河川水辺の国勢調査を用いたアユ・カマツカの生息環境評価における外れ値除去の適用可能性

Applicability of Noise Removal for the Habitat Assessment of Ayu and Kamatsuka using the National Census of River Environment

○小川洸生・福田信二
Kosei OGAWA, Shinji FUKUDA

1. 背景　河川魚類の保全には，現在の分布域とその変動を把握することが重要である．国土交通省は，1990 年から「河川水辺の国勢調査」[1] (以下，水国)として，全国の 1 級河川で魚類や底生動物などの調査を行っており，オープンデータとして公開している．水国には生物の個体数だけでなく，流速や水深などの調査地点の物理環境も併せて記載されている．しかし，大量のデータが含まれており，不正確なデータに基づく解析のリスクが指摘されている[2]．そこで本研究では，Isolation Forest を用いた外れ値除去によるデータの前処理を施し，Random Forests を用いた生息環境評価への影響を検討した．

2. 方法　本研究では，1993 年～2021 年の全国の水国データを用いた．外れ値除去には Isolation Forest[3]（IF）を使用した．IF は，大量のデータの中から外れ値を検出するための機械学習アルゴリズムで，あるデータと他のデータの距離に基づいて外れ値を識別する．本研究では，流速（cm/s）と水深（cm）に基づいて外れ値を除去した．IF の精度を高めるため，明らかに外れ値と思われる値（流速 3,040 cm/s，水深 86,400 cm など）や空白のデータが含まれるデータは，解析前に除外した．全データに対する外れ値の割合は 10%から 90%まで 10%間隔で設定し，除去前と除去後のデータに Random Forests[4]（RF）を適用して種分布モデルを構築した．種分布モデルの特徴量は，流速・水深に加えて，河口からの距離（km）と調査年の 4 変数とし，河口からの距離は対象魚種の地理的分布，調査年は分布の変遷を考慮するために採用した．個体数データにも誤りと思われるデータが含まれているため，魚類が捕獲されている地点を「在」，その他を「不在」データとし，それぞれ 1／0 として処理した．これをモデルの出力値として，尤度を計算した．RF のハイパーパラメータは Optuna を用いて最適化し，30 種類の乱数シードを用いた．また，Nested-Cross Validation を用いてモデルの汎化性能を評価した．モデルの精度評価として，モデルの Accuracy, Precision, Recall, F1 スコアを計算し，応答曲線から，アユ・カマツカの生息環境を評価する．

3. 結果及び考察　IF の解析結果を図 1 に示す．10%の除去により多くのデータが除去されていることが確認できる．種分布モデルの精度を表 1 に示す．Accuracy は，両種とも外れ値除去後にやや減少した．また，アユは，Precision の減少と Recall の増加がみられたが，カマツカは，Precision の増加と Recall の減少がみられた．F1 スコアは，アユでは減少したが，カマツカでは増減を繰り返した．よって，種分布モデルの精度は，外れ値除去により全体的に低下しており，特にアユに対する識別性能では Precision の低下と Recall の増加により，F1 スコアも減少した．一方，カマツカは，Precision の向上が見られるものの，

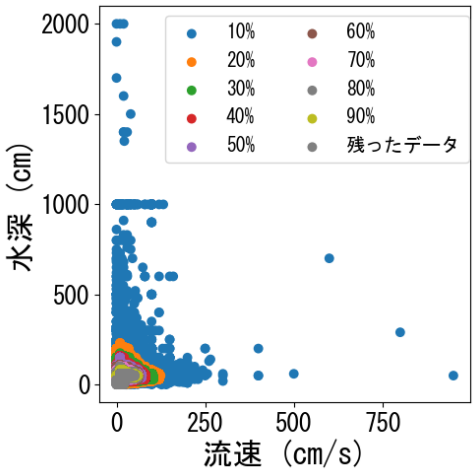

図 1 Isolation Forest による外れ値除去の結果．色の違いは除去割合を示す．

＊東京農工大学大学院農学府農学専攻食農情報工学コース
Environmental and Agricultural Engineering, Tokyo University of Agriculture and Technology
キーワード：生息環境評価，河川水辺の国勢調査，外れ値除去

一貫性に欠けるため，モデルの予測結果の一貫性の向上が課題であり，魚種の違いによる精度のばらつきが小さいモデルの構築が必要である．外れ値除去前後の各魚種の応答曲線を図2および図3示す．図2では，外れ値と思われる値の影響で，乱数ごとに応答曲線が大きく変動しており，水国を用いた生息環境評価には外れ値除去の必要性が確認された．図 3 上段のアユの応答曲線を見ると，流速が増加するほど尤度が増加している．明かに外れ値であると思われるデータは除去できたが，アユの最適流速は 100 cm/s から 140 cm/s とされており [5]，10%除去した際の流速の最大値は 124.1 cm/s であったため，過度な除去は避けるべきであることが示唆された．水深に関しては 30 cm 付近をピークに横ばいになっている．水深 100 cm を境に選好値が急減するとされている [6] が，本研究ではそのような傾向はみられなかった．河口からの距離と調査年は各除去割合とも同様の挙動を示しており，河口からの距離に関しては 200 km 付近で尤度が急増しており，陸封個体の影響が示唆された．調査年に関しては，1995 年付近で急減しており，生息地の減少が示唆された．図 3 下段のカマツカの流速に関する応答曲線をみると，増減を繰り返しているが，顕著な傾向はみられなかった．水深を見ると，水深が大きくなるにつれて尤度も増加している傾向が示された．カマツカは底生魚であるため，水深が大きい地点を選好するという妥当な結果が得られた．河口からの距離はアユと同じく 200 km 地点で急増しており，陸封個体の影響が示された．調査年をみると 2019 年から 2020 年にかけて急な増減を示しており，調査年による河川環境や採捕魚種の違いが影響している可能性が示唆された．

表 2．モデルの精度評価．
上段がアユ，下段がカマツカを示す．

	除去前	10%	20%	30%	40%	50%	60%	70%	80%	90%
Accuracy	0.681	0.681	0.680	0.677	0.683	0.674	0.667	0.673	0.663	0.668
Precision	0.489	0.494	0.491	0.488	0.494	0.490	0.485	0.494	0.495	0.495
Recall	0.827	0.823	0.827	0.828	0.832	0.830	0.820	0.827	0.836	0.800
F1 スコア	0.615	0.617	0.616	0.614	0.620	0.616	0.609	0.619	0.622	0.612

	除去前	10%	20%	30%	40%	50%	60%	70%	80%	90%
Accuracy	0.680	0.679	0.674	0.676	0.667	0.665	0.659	0.647	0.623	0.633
Precision	0.463	0.459	0.446	0.440	0.420	0.422	0.423	0.420	0.411	0.410
Recall	0.721	0.722	0.717	0.721	0.718	0.726	0.735	0.733	0.721	0.765
F1 スコア	0.564	0.561	0.550	0.546	0.530	0.533	0.537	0.534	0.524	0.534

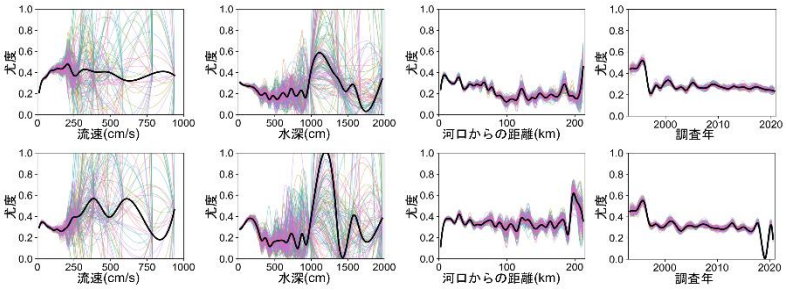

図 2 外れ値除去前のデータに基づくモデルの応答曲線．上段がアユ，下段がカマツカを示す．背景の線は，各乱数における計算結果，黒線は中央値を示す．

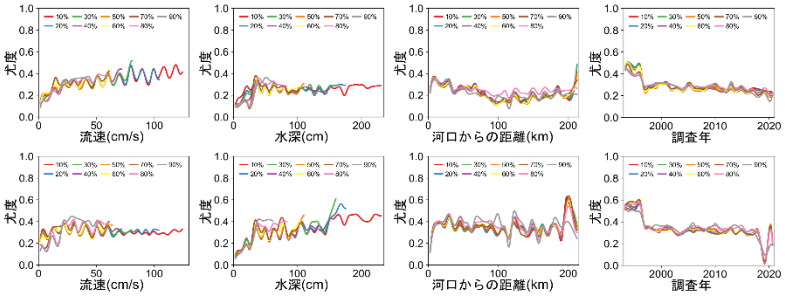

図 3 外れ値除去後のデータに基づくモデルの応答曲線．上段がアユ，下段がカマツカを示す．色の違いは除去割合の違いを示す．

4. 結論　本研究では，水国に基づくアユ・カマツカの生息環境評価における外れ値除去の適用可能性を検討した．結果として，外れ値除去の有効性が示され，両種の生息環境が評価できることが示された．一方で，種分布モデルの精度は両種とも低下したため，予測結果の一貫性を高め，異なる魚種においても精度のばらつきを抑えたモデルを構築する必要がある．また，除去率の増大に伴って特徴量の値域が限定されるため，予測結果の妥当性が低下した．今後は，「在」「不在」ではなく，個体数データを基にモデルを構築し，魚類の生息環境を評価する普遍的なモデルの構築を検討している．

《引用文献》(1)河川環境データベース：国土交通省　(2) 末吉ら(2016)：河川水辺の国勢調査を保全に活かす-データがもつ課題と研究例　保全生態学研究　21 巻　167-182　(3) F. Tony Liu ら(2008): Isolation Forest, ICDM, 978-0-7695-3502-9（4）L. Breiman（2001）: Random Forests, MACHINE Learning, Vol.32,5-32　（5）松原ら（2004）：魚のすみか数量化調査-Ⅲ（神流川における魚類生息場適正基準- Ⅰ）　群馬県水産試験場研究報告　第 10 号　15-20　（6）石川ら（1996）：河川における魚類生息環境評価（IFIM 適用）のための基礎調査　木更津工業高等専門学校紀要　第 29 号　23-32