

機械学習を用いた三重県におけるブリ類漁獲量の実用的予測モデルの構築
Development of Statistical Prediction Models for Yellowtail Catch in Mie
Prefecture for Operational Purposes Using Machine Learning

万田敦昌・小川翔大*・久野正博**・藤田弘一**・武田保幸***・御所豊穂***・
海野幸雄****・山田二久次†

(三重大学大学院・*元三重大学・**三重県水産研究所・
和歌山県水産試験場・*静岡県水産技術研究所)

Atsuyoshi MANDA, Shota OGAWA*, Masahiro KUNO**, Koichi FUJITA**,
Yasuyuki TAKEDA***, Toyoho GOSHO***, Yukio UNNO**** and Fukuji YAMADA†
(Mie University, Graduate school / *Former undergraduate student of Mie University
/ **Mie Prefecture Fisheries Research Institute / ***Wakayama Prefecture
Fisheries Research Institute / ****Shizuoka Prefecture Fisheries Research Institute)
E-mail : †fyamada@bio.mie-u.ac.jp

【要約】

実際の予報業務での使用を念頭におき、三重県のブリ類漁獲量を予測する実用的な統計モデルを構築した。1973年から2015年の気象、海洋、他魚種を含めた漁獲量等の多数のデータを用い、統計モデルとして、線形回帰、サポートベクター回帰、ランダムフォレスト回帰、ニューラルネットワーク、ディープラーニングの5種類を検討した。使用できるデータ期間が比較的短期間であることに主に起因して、ニューラルネットワークとディープラーニングはモデルのデータへの当てはまりが他のものよりも悪かった。サポートベクター回帰、ランダムフォレスト回帰はデータの当てはまりは他の手法よりも良かったが、漁獲最大年の漁獲量をうまく予測できなかった。線形回帰モデルは漁獲量の最大値、中央値、最小値を20%程度の相対誤差で推定しており、一番安定した予測結果が得られた。

【キーワード】

ブリ類、漁獲量、予測、機械学習、三重県

【abstract】

In this paper, we have developed statistical models to predict the catch of yellowtails in Mie prefecture, Japan for the operational purposes, using meteorological, oceanographical, and fish-catch data, including fish species other than yellowtail during the period from 1973 to 2015. Five statistical models are examined, namely linear regression (LR), support vector regression (SVR) and random forest regression (RFR),

neural network (NN), and deep learning (DL). NN and DL exhibit poor performance compared with other models, mainly because of relatively small number of samples. SVR and RFR show better predictive skills than NN and DL. Their skills degrade in the case that the models predict the catch in 2015 when a maximum catch was recorded during the period from 1973 to 2015. Throughout the tests conducted in this study, the LR model shows the best performance with an order of 20% relative error and the model is the most robust and stable.

1. はじめに

水産資源量は乱獲や海洋汚染等による漁場環境の悪化等の影響だけでなく、気象変動や海況変動等の自然環境の変化によって変動することが指摘されている。例えば、Beamish and Bouillon(1993) は北太平洋亜寒帯域のカイアシ類の現存量とアリューシャン低気圧の強さに正の相関があることを示し、サケの資源量の増加にアリューシャン低気圧の強化が関係していることを指摘した。また、川崎(1994)では、日本近海、カリフォルニア西岸、南米西岸を中心にそれぞれ分布する極東マイワシ、カリフォルニアマイワシ、チリマイワシの間に、位相の一致する長周期変動が存在し、グローバルな気象変動がマイワシの資源量に影響を与えることが指摘されている。

漁獲量の変動と環境要因の関連性に注目した研究、それらを基に漁獲量の予測を試みた研究はこれまで数多く報告されている。従来、環境要因に強く影響を受ける回遊魚等の漁獲は定性的に予測されてきたが、近年では重回帰モデル等の統計モデルを用いた定量的アプローチに変化してきている（例えば、為石ら(1997)、横田ら(1998)）。定性的な予測では予測者の勘や経験に予測結果が大きく依存することから、客観的な予報が求められている。また、湯ら(1992)では道東沖のマイワシの漁獲量予測にファジー推論が使用されているように、予測精度向上のために先端的統計モデルの利活用が進められている。

先端的統計モデルの一つに機械学習と呼ばれる手法があり、近年様々な分野で利用が進んでいる。機械学習は人間が持つ学習にあたる仕組みをコンピュータで実現する技術・手法の総称で、分類や回帰問題に主に用いられる。最近では、関数形を特定せずに非線形現象を表すことができ、複雑な動態に対して柔軟に対応し再現できるモデルに注目が集まっている。機械学習は情報、制御等をはじめとした多くの分野で利用されているが、水産分野でも同手法が活用されてきており、漁獲量の予測に活用された研究もいくつか報告されている。例えば、庄野ら(2014)は、1987年～2010年の24年間の鹿児島近隣県のクロマグロの漁獲量および種苗漁獲量、枕崎漁港のカツオの漁獲重量、フェリーの水温データ、黒潮離岸距離等を用いて、サポートベクター回帰による鹿児島県のヨコワ（クロマグロの幼魚）の漁獲量予測を行っている。また、馬場・松石(2015)は、1972年～2011年のサンマ

来遊量を「高位」、「中位」、「低位」で分類し、環境データにランダムフォレストを適用して前述の3分類のサンマ来遊量を判別、予測した。その結果、的中率が約62%となり、現況の定性的な予報とほぼ同じかやや高い結果が得られ、人間の勘や経験に基づかずに予測が可能であることを示した。

熊野灘沿岸に位置する三重県南部地域の主要漁獲物の一つにブリがある。三重県南部地域ではブリは主に春先に定置網で漁獲され、同地域に位置する尾鷲市では「市の魚」に指定されている。紀伊半島南部のブリの漁獲量は年によって大きく異なり、環境変動と漁獲量の大小は古くから注目されてきた。特に、黒潮流路の変動や黒潮系暖水の接近と漁獲量の関係が議論されてきた。例えば、阪本(1991)は和歌山県の海況と漁況の関係をモニタリング調査結果から整理し、黒潮が離岸して黒潮系暖水の波及がなく、且つ16°C以上の高温海況ではブリの入網がほとんどみられないこと、離岸年でも日本海低気圧通過の気象擾乱時には漁があることを指摘した。つまり、回遊性の魚種の漁獲量の大小は系群の資源量だけでなく、海域に回遊してくるかどうかも重要であり、ローカルな気象、海況変動とも関連していることが考えられる。

これらの背景から、本稿では統計モデルを用いた三重県のブリ類漁獲量の予測モデルの構築を目的とした。漁獲量の予報は漁業者説明会で公表されており、三重県の定置網の場合では毎年発表会が行われている。その際、予測者の勘や経験に予測結果が大きく依存しない客観的な評価は重要な情報となる。また、漁獲量を精度よく予測することにより網目選択等の漁具調整や取引価格への参考情報として活用できる可能性がある。予測する漁獲量をブリ類としたのは、三重県では主に定置網で漁獲されることから、漁獲努力量の影響が比較的小さく、それを考慮せずに予測できる可能性があることである。先行研究の結果から、三重県のブリ類漁獲量は気象変動や海況変動等の環境要因に大きく影響を受ける可能性が高い。よって、ブリ類の資源量や餌環境に関連するデータに加え、海洋、気象データも網羅的に収集、選択し、複数の統計モデルの結果を比較してモデルの予測性能について検討した。現場の予測業務を念頭においた場合、精度の高さだけでなく、「モデルの入力データが入手しやすいこと」の他に「特殊な自作ソフトウェア等を必要とせず、一般的に用いられているソフトウェアで計算可能であること」や「大型計算機を必要としない計算量」など、計算資源の制約に左右されないことが予測業務を円滑に実施する上で重要である。近年、気象・海洋とも再解析データ等の時空間的に密な大規模データが利用可能となったが、データ量が膨大であることに加え、バイナリ形式の特殊なファイルフォーマットで記述されていることが多く、計算機に関する専門的知識と技術、および気象・海洋学に特有の独自ソフトウェアが必要となる。以上の理由から、モデル構築の際に予測性能だけでなく、計算手法、データ量およびフォーマット、計算量も考慮した。次節では本研究で使用したデータおよび統計モデルについて述べる。

表 1 本研究で使用した全データ

データの種類	使用期間	N _{MAX}
月平均水温（伊東，稲取，雲見，下田，焼津，串本西，串本東，浜島）：月平均	前年と当年（1~3月まで）	15
上記 8 地点の冬季平均水温の長期平均からの偏差	当年 1~3 月	1
海面水温の長期変動傾向（関東東，関東南，四国・東海沖，沖縄東，東シナ海南部，日本全海域平均）：季節平均，年平均	前年（四季，年平均）と当年（冬季）	6
海面水温の長期変動傾向（北太平洋）：年平均	前年	1
黒潮流路（スプラインの補間あり）：月平均	前年と当年（1~3月まで）	15
黒潮流量：季節平均	前年（夏季・冬季）と当年（冬季）	3
月平均気温（静岡，網代，石廊崎，尾鷲，潮岬）：月平均	前年	12
AO(Arctic Oscillation:北極振動)：月平均	前年	12
NAO(North Atlantic Oscillation:北大西洋振動)：月平均	前年	12
PNA(Pacific North American:太平洋・北米パターン)：月平均	前年	12
WP(Western Pacific:西太平洋パターン)：月平均	前年	12
NPI(North Pacific Index:北太平洋指数)：年平均	前年	1
PDO(Pacific Decadal Oscillation:太平洋十年規模振動)：年平均	前年	1
SOI(Southern Oscillation Index:南方振動指数)：月平均	前年と当年（1~3月まで）	15
前年のブリ類漁獲量（三重県，和歌山県，静岡県）：年漁獲量	前年	1
前年の被食魚（イワシ類計，マイワシ，カタクチイワシ，ウルメイワシ，シラス，アジ類計，マアジ，ムロアジ類，サンマ，イカ類計）漁獲量（三重県，和歌山県，静岡県）：年漁獲量	前年	1

注：N_{MAX} は 1 つの予測値を算出するための最大データ数（説明変数として使用するデータ数）。

2. データ・分析方法

本研究で予測を試みる三重県のブリ類漁獲量は、農林水産省の漁業・養殖業生産統計年報を用いた⁽¹⁾。モデルの説明変数となるデータは、環境に関連する海洋、気象データと資源量や餌環境等に関連すると思われる生物的なデータを網羅的に収集した（表 1）。海洋環境に関するデータは、各種水温データと黒潮流軸位置、流量データを用いた。水温データは、静岡県、三重県、和歌山県の各水産試験場が行っている沿岸 8 点（伊東、稲取、雲見、下田、焼津、串本西、串本東、浜島）の定地水温データを月平均、冬季 3 か月平均にした

機械学習を用いた三重県におけるブリ類漁獲量の実用的予測モデルの構築

ものと気象庁による日本近海の6海域および日本全海域平均の季節平均、年平均海面水温偏差データ、北太平洋の年平均海面水温データの3種類のデータを使用した⁽²⁾。黒潮に関するデータは気象庁による月別黒潮流路データ⁽³⁾と季節別黒潮流量データ⁽⁴⁾を用い、黒潮流路の欠損値は馬場・松石(2015)と同様にスプライン補間で補った。気象データは上述の沿岸定地水温観測点周辺の5地点(静岡、網代、石廊崎、尾鷲、潮岬)におけるAMeDAS(Automated Meteorological Data Acquisition System)の月平均気温を使用した。さらに、馬場・松石(2015)と同様に、北半球、北太平洋及びエルニーニョ現象に関する大規模な気象、気候変動に関連する7つのインデックスも使用データとして加えた⁽⁵⁾。

生物的なデータでは、農林水産省の漁業・養殖業生産統計年報による前年のブリ類漁獲量とブリ類に被食されると思われる魚種の漁獲量データを用いた。また、周辺海域に生息するブリ類も次年の漁獲量に影響を及ぼすと考え、和歌山県及び静岡県における前年のブリ類漁獲量も使用データとして加えた。ブリは肉食性で、稚魚は動物プランクトンを主に捕食し、その後、全長27mm前後で魚食性が発現し、全長130mmで完全に魚食性に移行する(山本ら(2007))。そこで、本研究では、イワシ類計、マイワシ、カタクチイワシ、ウルメイワシ、シラス、アジ類計、マアジ、ムロアジ類、サンマ、イカ類計の合計10種をブリ類の被食魚漁獲量データとして加えた。これらデータも同様に、三重県、和歌山県、静岡県の3県における前年の漁獲量を用いた。

本研究では、以上のすべてのデータが利用可能な1973年から2015年までの43年間のデータを使用した。収集したデータのうち、目的変数となる当年の三重県のブリ類漁獲量を除いた変数は合計341となった(表1)。サンプルサイズに対し説明変数の数が多すぎることで、関連性の低い変数が多く含まれている可能性が高いことから、回帰モデルの説明変数に使用するものを選択した。説明変数の選択にはいくつかの方法があるが、目的変数との相関係数を元に選別した。相関係数を基準とした理由は、目的変数と関連する変数を客観的かつ一番単純な方法で判別できるからである。三重県のブリ類漁獲量と関連性の高い

表2 本研究で回帰モデルに使用したデータ

説明変数					
前年のブリ類漁獲量	三重県	◎	前年のイカ類漁獲量	和歌山県	○
前年のブリ類漁獲量	静岡県	◎	前年のサンマ漁獲量	静岡県	◎
前年のブリ類漁獲量	和歌山県	◎	黒潮流路	前年4月	○
前年のマイワシ漁獲量	三重県	○	黒潮流路	前年6月	◎
前年のカタクチイワシ漁獲量	三重県	◎	黒潮流路	前年11月	○
前年のウルメイワシ漁獲量	三重県	◎	下田水温	2月	◎
前年のシラス漁獲量	三重県	◎	串本西水温	前年3月	○
前年のシラス漁獲量	和歌山県	○	SOI	前年4月	○
前年のイカ類漁獲量	静岡県	○	WP	前年7月	◎

注：◎はV045のみで、○はV04でも使われたデータ。

変数を残すため、同漁獲量との相関係数の絶対値が 0.4 以上の変数（以降、V04 と呼ぶ）を説明変数とした。相関係数の絶対値 0.4 以上を閾値とした理由として、サンプルサイズが 43 の場合の 1% で有意な相関係数は約 0.39 であり、それより絶対値が大きいことが挙げられる。ただし、多重共線性の影響を避けるため、VIF（Variance Inflation Factor）が 10 以上の変数は除いた。同様に、相関係数の絶対値が 0.45 以上の変数（以降、V045 と呼ぶ）を説明変数としたモデルでも評価を行った。それら 2 つの説明変数の組み合わせを表 2 に示した。

本研究では統計モデルとして、線形回帰、サポートベクター回帰、ランダムフォレスト回帰⁽⁶⁾、ニューラルネットワーク型のモデルを使用した。ニューラルネットワーク型のモデルは、中間層の活性化関数をシグモイドの 1 層としたモデル（以降、ニューラルネットワークと呼ぶ）と中間層を正規化線形ユニット（ReLU）の 4 層からなるモデル（以降、深層学習と呼ぶ）を用いた⁽⁷⁾。これら 5 種類のモデルを作成し、それらの比較を行った。ここで、非線形モデルの構築とそれを用いた予測を行うには、3 種類のデータが必要となり、モデル自体を調整（学習）させるために使うデータ、モデルの学習結果を評価するために使うデータ、モデルの最終的な予測結果を評価するためのデータ（以降、それぞれ訓練データ、検証データ、試験データと呼ぶ）に、全体のデータを区分する必要がある。非線形のモデルは表現力が高いため、線形回帰のように訓練データと検証データを同じもので行うと、それに過剰に適合（過学習）してしまい、汎化性能（検証で使ったデータ以外の予測性能）が低下してしまうことが理由である。一方、使用できる最大データ数は 43 個とサンプルサイズが小さいため、データを有効利用するよう工夫した。試験データは 1 年分とし、残り 42 年分のデータを訓練データ、検証データに用いた。試験データは漁獲量が最大値（2015 年）、中央値（2001 年）、最小値（1977 年）をとる年に設定した。それにより、計算量や計算するパターンは 3 倍になるが、予測値が 1 つから 3 つに増えることで予測値を評価しやすくなることに加え、モデル間の比較を通して訓練データ、検証データ、試験データの違いが予測精度に与える影響を検討できるからである。

サポートベクター回帰、ランダムフォレスト回帰は、過学習の防止とデータの有効活用を考慮し、交差検証（cross validation）で訓練データと検証データを併用した。その際、42 の因数となるよう 6 fold と 7 fold のモデルを作成した。ニューラルネットワークと深層学習は、前述の手法と異なり訓練データと検証データを分ける必要があるため、訓練データを 28 年分、検証データを 14 年分とした。モデルは入力層から中間層をそのまま結合したものと全結合したものを作成し、モデル構造の違いによる影響を調べた。また、使用した訓練データと検証データの違いによる影響を評価するため、順に訓練データを 1973～2000 年、検証データを 2001～2014 年、試験データを 2015 年に設定したモデルと表 3 に示したように漁獲量に対して訓練データと検証データを比較的均等になるようにして試験データを 2015 年にしたモデルを作成した。さらに、訓練データと検証データを比較的

表3 各年の漁獲量の順位とニューラルネット型モデルで使したデータの種類

年	順位	種類	年	順位	種類	年	順位	種類	年	順位	種類
2015	1	T or P	1990	12	T	1976	23	T	1999	34	T
2011	2	V	1974	13	T	2004	24	V	1983	35	T
2010	3	T	2006	14	V	2002	25	T	1975	36	V
2007	4	T	1993	15	T	1996	26	T	1987	37	T
2012	5	V	2003	16	T	1988	27	V	1981	38	T
2014	6	T	2005	17	V	1998	28	T	1980	39	V
1973	7	T	1984	18	T	1995	29	T	1985	40	T
1989	8	V	2008	19	T	1979	30	V	1991	41	T
2000	9	T	1997	20	V	1982	31	T	1978	42	V
2009	10	T	1994	21	T	1992	32	T	1977	43	T or P
2013	11	V	2001	22	T or P	1986	33	V			

注：T、V、Pはそれぞれ訓練用、検証用、予測用データを示す。

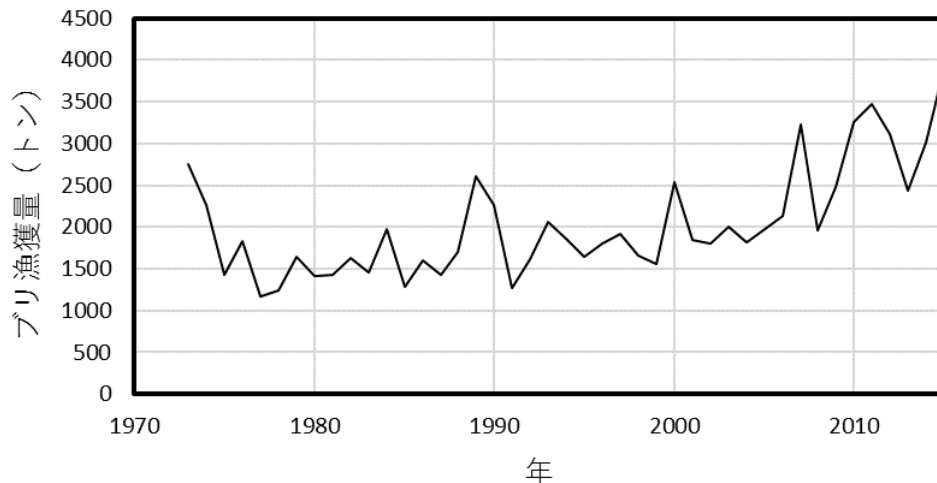


図1 1973年から2015年までの三重県のブリ類漁獲量の時系列

均等になるようにしたモデルでは、試験データを2001年、1977年としたものも行った。

3. 結果

3-1. 三重県のブリ類漁獲量の変動傾向

最初に、対象期間の三重県のブリ類漁獲量の変動傾向について概説する。1973年～2015年までの同漁獲量の変動を図1に示す。この期間の漁獲量最大年は2015年、最小年は1977年で、中央値となる2001年の漁獲量は1846トン、平均漁獲量は2033.7トン、標本標準偏差は648.3トンとなった。平均から標準偏差を超えて多く漁獲された年は、1973、2007、2010、2011、2012、2014、2015の7年で、漁獲量が3000トンを超えた年は1973年を除いた残り6年で、それぞれ3226トン、3259トン、3473トン、3106トン、3012トン、

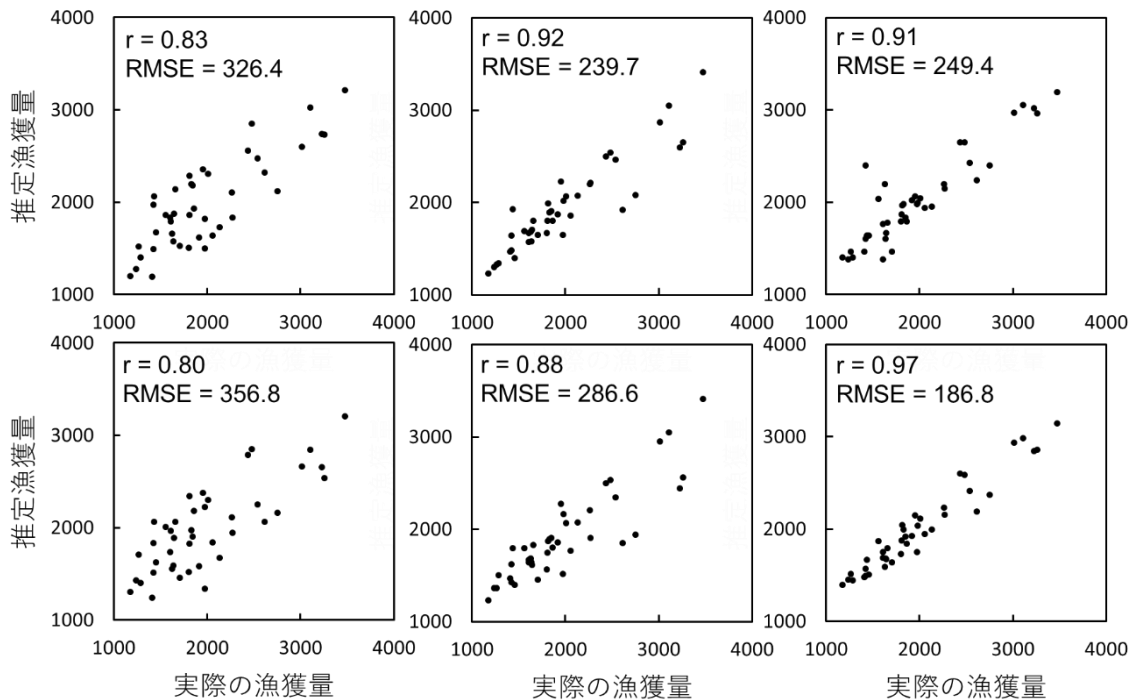


図2 統計モデルの学習結果

注：実際の三重県のブリ類漁獲量（トン）と検証データから各モデル（左：線形回帰、中：サポートベクター回帰、右：ランダムフォレスト回帰）で推定したブリ類漁獲量（トン）の関係、ただし、検証データの期間は1973年～2014年、上段の散布図はV04、下段はV045のデータを使用、非線形回帰の交差検証は6 foldで行った結果を示している。

3886トンとなった。逆に、標準偏差を超えて低い漁獲量の年は、1977、1978、1985、1991年の4年で、漁獲量はそれぞれ1174トン、1239トン、1284トン、1267トンとなる。図1からも明らかなように、三重県のブリ類漁獲量は概ね1000から3000トンで変動しながら、時に3000トンを超える大きな漁獲に恵まれる年が存在する。以下でモデルの結果を述べていくが、最初に訓練、検証データを長く使える線形回帰、サポートベクター回帰、ランダムフォレスト回帰、次にニューラルネットワークと深層学習の結果を示す。

3-2. 線形回帰、サポートベクター回帰、ランダムフォレスト回帰

最初に、2015年を試験データ（訓練、検証データは1973年～2014年）としたモデルの評価を行う。実際の三重県のブリ類漁獲量と検証データを用いてモデルから推定した漁獲量の関係を図2に示す。ここでは、サポートベクター回帰とランダムフォレスト回帰の交差検証は6 foldで行った結果を示している。全てのモデルで、実際の漁獲量と推定漁獲量の相関係数が0.8以上、root-mean-square error (RMSE) は最大で356.8トンとなり、

機械学習を用いた三重県におけるブリ類漁獲量の実用的予測モデルの構築

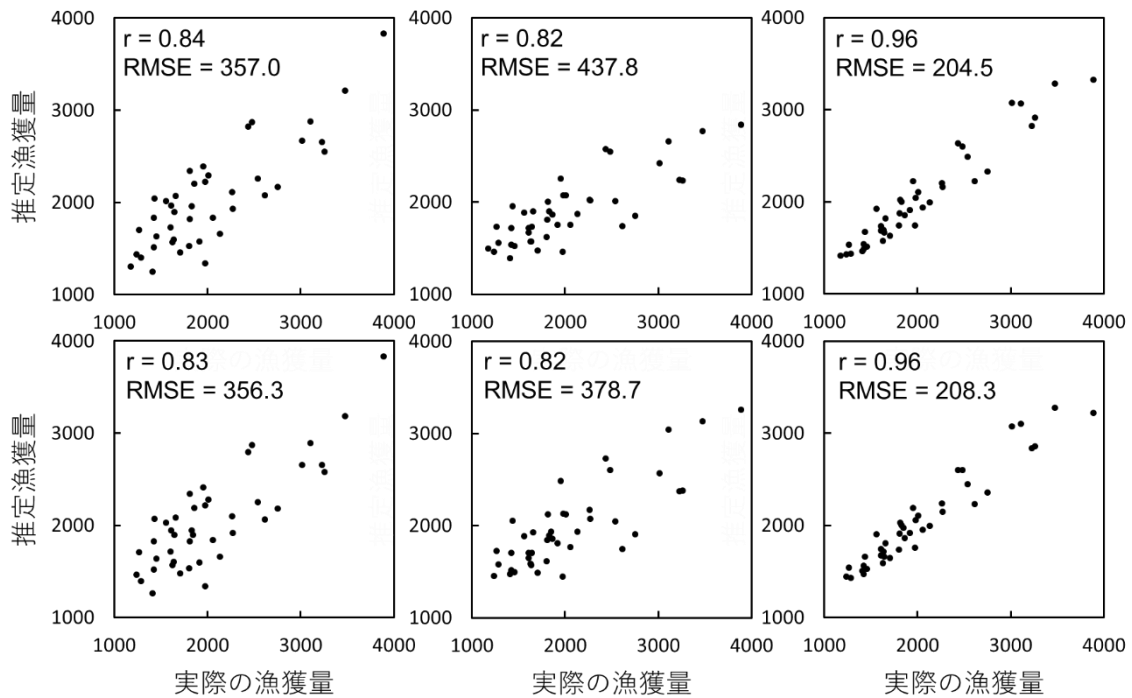


図3 図2と同様の散布図

注：ただし、説明変数は全てV045で、上段（下段）は2001年（1997年）以外の年を検証データに用いた場合の結果を示している。

実際の漁獲量がある程度再現できている。分析手法の違いに注目すると、V04の場合、線形回帰では相関係数が0.83、RMSEが326.4トンであるのに対し、サポートベクター回帰では相関係数が0.92、RMSEが239.7トンとなった。ランダムフォレスト回帰では、相関係数が0.91、RMSEが249.4トンとなり、手法間でモデルの当てはまりに違いがみられた。V04からV045へ説明変数を換えた場合、線形回帰では相関係数、RMSEがそれぞれ0.80、356.8トン、サポートベクター回帰ではそれぞれ0.88、288.6トンとなり、モデルの当てはまりが悪化した。ランダムフォレスト回帰では、相関係数、RMSEがそれぞれ0.97、186.8トンとなり、他の手法とは逆に当てはまりがよくなった。しかし、これらの変化は分析手法の違いと比べて比較的小さい傾向があった。

次に、学習期間の違いが学習結果に与える影響をみるため、試験データを漁獲量の中央値を取る2001年、漁獲量最少年の1977年とした結果を図3に示す。説明変数はV045を使用し、図2と同様に交差検証は6foldの結果である。まず、検証データに加えた2015年の漁獲量最大年の値（約4000トン）に注目すると、試験データを2001年、1977年とした場合に共通して、線形回帰ではよく推定されているのに対し、非線形回帰である2つのモデルでは、大きく過小評価されていることがわかる。次にモデルの当てはまりを示す相関係数とRMSEについてみると、試験データが2015年のもの（図2下段）と比べて、

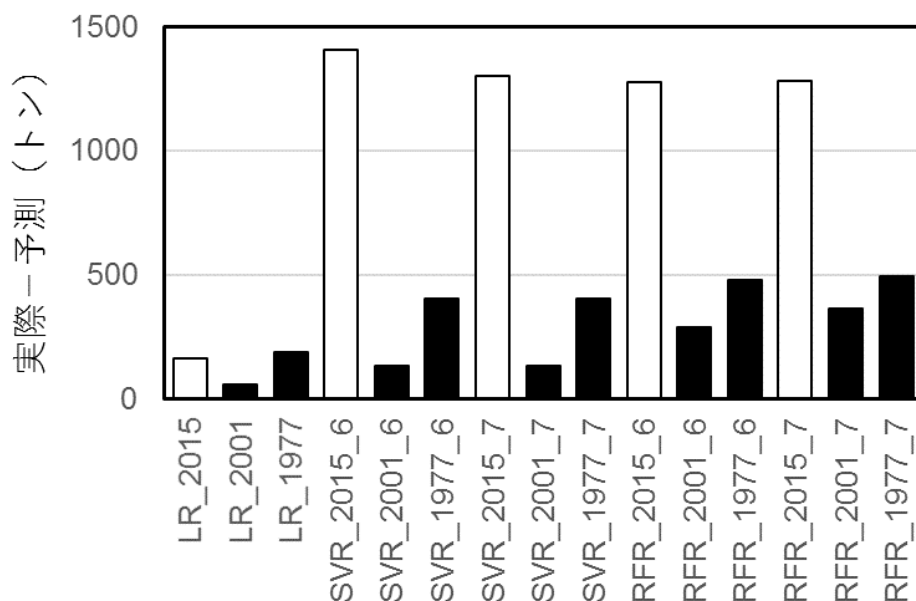


図4 見積もった予測漁獲量（トン）と実際の漁獲量（トン）の差

注：使用した説明変数は V045 のデータで、黒い（白い）バーは過大（過小）評価を示し、最初の英字は使用した統計モデル（LR：線形回帰、SVR：サポートベクター回帰、RFR：ランダムフォレスト回帰）、2番目の数字は予測年、SVR と RFR にのみ付した 3 番目の数字は使用した交差検証の種類（6 は 6 fold で 7 は 7 fold）を示している。

線形回帰では相関係数がわずかに上昇し、RMSE はほぼ同様の値となった。逆に、サポートベクター回帰では相関係数が低下、RMSE は上昇し、線形回帰と比較してもモデルの当てはまりは悪い結果となった。ランダムフォレスト回帰では、予測年の変更で両指数ともあまり大きな変化はみられなかった。

利用者への漁獲量の予測値の提供を念頭に置いて、試験データを用いて訓練、検証に使われていないデータからの予測も行った。モデルから見積もった予測漁獲量と実際の漁獲量の差を図 4 に示す。ここで、説明変数は図 3 同様 V045 で、サポートベクター回帰とランダムフォレスト回帰の結果は 6 fold、7 fold の両方の結果を載せてある。線形回帰の予測値はサポートベクター回帰やランダムフォレスト回帰と比べてずれが小さくなり、モデルの当てはまりが一番高かったランダムフォレスト回帰のずれが一番大きくなった。3 モデル共通して、漁獲量の中央値をとる 2001 年の予測は、実際の漁獲量と予測漁獲量の差が小さくなった。非線形の 2 つのモデルに共通して、交差検証の違いによる影響は小さく、漁獲量最大年である 2015 年の予測は大きく外れており、その差は 1000 トン以上となった。最後に、線形回帰、サポートベクター回帰、ランダムフォレスト回帰の 3 モデルで行った全ての評価結果を表 4 にまとめた。概ね、図 2 から図 4 の結果で述べたことと同様の傾向となっている。しかし、説明変数に V04 を使い、2001 年を試験データとした線形回

表4 線形回帰、サポートベクター回帰、ランダムフォレスト回帰の全ての評価結果

モデル	説明変数	交差検証	予測年			
			2015	2001	1977	
LR	V04		相関係数	0.83	0.87	0.86
			RMSE	326.4	318.9	326.6
			実際-予測	131.9	-651.9	-50.6
	V045		相関係数	0.80	0.84	0.83
			RMSE	356.8	357.0	356.3
			実際-予測	162.6	-58.7	-188.3
SVR	V04	6 fold	相関係数	0.92	0.81	0.84
			RMSE	239.7	442.5	396.1
			実際-予測	1359.3	-278.7	-372.8
		7 fold	相関係数	0.93	0.86	0.82
			RMSE	228.5	357.1	415.4
			実際-予測	1400.0	-403.8	-373.2
	V045	6 fold	相関係数	0.88	0.82	0.82
			RMSE	286.6	437.8	378.7
			実際-予測	1406.1	-134.7	-404.1
		7 fold	相関係数	0.78	0.81	0.80
			RMSE	411.8	452.7	440.4
			実際-予測	1303.3	-134.6	-405.2
RFR	V04	6 fold	相関係数	0.91	0.97	0.97
			RMSE	249.4	198.6	204.4
			実際-予測	1712.1	-293.6	-514.5
		7 fold	相関係数	0.97	0.97	0.97
			RMSE	166.0	205.7	202.9
			実際-予測	1751.4	-323.3	-504.6
	V045	6 fold	相関係数	0.97	0.96	0.96
			RMSE	186.8	204.5	208.3
			実際-予測	1276.5	-292.1	-480.0
		7 fold	相関係数	0.97	0.96	0.97
			RMSE	190.0	206.8	202.8
			実際-予測	1282.4	-367.3	-496.2

注：RMSE、実際-予測の単位はトンで表している。

帰では、実際の漁獲量と予測漁獲量の差が 651.9 トンとなり、V045 のものより大きくなった。他のモデルでも全般的に、説明変数に V04 を用いたものよりも V045 を用いたモデルのほうがよい評価となる傾向があった。

3-3. ニューラルネットワーク、深層学習

次に、訓練データと検証データを別に設定する必要があるニューラルネットワークと深層学習の結果について述べる。訓練データを 1973 年～2000 年の 28 年間、検証データを 2001 年～2014 年の 14 年間、試験データを 2015 年とした場合の両モデルの評価結果を表

5に示す。ここで、2種類の説明変数、入力層および中間層を全結合したものとそうでないもの、の4つのケースにまとめた。前述した3種類のモデル（線形回帰、サポートベクター回帰、ランダムフォレスト回帰）と比較して、どのモデルもモデルの当てはまりが悪く、相関係数が一番高いもので0.56、RMSEが一番低いもので507.8トンであった。また、訓練データ、検証データに含まれていない2015年の予測漁獲量と実際の漁獲量とのずれでは、ニューラルネットワークモデルではどのモデルでも1000トン以上の違いがある。深層学習モデルでは、説明変数がV045、中間層を全結合しないモデルで、実際の漁獲量とのずれが68.5トンと小さくなったが、説明変数がV04では382.5トンの過大評価、中間層を全結合した説明変数がV045のモデルでは1016.9トンの過小評価となり、安定したパフォーマンスは得られなかった。

モデルの当てはまりが悪かった理由はいくつか考えられるが、図1で示した実際のブリ類漁獲量をみると、訓練データである1973年～2000年は比較的漁獲量が低いのに対し、検証データの2001年～2014年は逆に漁獲量が高い傾向にある。そのため、訓練データと検証データの漁獲量の違いを考慮する必要がある。表3に示したように、実際の漁獲量に対して訓練データ、検証データを比較的均等になるように設定したモデルを作成した。この場合では予測年の変更に対して、検証データはまったく変わらず、訓練データは試験データが入れ替わる1年分だけ入れ替わることになる。ニューラルネットワークモデルの評価結果を表6にまとめた。試験データが同じ2015年の結果に注目すると、説明変数の違いや入力層および中間層の結合の違いに関わらず、相関係数、RMSE、予測漁獲量と実際の漁獲量とのずれの全てで改善が見られ、漁獲量に対し訓練データと検証データを均等に設定することが重要であることがわかった。予測年の違いについては、上述した訓練データ、検証データの均等化による違いと比較してそれほど大きな違いはみられなかった。また、深層学習のモデルでも訓練データ、検証データの変更による同様の改善がみられた（表7）。一番評価の高いモデルは、予測年が2001年、説明変数がV045、中間層を全結合するモデルであるが、相関係数、RMSE、実際の漁獲量と予測漁獲量の差はそれぞれ0.79、389.9トン、179.9トンとなり、同じ予測年、説明変数の場合の線形回帰と比較して、モデルの当てはまりが悪く、実際の漁獲量と予測値の差が大きい結果となった。

4. 考察

本稿では気象、海洋、他魚種を含めた漁獲量のデータを用いた機械学習モデルを作成し、三重県のブリ類漁獲量の予測を試みた。以降の部分で、複数のモデルの計算結果を比較し、適するモデルの選択と予測可能性を検討する。まず、訓練、検証データの使用期間、形式について注目すると、訓練、検証データを多く使える線形回帰、サポートベクター回帰、ランダムフォレスト回帰の方が、それらを多く使えないニューラルネット型のモデルより

表 5 2015 年を予測年とした場合のニューラルネット型モデルの評価結果

中間層の結合		単独		全結合	
説明変数		V04	V045	V04	V045
NN	相関係数	0.20	0.19	0.40	-0.08
	RMSE	586.3	598.2	586.7	816.0
	実際-予測	1130.4	1111.0	1048.4	1652.1
DL	相関係数	0.44	0.28	0.56	0.56
	RMSE	717.1	687.0	537.5	507.8
	実際-予測	-382.5	68.5	688.3	1016.9

注 1：訓練データは 1973 年～2000 年、検証データは 2001 年～2014 年。

注 2：NN、DL はそれぞれニューラルネットワーク、深層学習のモデルを示す。

注 3：RMSE、実際-予測の単位はトンで表している。

表 6 訓練・検証データを表 3 のものにした場合のニューラルネットワークの評価結果

説明変数		V04			V045		
予測年		2015	2001	1977	2015	2001	1977
単独	相関係数	0.74	0.77	0.76	0.75	0.79	0.78
	RMSE	426.9	404.2	413.8	435.2	394.3	403.8
	実際-予測	989.5	-115.0	-439.1	941.2	307.5	-389.5
全結合	相関係数	0.76	0.78	0.77	0.59	0.68	0.64
	RMSE	412.1	406.3	407.4	513.4	468.2	487.5
	実際-予測	684.1	413.8	423.2	1071.7	61.6	846.0

注：RMSE、実際-予測の単位はトンで表している。

表 7 訓練・検証データを表 3 のものにした場合の深層学習の評価結果

説明変数		V04			V045		
予測年		2015	2001	1977	2015	2001	1977
単独	相関係数	0.68	0.70	0.69	0.69	0.69	0.70
	RMSE	636.4	658.5	643.7	506.8	528.5	517.2
	実際-予測	766.5	-279.5	-274.4	620.3	502.5	-324.0
全結合	相関係数	0.79	0.78	0.78	0.76	0.79	0.79
	RMSE	391.6	398.8	397.2	414.6	389.9	391.4
	実際-予測	831.1	-447.6	-674.0	927.4	-179.9	-552.7

注：RMSE、実際-予測の単位はトンで表している。

も良好な結果が得られた。また、試験データおよび訓練、検証データをどのように設定するかで、モデルの当てはまりや実際の漁獲量と予測値の差が異なった。非線形のモデルでは 2015 年を予測年とした場合、実際の漁獲量と予測値の差が大きくなった。予測年を漁獲量の中央値が得られる 2001 年とし、訓練データに漁獲量最大年と最小年を含むようにしたモデルは、概ねよいパフォーマンスの予測となった。さらに、2015 年を試験データと

したニューラルネット型のモデルの結果を、訓練データを 1973 年～2000 年、検証データを 2001 年～2014 年としたもの（表 5）、実際の漁獲量に対して訓練データ、検証データを比較的均等になるように設定したもの（表 6、表 7）、で比較すると、前者よりも後者の方が明らかにモデルの当てはまりが高くなった。

次に、モデルや手法の違いに着目すると、上述したようにニューラルネット型のモデルは他の 3 モデルより、予測パフォーマンスが低かった。残りの 3 モデルでは、ランダムフォレスト回帰、サポートベクター回帰、線形回帰の順で、実際の漁獲量と検証データから推定した漁獲量の相関係数が高く、RMSE は低くなる。つまり、同順番でモデルの当てはまりが高くなる。一方、訓練、検証に使われていない試験データを用いた漁獲量予測では、逆にランダムフォレスト回帰、サポートベクター回帰、線形回帰の順で実際の値とのずれが大きくなった。使用した説明変数については、V04、V045 の間であまり大きな結果の違いはみられないが、2001 年を予測年とした線形回帰の予測値と実際の漁獲量の差は、651.9 トンと 58.7 トンで大きな違いがみられた。また、サポートベクター回帰、ランダムフォレスト回帰の交差検証の幅等でも違いがみられたが、モデルの種類による違いと比べて小さい傾向となった。一番予測パフォーマンスが高かった説明変数 V045 の線形回帰での相対的なばらつき（RMSE を平均漁獲量で除した数）をみると、2015 年、2001 年、1977 年がそれぞれ約 0.160、約 0.157、約 0.161 となった。同様に、相対的な誤差（実際の漁獲量と予測漁獲量の差の絶対値を実際の漁獲量で除した数）は、それぞれ約 0.042、約 0.032、約 0.160 となった。このことから、線形回帰モデルは漁獲量の最大値、中央値、最小値を 20% 程度の相対誤差で推定できると考えられる。

本報で使用したデータのように、43 年（1973 年から 2015 年）と比較的サンプルサイズが小さい場合には、訓練、検証データをできるだけ多く、且つそれらのばらつきが大きくなるように使用することが重要である。一方、サポートベクター回帰やランダムフォレスト回帰では、漁獲量最大年である 2015 年を訓練、検証データに含めず、同年を予測した場合に実際の漁獲量と値が大きくずれた。この原因はさらに詳細な検討の必要があるが、いくつかの理由が考えられる。サポートベクターやランダムフォレスト回帰はデータ数が少なくとも高性能を発揮する手法とされているが、本研究のように学習に使用できるデータ数が二桁程度の場合には、必ずしも予測誤差が一定値に漸近せず、本来のパフォーマンスが発揮できないケースがあることが示されている（Matykiewicz and Pestian (2012)）。また、予測する漁獲量が学習させた範囲を超えた場合、訓練、検証データに対して外挿する形になる。これらは非線形のモデルなので、そのことによりずれが大きくなった可能性が考えられる。本研究では広く使われている手法である交差検証を行い過学習の防止に努めたが、サンプル数不足等に起因して過学習の状態になっていた可能性も否定できない。特にランダムフォレスト回帰では、実際の漁獲量と予測した漁獲量が 1200 トン以上のずれがあるのに対して、相関係数が 0.9 以上、RMSE が V04、6 fold の場合で 249.4 トン、

それ以外では 200 トン未満と、データに非常によくフィットしていた。

実際の現場での予測業務を考えると、予測精度の大小だけでなく予測結果が大きく外れることが少なく安定していることも重要である。その点を考慮に入れると、データの違いに対して頑健な結果が得られた、説明変数に V045 を用いた線形回帰モデルが本稿のモデルの中で最も適している。しかしながら、今回のように使用するモデルに予測結果が比較的大きく依存するような場合、他のモデルも併用し、予測結果の不確実性を予測結果と合わせて定量化しておくことは、利用者が予測値の信頼度を判断する際に有益である (Lam *et al.* (2016))。サポートベクターやランダムフォレスト回帰は線形回帰と同様にモデルのデータへの当てはまりがよいことから、これらの手法と線形回帰モデルによる予測結果の相互比較を毎回の予報ごとに行っておくことは利用者によって有益であろう。

本報では、多くのデータを網羅的に収集、選択するとともに、複数のモデルの比較を行い、三重県のブリ類漁獲量を 20% 程度の相対誤差で予測することができるモデルを構築することに成功した。しかしながら、いくつかの改善すべき点が存在しており、それらを以下に述べる。一つはデータの使用期間である。本稿では、利用可能期間が一番短いものに合わせて解析したため、サンプルサイズが小さくなった。そのため、試験データを 1 年分としてモデルを作成し、3 パターンの予測年で実施した結果から予測モデルを評価した。しかし、この方法では 1 つ 1 つのモデルの汎化性能を評価することは困難である。実際に使用したデータのいくつかは 1973 年以前から存在しており、試験データを増やすことができればモデルごとの汎化性能の評価がより容易となる。また、増やしたデータを訓練、検証データに利用することにより、予測モデルが改善されることも考えられる。本研究では、実際の予報業務での使用を念頭に置いた実用的な統計モデルの構築を目的としたため、サポートベクター回帰やランダムフォレスト回帰における学習結果の検証には一般的な手法である交差検証を用いたが、上述のように過学習になっている可能性もある。訓練・検証データを増加させることと、**repeated cross validation** など、より高度な学習手法を採用することにより、過学習を抑え精度向上につながる可能性がある。

また、説明変数の選択についても検討の必要がある。本稿では単相関係数を基準として、V04 と V045 の 2 種類の説明変数の組み合わせを試した。ほとんどのケースで両者の間に大きな違いはみられず、説明変数を決定する閾値の設定がモデルの予測性能に大きな影響を与える可能性は低いと思われる。一方で、予測精度の向上を考えた場合、最適な説明変数の組み合わせは重要である。今後は、線形回帰では変数増減法など AIC を基準とした変数選択などを適用し、説明変数の選択についても改善をはかりたい。また、今回使用しなかったデータを用いることで精度が向上する可能性もある。例えば、黒潮の上流側の徳島県、高知県の漁獲量も有効な指標となる可能性があり、今後の検討課題としたい。

本稿で予測対象としたのは三重県全体の年間の漁獲量であり、特定の月、特定の定置網の予測は行われていない。漁業経営の観点からは、どの時にどれくらいの漁獲が期待でき

るのかをある程度の精度で予測できれば非常に有益である。また、小型のブリは脂がのっていないなど、サイズによって商品価値が異なることから、サイズごとの漁獲量がわかるとさらに有用な情報となる。本研究と同様の手法で定置網ごと、サイズごとの漁獲量を予測できる可能性があり、今後取り組んでいく必要がある。本研究では実際の予報業務を考慮して実用的な方法による予測に取り組んだため、近年急速に発展したデータ同化手法を活用した、時空間的に稠密な気象・海洋の再解析データは使用しなかった(淡路ら(2009))。更なる精度向上を考える場合、このような最新データの活用を検討する必要がある。これから本稿で明らかにできなかったことについては今後の課題としたい。

注

- (1) 山本ら(2007)によると、「漁獲統計上、ブリはカンパチ、ヒラマサと合わせてブリ類と計上されることが多いが、ブリが大きな割合を占めるので、ブリ類の漁獲量変動傾向をブリのそれと見ることができ」とある。データは以下の URL からダウンロードできる(2019年5月18日閲覧)。http://www.maff.go.jp/j/tokei/kouhyou/kaimen_gyosei/
- (2) 日本近海の6海域は、関東の東、関東の南、四国・東海沖、沖縄の東、東シナ海南部と日本全海域平均で、データの情報は以下の URL の末尾にカッコ内の英字をそれぞれ貼り付ければ参照することができる(海域区分図:[japan_warm.html](http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/a_1/japan_warm/)、関東の東データ:[cfig/data/areaL_SST.txt](http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/a_1/areaL_SST.txt)、関東の南データ:[cfig/data/areaM_SST.txt](http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/a_1/areaM_SST.txt)、四国・東海沖データ:[cfig/data/areaA_SST.txt](http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/a_1/areaA_SST.txt)、沖縄の東データ:[cfig/data/areaO_SST.txt](http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/a_1/areaO_SST.txt)、東シナ海南部データ:[cfig/data/areaB_SST.txt](http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/a_1/areaB_SST.txt)、日本全海域平均データ:[cfig/data/areaall_SST.txt](http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/a_1/areaall_SST.txt))。
http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/a_1/japan_warm/(上記 URL すべて 2019年5月18日に閲覧)
- (3) 黒潮流路データは、深さ 200m の水温資料、衛星の海面水温画像等から総合的に判断して決定された東海沖における黒潮流路の月ごとの最南下緯度データで、気象庁より公開され、以下の URL からダウンロードできる(2019年5月18日閲覧)。http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/b_2/kuroshio_stream/kuro_slat.txt
- (4) 黒潮流量データは気象庁により公開されている東経 137 度線を横切る黒潮流量である。夏季と冬季の気象庁海洋気象観測船の観測に基づく深さ約 1250m を基準とした地衡流量から再循環の流量を引いて正味の黒潮の東向き流量を算出したものである。データは以下の URL からダウンロードできる(2019年5月18日閲覧)。http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/b_2/kuroshio_flow/kt137.txt
- (5) これらの7つのインデックスデータのうち、AO、NAO、PNA、WP は NOAA の WEB サイトからダウンロードした。<https://www.cpc.ncep.noaa.gov/products/precip/CWlink/>の末尾に、以下で示す AO、NAO、PNA の URL を貼り付ければ、各インデックスを参照

できる。また、WP は以下の URL からダウンロードできる。NPI、PDO、SOI は以下の気象庁ホームページからダウンロードした (URL はすべて 2019 年 5 月 18 日に閲覧)。

AO : daily_ao_index/monthly.ao.index.b50.current.ascii、NAO : pna/norm.nao.monthly.b5001.current.ascii、PNA : pna/norm.pna.monthly.b5001.current.ascii、WP : ftp://ftp.cpc.ncep.noaa.gov/wd52dg/data/indices/wp_index.tim、NPI : <https://www.data.jma.go.jp/gmd/kaiyou/data/db/climate/pdo/npiwin.txt>、PDO : https://www.data.jma.go.jp/gmd/kaiyou/data/shindan/b_1/pdo/annpdo.txt、SOI : <https://www.data.jma.go.jp/gmd/cpd/data/elnino/index/soi.html>

- (6) データ解析環境 R を用いてこれら 3 モデルの計算を行った。サポートベクター回帰についてはパッケージ `e1071`、ランダムフォレスト回帰はパッケージ `randomForest` とパッケージ `caret` をそれぞれ使用した。これら 2 モデルでは演算能力向上のため、並列処理のためのパッケージ `doParallel` も併用した。サポートベクター回帰のカーネル関数は RBF カーネルを使用し、 ϵ 不感応関数は庄野ら(2014)にならって 0.1 (デフォルト) とした。RBF カーネル関数のチューニングパラメーター γ とスラック変数における係数 C は、交差検証とグリッドサーチを併用して最適値を決定、使用した。また、ランダムフォレスト回帰では、木の数を 500 本 (デフォルト) とし、サポートベクター回帰と同様に交差検証を用いて評価を行った。個々の木を作成する際の特徴量の数は最大でそれぞれの説明変数の数 (表 2) とし、最適数を用いた。サポートベクター回帰やランダムフォレスト回帰の仕組みの詳細は、赤穂(2008)、平井(2012)、ビショップ(2012)、波部(2012)を参照して欲しい。R のスクリプトについては、里(2014)や以下の URL が参考になる (2019 年 5 月 20 日閲覧)。<http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>、<http://topepo.github.io/caret/model-training-and-tuning.html>
- (7) 岡谷(2015)では、以前のニューラルネットワークモデルでは勾配消出問題によりうまく学習できず、2 層程度のモデルでしか期待通りの結果が得られなかったと記述されている。また、岡野原(2015)によると、ReLU を用いたニューラルネットワークでは、20 層近い深いニューラルネットワークであっても、特に工夫をしなくても学習することができ、勾配消出問題に対処できるとある。これら 2 モデルの作成と演算には、Sony Network Communications Inc. より提供されている Neural Network Console を使用した。合計 32 モデルの評価、予測を行うが、1 つのモデルに対して、学習回数は最大 3000 回とし、10 回ごとに検証を行った。モデル 1 つにつき、計 300 の結果が得られるが、その中から一番フィットしたモデルを採用した。使用方法は同アプリケーションのマニュアルや WEB サイト (<https://dl.sony.com/ja/>) を参照して欲しい。

参考文献

- [1] Beamish R. J. and D. R. Bouillon (1993) "Pacific salmon production trends in relation

- to climate” *Canadian Journal of Fisheries and Aquatic Sciences*, 50, 1002-1016.
- [2] Lam V. W. Y., W. W. L. Cheung, G. Reygondeau and U. R. Sumaila (2016) “Projected change in global fisheries revenues under climate change” *Scientific Reports*, 6, 32607.
- [3] Matykiewicz P. and J. Pestian (2012) “Effect of small sample size on text categorization with support vector machines” *Proceedings of the 2012 workshop on biomedical natural language processing (BioNLP 2012)*, Montreal, Canada, June 8, 2012, 193-201.
- [4] 赤穂昭太郎(2008)『カーネル多変量解析』、岩波書店。
- [5] 淡路敏之・蒲地政文・池田元美・石川洋一編著 (2009)『データ同化 観測・実験とモデルを融合するイノベーション』、京都大学学術出版会。
- [6] 岡谷貴之(2015)『深層学習 (機械学習プロフェッショナルシリーズ)』、講談社。
- [7] 岡野原大輔(2015)「大規模深層学習の実現技術」、神寫敏弘編『深層学習 Deep Learning (監修:人工知能学会)』近代科学社、pp.125-150。
- [8] 川崎健(1994)「浮魚生態系のレジームシフト(構造的転換)問題の10年-FAO 専門家会議(1983)から PICES 第3回年次会合(1994)まで」、『水産海洋研究』第58巻、pp.321-333。
- [9] 阪本俊雄(1991)「和歌山県沿岸域の漁海況」、『海と空』第66巻、pp.347-366。
- [10] 里洋平(2014)『戦略的データマイニング』、共立出版。
- [11] 庄野宏・堀江昌弘・井上あゆみ・東剛志(2014)「機械学習に基づく鹿児島近海に來遊するクロマグロ幼魚の漁獲量予測」、『計量生物学』第35巻、pp.1-15。
- [12] 為石日出生・花岡明・四之宮博(1997)「南下初期の操業データと暖水塊パラメータによるサンマ漁況予測」、『水産海洋研究』第61巻、pp.18-22。
- [13] 湯祖恪・桜本和美・和田時夫・北原武・原田泰志(1992)「道東沖マイワシ漁況のファジィ推論による予測」、『日本水産学会誌』第58巻、pp.1873-1881。
- [14] 馬場真哉・松石隆(2015)「ランダムフォレストを用いたサンマ來遊量の予測」、『日本水産学会誌』第81巻、pp.2-9。
- [15] 波部斉(2012)「ランダムフォレスト」、『研究報告コンピュータビジョンとイメージメディア』2012-CVIM-182巻、pp.1-8。
- [16] ビショップ C.M. (2012)『パターン認識と機械学習 下(ベイズ理論による統計的予測)』、丸善出版。
- [17] 平井有三(2012)『はじめてのパターン認識』、森北出版。
- [18] 山本敏博・井野慎吾・久野正博・阪地英男・檜山義明・岸田達・石田行正(2007)「ブリ (*Seriola quinqueradiata*) の産卵、回遊生態及びその研究課題・手法について」、『水産総合研究センター研究報告』第21号、pp.1-29。
- [19] 横田賢史・北田修一・鶴殿謙二郎・渡邊精一(1998)「富山湾におけるホタルイカの環境要因による漁獲量予測」、『日本水産学会誌』第64巻、pp.975-978。

機械学習を用いた三重県におけるブリ類漁獲量の実用的予測モデルの構築

[付記] 本研究の遂行にあたり JSPS 科研費 (JP16H01844, JP16H04046, JP17H02958, JP19H05697) の助成を受けた。気象庁より日本近海の 6 海域および日本全海域平均の季節平均・年平均海面水温偏差データ、北太平洋の年平均海面水温データ、月別黒潮流路データと季節別黒潮流量データ、AMeDAS データ、北半球、北太平洋及びエルニーニョ現象に関する大規模な気象、気候変動に関連する 7 つのインデックスの提供を受けた。