Horvitz-Thompson 推定量による河川負荷量の不偏推定
Unbiased load estimation by Horvitz-Thompson estimator

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

1.はじめに　年単位あるいはそれ以上の長期間を対象とした河川負荷量の正確な推定量は，農業や工業，あるいは生活排水などの人間活動が水環境に与える影響を評価するための基礎数値である。この量を比較的高頻度の河川流量と月 1 回程度の低頻度の水質観測値からどのように推定すべきか,著者らは長年検討を加えてきた。正確な推定を実現するために，特に負荷量の不偏推定法についてその区間推定量とともに検討を続けた。長年の検討の結果, 合計量の不偏推定ではポピュラーな方法である Horvitz-Thompson（HT）推定量（1952）が，様々なサンプリング法で得られた水質データに対し普遍的に不偏推定量を与えることが確認された。本発表ではその考え方と実際のデータへの適用課題について紹介する。

2.方法　河川負荷量の計算期間において，n 個の瞬間負荷量 l_i（水質試料採取時の濃度×流量）のデータ（i＝1～n）が確率 p で得られているものとする。このサンプリング確率 p は実際に標本を収集していない時間も含た母集団の N 個の単位時間すべてに対して定義され，j 番目の要素（単位時間）に対しては p_j と表記する（j＝1～N）。このとき，期間中の総流出負荷量の HT 推定量 L_{HT} は次式で与えられる。

$$L_{HT}=\sum_{i=1}^{n}\frac{l_i}{\pi_i}=\frac{1}{n}\sum_{i=1}^{n}\frac{l_i}{p_i}\quad\left(\sum_{j=1}^{N}\pi_j=n\quad or\quad p_j=\frac{\pi_j}{n}\right)\tag{1}$$

ここでπ_jは, 母集団の要素 j が n 個の標本中に含まれる割合の期待値で, 包含確率（inclusion probability）と呼ばれる。HT 推定量は非復元非均等確率抽出（unequal probability sampling without replacement，UPSWOR）データに対して用いられる。一般に水質モニタリング手法は同一標本を複数回抽出しないので UPSWOR である。HT 推定では, p_j が要素の大きさ（瞬間負荷量の大きさ l_j）に比例する時に，推定量の分散が最小となる。また HT 推定量はどのようなサンプリング法に対しても不偏である。つまり，水質試料のサンプリング方法は流出負荷推定量のばらつき（精度）にのみ影響を与え，不偏性には影響しない。

この分散最小の条件（要素の大きさに比例したサンプリング確率，sampling probability proportional to size，PPS）を復元非均等確率抽出（UPS with replacement，UPSWR）と組み合わせた推定量は Hansen-Hurwitz（HH）推定量（1943）と呼ばれ，Thomas（1985）により河川流出負荷量の推定に用いられたが, 現地モニタリングに採用されることはなかった。多田・田中丸（2021）による推定法である RCM using IS 法は PPS を UPSWOR と組み合わせた方法である。これらの方法と比べて，HT 推定量はサンプリング方法を選ばないより汎用的な不偏推定法であるが，推定量の精度はサンプリング確率に依存する。ここで重要なのは，HT 推定量の不偏性はサンプリング確率のみで決まることである。実際に米国の

（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）河川，流出負荷量，不偏推定，Horvitz-Thompson 推定量

複数の流域で得られた日単位の流量データ（USGS，2016）と水質データ（Heidelberg University, 2019）を用いて，25 の流域・年のデータセットに対して，月 1 回の定期調査に相当する（月内の採水日をランダムに選ぶ）方法で収集したデータに基づき，SS の年負荷量を推定した。信頼区間の構成は bootstrap-*t* 法に従った。推定量の分布は，標本を母集団から Monte Carlo 法でリサンプリング（繰り返し回数 20,000 回）して計算した。

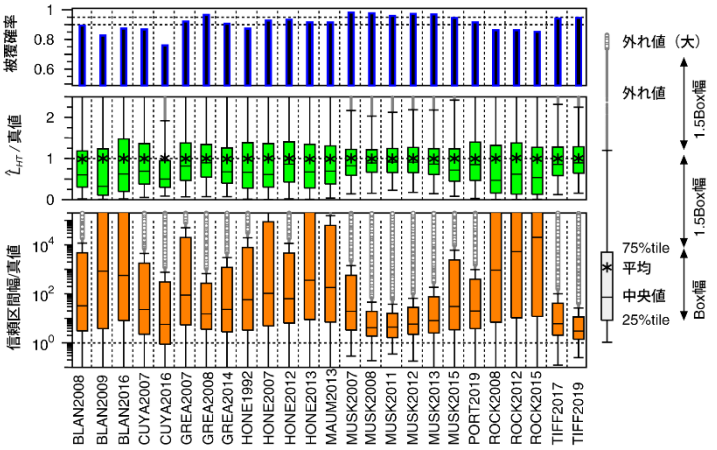

Fig.1　年単位 SS 負荷量の推定結果

3.結果と考察　Fig.1 に SS 年負荷量の推定量と中央 95%信頼区間の被覆確率を示した。この図からも明らかなように，HT 推定量はすべての流域に対して不偏推定量を与えている。一方で被覆確率は信頼水準の 95%をいくつかの流域で下回っている。推定量の偏りと異なり，区間推定は推定量の分布に基づくため，元々の母集団に極端に突出した大きな値が含まれていると被覆確率は低下する。実際に，被覆確率が低下しているデータの多くで，そのような極端に大きな値を持つデータが含まれていることが確認されている。

4.おわりに（HT 推定量の課題）　Fig.1 の例に顕著に示されるように，HT 推定量は水質項目の瞬間負荷量分布の分散が大きな時，サンプリング確率が PPS から離れるほど，（非常に）大きな不確かさを与える（unbiased but imprecise）。これを改善するためには，PPS に近づけるようサンプリング法の改善を行うか標本数を増やすしかない。一方，定期調査データに基づく場合，従来の LQ 式法などによる推定量は偏りを持つものの HT 推定量よりも精度が高いことが多い（biased but precise）が，偏りを制御できない推定法は誤った判断の原因になり，推奨されない。HT 推定量のもう一つの長所は既に収集されている過去のデータに基づいて負荷量の不偏推定量と信頼区間を構成できる点にある。しかし我が国のように年 12 個程度の低頻度調査データに基づく場合，HT 推定量は個々の年の河川負荷量に対しては非常に不確かな値しか与えないので，例えば 10 年分の 120 個のデータを用いて不確かさを減少させ，長期間の平均的な負荷量推定を行うのが現実的であろう。

参考・引用文献　Hansen, M., & Hurwitz, W. (1943). On the theory of sampling from finite populations. The Annals of Mathematical Statistics, 14(4), 333-362. Retrieved 18 March 2020, from https://www.jstor.org/stable/2235923

Heidelberg University. (2019). Heidelberg University National Center for Water Quality Research Tributary data download [data file]. Retrieved from https://ncwqr.org/monitoring/data

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 663-685. doi:10.1080/01621459.1952.10483446

Tada, A., & Tanakamaru, H. (2021). Unbiased estimates and confidence intervals for riverine loads. Water Resources Research, 57(3). doi:10.1029/2020wr028170

Thomas, R. B. (1985). Estimating total suspended sediment yield with probability sampling. Water Resources Research, 21(9), 1381-1388. doi:10.1029/WR021i009p01381

U.S. Geological Survey. (2016). National water information system data available on the World Wide Web (USGS water data for the nation) [Data file]. Retrieved from http://waterdata.usgs.gov/nwis/, https://doi.org/10.5066/F7P55KJN