

修 士 論 文

題 目 時系列データを使った漁獲モ
デル開発に関する研究

指導教官 小山田 耕二 教授

京都大学大学院工学研究科修士課程 電気工学専攻

氏 名 高見 円仁

平成28年2月3日

目次

第1章 序論	1
第2章 関連研究	3
2.1 漁場予測に関する研究	3
2.2 HSI モデル	3
2.3 漁場形成過程に関する研究	4
2.4 HSI モデルの構築手法に関する研究	5
第3章 手法	7
3.1 使用データ	7
3.1.1 海洋環境変数データ	7
3.1.2 過去のデータの計算	7
3.1.3 ルンゲクッタ法	8
3.1.4 漁獲実績データ	10
3.2 HSI モデルから漁獲量予測への変換	11
3.3 SI モデル	11
3.4 オーバーフィッティング	13
3.5 提案手法のアルゴリズム	17
3.6 IRLS	18
3.7 最適な罰則項の重みの決定	20
3.8 正答率	20
3.9 HSI モデル構築	20
3.10 別の HSI モデル構築手法	21
第4章 システム概要	23
4.1 システム概要	23
4.2 流跡線計算システム	24
4.3 最適化計算システム	25

4.4	地図描画システム	28
第 5 章	実験	32
5.1	使用データ	32
5.1.1	漁獲実績データ	32
5.1.2	海洋環境変数データ	32
5.2	過去のデータを用いることの有効性を確認する実験	33
5.2.1	実験方法	33
5.3	提案手法の有効性を確認する実験	34
5.3.1	実験方法	34
第 6 章	結果及び考察	35
6.1	実験結果及び考察	35
第 7 章	結論	36
7.1	結論	36
7.2	課題	36
	謝 辞	37
	参 考 文 献	38

第1章 序論

近年、日本南西部の気候が亜熱帯化しつつあると言われているように日本では温暖化が進んでいる現状がある。温暖化に代表される気候変動の影響により集中豪雨や土砂災害などだけではなく、日本近海でこれまで見られなかった漁場の変化が起こっている。その適応策として年単位のスケールでの水産資源変動推定が研究されてきている。しかし漁業ではそのような中長期的な予測だけではなく、日ごとに変動するような短期的な漁場予測が必要とされる¹⁾。日ごとに変化する短期的な漁場予測を行うために水産生物の資源分布予測モデルを構築する試みがいくつかある。そのようなモデルの一つとして、好的生息域 (HSI : Habitat Suitability Index) モデルが知られている²⁾。HSI モデルではまず、対象魚種の漁獲実績 (Catch Per Unit Effort) と各海域での海洋環境変数の関係を数式化した SI (Suitability Index) モデルを作成する。この SI モデルによって海水温や塩分濃度や流速などといった海洋環境変数ごとに、対象魚種にとっての生息しやすさが数値化される。海洋環境変数の数だけ作成された SI モデルを一つの式に統合することで、各海域がどの程度生息しやすいかを定式化した HSI モデルが構築できる。従来の研究では、深さの自由度を持つ各海洋環境変数と漁獲実績を散布図で表示し、各変数について一つ選択した深さの変数を平滑化スプラインで多項式近似することで各 SI モデルを構築する。そしてそれらを統合することで HSI モデルを構築している^{3) 4)}。

従来手法ではある一日の漁場予測を行う際、その日の海洋環境変数のみで SI モデルを構築している。しかし、実際の漁場は漁獲日当日の海洋環境だけではなく、漁獲日から遡った過去の海洋環境も影響するような因果関係が存在する。そのため、SI モデル構築に利用する海洋環境変数として、漁獲日当日の海洋環境変数データだけではなく、漁獲日から遡った過去の海洋環境変数データを利用したい。一方、従来行われてきた各 SI モデル構築ではパラメータをいくつか設定する操作を行うが構築したモデルについてオーバーフィッティングの問題があるので、適切なモデル構築ができていのかどうかを人が判断しモデル構築のためのパラメータを人が設定する必要がある。過去の海洋環境変数データを SI モデル構築に利用する場合、このパラメータ決定作業

は利用者の負担となる問題がある。

そこで本研究では、計算機を用いた HSI モデルの自動生成手法を提案する。またいくつかの実験を行い、過去のデータを利用することが漁場予測精度の向上につながることを確認し、提案手法が従来手法より精度よく漁場予測ができることを確認し、提案手法が有効であると結論付けた。

各章は以下のように構成される。2 章では本研究での関連研究を挙げ、3 章では従来手法及び提案手法について説明する。4 章ではプログラムをどのように実装したか説明し、5 章で実験を行い、6 章では考察を行う。7 章で本研究の結論と今後の展望について述べる。

第2章 関連研究

本章では漁場予測技術に関する研究及びその可視化技術に関する先行研究を述べる。また本研究では漁場予測の対象生物としてアカイカを用いた。そのアカイカの漁場形成に関する研究についても述べる。

2.1 漁場予測に関する研究

P.Lehodeyらはカツオやマグロ類の分布を予測するSEAPODYM(Spatial ecosystem and population dynamics model)を構築した¹⁰⁾。このモデルは海洋物理環境モデル、餌生物モデル、移動回遊モデルから構成されたモデルである。特徴として魚類の移動に影響する水温、流速、溶存酸素、基礎生産等の海洋環境要因を考慮した分布予測モデルである。

根本らはアカイカの卵稚仔が産卵から1ヶ月ほどは自走できず海流によって浮遊していることに着目し、その追跡シミュレーションシステムを構築した¹¹⁾。卵稚仔を粒子とみなし、海洋の流速データを用いて粒子を追跡シミュレーションしている。また単に追跡するだけではなく、海洋環境のトレースを行い卵稚仔が生存しているかどうかもシミュレーションし、それをもとにした漁獲量予測を行うことも可能である。

2.2 HSI モデル

HSI (Habitat Suitability Index) モデルとは、対象生物の生息地としての適性を0から1で数値化するモデルであり²⁾、各環境変数から算出される。HSIは一般に、単一の環境変数から算出されるSI (Suitability Index) を統合した値として計算される。単一の環境変数で生息地としての適性を0から1で数値化したモデルであるSIモデルを統合することによって、複数の環境変数で生息地としての適性を0から1で数値化したモデルであるHSIモデルを構築する。その統合手法には主に、算術平均法、幾何平均法、限定要因法、加算要因法がある。算術平均法は各SIモデルが算出した0から1の数値を相加平均

したもので SI 値が全て算出できなくても HSI 値を計算可能である。幾何平均は各 SI 値を相乗平均したものであり、複数の SI 値が対象生物の生息に適さない領域での値が小さくなる傾向にある。限定要因法は各 SI 値の中で最小値の SI 値を HSI 値とする統合手法である。植物の成長は必要とする栄養素の中で最も少ないものに影響されるというリービッヒの最小律¹²⁾と同じ概念で構築される。加算要因法は各 SI 値を足したもので算術平均法と違い、HSI 値が 1 より大きくなると 1 に制限される。このように HSI モデルによる数値化の目的によって各統合手法を使い分ける必要がある。

Tian らは本研究と同じくアカイカを対象とした HSI モデルを構築したが、その際に幾何平均法を用いている¹³⁾。また、五十嵐らもアカイカを対象魚種として同様に幾何平均法で HSI モデルを構築している⁴⁾。それらを踏襲し、本研究では幾何平均法を採用する。

HSI モデルはアサリ¹⁴⁾、メバル¹⁵⁾ など他の海洋生物だけでなく、オオタカ¹⁶⁾ などの陸上生物、海草¹⁷⁾ ¹⁸⁾ などの植物にも適用事例があり、広く一般に使われている指標である。また、対象生物にとっての生息しやすさを数値化することから、開発事業によってどれだけ生息しやすさが変化するかの環境アセスメントの評価¹⁹⁾ や生物多様性の評価²⁰⁾ にも使われる。

対象生物の生息地として適しているということと、対象生物の数は比例関係にあると仮定すると、HSI モデルは対象生物の多さを数値化したモデルであるとみなすことができる。

2.3 漁場形成過程に関する研究

本研究で漁場予測の対象とするアカイカは亜熱帯海域、温帯海域に広く分布するイカである。⁵⁾ 亜熱帯海域、温帯海域は太平洋ではカムチャッカ半島から台湾付近の南北幅、日本からアメリカまでの東西幅がある広い領域である。アカイカについての漁場形成過程は鈴木が過去に研究している⁶⁾。鈴木は日本近海の太平洋側におけるアカイカの魚群が暖流（黒潮）と寒流（親潮）の境目に集まる可能性について述べた。親潮は一般に栄養塩の濃度が高い。栄養塩とは窒素、リン、珪素の無機塩類の総称である⁷⁾。光と栄養塩は植物プランクトンの成長増殖にとって重要な因子である。その栄養塩を消費する植物プランクトンが親潮には少ないため栄養塩の濃度が高い。栄養塩は沈降す

るので特に海の深層に堆積する。

親潮と黒潮がぶつかる際に渦が発生し、その流れに乗って深層の栄養塩が深海から巻き上げられる。巻き上げられた栄養塩は深海に比べて太陽光が多い深度で植物プランクトンに消費され、その数が増える。植物プランクトンが増えることで、植物プランクトンを餌とする動物プランクトンも増える。さらに、動物プランクトンを餌とする小魚が集まり、小魚を餌とするアカイカが集まる。このような食物連鎖が起こることでアカイカの魚群が暖流と寒流の境目に集まるとされている。

栄養塩が巻き上げられてからアカイカが集まるまで 60 日かかるという専門家の意見がある。そのため、本研究では漁獲日から最大 60 日まで過去に遡ったデータを使用する。

村田らは北太平洋におけるアカイカの季節回遊性についてまとめた⁸⁾。アカイカは季節的に南北に回遊している。アカイカの産卵場は南の亜熱帯海域である。この海域で孵化したアカイカは 7 月頃に索餌域である亜寒帯海域付近まで北上し大きくなる。そして秋に再び亜熱帯海域へ南下し産卵する。また、アカイカを追跡することで日中は 300m 付近の深さにいて夜間は水面近くへ浮上するという、一日の中でも鉛直方向の周期的な移動を行うことがわかっている⁹⁾。漁場予測の研究では古くから、このような魚の生態をモデル化し漁場予測に取り入れる試みが行われてきた。

2.4 HSI モデルの構築手法に関する研究

漁師の経験や勘が示す場所が良い漁場であると考え、CPUE を予測するのではなく漁船がどこに集まるのかを予測するという研究も存在する。Tian らはアカイカの漁獲量予測を目的として塩分濃度や海水温などの海洋環境変数から HSI モデルを 2 種類構築した。それぞれ CPUE を予測するための HSI モデルと漁船がどこに集まるのかを予測するための HSI モデルである。そしてそれらの比較を行い、目的変数の異なるそれぞれのモデルによる傾向について述べている¹³⁾。

Chen らは環境変数に重みを付けて和を取る HSI モデルによって CPUE が説明できることを説明し、各海洋環境変数に対する重みの変化が HSI に与える影響を評価した²¹⁾。

櫛田らは HSI モデル構築のための支援システムを開発した³⁾。このシステムではシミュレーションから得られた海水温、塩分濃度、経度方向の流速、緯度方向の流速、深さ方向の流速、の 5 変数それぞれに SI モデルを構築し、その相乗平均を HSI モデルとしている。また、各 SI モデルを単一変数の平滑化スプラインによる多項式近似で構築している。海洋環境変数から漁獲実績を誤差なく説明する SI モデルを構築することもできるが、この場合オーバーフィッティングの問題から漁場予測の精度が悪くなる傾向にある。そのため、このシステムではモデルの滑らかさ (平滑化パラメータ) を設定できるようにしている。利用者は SI モデルを確認し、最適な滑らかさの係数を選択する必要がある。この HSI モデルでは漁獲日当日のデータのみを用いている。

本研究では、この負担をなくしさらに漁獲量と海洋環境変数の間の因果関係についてもモデル構築に組み込んだ手法を提案する。

第3章 手法

本章では、データの計算法や従来の HSI モデル構築手法及び提案する HSI モデル構築手法などの基礎技術について述べる。

3.1 使用データ

HSI 値を求めるためには海洋環境変数データ、対象魚種の漁獲実績データが必要となる。本研究で利用したこれらのデータについて述べる。

3.1.1 海洋環境変数データ

本研究で使用する海洋環境変数データは、気象庁気象研究所で開発された海洋データ同化システムで作成された MOVE (MRI Multivariate Ocean Variational Estimation) データである。この海洋データ同化システムは数値海洋モデル及び同化システムから成るシステムである²²⁾。気象研究所共用海洋モデル MRI.COM²³⁾ を用いて計算したシミュレーションデータと実測データを同化することで MOVE データは計算された。MOVE データでは海水温、塩分濃度、3次元方向の流速が緯度経度 0.1° ごとに計算されている。また、深さは全 54 段階で最大 6000m まで計算されている。MOVE データへのアクセスは OPeNDAP プロトコル²⁴⁾ を用いている。OPeNDAP は HTTP 経由でデータアクセスを行うためのプロトコルであり地球科学の分野でよく使われる。本研究で利用する MOVE データの場合は変数名、日付番号、深さインデックス、緯度経度の配列範囲を指定しリクエストすることで、緯度経度の配列番号と実際の緯度経度の対応、指定範囲の変数データなどが返ってくる。

3.1.2 過去のデータの計算

本研究では過去の海洋環境変数を利用することを提案している。ある地点での漁獲を過去の海洋環境変数から説明する場合その地点と海洋環境変数の間に時空間的な繋がりと考えられる。そのためその地点の海水がどこ

から来たものなのか計算し、その経路上でサンプリングした海洋環境変数のみを使用すべきである。本研究で使用する海洋環境データは3次元方向の速度も含んでいるので、海水を遡って計算することができる。そして深さごとに異なる速度を持っているので、それぞれの深さをスタート地点として遡ると全く別の地点にたどり着く。今回はより正確に過去の地点を遡るために4次のルンゲクッタ法を用いた。

3.1.3 ルンゲクッタ法

ルンゲクッタ法は常微分方程式の数値解を求める手法の一種である。同じように常微分方程式の数値解を求める手法として、オイラー法がある。海洋環境変数を使用することである地点から流れを遡る場合、その地点の速度にステップの時間幅をかけることで過去の地点を求める手法がオイラー法である（式3.1）。

$$(x^{(t-1)}, y^{(t-1)}, z^{(t-1)}) = (x^t, y^t, z^t) - (u^t, v^t, w^t) \cdot \Delta t \quad (3.1)$$

遡っているので速度が負になっている。オイラー法では速度はそのままその地点の速度を用いて、ステップの時間幅で速度変化がないものとして扱っている。しかし、実際には微小時間ごとに速度変化は起きている。ルンゲクッタ法はオイラー法の式である式3.1における u^t と v^t と w^t をより正確に求める手法である。そのため、速度変化が一切ない領域ではどちらの手法も同じ結果となるが、本研究で使用する海洋環境変数データはそのような領域がないため、ルンゲクッタ法による精度向上が期待できる。

4次のルンゲクッタ法は次のように計算する。

1. k1 から k4 を計算する
2. k1 から k4 を重みを付けて平均化する
3. それを速度としてステップの時間幅 Δt をかけることで、移動量とする
4. もとの位置に移動量を足すことで解とする

k1 は式3.2で計算する。

$$k1 = \mathbf{f}(\mathbf{X}, t) \quad (3.2)$$

ただし $\mathbf{f}(\mathbf{X}, t) = (x, y, z)$ であり、 $\mathbf{f}(\mathbf{X}, t)$ は時刻 t 、座標 (x, y, z) における速度ベクトルである。

$k2$ は式 3.3 で計算する。

$$k2 = \mathbf{f}(\mathbf{X} + k1 \frac{\Delta t}{2}, t - \frac{\Delta t}{2}) \quad (3.3)$$

$\mathbf{f}(\mathbf{X} + \frac{\Delta t}{2} k1) = (x + \frac{\Delta t}{2} k1_u, y + \frac{\Delta t}{2} k1_v, z + \frac{\Delta t}{2} k1_w)$ であり、 $\mathbf{f}(\mathbf{X} + \frac{\Delta t}{2} k1, t - \frac{\Delta t}{2})$ は時刻 $t - \frac{\Delta t}{2}$ 、座標 $(x + \frac{\Delta t}{2} k1_u, y + \frac{\Delta t}{2} k1_v, z + \frac{\Delta t}{2} k1_w)$ における速度ベクトルである。

$k3$ は式 3.4 で計算する。

$$k3 = \mathbf{f}(\mathbf{X} + k2 \frac{\Delta t}{2}, t - \frac{\Delta t}{2}) \quad (3.4)$$

$k4$ は式 3.5 で計算する。

$$k4 = \mathbf{f}(\mathbf{X} + k3 \Delta t, t - \Delta t) \quad (3.5)$$

$k1$ から $k4$ を用いて、ステップ時間前の座標は次式のようにになる。

$$\mathbf{X}^{(t-1)} = \mathbf{X}^t - \frac{\Delta t}{6} (k1 + 2k2 + 2k3 + k4) \quad (3.6)$$

本研究ではより正確に流跡線計算を行うために、8時間ごと（一日に3回）にルンゲクッタ法で流跡線計算を行う。この場合、式 3.3、式 3.4、式 3.5、式 3.6 での Δt は $\frac{1}{3}$ になり、これを3回計算することで一日分の流跡線が計算できる。

このようにして計算した流跡線上で海洋環境変数を一日ごとにサンプリングする。サンプリングは線形補間で行う。今回は3次元の座標 (x, y, z) での海洋環境変数値 $P(x, y, z)$ を線形補間するので近傍の8点 (X_i, Y_i, Z_i) , (X_{i+1}, Y_i, Z_i) , (X_i, Y_{i+1}, Z_i) , (X_{i+1}, Y_{i+1}, Z_i) , (X_i, Y_i, Z_{i+1}) , (X_{i+1}, Y_i, Z_{i+1}) , (X_i, Y_{i+1}, Z_{i+1}) , $(X_{i+1}, Y_{i+1}, Z_{i+1})$ とそこでの海洋環境変数値 $P(X, Y, Z)$ を使って次の式で補間できる。

$$\begin{aligned}
P_{y-z-} &= P(X_i, Y_i, Z_i) + \frac{x - X_i}{X_{i+1} - X_i} (P(X_{i+1}, Y_i, Z_i) - P(X_i, Y_i, Z_i)) \\
P_{y+z-} &= P(X_i, Y_{i+1}, Z_i) + \frac{x - X_i}{X_{i+1} - X_i} (P(X_{i+1}, Y_{i+1}, Z_i) - P(X_i, Y_{i+1}, Z_i)) \\
P_{z-} &= P_{y-z-} + \frac{y - Y_i}{Y_{i+1} - Y_i} (P_{y+z-} - P_{y-z-}) \\
P_{y-z+} &= P(X_i, Y_i, Z_{i+1}) + \frac{x - X_i}{X_{i+1} - X_i} (P(X_{i+1}, Y_i, Z_{i+1}) - P(X_i, Y_i, Z_{i+1})) \\
P_{y+z+} &= P(X_i, Y_{i+1}, Z_{i+1}) + \frac{x - X_i}{X_{i+1} - X_i} (P(X_{i+1}, Y_{i+1}, Z_{i+1}) - P(X_i, Y_{i+1}, Z_{i+1})) \\
P_{z+} &= P_{y-z+} + \frac{y - Y_i}{Y_{i+1} - Y_i} (P_{y+z+} - P_{y-z+}) \\
P(x, y, z) &= P_{z-} + \frac{z - Z_i}{Z_{i+1} - Z_i} (P_{z+} - P_{z-})
\end{aligned}$$

ルンゲクッタ法で補間を行う場合はさらに時間の補間も必要となる。これも先の式と同様にして次式で計算できる。

$$P(x, y, z, t) = P(x, y, z, T_i) + \frac{t - T_i}{T_{i+1} - T_i} (P(x, y, z, T_{i+1}) - P(x, y, z, T_i)) \quad (3.7)$$

3.1.4 漁獲実績データ

本研究で使用する漁獲実績データは、日本近海の太平洋沖におけるアカイカの漁獲実績データある。漁獲実績データには日付、緯度、経度、漁獲量、操業開始時間、操業終了時間、CPUE（Catch Per Unit Effort）値が記録されている。CPUE 値は単位釣り機が単位時間に捕獲したアカイカの重量を表している。漁船ごとに釣り機の性能や数は異なり、操業を行う人間の釣り機を扱う技能によっても CPUE は変化する。しかしすべての漁船に搭載される釣り機が同じ性能であり、操業を行う人間の技能も同じであると仮定すると、CPUE 値はその海域での魚群の密度を反映する値であることができる。このことから、CPUE 値は漁場予測において重要な値である。また、この漁獲実績データは、HSI モデリング結果を可視化する際に漁を行った地点をマーキングするためにも使用する。例えば、太平洋の漁獲実績を基に日本海での漁獲量を予測しようとしても、それぞれの漁獲に関連がないので、有効な予測をすることは

できない。漁を行った地点をマーキングすることで、HSI モデリングの有効な範囲を大まかに把握することができる。

3.2 HSI モデルから漁獲量予測への変換

先に説明したように HSI はその地点が対象生物の生息地として適しているかどうかを 0 から 1 で数値化したモデルである。一方、漁獲量を表す CPUE は単位釣り機が単位時間に捕獲した対象生物の重量である。生息しやすい環境に対象生物が集まる、生息しやすい環境では対象生物の生存確率が上がるので生息しやすい環境では対象生物の数が多くなると考えられる。対象生物の数が多くなると、一般には対象生物の CPUE が高くなる。そのため HSI 値と CPUE が比例関係にあると仮定することで漁獲予測に関する過去の研究が行われてきた。HSI=0 の時は対象生物がおらず CPUE=0 となり、HSI=1 の時は対象生物が最大数おり CPUE=max となるとすると、その変換は式 3.8 で行うことができる。

$$PredictedCPUE = (maxCPUE - 0) \times HSI \quad (3.8)$$

$$= maxCPUE \times HSI \quad (3.9)$$

PredictedCPUE は漁獲量予測、maxCPUE は観測された最大 CPUE である。本研究では HSI と CPUE の変換をこの式で行う。

3.3 SI モデル

漁獲量を単一の海洋環境変数から説明するモデルが SI モデルである。以下では提案手法での SI モデル構築手法について述べる。本研究で使用する海洋環境変数データは 5 種の変数、塩分濃度、温度、緯度方向の流速、経度方向の流速、深さ方向の流速からなる。また各変数は特定の深さごとに値を持っており、それらの線形和を使って SI モデル構築をする場合、次式で表すことができる。

$$SI = \sum_{i=0}^{depth} a_i x_i \quad (3.10)$$

x_i は漁獲点の鉛直方向深さ i での海洋環境変数の値、 a_i はその係数である。さらに過去のデータとして、先に説明した手法で計算した海洋環境変数データを用いると次式で SI モデルを表せる。

$$SI = \sum_{j=0}^{days} \sum_{i=0}^{depth} a_{ij} x_{ij} \quad (3.11)$$

x_{ij} は漁獲点の鉛直方向の深さ i を始点として j 日遡った地点での海洋環境変数の値、 a_{ij} はその係数である。

漁獲点における式 (3.10) の SI 値を漁獲量予測に変換した値と漁獲実績 (CPUE) との差の絶対値を最小化する a_{ij} を求めることで SI モデルを構築できる (式 3.12)。

$$\arg \min_{a_{ij}} (|CPUE - PredictedCPUE|) \quad (3.12)$$

$$= \arg \min_{a_{ij}} \left(\left| CPUE - (maxCPUE * \sum_{j=0}^{days} \sum_{i=0}^{depth} a_{ij} x_{ij}) \right| \right) \quad (3.13)$$

また、実測した漁獲点は複数存在するのでこれを行列で表すことができる (式 3.14)。

$$\arg \min_{\theta} ||\mathbf{Y} - \mathbf{X}\theta||_2^2 \quad (3.14)$$

\mathbf{Y} は $n \times 1$ 行列で、 i 行目は i 番目の漁獲点での CPUE となっている。 \mathbf{X} は k 番目の漁獲点で成立する式 3.11 の x_{ij} を k 行目に横に並べた $n \times m$ 行列である。 θ は $m \times 1$ 行列で、 \mathbf{X} における海洋環境変数 x_{ij} に対応する係数行列である。そのため m は x_{ij} の数と同じである。式 3.14 ではどの漁獲点でも利用可能なモデルを構築しようとしているので、 x_{ij} はどの漁獲点でも同じ係数となる。そのため、 θ は $n \times m$ 行列ではなく、 $n \times 1$ 行列となる。式 3.14 は機械学習

の分野で教師あり学習と呼ばれたり、回帰問題と呼ばれる問題で最小二乗法などの解法によって解くことができる。式 3.14 を解くと全ての変数に 0 でない係数が現れる。これは全ての深さの全ての変数が漁獲量に影響するということを意味するが、例えばある深さで対象生物を漁獲する際に、同じ日の全く別の深さの海洋環境変数とその漁獲量に影響するとは考えにくく、海洋の専門家も否定している。また、すべての変数を漁獲モデルに使用するとオーバーフィッティングの問題が発生する。そのため、漁獲モデルに全ての変数を利用するのではなく、一部を削減する必要がある。

3.4 オーバーフィッティング

構築したモデルが、構築に使用したデータに対してはよく適合しているが、未知のデータに対して適合していない状態がオーバーフィッティング（過学習）である。今回のように計算機による自動計算でモデルを構築する際によく問題となる状態である。本研究の場合であれば、実測した漁獲点で予測させた場合は実際の CPUE に近い値が出るが、その予測モデルを実際の予測に使用すると実際の CPUE とかけ離れたモデルができてしまう状態がオーバーフィッティングである。一般にオーバーフィッティングはモデルの自由度が高すぎる場合に発生する。そこで、式 3.11 のように使用可能な全ての変数を使用するのではなく、変数の数を減らす工夫をする。以下で説明する変数の数を減らすモデル構築手法はスパースモデリングとも呼ばれる。係数行列 θ の成分に 0 の項があれば、実質的に変数の数が減っていることになる。これがスパースモデリングの基本的な変数削減手法であり、その手法には Ridge 回帰²⁵⁾、Lasso 回帰²⁶⁾ など様々な手法がある。Ridge 回帰は古くに提案された変数削減手法であり次のように行う。式 3.14 で示した最小化式に罰則項として L2 ノルムを加えたものである。Lp ノルムは一般に $p \geq 1$ で定義されるノルムで、式 3.15 で表せる。

$$Lp\text{norm} = \|x\|_p = (|x_0|^p + |x_1|^p + |x_2|^p + \dots)^{\frac{1}{p}} \quad (3.15)$$

p=2 の場合が L2 ノルムであり、次式で示すように各係数の二乗和の平方根となる。

$$L2norm = \|x\|_2 = (|x_0|^2 + |x_1|^2 + |x_2|^2 + \dots)^{\frac{1}{2}} = \sqrt{\sum_i |x_i|^2} \quad (3.16)$$

これを罰則項とした上で罰則項そのものの重み λ も考慮し、式 3.14 に対し適用すると次式となる。

$$\arg \min_{\theta} \left(\|\mathbf{Y} - \mathbf{X}\theta\|_2^2 + \lambda \sqrt{\sum_i |\theta_i|^2} \right) \quad (3.17)$$

ただし、 θ_i は θ の i 番目の成分である。式 3.17 が Ridge 回帰である。式 3.17 は次のようにして解く。

式 3.17 は最小化を行いたいので θ で偏微分すると 0 になる（式 3.18）。

$$2X^T X \theta - 2X^T Y + 2\lambda \theta = 0 \quad (3.18)$$

これを整理すると、解くべき方程式は式 3.19 となる。

$$(X^T X + \lambda I) \theta = X^T Y \quad (3.19)$$

式 3.19 を満たす θ は $X^T X + \lambda I$ の逆行列を求めることで求めることができる。これによって Ridge 回帰は解くことができる。

このようにして求めた解はオーバーフィッティングの問題が発生しないようなスパースなモデルになっているだろうか。Ridge 回帰では罰則項が係数の二乗和の平方根であり、変数の数（係数の非 0 の項の数）については一切制約がない。しかし実際にはモデルに与える影響の小さい変数に対応する係数は 0 に近づき、閾値処理によって係数 0 にできる。これは次の理由で説明できる。まず、一般化した Ridge 回帰において、最小化式は式 3.20 となる。

$$\min \left(E(\mathbf{w}) + \lambda \sqrt{\sum_i |w_i|^2} \right) \quad (3.20)$$

この目的関数を w_i で微分すると次式になる。

$$\frac{\partial E(\mathbf{w})}{\partial w_i} + \lambda w_i \quad (3.21)$$

式 3.21 の第一項は特定の w_i で 0 になり、それ未満では正、それより大きい時は負となる。第二項は $w_i = 0$ で 0 になり、それ未満では負、それより大きい時は正となる。また、線形回帰の場合第一項は階段関数となる。そして第一項はその変数がモデルに与える影響の大小に比例する。そのため、影響が小さい変数の場合、第二項が係数 w_i 決定に支配的となり w_i は 0 に近づく。一方、影響の大きい変数の場合、第二項が w_i 決定にはあまり寄与しなくなり、 w_i はあまり 0 に近づかない。モデルに対する影響の大小によって罰則項の効果が変化し、モデル決定に対して影響力の小さい変数は w_i が 0 に近づく。しかし、この手法では w_i が 0 に近づくほど第二項の影響が小さくなるので w_i は完全には 0 にならず、閾値処理をしないと w_i を 0 にできない。

そこで、Lasso 回帰が考案された。Lasso 回帰は罰則項に L1 ノルムを用いた手法である。

$$L1norm = \|x\|_1 = (|x_0|^1 + |x_1|^1 + |x_2|^1 + \dots)^{\frac{1}{1}} = \sum_i |x_i| \quad (3.22)$$

L1 ノルムとはつまり、係数行列の絶対値の和である。こちらに変数の数（係数の非 0 の項の数）についての制約はない。しかし、実際には係数の一部が 0 になる。それを以下で説明する。

一般化した Lasso 回帰の最小化式は次式となる。

$$\min E(\mathbf{w}) + \lambda \sum_i |w_i| \quad (3.23)$$

式 3.23 は Ridge 回帰の時と違い単に偏微分の逆行列では解が出ない。Lasso 回帰を解くアルゴリズムは LARS²⁷⁾ や GPS²⁸⁾、CDA²⁹⁾ などが存在する。

式 3.23 を式 3.21 と同じように w_i で微分すると次式になる。

$$\frac{\partial E(\mathbf{w})}{\partial w_i} + \lambda \cdot \text{sgn}(w_i) \quad (3.24)$$

$\text{sgn}(x)$ は符号関数である。これも先と同様に考えると w_i は第二項を入れな
い場合と比較し、0 に近づく。しかし、符号関数は w_i の正負のみで w_i の大小
に関わらない値を持つ。そのため第一項が小さく、モデルに対する影響の小
さい変数は $w_i = 0$ に収束する。これにより、Ridge 回帰のような閾値処理を用
いることなく、係数が 0 となり変数が削減される方向に最小化式が変化して
いく。

Lasso 回帰は影響の小さい変数の係数が 0 に収束するが、あくまで罰則項は
係数の絶対値の和である。そのためある変数が別の変数と従属の関係にあり、
モデルに与える影響が大きい場合、どちらの変数もモデルに残ってしまう。つ
まりモデル構築に不要な変数が存在しているが、Lasso 回帰では削減できない
場合がある。

そこで罰則項として、係数行列 θ の L0 ノルムと重み λ の積を加えた次式を
考える。

$$\arg \min_{\theta} \left(\|\mathbf{Y} - \mathbf{X}\theta\|_2^2 + \lambda \sum_i \|\theta_i\|_0 \right) \quad (3.25)$$

式 3.15 は $p \geq 1$ の場合で定義されているが、それを $p=0$ の場合にも拡張す
ることで L0 ノルムは次式で表せる。

$$L0norm = \|x\|_0 = (|x_0|^0 + |x_1|^0 + |x_2|^0 + \dots)^{\frac{1}{0}} = \sum_i (|x_i|^0)^{\infty} \quad (3.26)$$

$0^0 = 0$ と定義した場合、 $x_i = 0$ のとき $0^{\infty} = 0$ 、 $x_i \neq 0$ のとき $1^{\infty} = 1$ とな
る。つまり、式 (3.26) は x_i の非ゼロ成分の数となる。この場合、罰則項は変
数の数に比例するので Lasso 回帰の問題のように変数同士が従属の関係にあ
る場合であっても変数の数が削減されるようになる。また、Lasso 回帰と比較
して変数の数が少なくなる傾向にあるのでよりオーバーフィッティングの問題
が発生しにくい。

本研究は以上の理由より、L0 ノルムを罰則項としてモデル構築に使用する。
次に、最適化式 3.25 をどのようにして解くか説明する。

3.5 提案手法のアルゴリズム

先に述べた計算手法を計算機上で実現するアルゴリズムについて説明する。式 3.25 において、求めたい係数行列 θ は $n \times 1$ 行列で、 n は海洋環境変数の数である。一方で θ を求めるために使うことができる式の数 Y の行数であり、つまり漁獲点の数である。海洋環境変数の数は深さインデックス \times 過去のデータのサンプリング数で、5 章で行う実験では 1200 種類ある。一方の漁獲点は一ヶ月分のデータで 180 点程しかない。これはつまり、1200 の未知数を求めたいのに対して、式が 180 しかないことになる。このような連立方程式を劣決定系という。劣決定系の線形連立方程式の解法は様々な手法があるが、例えばどの変数をモデルに入れるかを全てのパターン試すことは現実的ではない。ここで SI モデルが生物の生息環境としての適性を数値化したものであるということに立ち返ると、1200 種類全ての変数がある程度以上 SI モデルに影響を与えるとは考えにくい。また、1200 変数全てを用いたモデルはオーバーフィッティングの問題からも適切なモデルではない。そこで SI モデルがスパースなモデルであると仮定する。この仮定を置くことで圧縮センシングと呼ばれる枠組みを利用できるようになる。

解がスパースであることがわかっている劣決定系の線形連立方程式を解く枠組みとして圧縮センシングが存在する³⁰⁾。標本化定理では信号の最大周波数成分が W 以下の時、 $2W$ 以上の周波数でサンプリングすると完全に再構成可能となる。圧縮センシングでは信号がスパースであるとすることによって標本化定理で示されるサンプリング数より少ないサンプリング数で再構成ができるようになる。このことから可逆性を保ったままデータの通信量を減らす応用やより少ない測定回数で正確に測定する応用などが研究されている。本研究では少ないサンプリング数（実測した漁獲量）で原信号（海洋環境変数の方程式で表した漁獲量の分布）を再構成する問題なので圧縮センシングの考え方を利用できる。

本研究ではその中でも IRLS (Iteratively Reweighted Least Squares) と呼ばれるアルゴリズムを利用した。

3.6 IRLS

IRLS³¹⁾ は最小二乗法を繰り返し計算することで解を得る手法である。まず式 3.25 の変数を使って、行列 W を定義する。行列 W は $m \times m$ の対角行列で、 m は求めようとしている SI モデルで使用する全ての海洋環境変数の数である。その i 番目の対角成分 W_{ii} は式 3.27 で定義する。

$$W_{ii} = \frac{1}{|\theta_i| + \epsilon} \quad (3.27)$$

ただし ϵ は微小定数で、 θ の成分の内 0 でない任意の成分 θ_n に対して $\theta_n \gg \epsilon$ となる定数である。 θ の i 番目の成分が 0 のときは $\|W_{ii}\theta_i\| = \|\frac{0}{\epsilon}\| = 0$ となり、0 でない時は $\|W_{ii}\theta_i\| = \|\frac{\theta_i}{\theta_i}\| = 1$ となる。そのため、 $\|W\theta\|^2$ はちょうど L0 ノルムと同じものとなる。式 3.25 はこの W を使って次のように書ける。

$$\arg \min_{\theta} (\|\mathbf{Y} - \mathbf{X}\theta\|_2^2 + \lambda\|W\theta\|^2) \quad (3.28)$$

IRLS(Iteratively Reweighted Least Squares) は式 3.28 の θ と第二項 (罰則項) を交互に更新することで解に近づいていく手法である。

そのため、式 3.28 にステップ t を導入する。 t 回目の θ の更新と罰則項の中の W の更新をそれぞれ $\theta^{(t)}$ 、 $W^{(t)}$ と表す。 θ 、 W の順に更新するものとした場合、 $t+1$ 回目の θ の更新は式 3.29 のように書ける。

$$\theta^{(t+1)} = \arg \min_{\theta} (\|\mathbf{Y} - \mathbf{X}\theta^{(t)}\|_2^2 + \lambda\|W^{(t)}\theta^{(t)}\|^2) \quad (3.29)$$

式 3.29 の $\arg \min$ 内が極値を取るとき、その θ 偏微分は 0 となる。(式 3.30)

$$\frac{\partial \theta^{(t+1)}}{\partial \theta^{(t)}} = 0 \quad (3.30)$$

これを満たす θ が $\theta^{(t+1)}$ なので式 3.30 から次式が成り立つ。

$$X^T X \theta^{(t+1)} - XY + \lambda(W^{(t)})^2 \theta^{(t+1)} = 0 \quad (3.31)$$

$$(X^T X + \lambda(W^{(t)})^2) \theta^{(t+1)} = XY \quad (3.32)$$

式 3.31 の W は各成分が 0 または 1 の対角行列なので $(W)^2 = W$ となり、

$$(X^T X + \lambda W^{(t)}) \theta^{(t+1)} = XY \quad (3.33)$$

式 3.33 を θ について解くことで $\theta^{(t+1)}$ が計算できる。また $t+1$ 回目の W の更新は直前に更新した $\theta^{(t+1)}$ を使って式 3.34 のように書ける。

$$W_{ii}^{(t+1)} = \frac{1}{|\theta_i^{(t+1)}| + \epsilon} \quad (3.34)$$

ただし、 W は対角行列で W_{ii} はその i 行目 i 列目の成分である。

以上をまとめるとステップ t を導入した W と θ を使って、IRLS アルゴリズムは次のように表せる。

θ_i^0 を全て 1 にする

W_{ii}^0 を全て 1 にする

while until($\|\theta^{(t+1)} - \theta^{(t)}\| < tol$)

 式 3.29 を解くことで $\theta^{(t+1)}$ を求める

 式 3.34 を解くことで $W_{ii}^{(t+1)}$ を求める

 end for

$W_{ii}^{(t+1)}$ を統合して $W^{(t+1)}$ を求める

$t = t+1$ にする

end while

ただし、 tol は微小な定数。

3.7 最適な罰則項の重みの決定

先に説明した IRLS アルゴリズムによって、L0 ノルムを罰則項としたモデル構築が可能である。一方、モデルにはまだ罰則項の重み λ が自由度として残っている。重み $\lambda = 0$ の時は罰則項のない最小二乗法となり、重みが大きすぎると $\theta = \mathbf{0}$ が解となる。重みの決定については次のように行う。重みを変化させて、それぞれの重みでモデル構築をする。その後、各モデルについて次に説明する正答率を計算し、それが最大となる重みを選択することで重みを決定する。このようにして決定した重みを用いて構築した漁獲モデルが、その海洋環境変数の SI モデルとなる。

3.8 正答率

本研究では正答率を定義し、その大小によって罰則項の重みを決定する。正答率は機械学習の分類器などの評価でよく使われる指標である。モデル構築の時に使用していないデータ、つまり未知のデータを分類させた時にどれだけ正しく判断できたのかを意味する。本研究では次のようにして正答率を求める。まず実測した CPUE データ全体をテストデータと学習データに分割する。そして学習データのみでモデル構築を行う。テストデータの漁獲点座標を構築したモデルに入力して、CPUE の予測値を求める。また同時に学習データ内の実測値としての CPUE の平均値を求める。予測した CPUE と学習データの平均 CPUE を比較した時の大小を求める。次に、予測した漁獲点での実際の CPUE と学習データの平均 CPUE を比較した時の大小を求める。これらが一致するかどうかを全てのテストデータに対し求め、その率を正答率とした。

3.9 HSI モデル構築

このように従来手法ではある海洋環境変数の SI モデルを構築する際、当日データの中の一つの深さインデックスの海洋環境変数をスプライン曲線で近似することでモデル構築を行っている。一方で提案手法ではある海洋環境変数の SI モデルを構築する際、当日データと過去データから複数の海洋環境変

数を選びその線形和で近似することでモデル構築を行っている。このようにして計算した SI モデルを統合し HSI モデルとするがその方法は従来手法と提案手法で同じ方法を用いた。先程の SI モデル構築を各海洋環境変数ごとに適用することで、対象魚種にとって生息しやすい各海洋環境変数がわかる。本研究で漁場予測の対象とするアカイカについては過去にいくつか漁場予測を試みた研究がある。それらの研究では各 SI モデルの相乗平均を HSI モデルとしている^{4) 13)}。本研究ではそれを踏襲し、各 SI モデルの相乗平均を HSI モデルとした (式 3.35)。

$$HSI = \sqrt{\prod SI} = \sqrt{\prod \sum_{k=0}^{depth} a_k x_k} \quad (3.35)$$

3.10 別の HSI モデル構築手法

これまで説明したような、SI モデルを統合することで HSI モデルを構築する手法は過去の関連研究⁴⁾³⁾で行われてきた手法である。一方で SI モデルを構築することなく、HSI モデルを直接構築する手法もありえる。つまり式 3.11 は単一の海洋環境変数の総和で SI モデルを構築しているが、これを拡張して次式を考える。

$$HSI = \sum_{k=0}^{var} \sum_{j=0}^{days} \sum_{i=0}^{depth} a_{ijk} x_{ijk} \quad (3.36)$$

ただし k は塩分濃度や海水温などの海洋環境変数の種類である。

この手法が使われず、SI モデルを統合する HSI モデル構築手法が使われてきた理由の一つとして利用者の負担が軽減できるというものがある。例えば M 種類の海洋環境変数を使って SI モデルを統合する場合を考える。一つの SI モデルを構築する所要時間を N とすると、HSI モデル構築の所要時間は MN 時間となる。次に式 3.36 を使って直接 HSI モデルを構築する場合を考える。海洋環境変数全体から一つ変数を選んで HSI モデルを構築する場合でも所要時間は MN 時間となる。さらにこの場合は MN 時間必要であるにも関わらず、構築される関数の複雑さは先の SI モデル一つと同じ程度にしかない。そのため従来手法ではこの構築手法は使われてこなかった。

翻って本研究ではモデル構築を全自動化している。そのため直接 HSI モデル構築を行う場合でも利用者の負担は変わらない。以上のことから式 3.36 を用いて HSI モデルを構築する手法も提案する。どちらのほうがより漁獲予測精度が良いかについては 5 章で実験を行う。

第4章 システム概要

4.1 システム概要

本研究で提案する HSI モデルの新しい構築手法について先の章で説明した。この章ではこれをどのように実現したかについて説明する。

提案手法を実現するシステムの要件は以下のとおりである。

- (1) 公開してはいけない漁獲実績データを扱う場合がある。そのため、漁獲実績データを外部へ送信せずにモデル構築する必要がある。
- (2) 利用者として漁獲モデル構築の専門家だけではなく漁師も想定しているので、システムに対する理解が十分でなくても結果がわかるような可視化を行う必要がある。そのため地図上で漁場予測を表示することで本システムの結果を可視化する。

この要件を満たすよう設計されたシステム概要は図 4.1 のようになっている。

要件 (1) についてはモデル構築を手元の PC で行い、外部とのやり取りを MOVE データ取得のみに限定することで達成できる。要件 (2) を達成するために漁場予測の分布を地図上で可視化する。また利用者にとってのインタラクティブ性を向上させるために、地図の平行移動および拡大縮小も行えるようにした。本研究では提案手法を3つのサブシステムに分けて実装した。

- 流跡線計算システム
- 最適化計算システム
- 地図描画システム

各システムで利用する手法については3章で述べたが、それをどのように実現したのかを次節から説明する。

4.2 流跡線計算システム

流跡線計算システムではルンゲクッタ法を用いた流跡線計算を行い、計算結果に沿ったサンプリングを行う。

なお本研究では流跡線に沿ったサンプリングを 24 時間ごとに行うようにし、流跡線計算を 60 日まで行う。また、実測した漁獲点の中で最も海底が浅い点の最大深さインデックスが 20 なので、深さごとの流跡線計算の始点は深さインデックス 20 までにする。

このシステムでは 60 日分の流跡線計算をそれぞれの深さインデックス、それぞれの始点について順に行っている。各漁獲点での流跡線計算は他の点での流跡線計算と独立している。そして流跡線計算システムでは各点ごとの流跡線計算を順に行うので、始点情報を分割するだけで別の計算機による分散計算ができる。本研究で使用する流跡線計算システムは次のようになっている。

流跡線計算を行いたい始点の座標と日付を読み込む

for i=0;i<(流跡線計算を行いたい点の数);i++

for d=0;d<(深さインデックスの数);d++

i 番目の始点の d 番目の深さインデックスの地点を開始点にする

for t=0;t<(遡りたい日数);t++

時刻 t での 3 次元流速と時刻 t+1 での 3 次元流速を取得

ルンゲクッタ法を用いて 24 時間前の 3 次元座標を求める

時刻 t+1 での各海洋環境変数を取得

先のステップで求めた 3 次元座標での変数を補間し記録する

end for t

end for d

end for i

3章で説明したように、MOVE データを OPeNDAP で取得する際に特定の深さでの水平な格子データが一度に取得できる。そしてある深さでの値を補間するためにはその深さと隣接した2つの深さのデータから線形補間で計算を行う。これを3次元それぞれの速度で行うので、ある点での流速を補間するためには6回の MOVE データ取得が必要となる。3章で説明したようにこのアルゴリズムでは一日前の流跡線を計算するためにルンゲクッタ法による流跡線計算を3回行っておりそれぞれの流速は MOVE データから計算を行う。MOVE データでは一日ごとに各格子点での流速が計算されているので時間での補間は一日ごとの流速データから行う。そのため、一日前の流跡線計算には合計12回の MOVE データ取得が必要となる。そのようにして求めた一日前の点で流速の他に海水温と塩分濃度の2種類の海洋環境変数を線形補間する。そのため先の12回と合わせて合計16回の MOVE データ取得を行うことで単一の点の一日前の各海洋環境変数の計算ができる。単一の漁獲点ではこれを60日、20段階の深さインデックスについて行うので、単一の漁獲点から得られる海洋環境変数を全て計算するためには19200回の MOVE データ取得が必要である。本実験で使用する漁獲実績データでは合計で366点の観測データがあるので、すべての海洋環境変数の計算をするために702万7200回の MOVE データ取得を行った。

4.3 最適化計算システム

最適化計算システムでは流跡線計算システムの結果として計算された、過去のデータを含む海洋環境変数を入力とし HSI モデル構築に必要な各 SI モデル式を計算する。

IRLS アルゴリズムで解くべき式は式 3.33 および式 3.34 である。式 3.34 は式 3.33 の結果から一意に解くことができる。式 3.33 は各行列を次のように構成することができる。Y は $n \times 1$ 行列でその i 行目は i 番目の漁獲点で実測した漁獲量となっている。X は $n \times m$ 行列でその i 行目は i 番目の漁獲点の流跡線を計算することで得られる単一の海洋環境変数を全て横に並べた行列である。 λ は罰則項の重みである。

しかしこのままだと水平方向の流速についての SI モデルを構築する際に問題が生じることがある。水平方向の流速はおおよそ最大値が 100cm/s、最小

値が 0.05cm/s となっている。そのため式 3.33 の行列 X の各項は最大 1000 倍以上の差ができる。例えば流速の中で一つの変数が、SI モデルを構築する際は 0.05cm/s 付近の値を取り、未知の漁獲量を計算する際は 100cm/s 付近の値を取るような場合を考える。この場合 SI モデル内のその変数にかかる係数は大きくなる。さらに未知の漁獲量を計算する際には変数も大きな値を取ることで SI モデルではその変数が支配的になってしまう。このように SI モデルを構築した際に意図した各変数の影響の大きさと、実際の計算結果での各変数の影響の大きさが異なる SI モデルとなる。この問題を回避するために式 3.33 の変数 X に \log をかけた。こうすることで式 3.33 の行列 X の各項の差が小さくなり、未知の漁獲量を計算した結果で各変数の影響が変化しにくい。つまり、最適化計算システムでは $\log X = X^*$ を使い、式 3.33 ではなく式 4.1 を解いている。

$$(X^{*T}X^* + \lambda W^{(t)})\theta^{(t+1)} = X^*Y \quad (4.1)$$

一方で式 3.10 の両辺に \log をかけた場合、式 4.1 ではなく式 4.2 になる。

$$(X^{*T}X^* + \lambda W^{(t)})\theta^{(t+1)} = X^*Y^* \quad (4.2)$$

この方が解法としては自然であるが、式 4.2 では正答率を計算する際に問題が生じる。正答率を計算する時に漁獲量の平均を計算するが、例えば 10^3 と 10^1 の平均を求める場合を考える。算術平均では 505 となるが、 \log をかけた場合は平均が $10^2 = 100$ となる。漁獲予測としての正答率を求める際は算術平均が適切なので、正答率を計算する際に漁獲量の \log を元に戻す作業が必要となる。その処理を省略するために式 4.2 ではなく式 4.1 を解いている。

行列 X は流跡線上でサンプリングした海洋環境変数を並べた行列である。そのため X の一部の成分が 0 や負の数となる可能性がある。その場合 \log をかけると行列 X が特異行列となりモデル構築ができない。これを回避するためにまず 0 の場合は 10^{-8} に置き換える。次にすべての海洋環境変数は絶対値をとった上で \log をかけた。

本研究では 0 から 1 まで 0.01 ごとに用意したそれぞれの λ の場合で SI モデルを計算し正答率が最大かつ最も λ が大きくなる SI モデルを計算結果の SI モデルとして出力する。

最適化計算システムは matlab で実装した。IRLS アルゴリズムでは $\mathbf{Ax}=\mathbf{b}$ を満たす \mathbf{x} を計算する必要があるが、matlab ではこれを計算する関数として `mldivide` 関数がある。今回、 $\mathbf{Ax}=\mathbf{b}$ に相当する方程式は式 3.33 となる。この内、 \mathbf{A} に相当する $X^T X + \lambda W^{(t)}$ は $n \times n$ 行列となっている (n はモデル構築に使用する海洋環境変数の数)。この場合、`mldivide` 関数は $X^T X + \lambda W^{(t)}$ を下三角行列 \mathbf{L} と上三角行列 \mathbf{U} に分解する LU 分解を行うことで式 4.3 の変形をまず行う。

$$\mathbf{Ax} = \mathbf{LUx} = \mathbf{b} \quad (4.3)$$

式 4.3 は $\mathbf{L(Ux)} = \mathbf{b}$ となり、これを \mathbf{Ux} について解く場合はちょうどガウスの消去法の前進消去となるので \mathbf{Ux} が求まる。 \mathbf{Ux} は同様にガウスの消去法の後退代入となるので、 \mathbf{x} が求まる。このようにすることで、 $\mathbf{Ax}=\mathbf{b}$ から \mathbf{x} を求める場合、 \mathbf{A}^{-1} を求める必要はない。つまり、式 3.33 を θ について解く場合、 $(X^T X + \lambda W^{(t)})^{-1}$ を求める必要が無いということである。以上のようにして IRLS アルゴリズムでの θ の更新を行っている。

最適化計算システム全体のアルゴリズムは次のようになっている。

読み込んだ漁獲点を学習データとテストデータに分割する

for $\lambda = 0; \lambda < 1.0; \lambda = \lambda + 0.01$

IRLS アルゴリズムを用いてモデルを構築し係数行列 θ を求める

構築したモデルの正答率を求める

if (正答率 > 最大正答率)

θ と λ と正答率を暫定的に保持する

end if

end for

最大正答率となる θ と λ を出力する

これを各海洋環境変数ごとに行うことで各 SI モデルを構築できる。このようにして構築した SI モデルを次の地図描画システムに送ることで漁獲モデルによる予測の分布が可視化できる。このシステムでは各 SI モデル構築を独立して行うことができるので、別の計算機を用いた分散計算が可能となっている。

本研究で構築する SI モデルは式 3.10 の形になっている。そのため最適化計算システムでの計算結果は使用変数名、その係数の行列の形で表すことができ、その形で地図描画システムへデータを渡している。

4.4 地図描画システム

地図描画システムでは最適化計算システムの結果として得られた係数行列 θ を基に HSI モデルを構築し、地図に描画することで可視化を行う。そのため、漁獲量予測を行い、その正答率を求めて評価する部分はこのシステムが担当している。

最適化計算システムですでに各 SI モデルが構築できているので、3 章で説明したようにそれらの相乗平均をとることで HSI モデルを構築する。そしてその評価を行うために 3 章で説明した正答率をこの HSI モデルに対して求める。

構築した HSI モデルは 3 章で説明した手順で CPUE の予測として可視化する。本研究では過去の関連研究^{4) 13) 3)}と同様に、地図の海の部分を色分けすることで可視化した。

また、地図描画システムには JavaScript を使用し、地図描画に three.js ライブラリを使用した。three.js は WebGL を JavaScript 上で操作するライブラリで 3D の描画ができる。これを 2D 的に扱うことで地図を実現した。システムを実行するとまず陸地を描画する (図 4.2)。

陸地データとして <http://www.naturalearthdata.com/>にある 1:10,000,000 スケールの陸地データを用いた。このデータには陸地情報として陸地の各頂点が緯度経度の座標系で入っている。それを頂点とした MultiPolygon を three.js で描画し黒色で塗りつぶすことで陸地ができる。また、地図は一般になじみ深いメルカトル図法で描画している。メルカトル図法では地図上での経度の単位長さは一定だが、緯度の単位長さは緯度によって変化する。そのため、単に緯度経度に比例した座標変換ではメルカトル図法の地図が作れない。緯度

経度が東経 a 度、北緯 b 度の点は地図の座標系で次式の xy に変換する必要がある。

$$x = a \quad (4.4)$$

$$y = \ln(\tan(45^\circ + b^\circ)) \quad (4.5)$$

地図を描画する際の位置情報は全て緯度経度なのでこの式を全ての位置情報に適用して地図の座標系に変換する。

海の描画は three.js の三角形ポリゴンを描画する機能で実現している。本研究で計算した漁獲予測は計算に使用した MOVE データの格子点上で行うので、その解像度は MOVE データの解像度である緯度経度 0.1 度と同じである。格子は地図上では長方形であるが、これを 2 つの三角形に分割して描画する。そうすることで各頂点に与えた色を線形に補間できるようになる。海の解像度は陸地の解像度より低いので陸地と海の境界が詳細になるよう、陸地の背後に配置した。

以上のようにして陸地と海を構成した。これを通常の 3 次元空間の見え方と同じ見え方を実現する PerspectiveCamera で見ると、近くのものが大きく見え遠くのものが小さく見える。そのため陸地と海の境界付近で、本来海の描画結果を見たい部分が陸地に隠されてしまう。そこで OrthographicCamera を使用した。OrthographicCamera はカメラの射影面を決め、平面状の射影面の各ピクセルから射影面と垂直に直線を引き、その直線上でサンプリングすることでそのピクセルの色を決定するカメラである。一方の PerspectiveCamera はカメラの視点から単位距離離れた場所に視野の長方形を設定することで射影面を決め、視点から射影面の各ピクセルに向かって直線を引き、その直線上でサンプリングすることでそのピクセルの色を決定するカメラである。OrthographicCamera は射影面が平面であることから、視点からの距離によって物体の見え目の大きさが変化せず今回の地図描画に適している。

さらに漁獲予測分布を見る際にインタラクティブ性を高めるために、地図描画システムではマウスを用いた地図の拡大縮小機能と視点移動機能を追加した。これらの機能は全て OrthographicCamera のパラメータ変更によって実現できる。OrthographicCamera では視点の中心、視線の方向、描画する範囲を設定することで一つの見え方を決定できる。この中で視点の中心を、射影面

と水平に動かすとちょうど見え方も水平移動するようになる。例えばマウスで地図をドラッグした際に視点移動を実現したい場合は、マウスの移動に合わせて視点の中心を射影面と水平に移動することでできる。マウスのホイールを操作することで拡大する場合は、地図そのものの大きさは変化させずにカメラが描画する範囲を狭くすることで結果として地図の拡大が実現できる。縮小はちょうどその逆で地図全体の大きさを変化させずにカメラが描画する範囲を広くすることで実現できる。

また、漁獲予測分布のカラーマップも過去の関連研究³⁾と同様に、青→緑→赤の順に予測した CPUE が大きくなるように設定する。一般によく使われる RGB 色空間は赤緑青を成分とする色空間である。これを用いて青→緑→赤の滑らかな色変化を実現するためには、赤緑青の各成分を操作するために 3 つの非線形な変換式が必要になる。そのためこの色の変化を実現するために HSL 色空間に基づいたカラーマップを用いた。HSL 色空間は色相 (Hue) 彩度 (Saturation) 輝度 (Lightness) を成分とする色空間である。色相は 0 から 360 度までの角度で色合いを表し、彩度は 0 から 100% で色の鮮やかさを表し、輝度は 0 から 100% で色の明るさを表している。この中で、色相は 0 度から順に変化させることで赤→緑→青→紫→赤までなめらかに変化する。そのため彩度と輝度を一定にしたまま色相を 240 度から 0 度まで変化させることで青→緑→赤の色の変化を単一の線形な式で表現できる。本研究では純色で表すために彩度を 100%、輝度を 50% とした。そして CPUE の予測値と色相の変換は次式で行った。

$$Hue = 240 \cdot \left(1 - \frac{CPUE - CPUE_{min}}{CPUE_{max} - CPUE_{min}} \right) \quad (4.6)$$

この式で $CPUE_{min} = -1$, $CPUE_{max} = 1$ とした時の色の分布は図 4.3 となる。

先の式で用いた $CPUE_{min}$ と $CPUE_{max}$ は計算した全ての CPUE の最大最小とする方法や、特定の範囲の最大最小とする方法、CPUE の計算結果にかかわらず一定にする方法がありえる。各方法のどれが最適なのかは実際に可視化して比較する必要がある。そのため、地図描画システムでは可視化結果を見てから各方法で可視化し直すことができるようにした。

構築した HSI モデルの分布を描画するためにはある程度の範囲で流跡線計

算とサンプリングを行った点が必要である。しかし、ここまでの流跡線計算はすべて実測した漁獲点上を始点として行っている。そのためこれまでの計算とは別に、分布を知りたい範囲で流跡線計算を行う必要がある。そこで分布を知りたい範囲を格子状に区切りその頂点を記録する。それらを始点として、先の流跡線計算システムで流跡線計算と流跡線上でのサンプリングを予め行う。こうすることで利用者にとっての最終結果である CPUE 予測の分布を可視化できる。

地図描画システムのアルゴリズムは次のようになる。

1. 陸地を描画する
2. 最適化計算プログラムで計算した各 SI モデルデータを読み込む
3. 分布を描画したい範囲での各格子点を始点とした流跡線に沿ってサンプリングした海洋環境変数データを読み込む
4. 陸地の下になるように海を描画する

第5章 実験

本章では、過去のデータを用いることで漁獲モデルの精度が向上することを確認するための実験と提案手法の有効性を確認するための実験について説明する。

5.1 使用データ

この節では本実験で使用するデータについて説明する。

5.1.1 漁獲実績データ

本研究ではアカイカを対象とした漁獲実績データ（CPUE データ）を用いる。このデータは青森県八戸港に所属するイカ釣り漁船が、2006 年 1 月及び 2 月に行った操業を記録したものでそれぞれ 184 点、186 点の漁獲データがある。データには操業日、操業を行った緯度経度、CPUE が記録されている。本研究では緯度経度の情報は CPUE に影響を与えないと考えているので、本実験では緯度経度情報は使用しない。この漁獲実績データ全体を一回の実験で全て使用することはない。先の章で述べたように、提案手法と従来手法ともに SI モデルを構築する時に時刻の情報が失われるため、SI モデルの時間変化を考慮せずモデル構築を行う。そのため、このデータを月ごとに別個のデータとして扱うことで SI モデルの時間変化は無いものと仮定する。

5.1.2 海洋環境変数データ

本実験で使用する海洋環境変数として、気象庁気象研究所で開発された海洋データ同化システムで作成された MOVE(MRI Multivariate Ocean Variational Estimation) データを用いる。このデータでは海水温、塩分濃度、経度方向の流速、緯度方向の流速、深さ方向の流速が利用できる。また、各変数は 54 段階の深さを持っているが最大深さは 6000m にもなり、操業する地点ではそれほどの深さはない。そのため、本実験では全ての漁獲点で最も浅い 300m の深

さまで使用する。これは MOVE データでは 20 段階目の深さである。先の章で説明したように、過去のデータとして流跡線をたどりサンプリングしたデータを用いる。流跡線の開始点は漁獲点の鉛直方向で MOVE データの各段階ごとの深さの点である。そのため一つの漁獲点では 20 点の流跡線開始点が存在する。これをルンゲクッタ法によって 60 日遡り、一日ごとにサンプリングする。さらに各点では 5 種類の海洋環境変数を得ることができる。そのため、過去のデータまで利用する場合最大で 6000 種類の海洋環境変数が HSI モデルに影響することになる。一方当日のデータのみを利用する場合は 100 種類の変数までである。

5.2 過去のデータを用いることの有効性を確認する実験

過去のデータを用いることでそもそも漁獲モデルの精度が向上するのかどうかを確認する実験を行った。

5.2.1 実験方法

一ヶ月分の漁獲実績データを一回の実験で使用する。まず、2006 年 1 月の漁獲実績データのみを用いる。このデータを日時の順に並べ、奇数番目を学習データとして扱う。そして学習データのみでモデルを構築する。構築するモデルは当日のみのデータを用いた場合、10 日前までのデータを用いた場合、20 日前までのデータを用いた場合、30 日前までのデータを用いた場合、40 日前までのデータを用いた場合、50 日前までのデータを用いた場合、60 日前までのデータを用いた場合でそれぞれの場合について提案手法を用いて SI モデル構築を行う。これを各海洋環境変数に対し行うことで 5 種類の SI モデルが構築できる。SI モデルを統合した、最終的な漁獲モデルである HSI モデルの正答率で評価する。そのために、偶数番目の漁獲実績データをテストデータとし、これをもとにそれぞれの HSI モデルの正答率を求める。先に説明した正答率は SI モデルを評価するためのものだが、これはそのまま HSI モデルの評価にも利用できる。次に、学習データとテストデータを入れ替えて同じ実験を行った。さらに、2006 年 2 月の漁獲実績データを用いて同じように奇

数と偶数に分割して実験を行った。提案手法でモデル構築を行う際に式 3.34 の ϵ と IRLS アルゴリズムの手順 4 の tol を指定する必要がある。本研究では $\epsilon = 10^{-8}$ 、 $\text{tol} = 10^{-5}$ で全ての実験を行った。

5.3 HSI モデルの構築手法を比較する実験

3 章では HSI モデルの構築手法として SI モデルを統合する手法と直接 HSI モデルを構築する手法を提案した。これらの構築手法はどちらがより漁獲予測精度が良いか比較する実験を行った。

5.3.1 実験方法

先の実験と同じように漁獲実績データの 1 月分を分割し、奇数番目の漁獲実績データのみを使ってそれぞれの HSI モデル構築手法で HSI モデルを構築する。それぞれの手法で構築した HSI モデルを先の実験と同じように正答率で評価する。こちらも学習データとテストデータを入れ替えて同じ実験を行う。そして同様に漁獲実績データの 2 月分を用いて同じ実験を行った。

第6章 結果及び考察

6.1 実験結果及び考察

先の章で行った実験の結果を述べ、考察する。

第7章 結論

7.1 結論

従来の HSI モデル構築手法では CPUE と海洋環境変数データの間の因果関係を考慮することができない、HSI モデルを構築する際に利用者の負担が大きい、という問題があった。本研究では、それらの問題を解決するために最適化手法を用いた HSI モデル構築の新しい手法を提案した。また提案手法の有効性を確認するために2つの実験を行った。その結果、過去のデータを用いたほうがより正答率の高い HSI モデルを構築できることを確認した。また、提案手法のほうが従来手法と比べ、より正答率の高い HSI モデルを構築できることを確認した。

7.2 課題

一方で課題もある。従来手法ではスプライン曲線で SI モデルを構築しているが、提案手法ではより単純な線形回帰で SI モデルを構築している。一般にモデル構築を行う際は複雑な関数でモデル構築を行ったほうが精度が良くなる。そのためスプライン曲線を最適化することでより精度の高いモデル構築が可能になることが期待できる。

海洋環境変数データの中で、深さ方向の流速は南北方向の流速に比べ HSI モデルに対する寄与は小さいはずという専門家の意見がある。しかし本研究では各 SI モデルを統合し HSI モデルを構築する際、各 SI モデルを全く等価に扱っている。ここで各 SI モデルに重みを導入することによって、各海洋環境変数の HSI に対する寄与を変えることが可能である。これにより、例えば精度の悪い SI モデルの重みを減らし HSI モデルの精度に対する悪影響を減らすことが可能となる。

謝 辞

ここから謝辞の文
ここまで

参 考 文 献

- 1) 気候変動に伴う水産資源・海況変動予測技術の革新と実利用化, 気候変動適応研究推進プログラム平成24年度報告書, (2012), pp. 4–16.
- 2) 田中章, HEP 入門―“ハビタット評価手続き”マニュアル, (朝倉書店, 2006).
- 3) 櫛田将史, 上中隆史, 坂本尚久, 小山田耕二, 海洋データ同化結果を使ったアカイカ好適生息域モデル作成支援可視化環境の開発, 第27回数値流体力学シンポジウム講演論文集, (2013).
- 4) 五十嵐弘道, 淡路敏之, 石川洋一, 蒲地政文, 碓氷典久, 酒井光夫, 加藤慶樹, 齊藤誠一, 清藤真樹, 海洋再解析データを用いたアカイカ好適生息域モデルの構築とその利用 recca アカイカプロジェクトの事例, *JAMSTEC Rep. Res. Dev.*, Vol. 18, (2014), pp. 89–101.
- 5) Roper C.F.E., Sweeney M.J., Nauen C.E, Fao species catalogue: an annotated and illustrated catalogue of species of interest to fisheries, *FAO Fish. Synop.*, Vol. 3, (1984), pp. 175–176.
- 6) SUZUKI TSUNEYOSHI, Studies on the relationship between current boundary zones in waters to the southeast of hokkaido and migration of the squid, *ommastrephes sloani pacificus* (steenstrup), *MEMOIRS OF THE FACULTY OF FISHERIES HOKKAIDO UNIVERSITY*, Vol. 11, (1963), pp. 75–153.
- 7) 宮尾孝, 藤原弘行, 川村知裕, 北川隆洋, 永井直樹, 栄養塩分析の高精度化, 測候時報, Vol. 80, (2013), pp. 139–147.
- 8) 村田守, 北太平洋におけるいか流し網漁場の海洋環境及びアカイカの分布・回遊, 日本海ブロック試験研究集録, Vol. 17, (1990), pp. 144–148.
- 9) 中村好和, イカ類の野外行動観察の手法特にバイオテレメトリーを中心に, 平成4年度イカ類資源研究会議報告, (1994), pp. 118–128.

- 10) P.Lehodey, A spatial ecosystem and populations dynamics model (seapodym) - modeling of tuna and tuna-like populations, *Progress in Oceanography*, Vol. 78, (2008), pp. 304–318.
- 11) 根本利弘, 五十嵐弘道, 淡路敏之, 喜連川優, 水産資源研究のための粒子追跡シミュレーション: インタフェースの作成, 情報処理学会全国大会講演論文集, Vol. 72, (2010), pp. 41–42.
- 12) Justus von Liebig, *Die Grundsätze der Agricultur-Chemie mit Rücksicht auf die in England angestellten Untersuchungen*, (Friedrich Vieweg und Sohn Publisher, 1855).
- 13) Siquan Tian, Xinjun Chen, Yong Chen, Liuxiong Xu, Xiaojie Dai, Evaluating habitat suitability indices derived from cpue and fishing effort data for *ommatrephes bratramii* in the northwestern pacific ocean, *Fisheries Research*, Vol. 95, (2009), pp. 181–188.
- 14) 新保裕美, 田中昌宏, 池谷毅, 越川義功, アサリを対象とした生物生息地適性評価モデル, 海岸工学論文集, Vol. 47, (2000), pp. 1111–1115.
- 15) 日本環境アセスメント協会・研究部会自然環境影響評価技法研究会, メバル稚魚・幼魚 hsi モデル, 自然環境影響評価技法研究会報告書, (2004), pp. 143–162.
- 16) 大野剛, 藤原靖, Gis による自然度評価およびオオタカ hsi モデルからみた建設事業地選定の可能性について, 地球環境シンポジウム講演集, Vol. 17, (2009), pp. 103–108.
- 17) 田中昌宏, 上野成三, 林文慶, 新保祐美, 高山百合子, 沿岸自然再生の計画・設計を支援する環境評価手法に関する一考察, 土木学会論文集, Vol. 741, (2003), pp. 89–94.
- 18) 高山百合子, 上野成三, 勝井秀博, 林文慶, 山木克貝, 田中昌宏, 江奈湾の藻場分布データに基づいたアマモの hsi モデル, 海岸工学論文集, Vol. 50, (2003), pp. 1136–1140.

- 19) 田中章, 大澤啓志, 吉沢麻衣子, 環境アセスメントにおける日本初のH E P適用事例, ランドスケープ研究, Vol. 71, (2009), pp. 543–548.
- 20) 浜本洋, 風間聡, 澤本正樹, Hsl モデルを用いた流域生物多様性の評価, 環境工学研究論文集, Vol. 43, (2006), pp. 559–567.
- 21) C. Xinjun, T. Siquan, L. Bilin, C. Yong, Modeling a habitat suitability index for the eastern fall cohort of *ommastrephes bartramii* in the central north pacific ocean, *Chinese Journal of Oceanology & Limnology*, Vol. 29, (2011), pp. 493–504.
- 22) Norihisa Usui, Shiro Ishizaki, Yosuke Fujii, Hiroyuki Tsujino, Tamaki Yasuda, Masafumi Kamachi, Meteorological research institute multivariate ocean variational estimation (move) system: Some early results, *Advances in Space Research*, Vol. 37, (2006), pp. 806–822.
- 23) Hiroyuki Tsujino, Tatsuo Motoi, Ichiro Ishikawa, Mikitoshi Hirabara, Hideyuki Nakano, Goro Yamanaka, Tamaki Yasuda, Hiroshi Ishizaki, *Reference manual for the Meteorological Research Institute Community Ocean Model (MRI.COM) Version 3*, (Oceanographic Research Department, 2010).
- 24) P. Cornillon, J. Gallagher, T. Sgouros, Opendap: Accessing data in a distributed, heterogeneous environment, *Data Science Journal*, Vol. 2, (2003), pp. 164–174.
- 25) Arthur E. Hoerl, Robert W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, Vol. 12, (1970), pp. 55–67.
- 26) Robert Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, (1996), pp. 267–288.
- 27) BRADLEY EFRON, TREVOR HASTIE, IAIN JOHNSTONE, ROBERT TIBSHIRANI, Least angle regression, *The Annals of Statistics*, Vol. 32, (2004), pp. 407–499.

- 28) Jerome H. Friedman, Fast sparse regression and classification, *International Journal of Forecasting*, Vol. 28, (2012), pp. 722–738.
- 29) Jerome Friedman, Trevor Hastie, and Robert Tibshirani, Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, Vol. 33, No. 1, (2010), pp. 1–22.
- 30) 三村和史, 圧縮センシング：疎情報の再構成とそのアルゴリズム (時間周波数解析の理論とその理工学的応用), 数理解析研究所講究録, Vol. 1803, (2012), pp. 26–56.
- 31) Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, C. Sinan Gunturk, Iteratively re-weighted least squares minimization for sparse recovery, *Communications on Pure and Applied Mathematics*, Vol. 63, (2010), pp. 1–38.

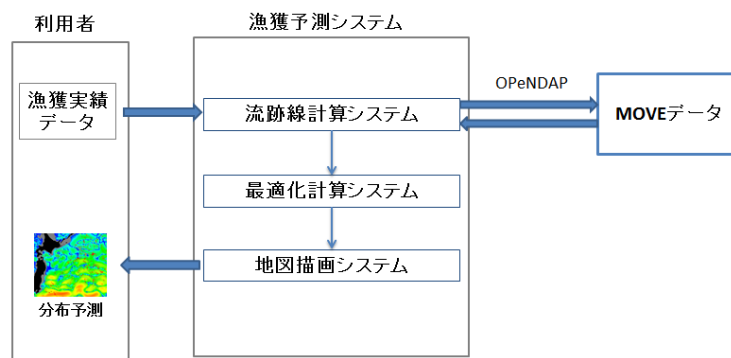


図 4.1: システム概要



図 4.2: 地図描画システムの初期状態。海の部分はまだ描画していない。メルカトル図法で地図を描画している。

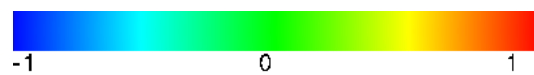


図 4.3: HSL 色空間を基にしたカラーマップでの色の分布