ランダムフォレストによる生息場モデリングにおける初期値依存性の影響
Effects of initial conditions on habitat modeling using Random Forests

○福田信二[*]・山口真理恵[**]・鬼倉徳雄[***]・中島 淳[****]・平松和昭[*]・原田昌佳[*]
FUKUDA Shinji[*], YAMAGUCHI Marie[**], ONIKURA Norio[***], NAKAJIMA Jun[****],
HIRMATSU Kazuaki, and HARADA Masayoshi[*]

1．はじめに

一般に，生物の生息環境を評価する際には，対象種の空間分布と環境条件の関係性解析が行われる．近年，その生息場モデリングに種々の機械学習が利用されるようになってきており，その一手法であるランダムフォレスト（RF: Breiman, 2001）が注目されている．RF の利点として，他の機械学習と比較して精度が高いこと，説明変数が多く，説明変数間に相関がある場合でも効果的に動作すること，目的変数に対する説明変数の重要度を評価できること，多クラス分類が可能であることが挙げられる（Cutler *et al.*, 2007）．このような利点がある反面，RF は学習の際に乱数を使用するため，構築モデルによる評価結果には初期値依存的なばらつきが生じると考えられるが，その影響について定量的に評価した事例はない．そこで本研究では，北部九州における在来タナゴ類の生息分布データを用いて，RF の初期値依存性が生息場モデルとその評価結果に及ぼす影響について評価する．

2．生息分布データ

使用データは，九州北部の 84 水系，1,064 地点における調査結果であり（Onikura *et al.*, 2012），同調査から，ヤリタナゴ *Tanakia Lanceolata*（Tla），アブラボテ *T. limbata*（Tli），セボシタビラ *Acheilognathus tabira nakamurae*（Atn），カネヒラ *A. rhoombeus*（Ar），ニッポンバラタナゴ *Rhodeus ocellatus kurumeus*（Rok），カゼトゲタナゴ *R. atremius atremius*（Raa）の 6 種の在来タナゴ類の生息が確認されている．生息場モデリングに供する環境変数は，流路延長（LMR: km），標高（ELEV: m），勾配（SLP），河川幅（WID: m），河川・水路結合数（RCC: 点），水路網指数（CNI），水田面積（PF: km^2），宅地面積（RA: km^2）の 8 つである．この 8 変数により，リーチスケールから流域スケールに至る複数スケールを考慮した解析が可能である．各タナゴ類の出現パターンは，出現を 1，非出現を 0 と定義した．

3．生息場モデリング

RF は，Breiman（2001）によって提案された先進的手法であり，ブートストラップサンプリングと多数の分類回帰木を駆使することにより，高精度かつロバストなモデリングを実現している．本研究では，統計ソフト R（R Core Team, 2012）のパッケージ「randomForest」（Liaw and Wiener, 2002）を使用し，8 つの環境変数を入力値とし，各タナゴ類の出現パターンを出力する生息場モデルを構築した．その際，RF に関するモデルパラメータにはデフォルト値を使用した．変数の重要度は，Mean Decrease in Accuracy を基準に評価した．また，構築モデルの再現精度の評価には，正答率および AUC（Area Under the receiver operating characteristic Curve）を使用した．

本研究では，RF 解析における初期値依存性を評価するために，初期乱数の種を 1 から 100 まで変化させ，それぞれについてモデルを構築し，変数の重要度と再現精度を計算した．ここでは，RF モデルの初期値依存性を解析結果のばらつきによって定量的に評価した．

*九州大学大学院農学研究院／Faculty of Agriculture, Kyushu University　**長崎市役所／Nagasaki City Government Office　***九州大学水産実験所／Fishery Research Laboratory, Kyushu University　****福岡県保健環境研究所／Fukuoka Institute of Health and Environmental Sciences
キーワード：生息場モデル，変数の重要度，再現精度，機械学習，初期乱数，ばらつき

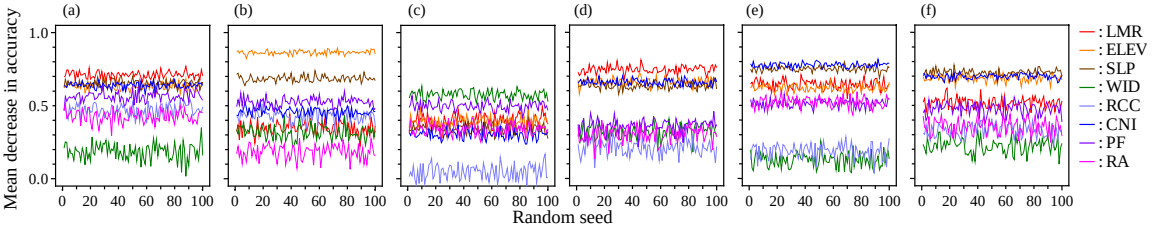

Fig. 1　Variable importance evaluated based on 100 different randomization sets: (a) Tla (b) Tli, (c) Atn, (d) Ar, (e) Rok and (f) Raa.

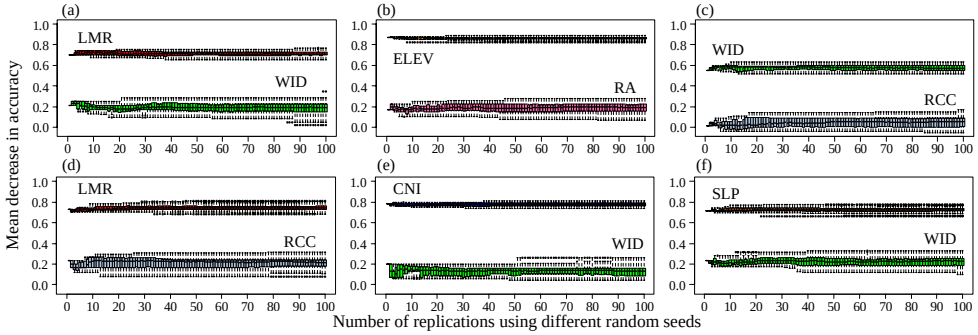

Fig. 2　Relationship between variable importance and the number of replicated randomization: (a) Tla (b) Tli, (c) Atn, (d) Ar, (e) Rok, and (f) Raa

4．結果と考察

構築モデルの再現精度は，全種ともほぼ 100%であり，RF モデルの高い有用性が示唆された．また，RF の初期値依存性の影響はほとんど見られなかった．乱数の種を 1 から 100 まで変化させて変数の重要度を評価した結果，**Fig. 1** のように，特に変数の重要度が低い場合にばらつきが大きくなり，場合によって順位が逆転するなど，RF の初期値依存性が評価結果に影響を及ぼしていることが明らかになった．なお，重要な変数は種ごとに異なっているが，概して，ELEV や SLP に加えて，流域規模を示す変数（LMR）や水田環境に関連する変数（CNI，PF 等）の重要性が高い．同様の知見は，タナゴ類の生態に関する既往の研究でも報告されており，RF による生息環境評価の妥当性が示唆された．乱数の種の変更回数と変数の重要度の評価結果のばらつきの関係を **Fig. 2** に示す（上段：重要度が最も高い変数，下段：重要度が最も低い変数）．変数の重要度の順位が高いほどばらつきが小さく，中央値が安定するまでに必要な変更回数が少ない．全体的には，乱数の種の変更回数が 20 回程度を超えると，変数の重要度の中央値がほぼ一定となっており，ニューラルネットワークでの事例と同様の結果になった．ただし，この変更回数は，データの質的要素（例えば，出現地点数の割合）から影響を受けるため，使用データに応じた検討が必要である．

5．おわりに

本研究により，先進的機械学習手法である RF の初期値依存性が構築モデルに及ぼす影響が定量的に示された．結果として，再現精度への影響は無視できる程度であったが，変数の重要度への影響は比較的大きいため，解析結果を解釈する際には，乱数の種を 20 回程度の変更し，RF の初期値依存性を考慮する必要があると考えられる．今後の課題として，生態水理学分野で広く使用されている生息場適性曲線の導出等が挙げられる．

謝辞　本研究の一部は，平成 24～26 年度日本学術振興会研究拠点形成事業（B.アジア・アフリカ学術基盤形成型）の支援を受けた．ここに記して感謝の意を表する．

参考文献　Breiman (2001): Random forests. Mach. Learn., 45, 5–32.　Cutler *et al.* (2007): Random forests for classification in ecology, Ecology, 88(11), 2783–2792.　Liaw and Wiener (2002): Classification and regression by random forest, R News, 2(3), 18–22.　Onikura *et al.* (2012): Predicting distributions of seven bitterling fishes in northern Kyushu, Japan, Ichthyol. Res., 59, 124–133.　R Core Team (2012): R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/