})$), ⑦は3次の精度 ($O(n^{-3/2})$)を有している。

4. 区間推定の検討・評価法

実際に3つの溶存態水質項目に対して, 前述の①, ②, ④~⑥の5つの方法について区間推定を実施した。

回帰モデルの bootstrap 法には, 説明変数 \mathbf{X} と目的変数 y の組 (\mathbf{X}_i, y_i) をリサンプリングする方法と, 回帰モデルの残差をリサンプリングする方法, すなわち $(\mathbf{X}_i, \hat{y}_i + \varepsilon^*_i)$ に基づいて bootstrap 複製標本集団を作成する方法がある。本報告では後者のリサンプリングに依った。10分間隔のデータに基づいて4日間隔の定期サンプリングを検討したので, 採水開始時刻の異なる独立なデータセットが $6 \times 24 \times 4 = 576$ 組得られる。各データセットは約80個の独立な観測データを有している。この約80個のデータから同じデータ数の80個のデータを復元抽出して, bootstrap 複製標本を2,000組作製し, そこから2,000組の流出負荷量や分散の bootstrap 推定量を得る。大雑把に言えば, ①~⑤までの区間推定は, こうして得られた bootstrap 推定量やジャックナイフ推定量に基づいて決定されている。

区間推定の良否は, 576組の区間推定結果のうち真値をその中に含んだ割合 (カバー率, 被覆確率) により判断した。

5. 結果と考察

表-1に各項目ごとの区間推定結果 (カバー率) を示した。表より, (1)式と smearing バイアス修正項の組み合わせでは, 期待した95%のカバー率を得られていないことがわかる。

1次の精度を有する①と②の方法に関しては, ②では項目間でのばらつきが大きく, 結果的に①の方が良好な区間推定結果となっている。2次の精度を有する④, ⑤, ⑥の方法では, ABC法の区間推定結果が最も低いカバー率となっている。原理が類似している⑤のBCa法と⑥のABC法での結果が異なるのは, BCa

法ではブートストラップ複製ごとにバイアス修正が実施されるが, ABC法では元の標本集団に対してのみバイアス修正が実行されるためであろう。

一般的に③の bootstrap-t 法はより広い信頼区間を与え, カバー率は大きくなる傾向にあるが, 計算労力のペナルティが大きい。この問題もあって④の分散安定化手法を採用したが, ⑤のBCa法よりも若干劣る結果となった。なお④の計算にあたり, 分散安定化関数にべき乗型の式を仮定したため, K^+ の計算に際していくつかのケースで計算が収束しない場合もあったが, その信頼区間が過小評価する傾向があることには変わりはなく, ⑤の結果の方が良好であった。

以上より, 結果的に①と⑤が同程度の信頼区間を与えているが, LQ式に基づく流出負荷量の区間推定には, より精度の高い⑤のBCa法を採用すべきことが確認できた。これまでの著者らの研究結果では, 十分な確からしさでの区間推定法の確立にまで未だ至っていないが, 今回の結果より, 検討すべき課題は区間推定法の選択でなく, よりバイアスの小さな負荷量算出法の選択とサンプリングにあることも確認できた。

最後に, ⑦のマルチスケールブートストラップ法について検討した結果, 計算労力がほかの方法と比較して大きなこと (計算時間が長いこと), およびスケール (サンプルサイズ) の変化に対して滑らかな bootstrap 確率の変化が得られないケースがあることがわかった。このため, 負荷量の区間推定にそのまま⑦の手法を適用することは現時点では困難であると考えている。

引用・参考文献

- 1) 北村源四郎・竹村彰通, 21世紀の統計科学III—数理統計の計算科学, 東京大学出版会, 2008,
- 2) 汪金芳・桜井裕仁, Rで学ぶデータサイエンス—ブートストラップ入門, 共立出版, 2011,
- 3) Efron, B. and Tibshirani, R. J., An Introduction to the Bootstrap, Chapman Hall/CRC, 1994

表-1 流出負荷量の95%区間推定結果 (%)
Table 1 Results of 95% interval estimation of diffuse loads (%)

Methods	K^+			Cl^-			Na^+		
	under	covered	over	under	covered	over	under	covered	over
basic	8.3	91.6	0.0	0.2	86.1	13.7	1.9	83.3	1.8
percentile	32.1	66.8	1.0	0.3	92.2	7.5	4.2	88.4	7.5
VS boot-t	14.4	81.8	1.7	2.3	84.5	12.8	3.5	82.8	13.5
BCa	8.2	91.8	0.0	0.2	86.6	13.2	1.7	82.8	15.5
ABC	20.5	75.7	3.8	2.3	80.0	17.7	5.7	71.7	22.6

*under ; true value > 95% confidence interval (C.I.), covered ; true value is included in 95% C.I., over ; true value < 95% C.I.