改良 IR 法による負荷量推定

Load estimation by improved importance resampling

○多田　明夫[*]・田中丸治哉[*]

○Akio TADA[*] and Haruya TANAKAMARU[*]

1. はじめに　1 年などの比較的長期間に集水域から河川を流下して下流の水体へと流入する汚濁物質などの総量である流出負荷量の不偏推定を実現するために、著者らはべき乗型 LQ 式による負荷量計算法（rating curve method，RCM）と重点的サンプリング（Importance sampling、IS 法）を組み合わせた推定法（RCM using IS 法）を開発した[1)]。IS 法では時刻 t での瞬間負荷量の期待値$\hat{l}(t)$の大きさに比例した確率（Probability proportional to size、PPS）での水質試料のサンプリングが必要であるが、この$\hat{l}(t)$は現地で事前調査をして LQ 式（rating curve）等を決定しないと得られない。しかし現実には、事前調査に続く負荷量の計算期間では事前調査とは異なる LQ 式を示し、推定量が偏ることもあり得る。この問題を解決するために、定期調査（採水）や流量比例サンプリングなど期間中に何らかの一貫したルールで収集された既存の水質標本集団から、事後的に PPS な標本集団をリサンプリング（再抽出）して概ね不偏な推定量を得る手法（Importance resampling 法，IR 法）も開発した[2)]。これまで筆者らはこの手法の改良を続けていたが、従来の IR 法では①元となる標本集団のサイズが小さいと推定に必要な最低個数（4 個）の標本を再抽出できない確率が高くなる、②95%信頼区間の被覆確率（信頼区間が真値を含む割合）が IS 法より低くなるという課題があった。ここではこの課題の改善法を提案する。

2. IS 法と SIR 法　2.1 IS 法　IS 法は効率的なモンテカルロ積分法で、これを期間［t_1、t_2］の瞬間負荷量 $l(t)$の積算値 L の推定に用いる場合、次式が基礎式となる。

$$L=\int_{t_1}^{t_2} l(t)dt=\int_{t_1}^{t_2}\frac{l(t)}{g(t)}g(t)dt=\hat{L}\int_0^1\frac{l(\eta)}{\hat{l}(\eta)}d\eta \quad , \quad g(t)=\hat{l}(t)\Big/\int_{t_1}^{t_2}\hat{l}(t)dt=\hat{l}(t)\Big/\hat{L} \qquad (1)$$

ここで$\hat{L}$は同じ期間での$\hat{l}(t)$の積分値で、$g(t)$は[0,1]の定義域の確率密度関数、$d\eta=g(t)dt$ である。$g(t)$に基づき逆関数法や棄却法でサンプリングをすれば、標本は$\hat{l}(t)$の PPS 標本となる。特に $l(t)/\hat{l}(t)$がほぼ一定値であれば、少ない標本数で(1)式の積分値の近似が可能となる。ここで重要なのは g の台が $l(t)$の台を含む（$\hat{l}(t)$の積分領域が $l(t)$の積分領域を含まむ）必要があることである。さもなければ、推定量が偏ることになる。なお筆者らは$\hat{l}(t)=\alpha q(t)^{\beta}$、$\alpha$・$\beta$は回帰係数、の利用を提案している。**2.2 (S) IR 法**　Sampling/Importance-resampling（SIR）法は、ある確率分布 p_0 に従う標本集団から別の確率分布 p_1 に従う標本集団を効率よく再抽出する手法であり[3)]、IR 法は SIR 法の別名でもある。例えば$\hat{l}_0(t)=\alpha_0 q(t)^{\beta_0}$を仮定して n 個の PPS 標本を得たが（事前分布）、得られた標本集団から求めた LQ 式が$\hat{l}_1(t)=\alpha_1 q(t)^{\beta_1}$だったとする。事後分布として$\hat{l}_1(t)$での PPS 標本集団を得るために、SIR 法では(2)式の重み $w(t)$に従って再抽出を行う。SIR 法でも n 個の標本が$\Sigma\hat{l}_1(t)$軸（定義域は［t_1、t_2］）上をカバ

（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）Sampling/imprtance-resampling，面源，負荷量，不偏推定，区間推定

$$w(t)=\left[p_1(t)\Big/p_0(t)\right]\Big/\sum^{n}\left[p_1(t)\Big/p_0(t)\right]\propto\left[q(t)^{\beta_1}\Big/\sum q(t)^{\beta_1}\right]\Big/\left[q(t)^{\beta_0}\Big/\sum q(t)^{\beta_0}\right]\propto q(t)^{\beta_1-\beta_0} \quad (2)$$

ーしているかどうかが重要である。この台のカバーの問題は、(1)式の積分軸η（$d\eta = g(t)dt$、すなわち[0,1］に基準化された定義域［t_1、t_2］での$\Sigma\hat{l}_1(t)$軸）の全領域を所定の標本数 n_{pps} で一様にカバーできているかという問題である。つまり SIR 法を有限個の事前標本集団に適用するためには、事前標本集団が既に$\hat{l}_1(t)$の PPS 標本集団であることが要求される。実際にはη軸上に（ランダムという意味では無く）無規則に分布する有限個のデータから、η軸上でランダムに一様分布するようなデータのみを再抽出することは困難である。この代替策として、適合度検定量である Anderson-Daring 検定量（AD 検定量、A^2）[4)]を利用して、η軸上で概ね等間隔するようなデータを再抽出することは可能であり、これが従来の IR 法である。なお、A^2値が小さな数列ほど、ランダム・一様分布では無く、より等間隔分布に近い分布となる。IR 法はη軸上での等間隔分布により PPS 標本の再抽出を行うので、再抽出の基準となる A^2 値は小さい方が良いが（従来法では非超過確率 $p=0.05$ に相当する A^2 値が抽出判断基準)、小さな A^2 値は再抽出できる標本数を小さくし、結果的に最低限必要な標本数が確保されない確率を高める。また最低必要個数に近い標本数では IR 法での推定量の偏りが大きく被覆確率が低下する問題が生じるので、いたずらに p を小さくできない。

3. IR 法の改良　上記の問題を解決するため、(1)元となる標本集団からの標本のランダム復元抽出により複数の複製標本集団を作成し、(2)それぞれの複製標本集団から $p=0.005$ に相当する A^2 値で再抽出を行い、このうち標本数が最大のものを選択する、の 2 点の改良を行った（Fig.1)。この結果、計算労力は増大したが、従来の IR 法と比べて標本集団からの負荷量推定成功率を同等以上に維持しつつ、推定量の偏りを低下させ、信頼区間の被覆確率も改善できた。結果の詳細は発表時に報告する。また SIR 法の解釈によれば、これまでの標本集団毎に p 値を最適に決定するような IR 法の改良には限界があることもわかった。

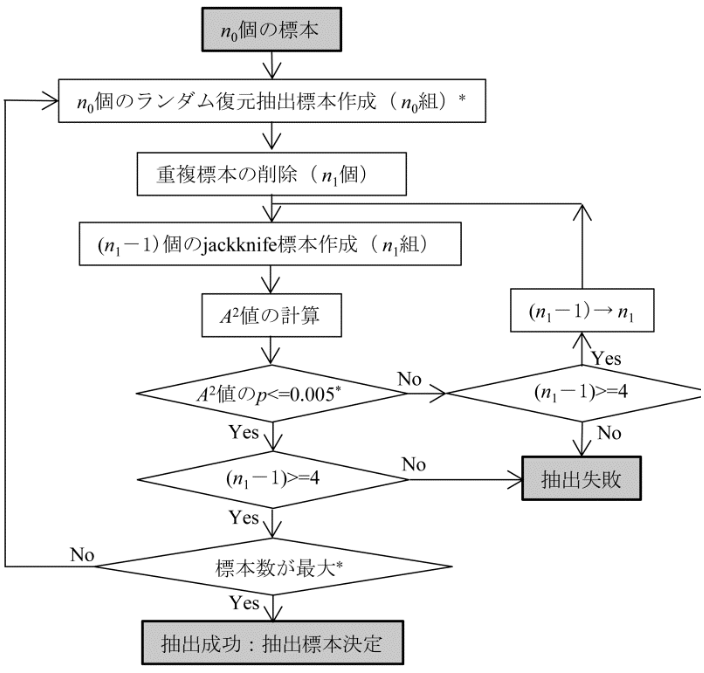

Fig.1 Improved IR method (*: modification/improvement)

参考・引用文献　1) 多田明夫, 田中丸治哉（2015)、流量比例コンポジットサンプルは正しい流量荷重平均濃度を与えるか？、平成 27 年度農業農村工学会大会講演会, 2) 多田明夫・田中丸治哉(2017), 流出負荷量の不偏推定法の一般化, 平成 29 年農業農村工学会大会講演会, 3) Tanner, M. A. (2006). *Tools for statistical Inference*, 3rd edition. New York, NY: Springer.、4)Stephens, M. A. (1986); *Goodness-of-fit techniques.* Series Statistics: Textbooks and Monographs, 68.