ランダムフォレストを用いたサンマ来遊量の予測

ランニングタイトル：ランダムフォレストによるサンマ来遊量予測

馬場真哉，[1]　松石隆[2*]

[1]北海道大学水産科学院

[2]北海道大学水産科学研究院

Pacific saury fishing forecast by using random forests

SHINYA BABA, [1] TAKASHI MATSUISHI[2*]

[1]Graduate School of Fisheries Sciences, Hokkaido University, 3-1-1 Minato-cho, Hakodate, Hokkaido, 041-8611, [2] Faculty of Fisheries Sciences, Hokkaido University, 3-1-1 Minato-cho, Hakodate, Hokkaido, 041-8611, Japan

——————————————————————————————————

ランダムフォレストを用いたサンマ来遊量の予測

馬場真哉，[1]　松石隆[1]

[1]北大院水

本研究では，ランダムフォレストを用いたサンマ来遊量予測モデルを作成し，その予測精度をモンテカルロリサンプリングにより評価した。応答変数はサンマ来遊資源量指数を 3 カテゴリに分けたものである。説明変数は 1972-2011 年の海洋環境など 22 項目 186 種類を使用した。変数選択の結果，4 種の変数のみが説明変数として選ばれ，説明変数の圧縮が可能となった。予測の的中率はおよそ 62%となり，現状の予測精度をやや上回った。

キーワード：機械学習，漁況予測，サンマ，モンテカルロリサンプリング，ランダムフォレスト

In Japan, Pacific saury fishing forecast is indispensable information for fisheries. In the current research, Pacific saury fish abundance in the fishing ground was predicted by using the random forests and the forecast was evaluated by Monte Carlo re-sampling method. Pacific saury abundance index with three categories was used for response variables. A total of 186 explanatory variables from 22 indices of marine environments from 1972 to 2011 were used. After the variable selection by random forests, only four kinds of variables were chosen as explanatory variables. The hitting ratio became approximately 62% and exceeded the present posted forecast accuracy. By using random forest, compression of the explanation variable can be enabled and could make accurate forecast by using many explanation variables.

漁況予報に関する研究は長く続けられてきており，実証研究は数多い。[1-3] 漁況予報を利用することは，計画的かつ効率的な操業に役立ち，値崩れなどが防止されて魚価が安定することによって，漁業者の経営も安定すると考えられる。[4] よって，来遊量を事前に予報することは，漁業管理上極めて重要である。[5]

現状では勘と経験に頼って予測を行うことが多く，現在予測精度が高かったとしても，将来的に予測を出す人が変われば予測の精度まで変わる可能性がある。将来へ向けての予測の改善として，経験豊富な人材がいなくても安定してよい予測が出せることが重要であり，予測の基礎を人間の勘と経験からコンピュータを用いた処理方式に移行させることが望まれている。[6] 日本ではサンマ *Cololabis saira* の漁況予報は漁業者の営漁方針を立てるうえで必要不可欠なものとなっている。[7] さらにサンマの漁海況予報に関する実証研究も多く，[4,8] 豊富な知見がそろっている。このようにデータが豊富である上に漁況予報のニーズが高いため，データをもとにして，人材や経験によらない安定した予測を行うことができる機械学習法を適用する対象として，サンマは適しているだろう。

本論文では，機械学習法の一種であるランダムフォレスト [9] を使用した来遊資源量予測モデルを提案する。ランダムフォレストの学習方式は計算速度が速く，外れ値やノイズに対して相対的に頑健であり，分類問題における性能が相対的に良いことが知られている。[10] ランダムフォレストには， (i) サンプルサ

イズより説明変数の種類数が多くても使用できる， (ii) 結果として予測に寄与しない変数が多数含まれていても予測能力が低下しにくい，(iii) オーバーフィッティングしにくい，などの特徴がある。[11)]

本研究では，ランダムフォレストを用いて，水温や気象インデックスを含む多種類の説明変数から，サンマ来遊資源量指数の水準を予測するモデルを構築した。そして，モンテカルロリサンプリング法を用いて予測分布を推定し，ランダムフォレストを用いた予測の精度評価を行うことを目的とした。

材料と方法

応答変数　本研究では東北区水産研究所が発表しているサンマの来遊資源量指数（旬別 30 分桝目の 1 網あたり平均漁獲量の累積値 [12)]）を来遊量の指標として用いた。本研究においては，来遊資源量指数を，「高位」「中位」「低位」の 3 つの水準に分けた。これは，現行の予報が来遊量を定量的に予測するものではなく「前年を上回る」「前年並み」など定性的な予測であることに合わせている。[12)] なお，予報の水準は，現行の予報とは異なる水準を採用した。その理由は，現行の予報では「前年を上回る」という言葉の明確な定義がなされていないためである(夏目，私信，2011 年)。また，「前年を上回る」という前年比較としての予報では，意思決定がしにくいため，来遊資源量指数の水準の予報とした。

具体的には「前年を上回る」という予報が出されたとしても，前年が高位である時，中位である時，低位である時に合わせて行動を変えていかなくてはならず，意思決定がより複雑になってしまうからである。

来遊資源量指数の水準は, 北海道立総合研究機構水産研究本部 (マリンネット北海道：http://www.fishexp.hro.or.jp/exp/central/kanri/SigenHyoka/index.asp， 北海道立総合研究機構水産研究本部，2012 年 7 月 7 日) を参考として分割した。o 年における来遊資源量指数の水準を，分割の基準パラメタ k を使って以下のように分けた。

$$\begin{aligned} &\text{高位} \quad \frac{Y_o}{\overline{Y}} - 1 > k \\ &\text{中位} \quad \left| \frac{Y_o}{\overline{Y}} - 1 \right| \le k \\ &\text{低位} \quad \frac{Y_o}{\overline{Y}} - 1 < -k \end{aligned} \tag{1}$$

ただし，$\overline{Y}$ は 1972 年から 2011 年までの来遊資源量指数の平均値である。北海道立総合研究機構水産研究本部では，$\overline{Y}$ を過去 20 年間の来遊資源量指数の平均値と定めていたが，本研究では水準を分ける基準を全ての年で統一させるため，このような定義とした。Y_o は，o 年の来遊資源量指数を表す。

本研究では，Baba and Matsuishi で示された予報の持つ情報量を最大にするという観点 [13] から, 各水準がほぼ 1:1:1 で出現するように, 情報エントロピー[14] を最大化する k を設定した。ここで，情報エントロピーは以下で計算される。

$$H(Level) = -\sum_{i=1}^{3} p(l_i)\log p(l_i) \tag{2}$$

ただし，$p()$は来遊資源量指数の水準が「高位」「中位」あるいは「低位」となる確率を表す。*Level* は，実際の来遊資源量指数の水準を示す$(l_i \in Level)$。(l_1, l_2, l_3)は，各々来遊資源量指数の水準が「高位」「中位」「低位」である状態として定義した。$H(Level)$は来遊資源量指数の水準の情報エントロピーであり，$p(l_i)$ が 0 の時 $p(l_i)\log p(l_i) = 0$ と定義される。対数の底として 2 を使用したため，情報量の単位は bit で表される。

情報エントロピーを最大にするパラメタ $k_{entropy}$ は，以下のように定義した。

$$k_{entropy} = \arg\max_{k}\left[-\sum_{i=1}^{3} p(l_{i|k})\log p(l_{i|k})\right] \tag{3}$$

なお, $l_{i|k}$ は分割の基準パラメタ k が設定された条件の元での各水準を表す。k は，0 から 1 の範囲内を 0.001 刻みで変化させた。本研究においては，$H(Level)$が同値になる k があった際には，その時の k の最小値を使用した。

説明変数　サンマの来遊量には，海洋環境が大きく影響すると考えられる。[15] そこで，本研究では「高位」「中位」あるいは「低位」という前年の来遊資源量指数の水準に加えて，21 項目 185 種類の環境データも説明変数として加えた（Table 1）。つまり使用した説明変数は，22 項目 186 種類である。なお，環境デ

ータは気象庁(気象庁ホームページ : http://www.jma.go.jp/jma/，気象庁，2012 年 7 月 10 日)と NOAA(Climate Prediction Center : http://www.cpc.ncep.noaa.gov/，NOAA，2012 年 7 月 12 日)で公開されている，広域の海洋，気象データのうち，南太平洋のみの水温など明らかにサンマの生息域と合致していないものを除いて全てを使用した。各変数の定義は気象庁及び NOAA の定義に従う。サンマの来遊量予報は，漁期に合わせて例年 7 月末に発表される。[12] そのため，本研究では前年と当年 6 月までの環境データのみを説明変数として使用した。ただし，PDO や北太平洋亜熱帯モード水の水温，NPI，親潮面積，親潮南限データは月平均値が公開されていないため，前年データのみを使用した。なお，本研究における年とは 1 月から 12 月の区間を指し，例えば，2000 年は 2000 年 1 月から 2000 年 12 月の期間であると定義した。

黒潮流路の月平均値と親潮南限の年平均値に関しては欠損値があったため，

Fig. 1

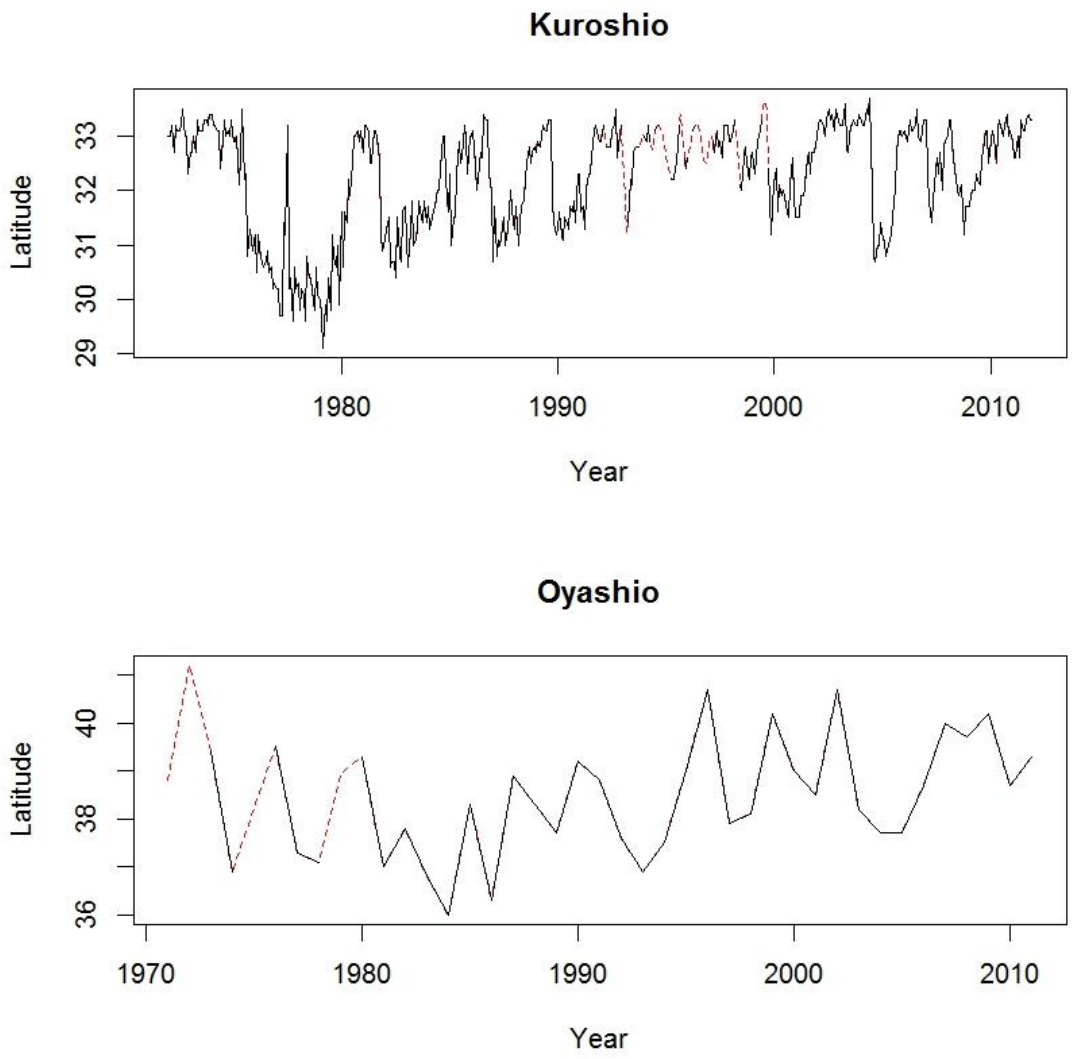


Fig. のように 3 次スプラインによる補間を行った。

Table 1

全ての説明変数及び応答変数は，1972 年から 2011 年まで 40 年間分を使用した。なお，1996 年の来遊資源量指数のデータが欠測であるため，サンプルサイズは 37 になる。

ランダムフォレスト　来遊量予測モデルとしてはランダムフォレスト[9]を使用した。ランダムフォレストのアルゴリズムについて，詳細は参考文献[16]に譲るが，その概要を解説する。ランダムフォレストは，分類回帰樹木法（Classification and Regression Trees: CART 法）を基礎に持つアンサンブル学習器の一種である。アンサンブル学習器は，多くの基本学習器の結果を統合して予測を出力する。本論文では CART 法を概説したうえでアンサンブル学習法であるバギングを解説したのち，ランダムフォレストの説明に移る。なお，本論文では全てカテゴリデータへの分類問題のみを扱うが，ランダムフォレストは回帰問題にも適用可能である。

CART 法[17]は，ある分岐規則を満たせば来遊資源量が「高位」になり，また別の分岐規則を満たせば「中位」あるいは「低位」に分類されるという分岐規則を策定することによって予測値を返す。ここで，分岐規則とはある変数の値が閾値以上か否かといった基準で分けることを意味する。このとき，分岐させた部分集合を「ふし」と呼ぶ。ふし t において来遊資源量指数の水準が l_i をとる確率を $p(l_i|t)$ と定義する。この時，ふし t における予測値は以下のようになる。

$$\hat{l}_i(t) = \arg\max_{l_i}\left[p(l_i|t)\right] \tag{4}$$

分岐の評価基準には，分岐されたそれぞれのふしの応答に対する不均一性の測

度が用いられる。不均一性の測度としてはGini係数を使用した。

$$r_c(t) = \sum_{i=1}^{3} p(l_i|t)(1 - p(l_i|t)) \tag{5}$$

上記の不均一性の尺度$r_c(t)$を用いると，ふしtでの不均一性$R(t)$は，以下のように定義される。

$$R(t) = p(t) r_c(t) \tag{6}$$

ここで，$p(t)$はサンプルサイズNに対するふしtに属するデータの割合であり，ふしtに属するサンプルサイズを$N(t)$とすると，$N(t)/N$で計算される。CART法は，このリスクの再代用推定値を最も大きく減少させる分岐規則を作成することによって計算される。

$$s_t^*(t) = \arg\max_{s_i \in S} [\Delta R(s_t, t)] \tag{7}$$

ただし，ふしtにおいて考えられうるすべての分岐規則s_tの集合をSとおく。ΔRは以下に示すように，ふしtでの不均一性の減少量を表す。

$$\Delta R(s_t, t) = R(t) - R(t_L) - R(t_R) \tag{8}$$

ここで，ふしtに属するデータは分岐規則s_tを満たせばt_Lに，満たさなければt_Rに送られると仮定している。

バギング[18]はアンサンブル学習法の一つである。樹木モデルを基本学習器に採用したアンサンブル学習法にはいくつかの種類があるが，本研究で使用され

たランダムフォレストは，バギングと同様にブートストラップ法にもとづく接近法として分類される。

バギングは，以下の 3 つのステップを踏んで計算される。

(i)　重複可能な抽出により，*B* 個のブートストラップ標本を作製する。

(ii)　各ブートストラップ標本に対して，CART 法により樹木を作成する。

(iii)　すべての樹木から得られた予測結果の多数決をとり，予測値とする。

バギングは多くの樹木を構築して統合することにより，単一の樹木よりも予測確度を上げることができる。[16)]

ランダムフォレストはバギングの流れをくんだ解析手法であるが，アルゴリズムに一部変更が加えられている。バギングは，標本選択においてブートストラップ標本を使用することによってランダム性を加えていたが，さらにランダムフォレストは分岐変数選択にもランダム要素を導入している。これにより分類木の相関が減り，推定の安定性が高まる。ランダムフォレストのアルゴリズムを概説すると以下のようになる。

(i)　重複可能な抽出により，*B* 個のブートストラップ標本を作製する。

(ii)　各ブートストラップ標本に対して，以下の規則 A，B を満たす CART 法により樹木を作成する。

A)　ふし t の分岐には，ランダムに選択された P 個の変数のみを用いる

B)　すべての末端のふし t 内のデータ数 $N(t)$が，あらかじめ規定された値よりも小さくなるまで分岐を続ける。

(iii)　すべての樹木から得られた予測結果の多数決をとり，予測値とする。

バギングと比べてランダムフォレストは予測結果が安定し，かつ相対的に高い予測精度を示す例も報告されている。[16)]

本研究では，Uriate and Andres [11)] にもとづきブートストラップ標本の個数 B を 5000，ふしの分岐に使用する変数の数 P を，モデル作成に使用された変数の数の平方根と定めた。また，なるべく複雑な樹木を作成して各樹木間の相関を減らすために，末端のふしのデータ数の最小値は 1 を指定した。

環境要因の可視化のために，部分従属プロットを作成した。[16)] 例として来遊資源量指数の水準が「高位」である場合(i=1)の部分従属性 PD_1^* の推定値を以下に示す。この場合，PD_1^* が大きい値を示すと，資源量指数の水準は「高位」になりやすくなる。

$$PD_1^* = \log\left[p_{l_1}(z_c)\right] - \frac{1}{3}\sum_{i=1}^{3}\log\left[p_{l_i}(z_c)\right] \tag{9}$$

興味のある変数が z_c であり，p_{l_i} は来遊資源量指数の水準 l_i に属する確率を表す。このように, 分類問題では来遊資源量指数の水準 l_i ごとに部分従属度が構成される。

ランダムフォレストによる変数選択　ランダムフォレストで計算された変数の重要度を使用して変数選択を行った。[11] 変数の重要度は，OOB（out-of-bag）データにおける予測精度から算出された。OOB データとは，樹木を作成する際のブートストラップ標本に含まれなかったサンプルを指す。[9] OOB データを使うことによって，モデルの作成時に使われていないデータへの予測精度を評価することができ，クロスバリデーションの代用となっている。変数の重要度は，以下のステップにより計算される。

(i) ブートストラップ標本に対して，OOB データへの予測誤差 err^{oob} を算出する。

(ii) 各 OOB データの p 番目の説明変数値 $x_{po(n)}$, n =1,2,…, N のみをランダムに入れ替えた OOB データを作成し，その時の予測誤差 $\widetilde{err_p^{oob}}$ を算出する。

(iii) 以下の計算により，説明変数 x_p の重要度を求める。

$$Imp_p = \widetilde{err_p^{oob}} - err^{oob} \tag{10}$$

なお，本研究では予測誤差として誤分類率を使用した。誤分類率は以下で定義される。

$$err^{oob} = 1 - p\left(\hat{l}_i(t)\middle|t\right) \tag{11}$$

変数選択は，以下のステップに従って実行された。[11)]

(i) 全ての変数を入れたランダムフォレストモデルを作成し，説明変数の重要度を算出する。同時に，OOB データを使用した誤分類率を求める。

(ii) 重要度が小さい説明変数を，使用された説明変数の数の 20%分だけ排除したデータセット(仮に 100 個の説明変数を使用していた場合は 20 個を排除する)を作成し，新たにランダムフォレストモデルを作成する。そして，OOB データを使用した誤分類率を算出する。

(iii) (ii) を繰り返す。なお, 重要度の過学習を避けるため, 重要度は常に(i) で算出されたものを使用する。これを説明変数が排除できなくなるまで繰り返す。

(iv) (iii) で作成されたすべてのモデルにおいて, OOB データを使用した誤分類を比較し，最も誤分類率が小さい説明変数の組み合わせを見つけ，そのモデルを採択する。

(v) また，誤分類率の標準偏差を算出し，1 標準偏差内に入る誤分類率であり，かつ最も誤分類率が小さかったモデルよりも説明変数の数が少ないモデルも候補として採択する。

この枠組みを使用することで，多くの説明変数を使用していても，その中から

予測をするのに必要となる変数だけを選び出すことが可能となる。また，誤分類率が最も小さくなる変数の組み合わせを採用するため，すくなくとも OOB データによる評価においては，全変数を使った場合よりも予測精度は高くなる。

モンテカルロリサンプリングによる予測精度の評価　予測分布は, 500 回のモンテカルロリサンプリングにより算出した。1 回のリサンプリングでは，全データ（N＝37）の 1 割にあたる 4 つのデータをテスト用データとして重複しないように抜いて，評価を行った。[19)]

計算にはすべてデータ解析環境 R version 3.0.1 (R Development Core Team, Wien, Austria)を使用した。ランダムフォレストモデルの作成には，パッケージ randomForest (http://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf, 2014 年 1 月 17 日)を用い, 変数選択にはパッケージ varSelRF[20)]を使用した。不均一性の尺度は, これらのパッケージが使用しているものをそのまま用いた。

結果

情報エントロピーを最大にする $k_{entropy}$ を探索した結果, $k_{entropy}$=0.23 と求められた。本研究では，k=$k_{entropy}$ として来遊資源量指数を水準に分離した結果を解析に用いた。このとき,「高位」水準は 12 データ,「中位」水準は 12 データ,「低位」

水準は 15 データとなった（Table 2）。 Table 2

変数選択の結果，Table 3 のようになった。前年 11 月の PNA（PNA.Nov）と前年 11 月の黒潮流路（Kurohio.path.Nov），前年 9 月の北太平洋水温（North.Pacific.Sep），前年 1 月の SOI（SOI.Jan）のみが選ばれた。各変数における推定された重要度の値は，およそ 34.5，30.3，21.6，14.9 となり，PNA が最も高くなった。 Table 3

部分従属プロットは，Fig. のようになった。水準が「高位」である時は SOI が負になっていた。それ以外の変数からは明瞭な関係は見いだせなかった。水準が「中位」である時 PNA は低く，黒潮は北偏しており，北太平洋水温はやや高い傾向が見られた。また，SOI が負である時には「中位」になりにくい傾向がみられた。水準が「低位」である時は，PNA が高く，黒潮は南偏しており北太平洋水温は低かった。SOI と「低位」には明瞭な関係が見いだせなかった。

Fig. 2

500 回ブートストラップによる予測精度評価の結果は，Table 4 のようになった。的中率はおよそ 62%であり，現行の予報（1972 年から 2009 年まで）の的中率である 58%[20]をやや上回った。

Table 4

考察

予報の的中率は 62%であり，現況の予報とほぼ同じかやや高い値となった。本手法は人間の勘や経験に基づかずに予測を出すことができる。また，サンマ漁業がおこなわれる以前に手に入る環境要因しか予報に使わず，東北区水産研究所が実施している漁期前調査の結果[12)]も参照していない。現行の予報は前年比を予測しているため，資源量水準を予測している本研究の予測とは形態が異なる。そのため厳密な比較を行うことは難しいと考えられるが，それでも現行の予報と比べて高い精度を有していた。

漁期前調査は 2003 年から開始されているためデータが少なく，モデルを作成する際には使用しなかった。今後調査データが蓄積されてモデルに組み込むことができれば，予測精度が向上する可能性もあるだろう。データが増えることによる予測精度の向上は今後の検討課題としたい。また，本研究で使用した来遊資源量指数は漁場の広さも考慮した指数であり，厳密な来遊量と乖離する可能性もある。データを整理し，より漁業者の意思決定に寄与できる応答変数を選ぶことも，今後の課題となろう。

本研究では，スプライン補間を行ったデータを一部使用した。補間された年数は多くないので大きな影響があるとは考えにくいが，今後欠損データを使用する必要がなくなるほどにデータが蓄積されれば，こういった前処理がもたら

す影響を評価することも可能となるだろう。

変数選択の結果選ばれたものは，PNA，北太平洋水温，黒潮流路，SOI の 4 種類だった。PNA は日本南東方海域の海面水温偏差に影響することが知られており，[21] 北太平洋水温とともにサンマの来遊量には水温が関わっていることが示唆された。黒潮流路も関与していた。栗田 [22] によると，秋季は黒潮域と混合域ともに産卵場となっており，黒潮域で一尾あたり産卵数が多くなっている。このため，サンマの再生産に黒潮の位置が関わっている可能性がありうる。SOI が低ければ来遊量は「高位」になりやすい傾向がみられた。田ら [15] によると，エルニーニョが発達している時期には，サンマ来遊量は増加することが示唆されており，本研究の結果もそれを支持した。

ランダムフォレストを用いることで，大量データの圧縮が容易になった。説明変数に 186 もの変数を使用したが，最終的に選ばれたのは 4 つの変数のみであった。未知データへの的中率を最大とするように変数選択されているため，モデル作成時のデータへ過剰適合しにくくなるように最適化された結果，変数が少なくなったものと考えられる。多くのデータの中から必要な変数を選ぶ際に，本手法は有用といえよう。しかし，乱数発生アルゴリズムを多用している関係上，変数選択の結果は頑健とは言い難い。[11] 本研究では，予測精度をモンテカルロリサンプリングにより正確に評価をすることに努めたが，あまり頑健

でないモデルである以上，選ばれたモデルに対してこのような厳密な評価手法をとることは重要と考えられる。また，ブートストラップによる評価を複数回行うことで，ブートストラップ結果の精度を検証することもできるだろう。

本研究ではランダムフォレストをカテゴリデータへの分類問題へ適用したが，回帰問題にも適用可能である。来遊量を連続変数として使用した方が，情報の損失を避けることが可能であるかもしれない。しかし，ランダムフォレストは分類問題において相対的に高い性能を示すことが知られている。[10] また，カテゴリ予測であれば，予測分布をモンテカルロリサンプリングなどのノンパラメトリックな手法において評価することが容易である。仮に連続データへの予測であれば，「来遊資源量指数が少ないと予報された時には過少推定されていることが多く，来遊資源量指数が中位程度と予報された時には過大推定されていることが多い」といった特徴をつかむことはより困難となるであろう。定量的な予測にするか，カテゴリ予測にするかは目的に応じて変えることが望ましいが，この点においてカテゴリ予報の方が定量的な予報よりも優れていると考えられる。

本研究ではサンマの来遊資源量指数の水準を予測するためにランダムフォレストを使用したが，ランダムフォレストと同じく CART 法を使用した樹木構造接近法の一種であるアダブースト [16]などの使用も検討できるだろう。アダブー

ストはランダムフォレストのように多数の CART 樹木を作成するが，ブートストラップ標本を作製する際に違いがみられる。すなわちランダムフォレストやバギングは過去に作成された樹木の影響を受けることのない独立な樹木をブートストラップ標本により作製する。一方アダブーストに代表されるブースティングは，誤分類率の大きさで重み付けをしてリサンプリングを繰り返すという特徴がある。学習能力を向上させるためにアダブーストなどのブースティングを使用した来遊量資源量指数の水準を予測するモデルの作成は今後の検討課題としたい。また，多くの手法を使用した際は，予測モデルの有意性検定や，予測モデルの経済価値の評価など新たな評価方法を導入して比較検討する必要があるだろう。

調査船の情報を用い，多くのベテランの研究者の意見を用いた現行の予測結果と同じかそれ以上の精度を出せたことは，本研究で用いたランダムフォレストの優秀性を示す結果と考えられる。また，数多くの環境要因を圧縮して少ない説明変数だけを使用したモデルを作成することができたこともランダムフォレストの成果と考えられる。変数選択で選ばれた説明変数は全て漁期の前年に入手できるものであるため,操業がおこなわれる年の 1 月にはデータがそろい,早期に予報を出すことも可能となるであろう。ランダムフォレストはノンパラメトリックな手法であるため，予測結果を計算する式を示すことは難しいが，

作成されたモデルに各説明変数の数値を代入することで，容易に予測結果を返すことができる。機械的な予測を行うことで，予報を出す人員の交代などの影響が少なくなることも，本手法を用いることの大きなメリットと言えるだろう。

予報を用いた意思決定を行う際には，Table 4 のような分割表形式で予報の精度が評価されていることが重要である。[23] 新しい予測手法を適用する際，厳密な精度評価をしたうえで，分割表形式での予報の精度を公開することで，操業計画を策定する際の手助けとなる情報を提供できるだろう。

謝辞

研究を進めるにあたり有益なコメントを賜った東北区水産研究所巣山哲博士，中神正康博士，柴田泰宙博士，統計数理研究所島谷健一郎准教授，北海道大学大学院水産科学研究院桜井泰憲教授，和田哲准教授に感謝の意を表します。

文献

1) 青木一郎, 小松輝久. ニューラルネットによるマイワシ未成魚漁獲量の予測. 水産海洋研究 1992; **56**: 113-120.

2) Watanabe K, Tanaka E, Yamada S, Kitakado T. Spatial and temporal migration modeling for stock of Pacific saury *Cololabis saira* (Brevoort), incorporating effect of sea surface temperature. *Fish. Sci*. 2006. **72**: 1153-1165.

3) Rupp DE, Wainwright TC, Lawson PW, Peterson WT. Marine environment-based forecasting of coho salmon (*Oncorhynchus kisutch*) adult recruitment. *Fish. Oceanogr.* 2012; **21**: 1-19.

4) 為石日出男，花岡明，四宮博．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 1997; **61**: 18-22.

5) 湯祖恪，桜本和美，和田時夫，北原武，原田泰志．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 1992; **58**: 1873-1881.

6) 青山恒雄.「短期漁海況予測手法開発検討委員会」における課題と展望．水産海洋研究 1984; **46**: 27-53.

7) 高杉知．聞き取り調査で得られたサンマ漁海況予報の利用状況．サンマ等小型浮魚資源研究会議報告．青森．1989; **37**:262-268.

8) 渡邉一功，斉藤克弥，為石日出男，小坂淳．時系列解析によるサンマ漁場の

水温分布予測．水産海洋研究 1999; **63**: 61-67.

9) Breiman L. Random forests. *Mach. learn*. 2001; **45**: 5-32.

10) 杉本知之，下川敏雄，後藤昌司．樹木構造接近法と最近の発展．計算機統計学 2007; **18**: 123-164.

11) Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. *BMC bioinformatics* 2006; **7**: 3.

12) 第 60，61 回サンマ等小型浮魚資源研究会議報告．東北区水産研究所資源海洋部．2012, 2013.

13) Baba S, Matsuishi T. Evaluation of the predictability of fishing forecasts using information theory. *Fish. Sci*. 2014; **80**: 427-434.

14) Shannon CE. A mathematical theory of communication. *Bell. Syst. Tech. J* 1984; **27**:379-423,623-656.

15) 田永軍，赤嶺達郎，須田真木．北西太平洋におけるサンマ資源の長期変動特性と気候変化．水産海洋研究 2002; **66**: 16-25.

16) 下川敏雄，杉本知之，後藤昌司．「樹木構造接近法：R で学ぶデータサイエンス 9」共立出版，東京．2013.

17) Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Wadsworth & Brooks, Monterey, 1984.

18) Breiman L. Bagging predictors. *Mach. learn*. 1996; **24**: 123-140.

19) Lee YW, Megrey BA, Macklin SA. Evaluating the performance of Gulf of Alaska walleye pollock (*Theragra chalcogramma*) recruitment forecasting models using a Monte Carlo resampling strategy. *Can. J. Fish. Aquat. Sci*. 2009; **66**: 367-381.

20) Diaz-Uriarte R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC bioinformatics* 2007; **8**: 328.

21) Hanawa K, Watanabe T, Iwasaka N, Suga T and Toba Y. Surface thermal conditions in the western North Pacific during the ENSO events. *J. Meteorol. Soc. Jpn.* 1988; **66**: 445-456.

22) 栗田豊．サンマの産卵場及び産卵量の季節変化．サンマ資源研究会議報告．青森．2001; **49**: 203-205.

23) 立平良三．「気象予報による意思決定 不確実情報の経済価値」東京堂出版，東京．1999.

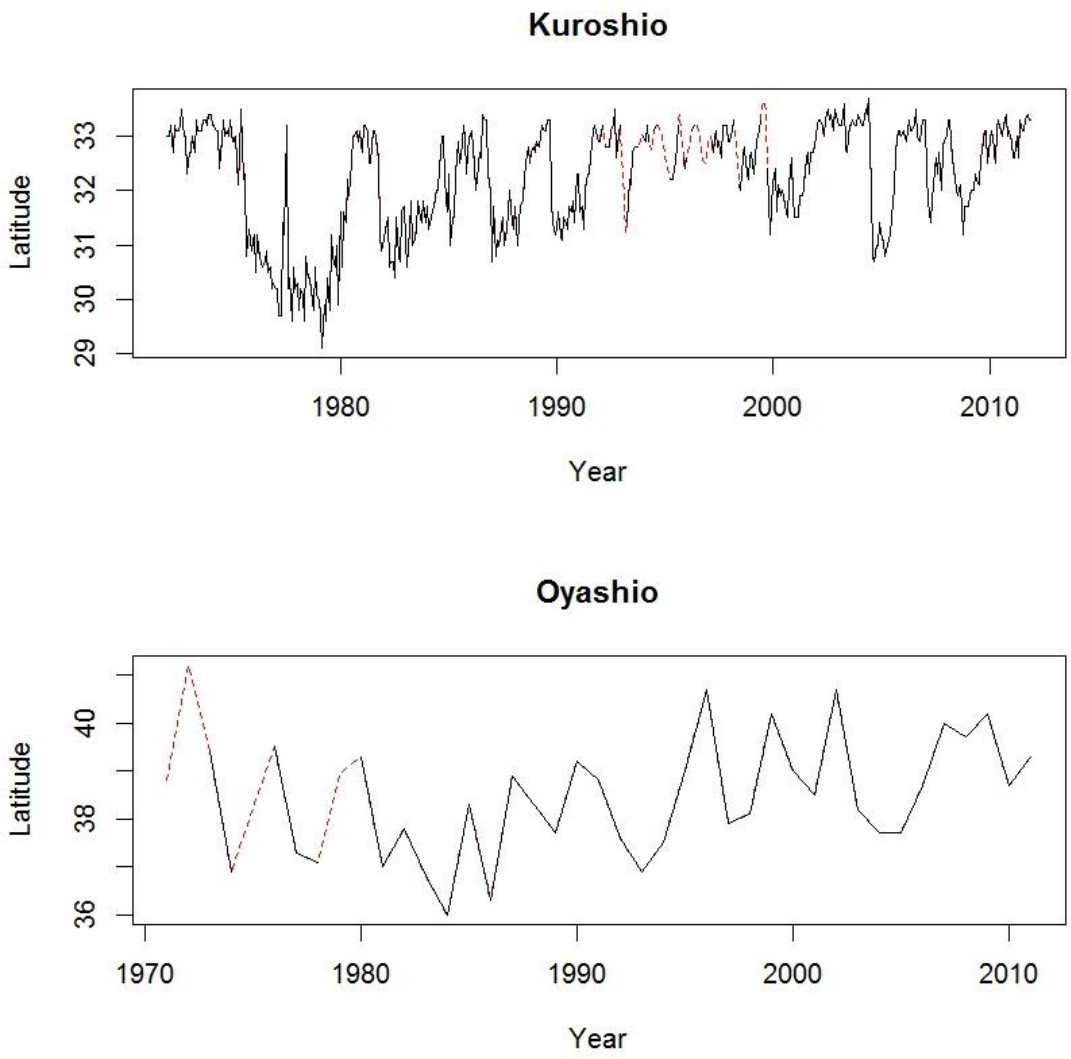
1

2

図説明

Fig. 1 Result of interpolation. A broken line is interpolated data. The definition of each indices were defined by Japan Meteorological Agency (http://www.jma.go.jp/jma/, 10 July 2012)

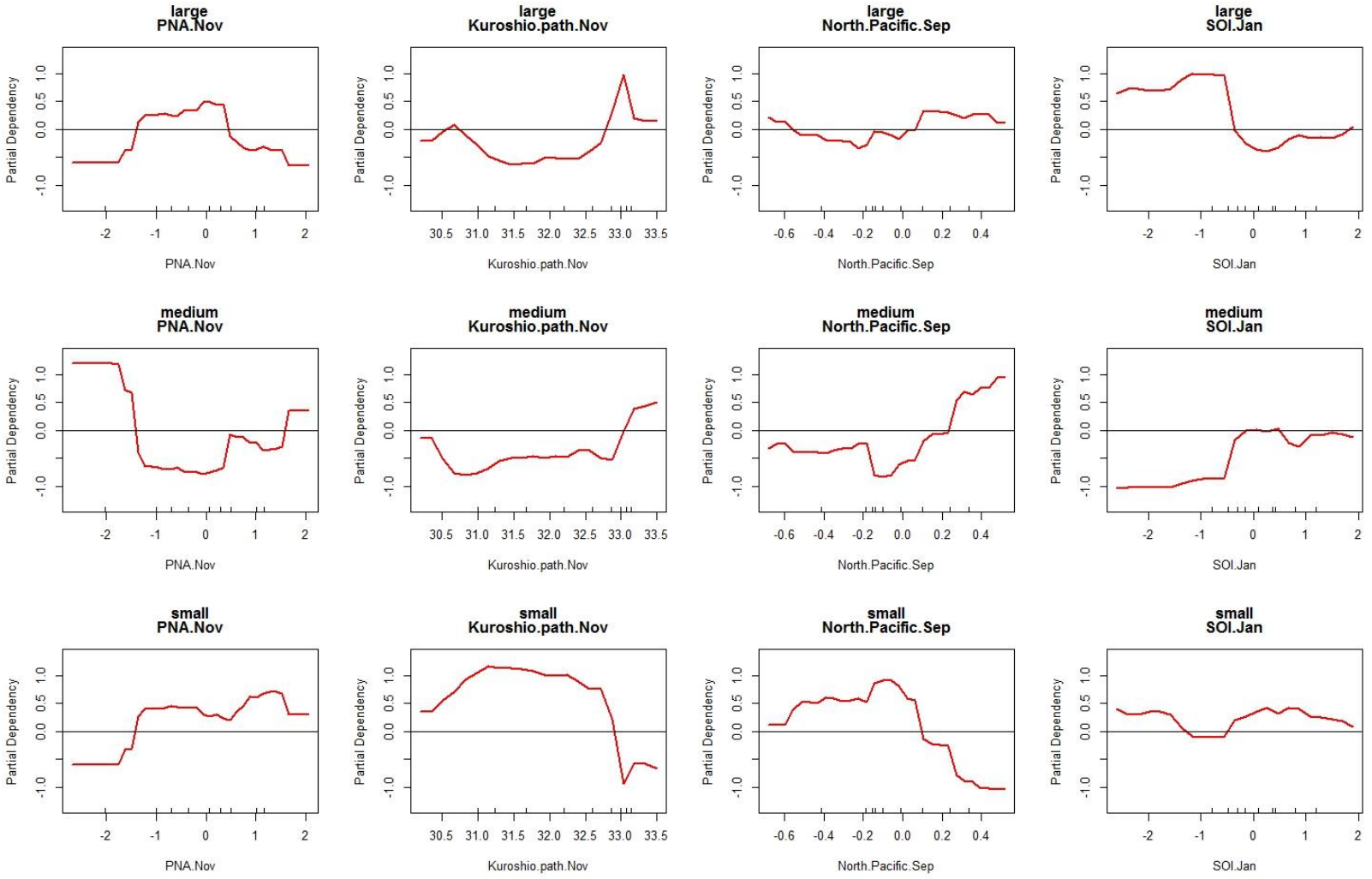
Fig. 2 Partial dependency plot. Partial dependency was calculated by Eq(9). Partial dependency expresses relative frequency of each fish abundance level. PNA.Nov means PNA index in November of the previous year. Kuroshio.path.Nov means interpolated Kuroshio path in November of the previous year. North.Pacific.Sep means SST in north Pacific in September of the previous year. SOI.Jan means SOI in January of the previous year. The definition of each indices were defined by Japan Meteorological Agency (http://www.jma.go.jp/jma/, 10 July 2012), and Climate Prediction Centre (http://www.cpc.ncep.noaa.gov/, 26 July 2012).


Kuroshio
Latitude
29
30
31
32
33
1980
1990
2000
2010
Year
Oyashio
Latitude
36
38
40
1970
1980
1990
2000
2010
Year

15

16 Fig. 1

17


large
PNA.Nov
large
Kuroshio.path.Nov
large
North.Pacific.Sep
large
SOI.Jan
medium
PNA.Nov
medium
Kuroshio.path.Nov
medium
North.Pacific.Sep
medium
SOI.Jan
small
PNA.Nov
small
Kuroshio.path.Nov
small
North.Pacific.Sep
small
SOI.Jan
Partial Dependency
PNA.Nov
Kuroshio.path.Nov
North.Pacific.Sep
SOI.Jan

18

19 Fig. 2

Table 1 List of data set which used in Random Forest

No	Names	Period	*n*	Note
1	PDO: Pacific Decadal Oscillation (annual average)	previous year	1	
2	SOI: Southern Oscillation Index (monthly average)	previous year and until this year June	18	
3	sea temperature of subtropical mode water in north Pacific (annual average)	previous year	1	
4	area of cross section of tropical water in north Pacific (seasonal average)	Summer and Winter in previous year and Winter in this year	3	
5	NPI: North Pacific Index (annual average)	previous year	1	
6	SST in north Pacific (monthly average)	previous year and until this year June	18	
7	SST in Kushiro area (seasonal average)	Winter, Spring, Summer and Autumn in previous year and Winter and Spring in this year	6	
8	SST in Sanriku area (seasonal average)	Winter, Spring, Summer and Autumn in previous year and Winter and Spring in this year	6	
9	SST in eastern Kanto area (seasonal average)	Winter, Spring, Summer and Autumn in previous year and Winter and Spring in this year	6	

10	SST in southern Kanto area (seasonal average)	Winter, Spring, Summer and Autumn in previous year and Winter and Spring in this year	6	
11	Kuroshio path (monthly average)	previous year and until this year June	18	interpolated
12	the flow of Kuroshio (seasonal average)	Summer and Winter in previous year and Winter in this year	3	
13	area of Oyashio (annual average)	previous year	1	
14	the southern limit of Oyashio (annual average)	previous year	1	interpolated (data in 1971 was used for interpolation)
15	TNH: Tropical North Hemisphere (monthly average)	January, February and December in previous year and January and February in this year	5	
16	AO: Arctic Oscillation (monthly average)	previous year and until this year June	18	
17	PNA: Pacific - North American pattern (monthly average)	previous year and until this year June	18	
18	WP: West Pacific (monthly average)	previous year and until this year June	18	
19	EP-NP: East Pacific - North Pacific (monthly average)	previous year and until this year June (without December)	17	
20	PT: Pacific Transition	previous year (only August and	2	

	(monthly average)	September)	
21	PE: Polar/ Eurasia	previous year and until this year June	18
	(monthly average)		

Note: The data from 1 to 14 lines were from Japan Meteorological Agency (http://www.jma.go.jp/jma/, 10 July 2012). The data from 15 to 21 lines were from Climate Prediction Centre (http://www.cpc.ncep.noaa.gov/, 26 July 2012). *n* was number of the kinds of each data.

Table 2 Separated fish abundance index

Year	Abundance	Year	Abundance
1972	medium	1992	medium
1973	large	1993	large
1974	medium	1994	large
1975	large	1995	medium
1976	small	1996	no-data
1977	medium	1997	large
1978	large	1998	small
1979	small	1999	small
1980	small	2000	small
1981	small	2001	medium
1982	small	2002	small
1983	small	2003	medium
1984	small	2004	medium
1985	small	2005	large
1986	medium	2006	medium
1987	small	2007	large
1988	large	2008	large
1989	small	2009	medium
1990	small	2010	large
1991	medium	2011	large

Note:

Table 3 Result of variable selection

Variable names	Importance index
PNA.Nov	34.5
Kuroshio.path.Nov	30.3
North.Pacific.Sep	21.6
SOI.Jan	14.9

Note: Importance index was calculated by Random Forest. PNA.Nov means PNA index in November of the previous year. Kuroshio.path.Nov means interpolated Kuroshio path in November of the previous year. North.Pacific.Sep means SST in north Pacific in September of the previous year. SOI.Jan means SOI in January of the previous year. The definition of each indices were defined by Japan Meteorological Agency (http://www.jma.go.jp/jma/, 10 July 2012), and Climate Prediction Centre (http://www.cpc.ncep.noaa.gov/, 26 July 2012).

35 Table 4 Result of forecast evaluation by using Monte Carlo resampling

		Actual fishing level			
		Large	Medium	Small	Sum
Forecast	Large	436	98	143	677
	Medium	55	251	115	421
	Small	138	212	552	902
	Sum	629	561	810	2000

36

37