国際漁業研究
第 20 巻，2021（早期公開版）
（早期公開 2021 年 7 月 13 日）

▸ 論　文 ◂

スパースモデリングによる三重県のブリ類漁獲量予測モデルの改良

Improvement of the Yellowtail Catch Prediction Model in Mie Prefecture Using a Sparse Modelling

山田二久次・大木里夏・久野正博*・吉田彰**・万田敦昌†
（三重大学大学院／*三重県水産研究所／**静岡県水産技術研究所）
Fukuji YAMADA, Rina OHKI, Masahiro KUNO*, Akira YOSHIDA**
and Atsuyoshi MANDA†
（Mie University, Graduate school／*Mie Prefecture Fisheries Research Institute／
**Shizuoka Prefecture Fisheries Research Institute）
E-mail：†am@bio.mie-u.ac.jp

【要約】

万田ら(2020)で提示された三重県におけるブリ類漁獲量予測のための線形重回帰モデルにおいて残された課題となっていた変数選択の手法を、スパースモデリングによって改良した。本研究ではスパースモデリングにおける代表的手法である Least Absolute Shrinkage and Selection Operator（LASSO）および Adaptive LASSO を使用し、これら 2 種類の手法に対し、それぞれ 2 種類の情報量基準を適用することで、合計 4 つの統計モデルを作成した。その結果、300 以上あった説明変数の個数は最大のモデルでも 19、最小のモデルでは 5 まで絞り込まれ､客観的根拠に基づき説明変数の個数を大幅に減ずることに成功した。また、実測の漁獲量と推定漁獲量の相関係数は最大のモデルで 0.93、最小のモデルでも 0.83 となり、三重県のブリ類漁獲量を精度良く推定することができた。説明変数の数が同じ場合、万田ら(2020)の線形重回帰モデルと同じ変数選択手法を採用した場合よりも、スパースモデリングを使用した場合の方がモデルの推定精度は向上した。

【キーワード】

ブリ類漁獲量、LASSO、Adaptive LASSO、三重県

【abstract】

This study aims to improve the statistical model for yellowtail catch in Mie Prefecture, Japan using a sparse modelling. Two methods in the sparse modelling have been tested, namely least Absolute Shrinkage and Selection Operator (LASSO) and Adaptive LASSO, with two different statistical criteria. These methods allow much reduction in a number of explanatory variables (from more than 300 to less than or equal to 19) and

exhibit good performance in estimating fish catches. Most of the new models show better performance than the model proposed in the previous study, even though the model has the same number of the explanatory variables as that in the previous study, suggesting that the sparse modeling provides more accurate and the cost effective statistical model for predicting the yellowtail catch in Mie Prefecture.

1. はじめに

資源量や漁獲量の変動予測は水産分野における重要課題の一つである。回遊魚等の漁獲は環境要因に強く影響を受けることもあり、従来は定性的な予測が行われてきた。しかしながら、定性的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、最近では客観的手法を用いた定量的予測にシフトしてきており（例えば、為石ら(1997)、横田ら(1998)）、予測精度向上のために先端的統計モデルの利用が進められている（例えば、湯ら(1992)）。先端的統計モデル構築のための手法の一つとして機械学習がある。機械学習は人間による学習にあたる仕組みをコンピュータ上で実現する手法の総称で、分類や回帰問題に主に用いられる。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利用されてきており、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)、馬場・松石(2015)）。

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り、この地域に位置する尾鷲市では「市の魚」に指定されている。三重県のブリは主に春先に定置網で漁獲される。紀伊半島南部のブリの漁獲量は年によって大きく異なることから、環境変動と漁獲量の関係が以前から注目されてきた。特に、日本南岸を流れる黒潮は紀伊半島の沖で大きく離岸、蛇行する大蛇行流路の存在が知られており、ブリの漁獲量変動と黒潮流路の変動や黒潮系暖水の接近と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℃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等の気象擾乱発生時には漁があることを報告しており、ローカルな海況変動の重要性が指摘されている。

ローカルな環境だけでなく、グローバルな気候変動も水産物の資源量に影響を与えることが知られている。マイワシの資源量では、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することが指摘されている（川崎(1994)）。さらに、北太平洋亜寒帯域の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が指摘されており（Beamish and Bouillon(1993)）、グローバルな気候変動は高次の栄養段階の生物にとっても大きな影響を与えることが示唆されている。ブリの漁獲量と気候レジ

ームシフトとの関連も指摘されていることから（久野(2004)）、(Tian *et al.* (2012))、三重県南部地域のブリ漁獲量変動にグローバルな気候変動とローカルな海況変動の両方が関連している可能性がある。

これらの背景から、万田ら(2020)は三重県のブリ類漁獲量の予測モデルを構築した。三重県のブリ類漁獲量と関連する可能性があるデータを網羅的に収集し、5 種類の機械学習モデルの比較から有効なモデルを選定した。その結果、サポートベクター回帰やランダムフォレスト回帰等の非線形のモデルよりも、線形重回帰モデルの方が予測年の漁獲量の大小にかかわらず、大きく予報結果が外れることのない安定した予測結果が得られることを示した。一方、万田ら(2020)の線形重回帰モデルの説明変数は、単純に目的変数である三重県のブリ類漁獲量の相関係数と Variance Inflation Factor（VIF）をもとに選択されている。このため、相関係数と VIF の閾値に結果が依存すること、相関係数の絶対値が小さくなる変数の有用性は考慮されていないこと等、変数選択の方法に課題が残されている。過剰適合によるモデル性能の過大評価を避けるためにも、客観的手法によって説明変数を選択することは極めて重要である。

近年の人工知能技術の興隆に伴い、機械学習およびその先進モデルの一つである深層学習（Deep Learning）に注目が集まっている。深層学習では、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できる利点がある。しかしながら、大量のデータをコンピュータに学習させて特徴を見出す方法であることから、精度向上に所謂ビッグデータが必要となること、学習や予測のための計算量が増大する等の欠点も存在する。一方、水産分野では本研究における漁獲量のように大量のデータが集められない問題も数多く存在する。このような学習データ不足を解決するための手段として、スパースモデリングと呼ばれる手法も注目されている(1)。スパースモデリングでは、必ずしも大量データを必要とせず、解析・予測のための計算時間を短縮することができる。さらに、同モデリングでは、客観的基準に基づいて関係性の強い要素を選択するとともに、不必要な要素を排除する。これによって、説明変数として重要なものだけを残し、モデルの過剰適合を避けることができる。

上記を踏まえ、本稿ではスパースモデリングの手法を用いることで、万田ら(2020)における線形重回帰モデルで課題となっていた説明変数の選択方法を改善し、新たに適用した変数選択手法によって改善された線形重回帰モデルの精度検証を行うことを目的とする。使用するスパースモデルは正則化項と呼ばれる罰則項を付けた線形回帰モデルである。非線形回帰手法であるサポートベクターやランダムフォレスト回帰については対象外とし、今回は線形重回帰のみを検討した。

本論文の構成は以下の通りである。本節に続く第 2 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、第 3 節では変数選択とモデルによる推定結果について述べる。第 4 節では本研究の主要な結果をまとめ、その内容について議論する。

2. データ・分析方法

本稿では万田ら(2020)と同様に、海洋、気象、前年漁獲量データから三重県のブリ類漁獲量[2]の予測を試みる。収集したデータを表 1 に示す（出典は注(3)、(4)、(6)、(7)に記載した)。海洋データは、静岡県・三重県・和歌山県の定地水温データ、気象庁による日本近海の海面水温偏差データ[3]、気象庁の黒潮流軸位置・流量データ[4]の 3 種類に大別される。現在継続中である海洋温暖レジームにおいて、青森県、北海道、岩手県など分布の北縁部での漁獲量の増加が報告されていることから[5]、本研究では万田ら(2020)で扱わなかった北方の海域も含め、日本近海全ての海面水温を対象データに加えた。気象データは北半球、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[6]を用いた。漁獲量に関しては、農林水産省の漁業・養殖業生産統計年報で報告されているブリ類、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、イカ類計のデータを用いた[7]。

前述のデータセットから、説明変数として使用するデータの種類と対象期間を検討する。万田ら(2020)では、目的変数である三重県ブリ類漁獲量の当該年前年からのデータを説明

表 1　本研究で収集したデータ

データの種類		開始年月	データの種類		開始年月
定地水温 月平均	伊東	1952 年 1 月	気象・気候変動に関連する指数	AO（月平均）	1950 年 1 月
	稲取	1968 年 1 月		NAO（月平均）	1950 年 1 月
	雲見	1970 年 1 月		PNA（月平均）	1950 年 1 月
	下田	1970 年 1 月		WP（月平均）	1950 年 1 月
	焼津	1971 年 1 月		SOI（月平均）	1946 年 1 月
	浜島	1954 年 1 月		NPI（年平均）	1959 年
	串本西	1967 年 5 月		PDO（年平均）	1901 年
	串本東	1967 年 5 月	黒潮	冬季流量	1967 年
日本近海海面水温 季節平均 年平均	釧路沖	1963 年		夏季流量	1972 年
	三陸沖	1950 年		流路	1961 年 2 月
	関東の東	1949 年	年漁獲量	ブリ類	1956 年
	関東の南	1949 年		イワシ類計	1956 年
	四国・東海沖	1949 年		マイワシ	1956 年
	沖縄の東	1949 年		カタクチイワシ	1956 年
	先島諸島周辺	1950 年		ウルメイワシ	1956 年
	東シナ海南部	1947 年		シラス	1956 年
	東シナ海北部	1952 年		アジ類計	1956 年
	黄海	2003 年		マアジ	1956 年
	日本海南西部	1956 年		ムロアジ類	1956 年
	日本海中部	1956 年		サンマ	1956 年
	日本海北東部	1953 年		イカ類計	1956 年
	全海域平均	1951 年			

表 2　回帰モデルの説明変数に使用したデータ

	データの種類	使用期間
定地水温	伊東，稲取，雲見，下田，串本西，串本東，浜島	前年 1 月~12 月、当年 1~3 月、当年 1~3 月の 3 ヶ月平均
日本近海海面水温	釧路沖，三陸沖，関東東，関東南，四国・東海沖，沖縄東，先島諸島周辺，東シナ海南部，東シナ海北部，日本海南西部，日本海中部，日本海北東部，全海域平均	前年（四季，年平均）と当年（冬季）
気象・気候変動関連指数	AO，NAO，PNA，WP，SOI SOI NPI，PDO	前年 1 月~12 月 当年 1~3 月 前年（年平均）
黒潮	流路（東海沖の黒潮流路の最南下緯度） 冬季流量	前年 1 月~12 月と当年 1~3 月 前年と当年
前年漁獲量	ブリ類，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，イカ類計の年漁獲量	静岡県，三重県，和歌山県

変数に使用し、対象期間を 1973 年から 2015 年までの 43 年間としている。本稿ではできる限り長期間のデータを利用できるようにデータの使用期間を延長した。表 1 にあるように、データの取得期間が短いものから並べると、黄海の海面水温、黒潮の夏季流量、焼津の水温、雲見および下田の水温の順になる。これらのうち、下田の水温は万田ら(2020)で最終的に回帰モデルの説明変数として用いられている。この点を考慮し、1971 年から 2018 年までの 48 年間を本研究の対象期間とした。回帰モデルの説明変数となるデータは表 2 のようになり、変数は合計 305 となった。サンプルサイズに対し説明変数の数が多すぎることから、回帰モデルの説明変数を選別する必要がある。前述のように、万田ら(2020)では目的変数との相関係数と VIF を併用して説明変数の選択を行っているが、本稿では前述のようにスパースモデリングを用いて、より適切な変数選択を試みる。

使用する回帰モデルは、代表的なスパースモデリングの一つである Least Absolute Shrinkage and Selection Operator（LASSO）（Tibshirani(1996)）とその改良版の Adaptive LASSO（Zou(2006)）である。LASSO は線形回帰モデルに正則化項と呼ばれる罰則項が付いたモデルで、同項の追加により過剰適合（過学習）を防ぐとともに最小二乗法（Ordinary Least Squares; OLS）と比較して安定した推定を行うことができる。さらに、LASSO では目的変数への寄与が無視できる説明変数を 0 と推定することによって、自動的に変数選択を行うことができる。以下、LASSO の概要について説明する[8]。

説明変数の数をp個、目的変数、説明変数の観測回数をn回、yを目的変数ベクトル、Xを説明変数行列、βを回帰モデルのパラメータベクトル、Sを誤差二乗和ベクトルとした場合について、OLS のパラメータ推定式とその推定量はそれぞれ以下の式になる。

$$\frac{\partial S}{\partial \beta} = -2X^T(y - X\beta) = 0$$

$$\beta = (X^T X)^{-1} X^T y$$

上式には逆行列$(X^T X)^{-1}$が含まれているが、「説明変数間の相関が非常に高い」、「説明変数の数がサンプルサイズに近いもしくは超えている」等の状況では、「逆行列が計算できない」もしくは「各要素の推定値が低いものになる」といった現象が起きてしまう。

正則化項が含まれる回帰の一つにリッジ回帰がある。OLS と同様に、リッジ回帰でのパラメータ推定式とその推定量は以下のようになる。

$$\frac{\partial S}{\partial \beta} = -\frac{1}{n} X^T(y - X\beta) + \lambda\beta = 0$$

$$\beta = (X^T X + n\lambda I)^{-1} X^T y$$

ここで、Iは$p \times p$の単位行列を示し、正則化パラメータλは任意に設定するハイパーパラメータである。上式で逆行列を示す項の中に$n\lambda I$が含まれることから、行列の対角成分に「尾根（リッジ）」を作った構造になる。この時、行列$X^T X + n\lambda I$は任意の$\lambda > 0$に対して正則となるため[9]、OLS より安定した推定を行うことができる。ここで、リッジ回帰の誤差二乗和は OLS の誤差二乗和に罰則項$\lambda\beta\beta^T$を加えた形になる。一方、LASSO では罰則項がβの二乗の形ではなく、λにβの絶対値を乗じた形になる。この罰則項を用いて LASSO は説明変数の選択を行う。

簡略化のために、説明変数が 2 つ（$p = 2$）の場合で考える。この場合について、リッジ回帰、LASSO を制約付き最小化問題として考えると、制約領域はそれぞれ以下のようになる（sは制約の強さを調整するためのパラメータ）。

$${\beta_1}^2 + {\beta_2}^2 \leq s$$

$$|\beta_1| + |\beta_2| \leq s$$

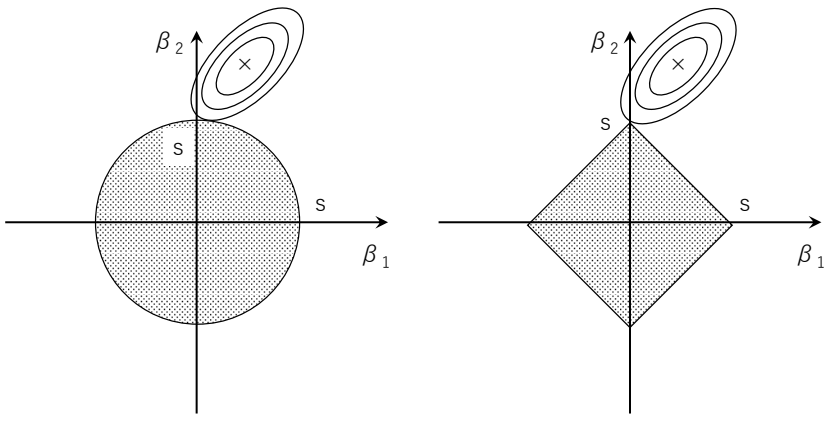

図 1　2 変数の場合のリッジ推定（左）と LASSO 推定（右）の模式図

求めたい推定値は制約領域内にある必要があるが、最小二乗推定値が制約領域内にある場合は最小二乗推定値が求めたい推定値になる。一方、最小二乗推定値が制約領域外にある場合は図 1 のようになる。この図は誤差二乗和をβ_1、β_2の関数としたグラフとして模式的に表したもので、sの値は固定されており、×は OLS で推定した最小二乗推定値の点、楕円は誤差二乗和の等値線(10)、陰部は制約領域をそれぞれ示している。加えて、最小二乗推定値はリッジ回帰（左)、LASSO（右）とも等しいものとする。推定値は制約領域内にある必要があるため、楕円と制約領域との接点が求めたい推定値となるが、LASSO ではβ_1が 0 となるのに対し、リッジ回帰ではβ_1の値は小さいもののぴったり 0 にならないことがわかる。

Adaptive LASSO は二段階で推定を行う LASSO で、適当な重みを置くことによって、ある正則条件下でオラクル性(11) と呼ばれる良い統計的性質を満たすことが知られている。Adaptive LASSO では罰則項が LASSO と異なり、以下の形となる。

$$\lambda \sum_{j=1}^{p} w_j |\beta_j|$$

ここで、w_jは既知の正の重みであり、事前に行った最小二乗推定量等(12)の逆数の絶対値を示している。事前推定量の絶対値が小さい変数ではw_jは大きくなり、正則化の度合いは強くなる。逆も同様である。つまり、このモデルでは、事前推定量の絶対値が小さい変数はモデルに含む必要性が低いため回帰係数を 0 に推定する方向に導き、事前推定量の絶対値が大きい変数は事前に行った最小二乗推定量等に近い推定量を採用することを意味している。本研究では、LASSO、Adaptive LASSO とも R のパッケージ msgps を用いて計算を行った(13)。また、正則化パラメータλの選択には、Mallows の C_p（Mallows(1973)）、Bayesian Information Criterion (BIC)の 2 基準を用いた。

3. 結果

解析結果に触れる前に、三重県のブリ類漁獲量の変動傾向を確認する。1971 年～2018 年までの同漁獲量の変動を図 2 に示す。この期間の平均漁獲量は 2236 トン、標本標準偏差は 1129 トン、中央値は 1888 トンとなった。時間の経過とともに漁獲量が増加する長期トレンドがあり、最近数年間は特に漁獲量が大きくなっている。漁獲量は 2017 年に最大値 7917 トンを記録したが、これは標準偏差の 5 倍を超える突出した値である。次いで 2018 年の 4646 トン、2015 年の 3886 トンが続く。最小値は 1977 年の 1174 トンであった。万田ら(2020) で解析に使用された 1973 年～2015 年の期間では、平均漁獲量と標本標準偏差はそれぞれ 2034 トン、648.3 トンで、漁獲量最大年は 2015 年、最小年は 1977 年であった。近年の漁獲量の増加を反映して、本研究で使用したデータ期間（1971 年～2018 年）

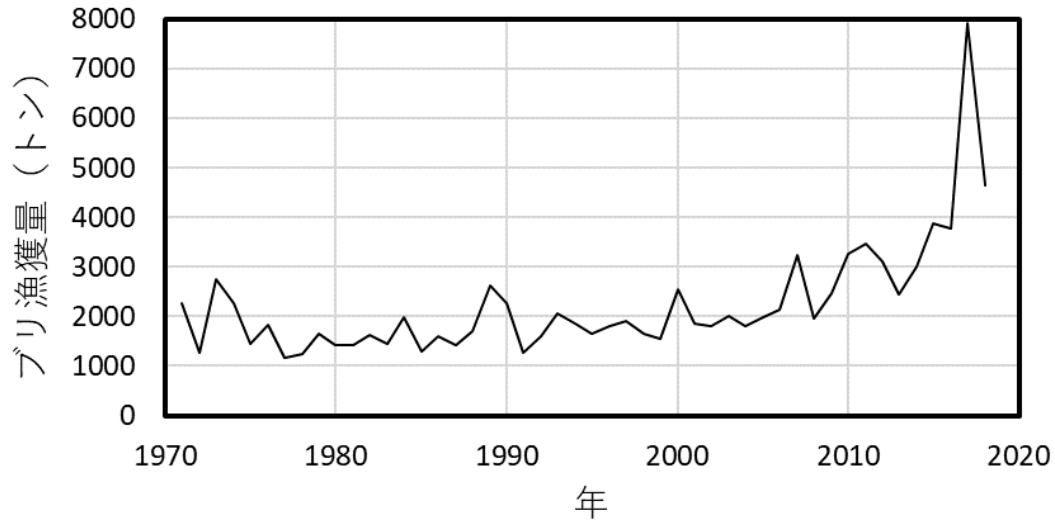

図 2　1971 年から 2018 年までの三重県のブリ類漁獲量の時系列

では最大値や標準偏差が大きくなった。

正規分布を仮定した統計学では、2017 年のようなデータが含まれている場合、そのデータを通常は異常値（外れ値）として分離して処理する。よって、本研究ではまず 2017 年のデータを除外してモデルを作成した（2017 年を含む結果については後述）。また、汎化予測性能の評価は行わず、各説明変数の選択の問題やモデルへの寄与の大きさ等に重点を置いて分析することとした(14)。2017 年を除いた場合では、平均漁獲量、標本標準偏差、中央値は、それぞれ 2115 トン、774.6 トン、1860 トンとなった。2017 年を除くことによって標準偏差は小さくなり、万田ら(2020)と同程度の値となった。

LASSO で推計した切片、回帰係数を表 3 にまとめた。ここで、係数間の比較を容易にするため、目的変数、説明変数とも標準化されたデータを使用した。また、「-」で示された箇所は係数が 0 で推計されたことを表している。C_p基準、BIC 基準で作成したモデルのどちらかで 0 でない値が得られた説明変数については、三重県のブリ類漁獲量との相関係数も示した。C_p基準、BIC 基準で、それぞれ 19 変数、9 変数が 0 でない係数として推計され、もともと 305 あった説明変数の数を大幅に削減することができた。目的変数との相関係数をみると、絶対値が一番大きいもので 0.71、逆に小さいものは 0.01 と、相関係数の絶対値の大きさに対して大きな偏りがみられた。すべての変数は標準化されていることから平均値は 0 となり、このことから切片は 0 となることが期待されるが、両基準とも推定された切片の絶対値は極めて小さく、モデルのバイアスは無視できる。

回帰係数の絶対値の大きさに注目すると、「前年の三重県ブリ類漁獲量」の係数が一番大きく、「前年の三重県ウルメイワシ漁獲量」、「前年の和歌山県イワシ類漁獲量」の順で係数の絶対値が大きくなった。イワシの漁獲量で、三重県では正の値を示すのに対し、和歌山県では負の値を示したが、この結果は万田ら(2020)と整合的である。推計した係数の絶対

表 3　LASSO の推計結果

説明変数		C_p	BIC	目的変数との相関係数
定地水温	稲取　前年 10 月	0.003	-	0.22
	雲見　前年 2 月	-0.008	-	-0.16
	下田　当年 1-3 月平均	0.018	-	0.32
	串本東　前年 7 月	-0.083	-0.021	-0.25
日本近海海面水温	釧路沖　前年夏季平均	0.044	-	0.42
	四国・東海沖　前年年平均	0.093	0.038	0.51
	日本海北東部　当年冬季平均	0.055	0.029	0.33
気象・気候変動関連指数	NAO　前年 6 月	0.056	-	0.01
	PNA　前年 6 月	0.019	-	0.10
	WP　前年 1 月	0.032	-	0.27
	WP　前年 7 月	-0.015	-0.008	-0.35
	WP　前年 9 月	-0.006	-	-0.26
黒潮	流路　前年 6 月	0.076	0.017	0.35
前年漁獲量	イワシ類　和歌山県	-0.162	-0.161	-0.64
	アジ類　和歌山県	-0.062	-	-0.24
	ブリ類　和歌山県	0.064	0.064	0.66
	ウルメイワシ　三重県	0.213	0.213	0.68
	マアジ　三重県	-0.025	-	-0.20
	ブリ類　三重県	0.282	0.281	0.71
切片		-2.74×10^{-11}	-3.04×10^{-11}	-

値が 2 モデル間であまり変わらない説明変数は、上述の 3 変数と「前年の和歌山県ブリ類漁獲量」で、三重県のブリ類漁獲量との相関係数の絶対値は 0.6 以上と高い値を示した。一方、「前年 7 月の串本東の水温」、「前年年平均の四国・東海沖海面水温」、「当年冬季平均の日本海北東部海面水温」、「前年 7 月の WP 指数」、「前年 6 月の黒潮流路」は、両モデルとも係数が 0 とならなかったが、モデル間で得られた係数の値に違いがみられた。これらの 5 説明変数と三重県のブリ類漁獲量との相関係数の絶対値では、「前年年平均の四国・東海沖海面水温」では 0.51 と比較的強い相関があるが、それ以外の 4 変数では 0.35～0.25 となり、弱い相関しか見られなかった。BIC 基準のモデルで回帰係数が 0 と推計された 10 変数では、「前年の和歌山県アジ類漁獲量」、「前年 6 月の NAO 指数」、「前年夏季平均の釧路沖海面水温」の回帰係数の絶対値は比較的高い値を取るが、それ以外の係数は低かった。三重県のブリ類漁獲量との相関係数の絶対値でも「前年夏季平均の釧路沖海面水温」は 0.4 以上の値が得られたが、それ以外での変数では相関関係は弱かった。

標準化した実際の三重県のブリ類漁獲量とモデルから推定した漁獲量の関係を図 3 に示す。C_p基準では、実際の漁獲量と推定漁獲量との相関係数、RMSE（Root Mean Square Error）、MAE（Mean Absolute Error）は、それぞれ 0.93、0.43、0.34 であり、実際の漁獲量に換算した RMSE と MAE はそれぞれ 331.2 トン、263.5 トンとなった。相関係数が

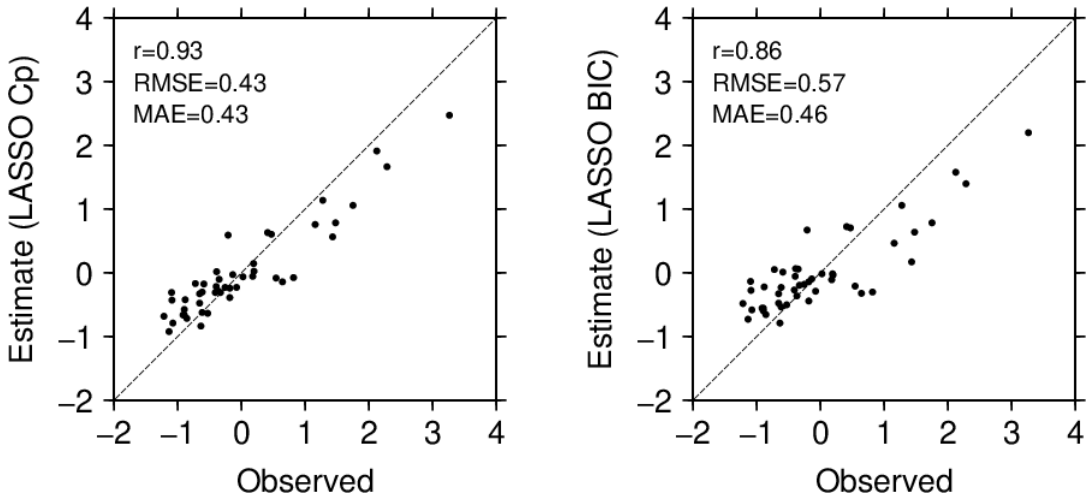

図 3　標準化した三重県のブリ類漁獲量と LASSO による推定漁獲量の関係
（左図はC_p基準、右図は BIC 基準）

表 4　Adaptive LASSO の推計結果

説明変数		C_p	BIC	目的変数との相関係数
定地水温	稲取　前年 10 月	0.026	-	0.22
	雲見　前年 2 月	-0.035	-	-0.16
	串本東　前年 7 月	-0.052	-	-0.25
日本近海海面水温	釧路沖　前年夏季平均	0.022	-	0.42
	日本海北東部　当年冬季平均	0.113	0.062	0.33
気象・気候変動関連指数	NAO　前年 6 月	0.083	-	0.01
	PNA　前年 6 月	0.100	2.60×10^{-4}	0.10
前年漁獲量	イワシ類　和歌山県	-0.224	-0.085	-0.64
	アジ類　和歌山県	-0.068	-	-0.24
	ウルメイワシ　三重県	0.264	0.252	0.68
	マアジ　三重県	-0.021	-	-0.20
	ブリ類　三重県	0.398	0.398	0.71
切片		-1.80×10^{-11}	-3.14×10^{-11}	-

高く、両者の間におおむね線型的な関係があることがわかる。BIC 基準でも実際の漁獲量と推定漁獲量との相関係数は 0.86 と比較的高い値を示したが、C_p基準と比べるとその値は減少し、さらに RMSE、MAE はそれぞれ 0.57、0.46 と増加した。RMSE と MAE の差はC_p基準で 0.09 であるのに対し、BIC 基準の方では 0.11 と差が大きくなった。RMSE は差の二乗平均の平方根なので外れ値の影響が大きくなる。つまり、実際の漁獲量と推定漁獲量との差がある程度大きいデータがいくつか含まれると、MAE と比べて RMSE の値の方が大きくなる傾向がある。散布図からも BIC 基準ではC_p基準よりも外れた点が多いこと

がわかる。上記のことから、LASSO を使用した場合、C_p基準の方が BIC 基準よりも推定精度が向上することが明らかとなった。

次に、Adaptive LASSO の結果を表 4 に示す。LASSO と同様、切片はほぼ 0 となった。LASSO の推計結果と同様に、「前年の三重県ブリ類漁獲量」の係数が一番大きく、「前年の三重県ウルメイワシ漁獲量」、「前年の和歌山県イワシ類漁獲量」の順で係数の絶対値が大きくなるとともに、三重県のウルメイワシでは相関係数は正、和歌山県のイワシ類では負の値を示した。C_p基準で推計された LASSO と Adaptive LASSO の 0 以外の回帰係数をとった説明変数を比較すると、Adaptive LASSO の同変数は LASSO から「当年 1-3 月平均の下田の水温」、「前年年平均の四国・東海沖海面水温」、「前年 6 月の黒潮流路」、「前年の和歌山県ブリ類漁獲量」と前年 1、7、9 月の 3 種 WP 指数の計 7 変数を除いた 12 変数となった。同様に BIC 基準同士の比較でも、0 でない係数が推計される変数は 9 変数から 5 変数に減少したが、C_p基準の場合と異なり「前年 7 月の串本東の水温」、「前年年平均の四国・東海沖海面水温」、「前年 7 月の WP 指数」、「前年 6 月の黒潮流路」、「前年の和歌山県ブリ類漁獲量」の 5 変数が除かれ、「前年 6 月の PNA 指数」が加わった。

実際の漁獲量と推定漁獲量の関係（図 4）に注目すると、LASSO と同様にC_p基準で作成したモデルの方が、BIC 基準で作成したモデルより精度が良くなった。また、C_p基準で作成した LASSO の方が同じ基準で作成した Adaptive LASSO より相関係数が高くなった。一方、MAE は Adaptive LASSO の方が若干小さくなり、RMSE は LASSO と Adaptive LASSO で同程度の値となった。

本研究で使用したデータセットに対して、万田ら(2020)と同様の線形重回帰分析を適用した結果を図 5 に示す。説明変数の選択手法は万田ら(2020)と同一（目的変数との相関係数の絶対値が 0.45 以上で VIF が 10 未満の説明変数を採用する）である。相関係数、RMSE 、MAE はそれぞれ 0.85、0.53、0.41 となった。この時、選択されたのは「前年年

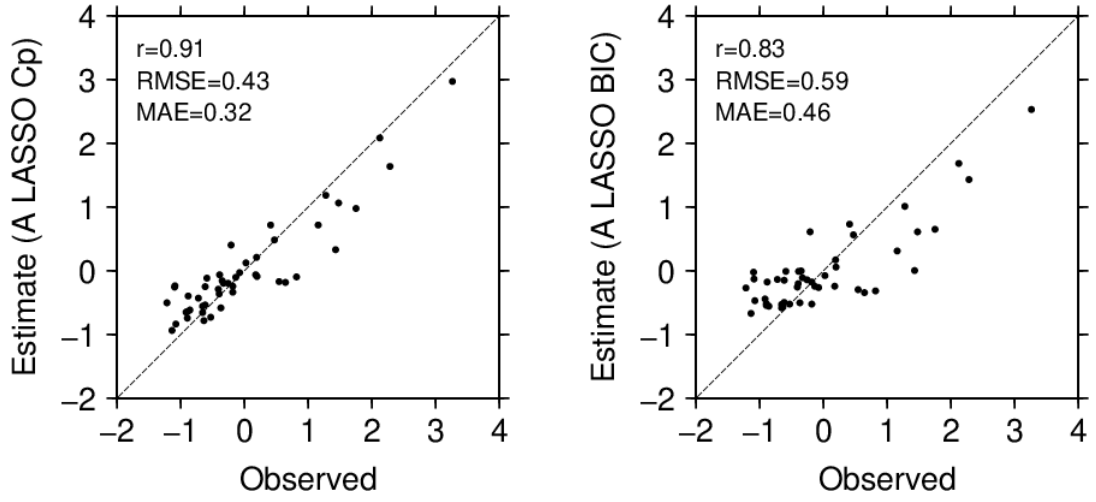

図 4　図 3 と同様（ただし Adaptive LASSO による推定）

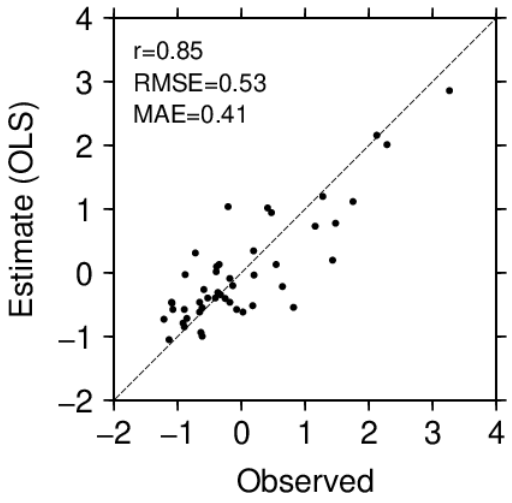

図 5　図 3 と同様（ただし重回帰による推定）

平均の四国・東海沖海面水温」、「前年冬季平均の東シナ海南部海面水温」、静岡県のサンマ、イカ類、ブリ類の前年漁獲量の 3 変数、和歌山県のシラス、イカ類、ブリ類の前年漁獲量の 3 変数、三重県のウルメイワシ、カタクチイワシ、シラス、ブリ類の前年漁獲量の 4 変数、の計 12 変数であった。C_p基準の Adaptive LASSO を使用した場合、万田ら(2020)の線形重回帰よりも、推定漁獲量と実際の漁獲量との相関係数は増加し、RMSE と MAE は減少した。また、C_p基準の Adaptive LASSO の説明変数は 12 であり、万田ら (2020)の線形重回帰を本データに適用した場合と同数となった（表 4)。説明変数が同数にもかかわらず推定精度が向上したことから、万田ら(2020)による主観的変数選択手法よりも、C_p基準の Adaptive LASSO によって変数選択を行った方が、より適切な変数が選択されることが示唆される。

4. まとめと考察

本研究では、万田ら(2020)で作成されたブリ類漁獲量の線形重回帰モデルにおいて課題となっていた説明変数の選択手法を、スパースモデリングを用いることで改良した。このことはモデルの過剰適合を防ぎ、モデル性能の過大評価を避ける上で重要である。また、スパースモデリングでは、変数の選択がモデル構築の際に自動的に行われるため、モデル作成の際の労力が大幅に削減できる。この点は、実際の予報業務での使用を念頭においたモデルを作成する場合、非常に重要な利点となる。

本研究では、LASSO および Adaptive LASSO のそれぞれについてC_pおよび BIC 基準を用いることで、合計 4 種の統計モデルを作成した。最初に選択された説明変数の個数に注目すると、C_p基準、BIC 基準の LASSO でそれぞれ 19 変数、9 変数、同 Adaptive LASSO ではそれぞれ 12 変数、5 変数となり、客観的基準に基づいて説明変数の個数を大幅に減少

させることに成功した。モデルの説明変数の数を作成されたモデルの回帰係数について述べると、4 モデル全てで「前年の三重県ブリ類漁獲量」、「前年の三重県ウルメイワシ漁獲量」、「前年の和歌山県イワシ類漁獲量」の順で係数の絶対値が大きくなった。これらの変数は目的変数である三重県のブリ類漁獲量と相関の高い変数で、その絶対値は 0.6 以上の高い値を示した。それ以外には「当年冬季平均の日本海北東部海面水温」が全てのモデルで選択された。

今回作成した 4 モデルとも推定漁獲量と実際の漁獲量との相関係数は 0.8 以上となり、特にC_p基準のモデルでは 0.9 以上の強い相関を示した。RMSE、MAE に関しては、BIC 基準のモデルでそれぞれ 0.59、0.46、C_p基準のモデルでそれぞれ 0.43、0.34 と値が減少した。また、C_p基準の Adaptive LASSO では、万田ら(2020)と同様の手法で説明変数を選択した重回帰モデルよりも相関係数が増加するとともに RMSE、MAE は減少することが示され、より高精度のモデルを構築することに成功した。実際の漁獲量との相関係数が最も大きくなったC_p基準の LASSO における RMSE、MAE を次元量に換算すると、それぞれ約 331 トン、約 264 トンとなる。三重県のブリ類漁獲量の標本標準偏差が約 775 トンであることから、平均的には目的変数の標準偏差の半分以下の誤差で推定されている。

本稿では、異常値の可能性のあるデータを除外するという観点から、突出して漁獲量の高かった 2017 年の漁獲量の推定を行わなかった。一方で実際の予測を考えた場合、2017 年のような極値に関してもその推定が可能かどうか検証することは重要である。この点を考慮し、あえて 2017 年も含めて推計したC_p基準の LASSO、Adaptive LASSO および万田ら (2020) の線形重回帰による推定漁獲量の時系列を図 6 に示す。2017 年の漁獲量 7917

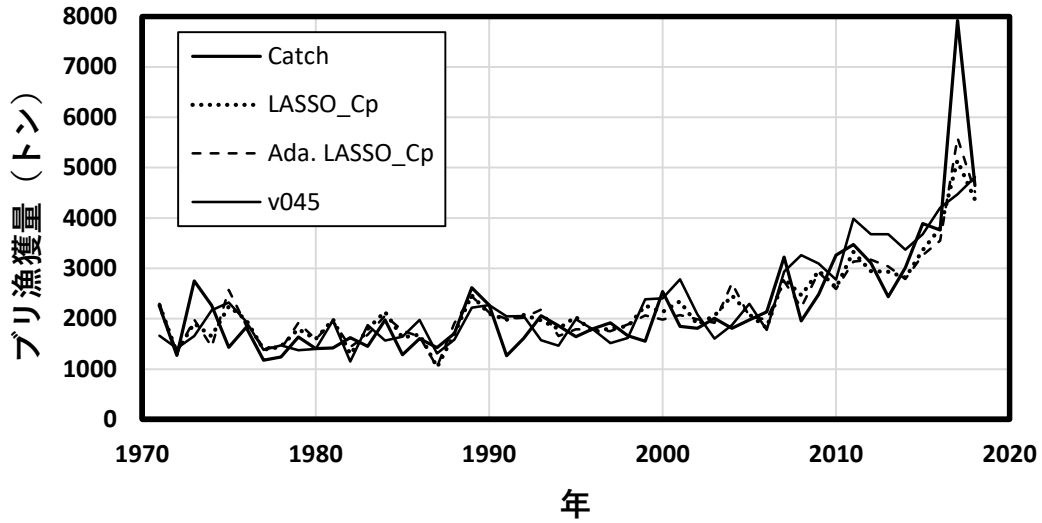

図 6　1971 年から 2018 年までの三重県のブリ類漁獲量（太線）とC_p基準の LASSO（点線）、Adaptive LASSO（破線）、線形重回帰分析（細線）で見積もった同推定漁獲量の時系列

トンに対し、LASSO および Adaptive LASSO の推定漁獲量はそれぞれ約 5141 トン、約 5593 トンと過小評価されたが、同年の推定漁獲量は推定期間中の最大値となった。さらに、両モデルとも推定漁獲量と実際の漁獲量との相関係数は 0.90 と高い値を示しており、漁獲量の変動を大まかには再現できていると考えられる。一方、万田ら(2020)と同様の線形重回帰モデル（目的変数との相関係数の絶対値が 0.45 以上で VIF が 10 未満の説明変数を採用する）では、2017 年の漁獲量が他の手法に比べ大幅に過小評価となるだけでなく、この年の漁獲量が極大となることを再現できなかった。今回はデータをモデル作成のみに用いているため、モデルの汎化予測性能を評価することはできない。しかし、上述の結果から、2017 年と類似の豊漁年のデータを蓄積することによって、極値をとるような年の漁獲量も予測できる可能性がある。

本稿では、LASSO、Adaptive LASSO の 2 種類のモデルに対し、C_p と BIC の 2 つの基準でモデルを作成した。客観的な手法によって、多数の説明変数から重要な変数を選別し、推定精度の高いモデルの構築に成功した。しかしながら、今回検討したのはスパースモデリングにおける代表的な手法である LASSO、Adaptive LASSO の 2 種類のみである。最近提案されたより高度な手法である、罰則項を Smoothly Clipped Absolute Deviation とした SCAD（Fan and Li (2001)）や、罰則項を Minimax Concave とし、計算アルゴリズムに罰則付き線形不偏選択アルゴリズムを用いた MC+（Zhang (2010)）等については検討していない。

一方、さらなるモデル改良の方向性として、スパースモデリング以外の手法を組み合わせることも考えられる。例えば、LASSO を用いて重要な説明変数の最初のスクリーニングを行い、その後比較的簡便な線形重回帰モデルに対して AIC を基準としたステップワイズ法等を適用して、簡便かつ精度の高いモデルを構築するようなアプローチである。

2016 年以降、突出した漁獲量を示した 2017 年に代表されるように、それ以前とは漁獲量の変動特性が変化し、非線形性が顕著になりつつある可能性もある（図 2)。このような場合、線形重回帰を基礎とするスパースモデリングには限界がある。また、資源量のレジームシフトが起こった場合にも、線形重回帰の適用が困難となる可能性がある。万田ら(2020)では過剰適合の傾向にあるとされたサポートベクター回帰やランダムフォレスト回帰のような非線形回帰手法が線形手法よりも今後有効性を増すことも考えられる。非線形回帰における変数選択手法には課題も多く存在するが、近年の漁獲量の変動を注視しつつ、さらなるモデルの改良に取り組んでいきたい。

注

(1) 日高ら(2017)では、科学研究費の新学術領域「スパースモデリングの深化と高次元データ駆動科学の創生」で取り組まれている研究が紹介されている。同モデリングの理論の紹介だけでなく、医療（MRI 等を用いた脳卒中、心筋梗塞、ガン等の画像解析）、生命科学（NMR

計測・解析の高速高精度化)、脳科学(脳における視覚物体像の時空間表現)、地球科学(津波堆積物判別の高精度化、物理化学的挙動の解明)、惑星科学(隕石分析、小惑星観測データのデータベース構築と統合)、天文学(超巨大ブラックホールの直接撮像)等、多様な分野における応用例が紹介されている。また、伊庭ら(2017)では、画像処理、マーケットシェア、スーパーマーケットの購買データにスパースモデリングを適用した例が示されている。

(2) 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る」とある。

(3) 日本近海の 13 海域およびその全海域平均(各海域の面積に応じて重みを付けて平均した値)の海面水温偏差データは、以下 URL にある「日本近海の海域平均海面水温の上昇率」図の該当海域をクリックすることで参照できる(2020 年 7 月 25 日閲覧)。http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/a_1/japan_warm/japan_warm.html

(4) 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ある。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、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。両データともに気象庁より公開され、以下の URL からそれぞれダウンロードできる(2020 年 7 月 25 日閲覧)。

黒潮流路データ:http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/b_2/kuroshio_stream/kuro_slat.txt

黒潮流量データ:http://www.data.jma.go.jp/gmd/kaiyou/data/shindan/b_2/kuroshio_flow/kt137.txt

(5) 例えば、桑原ら(2006)では、気象庁の水温データから将来の日本近海の海水温変化を「短期:現状+1.0℃」、「中期:気象庁の長期水温値」、「長期:同長期水温値+1.5℃」として水産有用魚種の影響を予測しており、日本海北区(青森~石川)の短期~長期、太平洋中区(千葉~三重)の短期~中期で、ブリの漁獲量が増大傾向となることを示している。また、宍道ら(2016)では、ブリ類漁獲量から分布域の重心を求め、同重心は海面水温が上昇する温暖期には北東方向にシフトし、ブリの分布域が広がることが報告されている。

(6) 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 2020 年 7 月 25 日閲覧)。

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

(7) 魚種別の漁獲量データは農林水産省による以下の URL からダウンロードできる（2020 年 7 月 25 日閲覧）。https://www.maff.go.jp/j/tokei/kouhyou/kaimen_gyosei/index.html

(8) アルゴリズム等、LASSO の詳細については、冨岡 (2015)、川野ら(2018)、Irina Rish ら(2019)が詳しい。

(9) リッジ推定量が正則となる理由については、川野ら(2018)の付録 A.2 を参照してほしい。

(10) 川野ら(2018)の付録 A.3 に誤差二乗和の等高線が楕円になる理由が解説されている。

(11) オラクル性についての解説は、荒木(2013)、川野ら(2018)、梅津ら(2020)で述べられている。具体的には、0 でない係数を持つ説明変数に対し、サンプルサイズが大きくなるにつれ同変数が正しく選択される確率が 1 に近づく性質（変数選択の一致性）と同変数の推定量が漸近的に正規分布に従う性質（漸近正規性）を併せ持つ性質を指す。

(12) 最小二乗推定値が得られない場合は、リッジ推定値等を用いる。ただし、リッジ推定値を用いた場合にはオラクル性は満たさない。

(13) パッケージ msgps の使用方法については、CRAN ホームページに掲載されているマニュアル（以下の URL）を参照してほしい。また、荒木(2013)では罰則付き回帰に関連する R パッケージの一覧がまとめられている。本稿と別のパッケージを使用した例ではあるが、伊庭ら(2017)では LASSO、川野ら(2018)、梅津ら(2020)では LASSO と adaptive LASSO の計算をするスクリプトが掲載されている。

https://cran.r-project.org/web/packages/msgps/msgps.pdf（2020 年 7 月 24 日閲覧）

(14) 突出した漁獲量の予測は重要であるが、極値をとる漁獲量データは 2017 年の 1 年だけであるので、モデルの構築に同データを使用すると、極値の汎化性能を評価できない。逆も同様であるので、汎化性能の評価は困難と判断した。

参考文献

[1] Beamish R. J. and D. R. Bouillon (1993) “Pacific salmon production trends in relation to climate,” *Can. J. Fish. Aquat. Sci.*, 50, 1002-1016.

[2] Fan J. and R. Li (2001) “Variable selection via nonconcave penalized likelihood and its oracle properties,” *J. Amer. Statist. Assoc.*, 96, 1348-1360.

[3] Mallows C. L. (1973) “Some comments on C_p,” *Technometrics*, 15, 661-675.

[4] Tian Y., H. Kidokoro, T. Watanabe, Y. Igeta, H. Sakaji and S. Ino (2012) “Response of

yellowtail, *Seriola quinqueradiata*, a key large predatory fish in the Japan Sea, to sea water temperature over the last century and potential effects of global warming," *J. Mar. Sys.*, 91, 1-10.

[5] Tibshirani R. (1996) "Regression shrinkage and selection via the lasso," *J. R. Statist. Soc. B*, 58, 267-288.

[6] Zhang C.-H. (2010) "Nearly unbiased variable selection under minimax concave penalty," *Ann. Stat.*, 38, 894-942.

[7] Zou H. (2006) "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc.*, 101, 1418-1429.

[8] 荒木孝治 (2013)「罰則付き回帰とデータ解析環境 R」、『オペレーションズ・リサーチ』第 58 巻、pp.261-266。

[9] 伊庭幸人・池田思朗・麻生英樹・井出剛・本谷秀堅・日野英逸・尾崎隆 (2017) 「スパースモデリングと多変量データ解析」、岩波データサイエンス刊行委員会編『岩波データサイエンス Vol.5』、岩波書店、pp.4-115。

[10] Irina Rish・Genady Ya. Grabarnik・竹澤邦夫・大関真之・高橋茶子・竹田晃人・徳田悟・藤本晃司・安田宗樹(2019)『スパースモデリング 理論、アルゴリズム、応用』、ジャムハウス。

[11] 梅津佑太・西井龍映・上田勇祐(2020)『スパース回帰分析とパターン認識 (データサイエンス入門シリーズ)』、講談社。

[12] 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。

[13] 川野秀一・松井秀俊・廣瀬慧(2018)『スパース推定法による統計モデリング』、共立出版。

[14] 久野正博(2004)「ブリ資源の長期変動特性と気候のレジームシフト」、『黒潮の資源海洋研究』第 5 号、pp.29-37。

[15] 桑原久実・明田定満・小林聡・竹下彰・山下洋・城戸勝利(2006)「温暖化による我が国水産生物の分布域の変化予測」、『地球環境』第 11 巻、pp.49-57。

[16] 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。

[17] 宍道弘敏・阪地英男・田永軍(2016)「魚獲量重心の変動からみたブリ類の漁獲量変動」、『水産海洋研究』第 80 巻、pp.27-34。

[18] 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来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。

[19] 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。

[20] 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。

[21] 冨岡亮太(2015)『スパース性に基づく機械学習（機械学習プロフェッショナルシリーズ）』、講談社。

[22] 馬場真哉・松石隆(2015)「ランダムフォレストを用いたサンマ来遊量の予測」、『日本水産学会誌』第 81 巻、pp.2-9。

[23] 日高昇治・松下亮祐・楠田哲也(2017)『スパースモデリングって何だ？―データ構造を解き明かす先端技法』、カットシステム。

[24] 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(2020)「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」、『国際漁業研究』第 18 巻、pp.1-19。

[25] 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。

[26] 横田賢史・北田修一・鵜殿謙二郎・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。

[**謝辞**]本研究の遂行にあたり JSPS 科研費（JP16H01844、JP17H02958、JP19H05697）の助成を受けた。また、注(3)、(4)、(6)、(7)に記載したデータを使用した。データ提供機関に対して記して御礼申し上げる。

（受理日：2021 年 2 月 22 日）