統計モデルとデータマイニング手法の水産資源解析への応用

水研センター研報，第22号，1－85，平成20年
Bull. Fish. Res. Agen. No. 22, 1-85, 2008

博士号論文

統計モデルとデータマイニング手法の水産資源解析への応用[*1]

庄野　宏[*2]

Application of statistical modeling and data mining method to the fish stock analyses

Hiroshi SHONO

Abstract：In this thesis, we focused on the various problems in the filed of fish population analysis, especially regarding the analyses of CPUE (catch per unit effort) which shows the relative abundance. We suggested several techniques to solve these issues by the statistical modeling and approaches for data mining using the actual fishery data on tuna and its related species, and computer simulation experiments.

Catch per unit effort (CPUE) is an important concept which is corresponding to the relative stock size and is proportional to the stock abundance. However, because the nominal CPUE may include various spatiotemporal and environmental effects except for stock density such as area, season and fishing gears, we need to remove these effects to grasp the annual variation of the stock. Therefore, it has been traditionally carried out to estimate the factorial effect of year using analysis of covariance (ANCOVA) model (e.g. CPUE Log-Normal model) where natural logarithm of CPUE is set to the response variable and assumed factorial effects are incorporated into the model as explanatory variables under the normal error, and generalized linear model (GLM) (e.g. Catch Poisson model, Catch Negative-Binomial model) in which catch, discrete variable, is set to the response one and Poisson or negative binomial distribution and so on is assumed. Such work is called CPUE standardization and approaches for data mining such as tree-regression model and neural networks have been recently utilized for it in addition to the statistical modeling.

In this study, we dealt with the CPUE standardization, major issue in the fish population analysis, as main theme of this paper and discuss in detail three problems about CPUE analysis as follow:

1) Choice of the factorial effects, performance evaluation of the model through the various information criteria and stepwise test in the ANOVA type model supposing the CPUE standardization (Chapter 3)
2) Approach of CPUE prediction and the simple method for attribution analysis (i.e. method for extracting CPUE year trend) in the time-space without operation for southern bluefin tuna by the neural networks (Chapter 4)
3) Performance evaluation of Tweedie model if it includes many zero-catch and comparison of Tweedie distribution and the traditional methods (ad hoc ANCOVA method, Catch model) (Chapter 5)

Chapter 1 becomes an introduction and describes the background, purpose of this research and composition of this thesis. In chapter 2, we outlined CPUE standardization from

2007年5月16日受理 (Received on May 16, 2007)
[*1] 筑波大学審査学位論文（掲載に際し投稿規定に沿って一部修正した）
[*2] 遠洋水産研究所 〒424-8633 静岡市清水区折戸5-7-1 (National Research Institute of Far Seas Fisheries, 5-7-1, Orido, Shimizu, Shizuoka 424-8633, Japan)

the viewpoint of the statistical modeling, approach for data mining, proper problems of fish stock and reviewed several related issues, especially main three problems to be coped with in this study.

In chapter 3, we performed the model selection by various information criteria (AIC, BIC, CAIC, c-AIC, HQ, TIC etc.) using the generalized linear models corresponding to the CPUE standardization through several cases such as in small samples, large samples. It is also presented that the result of model selection may be different depend on the used information criteria in actual fishery data. We evaluated the selection performance of these information criteria using the computer simulation in which we calculated the selection performance to choose the true model among several candidate models generated random numbers from the true model. We also compared the performance of information criteria and stepwise test by computer experiments because some stepwise test such as chi-square or F test can be applied in the nested model. The variable selections are an important and essential issue in terms of selecting the factorial effects statistically to affect the CPUE. In addition, the results of model selection based on the information criteria and stepwise test may cause the difference of the attribution analysis (i.e. estimated CPUE year trend), which may lead to the big difference of estimated absolute abundance in the model where CPUE year trends are included as the tuning indices. Specific study results in this chapter are as follow:

- It was found that the result of model selection in small samples and in the case that there are many parameters compared to the sample size by c-AIC, which is a finite correction of AIC, is different from that by AIC and the selection performance of c-AIC is better than that of AIC through the ANOVA-type simulation in such cases.
- It was shown that AIC may have a bias in large sample, the result of model selection is different depend on the information criteria utilized and the consistent information criteria (BIC, HQ and CAIC) is superior to AIC as a whole through the analysis by actual fishery data and simulation by linear regression, respectively. We also suggested the recommendation value and formula of the constant term in the consistent information criterion, HQ.
- It was proofed that the expectation of TIC, which is known as having good performance traditionally in the nested model, becomes theoretically equal to that of AIC in the generalized linear model with having normal error and identity link function, and the selection performance of TIC is almost the same as that of AIC by the computer simulation.
- In the nested model, we found that the information criteria is generally a little superior to stepwise test by our computer experiments and the simple model with a few parameters tend to be selected if the significance level is low in the stepwise test.

In Chapter 4, we focused on the issue of CPUE interpretation of southern bluefin tuna, the problem of CPUE prediction in the spatiotemporal cells without observation, and carried out the CPUE analysis using the neural networks. In terms of the relative abundance, it is reasonable to define the CPUE as multiplying standardized CPUE by relative area size and which is called abundance index (AI). In the stock of southern bluefin tuna, because the fishing ground has shrunk from past to present, it has influenced on the abundance index that the assumption of CPUE in the cell with operation in the past and without one now, that is whether CPUE in these cells is assumed to the same as that in the surrounding areas or 0. This cause the difference of CPUE year trend obtained from the abundance index.

Therefore, in this paper, we predicted the CPUE in such missing cells using the error back propagation method, which is a typical algorithm in the supervised neural networks, and suggested the simple way of attribution analysis to extract the CPUE year trend. We compared to the MCMC method based on the EM algorithm in same conditions by cross-

validation to evaluate the accuracy of the neural networks. Performance check and comparison of the models were carried out using the n-fold cross-validation based on the correlation coefficient between observed and predicted values and mean squared error (MSE).

As a result, the ratio of CPUE without operations over with ones based on the CPUE predicted values by the neural networks showed the range of 0.8 to 1.0. This does not imply extreme contradiction with the CPUE ratio in the Japanese Experimental Fishing Program (EFP) which was locally done for 1998 to 2000, where CPUE ratio was recorded about 0.7 although year, season and area of the experiment were very limited. Predicted performance of CPUE by the neural networks is rather superior to that by MCMC method based on the EM algorithm in the same situation as the neural networks and the CPUE year trend calculated from the predicted CPUE is very similar to that by generalized linear model including the ANCOVA. The results suggest the excellence of the predicted performance of the neural networks and the validity of the simple method of the attribution analysis proposed.

In Chapter 5, we discussed in detail the issue where the ANCOVA model (in which the natural logarithm of CPUE is set to the response variable) can not be applied if it includes the data that catch is zero called zero-catch problem, supposing the shark species caught by tuna longline fishery. We carried out the CPUE standardization for yellowfin tuna in the Indian Ocean caught by the Japanese commercial longline fishery in which the ratio of zero-catch is low about 10% and silky shark in the North Pacific Ocean by Japanese training vessels (for silky shark where the zero-catch ratio is high more than 80%) using the so-called Tweedie distribution which is an extension of compound Poisson model and can be uniformly dealt with the zero data. Actually, we compared the CPUE year trends obtained from the Tweedie model, ad hoc ANCOVA model to add the constant term to all CPUE and Catch Negative-Binomial model. As a result, there is no extreme difference of year trends between the Treedie model and ad hoc method for yellwofin tuna in the Indian Ocean, a target species with low zero-catch rate. On the other hand, CPUE year trend obtained from the Tweedie model is different from that based on the Catch model and ad hoc method for silky sharks in the North Pacific Ocean, a by-catch species with high zero-catch ratio..

Accuracy of the Tweedie distribution is higher in each case judging from the performance check of the candidate models based on the both indicators, correlation coefficient between observed and predicted values and MSE, using n-fold cross-validation as well as our analysis by the neural networks. As a result of cross-validation, the superiority of the Tweedie model does not appear so clearly if the rate of zero-catch is low and it has few problems to apply the ad hoc method practically. On the contrary, if the ratio of zero-catch is high, then the superiority of the correlation coefficient and MSE is the order of the Tweedie model, Catch model, ad hoc method and Tweedie model, ad hoc method, Catch model, respectively. However, the ad hoc method has a large bias because almost all of the estimated CPUE show extreme low regardless of the magnitude of the observed CPUE values. Therefore, we concluded that it is not adequate to apply the ad hoc method in the case that the ratio of zero-catch is high such as shark species.

The last Chapter 6 shows the conclusion of this thesis. We methodically described the study results of three issues which were dealt with in this paper from the viewpoint of fish population analysis, applied statistics and research problem for the future.

Key works: CPUE standardization, data mining, generalized linear model, model selection, Tweedie distribution

目　次

第1章　序　　論

1-1. 本研究の背景と目的

水産資源解析における1つの重要な概念として，CPUE（catch per unit effort：単位努力当たり漁獲量）（Russell，1931）が挙げられる。この指標は，一般の漁業では

$$\mathrm{CPUE}=\frac{\mathrm{Catch}}{\mathrm{Effort}} \tag{1.1}$$

と定義される。Catch（漁獲量）は単一魚種の漁獲重量や漁獲尾数で考えることが多く，努力量としてはえ縄船では針数を，まき網船では操業日数や操業回数などを用いる。一般にCPUEは資源密度に比例していると考えられており，特別な計算を行わなくともCPUEの年トレンドを見るだけで相対資源量の増減傾向を把握することが可能である。また，CPUEは微分方程式をベースに個体群動態を表したプロダクションモデル（Pella and Tomlinson，1969）や年別年齢別漁獲量を使用したVPA（virtual population analysis）（Gavaris，1988）などの資源評価モデルにおけるチューニングインデックスとして用いられることも多く，水産資源評価において非常に重要な意味を持っている。

しかし，商業船などの操業データによる加工していないCPUEは，資源密度以外の様々な要因（季節，海区，漁具など）も含んでおり，資源の変動を正確に知るためにはこれらの影響を取り除く必要がある。Gavaris（1980）などは，このような資源の年変動に対応する部分を取り出す作業をCPUE標準化と呼んだ。本研究においてもこのような計算をCPUE標準化またはCPUE解析と呼称することにしたが，CPUE標準化の主な目的を挙げると以下のようになる。

1．資源の年トレンド（相対資源量）の効果の抽出
2．海区・季節などの時空間的な要因や，漁船に装備されているソナーやバードレーダーなどの探索機器，パワーブロックやパースウィンチなどの操

業機器，表面水温や塩分濃度などの環境要因等がCPUEに与える影響の統計的な測定

3．年別漁獲重量を使用したプロダクションモデル，年別年齢別漁獲尾数などに基づくVPA（virtual population analysis）等の資源の絶対量を推定するための評価モデルにおけるチューニングインデックスとしての有用性の向上

CPUE標準化に関する一般的な事項については平松（1995）の概説論文やHilborn and Walters（1992），Quinn and Deriso（1999）などの水産資源学に関する一般的なテキストに記載されているが，現状ではGLM（generalized linear models：一般化線形モデル）（Dobson, 1990; McCullagh and Nelder, 1989）などの統計モデルを使用して，資源密度以外の要因効果を除去した何らかの基準化を行うことが多い。具体的には，分散分析や共分散分析に代表される線形モデル，すなわち一般化線形モデルの枠組みで言えばリンク関数が恒等写像でかつ誤差項が正規分布に従うモデル

$$\begin{aligned}\mathrm{Log}\,(\mathrm{CPUE}) &= (\mathrm{Intercept}) + (\mathrm{Year}) + (\mathrm{Season}) + (\mathrm{Area}) + (\mathrm{EMT}) \\ &\cdots + (\mathrm{Two\text{-}way\ Interactions}) + \mathrm{error},\ \mathrm{error} \sim N\,(0,\ \sigma^2)\end{aligned} \tag{1.2}$$

を利用することが多い。

また，離散変量である漁獲量を応答変数に設定し，Poisson分布や負の二項分布（negative binomial, NB）を仮定した一般化線形モデル

$$\begin{aligned}E\,[\mathrm{Catch}] = \mathrm{Effort}^{*}\exp\{&(\mathrm{Intercept}) + (\mathrm{Year}) + (\mathrm{Season}) + (\mathrm{Area}) + (\mathrm{EMT}) + (\mathrm{Two\_wayInteractions})\},\ \mathrm{Catch} \sim \mathrm{Po}\,(\lambda) \\ &\text{または } \mathrm{NB}\,(\alpha,\ \beta)\end{aligned} \tag{1.3}$$

などを使用することも多い（Read, 1996）。

（1.2）式や（1.3）式での（Year），（Season），（Area）はそれぞれ年，季節（月または四半期など），緯度や経度などを元に定義されることが多い海区の要因効果を表し，これらは順序のないカテゴリカル変数と定義することが一般的である。（EMT）は表面水温，塩分濃度などの環境要因や漁船に装備されている装置類，漁具などの機器効果を総称しており，連続変量を使用することもある。

最近では，共分散分析などを含めた一般化線形モデルに加えて，樹形モデルやニューラルネットワークなどのデータマイニング的なアプローチがCPUE標準化に対して用いられることも多く（Watters and Deriso, 2000），これらの手法も含めて第2章のレビュー（既存研究のサーベイ）部分で詳しく記述する。

CPUE標準化の主な目的として挙げた1.の相対資源量に対応する年トレンド抽出は，一般化線形モデルにおいては（1.2）式や（1.3）式での年（Year）効果を取り出すことによって可能になる。場合によっては，（Year）と他の要因の交互作用を取り入れることもあるが，その際にはLSMEANS（least squared means：高橋ほか, 1989）を利用することが一般的であり，LSMEANSについては2-2-3節で詳しく取り上げる。

目的2.で述べた，CPUEに影響を与えている可能性がある時空間的な要因や環境要因，漁船に装備された機器類の要因効果の分析についても，一般化線形モデルを利用した場合には，該当する説明変数の主効果，もしくはLSMEANSを計算することにより推定可能である。ただし，ニューラルネットワークなどのデータマイニング的なアプローチの場合には，要因分析が難しい場合もあり，本論文の第4章ではニューラルネットワークによる得られた予測値を元にしたシンプルな年トレンド抽出法，すなわち簡便な要因分析法の提案を行った。

CPUE解析における目的3.は，標準化されたCPUEを入力（チューニングインデックス）として用いて資源の絶対量推定を行う場合の利用法である。実際，使用するCPUE年トレンドのわずかな違いが，推定された資源尾数ないし資源重量の大きな差異をもたらす場合もあり，細かな注意が必要である。このことについては，資源評価に用いられている代表的なモデルと合わせて，第2章（2-4節）で再度議論する。

本研究では，水産資源解析に関係する実用的な観点から，CPUE解析（CPUE標準化）における様々な問題を取り上げ，統計モデルやデータマイニング手法を応用して解決することを目的とする。具体的には，

1）モデル選択の問題

2）資源量指数の問題

3）ゼロ・キャッチの問題

の3つについて取り上げ，詳細な検討を行った。

1）は一般化線形モデルの枠組みを用いて，時空間的な要因，季節要因，環境要因など，どのような説明要因がCPUEに対して影響を与えているか否かを統計学的に判断するという，CPUE解析における重要な変数選択の問題であり，本研究では，AIC（Akaike's information criterion, 赤池情報量規準）（Akaike,

1973）や BIC（Bayesian information criterion, Bayes 情報量規準）（Schwarz, 1978）などに代表される情報量規準，および F 検定やカイ二乗検定などに代表される stepwise 検定を利用し，モデル選択の良さについて比較検討を行った。

具体的には，CPUE 標準化を中心に，成長曲線の推定や体長組成の年齢分解など水産資源解析における典型的な問題に関する実際の漁業データや仮想データを用いて様々な情報量規準や統計的検定による変数選択，モデル比較を行うとともに，小標本の場合，大標本の場合，ネスト構造を持つ場合，正規混合分布などの CPUE 標準化を想定したモデルを仮定してシミュレーション実験を行い，これらの指標のパフォーマンスを詳細に検証した。これらモデル選択の議論は，CPUE 標準化に限ってみても，CPUE に影響を与えていると予想される様々な説明要因が，本当に影響を与えているか否かを統計的に検証するという意味において，極めて重要かつ本質的な問題である。

2）の資源量指数とは，CPUE に該当部分の相対面積サイズを掛け合わせたものであり，標準化された CPUE に対してこのような加工を行うことによって相対資源量に対応する年トレンドなどが抽出されると考えられている。一般に CPUE は資源密度を表わすと考えられており，想定する海域でのまぐろ類の相対資源量を算出するためには，面積指数による重み付けが必要不可欠である。しかし，このような計算による資源量指数の算出方法では，解析エリアを複数のサブエリアに分割した場合も含めて，想定するエリアでの CPUE（資源密度）は一定であると仮定しており，仮に過去から現在にかけて魚の分布が縮小して漁場も縮小している場合には，資源の過大推定になっていると思われる。

ミナミマグロ資源では，緯度経度が 5 × 5 で月別に集計された漁獲量・努力量データを用いて CPUE 解析を行っているが，過去に操業があり現在無くなっているセル（5 × 5 ブロック）が多く存在し，その部分の CPUE を前述のように周辺セルと同じと仮定するかそれともゼロと仮定するかにより，近年，特に1990年代以降の CPUE 年トレンドが大きく異なり，それをチューニングインデックスとして使用した資源の絶対量推定の結果も大きく異なってしまった。これは，一般紙で大きく取り上げられた日本と豪州によるミナミマグロ裁判の主要な原因と考えられている。そこで，本研究では出力信号が存在しない部分の CPUE をニューラルネットワークによって推定し，予測精度をクロス・バリデーションにより評価した。また，ニューラルネットワークによる予測結果から年トレンドを抽出することを目的とし，合わせて簡便な要因分析法の提案を行った。

3）のゼロ・キャッチ問題とは，努力量があり，すなわち操業を行って漁獲が無かったデータが存在する場合に，Catch がゼロ，言い換えれば CPUE がゼロとなり，(1.2) 式で表現されるような CPUE の自然対数を応答変数にした共分散分析モデルが使用出来なくなる現象である。そこで，通常は全ての CPUE に一定量（微少量）を足し込む ad hoc な解決法（(1.4) 式参照）が取られることが多いが，CPUE の区間推定値のみならず点推定値に対しても偏りが生じる可能性が高く，特にゼロ・キャッチの割合が多いケースでは深刻な問題になる。また，Catch を応答変数に一般化線形モデルを使用する方法（(1.3) 式）など，他の回避策に関して，モデルがデータとマッチしていない場合が多く，いずれも一長一短である。

$$\begin{aligned}\mathrm{Log}\,(\mathrm{CPUE}+\mathrm{constant\_term}) &= (\mathrm{Intercept}) + (\mathrm{Year}) \\ &+ (\mathrm{Season}) + (\mathrm{Area}) + (\mathrm{EMT}) \\ &\cdots + (\text{Two-way Interactions}) + \text{error, error} \sim \\ &\mathrm{N}\,(0,\ \sigma^2) \qquad (1.4)\end{aligned}$$

そこで，本研究ではゼロ・データを統一的に取り扱える，複合 Poisson 分布の概念を拡張した Tweedie 分布（Tweedie, 1984: Jorgensen, 1997）を用いて，まぐろ類の漁獲データやゼロ・キャッチ率が高い混獲データの CPUE 標準化を行い，(1.4) 式の ad hoc な方法や Catch モデルとの比較をクロス・バリデーションにより試みた。実際には，n-fold cross-validation と呼ばれるデータをランダムに n- 分割する方法を使用し，モデル評価の基準として，主に観測値と予測値の Pearson's 相関係数および MSE（mean squared error：平均二乗誤差）を使用した。

1-2. 本論文の構成

本論文では，水産資源解析に特有の種々の問題について，特に CPUE 解析における幾つかの問題について統計科学的な側面から捉え，これらの問題に様々な統計モデルやデータマイニング的なアプローチを適用することを目的にしている。本論文の構成は以下の通りである。

第 1 章では，本論文の目的および背景を示した。

第 2 章では，CPUE 解析の現状と課題に関して，まぐろ類の資源に関する題材を例として主に方法論に焦点を当ててレビューし，合わせて問題点について整理することを目的とする。CPUE 標準化に広く使用されている一般化線形モデルなどを含む統計モデル，最近

適用例が増えつつあるデータマイニング手法，CPUE標準化における漁業資源特有の問題について，背景も含めて概説する。

そして，これらのレビューから得られた水産資源解析（CPUE標準化）における主要な3つの問題（5頁の1)-3)）を本論文のメインテーマに据え，第3章から第5章にかけて詳しく検証した。

第3章では，CPUE標準化におけるモデル選択について，様々な情報量規準やstepwise検定を取り上げて比較し，小標本の場合，大標本の場合，ネスト構造を持つ場合，正規混合分布など様々なケースを仮定して，現実の漁業データおよびシミュレーション実験を通して，モデルの良さについて検討を行う。これらのモデル選択は，CPUEに影響を与えていると予想される様々な変量（時間的,空間的な要因,季節的な要因，環境要因，漁船に装備されている機器類の要因等）の効果が，統計的に有意であるか否かを判断するという意味において，本質的な問題である。

第4章では，ミナミマグロ資源のCPUE解釈の問題について議論する。ミナミマグロ漁業では，過去から現在にかけて漁場が縮小しており，過去に漁獲があり現在漁獲が行われていないエリアのCPUEを，周りの海域と同じと考えるかゼロと仮定するかにより，資源量指数（資源密度に相当すると考えられているCPUEに相対的な面積指数を掛け合わせたもの）の値がかなり変わってくる。そのため，該当するエリアをニューラルネットワークにより推定し，クロス・バリデーションによる精度評価を行い，ニューラルネットワークによる予測値を元にCPUE年トレンドを抽出するための簡便な要因分析手法を提案する。

第5章では，ゼロ・キャッチと呼ばれる，漁獲がゼロであるデータが含まれる場合に，CPUEの自然対数を応答変数とした共分散分析モデルが適用できない問題を取り上げる。ゼロ・データを統一的に取り扱えるTweedie分布モデルと既存の回避法のモデル比較を目的として，実際の漁業データ，特にゼロ・データを多く含む混獲データによる解析を行い，複数モデルの精度評価を行った。

第6章では，本研究によって得られた成果を要約するとともに，今後の課題について述べる。

第2章　水産資源解析におけるCPUE標準化の現状：既存研究のレビュー

2-1. はじめに

本章では，CPUE標準化の現状と課題に関して，庄野（2004）の内容を主に取り上げて説明する。具体的には，まぐろ類の資源に関する題材を例として主に方法論に焦点を当ててレビューを行い，合わせて問題点について整理することを目的とする。

2-2節では，CPUE標準化に広く使用されている統計モデル，その中でも一般化線形モデルに焦点を当て，モデルの仮定を含めた標準化の現状およびモデル選択や要因分析の問題点について述べる。2-3節では，最近適用例が増えてきているデータマイニング手法，その中でも樹形モデルやニューラルネットワーク，一般化加法モデルを主に取り上げ，解析の現状とCPUE年トレンドの抽出などの問題点について記述する。2-4節では，その他のCPUE標準化特有の問題点を紹介し，ゼロ・キャッチ問題と呼ばれる事項に関する対処法と相対面積指数による重み付けを行った資源量指数を中心に取り上げる。また，2-4節では，努力量の定義や不均一性，CPUE標準化に使用されるデータの性質やハビタットモデルなど，本研究で詳しく議論出来なかった話題について簡単に触れる。

2-5節は，本章のまとめ部分であり，レビューを通じて明らかになった問題点の中で，特に水産資源解析の観点から重要な課題について記述し，本研究のメインテーマとして，第3章以降で詳しく検討する。具体的には，モデル選択，資源量指数，ゼロ・キャッチ問題の3つであり，それぞれ本論文では1章ずつ割り当てている。

情報量規準やステップワイズ検定を利用したモデル選択は，CPUE標準化においてどの要因がCPUEに影響を与えているかを考える意味で非常に重要であり，本論文のメインテーマとして第3章で詳細に議論する。また，資源量指数の問題とゼロ・キャッチ問題は本研究の主要なテーマの1つであり，第4章および第5章で詳しく述べる。

2-2. 統計モデル

CPUE標準化では,CPUEに影響を与える要因（年・季節・海区・漁船に装備されている操業機器・環境要因など）を説明変数に，CPUEあるいはCatchを応答変数とした一般化線形モデル（回帰分析や分散分析・共分散分析モデルを含む）が伝統的に使用されている。代表的なモデル（Gavaris, 1980；Large, 1992; Reed, 1996）としては，以下のCPUEモデルとCatchモデルが挙げられるが，いずれもCPUEが各々の要因について効果の積の形で表されており，乗法モデル（multiplicative model）と呼ばれている。本章ではこれらの2つのモデルを中心に取り上げて，要因分析やモデル選択の現状について述べる。また，最近多く用いられつつある混合効果モデルについても簡単に触れ

る。

2-2-1. CPUE モデル

Robson（1966）以来，CPUE に関して対数正規誤差を仮定した，いわゆる CPUE-LogNormal モデルが使用されることが多く，このモデルは（2.1）式のように表現される。データ解析において良く使用される i.i.d.（independent and identical distributed：独立で同一な分布に従うという意味）の条件を CPUE 標準化においても仮定することが一般的であり，本論文についても特に断わらない限りこの i.i.d. の仮定を置くこととする。

$$\mathrm{E}[\mathrm{Log}(\mathrm{CPUE})] = (\mathrm{Intercept}) + (\mathrm{Year}) + (\mathrm{Area}) + \cdots + (\mathrm{EMT}) + (\mathrm{Interactions}) \quad (2.1)$$

但し Log（CPUE）～ N（μ，σ^2）とし，E [ ] は期待値を表す。μ，σ^{2i} はスカラーの未知母数であり，それぞれ Log（CPUE）が従う正規分布の平均，分散を表す。

（2.1）式での（Year）と（Area）は年や海区の効果を表し，（EMT）は漁船に整備されている装置類や環境要因等の効果を総称して表現している。CPUE 標準化においては分布や回遊，系群などの情報をもとに海区分けを行うことが一般的であるが，区分された各々の海区内では資源が均一に分布していることを仮定している。これらの要因はカテゴリカル変数として扱われることが多いが，（EMT）については連続変数として用いられることもある。説明変数がカテゴリカル変数のみの場合と連続変数のみの場合は，それぞれ分散分析モデル・回帰分析モデルに対応しており，両者が混在している場合には共分散分析モデルとなる。

このモデルは取り扱いが容易なことと，観測誤差が CPUE の絶対値に比例するという仮定が合理的であると考えられていることもあり，まぐろ類の CPUE 解析において様々な国際漁業委員会等で広く使用されている。CPUE-LogNormal モデルを用いた CPUE 標準化の例は，古くは Robson（1966）に始まり数え切れないほど多く存在する。著者もこの CPUE-LogNormal モデルを使用して幾つか CPUE 標準化を行っている（Shono and Ogura, 1999；Shono *et al.*, 2000：2002）。

なお，ゼロキャッチ（CPUE=0となるデータ）が存在する場合には CPUE の自然対数を取ることが出来ず，(2.1) 式のままでは計算が不可能である。そこで，解析を行うための幾つかの方法が提案されており，この問題については2-4節で議論する。

例2-1. CPUE モデルの例（Shono *et al.*, 2002）

CPUE-LogNormal モデルの一例として，日本のはえ縄船によるインド洋キハダのデータを用いた CPUE 標準化モデルを取り上げる。この例では，緯度と経度を 5 度毎に区切ったセルを 1 つの単位とし，月毎に集計したデータを使用している。

$$\begin{aligned}\mathrm{Log}(\mathrm{CPUE}_{ijkl}+0.1) &= (\mathrm{Intercept}) + (\mathrm{Year})_i + (\mathrm{Month})_j \\ &+ (\mathrm{Area})_k + (\mathrm{Gear})_l + (\mathrm{SST}) + (\mathrm{SOI}) + (\mathrm{Year}^*\mathrm{Area})_{ik} \\ &+ (\mathrm{Month}^*\mathrm{Area})_{jk} + (\mathrm{Month}^*\mathrm{Gear})_{jl} + (\mathrm{Area}^*\mathrm{Gear})_{kl} \\ &+ (\mathrm{Area}^*\mathrm{SST})_k + (\mathrm{Area}^*\mathrm{SOI})_k + \mathrm{Error}_{ijkl} \end{aligned} \quad (2.2)$$

但し，$\mathrm{Error}_{ijkl} \sim N(0,\sigma^2)$とする。

(2.2)式での各々の変数の定義は以下のようになり，SST と SOI は連続変数として，その他の説明要因はカテゴリカル変数としてモデルに組み込んでいる。また，全ての CPUE に対して微小量（0.1）を加えているのは，ゼロ・キャッチデータに対して自然対数を取ることが出来ない欠点を回避するためのものである（2-4節）。

CPUE：Catch（キハダの漁獲尾数）/ Effort（はえ縄の針数：1000本を 1 と換算）
Intercept：切片項
Year：年の効果
Month : 月の効果
Area : エリア（海区）の効果
Gear : 枝縄数（number of hooks between floats：NHF）の効果
SST : 表面水温（sea surface temperature）の効果
SOI : 南方振動指数[ii]（southern oscillation index）の効果
Year*Area: 年とエリアの交互作用（以下同様にして記号 * は交互作用を表す。）

また，各々の添字の定義は以下のようになり，月毎の41年分のデータに対して操業海域全体を 6 つに分けたサブエリアを使用しており，枝縄数についても 4 つのクラスに分類している。

i（Year）: 1960-2000,
j（Month）: 1-12,
k（Area）: 1-6,
l（Gear）: 1-4（class 1: 5-6, class 2: 7-10, class 3: 11-14, class 4: 15-24）.

なお，（2.2）式のモデルは解析の最初に仮定したものである。実際には統計的な方法に基づいて各々の要因効果の取捨選択を行う必要があり（2-2-4節），説明変数を 1 つずつ減らしていくバックワードなステップ

ワイズ検定（2-2-4節）により最終的に選択されたモデルは，(2.2) 式から SOI 指標の主効果のみが除かれ，(2.3) 式のようになった。

$$\mathrm{Log}(\mathrm{CPUE}_{ijkl}+0.1)=(\mathrm{Intercept})+(\mathrm{Year})_i+(\mathrm{Month})_j+(\mathrm{Area})_k+(\mathrm{Gear})_l+(\mathrm{SST})+(\mathrm{Year}^*\mathrm{Area})_{jk}+(\mathrm{Month}^*\mathrm{Area})_{jk}+(\mathrm{Month}^*\mathrm{Gear})_{jl}+(\mathrm{Area}^*\mathrm{Gear})_{kl}+(\mathrm{Area}^*\mathrm{SST})_k+(\mathrm{Area}^*\mathrm{SOI})_k+\mathrm{Error}_{ijkl} \quad (2.3)$$

2-2-2. Catch モデル

Reed (1996) をはじめとして，Catch に対して Poisson 分布もしくは負の二項分布（negative binomial: NB）を仮定したモデルが近年用いられており，(2.4) 式のように定式化される。

$$E[\mathrm{Catch}]=(\mathrm{Effort})^*\exp\{(\mathrm{Intercept})+(\mathrm{Year})+(\mathrm{Area})+\cdots+(\mathrm{EMT})+(\mathrm{Interactions})\} \quad (2.4)$$

但し Catch ～ Poisson（λ）あるいは Catch ～ NB（α，β）とする。ここで，漁獲尾数は努力量に比例することを仮定しており，この（Effort）の項は offset 変数[iii]と呼ばれる。

このモデルでの応答変数は整数値を取る必要があることから，Catch としては重量ではなく尾数が用いられる。1990年代後半までは Catch-Poisson モデルが主流であったが，最近は負の二項分布モデルが使用されることも多い。この理由としては，統計パッケージ SAS[iv] の最新バージョン（Ver.8.2）で負の二項分布が標準装備されたことが挙げられる。

Poisson 分布モデルでは，平均と分散が同じであるという仮定が現実の状況にそぐわない場合もあり，多くは over-dispersion parameter ϕ を使用している (Okamoto *et al.*, 2003)。すなわち，確率変数 X がパラメーター λ を持つ Poisson 分布に従う場合には，その平均と分散をそれぞれ λ，$\phi\lambda$ とする必要があった。しかし，over-dispersion parameter を導入すると想定する確率分布を何にしたかということを表現出来ないため，通常の最尤推定は使用出来ない。

そこで，疑似尤度（Quasi-Lilelihood）と呼ばれるフレームワークによりパラメーター推定が行われているが（Bayler, 1993），理論的にはかなり複雑になる（Wedderburn, 1974; 椿，1988）。このパラメーター推定の問題を回避するための１つのオプションとして，over-dispersion Poisson モデルの代わりに（平均と分散が異なる）負の二項分布モデルを使用する方法が挙げられる。

2-2-3. CPUE の年トレンド抽出と要因分析

CPUE 標準化における主な目的は，資源量（資源密度）の年変動に対応する部分を取り出すことにある。交互作用を含まない場合には，CPUE モデルと Catch モデルのいずれにおいても年（Year）効果の推定値をそのまま取り出せば良い。しかし，年効果の交互作用を含む場合には，その解釈が難しい。そこで，交互作用を含む場合における要因分析のひとつの考え方として，LSMEAN (least squared mean) が挙げられる。まぐろ類の CPUE 標準化では，通常は年効果の LSMEAN を計算することによって CPUE 年トレンドを抽出しており（Shono *et al.*, 2002 : Matsunaga *et al.*, 2003），定義は（2.5）式で与えられる。

$$\mathrm{CPUE}_i=\exp\{(\mathrm{Intercept})+(\mathrm{Year})_i+\overline{(\mathrm{Year}^*\mathrm{Area})}_i+\overline{(\mathrm{Year}^*\mathrm{Season})}_i+\ldots\} \quad (2.5)$$

但し

$$\overline{(\mathrm{Year}^*\mathrm{Area})}_{i\cdot}=\frac{1}{N_j}\sum_{j=1}^{N_j}(\mathrm{Year}^*\mathrm{Area})_{ij},\ \overline{(\mathrm{Year}^*\mathrm{Season})}_{i\cdot}=\frac{1}{N_k}\sum_{k=1}^{N_k}(\mathrm{Year}^*\mathrm{Season})_{ik}$$

などとする。これらの平均化された項は，単純平均ではなく各々のセルに属するデータ数に応じた重み付け平均とすることもある。

なお，(2.5) 式では変数（Year）の主効果とそれを含む交互作用についてのみ考慮すれば良い。また(2.6)式で表される（Year*Area）の LSMEAN（高橋ほか，1989）を考えることにより海区別の CPUE 年トレンドを抽出している（Shono *et al.*, 2002）。

$$\mathrm{CPUE}_{ij}=\exp\{(\mathrm{Intercept})+(\mathrm{Year})_i+(\mathrm{Area})_j+(\mathrm{Year}^*\mathrm{Area})_{ij}+\overline{(\mathrm{Area}^*\mathrm{Season})}_{j\cdot}+\ldots\}$$

但し $\overline{(\mathrm{Area}^*\mathrm{Season})}_{j\cdot}=\frac{1}{N_k}\sum_{k=1}^{N_k}(\mathrm{Area}^*\mathrm{Season})_{jk}$

などとする。 (2.6)

LSMEAN の考え方を利用すると，交互作用を含む場合にも（2.5）式を用いて漁船に装備されている操業機器や環境要因などの効果が検出可能であり，要因分析が行える。(Shono *et al.*, 2000)。

$$\frac{CPUE_2}{CPUE_1} = \frac{\exp\{EMT_2 + \overline{(Area*EMT)}_{\cdot 2} + \overline{(Season*EMT)}_{\cdot 2} + \ldots\}}{\exp\{EMT_1 + \overline{(Area*EMT)}_{\cdot 1} + \overline{(Season*EMT)}_{\cdot 1} + \ldots\}} \quad (2.7)$$

但し

$$\overline{(Area*EMT)}_{\cdot 1} = \frac{1}{N_j}\sum_{j=1}^{N_j}(Area*EMT)_{j1}, \overline{(Season*EMT)}_{\cdot 1} = \frac{1}{N_k}\sum_{k=1}^{N_k}(Season*EMT)_{k1}$$

などとする。(2.7) 式においても変数（EMT）の主効果とそれを含む交互作用についてのみ考えれば良い。

なお，(2.7) 式では説明変数（EMT）が離散の場合を考えているが，連続変量の場合は推定されたパラメーターが 1 単位当たりの CPUE に対する変化量ととらえることが可能であり，回帰分析などの場合と同様にして要因分析を行うことが可能である（Shono *et al.*, 2002）。

2-2-4. 変数選択とモデル比較

本節では，同一モデルにおける変数選択（要因の取捨選択）と複数モデル間におけるモデル比較について取り上げる。

前者（変数選択）は，モデルをひとつ固定したときに各々の説明要因が CPUE に影響を与えているか否かを統計的な手法に基づいて判断する作業であり，ステップワイズ検定や情報量規準を用いて行われることが多い。例えば，例2.1では SOI 指標の主効果が CPUE に影響を与えていないことがステップワイズ検定により確かめられたため，この項を残しておくことは統計的には意味を持たないことになる。このようにして，応答変数である CPUE に影響を与えている要因効果のみを最終的に説明変数として含めることが一般的であり，本研究では同一モデルにおける変数選択と呼ぶことにする。

後者（複数モデル間におけるモデル比較）は，候補となる複数のモデルを比較して一番良いモデルを選択する作業であり，CPUE モデルと Catch モデルの比較に代表される統計モデル間の比較のみならず，一般化線形モデルとニューラルネットワークの比較など統計的なアプローチとデータマイニング的なアプローチの比較を考える場合も含まれる。本報告では主に統計モデル間の比較（CPUE モデル vs. Catch モデル）を想定しているが，現状ではこれといった有効な方法がなく，残差プロットなどを見て主観的に判断することが多い。

最初に変数選択に際して，次の 2 つの方法が広く用いられている。1 つは，階層構造を持つモデルに対して使用可能な，尤度比に基づいた Deviance や Pearson カイ二乗統計量（Dobson, 1990）に基づくステップワイズ検定であり，CPUE モデルではＦ検定が，Catch モデルではカイ二乗検定が用いられる（Shono and Ogura, 1999 ; Okamoto *et al.*, 2003）。ステップワイズ検定は総当たり法でないため，変数の数が多い場合にも計算の手間が比較的少なくて済む一方，検定のパスによって最終的な結果が異なる可能性がある（例2.2参照）（庄野，2000）。

もう 1 つは情報量規準による方法であり，水産資源解析においては AIC（Akaike's information criterion: Akaike, 1973）のみが長い間使用されてきた。しかし近年では一致性を持つ BIC（Bayesian information criterion: Schwarz，1978）や AIC に有限修正を施した c-AIC（finite correction of AIC: Sugiura，1978）などの情報量規準も状況に応じて用いられるようになってきている。特に c-AIC は小標本の場合のパフォーマンスが良いこともあり（Shono, 2000），国際委員会の ICCAT で良く使用されている。また，他の一致性を持つ HQ（Hannan and Quinn, 1979），BIC と漸近的に同等である MDL 規準（minimum description length criterion: Rissanen, 1983），真のモデルが候補となるモデルを含まない場合の選択パフォーマンスを改善する TIC（Takeuchi's information criterion: 竹内，1976），AIC におけるペナルティ項の係数を変更した MAIC（Bozdogan, 1987）などの情報量規準も今後使用される可能性があると考えられる。なお，庄野（2001）は正規誤差を持つ一般化線形モデル（回帰分析モデルや分散分析モデルを含む）において，TIC が AIC と漸近的に同等になることを証明した。

情報量規準による変数選択は，総当たり法で考えると一番良いモデルを一意に決めることが可能である。その一方，説明変数の数（主効果や交互作用など全て含む）が多い場合には，計算量が膨大となる欠点を持つ。また，over-dispersion Poisson モデルなど疑似尤度の枠組みで考える場合には，AIC など通常の情報量規準は使用することが出来ない。疑似尤度に関する情報量規準として，Burnham and Anderson (1998) は Q-AIC（Quasi-AIC）を提案したが，その信頼性や妥当性には疑問の余地が残っていることもあり，広くは知られていない。そのため，over-dispersion parameter を仮定した場合にはフルモデルと候補となるモデルの対数尤度比統計量をベースにした Deviance（逸脱度）や Pearson 統計量に基づくカイ二乗ステップワイズ検定を用いるのが一般的である。

なお，多くの主効果や交互作用を含む場合，ステップワイズ検定や情報量規準の値に基づいて主効果を取り除いた際に，関連する交互作用も一緒に取り除くべきか含めるべきかについては議論の分かれるところであり，ケース・バイ・ケースで判断することが多い。一般には，医薬品統計分野などでは薬の副作用の観点から残しておくことが多く，工学分野などでは推定精度の観点から取り除くことが多いと思われる。また，CPUE-LogNormal モデル（共分散分析型モデル）におけるステップワイズ F 検定と情報量規準 AIC との間には漸近的な関係が成り立っている（庄野，2000）。

次に複数モデル間におけるモデル選択の問題として，CPUE 標準化のための代表的なモデルである CPUE-LogNormal モデル（式（2.1））と Catch-Poisson モデルまたは Catch-NegativeBinomial（式（2.4））を比較する問題が挙げられる。現実問題としては，CPUE-LogNormal モデルと Catch-Poisson モデルの両方を使用した CPUE 標準化の計算を行い，標準化残差の傾向をチェックすることが多い。しかし，その比較検討はあくまで主観的なものに過ぎない。そこで，Shono（2001）は CPUE-LogNormal モデルの応答変数を Catch に変更することにより応答変数を揃え，両者の情報量規準による比較を可能にした。

$$E[\log(\mathrm{Catch})] = \log(\mathrm{Effort}) + (\mathrm{Intercept}) + (\mathrm{Year}) + \cdots + (\mathrm{EMT}) + (\mathrm{Interactions}) \quad (2.8)$$

但し log（Catch）～ N（μ，σ^2）とする。

ただし，over-dispersion parameter を持つ Catch-Poisson モデルでは AIC などの一般的な情報量規準が使用出来ない。また Q-AIC と AIC との比較は意味を持たないため，このままでは応答変数を変更した上の CPUE-LogNormal モデルとの比較が難しい。その場合には，Poisson 分布モデルの代わりに負の二項分布を用いることが 1 つの解決策になる。なお，ゼロ・キャッチが存在する場合には，応答変数に対して一律に定数項を足し込む方法では（2.8）式に対応するモデルが複数考えられるため，2-4-1節で再び取り上げる。

例2-2. CPUE モデルにおける変数選択の例（庄野（2000）から抜粋して引用）

本報告では，Hilborn and Walters（1992）による CPUE 標準化の仮想例（Table 2-1, データは多少加工している）を取り上げて，2 元配置分散分析型のシンプルな CPUE モデルにおける変数選択の実際について述べる。

まず，候補となる（2.9）式の 4 つのモデル（Model-1）～（Model-4）を考える。本来であれば，その他に（2.10）式で表現される候補モデル（Model-5）を考えることが可能である。しかし，（Model-5）ではデータ数よりもパラメーター数の方が多くなってしまって推定が不可能なため，ここでは候補となるモデルから外した。

Model-1: Log（CPUE）= Intercept + Error
Model-2: Log（CPUE）= Intercept + Year + Error
Model-3: Log（CPUE）= Intercept + Class + Error （2.9）
Model-4 Log（CPUE）= Intercept + Year + Class + Error
但し Error ～ N（0, σ^2）とする。
Model-5: Log（CPUE）= Intercept + Year + Class + （Year*Class） （2.10）

（2.9）式の 4 つのモデル間の包含関係は式（2.11）のようになり，Backward に変数を減らしていく形のステップワイズ検定におけるパスは 2 通り考えられる（（2.12）式）。

Table 2-1. Virtual data for CPUE standardization（loosely based on the data. Hilborn and Walters, 1992）. The values show catch rate（tons per hour）for three classes of vessel in four different years

Year	Class-1	Class-2	Class-3
1	0.63	1.03	1.22
2	0.48	0.56	1.26
3	0.33	0.67	0.89
4	0.54	0.48	1.01

(Model-1) ⊂ (Model-2)
　∩　　∩
(Model-3) ⊂ (Model-4)　　(2.11)

Path-1: (Model-4) → (Model-2) → (Model-1)
Path-2: (Model-4) → (Model-3) → (Model-1)
　　(2.12)

ステップワイズ検定の具体的な手順として，(Path-1) では最初に (Model-2) が真という帰無仮説に対して (Model-4) が真という対立仮説を考え，帰無仮説が棄却されたら (Model-4) を選択し，棄却されなければ次のステップに進む。次に (Model-1) が真という帰無仮説に対して (Model-2) が真という対立仮説を考え，帰無仮説が棄却されたら (Model-2) を選択し，採択されたら (Model-1) を選択する。

このようにして，有意水準を1％としたステップワイズF検定により変数選択を行うと，Path-1では (Model-1) が選択された（庄野，2000)。(Model-1) は年によるトレンドの違いを持たない非常に単純なモデルである。Path-2についても同様にして考えれば良く，結局 (Model-3) が選択された（庄野，2000)。例2-2では，検定のパスによって最終的な変数選択の結果が異なってしまう。このようなステップワイズ検定におけるパスの一意性が問題になるケースでは，AIC に代表される情報量規準を用いることが1つの解決策になる。

情報量規準による変数選択では，候補となる全てのモデルに対して AIC など情報量規準の値を計算し，その値が一番小さくなるモデルを選択すれば良い。Table 2-2に示した AIC の値から判断すると，(Model-4) が最終的に選択された（庄野，2000)。今回の例では，ステップワイズF検定による結果と情報量規準 AIC を用いた結果が一致しないが，一般にはこのような手順にてステップワイズ検定や情報量規準によるモデル選択（変数選択）を行うことが可能である。

Table 2-2. Result of the model selection using AIC on the data of Table 2-1

Model	AIC
1	16.445
2	20.222
3	6.273
4	2.666

2-2-5. 混合効果モデル

CPUE-LogNormal モデルや Catch-Poisson モデルにおける各々の要因は，通常は固定効果として扱われる。しかし，場合によっては (Year)，(Area) などの主効果や交互作用を変量効果 (random effect) と考える場合もある (Verbeke and Molenberghs 1997; Little *et al.*, 1996; Searle *et al.*, 1992)。変量効果とは，該当する説明要因がある母集団からランダムに選択された標本であると考える効果であり，要因効果自身を確率変数と見なしている。これに対して，通常の分散分析型モデルでは，解析するデータ（要因実験でとられた因子水準）のみに推論の興味があり，この場合に含める主効果や交互作用を固定効果 (fixed effect) と呼んでいる。固定効果と変量効果は厳密に区別出来るものではなく，解析者が推論するときの関心によって仮定されることがある。なお，要因のいくつかが固定効果であり，それ以外のものが変量効果であるときのモデルを，一般には混合効果 (mixed effect) モデルと呼んでいる。要因のすべてを固定効果と設定した分散分析型モデルではカテゴリカル変数の水準数が大きくなることもあり，パラメーター数が多いゆえに推定精度が悪くなりがちであるが，変量効果を取り入れることによりパラメーター数が減少し，推定結果が安定

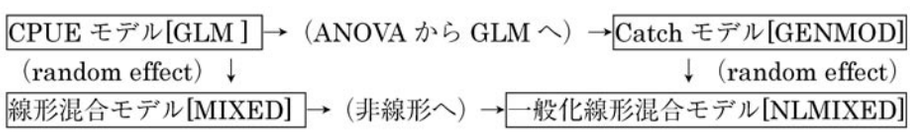

Fig. 2-1. Relationship among the models used for CPUE analysis and the corresponding SAS/STAT procedure.

することも多い。

まぐろ類のCPUE標準化に変量効果を導入する理由としては，要因効果自身を確率変数と見なすという本来の考え方ではなく，実用的な部分が大きいと考えられる。すなわち，漁業の分布などに基づいて分けられた海区を表すカテゴリカル変数と年の効果との交互作用（(Area)*(Year)）の存在が，まぐろ類の広い範囲に及ぶ時空間的な移動が認められることから一般に強く示唆されるにもかかわらず，データの欠損ゆえに計算出来ない場合（分けられたすべてのセルに対してデータが存在しない場合）に，これらの要因に関わる主効果や交互作用を混合効果に設定することが見受けられる（Yokawa and Shono, 2000）。

最近ではCatch-Poissonモデルや Catch-NBモデルなどの説明変数に対して変量効果を仮定するモデル，すなわち連結関数を恒等写像以外に設定して正規分布以外の確率分布を仮定するという，いわゆる一般化線形モデルのフレームワークの中での混合効果モデル（Fahrmeir *et al.,* 2001）も使用されるようになってきている。なお，固定効果モデルと混合効果モデルの関係を図示すると，Fig. 2-1のようになる。

2-2-6. その他の問題

CPUE標準化への応用が考えられるその他の統計手法としては，構造方程式モデリング（共分散構造分析）（Bollen, 1989；狩野，三浦，2002；豊田，1998a; 2000; 2003）が挙げられるが，現在のところこれら手法の水産分野への適用例はほとんど存在しない。著者らは現在構造方程式モデリングを利用して漁船に装備されている装置類の効果推定に取り組んでおり，（操業機器・探索機器・漁船情報の３つの潜在因子を設定して）どの装置類がCPUEに影響を与えているかを測定している。なお，Bayes流アプローチを利用した一般化線形モデルのCPUE標準化への適用事例も報告されている（Badcook and Mcallister, 2001）。

また，著者らのCPUE標準化では漁船に装備されているソナーなどの探索機器や低温畜養装置などの操業機器，漁船や漁具の規模，表面水温や塩分濃度などの環境要因などがCPUEに与える影響について検討を行ってきた（Shono and Ogura, 1999, 2000; Shono *et al.,* 2000）。その結果，多くは過去の知見と一致したが，一部は機器を装備しない場合のCPUEが高くなるという負の効果も認められ，不自然な部分も残ってしまった。これは，ある意味で一般化線形モデルの限界を示していると考えられないこともない。そこで，このような過去の知見と矛盾する負の効果の問題を解決する手段として，漁船の番号を（カテゴリカル変数として）要因に取り込むことや一般化線形モデルの代替としてデータマイニング手法が考えられており，後者については2-3節で取り上げる。

2-3. データマイニング手法

これまでの一般化線型モデルに代表される統計手法に加えて，近年樹形モデルやニューラルネットワークなどのデータマイニング手法（Berry and Linoff, 1997; Hastie *et al.,* 2001; Witten and Frank, 2000; 内田，2002）が，Watters and Deriso (2000) をはじめとしてCPUE標準化に使用されるようになってきている。その理由として，数学・物理学・情報科学などから水産資源解析に転向した研究者によってこれらの手法が紹介されたことが挙げられるが，統計モデルによる標準化された年トレンドなどが現状にそぐわないことも多く，観測誤差などのバイアスの大きいデータに対応する斬新かつ有効な手法の登場が期待されていたという背景も一因であると考えられる。

なお，本博士論文でのデータマイニング手法の定義として，与えられた大量のデータの中から有用な規則や仮説を発見するために行われる探索的な方法を表すこととし（寺野，2002），統計モデルを構築してからデータをモデルに当てはめるという検証的なアプローチと区別することとする。この寺野(2002)による分類，すなわち検証的な手法を統計モデルとし，探索的な方法をデータマイニングと分類する定義は一般的なものではなく，あくまで人工知能的な立場からの１つの考え方である。異なる考え方の一例として，Tukey(1977)が提案したEDA（exploratory data analysis：探索的データ解析）が挙げられる。EDAはモデルを仮定する前に現実的な立場でデータの示唆する情報を多面的に捉えるという頑健な探索的手法であるが，統計モデルに分類可能である。しかし，本博士論文では，検証的，探索的な方法をそれぞれ統計モデル，データマイニングと分類する立場に基づいて議論を展開する。なお，データマイニング的なアプローチを大別すると教師付き学習と教師無し学習[v]に分けられるが，本節では教師付き学習について主に取り上げる。

本節では，まずCPUE標準化への応用例が存在するデータマイニング手法である樹形モデル（tree-regression models: TRM）とニューラルネットワークについて取り上げて，CPUE標準化における適用の現状と問題点について説明する。また，一般化線形モデルとの比較も含めて議論されることの多い一般化加法モデル（generalized additive models: GAM）について，先行研究などの紹介を行う。なお，今後の使用が考えられる可能性のあるデータマイニング手法

についても，本節の最後で簡単に取り上げる。

2-3-1. 樹形モデル

樹形モデルは，ある基準に従ってデータセットを複数のサブセットに次々と分割していく手法であり，応答変数が単一である場合の分類問題（判別問題）や回帰問題[vi]に有用である（Kass, 1980）。樹形モデル分析によって得られたルールは，一般的な高級言語やSQLのようなデータベース言語によって簡単に表すことが出来る。

樹形モデルの代表的な計算アルゴリズムとしては，CART（Breiman *et al.,* 1983; 大滝ら，1998），C5.0（Quinlan, 1993），CHAID（Hartigan, 1975）などが挙げられるが，大別するとCART系（CART, C5.0など）とCHAID系（CHAIDなど）に分けられる。これらの主な違いとしては，CART系では量的変数を連続のままで処理を行って最終的な分岐後に枝刈りを行うのに対し，CHAID系では量的変数を内部でカテゴリカル変数に変換してから計算を行って分岐する際にカイ二乗検定を用いてその必要性を判断することが挙げられる。CART系アルゴリズムでは分岐の際の多重比較の問題が未解決のままであるため，統計研究者は一般にCHAID系アルゴリズムを好む傾向にある。

この樹形モデルを一般化線形モデルに代表される統計モデルの代替としてCPUE解析に用いることに対しては，次のようなメリットがあると考えられる。

- いかなる統計分布も仮定することなく，柔軟なモデリングが可能である。
- 重要な要因（説明変数）を自動的に抽出して，グループ化を行う。
- 欠損データに対してロバストである（まぐろ類の広範な時空間的な分布パターンの変化ゆえに極めて有意であると考えられる（Year)*（Area）の交互作用を一般化線形モデルでは欠損データのために検出できないことが見受けられるが，樹形モデルでは原則として交互作用について考慮する必要がない）。
- 一般化線形モデルが持つ技術的な問題（ゼロ・キャッチの問題・モデル比較の問題（CPUE-LogNormal vs. Catch-Poisson など））が解決される。
- 統計分析を実行する前の探索的段階で使用した場合に，データの非線形性や交互作用が検出可能である。具体的には，CPUE解析を行う前にエリア分けに対して樹形モデルが適用可能であり，その場合にはCPUEでなく努力量を出力変数に設定することも可能である（Shono *et al.,* 2001）。

そのため，Watters and Deriso（2000）に始まり，樹形モデルを用いたCPUE標準化は，これまでにもいくつかの解析例が紹介されている（Shono *et al.*, 2001; Venables and Toscas, 2002）。しかし，上のような長所が存在する一方，使用するアルゴリズム（CHAID vs. CART）によるCPUE年トレンドの違い（Shono *et al.,* 2001），どのようにしてCPUEの年トレンドを抽出するか，量的な応答変数に対して推定値が離散になってしまうなどの問題点も残っている。

離散推定値の問題点を克服するために，樹形モデルに対するバギングの使用（櫻井，1998）やファジーメンバーシップ関数を用いた分類（竹澤，1999）など新しい試みも見られる。また，Venables and Toscas（2002）はバギングを用いたミナミマグロCPUEの標準化を行っており，今後は要因分析の計算方法も含めて，より良いモデルの開発ならびに改良が望まれる。

2-3-2. ニューラルネットワーク

ニューラルネットワーク（麻生，1988; Haykin, 1994; Smith, 1996; 吉富，2000）とは，複雑な環境をモデル化する際に良く使用されている脳や神経系の仕組み（生物の神経細胞の回路で行われている情報処理システム）をモデル化したものであり，現在では様々な社会問題や工学上の問題などに適用されている。具体的には，多数の素子が互いに適当な重みを持って結合されたモデルであり，結合の種類によって相互結合型ニューラルネットワーク（各素子が全て相互に結合されたもの）と階層型ニューラルネットワーク（素子が幾つかの層をなして配列されており，各層間の素子は互いに結合されているが，層内の素子は結合されていないもの）に大別出来る。

ユーザーの立場からすれば，教師無し学習アルゴリズムと教師付き学習アルゴリズムとで分類する方が理解し易く，前者の代表的な例としてSOM（self-organization maps，自己組織化マップ：Kohonen, 1989）が有名であり，後者の代表的な例としてはRBFN（radial basis function network, 動径基底関数ネットワーク：Broomhead and Lowe, 1988）やMLEBPN（multi layer error back propagation network, 多層誤差逆伝搬ネットワーク：Rumelhart *et al.,* 1986）が挙げられる。SOMは低次元の格子によって構成された出力層が存在しないニューラルネットワークであり，クラスター分析の代わりに使用されることも多い。RBFNはデータが似たような性質を持つ幾つかのクラスターに分割出来るような場合に有効な方法であり，多変量正規混合分布モデルや多変量正規密度をカーネル関数とする確率密度関数推定とも密接な関係がある。しかし，ソフトウェアなどの対応状況などもあり，教師付き学習アルゴリズムにおいて一般

に広く用いられているのはMLEBPNモデルである。

MLEBPNとは入力層と出力層の間に任意の個数の中間層（隠れ層）を設ける方法であり，入力に対する出力と望ましい出力（教師信号）との誤差を減らすようにニューラルネットワークの各素子の結合加重を修正していく。その際に情報が出力層から入力層に向かって逆方向に伝わっていくことが特徴的である。MLEBPNでは，中間層の数を1つに設定することが多く，その中間素子数を変更させることによって非線形関数に対する高い関数近似能力を持つ。その上，入出力データに対する制限が少なく（入力・出力ともに離散変数と連続変数の両方が扱える）回帰問題と分類問題の両方が計算可能であり，多くのソフトウェアが開発されていて実行も容易である。後述する水産分野への応用のほとんどがMLEBPNを使用しており，本研究ではこのアルゴリズムのみを取り上げることにする。なお，ニューラルネットワークと統計モデルとの関係については，佐藤（1996）や豊田（1998b）などに詳しく述べられている。

ニューラルネットワークの水産分野への応用としては，加入量・資源量予測や魚種判別の適用例が幾つか見受けられるが（青木，小松，1992; Haralabous and Georgakarakos, 1996; Chen and Ware, 1999など），CPUE標準化への適用例はそれほど多くない。Shono（2002）は，ミナミマグロ資源における過去に漁業が存在して現在は存在しない部分のCPUE予測を3層のエラーバックプロパゲーションを仮定した教師付き学習の典型的なニューラルネットワーク（MLEBPN）を用いて行っている。

ニューラルネットワーク（MLEBPN）においては，入力（原因）と出力（結果）の関係を表す式を記述する必要がなく，非線形かつ柔軟なモデリングが可能であるが，一方では要因分析や（中間層の数の設定をも含めた）モデル選択の難しさなどの問題があり，ブラックボックス的な側面も併せ持っている。また，学習速度が遅く，過学習（over-fitting）が起こりやすいことも指摘されている（Repley, 1994）。要因分析については，説明変数の数が少なくかつカテゴリカル変数の水準数が少ない場合（カテゴリカル変数を想定しているが，連続変数の場合にはカテゴリーに置き換える必要がある）には，添字を固定して平均的な予測値を比較することが可能であるが，一般には難しい。ルール抽出に基づく方法も一部のソフトウェアで実装されているが，実用化の域には達していない感もある（Tsukimoto, 2000; 月本，森田，2000）。そのため，CPUE標準化の主目的である年トレンドの抽出に対する定まった方法が存在しないという問題は残ったままである。

なお，中間層の数の推定に関して，認定可能性の問題ゆえに最尤推定量の漸近的な性質が成立しないため，情報量規準を用いたモデル選択が理論的には不適切である。そのため，クロスバリデーションなどに基づく有効な推定方法の開発が望まれる。

2-3-3. 一般化加法モデル

CPUE標準化での統計モデルの使用は，1990年代中頃まで一般化線形モデル（GLM）一辺倒であった。しかし，最近になって平滑化スプライン（smoothing-spline: Whittaker, 1923）などの一般化加法モデル（GAM）（Hastie and Tibshirani, 1990; Simonoff, 1998）もCPUE標準化に使用されるようになってきている（Wise *et al.*, 2002）。一般化加法モデルは，一般化線形モデルにおける各々の要因効果と応答変数の線形な関係を非線形関数で置き換えたものと解釈することが可能であり，smoothing-splineや局所重み付き多項式回帰（locally weighed polynomial regression: Clevelend and Grosse, 1991）など多くの計算方法が提案されている。

一般化加法モデルを使用する利点としては，何と言っても柔軟なモデリングが可能となることが挙げられる。一般化線形モデルでは，CPUEと要因（説明変数）の間に直線関係あるいは指数関係（応答変数をLog（CPUE）と考えることが多いため）が仮定されるのに対して，一般化加法モデルでは非線形の関係を記述することが可能である。そのため，環境要因などに対して平滑化スプラインを用いることが有効である場合が多い。例えば，ある一定の水温帯でCPUEが高くなっておりその周りの水温帯でCPUE低くなっている場合には，その関係を一般化線形モデルで記述することは難しいが，表面水温に対してスプライン関数を使用することによって一般化加法モデルによる定式化が可能となる。一般に環境要因などに対して平滑化スプラインなどの手法を用いることが有効である場合が多い。その一方で，一般化加法モデルにおいては自由度の指定方法などの問題も残っている（竹澤，2001）。SASやS-Plusなどの統計パッケージによる計算も一般化線形モデルと比較して複雑であることから，理論と応用の両面においてこれらの問題点の改良が望まれる。

なお，一般化線形モデルにおいて従来カテゴリカル変数として組み込んでいた要因に対して非線形な仮定を取り入れること（例えば漁業海区を便宜的に区分したカテゴリカル変数の代わりに緯度・経度のスプライン関数を用いることなど）が場合によっては有効であ

ると思われる。しかし，水産資源分野での実際例はほとんど存在しない。

2-3-4. その他の手法

その他にCPUE標準化への応用が考えられるデータマイニング手法としては，グラフィカルモデリング，ベイジアンネットワークなどが挙げられる（Jenson, 2001; Edwards, 2000; 宮川，1997）。また，2値分類問題での有効な手法であり，多値判別問題や回帰問題に拡張されているSVM（Support Vector Machine，サポートベクターマシーン：Vapnik, 1998; 麻生ほか，2003）も将来的にはCPUE標準化に対する有効な方法の1つになるのではないか，と著者は考えている。現在のところこれらの手法の水産資源解析への適用例は少ないが，因果関係や相関関係を検出するためのデータマイニング手法の今後の発展が期待される。

2-4. 水産資源に特有の問題

本節では，CPUE標準化における漁業資源特有の問題について論じる。主に観測値がゼロであるデータが含まれている場合の取り扱いとエリアサイズによる重み付けの問題であり，その他としてまき網漁業での努力量の定義や説明変数として標準化に含める要因，ハビタットモデルなどについて取り上げる。

2-4-1. ゼロ・キャッチの取り扱い

CPUE標準化において広く用いられてきたCPUE-LogNormalモデルについて，応答変数であるCPUEに対して自然対数を取っていることから，CPUEがゼロとなるデータはLog（CPUE）＝－∞となってしまうのでそのままでは取り扱うことが出来ない。このことをゼロ・キャッチ（CPUE=0とCatch=0は同じ意味であることからこのような名前が付いたと考えられる）の問題と呼んでおり，標準化の計算を行うために，大きく分けて以下の2つの方法が使用されている。

1） 全てのデータに対して一定量（定数項）を足し込む方法
2） CPUEがゼロか否かを分けてからゼロキャッチ率をlogitモデル（（2.13）式）によって推定し，ゼロでない部分のみに対して通常のモデル（CPUEモデルやCatchモデルなど）を適用する方法（Delta型2段階法）

$$E[\log\{R/(1-R)\}] = (\text{Intercept}) + (\text{year}) + (\text{area}) + \cdots + (\text{interactions}) + (\log(\text{effort}))$$
但しR（ゼロ・キャッチ率）～Binomial（p）とする。 (2.13)

1）はユーザーに取って扱い易い反面，区間推定における偏りの原因になってしまう（点推定に関しては，推定値からこの定数項を差し引くことによって偏りを防ぐことが出来る）。また，一定量としてどのような値を取れば良いのか，という問題もある。現状では1）の方法が多く使用されており，国際委員会のICCATなどではCPUEに足し込む一定量として平均CPUEの10%が用いられている（ICCAT, 1997）。しかし根拠は不明であり，長所も感じられない。一般には微少量の方が区間推定への影響が少ないと感じられるが，データに依存しない形でこのことを示すことは困難である。

Shono（2001）では，CPUE-LogNormalモデルに対する（全てのCPUEデータに一律に足し込むための）値を，複数の候補の中から情報量規準を用いて決定出来ることを示した。また，尤度関数の形に着目してこの定数項の数値的最適化が可能であることを指摘した。この計算は，例えば統計パッケージのSASにおいてはORと呼ばれるprocedureの中のNLPステートメントを使用することによって実行可能となる。

$$E[\log(\text{CPUE}+k)] = (\text{Intercept}) + (\text{Year}) + (\text{Area}) + \cdots + (\text{EMT}) + (\text{Interactions}) \quad (2.14)$$
但し $\text{Log}(\text{CPUE}+k) \sim N(\mu, \sigma^2)$ とする。
$$E[\log(\text{Catch}+k^{*}\text{effort})] = \log(\text{effort}) + (\text{Intercept}) + (\text{Year}) + \cdots + (\text{Interaction}) \quad (2.15)$$
$$E[\log(\text{Catch}+\text{const.})] = \log(\text{effort}) + (\text{Intercept}) + (\text{Year}) + \cdots + (\text{Interactions}) \quad (2.16)$$

ただし，ゼロ・キャッチが含まれる場合のCPUEモデルとCatchモデルの比較（2-2-4節）に際しては，応答変数を揃えるに当たって注意が必要である。

すなわち，（2.14）式においてCPUE（=Catch/Effort）の定義を代入すると（2.15）式で表されるが，このモデルではCPUE年トレンドの抽出が難しい面も持っている。そのため，（2.14）式のCPUEをダイレクトにCatchに置き換えたモデル（（2.16）式）を用いることも1つの手段であるが，このモデルの仮定は（2.14）式で表されるモデルと異なっている。その上，Catchモデル（（2.4）式）では応答変数に定数項が含まれていないのに対し，CPUEモデルの変形版（（2.15）式または（2.16）式）では含まれておりモデルの微妙

な違いも認められる。

そのため，このゼロ・キャッチの問題に対して，出来る限り（以下に説明する）2）の Delta 型 2 段階法を用いるべきと著者は考えており，2）の方法を用いることが定数項を含まない形での CPUE モデルと Catch モデルとの比較を可能にする。

2)の方法では，最初の logit モデル（(2.13）式）と次の通常モデル（CPUE モデルまたは Catch モデル）を別々に計算することが CPUE 解析において多く行われており，Delta 型モデル（あるいは 2 段階型モデル）と呼ばれている（Lo, 1992 ; Stefansson, 1996)。なお，logit モデルのみを用いて CPUE 解析を行った例としては，Miyashita *et al.*（2000）が挙げられる。

この Delta 型アプローチは実務家にとって理解し易い部分があり，SAS などの統計パッケージによる解析も比較的容易に行われる。その反面，区間推定が難しい面があり，２つのモデルにおいて有意と認められた要因（説明変数）が異なる場合には，CPUE 年トレンドの推定がかなり複雑になる欠点も併せ持つ。２つのモデル（logit モデルと通常モデル）の尤度を１つに書き下して同時に推定することも理論的には可能であり，Zero-Inflat モデルと呼ばれている（Lambert, 1992; Ridoud *et al.*, 2001)。しかし，ソフトウェアによる計算手順が複雑なこともあり，水産資源分野における適用例はほとんど見受けられない。

2-4-2. 面積指数での重み付け

CPUE 標準化の第 1 の目的は年効果の推定であり，抽出された年トレンドは相対的な資源の増減傾向を表している。なぜなら，CPUE は資源量に比例すると考えられているからである。(CPUE と資源量の比例関係が線形か非線形か，あるいはどのような関数形で表すべきかという問題は古くから議論されているが，CPUE 標準化とは直接関係がないこともあり，本報告では割愛したい。)

相対資源量という立場で CPUE を捉えた場合，エリア分けが全て等しいサイズの場合には CPUE と相対資源量は同じことを表すが，個々の海区の大きさが異なり，なおかつ年とエリアの交互作用が認められる場合には，エリアサイズを考慮した補正が必要である。すなわち，推定された CPUE の年効果に相対的なエリアサイズを掛け合わせたものを資源量指数（abundance index: AI）と呼んでおり，通常は CPUE 推定値に基づいた資源量指数が相対資源量に対応すると考えられている（能勢ほか，1988; 山田，田中，1999)。

$$\mathrm{AI}_{ij} = w_j \mathrm{CPUE}_{ij}$$
$$\left(\text{但し}\sum_j w_j = 1 \quad (i:\mathrm{YEAR}、j:\mathrm{AREA}、w_j:\mathrm{AREA}(j)\text{の相対面積指数})\text{とする}\right) \tag{2.17}$$

この相対面積指数は，式（2.17）のように年に依存しないと考えることが多い。しかし，一般にまぐろ類は広範囲の時空間的な移動を行い，それに合わせて漁場が変化する場合も多い。そのため，年が経つにつれて漁場が縮小しているような場合に（2.17）式を用いて資源量指数を計算すると，資源の過大評価につながる恐れもある。

この問題に対して，ミナミマグロ漁業を取り扱っている国際委員会 CCSBT では，(2.17）式の方法の他に1990年代初めから（2.18）式のような相対面積指数が年に依存するという考え方を取り入れている (CCSBT, 1998)。すなわち漁業の変化に対してエリアサイズも変化させるという方法である。この考え方は (2.17）式の CS 仮説（constant square 仮説）に対して VS 仮説（variable square 仮説）と呼ばれている。

$$\mathrm{AI}_{ij} = w_{ij} \mathrm{CPUE}_{ij} \tag{2.18}$$
$$\left(\text{但し}\sum_j w_{ij} = 1\ (\mathrm{w_{ij}}:\mathrm{YEAR}_{(i)}\ \text{かつ}\ \mathrm{AREA}_{(j)}\ \text{の相対面積指数})\text{とする}\right)$$

現実問題として，ミナミマグロ資源の CPUE 標準化に際していずれの仮説を用いるかにより，1990年以降の相対資源量の年トレンドのみならずこれらをチューニングインデックスとして使用している資源評価モデルの結果が全く異なるものになってしまった。すなわち，CS 仮説によると資源状態は上向きであり，VS 仮説に基づくと資源は下方に向かっているという結論になる（Takahashi *et al.,* 2001)。

いずれの仮説が現実の状況に良く当てはまっているか, という問題については多く議論が行われているが, 本質は漁場の縮小が乱獲によって魚の存在するエリアが狭くなったのか，あるいはその他の原因によるものなのか（規制により操業エリアや漁船の数が減少したからなのか，(漁獲効率のみならず気象条件なども含めて）良い条件の場所での操業が多くなった結果として操業海域が狭くなったのか)，いずれに多く起因する問題なのか，ということに尽きる。

この問題に対して，Toscas and Thomas (1998）は混合効果モデルの一種である repeated measure の考え方に基づいて，過去に漁獲があって現在は漁獲

が無いエリアのCPUEを推定しており（論文では空間統計学の手法を用いて補間しているが，repeated measureによる推定と読み替えることが可能である），Shono (2002) はニューラルネットワークを用いて操業がないセルのCPUE予測を行った。

2-4-3. その他の話題

本節では，これまでに述べられなかった水産資源特有のCPUE標準化におけるその他の問題について，簡単に整理する。

・努力量の定義

まぐろ漁業における努力量として，はえ縄漁業や竿釣り漁業では針数（1,000hookを1単位とする場合が多い）や竿数を使用することが多く，国際委員会等においてもこれらを努力量の定義とすることに対しておおよそのコンセンサスが得られている。しかし，まき網漁業については，探索時間をどのようにカウントするか，あるいはFADsと呼ばれる魚を集めるための人工浮き漁礁の影響をどのように測定するか，など多くの問題点が指摘されており，努力量の定義自体が難しい側面を併せ持つ。実際問題として，まき網漁業のCPUEにおいては操業日数や操業回数を使用すること（O'Brien *et al.,* 1997; Shono *et al.,* 2000; Soto *et al.,* 2000; Soto *et al.*, 2002）が多く見受けられるが，これらの努力量の定義が漁業の現状にマッチしているかどうかについては注意深く検討していく必要がある。

・努力量の不均一性

一般に努力の絶対量が多い部分と少ない部分でその推定精度に差が出るのは，統計学的に考えれば当然のことである。操業の少ない部分を解析から除くことも一案であるが，この努力量の不均一性を解決するための1つのオプションとして，LSMEANを計算して要因効果を抽出する際にデータ数に応じて重み付けすること[vii]が考えられる（(2.2) 式）。

・説明変数として取り入れる要因の検討

一般化線形モデルなどを用いたCPUE解析においては，多くの説明変数が要因として組み込まれている。現状では，年（Year），海区（Area），季節（Season）（四半期もしくは月を単位と考える場合が多い）などに加えて，表面水温などに代表される環境要因，漁船に装備されている操業機器や探索機器などの装備類や漁船の規模などの要素，（はえ縄船における）枝縄数などの情報（Yokawa and Shono, 2000），あるいはターゲット種以外の漁獲の有無などを使用することもある（Shono and Ogura, 2000; Shono *et al.,* 2000）。ややもすると多重共線性などの問題が生じるため，モデルの初期設定（変数選択を行う前に仮定したモデル）の際には，説明変数間の相関関係について考慮すべきである。また，最初に仮定する要因効果の取捨選択に関しては，主効果のみならず交互作用の条件設定についても十分な注意が必要である。また，統計モデルにおける要因分析に際し，その単調性が想定される効果に関しては説明変数としての組み込み方，すなわち（連続値を有限個のカテゴリーに分解して）カテゴリカル変数と設定するか，それとも連続変数と指定するかについて，十分な検討を要する。

・CPUE解析に使用されるデータの種類

まぐろ類の商業船による操業データのCPUEは，通常は漁法・魚種・海域（系群）別に標準化されることが多い。これらの漁業データは大きく分けてshot-by-shotと呼ばれる操業毎のデータと（5×5度/月別などに）集計を施したデータに分類されるが，データの与える重みなどの問題もあり，一般には操業毎のデータが好ましいと考えられている。しかし，膨大な件数になることもあり，場合によってはデータの入力や加工段階で集計されることも多いため，操業毎のデータと集計されたデータがともに広く使用されている。通常は操業毎のデータが存在する場合にこれを集計する必要はないが，1つの入力セット（同じ添字を持つ説明変数の組）に対して複数の異なる出力（CPUE）が存在する場合もあるため，ニューラルネットワークなどにおいては注意が必要である。

・ハビタットモデル

1990年代後半から，ハビタットモデル（Hinton and Nakano, 1996）と呼ばれるはえ縄漁具の鉛直分布パターンと漁獲対象種の鉛直分布パターンを用いて，（水域別時期別の魚の鉛直分布確率と漁具のそれとの積の形で）有効努力量を直接推定するCPUE標準化の方法が，遠洋はえ縄漁業における一部の魚種・海域（南西太平洋や東部太平洋のまぐろ・かじき類など）で使用されている。具体的には，漁具の鉛直分布パターンをはえ縄に装着した小型水深水温計などのデータを基に，漁獲対象種の鉛直分布パターンは魚に装着したポップアップアーカイバルタグなどのデータを基にして推定されることが多い。この方法はかじき類など魚の鉛直方向の分布が非常に浅いところに偏っている場合などに有効であると考えられており，生物データを直接取り扱えるという長所がある。その一方，漁獲対象種の鉛直分布パターンのモデル化が難しいという欠点や，ハビタットモデルにより推定されたCPUEの年トレンドが一般化線形モデルによるそれと異なるケースも存在する（Yokawa *et. al.,* 2002）。そのため，今後ハビタットモデルと一般化線形モデルとの比較を含めた検討が必要である。

2-5. まとめ

本節では，以上のレビューを踏まえ，本研究で取り上げる重要な 3 つの問題について，水産資源解析，特に CPUE 標準化の観点から概説する。

最初は，CPUE に影響を与えている説明要因（主効果および交互作用）の中でどの変数が統計的に影響を与えているかどうかを判断するという問題であり，(2.19) 式で表現される CPUE-LogNormal モデル（共分散分析モデル）や (2.20) 式の Catch-Poisson（or Catch-NegativeBinomial）モデル（一般化線形モデル）を利用する場合には，AIC（Akaike's Information Criterion：赤池情報量規準）に代表される情報量規準が使用可能である。また，モデルがネスト構造を持つ場合には，F 検定やカイ二乗検定に代表される stepwise 検定も利用出来うる。

$$\begin{aligned}&\mathrm{Log}\,(\mathrm{CPUE}) = (\mathrm{Intercept}) + (\mathrm{Year}) + (\mathrm{Season}) + (\mathrm{Area}) + (\mathrm{EMT})\\&\cdots + (\mathrm{Two\text{-}way\ Interactions}) + \mathrm{error},\ \mathrm{error} \sim N(0,\ \sigma^2)\end{aligned} \tag{2.19}$$

$$\begin{aligned}&E\,[\mathrm{Catch}] = \mathrm{Effort}^{*}\exp\{(\mathrm{Intercept}) + (\mathrm{Year}) + (\mathrm{Season}) + (\mathrm{Area}) + (\mathrm{EMT}) + (\mathrm{Two\_wayInteractions})\},\\&\mathrm{Catch} \sim \mathrm{Po}(\lambda) \text{ または } \mathrm{NB}(\alpha,\ \beta)\end{aligned} \tag{2.20}$$

水産資源解析分野では，情報量規準 AIC のみが伝統的に使用されており，他の情報量規準の例は，これまであまり見受けられなかった。しかし，使用する情報量規準によって最終的なモデル選択（変数選択）の結果が異なることは多くあり，CPUE に影響を与える説明要因の解釈も異なってしまう。第 3 章では，まず実際例，漁業データを用いて，様々なケースにおける複数の情報量規準によりモデル選択を行い，説明要因の取捨選択結果が異なることを例示した。

さらに，CPUE 標準化に使用される一般化線形モデル（回帰分析モデルや分散分析モデルを含む）においても，小標本の場合や大標本の場合など，想定するケースに応じて適切な情報量規準を使用することが必須であると考えている。そこで，主に共分散分析モデルを用いた計算機シミュレーションを行い，具体的には複数の候補モデル（主効果や交互作用など説明要因を組み合わせたもの）の中から真のモデルを設定し，この正しいモデルに基づいて乱数を発生させる実験により，真のモデルを選ぶという選択パフォーマンスを，様々なケースについて測定した。また，ネストモデルにおける情報量規準と stepwise 検定の比較シミュレーションを行い，情報量規準の良さや stepwise 検定の有意水準について，詳しく検討した。

なお，これらのモデル選択が水産資源解析に与える影響としては，CPUE 標準化の主目的である年トレンド抽出が挙げられ，この問題は要因分析の問題として捉えられる。実際には，(2.21) 式で表現される LSMEANS (least squared means) などを利用して要因分析を行うことが多いが，使用する情報量規準（例えば AIC と BIC など）によって抽出された CPUE 年トレンドの微妙な違いが生じる場合もある。実際，この違いが，プロダクションモデルや VPA (virtual population analysis) などの資源評価モデルに標準化された CPUE をチューニングインデックスとして用いた場合に資源の絶対量推定結果の大きな違いとなって表れることが多く（Fig. B-2, p.85），このことが CPUE 標準化におけるモデル選択（説明変数の取捨選択）の重要性をさらに高めていると考えられる。CPUE 標準化におけるモデル選択の問題については，第 3 章で詳しく議論する。

$$\mathrm{CPUE}_i = \exp\{(\mathrm{Intercept}) + (\mathrm{Year})_i + \overline{(\mathrm{Year}^{*}\mathrm{Area})}_i + \overline{(\mathrm{Year}^{*}\mathrm{Season})}_{i\cdot} + \ldots\} \tag{2.21}$$

但し

$$\overline{(\mathrm{Year}^{*}\mathrm{Area})}_{i\cdot} = \frac{1}{N_j}\sum_{j=1}^{N_j}(\mathrm{Year}^{*}\mathrm{Area})_{ij},\ \overline{(\mathrm{Year}^{*}\mathrm{Season})}_{i\cdot} = \frac{1}{N_k}\sum_{k=1}^{N_k}(\mathrm{Year}^{*}\mathrm{Season})_{ik}$$

などとする。これらの平均化された項は，単純平均ではなく各々のセルに属するデータ数に応じた重み付け平均とすることもある。

次に，第 4 章ではミナミマグロ資源を取り上げ，ニューラルネットワークによる CPUE 予測と要因分析について論じる。2-4-2節で述べたように，CPUE を相対資源量として捉えた場合，CPUE に相対的なエリアサイズを掛け合わせた資源量指数（abundance index: AI）をその単位とすることは自然に思われる。

その場合に，(2.22) 式（ミナミマグロ資源では CS（constant square）仮説と呼ばれる）を用いることが多く行われている。この仮説では便宜的に区分されたサブエリア内で資源密度が一定と仮定しており，ミナミマグロ資源のように過去から現在にかけて漁場が縮小しており，仮に魚がいなくなったことが原因で起こっているのであれば，資源量指数を過大推定していることになる。なぜなら，この CS 仮説では，操業がなくなった時空間の CPUE を周りの漁業があるそれと同じ，すなわち操業がないセルの CPUE と操業があ

るセルのCPUEの比が1と仮定しているからである。

$$\mathrm{AI}_{ij} = w_j \mathrm{CPUE}_{ij}$$
$$\left(\text{但し}\sum_j w_j = 1 \quad (i:\mathrm{YEAR}、j:\mathrm{AREA}、w_j:\mathrm{AREA}(j)\text{の相対面積指数})\text{とする}\right) \tag{2.22}$$

また，漁場の相対的なエリアサイズが年によって変化する，次の（2.23）式

$$\mathrm{AI}_{ij} = w_{ij} \mathrm{CPUE}_{ij}$$
$$\left(\text{但し}\sum_j w_{ij} = 1 \quad (w_{ij}:\mathrm{YEAR}(i)\text{かつ}\mathrm{AREA}(j)\text{の相対面積指数})\text{とする}\right) \tag{2.23}$$

（VS（variable square）仮説）を使用する場合もあり，漁場縮小の原因が，魚がいなくなったことによるものであるなら，こちらの仮説の方が計算式として適切である。このVS仮説では，操業のないセルのCPUEを周りと同じではなく0と考えており，もし魚がいなくなったことではなく，操業条件の変化，すなわちより漁獲効率が高いエリアに操業が集中するためであるなら，この仮説ではなく，（2.22）式のCS仮説がふさわしい。

CS仮説（日本が提唱）とVS仮説（豪州が主張）の対立は，ハーグの国際司法裁判所に提訴されたミナミマグロ裁判の主要な原因と1つである。日本は，操業がない時空間に魚がいるのかいないのか，というこの問題に対して，1997年から1999年にかけて調査漁獲を行い，操業がないセルと操業があるセルとのCPUE比を0.7程度と推定した。しかし，この調査漁獲はあくまで限定した季節およびエリアで行っており，全体としての結論になりうるかは疑問が残る。

そこで，本研究では教師付きニューラルネットワークを用いて操業がない時空間のCPUE予測を行い，得られたCPUE予測値を使用して要因分析（CPUEの年トレンド抽出）を行うための簡便な分析法を提案する。

最後は，ゼロ・キャッチ問題と呼ばれる，キャッチ，すなわちCPUEがゼロとなるデータが含まれる場合に，（2.19）式のCPUE-LogNormalモデル（共分散分析モデル）が使用出来なくなる問題であり，第5章で詳しく論じる。

この問題に対する現状での回避法は，応答変数である全てのCPUEに微小量（定数項）を足し込み共分散分析を使用するad hocな方法

$$\begin{aligned}&\mathrm{Log}\,(\mathrm{CPUE}+(\mathrm{constant\_term}))=(\mathrm{Intercept})+\\&\quad(\mathrm{Year})+(\mathrm{Season})+(\mathrm{Area})+(\mathrm{EMT})\\&\quad\cdots+(\text{Two-way Interactions})+\mathrm{error},\ \mathrm{error}\sim\\&\quad N(0,\sigma^2)\end{aligned} \tag{2.24}$$

や，（2.20）式で表現されるCatchモデル，2-4-1節で記述したゼロ・キャッチ率をlogitモデルやprobitモデルにより推定し，非ゼロ部分にCPUE（共分散分析）モデルやCatchモデルを使用するDelta型2段階法などが利用されている。

しかし，（2.24）式のad hocな方法は係数の点推定や区間推定に偏りを生じ，Catchモデルではゼロ・キ

---

i 正確には，$\sigma^2 \mathrm{I}_n$（但し I_n: n 次の単位行列，n: データ数）と表記される。

ii エル・ニーニョやラ・ニーニャなどの度合いを表す指標の一つ。太平洋東部のタヒチと太平洋西部のダーウィンの2地点の大気圧の差（両者は逆相関を持つ）に基づいている。

iii offset 変数は線形予測のための1つの項として使用されるが，係数パラメーターが導入されない点が通常の変数と異なる，(2.4) 式では連結関数（link function）を自然対数に設定しているため，$(\mathrm{Effort})^{\alpha}$ の形を仮定して，すなわち α *log（Effort）の形において α = 1 に固定したものと考えれば良い。なお，場合によってはこの指数部分のパラメーター α を推定することも可能である。

iv 現在はSASやS-Plus，SPSSなどの統計パッケージによる計算が主流になっているが，中でもSASは大規模データの定型的な処理に向いていることもあり，細かいカスタマイズが得意なS-PlusやRと並び，水産資源解析分野では多く使用されている。

v 教師付き学習とは，原因（要因効果）を表す入力変数（説明変数）に対応する出力変数（応答変数）が存在する場合における学習方法であり，データの因果関係を調べることを主な目的とする．教師無し学習とは，結果系の変数が存在しない場合における学習方法であり，データの相関関係を調べることを主な目的とする，統計モデルとの関連で言うと，前者に対応する手法として重回帰分析や分散分析，一般化線形モデルなどが，後者に対応する方法として因子分析や主成分分析，クラスター分析などが挙げられる。

vi 一般に，分類問題（判別問題）とは原因に対する結果系変数が離散である場合の推定問題を指し，回帰問題とは結果系変数が連続である場合の推定問題を指す。

vii 各セルの重みを同じとした通常の平方和（Type Ⅲ）と並んで，データ数に応じた重み付けを行った平方和（Type Ⅱ）もSASやS-Plusなどの統計ソフトウェアに装備されている．しかし，データ数の不均一性も含めた様々な要因効果の除去というCPUE標準化の目的からするとType Ⅱ平方和の多用は問題があると著者は考えており，まぐろ類のCPUE解析においても一般にType Ⅲ平方和が広く使用されている。

ャッチ率が極端に高い場合や裾が広い fat tail な分布（ゼロ・データが多い一方で多数獲れているケースも存在する場合）では，確率関数形が現実とマッチしないことが多い。また，Delta 型 2 段階モデルやその尤度関数を数珠繋ぎにしてパラメーターを同時推定した Zero-Inflated モデルでは計算が複雑な上，２つのステップでのモデル選択結果が異なる場合もあり，欠点は解消されていない。

そこで，第 5 章では Tweedie 分布（Tweedei, 1984: Jorgensen, 1997）と呼ばれる確率過程の一種である複合 Poisson 分布の概念を拡張したモデルを用いて，ゼロ・キャッチが含まれるデータの解析を行った。Tweedie モデルは，ゼロ・データを統一的に取り扱える特徴がある。

本博士論文では，ゼロ・データが10% 程度のインド洋における日本のはえ縄商業船によるキハダ資源，およびゼロ・データが80% を超える北太平洋における日本のはえ縄公庁船によるクロトガリザメ資源を例として，CPUE 標準化を行った。現実の漁業データゆえに，n-fold cross-validation というデータをランダムに n- 分割して一部のサブセットにおける CPUE の値を予測する手法を用いて，Tweedie モデルと従来法（ad hoc な方法や Catch-NegativeBinomial モデル）と比較を行った。比較のための尺度としては，観測値と予測値の Pearson's 相関係数および平均二乗誤差（MSE: mean squared error）を使用したが，このフレームワークは第 4 章のニューラルネットワークの性能評価に対しても使用している。

第 3 章　CPUE 解析における統計モデル選択情報量規準とステップワイズ検定の取り扱い

3-1. はじめに

第 3 章では，遠洋域に生息するまぐろ・かつお類の CPUE 解析に多く用いられている CPUE-LogNormal モデル（共分散分析モデル）（式（3.1））を主に取り上げて，水産分野で広く知られている AIC（Akaike's information criterion, 赤池情報量規準）（Akaike, 1973）に代表される様々な情報量規準や，F 検定やカイ二乗検定などを含めた stepwise な統計的検定によるモデル選択, すなわち説明変数の取捨選択について，詳細に検討する。

$$\log(\mathrm{CPUE}) = (\mathrm{Intercept}) + (\mathrm{Year}) + (\mathrm{Area}) + (\mathrm{Season}) + \cdots + (\mathrm{Two\_way\ Interactions}) + (\mathrm{Error}),\quad \mathrm{Error} \sim N(0, \sigma^2) \qquad (3.1)$$

注）ここでは（Two_way Interactions）が 2 要因の積の形で表現される交互作用（e.g.（Year)*（Area））を表すこととする。

2-2節でも統計的モデル選択の現状について紹介したが，情報量規準や stepwise 検定などによる統計的なモデル選択は，CPUE 標準化において CPUE に影響を与えていると考えられる要因の主効果および交互作用が統計的に判断して本当に影響を与えているか否かを判断するという重要な役割を担っている。また，このモデル選択の結果は，CPUE 標準化の主目的である年トレンド推定にも大きな影響を与えており，どのモデルを使用するによって，LSMEANS（least square mean）などによる抽出された CPUE 年トレンドが大きく異なることも珍しくない。現状では，情報量規準 AIC もしくは Backward に変数を減らしていく stepwise 検定がまぐろ類の CPUE 解析において広く利用されているが，以下に示すような誤った使い方も多く，非常に問題である。

まず，情報量規準については，AIC のみが伝統的に使用されており，経済分野などで多用されている BIC（Bayesian information criterion, ベイズ情報量規準）（Schwarz, 1978）などの使用例は，非常に少ないと考えられる。AIC は条件によっては，真のモデルよりもパラメーター数が多い複雑なモデルを選ぶ傾向のあることが知られており，真のモデルを選ぶという選択パフォーマンスの観点から様々な条件下で，様々な情報量規準を用いたシミュレーションによる精度検証が，本章の主要なテーマである。また，基本的に情報量規準は候補となるモデルの中で総当りを行い，その値が一番小さいモデルを選択する必要があるが，一部のモデルのみを用いて stepwise に選択することも現に行われており，（統計パッケージの stepwise AIC などの機能を使用）注意が必要である。

一方，stepwise な統計的検定については，回帰分析や分散分析を含めた共分散分析モデルでは F 検定を，誤差構造として正規分布以外の確率密度関数を仮定した一般化線形モデルではカイ二乗検定を使用することが一般的である。

Backward に変数を減らしていく場合を含め，検定のパスによって最終的なモデル選択結果が異なる場合もあり注意を要するが（2-2節），実際には統計パッケージの出力結果から P- 値の一番大きい変数を順に除去していくというパスを全てチェックしない方法や，時には P- 値が有意水準を越えている値を一度に複数 delete するなどの乱暴な方法が使用されており，警告に値すると感じる。

本章では，まぐろ類のCPUE標準化を想定した，小標本の場合（3-2節），大標本の場合（3-3節），ネスト構造を持つ場合（3-4節および3-5節），正規混合分布モデル（3-6節）などにおける様々な情報量規準の良さについて，理論的な考察や実際例の解析，計算機シミュレーション実験などを通じて，stepwise検定との比較も交えて詳細に検討する。特にシミュレーションに関して，共分散分析モデルなどのフレームワークを仮定して真のモデルに基づく乱数を発生させ，説明要因の組み合わせを変化させた複数の候補モデルの中から正しいモデルを選ぶという選択パフォーマンスの観点から，情報量規準やstepwise検定の良さについて検証することを目的とする。

この真のモデルを選ぶ確率を最大にするという考え方（事後確率最大化）はBIC系統の情報量規準導出のコンセプトになっている。一方で，AIC系統の情報量規準は，観測値と推定値の間の差異を少なくするという考え方（予測誤差最小化）に基づいて導出されている。第3章では，CPUE標準化の究極の目的，すなわち現実の漁業の状況を把握するという目標に照らし合わせて，真のモデルを選ぶ選択パフォーマンスのチェックによる計算機実験の方法が，予測誤差最小を目的にした方法よりも現状にマッチしており適切と考えたため，このような事後確率最大化に基づくシミュレーションを行った。実際，まぐろ類に関する国際漁業委員会での資源解析では，外交および国際漁業交渉の観点から，水産資源管理における政策決定に使用するためには，候補に挙げられたどのモデルが正しいのか，すなわち漁業の現状をより正確に規定しているか，ということが極めて重要であり，予測よりも真のモデルの選定に主眼が置かれている。

また，本章での複数のモデルの中から正しいモデルを選ぶ選択パフォーマンスのチェックに関する計算機実験では，候補となるモデルの中に真のモデルが含まれていることを前提としている。CPUE解析の多くは，真実をモデル化出来ないまでも，現実の漁業の状況にマッチしたモデルが構築可能であると考えられるためであるが，この前提条件が崩れた場合には，BIC系統の情報量規準が持つ良い性質である一致性（第3-3節参照）を満たさなくなるため，これらのシミュレーションにおける仮定の妥当性をも含めた詳細な検証が必要である。

なお，これらの解析結果は，主に庄野（2000），Shono（2000），庄野（2001），Shono（2005），庄野（2006）の5つの論文から引用している。

3-2. 小標本におけるAICの有限修正の有効性

AICは導出過程において最尤推定量の一致性や漸近正規性など漸近理論を用いているため（庄野，2000），小標本の場合には大きなバイアスが生じる。Sugiura（1978）はAICの導出で最尤推定量の漸近正規性を用いてカイ二乗近似を行った部分に工夫を凝らして，exactにカイ二乗分布に従う統計量を用いることにより漸近理論の使用を回避した情報量規準c-AIC（finite correction of Akaike's information criterion: AICの有限修正）を提案しており，その形は

$$\text{c-AIC} = -2*l(\hat{\Theta}) + \frac{2np}{n-p-1} = \text{AIC} + \frac{2p(p+1)}{n-p-1} \qquad (3.2)$$

と表される。但し，n：標本数，p：未知パラメーター数，$l(\cdot)$：対数尤度関数，Θ：未知パラメーターベクトル，$\hat{\Theta}$：Θの最尤推定量，を表す。以下同様。

このc-AICは一般化線形モデルにおける連結関数が恒等写像でかつ正規誤差を持つ場合にのみ適用可能であるが，小標本の場合，特に標本数が25ないし50以下の場合には極めて有効である（坂本ほか，1983）。

また，このc-AICはパラメーター数pの標本数nに占める割合p/nが大きい場合，すなわちp/nが1/3ないし1/4を超えるような場合に有効である。AICではパラメーター数を標本数に近づけていったときに，望ましい性質である一致性（一致性については3-3節に記載）が成立しないために偏りが生じるが（Shibata, 1976），小標本の場合と同じく，c-AICを用いることによってモデルの選択パフォーマンスが改善する（Sugiura, 1978: Hurvich and Tsai, 1989, 1991）。

なお，c-AICは漸近的にはAICと同等であり，正規誤差を持つ一般化線形モデル（重回帰分析モデル・分散分析モデル・自己回帰モデルなどを含む）の場合にのみ適用可能であるため注意を要するが，水産資源解析におけるCPUE標準化に多く使用されるCPUE-LogNormalモデルには利用可能である。

本節では，Shono（2000）に倣い，サクラマスの成長曲線を扱った実際例において使用する情報量規準によりモデル選択結果が異なることを示すとともに，CPUE標準化を想定した小標本の場合（そしてp/nが大きい場合）における二元配置分散分析モデルを使用して計算機シミュレーションを行い，c-AICの真のモデルを選ぶ選択パフォーマンスがAICのそれに比べて高いことを示す。

3-2-1. サクラマスの成長曲線による様々な情報量規準を使用したモデル選択

ここでは，Table 3-1に示す小標本（サンプルサイズが18）におけるサクラマスの年齢・成長データ（Kiso *et al.,* 1992）を使用し，成長曲線の推定を通じて，c-AIC を含む3つの情報量規準（AIC, BIC and c-AIC）によるモデル選択の結果比較を行った。

成長式としては，水産生物に広く適用されている von Bertalanffy, Gompertz, logistic の3種類(式(3.3))を使用し，季節変化のパターンとして4種類（式（3.4）の Base, 1, 2, 3），計12種類の成長式に対して対数正規誤差構造を仮定し，3つの情報量規準によりモデル選択を行ったところ，Table 3-2の結果が得られた。

Table 3-1. Data for fluviatile masu salmon in rivers of the southern Sanriku district (Kiso *et al.*, 1992)

Month of age	Year of age	Mean length
2	0.17	1.79
3	0.25	2.33
4	0.33	2.51
5	0.42	2.97
6	0.50	4.49
7	0.58	6.93
8	0.67	7.33
9	0.75	8.00
10	0.83	10.81
11	0.92	11.68
12	1.00	12.53
13	1.08	13.90
14	1.17	12.30
15	1.25	13.10
16	1.33	15.59
17	1.42	15.60
18	1.50	16.08
19	1.58	16.80

growth formulae：

$$L(t)=L_\infty[1-\exp\{-K(t-t_0)\}]\quad(Von\ Bertalanffy)$$
$$L(t)=L_\infty\exp[-\exp\{-K(t-t_0)\}]\quad(Gompertz)$$
$$L(t)=\frac{L_\infty}{1+\exp\{-K(t-t_0)\}}\quad(Logistic)\qquad(3.3)$$

patterns of seasonal change : $t\rightarrow F(t)$

$$F(t)=t\quad(Basic\text{-}type),$$
$$F(t)=t+\frac{A}{2\pi}\sin 2\pi(t-t_1),\qquad where A\geq 0\qquad(type-1)$$
$$F(t)=t+\frac{A}{2\pi}\sin 2\pi(t-t_1)+\frac{B}{4\pi}\sin 4\pi(t-t_1),\qquad where A\geq 0\quad(type-2)$$
$$F(t)=t+\frac{A}{2\pi}\sin 2\pi(t-t_1)+\frac{C}{6\pi}\sin 6\pi(t-t_1),\qquad where A\geq 0\quad(type-3)$$
$$(3.4)$$

この結果，いずれが現実に近いかは判断不能であるが，AIC は BIC や c-AIC に比べて季節変化部分のパラメーター数が多い複雑なモデルを選択しており，オリジナル論文では AIC による選択のみ行っていたが，使用する情報量規準が成長曲線の選択という本質的な問題になりうることを，示唆している。

3-2-2. CPUE 標準化を想定した二元配置分散分析モデルを使用した計算機実験

ここでは，小標本の場合や p/n が大きい場合に対応する計算機実験を分散分析型モデルによって行い，情報量規準 c-AIC の優位性を示す。繰り返し数2の2元配置分散分析型モデルを仮定して式（3.5）の

Table 3-2. Results of the model selection by three information criteria (AIC, BIC and c-AIC) using the data for mean length of fluviatile masu salmon from Table 3-1

Formulae	Type	n	p	AIC	BIC	c-AIC
Bertalanffy	Basic	18	2	134.75	136.53	135.82
Gompertz	Basic	18	2	49.49	51.27	50.56
Logistic	Basic	18	2	52.98	54.76	54.05
Bertalanffy	1	18	4	15.48	19.04*	18.56*
Gompertz	1	18	4	16.20	19.76	19.28
Logistic	1	18	4	17.94	21.50	21.02
Bertalanffy	2	18	5	16.03	20.48	21.03
Gompertz	2	18	5	15.88	20.33	20.08
Logistic	2	18	5	19.52	23.97	24.52
Bertalanffy	3	18	5	16.30	20.75	21.30
Gompertz	3	18	5	14.69*	19.14	19.69
Logistic	3	18	5	19.25	23.70	24.25

* The minimum value in each information criterion.

Model-Ⅲを用いてデータを発生させ，3つの情報量規準（AIC, BIC, c-AIC）によってモデル選択を行ったときに，式（3.5）の4つのうちどのモデルが選択されるかについての計算機実験を，それぞれ100回ずつ行った（Shono, 2000）。

Model-Ⅰ： Log (CPUE) = Intercept + Error（但し Error ～ N (0, σ^2)）
Model-Ⅱ： Log (CPUE) = Intercept + Year + Error　(3.5)
Model-Ⅲ： Log (CPUE) = Intercept + Year + Area + Error（True model）
Model-Ⅳ： Log (CPUE) = Intercept + Year + Area + (Year*Area) + error

ちなみに，今回の実験での標本数は24（=12*2），パラメーター数は最大で12となる。また，計算機実験に使用した誤差項の分散の値（3種類）及びデータ発生に用いた真のモデルでの応答変数の値（2種類：Table 3-3）を組み合わせたシミュレーションでのシナリオは次のようになる。また，結果は Table 3-4の通りだが，すべてのシナリオにおいて c-AIC の選択パフォーマンスは AIC や BIC のそれと比べて良くなっている。ただし，c-AIC と BIC は AIC に比べてパラメーター数の少ない単純なモデルを選択しやすい傾向にあるため，パラメーター数の多い複雑なモデルを想定したケース等，場合によっては注意を要する。

―6種類の計算機実験シナリオ―

Case-1: Dataset-1 and σ^2=0.10
Case-2: Dataset-1 and σ^2=0.15
Case-3: Dataset-1 and σ^2=0.20
Case-4: Dataset-1 and σ^2=0.25.
Case-5: Dataset-2 and σ^2=0.05.
Case-6: Dataset-2 and σ^2=0.10.

3-3. 大標本における一致性を持つ情報量規準

情報量規準 AIC は前節で述べたような小標本の場合のみならず，大標本においてもパラメーター数の多い複雑なモデルを選ぶ可能性が高い。具体的には，

$$\mathrm{AIC} = -2*l(\hat{\Theta})+2p \approx -2*l(\hat{\Theta}) \quad (\text{as } n \to \infty) \qquad (3.6)$$

と書けることから，標本数が大きくなると対数尤度関数の値に比べて定数項の占める割合が小さくなり，このペナルティ項の効力がほとんどなくなってしまうからである。水産資源解析分野におけるまぐろ類の

Table 3-3. Dataset on the model of CPUE standardization for computer simulation, which corresponds to the analysis of variance model with two-way layout and two replicates

Dataset-1

Year	Area-1	Area-2	Area-3
1	1.0, 1.0	0.8, 0.8	1.2, 1.2
2	0.9, 0.9	0.72, 0.72	1.08, 1.08
3	1.1, 1.1	0.88, 0.88	1.32, 1.32
4	0.8, 0.8	0.64, 0.64	0.96, 0.96

Dataset-2

Year	Area-1	Area-2	Area-3
1	1.0, 1.0	0.9, 0.9	1.1, 1.1
2	0.95, 0.95	0.855, 0.855	1.045, 1.045
3	1.1, 1.1	0.99, 0.99	1.21, 1.21
4	0.9, 0.9	0.81, 0.81	0.99, 0.99

Table 3-4. Summary of computer experiments (frequency of selecting the correct model of an ANOVA in 100 replicates of the simulation)

Model	AIC	BIC	c-AIC
Case-1			
Ⅰ	0	0	0
Ⅱ	0	0	0
Ⅲ(true)	65	89	100
Ⅳ	35	11	0
Case-2			
Ⅰ	0	1	1
Ⅱ	0	1	1
Ⅲ(true)	65	90	98
Ⅳ	35	8	0
Case-3			
Ⅰ	0	12	13
Ⅱ	1	1	3
Ⅲ(true)	65	76	84
Ⅳ	34	11	0
Case-4			
Ⅰ	6	28	31
Ⅱ	1	1	2
Ⅲ(true)	57	64	67
Ⅳ	36	7	0
Case-5			
Ⅰ	0	0	0
Ⅱ	0	0	0
Ⅲ(true)	63	88	100
Ⅳ	37	12	0
Case-6			
Ⅰ	0	6	7
Ⅱ	1	2	2
Ⅲ(true)	58	78	91
Ⅳ	41	14	0

CPUE 標準化では，shot-by-shot と呼ばれる操業毎のデータを取り扱うことが多く，標本数が数十万件になることも少なくないため，慣用的に AIC を使用することに関しては十分な注意が必要である。

また，大標本における漸近的な性質の良さを表す一致性（consistency）は，情報量規準について，(3.7) 式で表される。

$$\Pr[IC(Mj)=IC(M)] \rightarrow 1 \quad (\text{as } n \rightarrow \infty) \tag{3.7}$$

但し M: 真のモデル，Mj: 候補となるモデル，IC: 情報量規準 とする。

AIC の一致性は p/n が一定で n →∞のときには満たされる（Shibata, 1976）。しかし，p を固定したまま n →∞としたときには成立しないため，前節で述べた大標本において偏りが生じる原因になりうる。

この欠点を修正した式（3.7）の一致性をもつ規準として，BIC（ベイズ情報量規準，Bayesian information criterion：Schwarz, 1978）や HQ（Hannan and Quinn, 1979），CAIC（Consistent AIC：Bozdogan, 1987） などが提案されており式 (3.8) のような形になる。但し，HQ における c は 2 より大きい任意定数とする。

$$\begin{aligned} \mathrm{BIC} &= -2*l(\hat{\Theta}) + p\log n \\ \mathrm{HQ} &= -2*l(\hat{\Theta}) + cp\log\log n \\ \mathrm{CAIC} &= -2*l(\hat{\Theta}) + p\log n + p \end{aligned} \tag{3.8}$$

BIC は n →∞の場合に一致性を持つものの，漸近理論による導出過程で評価の甘さもあり，真のモデルよりパラメーター数の少ない単純なモデルを選択しやすい傾向が認められる。精確には，Neath and Cavanaugh（1997）による

$$\mathrm{CAICF} = -2\log L(\hat{\Theta}) + p\{\log(n)+2\} + \log|I(\hat{\Theta})| \tag{3.9}$$

$$\left(\text{但し} I(\Theta) := E\left[\frac{\partial l(\Theta)}{\partial \Theta} \frac{\partial l(\Theta)}{\partial \Theta^t} \right] (\Theta\text{のFisher情報行列}), \quad |\ |\text{:行列式, とする} \right)$$

が正しいと思われるが，Fisher 情報行列の行列式の計算は困難なケースが多い。

なお，HQ は c が 2 より大きい定数という条件になっているため，現実のモデルに適用するためには，何らかの方法でこの定数 c の値を設定してやらなければならない。また，CAIC は共分散構造分析などで多く用いられているが，ペナルティ項の重みは BIC よりも大きいことには，注意が必要である。

そこで，本節では最初に大規模まぐろ漁業データによる CPUE 標準化（共分散分析モデル）を通じて，AIC および 3 つの一致性を持つ情報量規準（BIC, HQ and CAIC）によるモデル選択を行い, 比較検討する。次に，回帰分析型の計算機シミュレーション実験を行い，一致性を持つ規準の選択パフォーマンスを詳しくチェックする。また，これら 4 つの情報量規準（AIC, BIC, HQ and CAIC）のペナルティ項の大きさに着目して理論的・数値的な考察を行うとともに，HQ における定数項 c の定め方についても検証する。

3-3-1. インド洋キハダ資源の CPUE 標準化（情報量規準によるモデル選択）

ここでは，大標本（サンプルサイズが41,504）におけるインド洋キハダ資源の漁獲量・努力量データ（Shono *et al.*, 2002）を使用し，共分散分析モデルによる説明要因効果の推定を通じて，AIC および一致性を持つ 3 つの情報量規準（BIC, HQ and c-AIC）によるモデル選択の結果比較を行った。HQ の定数項は2.01及び2.71を使用したが，前者はペナルティ項を考えうる範囲で出来るだけ小さく，後者はペナルティ項の重みを BIC とほぼ同等にする，という考え方に基づいている。以下，Table 3-5，Table 3-6，Fig 3-1 にそれぞれ候補となるモデル（説明要因の組み合わせ）の概要，モデル選択の結果，抽出された CPUE の年トレンドを示す。

Table 3-6のモデル選択結果を見ると，AIC はパラメーター数の多い一番複雑なモデル（Model-13）が，BIC および CAIC は比較的パラメーター数の少ない単純なモデル（Model-9）が，HQ は c が2.01と2.71のいずれに設定した場合にも両者の中間のモデル（Model-12）が選択された。

水準間の重みを一定とした Type Ⅲ型の LSMEANS（least squared means）に基づいて抽出された CPUE 年トレンドは，全体的には良く似ているが，1970年以降に多少の違いも認められるため，CPUE 年トレンドの微妙な違いが資源推定の絶対値の大きな違いになって表れることも多くあるため（付録 B：例 B-1, p.84），資源評価モデルのチューニング・インデックスとして使用する場合には，注意が必要である。

3-3-2. 大標本における回帰分析型シミュレーションによる情報量規準の比較

次に，大標本における回帰分析型シミュレーションを通じて，一致性を持つ 3 つの情報量規準及び AIC による（正しいモデルを選ぶという観点から）選択パ

Table 3-5. ANOVA models for CPUE standardization of yellwofin tuna in the Indian Ocean (Shono *et al.*, 2002)

Model	Intercept	Effect			Interaction		
		Year	Month	Area	Year*Month	Year*Area	Month*Area
1	○	○					
2	○	○	○				
3	○	○		○			
4	○	○	○		○		
5	○	○		○		○	
6	○	○	○	○			
7	○	○	○	○	○		
8	○	○	○	○		○	
9	○	○	○	○			○
10	○	○	○	○	○	○	
11	○	○	○	○	○		○
12	○	○	○	○		○	○
13	○	○	○	○	○	○	○

Model-1: $\log(CPUE+k)_i = (Intercept)+(Year)_i+(error)_i$,

Model-2: $\log(CPUE+k)_{ij} = (Intercept)+(Year)_i+(Month)_j+(error)_{ij}$,

Model-3: $\log(CPUE+k)_{ik} = (Intercept)+(Year)_i+(Area)_k+(error)_{ik}$,

Model-4: $\log(CPUE+k)_{ij} = (Intercept)+(Year)_i+(Month)_j+(Year*Month)_{ij}+(error)_{ij}$,

Model-5: $\log(CPUE+k)_{ik} = (Intercept)+(Year)_i+(Area)_k+(Year*Area)_{ik}+(error)_{ik}$,

Model-6: $\log(CPUE+k)_{ijk} = (Intercept)+(Year)_i+(Month)_j+(Area)_k+(error)_{ijk}$,

Model-7: $\log(CPUE+k)_{ijk} = (Intercept)+(Year)_i+(Month)_j+(Area)_k+(Year*Month)_{ij}+(error)_{ijk}$,

Model-8: $\log(CPUE+k)_{ijk} = (Intercept)+(Year)_i+(Month)_j+(Area)_k+(Year*Area)_{ik}+(error)_{ijk}$,

Model-9: $\log(CPUE+k)_{ijk} = (Intercept)+(Year)_i+(Month)_j+(Area)_k+(Month*Area)_{jk}+(error)_{ijk}$,

Model-10: $\log(CPUE+k)_{ijk} = (Intercept)+(Year)_i+(Month)_j+(Area)_k+(Year*Month)_{ij}+(Year*Area)_{ik}+(error)_{ijk}$,

Model-11: $\log(CPUE+k)_{ijk} = (Intercept)+(Year)_i+(Month)_j+(Area)_k+(Year*Month)_{ij}+(Month*Area)_{jk}+(error)_{ijk}$,

Model-12: $\log(CPUE+k)_{ijk} = (Intercept)+(Year)_i+(Month)_j+(Area)_k+(Year*Area)_{ik}+(Month*Area)_{jk}+(error)_{ijk}$,

Model-13: $\log(CPUE+k)_{ijk} = (Intercept)+(Year)_i+(Month)_j+(Area)_k+(Year*Month)_{ij}+(Year*Area)_{ik}+(Month*Area)_{jk}+(error)_{ijk}$,

where $(error)_{ijk}$ is normally distributed with mean 0 and variance σ^2; *CPUE* is the nominal CPUE (number of yellowfin catch per 1000 hooks); *Year*, 1960-2000; *Month*, 1-12; *Area*, 1-6; k, constant term (we set to 0.1 here)

Table 3-6. Results of model selection by four information criteria (AIC, BIC, CAIC and HQ) using the data for IOTC yellowfin tuna (Shono *et al.*, 2002)

Model	n	p	AIC	BIC	HQ(c=2.01)	CAIC	HQ(c=2.71)
1	41504	41	153084.3	153438.2	153197.1	153479.2	153264.9
2	41504	52	151898.9	152347.9	152042.0	152399.9	152128.0
3	41504	46	144724.9	145122.0	144851.4	145168.0	144927.6
4	41504	492	150779.2	155026.9	152133.0	155518.9	152947.2
5	41504	246	143243.4	145367.3	143920.3	145613.3	144327.4
6	41504	57	143212.2	143704.3	143369.1	143761.3	143463.4
7	41504	437	141981.5	145754.4	143184.0	146191.4	143907.1
8	41504	257	141615.2	143834.0	142322.3	144091.0	142747.6
9	41504	112	141863.5	**142830.5**	142171.7	**142942.5**	142357.1
10	41504	697	140492.7	146510.3	142410.6	147207.3	143564.0
11	41504	552	140870.9	145636.6	142389.8	146188.6	143303.3
12	41504	312	140216.1	142909.8	**141074.7**	143221.8	**141591.0**
13	41504	752	**139230.8**	145723.2	141300.0	146475.2	142544.4

Remark) Bold type shows the minimum value in each of information criterion.

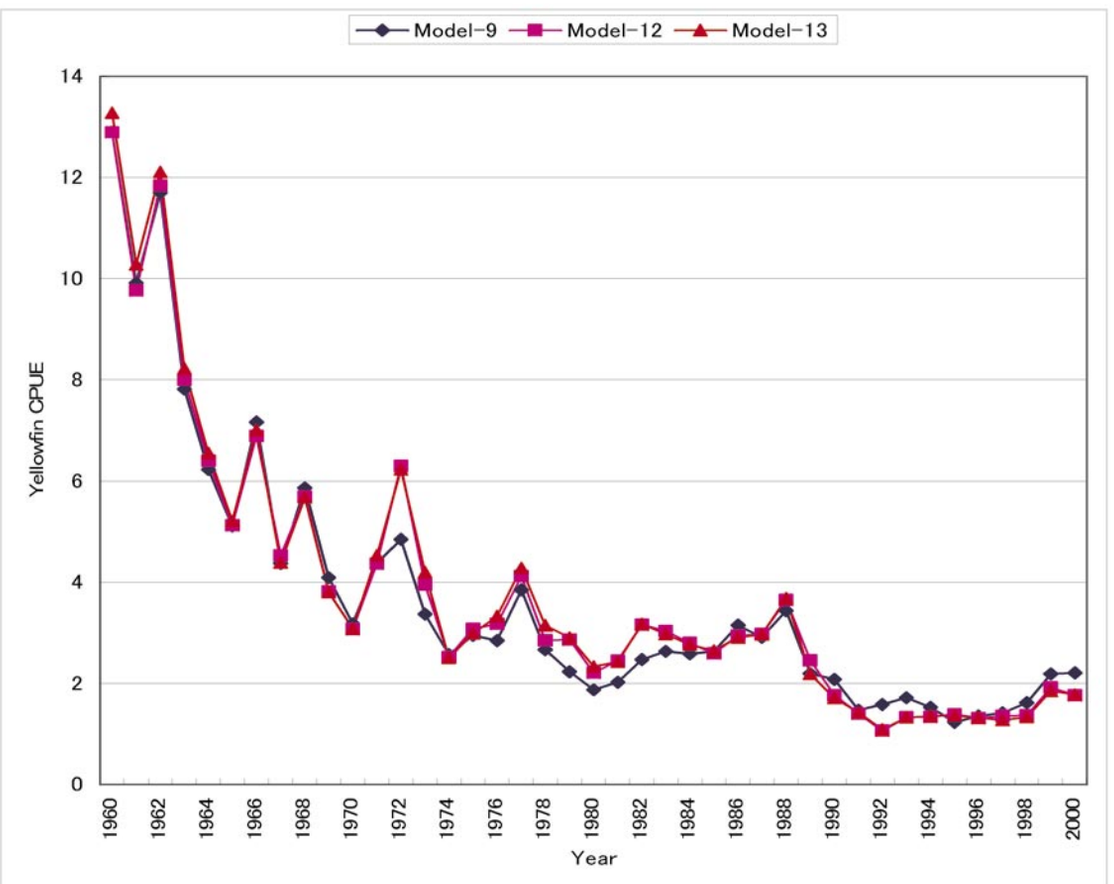

Fig. 3-1. Three catch per unit effort (CPUE) year trends for yellowfin tuna in the Indian Ocean obtained from the best models (Model 9, Model 12 and Model 13) based on each information criterion: Akaike Information Criterion (AIC), Baysian Information Criterion (BIC), consistent AIC (CAIC), Hannan-Quinn (HQ; c=2.01) and HQ (c=2.71).

フォーマンスの評価を行った。

シミュレーションの概要は以下の通りであるが，説明変数の分布形としては次の３つを取り入れた(特に，Case-3は変数間に相関を持つ場合に対応)。

・候補となるモデル

Model-b: Z = A + B*X + e
Model-c: Z = A + B*X + C*Y + e
Model-d: Z = A + B*X + C*Y + D* (X*Y) + e (3.10)

注）誤差項 e は，平均0, 分散 σ^2を持つ正規分布に従うことを仮定した。

・真のモデル（Model-c）

$Z = 1 + X + Y + e,\ e \sim N(0, \sigma^2)$ (3.11)

注)True model における回帰係数（A, B, C）の値はすべて1.0と設定した。

・説明変数の仮定された確率分布

Case-1 : $X_i, Y_i \sim$ i. i. d. U (0, 1)（一様分布）
Case-2 : $X_i, Y_i \sim$ i. i.d. N (0, 1)（正規分布）
Case-3 : $(X_i, Y_i) \sim$ N (0, 0, 1, 1, (p=) $1/\sqrt{3}$)（２変量正規分布）

注)i.i.d.: 互いに独立同一分布に従うという意　ρ : 確率変数 X と Y の相関係数

また，その他の条件は次の通りであり，シミュレーションにおけるシナリオの組み合わせを Table 3-7に，真のモデルを選ぶ選択パフォーマンスに関するモデル選択結果を Table 3-8に示した。

・標本数　Set-A: n=1000,　Set-B: n=500
・繰り返し数（i.e. データセット): 1000
・残差分散（σ^2）の大きさ
　Option-1:　分散が小さい場合（一様乱数（Case-1）ではσ^2=1とし，正規乱数（Case-2 and Case-3）ではσ^2=2.89と設定）
　Option-2:　分散が大きい場合（一様乱数（Case-1）ではσ^2=2とし，正規乱数（Case-2 and Case-3）ではσ^2=7と設定）

・情報量規準 HQ における定数項 c の仮定

HQ-1: 2.01,　　HQ-2: 2.71

Table 3-8から判断する限り，一致性を持つ情報量規準（BIC, HQ and CAIC）の選択パフォーマンスは一致性を持たない情報規準である AIC のそれと比べて全般的に極めて良くなっており，標本数500ないし1,000程度でもその良さが確認された。

ただし，説明変数間に相関が認められる場合（特に標本数が500のシナリオ LN2B（No.12)）には一致性を持つ情報量規準の優位性が見られないこともあり，共変量間の相関を持つ場合の影響を評価するため，2変量正規分布を仮定した Case-3における追加シミュレーションを行った。その概要は次のようになる。

・真のモデル（Model-b）

$Z = 1 + X + e,\ e \sim N(0, \sigma^2)$ (3.12)

・候補となるモデル

Model-a: Z = A + e,
Model-b: Z = A + B*X + e
Model-c: Z = A + B*X + C*Y + e (3.13)

注）他の条件はオリジナルセット（SN2A, LN2A, SN2B, LN2B）と同じに設定

追加シミュレーションの具体的な４つのシナリオ

Table 3-7. Summary of the scenarios for these computer simulations

No.	Name	Sample size	Replications	Residual variance	Distribution
1	SU1A	1,000	1,000	1	Uniform
2	SN1A	1,000	1,000	2.89	Normal
3	SN2A	1,000	1,000	2.89	2-Normal
4	LU1A	1,000	1,000	2	Uniform
5	LN1A	1,000	1,000	7	Normal
6	LN2A	1,000	1,000	7	2-Normal
7	SU1B	500	1,000	1	Uniform
8	SN1B	500	1,000	2.89	Normal
9	SN2B	500	1,000	2.89	2-Normal
10	LU1B	500	1,000	2	Uniform
11	LN1B	500	1,000	7	Normal
12	LN2B	500	1,000	7	2-Normal

Remark)　"2-Normal" shows the bivariate normal distribution.

Table 3-8. Summary of computer simulations (Frequency of selecting the correct model of a linear regression model in 1000 replications of the simulation)

				Model		Proportion of		
No.	Name	Criterion	b	c (True)	d	underfitting	correct	overfitting
1	SU1A	AIC	0	844	156	0	0.844	0.156
		BIC	0	992	8	0	0.992	0.008
		HQ(c=2.01)	0	956	44	0	0.956	0.044
		CAIC	0	**998**	2	0	0.998	0.002
		HQ(c=2.71)	0	977	23	0	0.977	0.023
2	SN1A	AIC	0	833	167	0	0.833	0.167
		BIC	0	991	9	0	0.991	0.009
		HQ(c=2.01)	0	955	45	0	0.955	0.045
		CAIC	0	**995**	5	0	0.995	0.005
		HQ(c=2.71)	0	982	18	0	0.982	0.018
3	SN2A	AIC	0	843	157	0	0.843	0.157
		BIC	0	992	8	0	0.992	0.008
		HQ(c=2.01)	0	948	52	0	0.948	0.052
		CAIC	0	**997**	3	0	0.997	0.003
		HQ(c=2.71)	0	976	24	0	0.976	0.024
4	LU1A	AIC	0	840	160	0	0.84	0.16
		BIC	27	**966**	7	0.027	0.966	0.007
		HQ(c=2.01)	3	946	51	0.003	0.946	0.051
		CAIC	43	955	2	0.043	0.955	0.002
		HQ(c=2.71)	12	958	30	0.012	0.958	0.03
5	LN1A	AIC	0	836	164	0	0.836	0.164
		BIC	9	**985**	6	0.009	0.985	0.006
		HQ(c=2.01)	1	954	45	0.001	0.954	0.045
		CAIC	12	**985**	3	0.012	0.985	0.003
		HQ(c=2.71)	5	980	15	0.005	0.98	0.015
6	LN2A	AIC	5	857	138	0.005	0.857	0.138
		BIC	117	875	8	0.117	0.875	0.008
		HQ(c=2.01)	29	**927**	44	0.029	0.927	0.044
		CAIC	152	842	6	0.152	0.842	0.006
		HQ(c=2.71)	56	923	21	0.056	0.923	0.021
7	SU1B	AIC	0	847	153	0	0.847	0.153
		BIC	0	988	12	0	0.988	0.012
		HQ(c=2.01)	0	939	61	0	0.939	0.061
		CAIC	0	**992**	8	0	0.992	0.008
		HQ(c=2.71)	0	974	26	0	0.974	0.026
8	SN1B	AIC	0	852	148	0	0.852	0.148
		BIC	0	985	15	0	0.985	0.015
		HQ(c=2.01)	0	954	46	0	0.954	0.046
		CAIC	0	**990**	10	0	0.99	0.01
		HQ(c=2.71)	0	975	25	0	0.975	0.025
9	SN2B	AIC	0	839	161	0	0.839	0.161
		BIC	1	986	13	0.001	0.986	0.013
		HQ(c=2.01)	0	948	52	0	0.948	0.052
		CAIC	1	**992**	7	0.001	0.992	0.007
		HQ(c=2.71)	0	972	28	0	0.972	0.028
10	LU1B	AIC	26	813	161	0.026	0.813	0.161
		BIC	195	798	7	0.195	0.798	0.007
		HQ(c=2.01)	70	**883**	47	0.07	0.883	0.047
		CAIC	244	750	6	0.244	0.75	0.006
		HQ(c=2.71)	130	848	22	0.13	0.848	0.022
11	LN1B	AIC	35	814	151	0.035	0.814	0.151
		BIC	257	731	12	0.257	0.731	0.012
		HQ(c=2.01)	101	**846**	53	0.101	0.846	0.053
		CAIC	304	690	6	0.304	0.69	0.006
		HQ(c=2.71)	166	806	28	0.166	0.806	0.028
12	LN2B	AIC	116	**745**	139	0.116	0.745	0.139
		BIC	485	508	7	0.485	0.508	0.007
		HQ(c=2.01)	263	702	35	0.263	0.702	0.035
		CAIC	557	438	5	0.557	0.438	0.005
		HQ(c=2.71)	381	607	12	0.381	0.607	0.012

Remark) Bold type shows the information criterion with highest selection performance in each experiment.

は Table 3-9のようになり，モデル選択結果は Table 3-10の通りである。これを見る限りでは，一致性を持つ情報量規準（BIC, HQ and CAIC）の選択パフォーマンスは，全てのシナリオで AIC のそれよりも良くなっている。

次に，一致性を持つ３つの規準の間でのペナルティ項の重みについて検討する。実際にシミュレーション実験ではこれらの情報量規準の中でどれを使用すべきかは，仮定に大きく依存するため判断が難しい。そこで，４つの情報量規準間でペナルティ項の大きさの対比較を grid search により行い，その条件を列挙してみた（式（3.14））。

$$\begin{aligned} &\mathrm{CAIC} - \mathrm{BIC} = p > 0 \ (\text{for all } n) \\ &\mathrm{CAIC} - \mathrm{AIC} = p*\{\log(2/e)\} > 0 \ (\text{if } n>3) \\ &\mathrm{BIC} - \mathrm{AIC} = p*\{\log(2/e^2)\} > 0 \ (\text{if } n>8) \\ &\mathrm{HQ} - \mathrm{AIC} = p*[\log\{c*\log(n)/e^2\} > 0 \ (\text{if } n>67) \end{aligned} \tag{3.14}$$

但し c は２より大きい任意定数とする。

注）（3.14）の見方として，例えば AIC と BIC の比較では n>8のときに BIC>AIC となり，BIC のペナルティ項は AIC のそれよりも大きくなる。すなわち，n>8の場合は BIC の方がパラメーター数の少ない単純なモデルを選ぶ傾向にある。

一般に，AIC はパラメーター数の多い複雑なモデルを，BIC はパラメーター数の少ない単純なモデルを選ぶ傾向があるため，ここでは HQ での定数項 c の値をうまく設定して，AIC と BIC の中間に持ってくることを目標としたい。

そのためには，４つの情報量規準の値が

$$\mathrm{AIC} < \mathrm{HQ} < \mathrm{BIC} < \mathrm{CAIC} \tag{3.15}$$

となるようにすれば良いため，（3.16）式の grid search の結果から，２＜ c <2.71と設定すれば良いことが分かる。これが漁業データの解析や計算機実験

Table 3-9. Summary of the options for additional simulations

No.	Name	Original	True Regression Model	Candidate Models
A1	SN2Aop	SN2A	Model-b	Model-a,b,c
A2	LN2Aop	LN2A	Model-b	Model-a,b,c
A3	SN2Bop	SN2B	Model-b	model-a,b,c
A4	LN2Bop	LN2B	Model-b	Model-a,b,c

Table 3-10. Summary of additional computer simulations (Frequency of selecting the correct model of a linear regression model in 1000 replications of the simulation)

			Model			Proportion of		
No.	Name	Criterion	a	b (True)	c	underfitting	correct	overfitting
A1	SN2Aop	AIC	0	849	151	0	0.849	0.151
		BIC	0	991	9	0	0.991	0.009
		HQ(c=2.01)	0	962	38	0	0.962	0.038
		CAIC	0	**998**	2	0	0.998	0.002
		HQ(c=2.71)	0	983	17	0	0.983	0.017
A2	LN2Aop	AIC	1	846	153	0.001	0.846	0.153
		BIC	22	**971**	7	0.022	0.971	0.007
		HQ(c=2.01)	6	944	50	0.006	0.944	0.05
		CAIC	30	965	5	0.03	0.965	0.005
		HQ(c=2.71)	12	967	21	0.012	0.967	0.021
A3	SN2Bop	AIC	0	830	170	0	0.83	0.17
		BIC	0	989	11	0	0.989	0.011
		HQ(c=2.01)	0	939	61	0	0.939	0.061
		CAIC	0	**995**	5	0	0.995	0.005
		HQ(c=2.71)	0	976	24	0	0.976	0.024
A4	LN2Bop	AIC	27	799	174	0.027	0.799	0.174
		BIC	203	783	14	0.203	0.783	0.014
		HQ(c=2.01)	90	**851**	59	0.09	0.851	0.059
		CAIC	269	723	81	0.269	0.723	0.081
		HQ(c=2.71)	141	827	32	0.141	0.827	0.032

Remark)　Bold type shows the information criterion with highest selection performance in each experiment.

でc=2.71と置いた1つの根拠である。HQの値をAICとCAICの間に持ってくるためには，同様の考察（式(3.17)）によりcを3.59未満とする必要がある。

$$BIC-HQ=p(\log(n)-c*\log(\log(n)))>0 \text{ (for all n)} \rightarrow c<2.71 \quad (3.16)$$

$$CAIC-HQ=p(\log(n)+1-c*\log(\log(n)))>0 \text{ (for all n)} \rightarrow c<3.59 \quad (3.17)$$

最後に，すなわちHQの定数項cの設定方法についての考え方を発展させて

（AICのペナルティ項）＜（HQのペナルティ項）＜（BICのペナルティ項） (3.18)

となるようなHQのペナルティを仮定した上で

$$c=2+k\,[\log n/\{\log(\log(n))\}-2] \quad (k\text{：任意定数，}0<k<1) \quad (3.19)$$

の形の定数を考え，kの値を計算機実験によって設定することを試みる。

重回帰分析モデルによる計算機実験の結果（3-3-3節），あくまで一例に過ぎないがk=0.05～0.1程度が望ましいという結論が得られた。

3-3-3. 大標本における回帰分析型シミュレーション（HQにおけるkの設定）

ここでは，大標本の場合に対応する計算機実験を3-3-2節に倣い回帰分析型モデルによって行い，一致性を持つ情報量規準であるBIC, HQの選択パフォーマンスを調べるとともに，(3.19）式の形を仮定した上でHQにおける適切な定数cの値（すなわち（3.19）式における定数kの値）を求めることを目標とする。

次式（3.20）のModel-Bを真と考え，X, Yがそれぞれ一様分布U(0,1）に従う確率変数とし，a=b=c=d=1と仮定する。この例3.3とほぼ同様の回帰分析型シミュレーションモデルでは，標本数を10000に設定し，乱数を500回発生させる。そして，3つの情報量規準AIC，HQ（本文中の（3.19）式においてk=0.05, 0.1, 0.3, 0.5と変化させる），BIC（HQでk=1としたときと一致）を使用してモデル選択を行った。その結果，あくまで一例ではあるが，HQにおいてk=0.05～0.1程度に設定した場合の選択パフォーマンスが良いことが示された（Table 3-11）。

Model-A : $Z = a+bX+e$（但し $e \sim N(0, \sigma^2)$，$\sigma=2.9$ と設定する）
Model-B : $Z = a+bX+cY+e$
Model-C : $Z = a+bX+cY+d(X*Y)+e$ (3.20)

3-4. ネストモデルにおける情報量規準TIC

本節では，ネスト構造を持つモデルについて述べる。ネスト構造を持つモデルとは，候補となる2つのモデルA，Bの未知パラメーターベクトルをΘ_A, Θ_Bとしたとき，両者に$\Theta_A \subset \Theta_B$という関係が成立しているモデルのことであり，例えば

$$\Theta_A=(\theta_1,\cdots\theta_q),\quad \Theta_B=(\theta_1,\cdots\theta_q\cdots\theta_p)\ (q<p) \quad (3.21)$$

とした場合がこれに該当する。ネスト構造を持つモデルにおいてはモデル選択の問題は変数選択の問題へと読み替えられるが，候補となるモデルが真のモデルを含まない場合，すなわち上の例で言うとモデルBが真でありモデルAが候補となる場合には，AICが偏りを持つことが知られている。

このような場合には，AICではその導出過程でexactに評価すべき項を期待値で置き換えているがゆえに偏りが生じるが，竹内（1976）による情報量規準TIC（Takeuchi's Information Criterion）を用いることにより修正可能であり，その式の形は（3.22）式のようになる。

$$TIC=-2*l(\hat{\Theta})+2\hat{t} \quad (3.22)$$

但し，$\hat{t}$は，trace $\{J(\Theta)^{-1}I(\Theta)\}$をその推定量で置き換えたものであり，

$J(\Theta)=-E\left[\frac{\partial}{\partial\Theta\partial\Theta^t}l(\Theta)\right]$である。

TICの導出も厄介な場合が多いが，例えば分散分

Table 3-11. Simulation results changing the value of the constant term in HQ

Model	AIC	HQ1(k=0.05)	HQ2(k=0.1)	HQ3(k=0.3)	HQ4(k=0.5)	BIC(HQ:k=1)
A	8	40	45	69	93	164
B (true)	423	446	447	428	407	336
C	69	14	8	3	0	0

析モデルなどでは

$$\mathrm{TIC}= -2*l(\hat{\Theta})+2p-3+\frac{n\hat{\mu}(4)}{(\mathrm{RSS})^2}$$

（$\hat{\mu}(4)$：平均まわりの４次モーメント）　　(3.23)

と簡便な形で書き表せる。しかし，この分散分析モデルのケースでは，シミュレーション実験における真のモデルを選ぶセレクションパフォーマンスがAICと比較してもさほど良くはなっていない（3-4-1節）。その理由としては，分散分析モデル等におけるTICとAICの差の理論値（期待値）が０となることが挙げられる（庄野，2001）。なお，分散分析モデルにおけるTICの導出，および理論値のAICとTICの比較について，付録Aに示す。

3-4-1. ネストモデルに対する計算機実験（TICの選択パフォーマンス評価）

ここでは，候補となるモデルが真のモデルを含まない場合に対応する計算機実験を分散分析型モデルによって行い，従来理論的に良いと考えられてきた情報量規準TIC選択パフォーマンスとAICのそれとの比較を行う。3-2-2節と全く同じ２元配置分散分析型モデルを仮定し（Table 3-12），式（3.24）のModel-Ⅲを用いてデータを発生させ，４つの情報量規準（AIC, BIC, c-AIC, TIC）によってモデル選択を行ったときにどのモデルが選択されるかについての計算機実験を1,000回行った（庄野，2001）。今回の実験においては繰り返し数を８と設定したため標本数は96（=12*8）である。また，σ^2（正規誤差の分散）=0.5と設定した。結果はTable 3-13の通りだが，TICの選択パフォーマンスはAICやc-AICのそれと大差ないと言える。

Table 3-12. Dataset used for the simulation of ANOVA type in the nested model

Year	Area1	Area2	Area3
1	1	0.8	1.2
2	0.9	0.72	1.08
3	1.1	0.88	1.32
4	0.8	0.64	0.96

―ネストモデルの計算機実験に用いた真のモデルおよび候補モデル―

Model-Ⅰ: Log (CPUE)= Intercept + Error
（但し Error ～ N (0, σ^2)）
Model-Ⅱ: Log (CPUE)= Intercept + Year + Error　　(3.24)
Model-Ⅲ: Log(CPUE)= Intercept + Year + Area + Error (True model)
Model-Ⅳ: Log (CPUE)= Intercept + Year + Area + (Year*Area) + error

3-5. ネストモデルにおける情報量規準とstepwise検定の比較

ネスト構造を持つモデルにおいては，情報量規準以外にstepwise検定も適用可能である。例えば，前述のパラメーターベクトル

$\Theta_A=(\theta_1,\cdots,\theta_q)$, $\Theta_B=(\theta_1,\cdots,\theta_q,\cdots\theta_p)$ $(q<p)$　　(3.25)

を持つモデルAとBを想定したとき，パラメーター数の差$p-q$が小さい場合（精確に言うと有意水準５％のときは７以下，１％のときは15以下）には検定の方が，それより大きい場合にはAICの方が複雑なモデルBを選ぶ傾向にある（庄野，2000）。

また，候補となるモデルが多数存在する場合には，有意水準の設定以外に検定のパス（順序）も問題となる。例えば，CPUE標準化などに用いられる分散分析モデルにおいて，次の５つの候補となるモデル

Model-1 : Log (CPUE)= Intercept + Error
Model-2 : Log (CPUE)= Intercept + Year + Error
Model-3 : Log (CPUE)= Intercept + Area + Error
Model-4 : Log (CPUE)= Intercept + Year + Area + Error
Model-5 : Log (CPUE)= Intercept + Year + Area + (Year*Area) + Error　　(3.26)

Table 3-13. Results of model selection using the data from Table 3-12 (simulation of ANOVA type in the nested model)

Model	AIC	BIC	c-AIC	TIC
Ⅰ	79	643	125	78
Ⅱ	51	50	62	51
Ⅲ(true)	776	307	778	781
Ⅳ	94	0	35	90

但し Error ～ N (0, σ^2) とする。

を想定して Backward に変数を減らしていく stepwise 検定を考えると，(3. 27) 式の２つのパスが存在する。そこで，このネストモデルを用いて計算機シミュレーションを行い，情報量規準と stepwise 検定の選択パフォーマンスを比較した。以下，3-5-1節でシミュレーション実験の手順について述べる。

Path1
(Model-1) ← (Model-2)
↑ Path2 ↑ Path1 (3.27)
(Model-3) ← (Model-4) ← (Model-5)
Path2 Path1&2

注）(Model-5) から (Model-4) のパスは Path1と Path2で共通

3-5-1. ネストモデルにおける情報量規準と stepwise 検定との比較実験

ここでは，ネスト構造を持ち検定のパスが複数存在する状況を考え，分散分析型モデルを仮定したシミュレーションにより情報量規準と stepwise 検定との比較を行った。繰り返し数２ (i. e. 標本数24) の２元配置分散分析型モデルを仮定して，上のパス図 (3.27) の５つのモデルのうち Model-3を真と考え (Table 3-14)，乱数を1,000個発生させる。そして２つの情報量規準 (AIC, BIC) と，式 (3.27) の２つの異なるパスを持つ stepwise 検定において有意水準を１% から 10% まで５段階に変化させたときにどのモデルが選ばれるのか，という選択パフォーマンスを調べた。この実験において，stepwise 検定では Backward に変数を減らしていく方法を採用したが，２つのパスによる検定結果が一致しない場合には次の２通りの判断を考える。判断１ではその場合に即矛盾としてしまうのに対して，判断２では新たに検定を行って矛盾かどうかを決定する，とする点が異なっている。なお，この一連の実験においては，σ =0.45と仮定した。

具体的な手順として，最初に Model-4が真という帰無仮説に対して Model-5が真という対立仮説を考え，帰無仮説が棄却されたら Model-5を選択し，棄却されなければ次のステップに進む(ここまでは両者共通)。次に Path1では最初に Model-2が真という帰無仮説に対して Model-4が真という対立仮説を考え，帰無仮説が棄却されたら Model-4を選択し，棄却されなければ次のステップに進む。その場合には Model-1が真という帰無仮説に対して Model-2が真という対立仮説を考え，帰無仮説が棄却されたら Model-2を選択し，採択されたら Model-1を選択することになる（Path2についても同様である)。

Table 3-14. Dataset used for the simulation of ANOVA type in the nested model

Year	Area1	Area2	Area3
1	1	0.8	1.2
2	1	0.8	1.2
3	1	0.8	1.2
4	1	0.8	1.2

実際，Path1と Path2で異なる検定結果が生じる場合もあり（庄野，2000)，解釈が難しいケースも存在する。我々は，そのような場合に Table 3-14のような２つの判断に従って選択パフォーマンスを調べる計算機実験を行った。例えば，Path1による stepwise 検定では一番単純な Model-1が選択されて，Path2によるそれでは Model-4が選ばれた場合，No.6では即矛盾と判断してしまうのに対し，No.8では実際に検定を行っていない Path2における Model-1 vs. Model-3の検定を行い，Model-1が選ばれた場合には最終的な結果が一致したと考えてこのモデルを選択し，Model-3が選択された場合には，ここではじめて矛盾と判断する。

結果は Table 3-15のようになる。全体として，情報量規準 BIC の選択パフォーマンスが stepwise 検定のそれに比べて多少良くなった。また，有意水準の値を極めて小さくすると変数の少ないモデルが選ばれる傾向があることも示されたため，このような stepwise 検定を用いる際には，十分に注意が必要である。

― Stepwise 検定における２つの判断―

No.	Path1による検定結果	Path2による検定結果	判断
1	Model-2	Model-3	矛盾
2	Model-2	Model-4	Model-2
3	Model-4	Model-3	Model-3
4	Model-1	Model-3	矛盾
5	Model-2	Model-1	矛盾

注）1-5のケースでは（判断１と判断２を区別せず）共通の判断を行なう。

判断１

No.	Path1による検定結果	Path2による検定結果	判断
6	Model-1	Model-4	矛盾
7	Model-4	Model-1	矛盾

判断２

（6，7のケースで即に矛盾とせず，検定を行った上で矛盾かどうかを判断）

Table 3-15. Results of model selection using the data from Table 3-14 (simulation of ANOVA type in the nested model)

Decision-1

Model	AIC	BIC	F(0.01)	F(0.03)	F(0.05)	F(0.07)	F(0.10)
1	4	18	148	59	36	17	13
2	0	1	0	1	1	1	1
3 (true)	637	887	795	870	871	861	826
4	132	55	8	24	42	56	78
5	227	39	10	22	40	56	81
discrepancy			39	24	10	9	1

Decision-2

Model	AIC	BIC	F(0.01)	F(0.03)	F(0.05)	F(0.07)	F(0.10)
1	4	18	161	67	37	18	13
2	0	1	0	1	1	1	1
3 (true)	637	887	795	870	871	861	826
4	132	55	8	24	42	56	78
5	227	39	10	22	40	56	81
discrepancy			26	16	9	8	1

No.	該当するケース	行動	結果	判断
8	No.6	Model-1 vs. Model-3	Model-1	Model-1
			Model-3	矛盾
9	No.7	Model-1 vs. Model-2	Model-1	Model-1
			Model-2	矛盾

3-6. 正規混合分布におけるモデル選択

本節では，水産資源分野における体長組成データの年齢分解などに使用される正規混合分布モデルのコンポーネント数（山の数）の推定方法について議論する。具体的には（3.28）式のように定式化出来，Fig. 3-2のような正規分布が複数混合された分布において，コンポーネント数 m と（3m-1）個のパラメーター α_i, μ_i, σ_i^2が推定目標となる。

$$Model(X)=\sum_{i=1}^{m}\alpha_i f(x\mid\mu_i,\sigma_i^2) \tag{3.28}$$

但し $X=(X_1,\cdots,X_n)$: 標本ベクトル，n: 標本数，$f(x\mid\mu_i,\sigma_i^2)$: 正規分布 $N(\mu,\sigma^2)$ に従う確率密度関数，α_i: コンポーネント i の混合比率（$\sum_{i=1}^{m}\alpha_i=1$）と仮定する。

このケースでは最尤推定量の漸近的性質（一致性・漸近正規性）が成り立たないためにカイ二乗検定を用いることが出来ない。なぜなら，

$$\begin{aligned}&X_1,\ldots,X_n(i.i.d)\sim P(x\mid\alpha,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2)\\&=\alpha f(x\mid\mu_1,\sigma_1^2)+(1-\alpha)f(x\mid\mu_2,\sigma_2^2)\end{aligned} \tag{3.29}$$

のような簡略化されたモデルにおいて帰無仮説を $\alpha=1$，対立仮説を $0\leqq\alpha<1$ とおくと，帰無仮説の下で真値 $\alpha=1$ はパラメーター空間の内点にならないことや（Charnoff, 1954），帰無仮説の下で Fisher 情報行列が正則にならないがゆえに，最尤推定量の漸近的性質が満たされないのである。その他，パラメーターの認定可能性の問題も生じている。

そこで，式（3.28）のモデルに Leroux（1992）が提案した罰金付き尤度法を用いてコンポーネント数やパラメーターの推定を行う。具体的には，

$$L=-2l(\hat{\Theta}\mid X)+2a(m,n) \tag{3.30}$$

とおいて，コンポーネント数 m の推定量 $\hat{m}$ を

$$\hat{m}=\min_m\{\min_\Theta L\} \tag{3.31}$$

の最小化問題の解として求める。

但し $l(\Theta\mid X)$ は対数尤度関数，$a(m,n)$ は

$$a(m,n)>0,\ a(m+1,n)>a(m,n),\ a(m,n)/n\to 0\quad(\text{as } n\to\infty) \tag{3.32}$$

を満たす実数列を表わし，その定め方によって多くの罰金形を表現出来る。例えば $a(m,n)=3m-1$ とおくと AIC に，$a(m,n)=(3m-1)\log n/2$ とおくと BIC になる。この罰金付き尤度法は漸近的な場合にコンポーネント数を過小推定しないことが証明されているが，過大推定の可能性については残されたままであ

るため，欠点の存在には注意が必要である。

なお，この漸近的な場合におけるコンポーネント数の過大推定の可能性に関する欠点に対処するため，フルモデル（制約無しのモデル）と構造モデル（制約付きのモデル：体長年齢分解においては von Bertalanffy の成長曲線などを仮定）の利点を併せ持つ，新しい推定方法（Eguchi and Yosioka, 2001）の水産資源分野への適用可能性について，統計数理研究所の研究者と共同研究を行っている。

本報告では，AIC，BIC と混合比率の事前分布として Dilichlet 分布

$$\pi(\alpha)=c\prod_{i=1}^{m}\alpha_i^{-1/2}\text{（但し c=}\Gamma(\mathrm{m}/2)/\pi^{m/2};\ \alpha_m=1-\sum_{i=1}^{m-1}\alpha_i \text{ とおく）} \tag{3.33}$$

を仮定した３つの規準を用いて真のコンポ－ネント数を４としたモデルから乱数を100回発生させて計算機実験を行った結果，Dilichlet 事前分布による Bayes 型規準の真のモデルを選ぶ選択パフォーマンスが AIC や BIC に比べてわずかに良くなった（Table 3-16）。しかしこの Bayes 規準については（3.32）式を満たすという数学的な証明が出来ていないため，理論的正当性は得られていない。

また，AIC などの代表的な情報量規準の場合と同様に，コンポーネント数の過大推定の可能性が残っていることにも注意が必要である。なお，今回の計算機シミュレーションでは，Dilichlet 分布

$$\pi(\alpha\mid\beta)=\frac{\Gamma(\sum_{i=1}^{m}\beta_i)}{\prod_{i=1}^{m}\Gamma(\beta_i)}\prod_{i=1}^{m}\alpha_i^{\beta_i-1} \tag{3.34}$$

において $\beta_1=\cdots=\beta_m=1/2$と仮定した無情報事前分布を使用したが，この超パラメーターを経験 Bayes 法などによって推定することも可能である。これらの指標は時系列データ（例えば月毎や４半期毎の体長組成データでそれぞれコンポーネント数が異なる場合）にも適用可能であり，実用性は高いと考えられる。

なお，3-6-1節で正規混合分布モデルにおける計算機シミュレーションの概略について述べる。モデル式などは次節に示すが，確率密度関数は Fig. 3-2のものを用いる。

3-6-1. 正規混合分布における計算機実験（情報量規準のパフォーマンス評価）

ここでは，正規混合分布モデルにおける計算機シミュレーション実験を考え，コンポーネント数（山の数）とパラメーター（各々の正規分布の平均と分散）の値の推定を情報量規準によって行った。コンポーネント数が４であり，一見したところ各々の正規分布の識別が難しいような分布（Fig.3-2の確率密度関数）

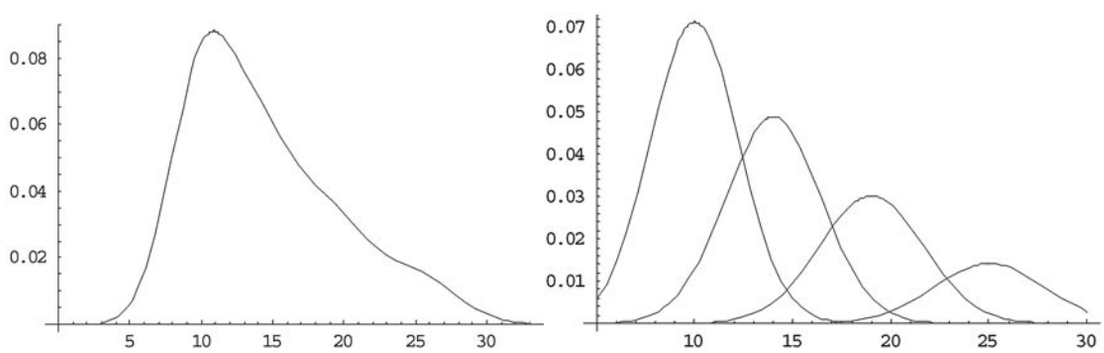

Fig. 3-2. Probability density function for the simulation studies using Normal mixture distribution.

Table 3-16. Summary of the model selection by Normal mixture distribution

Case-1(sample size-3000)

m	MLL	AIC	BIC	Dilichlet
2				
3			11	5
4 (true)	66	90	89	94
5	1	9		1
6	33	1		

Case-2(sample-size-7000)

m	MLL	AIC	BIC	Dilichlet
2				
3			2	
4 (true)	93	98	98	100
5		1		
6	7	1		

$$X_1, \ldots, X_n \sim g(x)=0.4\, f(x \mid 10, 5)+0.3\, f(x \mid 14, 6)+0.2\, f(x \mid 19.7)+0.1\, f(x \mid 25, 8) \quad (3.35)$$

但し，$f(x \mid \mu, \sigma^2)$ は平均 μ，分散 σ^2 を持つ正規分布の *p. d. f*（確率密度関数）を表わすこととする。から乱数を100回発生させて，4つの規準（MLL（最大対数尤度），AIC，BIC と Dilichlet 事前分布を用いた Bayes 型規準）により，この正規混合分布のコンポーネント数 m が 4 であると正しくと推定される回数を計算した。

結果は Table 3-16の通りだが，ケース 1（標本数3,000）とケース 2（標本数7,000）のいずれにおいても，Dilichlet 事前分布による Bayes 型規準と AIC，BIC は真のモデルを選ぶ選択パフォーマンスがかなり良く，大きな違いはみられなかった。このことから，理論的な正当性はともかくとして，実際上はこれらの情報量規準（AIC 及および BIC）は正規混合分布モデルにおけるコンポーネント数の推定に使用可能であると考えられる。ただし，この計算機実験から判断する限り，Dilichlet 事前分布による Bayes 型情報量規準の利用を推奨する次第である。

3-7. まとめ

第 3 章では，主に CPUE 標準化を想定した分散分析モデル, 共分散分析モデルを用いて, 小標本の場合，大標本の場合，ネスト構造を持つモデルの場合，正規混合分布モデルの場合など，様々なケースを取り上げた。水産資源分野で広く知られている情報量規準である AIC の他に，BIC, CAIC, c-AIC, HQ, TIC などを使用し，実際例を用いて利用する情報量規準によってモデル選択結果が異なること，及び複数の候補モデルの中から定めた真のモデルから乱数を発生させて正しいモデルを選ぶという選択パフォーマンスをシミュレーション実験により計算し，情報量規準の良さを詳しく評価, 検討した。なお, ネスト構造を持つモデルでは，F 検定やカイ二乗検定に代表される stepwise 検定も使用可能であり，計算機シミュレーションを通じた情報量規準と stepwise 検定の比較も合わせて実施した。以下, 本章で得られた各節毎の結果を要約して述べる。

最初に3-1節では，CPUE に影響を与えている要因効果を統計的に取捨選択するという観点から，CPUE 解析におけるモデル選択（変数の取捨選択）の重要性について述べた。

3-2節では，小標本の場合や未知パラメーター数の標本数に占める割合が高い場合に，c-AIC（finite correction of AIC）によるモデル選択結果が AIC などによるそれと異なることをサクラマスの成長データを用いて例示した。さらに，分散分析型のシミュレーションにより，c-AIC の選択パフォーマンスが AIC のそれに比べてかなり高いことを証明した。

3-3節では，大標本の場合にも AIC が偏りを持つ可能性があることを理論的に説明し，使用する規準によりモデル選択結果に差が生じることをインド洋におけるキハダマグロ CPUE 解析の実際例により例示した。さらに，漸近的に望ましい性質である一致性を持つ情報量規準（Baysian Information Criterion（BIC), Hannan and Quinn（HQ）and Consistent AIC（CAIC））が AIC に比べて全体として優れていることを，回帰分析型の計算機実験を通じて示した。

合わせて，HQ における定数項 c の定め方について議論し，c=2.01ないしは2.71とおくこと，あるいは $c = 2+k\,[\log n/\{\log(\log(n))\}-2]$（k：任意定数，0<k<1）とおき k を0.05～0.1程度に設定することの妥当性について，情報量規準のペナルティ項の比較や回帰分析型シミュレーション，grid search を用いて検証した。

3-4節では，大部分の CPUE 標準化も含まれるネスト構造を持つモデルにおいて，従来性能が良いと言われてきた TIC（Takeuchi's information criterion）が，正規誤差を持ちかつ連結関数が恒等写像であるような一般化線形モデルでは AIC と漸近的に同等になることを理論的に証明し，合わせて TIC と AIC の選択パフォーマンスにほとんど差がないことを，計算機実験により示した。

3-5節では，同じくネストモデルについて，使用可能な stepwise 検定と情報量規準について，適用上の長所と短所を整理し，特に stepwise 検定における検定のパスに関して，モデル選択の観点から詳しく議論した。情報量規準では総当り法でその値を調べる必要があり，候補モデルが多い場合には複雑なことから stepwise 検定を使用する選択も一案であるが，stepwise 検定ではその順序により矛盾した結果が得られることもあり，そのような場合の判断も含めて比較検討した。

具体的には，計算機シミュレーション実験を通じて情報量規準と stepwise 検定のパフォーマンスの比較を行い，一般に前者が多少優れていること，及び後者で有意水準を小さく設定した場合にパラメーター数が少ない単純なモデルが選ばれがちであることを示した。

3-6節では，体長組成データの年齢分解（年齢査定）に使用可能な正規混合分布モデルにおけるコンポーネント数の推定問題を取り上げ，カイ二乗値に基づく検

定が理論的に使用出来ないことを示し，計算機実験を通して，AIC や BIC に代表される情報量規準が実際上は利用可能なこと，および Bayes 型のディリクレ事前分布を用いた情報量規準が優れていることを合わせて実証した。

第 4 章　ニューラルネットワークによる CPUE 予測と要因分析ミナミマグロ資源への適用

4-1. はじめに

本章では，CPUE 標準化に関する水産資源に特有の問題である資源量指数に焦点を合わせており，操業がない時空間のミナミマグロ CPUE をニューラルネットワークにより予測し，要因分析（CPUE 年トレンドの抽出）を試み過去の知見と比較した。そして，ミナミマグロの資源研究において長年議論となっていた操業がない時空間の CPUE に関する一定の知見を得た。このミナミマグロの資源量指数の問題については，次の4-2節で詳しく述べる。

さらに，ニューラルネットワーク及び欠測値の問題を EM アルゴリズムにより補間した方法における予測値の精度検証を行った。解析には実際の漁業データ（日本のはえ縄船による5×5/ 月別に集計されたミナミマグロの漁獲量および努力量データ）を利用しているため，モデルの評価（予測値の精度検証）は n-fold cross-validation と呼ばれるデータをランダムに n- 分割して，故意に隠したサブセットの予測値を順番に推定する手法を使用した。

4-2. ミナミマグロに関する資源量指数の問題（漁獲がない時空間の取り扱い）

CPUE 標準化の主目的は年効果の推定であり，抽出された年トレンドは相対的な資源の増減傾向を表している。しかし，多くの CPUE 解析において便宜的に区分されたエリア分けを使用していることが多く，個々のサブエリアの大きさが異なり，かつ年と海区の交互作用が認められる場合には，相対的なエリアサイズによる補正を必要とする。すなわち，推定された CPUE の年効果に相対的なエリアサイズを掛け合わせたものを資源量指数（abundance index, AI）と呼んでおり，通常は CPUE 推定値に基づいた資源量指数が相対資源量に対応すると考えられている（能勢ほか，1988）。

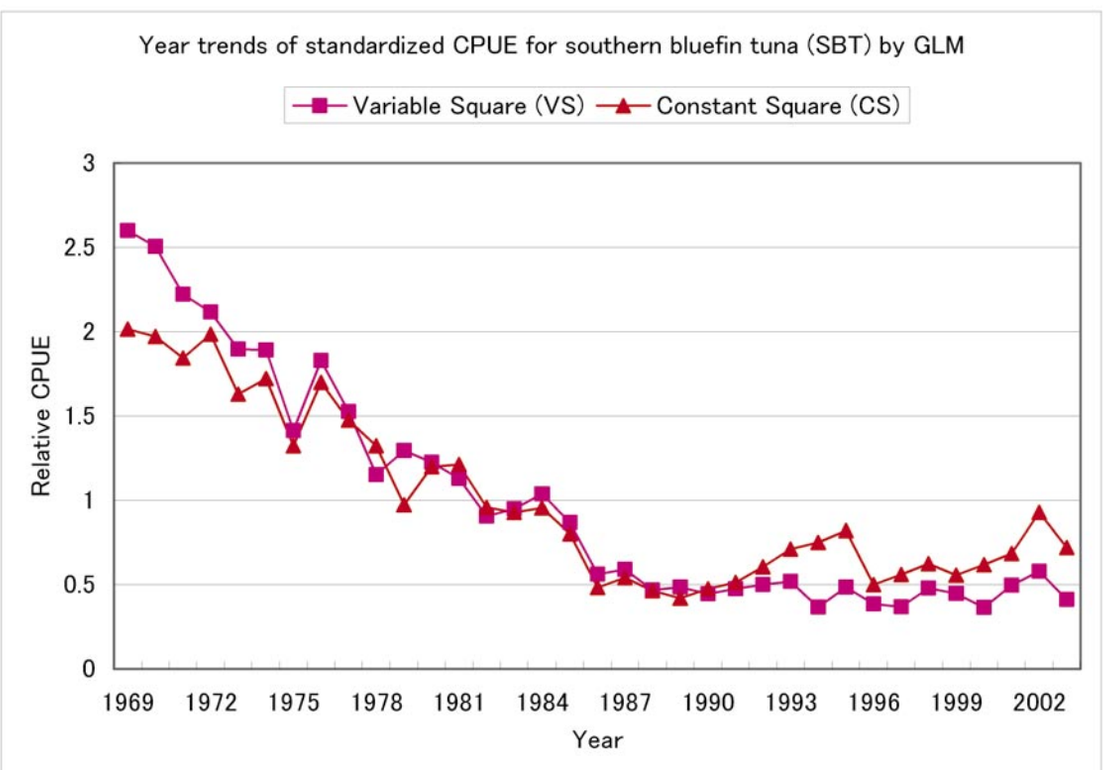

Fig. 4-1. Year trends of standardized CPUE based on the assumption of constant square (CS) and variable square (VS) obtained from the ANCOVA model.

$$AI_{ij} = w_j CPUE_{ij} \quad (4.1)$$

（但し$\sum_j w_j = 1$　（i: YEAR、j: AREA、w_j: AREA(j)の相対面積指数）とする）

この相対面積指数は，式（4.1）のように年に依存しないと仮定することが一般的である。しかし，まぐろ・かつおなどの魚は広範囲の時空間的な移動を行い，それに合わせて漁場が変化する場合も多い。そのため，年が経つにつれて魚の分布が縮小し，なおかつ漁場が縮小しているような場合に，（4.1）式を用いて資源量指数を計算すると，資源の過大評価につながる恐れがある。

この問題に対して，ミナミマグロ漁業を取り扱っている国際委員会CCSBTでは，（4.1）式の方法の他に1990年代初めから（4.2）式のような相対面積指数が年に依存するという考え方を取り入れている（CCSBT, 1998）。すなわち漁業の変化に対してエリアサイズも変化させるという方法である。この考え方は（4.1）式のCS仮説（constant square仮説）に対してVS仮説（variable square仮説）と呼ばれている。

$$AI_{ij} = w_{ij} CPUE_{ij} \quad (4.2)$$

（但し$\sum_j w_{ij} = 1$　（w_{ij}: YEAR(i)かつAREA(j)の相対面積指数）とする）

ミナミマグロ資源に対して（4.2）式のようなVS仮説の考え方が用いられる理由としては，定義されている時空間の中で過半数を占める操業のないセルの取り扱いが，資源量指数の年トレンドに大きな影響を与えているためと思われる。

日本の遠洋延縄漁船によるミナミマグロのCPUE（年齢別漁獲尾数・努力量）データは月毎に，そして海域毎（緯度・経度：5×5度のセルで総計）に集計された1960年から2003年まで40年余りのデータを使用しているが，計算に用いた過去に一度以上漁獲努力があるセル（5×5/月，全29,820レコード）のうち4分の3余りが操業のないセルとなっており，この過去に操業があり現在操業がない部分（以後セルと呼称する）の取り扱いが重要な問題となっている。

実際問題として，一般化線形モデルを利用した解析では，各々の説明要因のパラメーター推定値によるType Ⅲの平方和に基づくLSMEANS（least squared means）と呼ばれる値を用いて要因分析（CPUE年トレンド抽出等）を行うことが多い。ミナミマグロのCPUE解析においても（4.3）式および（4.4）式で表現される，それぞれ年効果および年とエリアの交互作用効果に関するLSMEANSを利用して，CPUEの年トレンドを抽出している。

$$CPUE_i = \exp\{(\text{Intercept}) + (\text{Year})_i + (\overline{\text{Area}}) + (\overline{\text{Season}}) + (\overline{\text{Year*Area}})_{i\cdot} + (\overline{\text{Year*Season}})_{i\cdot} + (\overline{\text{Area*Season}}) \ldots\} - (\text{constant\_term}) \quad (4.3)$$

但し，$\overline{\text{Area}} = \frac{1}{N_j}\sum_{j=1}^{N_j}(\text{Area})_j$，$(\overline{\text{Year*Area}})_{i\cdot} = \frac{1}{N_j}\sum_{j=1}^{N_j}(\text{Year*Area})_{ij}$

などと定義する。

$$CPUE_{ij} = \exp\{(\text{Intercept}) + (\text{Year})_i + (\text{Area})_j + (\overline{\text{Season}}) + (\text{Year*Area})_{ij} + (\overline{\text{Year*Season}})_{i\cdot} + (\overline{\text{Area*Season}})_{j\cdot} \ldots\} - \text{constant\_term} \quad (4.4)$$

（$(\overline{\text{Season}} = \frac{1}{N_k}\sum_{k=1}^{N_k}(\text{Season})_k$，$(\overline{\text{Year*Season}})_{i\cdot} = \frac{1}{N_k}\sum_{k=1}^{N_k}(\text{Year*Season})_{ik}$

等と定義する）

ただし，ミナミマグロ資源においては，標準化されたCPUEから相対的なエリアサイズを考慮した資源量指数に変換する過程で，操業のない欠測セルを区分されたサブエリアで一定と仮定した場合（CS仮説）とゼロと置いた場合（VS仮説）で近年のCPUE年トレンドが異なっており（Fig. 4-1），これらをチューニング指数として使用したVPA（virtual population analysis）などの評価モデルにより推定された資源量の絶対値も極端に異なる結果になってしまった。

そこで，本論文ではニューラルネットワークにより（ミナミマグロ漁業データの過半数を占める）操業がないセルのCPUE予測を行い，予測された値を用いた簡便な要因分析手法（CPUE年トレンドの抽出方法）を提案する。

4-3. 解析手法

本研究では，計算での入力変数と出力変数を次のように設定し，教師付きニューラルネットワークの典型的なアルゴリズムである誤差逆伝搬法を使用した。

出力変数（応答変数）―日本のはえ縄船による4歳以上のミナミマグロCPUE（=Catch/Effort）Catch：漁獲尾数，Effort：1000 hooks.

入力変数（説明変数）― Year　：年（1969-2003）
Month　：月（4-9）
Latitude　：緯度（30S-50S, 5度刻み）

Table 4-1. Dataset of southern bluefin tuna used for 5-fold cross-validation

Sub-set	No.of data	CPUE	Scenarios					
			Base case	I	II	III	IV	V
1	1,539	○	Rule	C.V.	Rule	Rule	Rule	Rule
2	1,540	○	Rule	Rule	C.V.	Rule	Rule	Rule
3	1,540	○	Rule	Rule	Rule	C.V.	Rule	Rule
4	1,539	○	Rule	Rule	Rule	Rule	C.V.	Rule
5	1,539	○	Rule	Rule	Rule	Rule	Rule	C.V.
Others	22,123	×	Pred	Pred	Pred	Pred	Pred	Pred

Remark) Rule, C.V. and Pred show the sub-dataset for rulemaking, cross-validation and prediction, respectively. (○-supervised data, ×-unsupervised data)

Table 4-2. Pearson's correlation coefficient of observed and predicted CPUE in each sub-dataset by neural networks

Sub-set	1	2	3	4	5
Correlation coefficient	0.5905	0.6019	0.5626	0.5857	0.6259

Table 4-3. Pearson's correlation coefficient of observed and predicted CPUE in each sub-dataset by MCMC method (EM algorithm)

Sub-set	1	2	3	4	5
Correlation coefficient	0.2867	0.3132	0.3205	0.3503	0.3246

Longitude：経 度（20W-0-90E-180-175W, 5 度刻み）

注）入力変数はすべて順序のないカテゴリカル変数とする

ニューラルネットワークによる予測と要因分析は以下の手順にて行ったが,

用いたセル数（添字の組合せ）は教師付きと教師無し合計で29,820件となる。

Step-1（CPUE 予測）

漁業（操業）があるセル（5×5/ 月）のデータ（教師付き：全7,697件）を用いて，ニューラルネットワークによるルール作成を行い，対応するセル（5×5/ 月：教師付き：全7,697件）のCPUE 予測値を算出し，これをベースケースとする。次に作成されたルールを使用して操業（漁獲）がないセル（5×5/ 月：教師無し：全22,123件）の CPUE の予測値を推定する。

Step-2（要因分析）

Step-1で計算された全てのセル（5×5/ 月：全29,820件）の予測値を用いて CPUE の年トレンドを求める。具体的には, 最初に月（month）による平均（もしくは総和）を計算し, 次に緯度および経度（latitude・longitude）による平均（もしくは総和）を計算する。これらのプロセスを式で表すと以下のようになる。

$$CPUE^{pre}_{year} = \frac{1}{N_L}\left[\sum_{Longitude}\left\{\frac{1}{N_l}\sum_{latitude}\left(\frac{1}{N_m}\sum_{month} CPUE^{pre}_{year,month,latitude,Longitude}\right)\right\}\right] \quad (4.5)$$

但し，（4.5）式は月による平均を取り緯度・経度による平均を取った場合を表しており，N_m, N_l, N_L はそれぞれ月，緯度，および経度の要因の水準数を表す。

CPUE 予測に関する教師付きニューラルネットワークでは，中間層を１つに固定した上で誤差逆伝播アルゴリズムを使用し（中間素子数は逐次的に決定），具体的な計算条件は以下の通りであるが，解析にはデータマイニングツール KINO suite-PR（Version 3.20, 東芝）を使用した（Tsukimoto, 2000）。

- 教師付きデータを学習用（85%）と検証用（15%）にランダム分割する
- 誤差逆伝搬アルゴリズムを利用し，関数形として

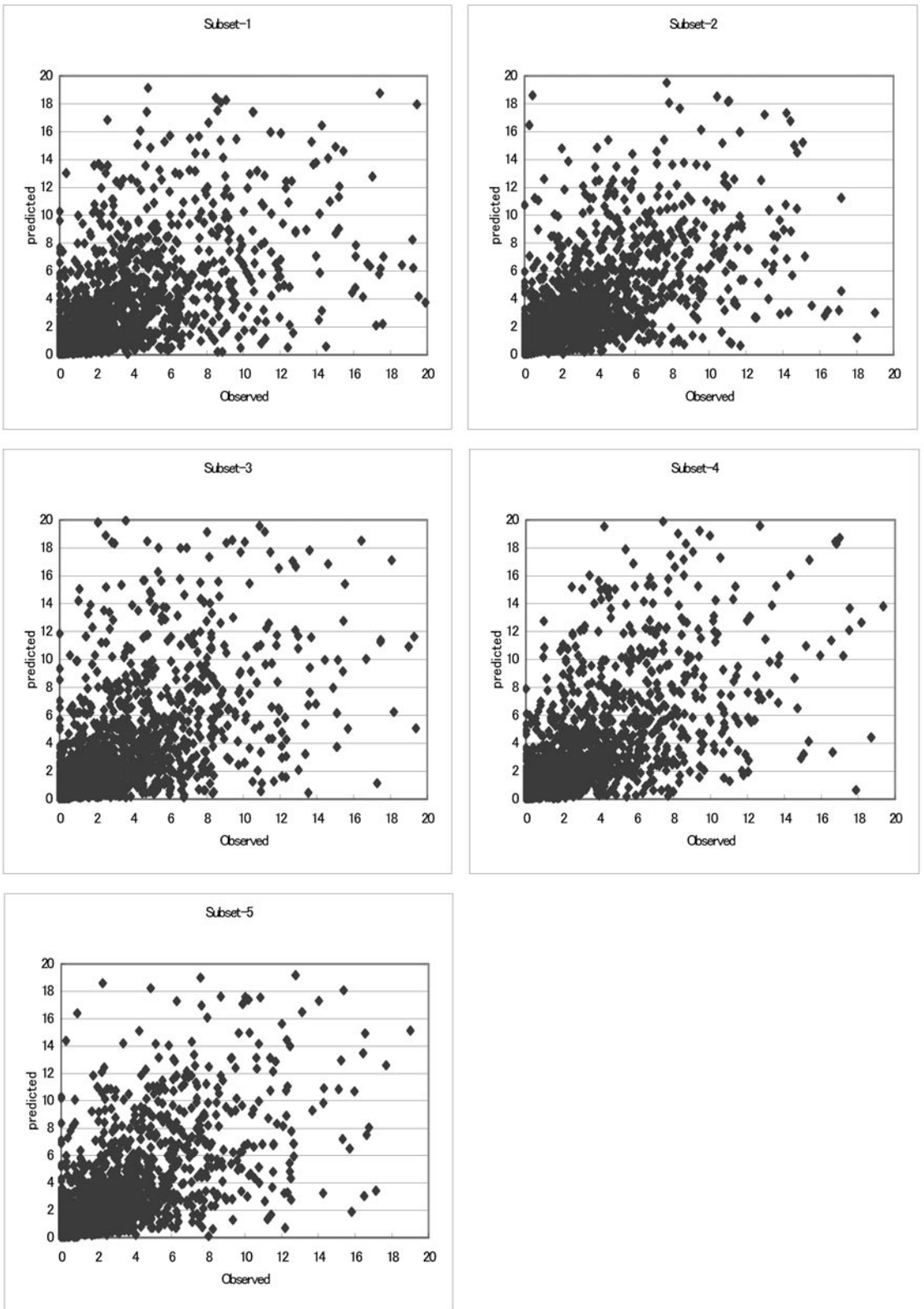

Fig. 4-2. Correlation plots of the observed and the predicted CPUE in the neural networks for southern bluefin tuna in each sub-dataset used for 5-fold cross-validation.

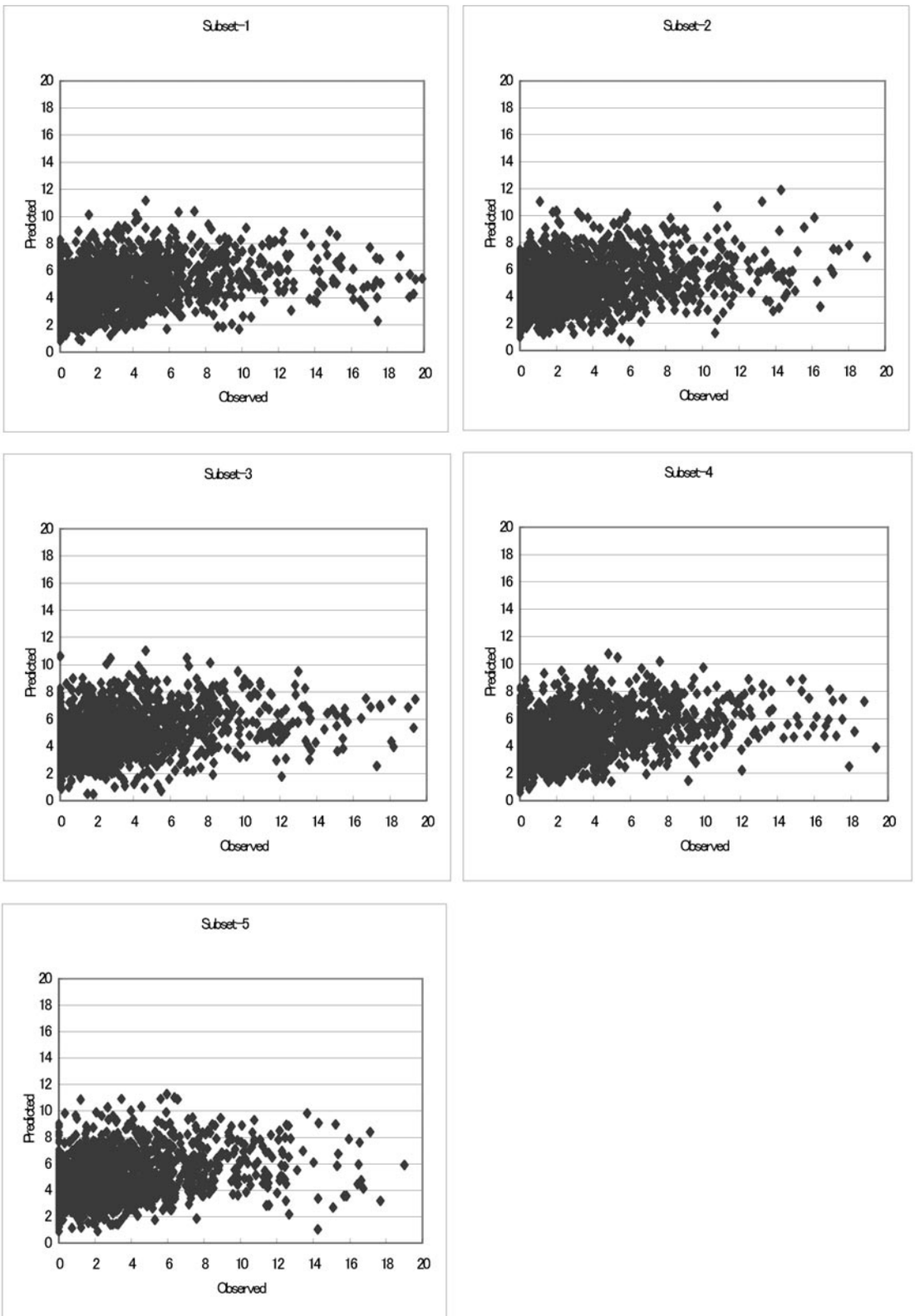

Fig. 4-3. Correlation plots of the observed and the predicted CPUE in the MCMC method based on the EM algorithm for southern bluefin tuna in each sub-dataset used for 5-fold cross-validation.

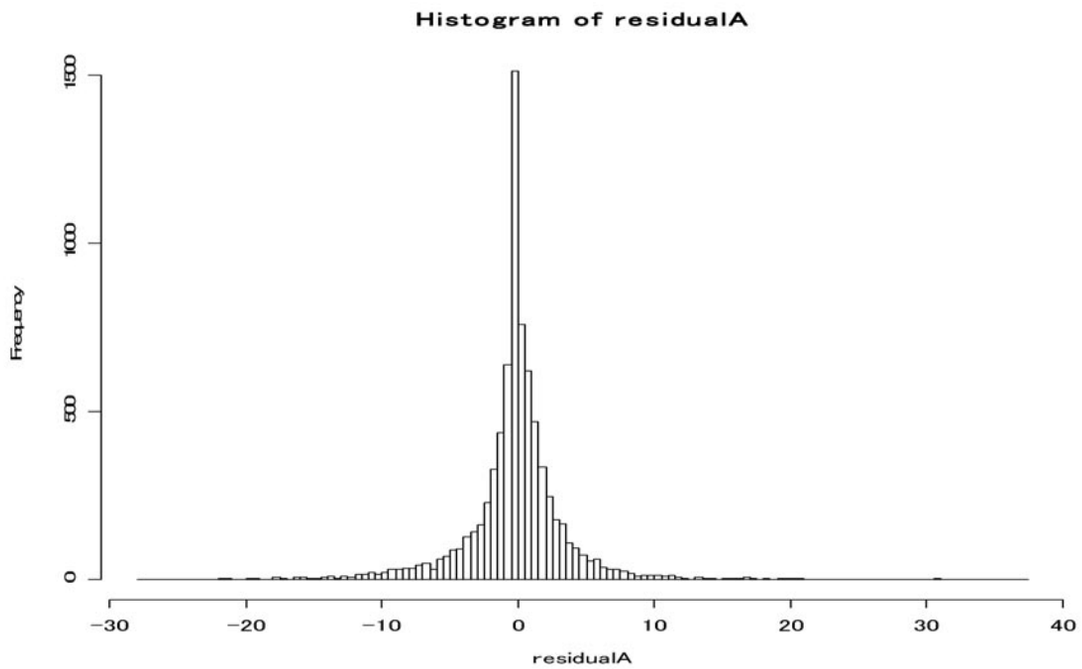

Fig. 4–4. Residual plots based on the observed CPUE and the corresponding predicted one for southern bluefin tuna obtained from the neural networks.

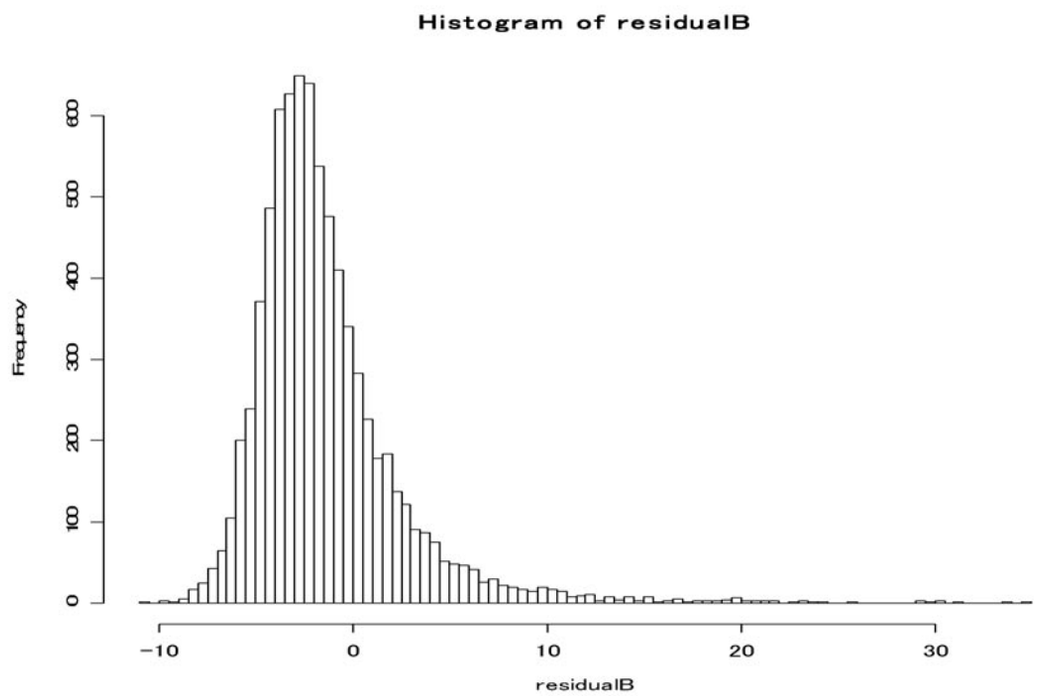

Fig. 4–5. Residual plots based on the observed CPUE and the corresponding predicted one for southern bluefin tuna obtained from the MCMC method based on the EM algorithm.

Table 4-4. Median, mean, minimum and maximum values of the absolute error in each sub-dataset by neural networks

Sub-set	1	2	3	4	5
Median	1.30983	1.08844	1.23445	1.09875	1.13160
Mean	2.41829	2.06122	2.42284	2.17452	2.03043
Minimum	0.00017	0.00005	0.00011	0.00046	0.00044
Maximum	22.7685	37.4014	29.1902	34.0170	30.5308

Table 4-5. Median, mean, minimum and maximum values of the absolute error in each sub-dataset by MCMC method (EM algorithm)

Sub-set	1	2	3	4	5
Median	2.74534	2.92685	2.74620	2.59463	2.88151
Mean	3.23935	3.20927	3.12462	2.99480	3.12390
Minimum	0.00181	0.00005	0.00032	0.00647	0.00649
Maximum	33.9628	34.7273	29.4821	31.4368	25.5896

Table 4-6. Median, mean, minimum and maximum values of the absolute error in the 5-fold cross-validation (Comparison between neural networks and MCMC method)

Model	Neural Networks	MCMC (EM algorithm)
Median	1.164076	2.781106
Mean	2.221564	3.138392
Minimum	0.000005	0.000315
Maximum	37.401410	34.72727

Table 4-7. Pearson's correlation coefficient of observed and predicted CPUE in the 5-fold cross-validation (Comparison between neural networks and MCMC method)

Model	Neural Networks	MCMC (EM algorithm)
Correlation coefficient	0.59046	0.31644

シグモイドを使用する

- 中間層は1つに固定し，中間素子数は以下の条件により逐次的に決定する
  - 中間素子数は2から1ずつ増加させ，以下の条件を満たしたら停止する
  - 入力変数は全て正規化（0-1変換）し，出力変数における観測値と予測値の最大絶対誤差が，学習用および検証用の両方で0.01以下になったら停止させる（順序のないカテゴリカル変数ゆえに，ダミー変数を作成）
  - 各々のケース（中間素子数を固定した場合）における学習は100,000回で停止させる

注）上の停止規則を適用し，ベースケースでの中間素子数は36と推定された。

なお，4-4節で述べる予測値の精度検証を目的として，ニューラルネットワークによる解析と全く同じデータを使用して，EMアルゴリズムに基づくMCMC法（Markov Chain Monte Carlo method）による計算を行った（Little and Rubin, 2002）。具体的には，共分散分析モデルにおける平均ベクトル及び分散共分散行列をJeffreysの事前分布を仮定してシミュレートするBayes的な方法を使用しており，多重代入法により5回の試行における平均値をCPUEの欠測部分に補完したが，解析にはSAS（Version 9.1.3, SAS Inc.）を利用した。

4-4. 予測値の精度検証

本節では，ニューラルネットワークに予測された

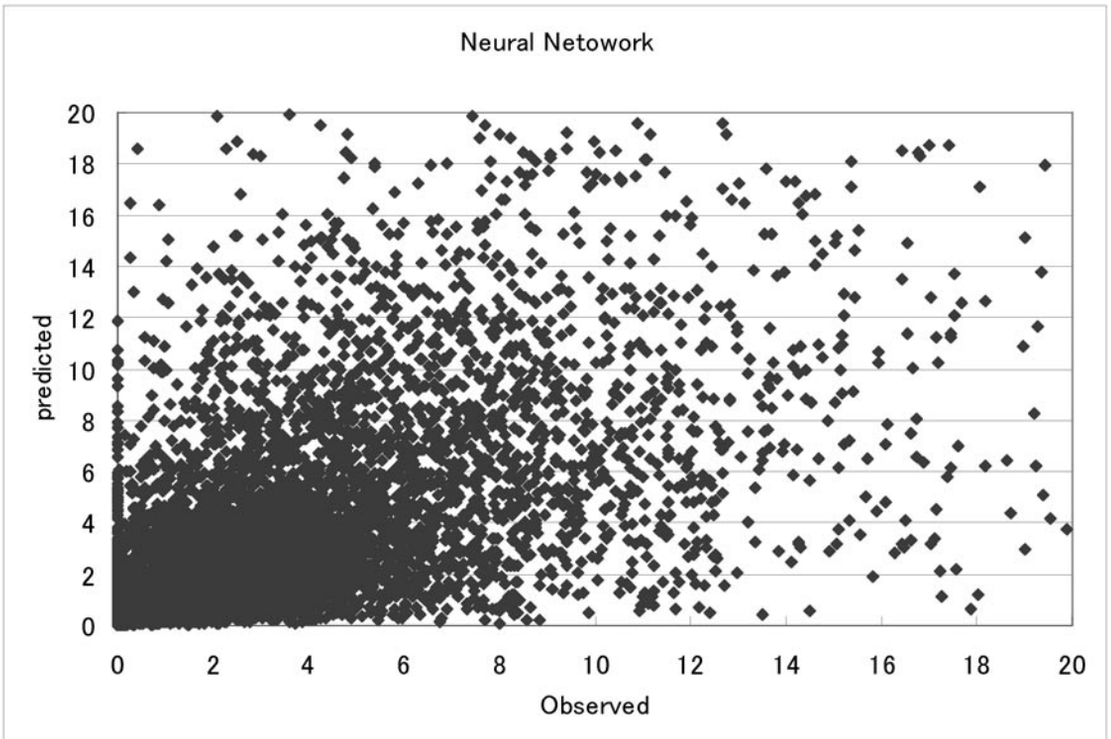

Fig. 4–6. Overall correlation plots of the observed and the predicted CPUE in the neural networks for southern bluefin tuna.

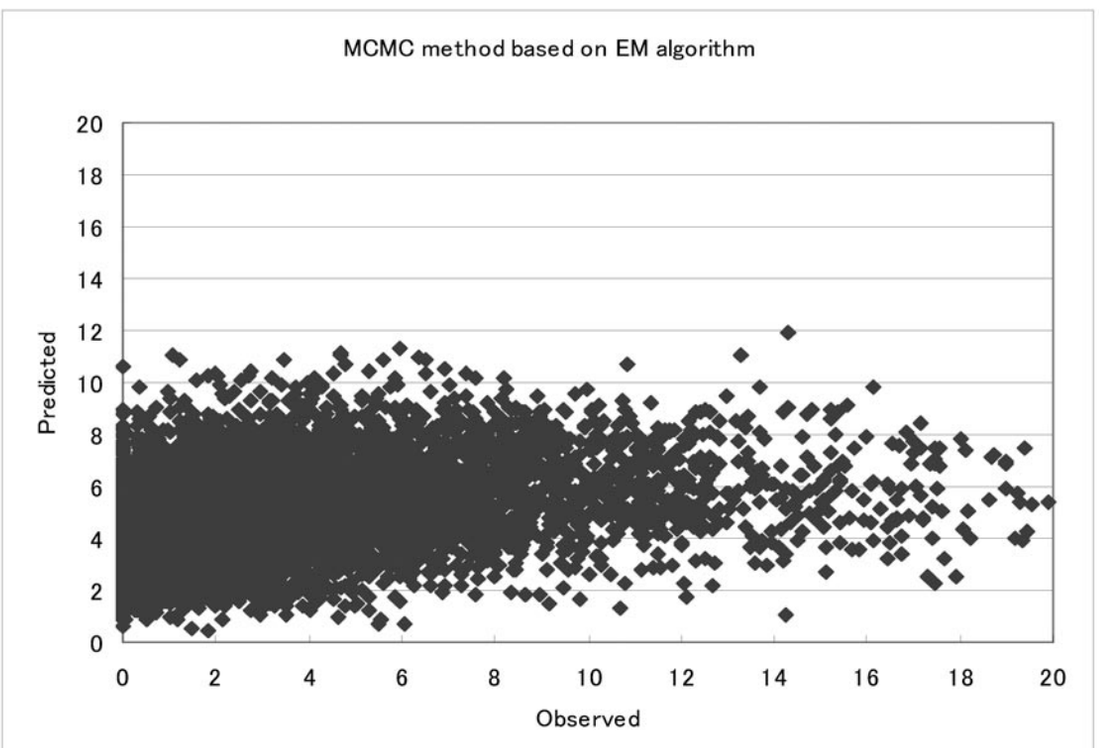

Fig. 4–7. Overall correlation plots of the observed and the predicted CPUE in the MCMC method based on the EM algorithm for southern bluefin tuna

CPUE の精度について検討する。原則として，操業のないセル（教師無しデータ）による精度は計算出来ないため，n-fold cross-validation と呼ばれる正解の一部を故意に隠して予測を行い，観測値と比較することによって精度の検証を行った。ここでは，操業があるセルである教師付きデータをランダムに 5 分割し，それぞれ 1 つずつのサブセットを隠して，すなわち教師無しとみなしてニューラルネットワークによるルールを作成した。全部で 5 通り（Table 4-1, Ⅰ－Ⅴ）の計算を行い，比較検討したが，精度評価のための5-fold cross-validation の概略を Table 4-1に示す。

なお，精度検証のための 5 つのニューラルネットワーク（1-5）の計算条件は，全てベースケースと同じくし，もちろん教師付きデータのランダム分割を行った。そして，5-fold cross-validation を用いた精度検証には，観測値と予測値の標本相関係数および絶対誤差の平均値および中央値を利用した。観測値と予測値の平方誤差でなく絶対誤差を使用した主な理由は，ニューラルネットワークにおける中間素子数決定の停止規則との整合性を保つためである。

標本相関係数の分割されたサブセット毎の値およびそれらのプロットは，Table 4-2および Fig. 4-2のようになる。全てのサブセット（1-5）について標本相関係数の値は比較的に高くなっており，そのことは相関プロットからも見て取れる。

これと全く同じデータセットを用いて（5 つに分割されたサブセットも全く同一のものを使用），欠測値解析のための統計手法（EM アルゴリズムで補間した MCMC 法 : 分散分析型モデルを仮定して，多重代入法により 5 回の試行の平均値を補完した）により計算された標本相関係数（Table 4-3）と比較すると，ニューラルネットワークに基づく CPUE 予測値により算出された相関係数の値は，MCMC 法によるそれに比べて 2 倍程度と，かなり高くなっている。

また，それらの相関プロット（Fig.4-3）から，極端に小さな観測値を多少大きく，かなり大きな観測値を小さめに推定するという中心への回帰傾向が，EM アルゴリズムで補間した MCMC 法の場合に認められる。

EM アルゴリズムは欠測値補間の代表的なアルゴリズムであり，多変量正規分布を仮定した線形モデルと位置付けられる。実際には，観測データと繰り返しが t 回目のパラメーター推定値が与えられた下で完全データの対数尤度（観測データの対数尤度と欠測部

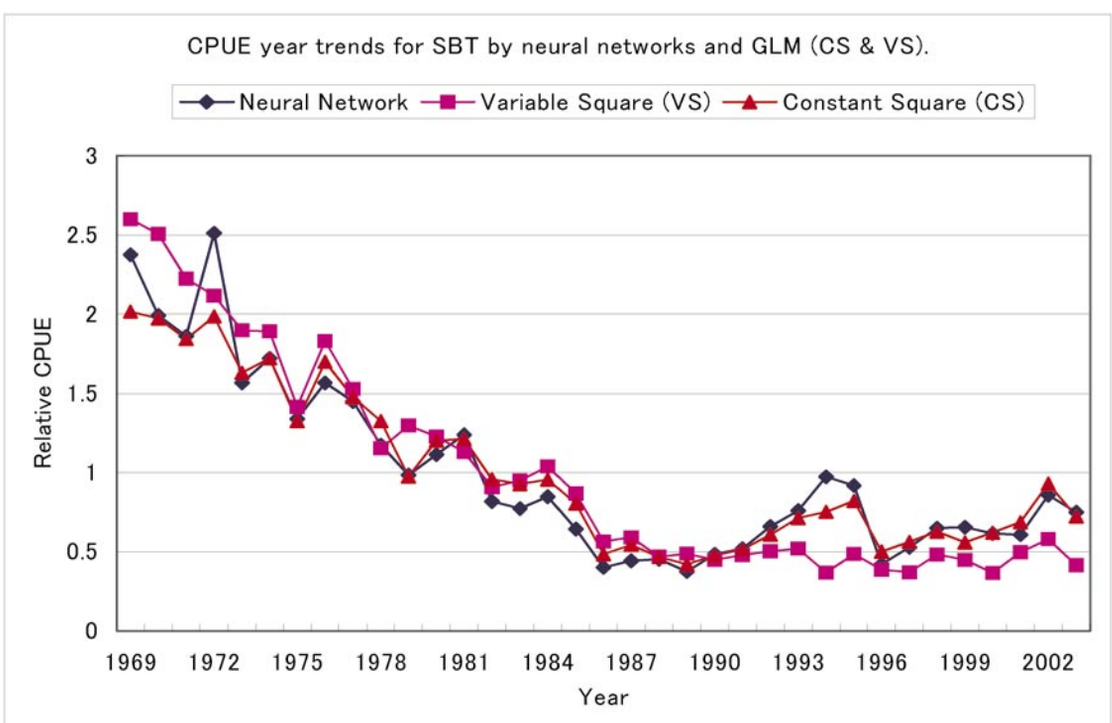

Fig. 4-8. Comparison among the results of attribution analysis (i.e. year trend of CPUE) based on the predicted CPUEs for southern bluefin tuna obtained from the neural networks and ANCOVA model.

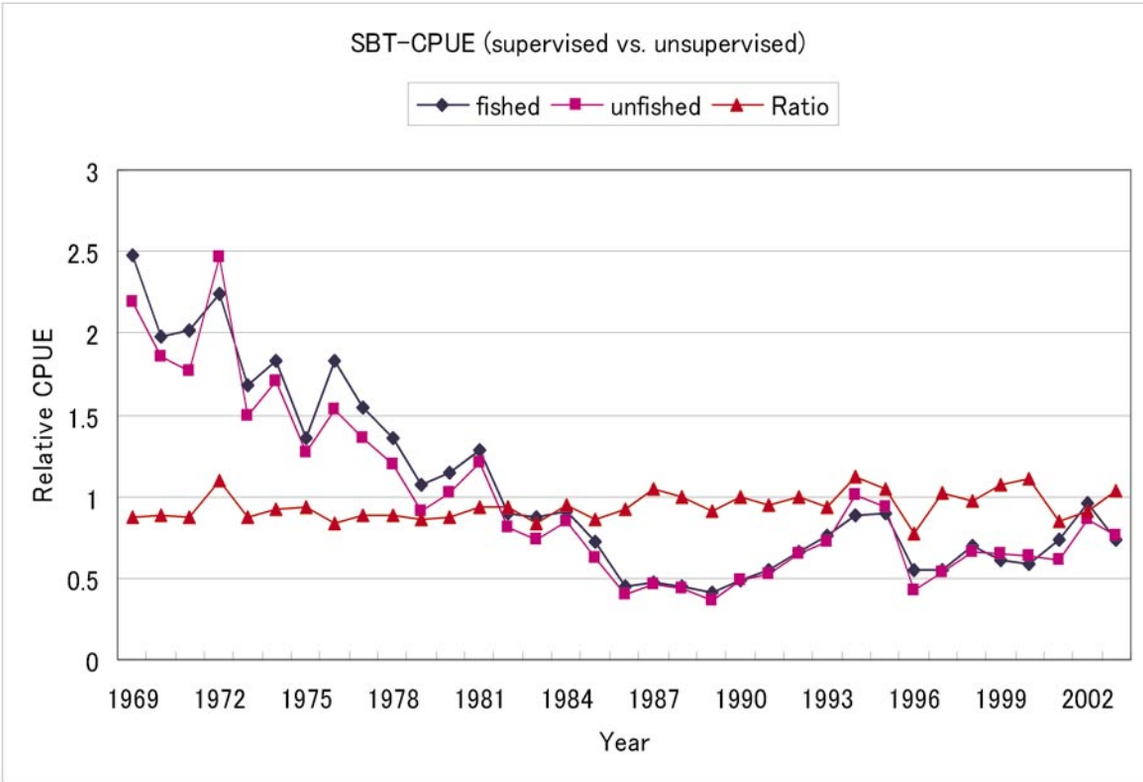

Fig. 4-9. Comparison of CPUE year trends between the supervised (i.e. with observations) and unsupervised (i.e. without observations) data based on the predicted CPUE for southern bluefin tuna.

分の予測分布，すなわち観測値とパラメーターを所与とした欠測値の対数尤度，の和の形で表現出来る）の条件付き期待値を計算するEステップと，Eステップで求めた対数尤度の期待値を最大化してパラメーターを求めるMステップから成り立つ。具体的には，対数尤度関数の変化率が微小になるまで，EステップとMステップを交互に繰り返す。解析では，Jeffreys prior と呼ばれる Fisher 情報行列の行列式の平方根に比例させた分布を事前情報として利用し，多重代入法による5回の試行の平均値を欠測値の補間に使用した。

次に，ニューラルネットワークと MCMC 法により計算された CPUE 予測値，および対応する CPUE の観測値を用いて，残差（= 観測値－予測値）を図示するとそれぞれ，Fig. 4-4および Fig. 4-5のようになる。

そして，ニューラルネットワークにより推定された CPUE 予測値を用いて分割されたサブセットごとに観測値と予測値の絶対誤差の平均値，中央値及び最大最小値を計算すると Table 4-4のようになる。なお，MCMC法により補完された予測値に基づく統計量を，Table 4-5に示す。

注）2つの表では Maximum observed CPUE が 39.13734, Minimum observed CPUE が 0 となり，絶対誤差は全て正解を隠した場合の汎化誤差を表している。

標本相関係数と同様，ニューラルネットワークによる観測値と予測値の絶対誤差の統計値の方が，MCMC 法のそれらよりも小さい値になっている。しかし，ニューラルネットワークから得られた絶対誤差は，5つ全てのサブセットにおいて中央値と平均値の差が大きくなっている。これは分布形が正規分布のように左右対称でなく，対数正規分布のような形であるためと考えられる。

本節の最後に，サブセット毎ではなく，全体として見た場合のニューラルネットワークと EM アルゴリズムを利用した MCMC 法との5-fold cross-validation によるモデル比較結果について述べる。Fig.4-6および Fig. 4-7はそれぞれのモデルにおける観測値と予測値の相関プロットを，Table 4-6および Table 4-7はそれぞれ2つのモデル（ニューラルネットワーク vs. EM algorithm による MCMC 法）における予測値に関する統計量および観測値と予測値の相関係数値を表している。

EM アルゴリズムを利用した MCMC 法は，絶対誤差の中央値と平均値の乖離傾向がさほど見られないが，しかしこれらの値はニューラルネットワークの場合に比べてかなり大きくなっている。この現象からもニューラルネットワークによる予測値の精度は MCMC 法のそれらと比較して良いことが確認された。

しかし，ニューラルネットワークにおいても，Pearson's 相関係数の値は0.6程度があり，必ずしも高いとはいえない。その上，MCMC 法のよる予測ほどではないが，ニューラルネットワークを用いても中心への回帰傾向が依然として認められるため，注意を要する。逆に言えば，EM アルゴリズムによる MCMC 法の予測性能が非常に悪く，推測機能を果たしていないとも考えられる。

4-5. 抽出された CPUE 年トレンド（要因分析結果）

ニューラルネットワークによる予測値をベースにした CPUE 年トレンド抽出は（4.5）式において月・エ

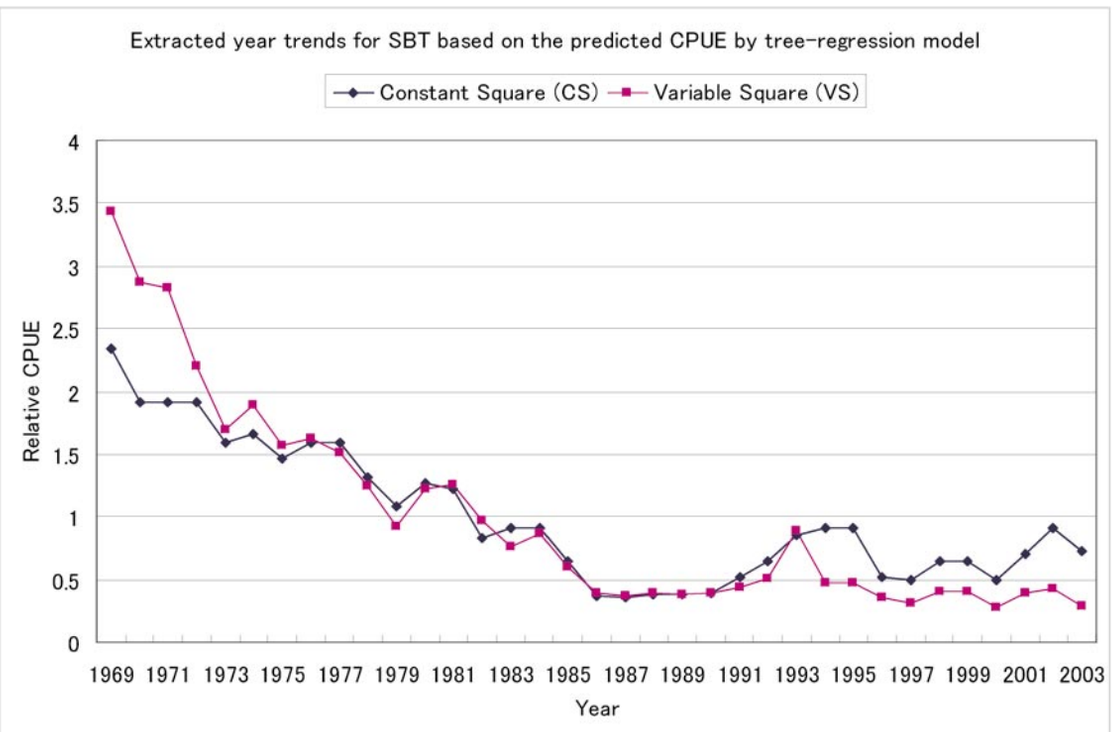

Fig. 4-10. CPUE year trends based on the estimated values by the tree regression model (Constant Square and Variable Square).

Table 4-8. Pearson's correlation coefficient of estimated (i.e. supervised) CPUE by tree-regression model and predicted CPUE by the neural networks

Hypothesis	constant square (CS)	variable square (VS)
Correlation coefficient	0.91747	0.56288

Table 4-9. Median, mean, minimum and maximum values of the absolute error by the neural networks

Hypothesis	constant square (CS)	variable square (VS)
Median	0.512460	0.0991
Mean	0.713396	0.881654
Minimum	0	0
Maximum	8.033459	16.6517

Remark) Maximum supervised CPUE: 16.91, Minimum supervised CPUE: 0

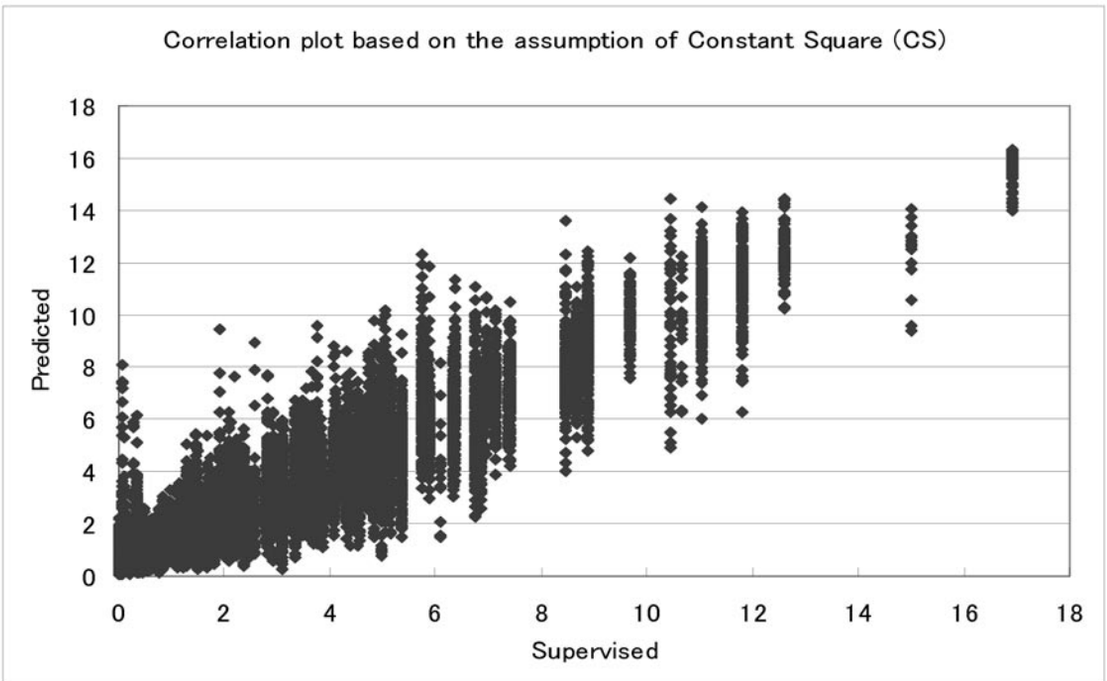

Fig. 4–11. Correlation plots of estimated values by the tree regression model (i.e. supervised data) and the corresponding predicted ones by the neural networks based on the assumption of constant square (CS).

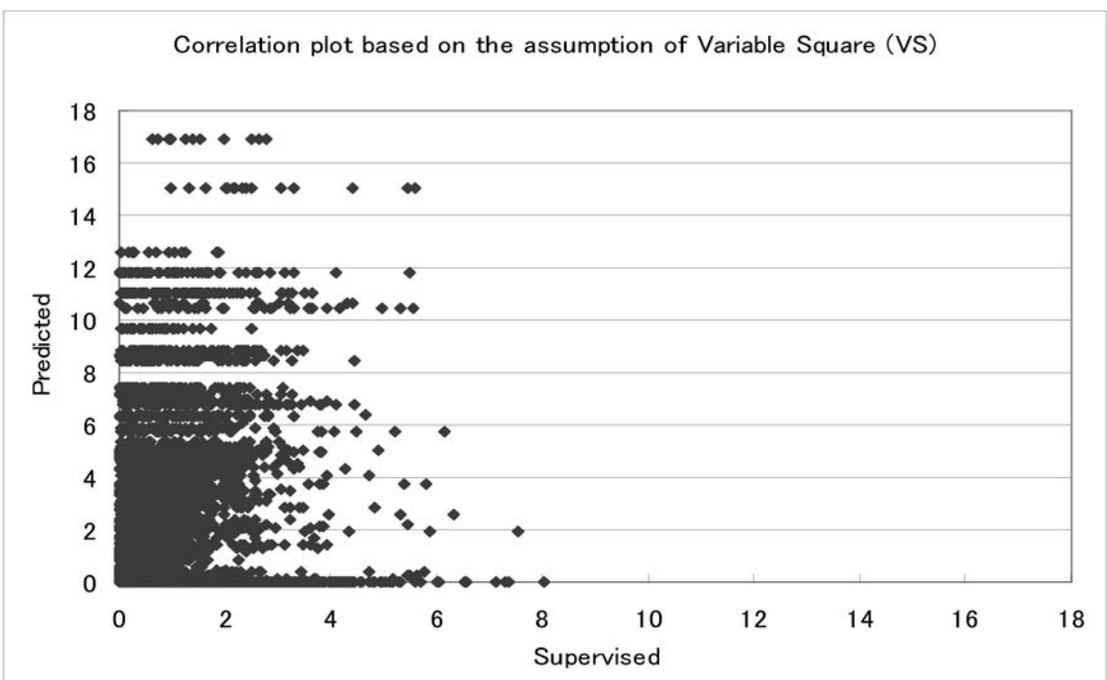

Fig. 4–12. Correlation plots of estimated values by the tree regression model (i.e. supervised data) and the corresponding predicted ones by the neural networks based on the assumption of variable square (VS).

リア（緯度・経度）ともに平均値を使用した。実際には，月やエリアの総和を取る場合も含め 4 通り全て年トレンドを計算したが，その傾向にさほど違いがなかったため，月・エリアともに平均値を使用した。

Fig.4-8は，この方法により抽出されたニューラルネットワークによる CPUE 年トレンドと，共分散分析モデルによるそれらとの比較を示している。その結果，ニューラルネットワークによる CPUE 年トレンドは，共分散分析による CS 仮説（サブエリアで一様に分布すると考えた仮定）に基づく推定値から得られたそれと比較的良く似ており，VS 仮説（操業がないセルの CPUE をゼロと置く仮定）に基づく推定値による CPUE 年トレンドとはかなり異なっている。

なお，CPUE 年トレンドの推定自体は両仮説と関係しないが，年トレンドの抽出に用いるセル毎の CPUE 推定値が CS 仮説と VS 仮説で異なるために，最終的に得られた CPUE 年トレンドも VS 仮説と CS 仮説で異なったものになる。

それを傍証するように，操業がある部分（教師付きデータ）と無い部分（教師無しデータ）に分けた場合の要因分析結果，すなわち抽出された CPUE 年トレンドを比較すると，全体として教師無し部分の CPUE が教師付き部分の CPUE の 7 割ないし 9 割程度の値を示している。Fig. 4-9では，教師付き部分および教師無し部分のそれぞれの CPUE 年トレンドとそれらの CPUE 比（(操業が無いセルの CPUE) / (操業があるセルの CPUE)）の年トレンドを表している。

4-6. 現状に合わせたニューラルネットによる予測値のバリデーション

本節では，CPUE データや設定条件を出来るだけミナミマグロの状況に近い形式にして，ニューラルネットワークによる計算機シミュレーションを行う。解析手法は4-3節での手順を踏襲するが，教師付きデータと教師無しデータの割合を 4 対 1 から現実の7,697対22,123（ほぼ 1 対 3）に変更し，CS 仮説および VS 仮説に基づく CPUE データ（樹形モデルによる推定値）を仮定して，CS 仮説および VS 仮説が再現，そして判別出来るか否かのバリデーションを行った。モデルの比較基準としては樹形モデルによる推定値とニューラルネットワークによる予測値の絶対誤差および相関（Pearson's 相関係数と相関プロット）を使用し，合わせて両者による抽出された CPUE 年トレンドを比較検討した。解析の概略および手順は以下の通りである。

4-3節の解析と同じデータ（教師付きと教師無し合計で29,820件）を用いる。漁業（操業）があるセル（5×5/ 月）のデータ（教師付き：全7,697件）を用いて，樹形モデルの代表的なアルゴリズムである CHAID（Hartigan, 1975）によるルール作成を行い，このルールを使用して操業（漁獲）がないセル（5×5/ 月：教師無し：全22,123件）の CPUE の予測値を推定する。その際に，CS 仮説に基づく場合には操業がないセルに対して樹形モデルによる CPUE 推定値を利用するが，VS 仮説に基づく場合には一律に 0 を対応させる。操業があるセルではすべて対応する樹形モデルによる CPUE 推定値を用いるため，CS 仮説と VS 仮説の教師信号の違いは操業がないセルにのみ表れる。いずれにせよ，本節の計算機実験では，操業がない時空間のみならず操業がある時空間も含めた全29,820件の教師信号を使用する。

次にこの 2 つのデータセット（CS 仮説および VS 仮説に基づくケース）について，それぞれデータ件数が7,697対22,123になるような 2 つのサブセットにランダム分割する。前者を教師付きとみなしてニューラルネットワークの学習に使用し，後者を故意に教師なしと考えて予測し，樹形モデルによる CPUE の推定値と比較する。すなわち，4-3節での観測値および予測値が，本節ではそれぞれ教師データである樹形モデルによる CPUE 推定値（以下，樹形 CPUE 推定値と表記する）およびニューラルネットワークによる CPUE 予測値（以下，ニューラル CPUE 予測値と略す）に対応している。

なお，バリデーションでのデータの 2 分割について，本節で用いたランダム分割の条件として操業位置もランダムに分布している必要があり，特に VS 仮説の場合に問題となりうる。ただし，ミナミマグロ資源に対する努力量としては，極端に偏った集中分布ではなく，（CS 仮説のみならず VS 仮説においても）どちらかと言えばランダムに近い形になっている。そこで，現状とかけ離れた仮定ではないと判断し，今回の計算機実験ではランダム分割を使用した。

この現実の状況にある程度即したニューラルネットワークによるバリデーションにおける教師付きデータとして，樹形モデルによる CPUE 推定値を利用した理由としては，CS 仮説および VS 仮説に対応するデータセットへの変換が容易であること，似たような性質を持つ要因をグルーピングする樹形モデルの特徴が条件（特に分割されたサブエリア内で一様に分布するという CS 仮説の仮定）にマッチしていることに加え，これら（CS 仮説と VS 仮説に対応する樹形 CPUE 推定値）に基づいて抽出された CPUE 年トレンドの違いが顕著に表れていること（Fig. 4-10）が挙げられる。

CS 仮説および VS 仮説を想定した場合の樹形 CPUE 推定値とニューラル CPUE 予測値の Pearson's

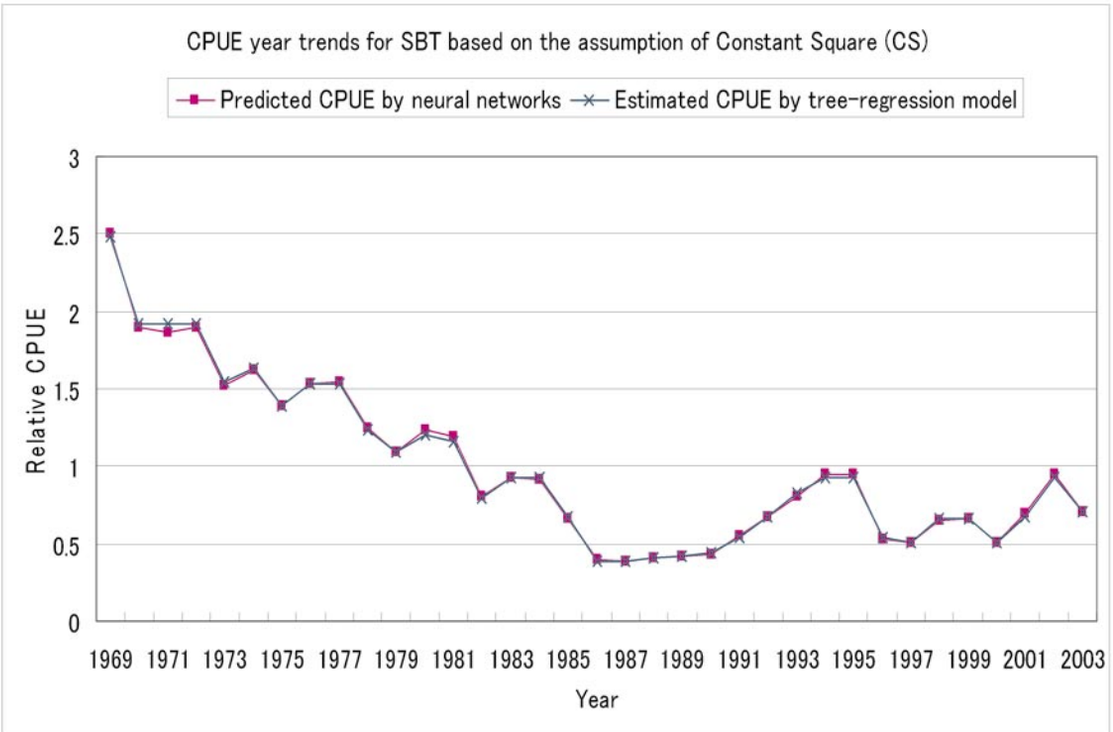

Fig. 4-13. Comparison between CPUE year trends extracted from the estimated values by the tree regression model (i.e. supervised data) and the corresponding predicted ones by the neural networks based on the assumption of constant square (CS).

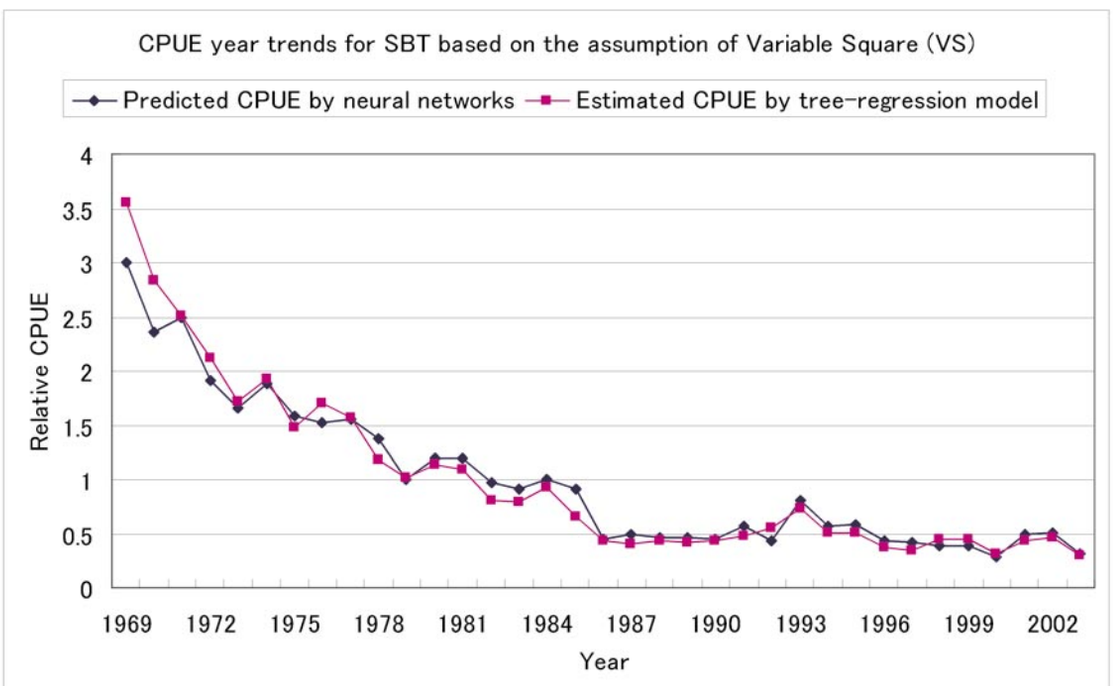

Fig. 4-14. Comparison between CPUE year trends extracted from the estimated values by the tree regression model (i.e. supervised data) and the corresponding predicted ones by the neural networks based on the assumption of variable square (VS).

相関係数の値，絶対誤差の平均値および中央値，最大最小値を Table 4-8と Table 4-9に示す。また，両方の仮説に対応する樹形 CPUE 推定値とニューラル CPUE 予測値の相関プロットは Fig. 4-11および Fig. 4-12のようになる。

これらの図表から判断すると，CS 仮説における数値指標は VS 仮説の場合のそれに比べてかなり良くなっている。このことは相関プロットからも読み取れる。特に CS 仮説に対応する場合の（樹形 CPUE 推定値とニューラル CPUE 予測値の）相関係数は0.91と非常に高くなっており，相関プロットもきれいな形になっている。一方，VS 仮説を想定したケースの相関係数値は CS 仮説のそれよりも低くなっており，教師データよりも大きく予測したデータセットが影響していると考えられる。しかし，絶対誤差の中央値は非常に小さい値になっており，VS 仮説で 0 と設定した操業がないセルのニューラルネットワークによる予測は，CS 仮説ほどでないとはいえ，比較的うまくいっていることが分かる。

次に，（4.5）式の方法により抽出された CPUE 年トレンドを Fig. 4-13（CS 仮説の場合）および Fig. 4-14（VS 仮説の場合）に示す。ここでは，教師付きデータである樹形 CPUE 推定値から抽出された CPUE 年トレンドと，バリデーションにおけるニューラル CPUE 予測値から推定されたそれとの比較を行っている。

その結果，予測精度が高い CS 仮説の場合のみならず，VS 仮説に対応するケースにおいても，樹形 CPUE 推定値とニューラル CPUE 予測値に基づく CPUE 年トレンドは非常に良く似ている。出来るだけ現状に合わせたシミュレーションの結果から判断して，CS 仮説と VS 仮説の再現および判別がある程度は可能であると思われる。

4-7. まとめ

本節では，ミナミマグロ資源への適用を通じたニューラルネットワークによる CPUE 予測と要因分析についてのまとめと今後の課題について，整理する。

第 4 章では，教師付きニューラルネットワークを使用して操業がない時空間のミナミマグロ CPUE を予測し，これらの予測値に基づいて CPUE 年トレンドを推定するための簡便な要因分析手法を提案した。この方法により抽出されたニューラルネットワークによる CPUE 年トレンドと共分散分析によるそれらとの比較を行った結果，ニューラルネットワークによる CPUE 年トレンドは，共分散分析に基づく CS 仮説のそれと比較的良く似ており，VS 仮説の年トレンドとはかなり異なっていた。このことは，ミナミマグロ資源解析上の問題となっていた操業がない時空間の CPUE 解釈について，過去から現在への大幅な漁場の縮小を踏まえ，水産学的な見地から一定の裏付けを与えた研究であると考えられる。すなわち，操業がなくなった時空間においてもミナミマグロ漁業が存在するセルの周辺とほぼ同等の資源密度になっていることを示唆している。

さらに，n-fold cross-validation による予測値の精度評価を全く同じデータセットを用いてニューラルネットワークおよび EM アルゴリズムに基づく MCMC 法を利用して行ったところ，ニューラルネットワークによる予測性能は，MCMC 法によるそれよりもかなり良いことが確かめられた。このことから，今回の解析における，ニューラルネットワークによるミナミマグロの操業がないセルの CPUE 予測精度が，他の方法と比べてかなり高くなっていることが確認された。

なお，実データを利用した現状にマッチした形でのニューラルネットワークによる CPUE 予測値をバリデーションしてみたところ，CS 仮説の方が VS 仮説に比べて精度が高かった。抽出された CPUE 年トレンド等から判断すると，実際上 CS 仮説と VS 仮説の判別および再現がある程度可能であると思われる。

以下，解析上の問題点および今後の課題について，項目ごとに整理する。

・予測に関するクロスバリデーションの方法

今回は，教師付きデータをランダムに 5 分割したサブセットを用いて，標本相関係数と5-fold cross-validation により予測値の精度検証を行ったが，クロスバリデーションの際に用いる選択の指標は，本質的かつ重要な問題である。本研究では，ニューラルネットワークの学習停止規則との整合性を保つため，観測値と予測値の絶対誤差を主に使用したが，平均二乗誤差などの他の指標も試したところ，ニューラルネットワークによる値が MCMC 法によるそれらよりも良くなるという同様の傾向が得られた。なお，今回はデータをランダムに分割したが，系統的な誤差に対処するためには規則的なデータの分割を行う必要があるため，現実の漁業状況に合わせた仮定を今後広く取り入れていきたい。

特に操業位置，すなわち努力量が極端に偏って集中分布している場合（特に VS 仮説で起こりうる可能性がある）には，今回使用した現状に即したバリデーション実験（4-6節）におけるデータのランダム分割の仮定が適切ではないため，今後傾向を持つ場合等も含めたデータ分割の方法を検討していきたい。

・要因分析における連続変量の取り扱い

今回の解析では，全ての説明要因が順序のないカテゴリカル変数であったため，簡便な方法を利用して要因分析を行い，ニューラルネットワークによって得られた予測値をベースにしてCPUEの年トレンドを抽出することが出来た。しかし，連続変量を含む場合には，取り扱いが難しい側面を持つ。カテゴリカル変数に変換するのが近道だが，連続変量をそのままで扱える要因分析手法の開発も検討していきたい。

・要因分析手法の検討

本研究では，緯度・経度が5×5のセルおよび月ごとに集計されたCPUEデータを用いているため，すなわち１つの入力データセット（i.e. 添字の組）に対して教師付き部分の出力が必ず１つ存在するため，要因分析の際に，月および緯度・経度について平均を取った場合と総和を取った場合とで，抽出された年トレンドは非常に良く似ている。しかし，データのバランスが取れていない場合，例えば同一の添字に対する複数出力が存在する場合には，この違いは非常に大きいと考えられる。その際に，データによる重み付けをする場合には総和を取り，セルの重みを等しいと置く場合には平均を取るのが妥当であると考えられ，これらは分散分析等における type-2と type-3の平方和の違いに対応している。今後は，このような計算過程での平均化と総和化の比較検討を行っていきたい。もっとも，１つの入力セット（i.e. 添字の組）に対する複数の出力が存在する場合は，矛盾が生じるゆえにニューラルネットワークによる学習そのものが不適切であるとの考え方も存在するため，適用に際しては細かな注意が必要である。

第５章　Tweedie modelのCPUE解析への応用ゼロ・データの統一的な取り扱い

5-1. はじめに

CPUE標準化において，伝統的に広く利用されてきたCPUE-LogNormalモデル（共分散分析モデル）は，前述の通り，（5.1）式のような定式化がなされる。

$$\mathrm{Log(CPUE)=(Intercept)+(Year)+(Area)+(Season)+(EMT)+\cdots+(Interactions)+(Error),\ Error}\sim N(0,\ \sigma^2) \quad (5.1)$$

注）（EMT）は環境要因や漁船に装備されている装置類の効果などの総称である

このモデルでは応答変数であるCPUEに対して自然対数を取っていることから，Catchがゼロすなわち CPUE（=Catch/Effort）=0となるデータではLog（CPUE）=－∞となってしまうため，そのままでは取り扱うことが出来ない。

このことをCPUE解析におけるゼロ・キャッチの問題と呼んでおり，標準化の計算を行うために，大別して以下のいずれかの方法が使用されることが多い。

1） 応答変数である全てのCPUEに対して一律に微少量（定数項）を足し込む方法（ad hoc な方法）
2） 離散変量であるCatchを応答変数に設定し，Poisson分布や負の二項分布を仮定した，Catch-Poisson あるいは Catch-Negative-Binomial モデルを用いる方法（Catch型モデル）
3） CPUEがゼロか否かを分けた上で，ゼロ・キャッチ率を logitモデルや probitモデルなどで推定し，ゼロでない部分にのみ CPUE-Lognormal モデルや2）の Catch型モデルを適用する方法（Delta型２段階法）（Lo, 1992など）
4） 3）の Delta型２段階法において尤度関数を数珠繋ぎに表現し，パラメーターを同時推定する方法（Zero-Inflated モデル）（Lambert, 1992など）

これまでは，（CPUE=0となるデータが存在する場合には）応答変数であるCPUEに対して一律に微少量を足し込む１）の方法が伝統的に多く用いられてきた（（5.2）式参照）。

$$\mathrm{Log\,(CPUE+constant\_term)=(Intercept)+(Year)+(Area)+(Season)+(EMT)+\cdots+(Interactions)+(Error),\ Error}\sim N(0,\ \sigma^2) \quad (5.2)$$

この方法は，解析者に取って扱いやすい反面，主に区間推定における偏りの原因となる。点推定に関しては，理論的には推定値からこの定数項を差し引くことによって偏りを防ぐことが出来るが，実際上は微少量を加えた状態でパラメーター推定を行っており，やはりバイアスが生じる。また，一定量としてどのような値を取れば良いのかという問題もあり，国際漁業委員会のICCATなどでは平均CPUEの10%が用いられているが（ICCAT, 1997），根拠は不明であり，特に長所も感じられない。

2）の離散変量であるCatchを応答変数にしたPoissonモデルや負の二項分布モデルは，一般化線形モデルの枠組みで取り扱えることもあり，近年多く用いられつつある。当初はCatch-Poissonモデルが使用されていたが，平均と分散が同じという制約が強く，観測されたCPUEデータの現状にマッチ

していないのではないかと感じられるため，徐々にCatch-Negative-Binomialモデルに移行していった経緯がある。水産資源解析分野で広く利用されている統計パッケージSASのGEMMOD procedure（一般化線形モデルのコンポーネント）において，負の二項分布モデルがVersion 8.2からデフォルトで装備されたことも，Catch-Negative-Binomialモデルが普及した一因である，と考えられる。

なお，最近ではCPUEがゼロか否かを分けてゼロ・キャッチ率を（5.3）式のlogitモデルやprobitモデルで推定し，ゼロでない部分にのみ共分散分析モデル（CPUE-LogNormalモデル）などを適用するDelta型2段階モデルが用いられつつある。このモデルでは，最初のゼロ・キャッチ率推定のプロセスと次の非ゼロ部分に対する一般化線形モデルの計算において，統計的に有意な説明要因の効果が異なることも多く見受けられ，その解釈を複雑にする側面がある。

$$\mathrm{E}[\mathrm{Log}\{\mathrm{R}/(1-\mathrm{R})\}]=(\mathrm{Intercept})+(\mathrm{Year})+\cdots+(\mathrm{Interactions})+(\mathrm{Log}(\mathrm{Effort})) \quad (5.3)$$

但しR（ゼロ・キャッチ率）～Binomial（p），すなわち二項分布に従うとする。

また，2つのモデルの尤度関数を1つに書き下して同時推定することも可能であり，Zero-Inflatedモデルと呼ばれている。実際，3）のDelta型2段階モデルと4）のZero-Inflatedモデルの尤度関数は基本的に同じ形をしており，パラメーター推定値も理論的には同じになる。しかし，このモデルはかなり複雑で，定義方法によってはCPUE年トレンドが一意に推定出来ない場合もあるため，水産資源解析において広くは普及していない。

5-2. Tweedieモデルの漁業データへの適用

前節のゼロ・キャッチ問題における現状を鑑み，本報告ではデータを区別することなく統一的に取り扱えるTweedieモデル（Tweedie, 1984）と呼ばれる各々の要素がGamma分布に従い，計数値がPoisson分布に従う確率過程から導かれたモデルに焦点を当てて，詳しく論じる。このモデルは，観測値Yがゼロにmass pointを持ち正の部分で絶対連続となる確率分布であり，式（5.4）のようにして定式化可能である。

$$Y=\begin{cases}\sum_{i=1}^{N}X_i & (N=1,2,3,\ldots)\\ 0 & (N=0)\end{cases} \quad (5.4)$$

但し，$X_1\cdots X_N$は平均μ，分散$\phi\,\mu^2$となるGamma分布に互いに独立に従い，Nは平均λのPoisson分布に従うこととする。（5.4）式での$X_1\cdots X_N$, Nはそれぞれイベントとそれが起こる回数を表し，計数値Nを制御することによりゼロ・データを統一的に取り扱うことが可能になる。なお，後述の5.3節および5.4節の例では，観測されたゼロを多く含むCPUEの値が式（5.4）のYに対応する。

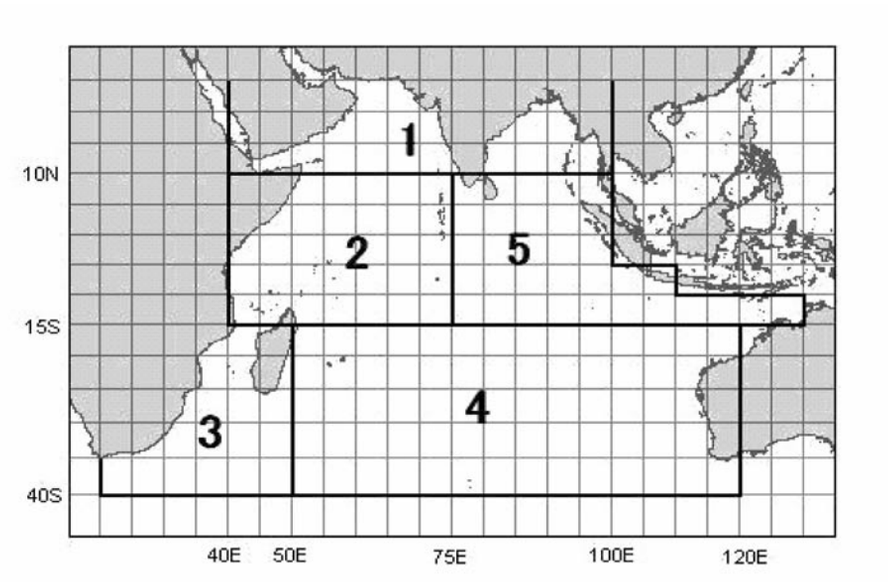

Fig. 5-1. Area stratification used for CPUE standardization of yellowfin tuna in the Indian Ocean caught by the Japanese longline commercial fishery. (Shono, et al., 2005)

このTweedieモデルは適当な変数変換を施すことにより指数型分布属に帰着出来，一般化線形モデルの中での疑似尤度のフレームワークでの論理展開が可能となるため，最尤推定量が標本算術平均で表現される，という意味において，Tweedieモデルによる推定量は漸近的に良い性質を持つことが知られている（Jorgensen, 1997）。そのため，実際に降水量予測などの問題に適用されており，今後は官庁統計分野（消費支出の分野など）への応用も期待される。

しかし，その一方で他の統計モデルとの比較が個々の推定量に限定され，モデル全体としての比較が困難であるという欠点が存在することには，注意を要する。

Tweedieモデルは，通常の疑似尤度のフレームワークにおける平均―分散関係，すなわち後述の式（5.8）で定義された分散関数の冪乗数を連続変数に拡張したモデルとも考えられ，この冪係数が最尤法により推定可能なことが特徴的である。また，回帰係数の推定には疑似尤度の枠組みを使用しているため，AICに代表される情報量規準が，モデル選択に際しては理論的には利用出来ない。変数選択にはDevianceなどに基づくstepwiseカイ二乗検定を用いることが一般的であり，Q-AIC（quasi-AIC）と呼ばれる情報量規準も提案されているが（Burnham and Anderson, 1998），理論的妥当性については議論の余地がある。

本研究では，２つの漁業データの解析例を通じて，Tweedie分布モデルと共分散分析モデルを使用した全てのCPUEに微少量を足し込むad hocな方法，さらに一部はCatchを応答変数にした負の二項分布モデルの比較検討を行った。

また，n-fold cross-validationと呼ばれる，データをランダムにn分割し，出力の一部，すなわちn分割された１つのサブセットにおけるCPUEを故意に隠した予測を全てのサブセットに関して順番に行う方法を通じて，各々のモデルの精度評価を行った。精度評価の指標としては観測値と予測値のPearson's相関係数および平均二乗誤差（mean squared error: MSE）を使用し，相関プロットを用いて標準化残差の傾向を調べた。

なお，5-3節では日本のはえ縄商業船によるインド洋におけるキハダ資源を，5-4節では日本のはえ縄公

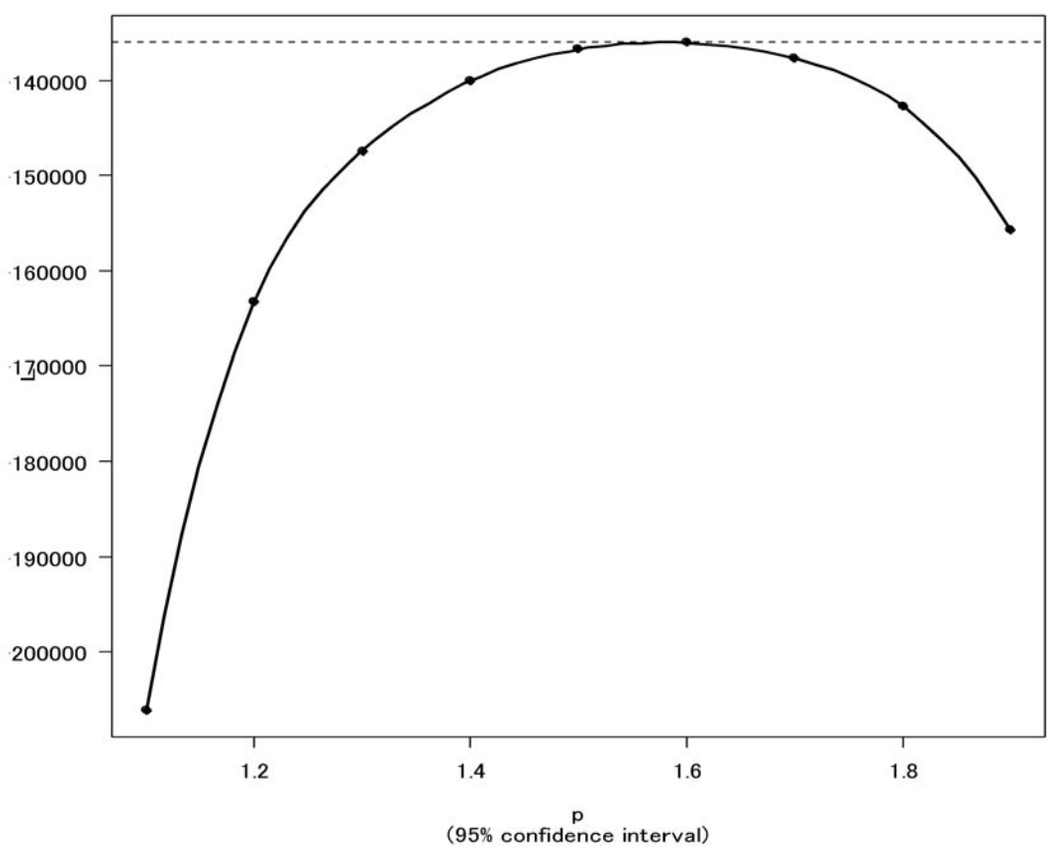

Fig. 5-2. Value of log-likelihood function (L) changing the power-parameter (p) of the Tweedie model for CPUE standardization of yellowfin tuna in the Indian Ocean caught by the Japanese longline commercial vessels.

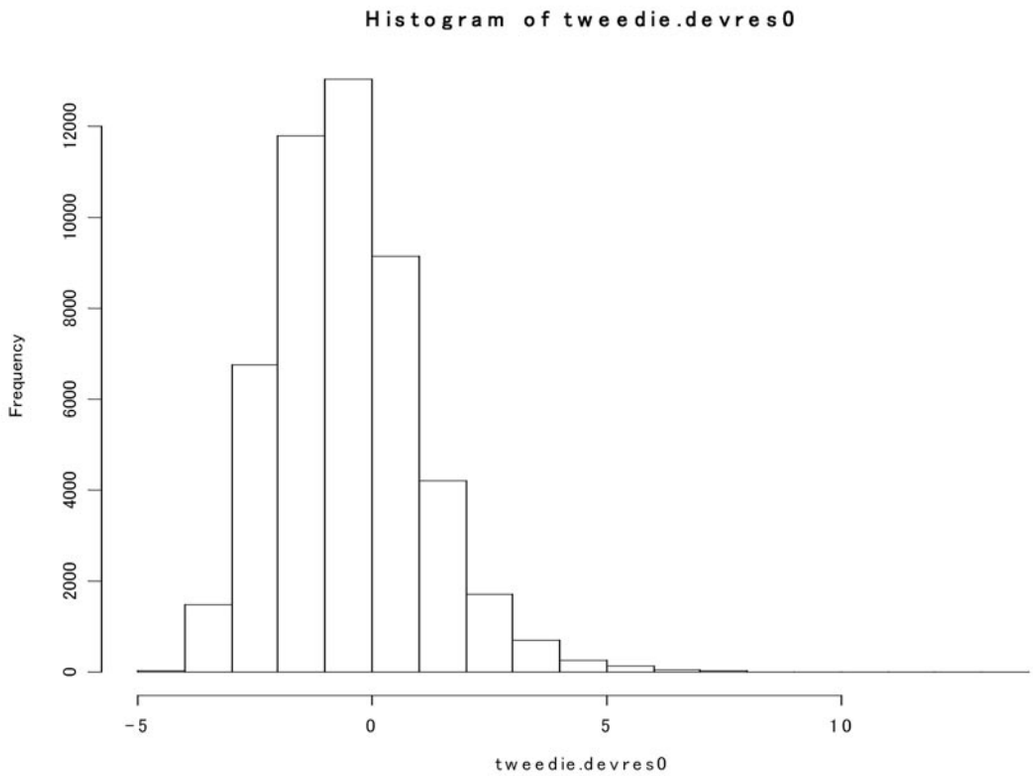

Fig. 5–3. Plots of the standard residual in the Tweedie model for yellowfin tuna.

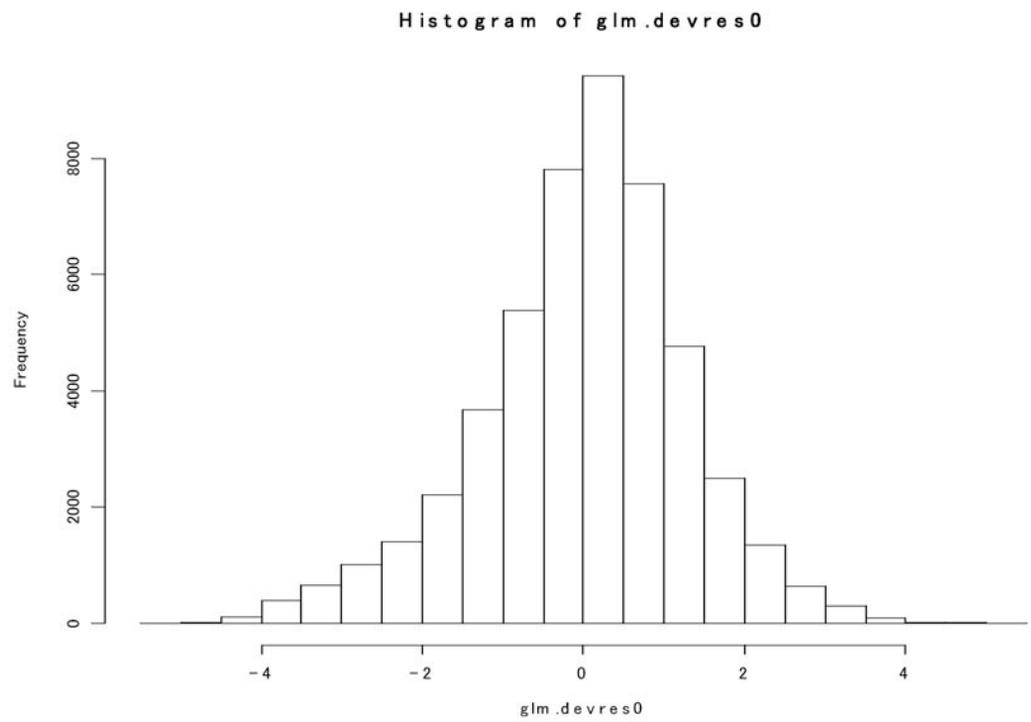

Fig. 5–4. Plots of the standard residual in the ad hoc method for yellowfin tuna.

庁船による北太平洋におけるクロトガリザメ資源を取り上げ，実際の漁業データによるCPUE解析例として，詳細に検証した。なお，5-5節は本章のまとめである。

5-3. 適用例1：日本のはえ縄商業船によるインド洋キハダ資源のCPUE解析

本節では，日本のはえ縄商業船によるインド洋キハダ資源の漁獲量・努力量データを使用したCPUE標準化を行い，CPUE-LogNormalモデル（共分散分析モデル）において，応答変数である全てのCPUEに対して微少量（一定量）を足し込むad hocな方法とTweedie分布モデルの比較を目的とする。緯度・経度が5度刻みで月毎に集計された漁獲データを使用し，解析に用いた応答変数と説明要因は，以下の通りである。なお，ゼロキャッチ・データの占める割合は10%程度と比較的低いが，ターゲット種のデータと

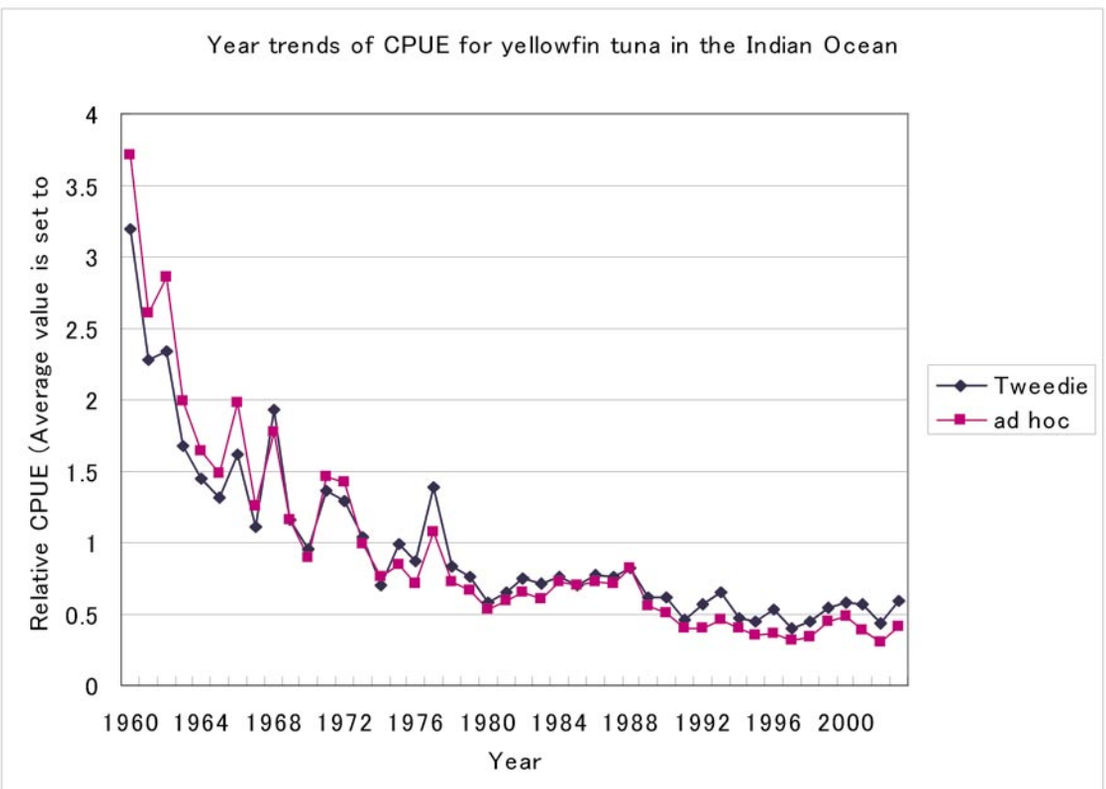

Fig. 5-5. Year trends of standardized CPUE obtained from the Tweedie distribution model and ad hoc method for yellowfin tuna in the Indian Ocean.

Table 5-1. Dataset used for 5-fold cross-validation in the example of yellowfin tuna in the Indian Ocean caught by the Japanese longline commercial fishery

Sub-set	No.of data	Scenarios					
		Base case	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
1	9,871	Rule	C.V.	Rule	Rule	Rule	Rule
2	9,872	Rule	Rule	C.V.	Rule	Rule	Rule
3	9,871	Rule	Rule	Rule	C.V.	Rule	Rule
4	9,872	Rule	Rule	Rule	Rule	C.V.	Rule
5	9,871	Rule	Rule	Rule	Rule	Rule	C.V.

*Rule and C.V. show the sub-dataset for rule making and cross-validation, respectively.

Table 5-2. Model comparison based on the results of 5-fold cross-validation for the example of yellowfin tuna

Candidate model	Pearson's correlation	Mean squared error
Tweedie model	0.493920	3,749,210
Ad hoc method	0.468231	4,222,409

Table 5-3. Pearson's correlation coefficient in each sub-dataset by 5-fold cross-validation for the example of yellowfin tuna

Correlation	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
Tweedie model	0.532	0.473	0.486	0.509	0.486
Ad hoc method	0.508	0.437	0.458	0.498	0.461

Table 5-4. Mean squared error in each sub-dataset by 5-fold cross-validation for the example of yellowfin tuna

MSE	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
Tweedie model	578,187	864,574	798,697	615,102	902,651
Ad hoc method	652,065	972,004	910,298	685,552	1,002,491

しては一般的である。

応答変数―日本のはえ縄商業船によるインド洋におけるキハダ資源の CPUE (=Catch/Effort)Catch：漁獲尾数，Effort：針数（1000本単位）

説明要因―Year ：年（1960-2003）
Month：月（1-12）
Area ：海区（1-5，インド洋全体を5つに区分（Fig. 5-1））
Gear ：枝縄数（number of hooks between float, HBF）
SST ：表面水温(sea surface temperature)
MLD ：混合層深度(mixed layer depth)

注)Year, Month, Area は順序のないカテゴリカル変数，他は連続変数と設定。

最初に，Tweedie モデルの比較対象である ad hoc な方法において，応答変数に一律に足し込む定数項を0.1と設定し，全ての主効果と交互作用を含めたフルモデル（(5.5) 式）から年（Year）の主効果のみを含むヌルモデル（(5.6) 式）まで各々の変数を全ての組合せにわたって取捨選択した候補モデルの中から，情報量規準 BIC (Baysian information criterion: Schwarz, 1978) により選択されたモデルを最終モデルに設定した（(5.7) 式）。

注)CPUE 標準化における主な目的は年トレンドの抽出であることから，ヌルモデルにおいても，年（Year）の主効果は説明要因として含めた。

$$\log(CPUE_{ijkl}+0.1)=Intercept+YEAR_i+MONTH_j+AREA_k+GEAR+SST+MLD+(YEAR*MONTH)_{ij}+(YEAR*AREA)_{ik}+(MONTH*AREA)_{jk}+(YEAR*SST)_i+(YEAR*MLD)_i+(AREA*GEAR)_k+(AREA*SST)_k+(AREA*MLD)_k+(MONTH*GEAR)_j+(MONTH*SST)_j+(MONTH*MLD)_j+(SST*MLD)+ERROR_{ijk},\ ERROR_{ijk}\sim N(0,\sigma^2) \quad (5.5)$$

$$\log(CPUE_{ijkl}+0.1)=Intercept+YEAR_i+ERROR_{ijk},\ ERROR_{ijk}\sim N(0,\sigma^2) \quad (5.6)$$

― Final model ―

log (CPUE+0.1)=Intercept + Year + Month + Area + Gear + SST + MLD + Area*MLD + error,
error ～ N (0, σ^2) (5.7)

次に，この ad hoc な方法での最終モデルと同じ要因効果の組合せを使用して，Tweedie モデルによる推定を行った。最初に分散関数の冪係数を最尤法により推定し，次に回帰係数の各パラメーターを疑似尤度の枠組みを利用して推定した。計算には，フリーの統計パッケージである R (Version 2.2.0) を用いた。

結果については，Tweedie モデルと ad hoc な方法の比較という観点から，Tweedie モデルにおける分

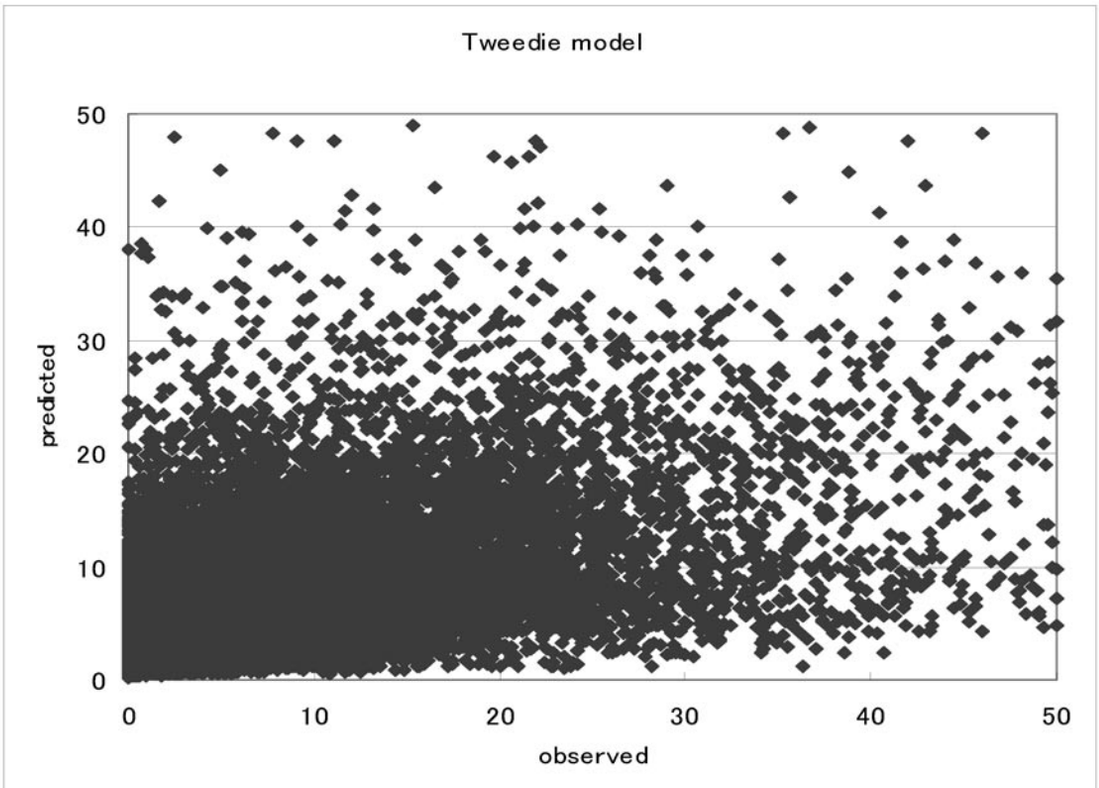

Fig. 5–6. Overall correlation plots of the observed and the predicted CPUE in the Tweedie model for yellowfin tuna.

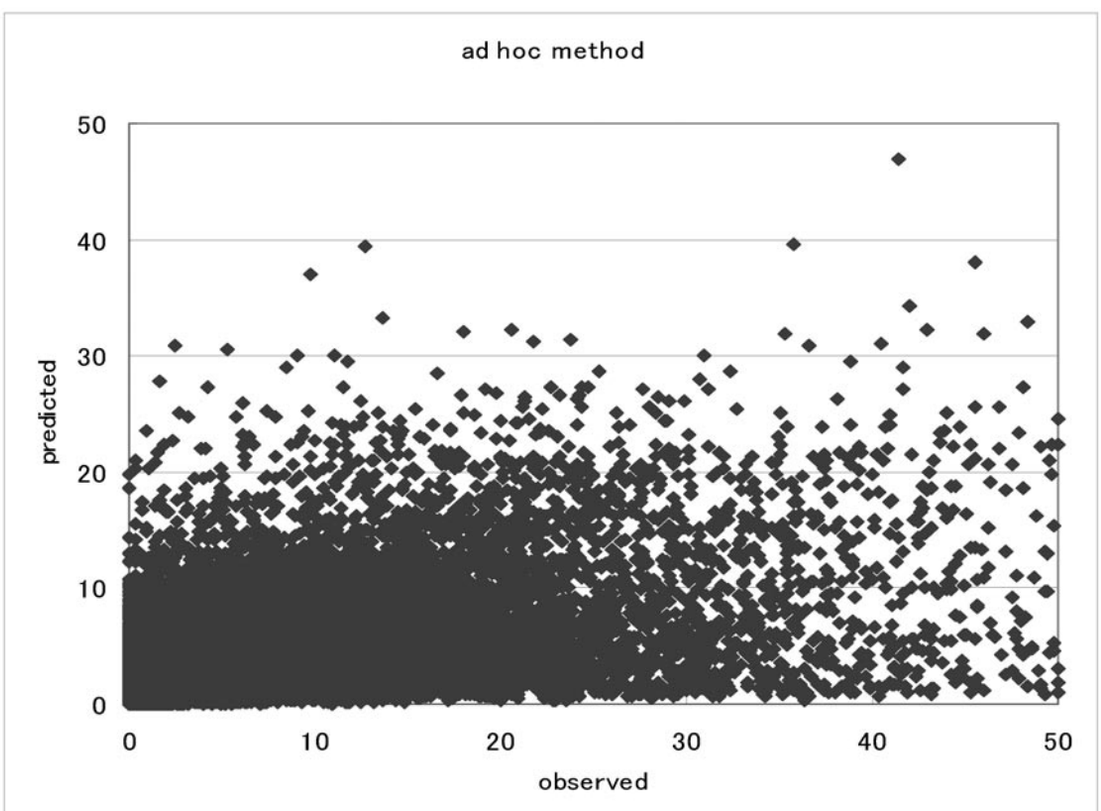

Fig. 5–7. Overall correlation plots of the observed and the predicted CPUE in the ad hoc method for yellowfin tuna

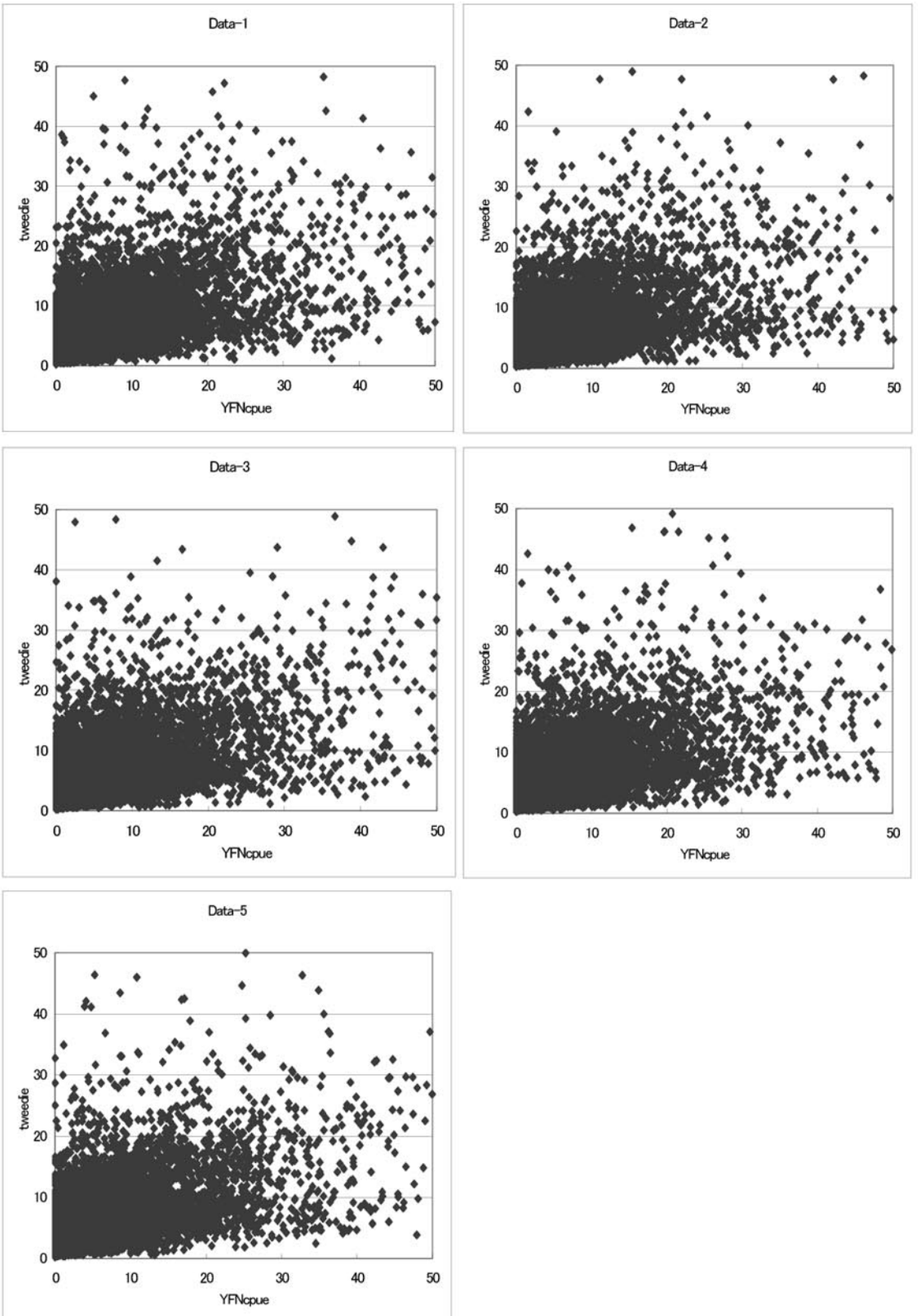

Fig. 5-8. Correlation plots of the observed and the predicted CPUE in the Tweedie model for yellowfin tuna in each sub-dataset used for 5-fold cross-validation.

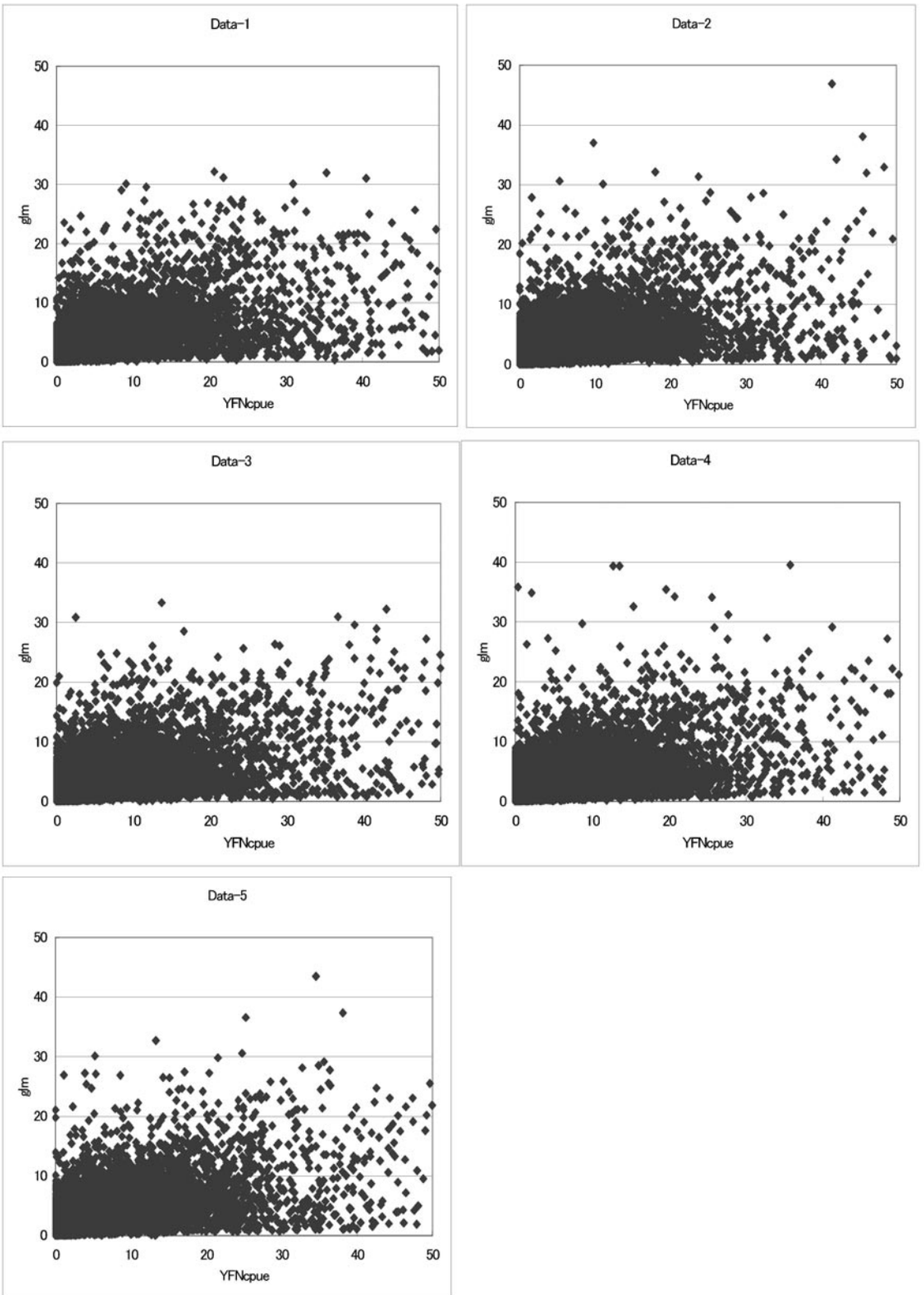

Fig. 5–9. Correlation plots of the observed and the predicted CPUE in the ad hoc method for yellowfin tuna in each sub-dataset used for 5-fold cross-validation.

散関数の冪係数推定や残差分析，２つのモデルにおける LSMEANS による抽出された CPUE 年トレンド，5-fold cross-validation により得られた予測値と観測値の相関プロット，Pearson 相関係数，平均二乗誤差などを利用したモデル比較，の順番に述べる。

まず，(5.4) 式での Tweedie モデルの定式化において適当な変数変換を行い指数型分布族化し，分散関数の冪係数 p を推定する。

$$\mathrm{Var}[\mu] = \mu^{p} \qquad (5.8)$$

注）p が1, 2, 3の時それぞれ Poisson 分布, Gamma 分布，逆正規分布を表す。

この冪係数 p は，0<p<1を除く実数で定義されるが，興味の対象は1<p<2の範囲であるため，この範囲で尤度関数が最大になる p を求めたところ，その値はおよそ1.58と推定された（Fig. 5-2：x 軸が冪係数，y 軸が尤度を表す）。

次に，この分散関数の冪係数の推定値を固定し，疑似尤度の枠組みを使用して回帰係数など他の母数を推定したが，モデル比較の観点から，Tweedie モデルにおいても ad hoc な方法と同じ説明変数のセット（式 (5.7)）を使用した。

さらに，Deviance に基づく標準化残差について，そのプロットを Fig. 5-3と Fig. 5-4に示した。図を見る限りでは，Tweedie モデルと ad hoc な方法とでは特段の差異は認められない。そして，Type Ⅲの平方和に基づく LSMEANS（least squared means）による CPUE 年トレンドを計算したところ，傾向にほとんど違いは見られなかった（Fig. 5-5: 平均 CPUE を1と仮定した相対値を表す）。この理由としては，ゼロ・キャッチの割合が10% 程度と少ないため，ゼロ・データの取り扱いに関する Tweedie モデルの特徴が表れていないことが挙げられる。

最後に，5-fold cross-validation の概要について述べる。ランダムに分割されたデータ・セットは Table 5-1の通りであるが，表中で Base 以外の５つのケース（Ⅰ－Ⅴ）に関して，“Rule”は Tweedie モデルもしくは ad hoc な方法でのパラメーター推定を行うために使用したデータを表し，“C.V.”は正解を故意に隠して cross-validation に使用するための教師無しデータを表している。

まず，全体としてみた場合，すなわち上のⅠからⅤまでの cross-validation の結果を繋ぎ合わせた場合の観測値と予測値の Pearson's 相関係数および平均二乗誤差（MSE: mean squared error）は Table 5-2の通

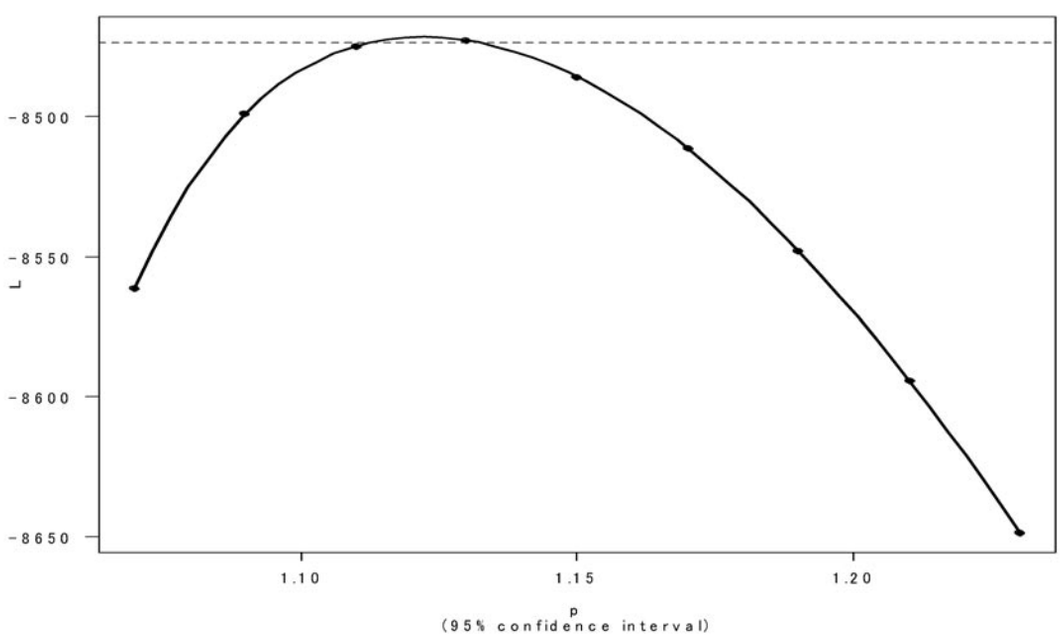

Fig. 5-10. Value of log-likelihood function (L) changing the power-parameter (p) in the Tweedie model for CPUE standardization of silky shark in the North Pacific Ocean caught by the Japanese longline training vessels.

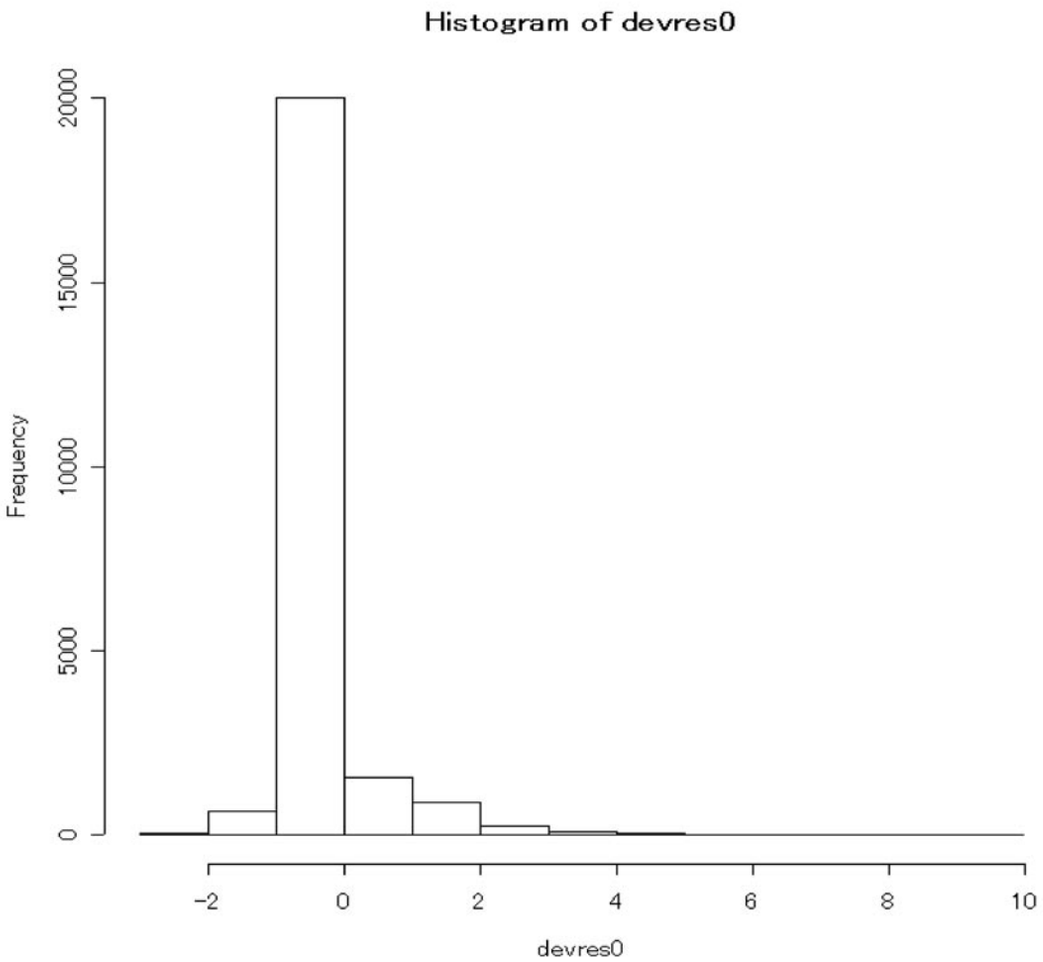

Fig. 5–11. Plots of the standard residual in the Tweedie model for silky shark.

りである。いずれも，Tweedie モデルの方が ad hoc な方法に比べて若干良くなっている。また，５つのケース（Ⅰ－Ⅴ）毎，すなわち正解を隠したサブセット毎の Pearson's 相関係数および平均二乗誤差の値は，Table 5–3および Table 5–4のようになる。

これを見ると，全てのケース（サブ・セット）において，相関係数および MSE のいずれも Tweedie モデルの方が ad hoc な方法に比べて良くなっている。

なお，Tweedie モデルおよび ad hoc な方法での観測値と予測値の相関プロットは，全体としては Fig. 5–6および Fig. 5–7のようになり，それぞれのサブセットにおけるプロットは Fig. 5–8および Fig. 5–9の通りである。

5–4 日本のはえ縄公庁船による北太平洋クロトガリザメ資源の CPUE 解析例

本節では，日本のはえ縄公庁船による北太平洋クロトガリザメ資源の漁獲量・努力量データを使用した CPUE 標準化を行い，Tweedie 分布モデルと Catch-NB（Negative-Binomial: 負の二項分布，離散変数の Catch を応答変数に設定して負の二項誤差を用いた一般化線形モデル），CPUE-LogNormal モデル（共分散分析モデル）において，応答変数である全ての CPUE に対して微少量（一定量）を足し込む ad hoc な方法との比較を目的とする。Shot-by-shot と呼ばれる操業毎の漁獲データを使用し，解析に用いた応答変数と説明要因は次の通りである。なお，ターゲット種でない混獲種であるため，ゼロキャッチ・データの占める割合は80% 以上と非常に高い。また，公庁船は主に水産高校の実習船であり，一般に商業船よりもデ

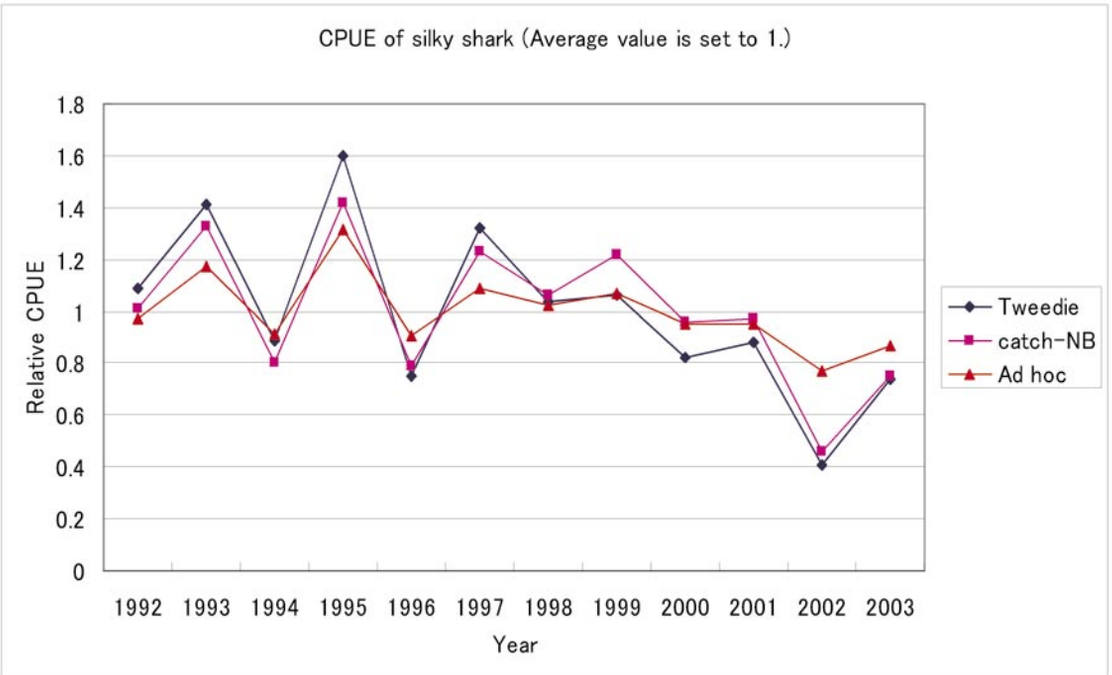

Fig. 5-12. Year trends of CPUE obtained from the Tweedie model, ad hoc method and Catch-NB model for silky shark in the North Pacific Ocean.

Table 5-5. Dataset used for 5-fold cross-validation in the example of silky shark in the North Pacific Ocean caught by the Japanese longline training fishery

Sub-set	No.of data	Scenarios					
		Base case	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
1	4,688	Rule	C.V.	Rule	Rule	Rule	Rule
2	4,687	Rule	Rule	C.V.	Rule	Rule	Rule
3	4,688	Rule	Rule	Rule	C.V.	Rule	Rule
4	4,687	Rule	Rule	Rule	Rule	C.V.	Rule
5	4,688	Rule	Rule	Rule	Rule	Rule	C.V.

*Rule and C.V. show the sub-dataset for rulemaking and cross-validation, respectively.

Table 5-6. Model comparison based on the results of 5-fold cross-validation for the example of silky shark

Candidate model	Pearson's correlation	Mean squared error
Tweedie model	0.502957	6761.768
Catch-NB model	0.450111	11432.45
Ad hoc method	0.446779	8814.842

Table 5-7. Pearson's correlation coefficient in each sub dataset by 5-fold cross-validation for the example of silky shark

Correlation	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
Tweedie model	0.513	0.464	0.558	0.504	0.497
Catch-NB model	0.472	0.431	0.486	0.453	0.427
Ad hoc method	0.456	0.422	0.47	0.46	0.443

Table 5-8. Mean squared error in each sub data-set by 5-fold cross-validation for the example of silky shark

MSE	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
Tweedie model	1,277	1,493	983	1,166	1,842
Catch-NB model	2,035	2,465	1,911	2,259	2,762
Ad hoc method	1,688	1,846	1,375	1,517	2,389

ータの精度が高いことが特徴的である。

応答変数―日本のはえ縄公庁船による北太平洋クロトガリザメ資源の CPUE (=Catch/Effort)Catch：漁獲尾数, Effort：針数（1,000本単位）

説明要因― Year：年（1992-2003）

Quarter：四半期（1：Jan.-Mar., 2:Apr.-Jun., 3：Jul.-Sep., 4： Oct.-Dec.）

Area：海区（1-4, 北太平洋全体（0° N～40° N）を緯度で４つに区分
1: 赤道以北で20度未満，2: 20度以北で30度未満
3: 30度以北で40度未満，4: 40度以北で50度未満）

Gear： 枝 縄 数（number of hooks between float, HBF）

注)Year, Quarter, Area は順序のないクラス変数, Gear は連続変量と仮定

今回の解析では，上記の要因のうちモデルに組み込める主効果および交互作用を全て含めたフルモデルから（Year）の主効果のみ含むヌルモデルまでの候補の中から，情報量規準 BIC によりモデル選択を行ったところ，ad hoc な方法（共分散分析モデル）と Catch-NB モデルで同じ説明変数の組み合わせが選択されたことから，この要因効果のセットを Tweedie モデルでも使用し，３つのモデルを比較検討した。具体的には, 次の式（5.9）及び式（5.10）で表現される。

Ad hoc method

$$\mathrm{Log}(\mathrm{CPUE}+0.01) = \mathrm{Intercept} + \mathrm{Year} + \mathrm{Area} + \mathrm{Quarter} + \mathrm{Gear} + \mathrm{Area}^{*}\mathrm{Gear} + \mathrm{error},\ \mathrm{error} \sim N(0,\ \sigma^2) \quad (5.9)$$

Catch-NB model

$$E[\mathrm{Catch}] = \mathrm{Effort}^{*}\exp(\mathrm{Intercept} + \mathrm{Year} + \mathrm{Area} + \mathrm{Quarter} + \mathrm{Gear} + \mathrm{Area}^{*}\mathrm{Gear},\ \mathrm{Catch} \sim \mathrm{NB}(\alpha, \beta)) \quad (5.10)$$

注）式（5.9）の右辺の Effort はデータから与えられるため，offset と仮定した。

具体的な計算手順，すなわち n-fold cross-validation の手順および使用した評価指標（相関係数と平均二乗誤差）は，モデルが３つ（ad hoc な方法，Catch-NB モデル, および Tweedie モデル）に増えたこと以外，基本的には5.3節のインド洋キハダ資源の例と同様であり，前節に倣って Tweedie モデルにおける分散関数の冪係数推定や残差分析，３つのモデル候補における LSMEANS による抽出された CPUE 年トレンド，5-fold cross-validation により得られた予測値と観測値の相関プロット，Pearson 相関係数，MSE（平均二乗誤差）などを利用したモデル比較，の順に述べる。なお，解析にはフリーパッケージ R（Version 2.2.0）に加え，SAS（Version 9.1.3）を用いた（主に Catch-NB モデルの計算）。

分散関数の冪係数推定値は約1.12となり，Poisson 分布に近い形状を示す（Fig. 5-10）。また，Tweedie

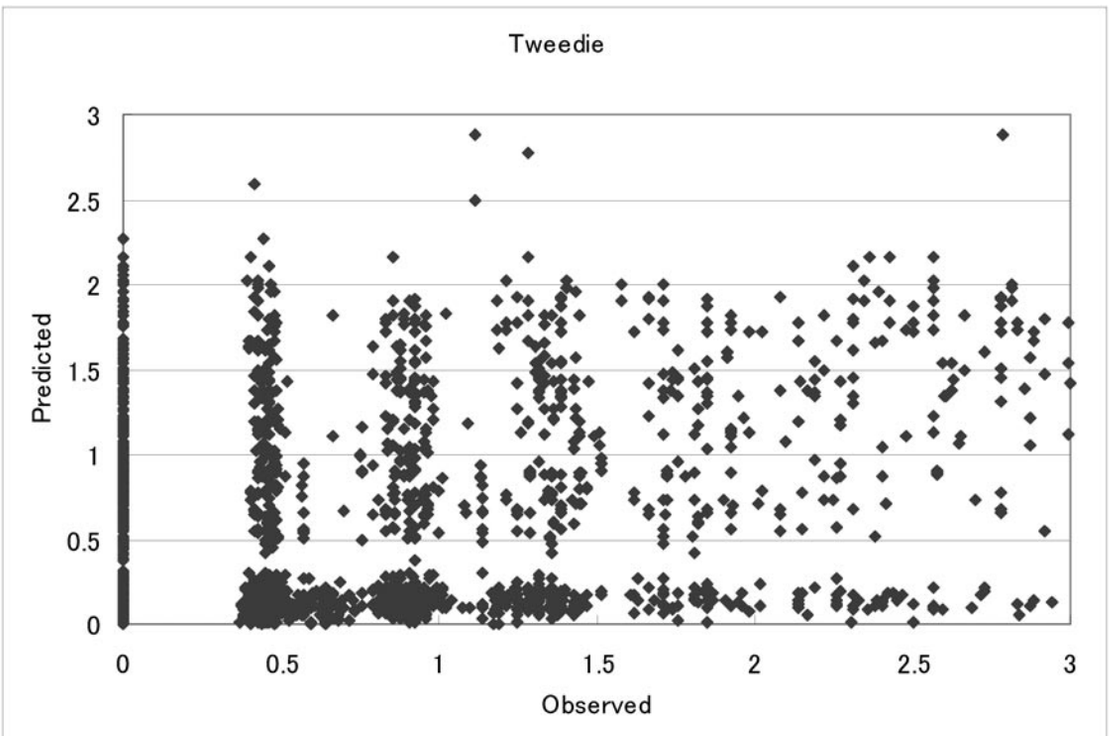

Fig. 5-13. Overall correlation plots of observed and predicted CPUE in the Tweedie model for silky shark.

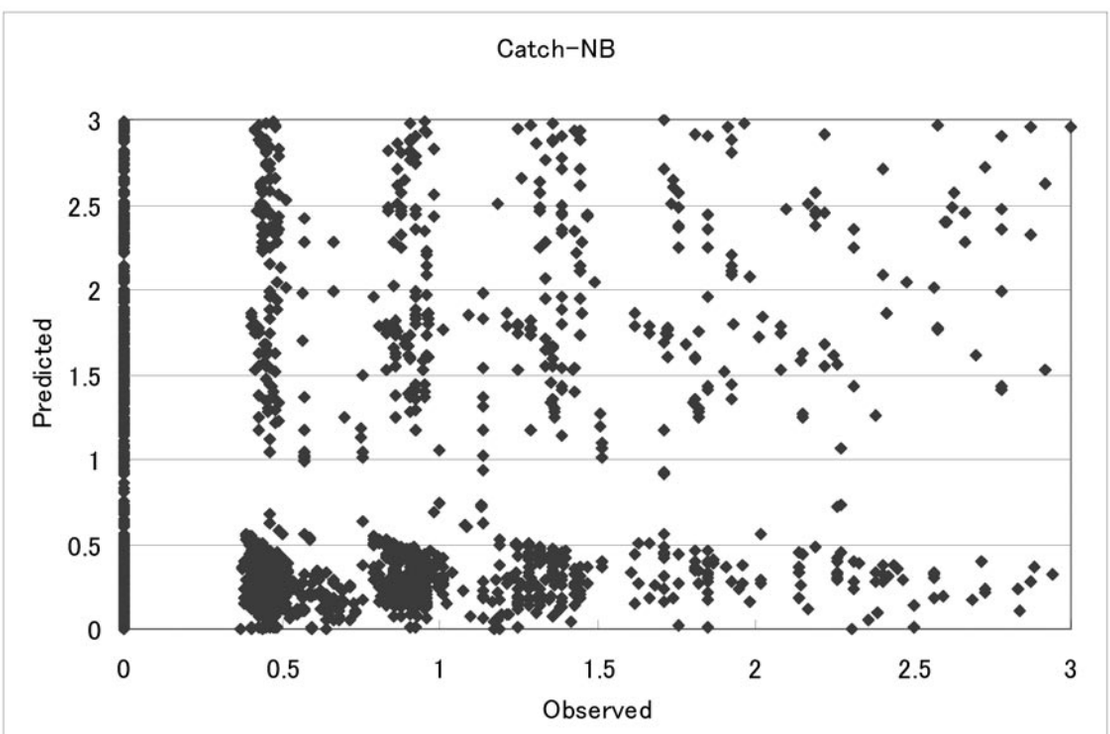

Fig. 5-14. Overall correlation plots of observed and predicted CPUE in the Catch-NB model for silky shark.

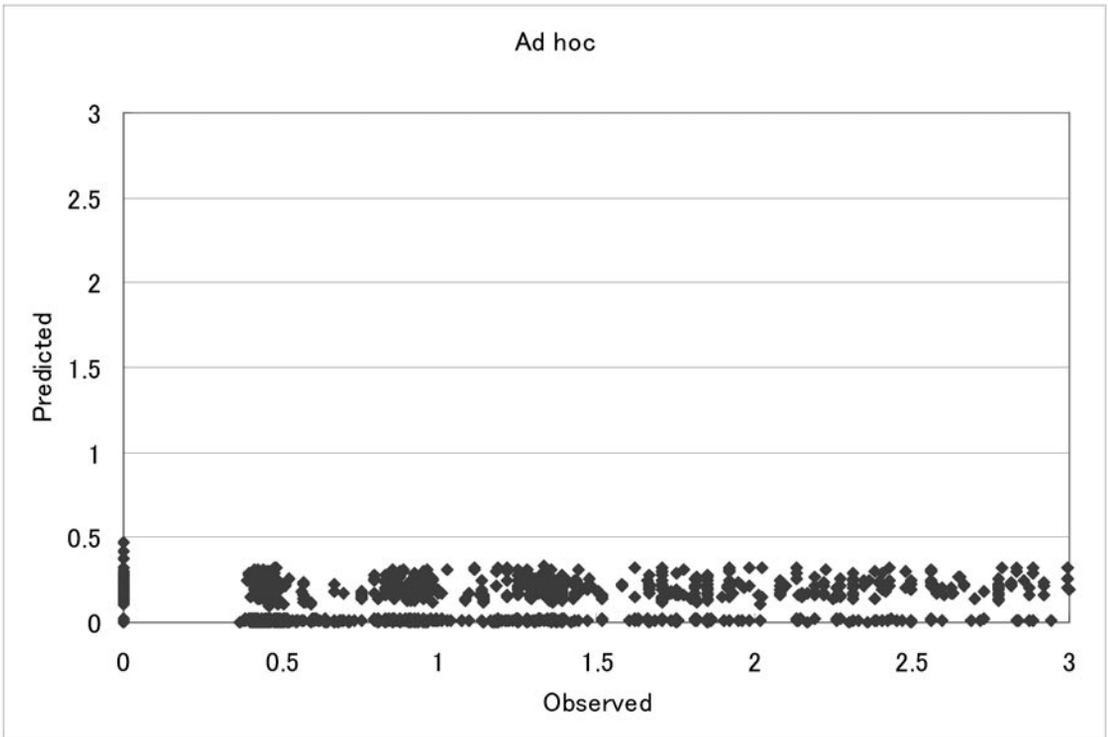

Fig. 5–15. Overall correlation plots of observed and predicted CPUE in the ad hoc method for silky shark.

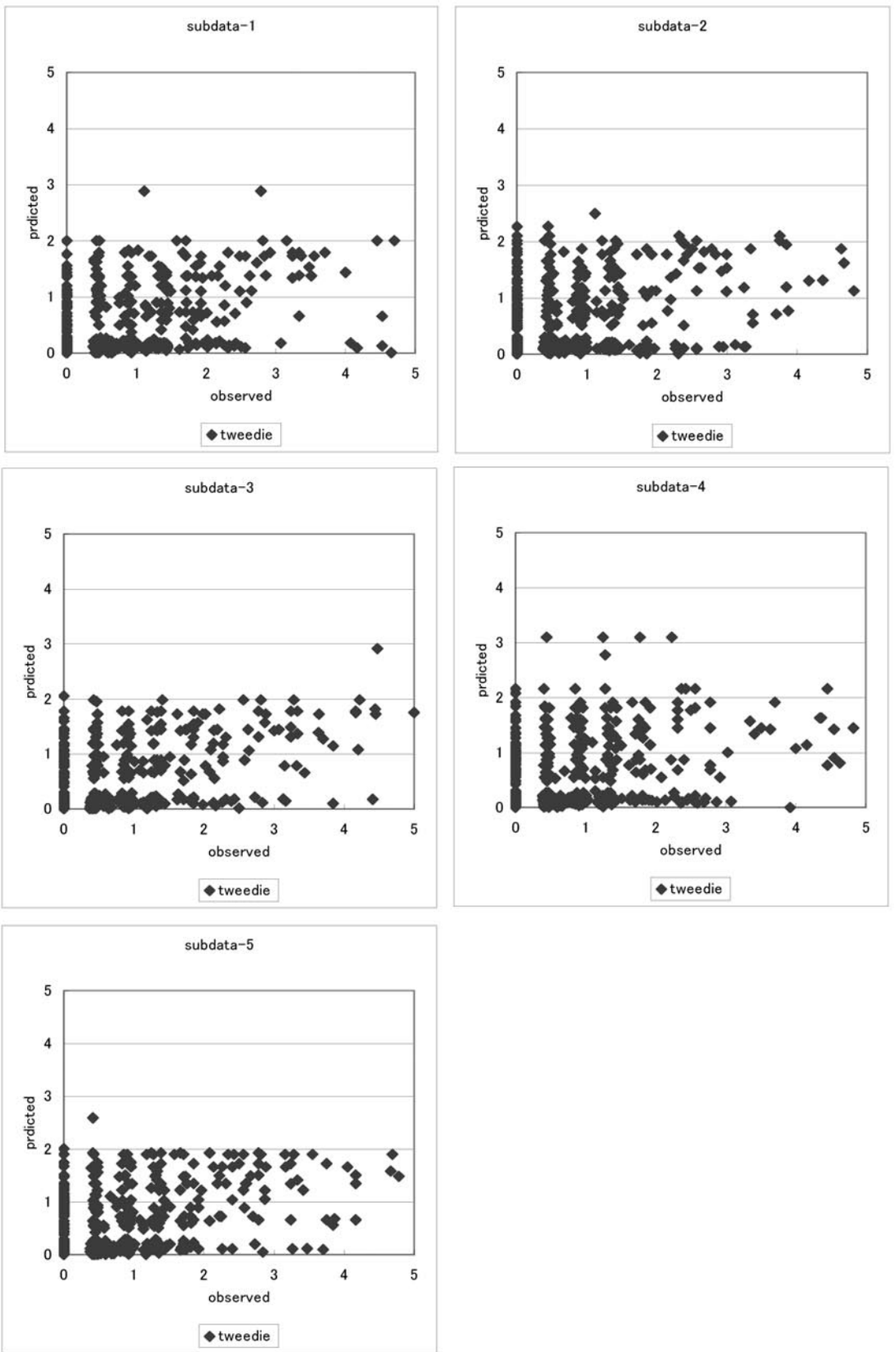

Fig. 5–16. Correlation plots of the observed and the predicted CPUE in the Tweedie model for silky shark in each sub-dataset used for 5-fold cross-validation.

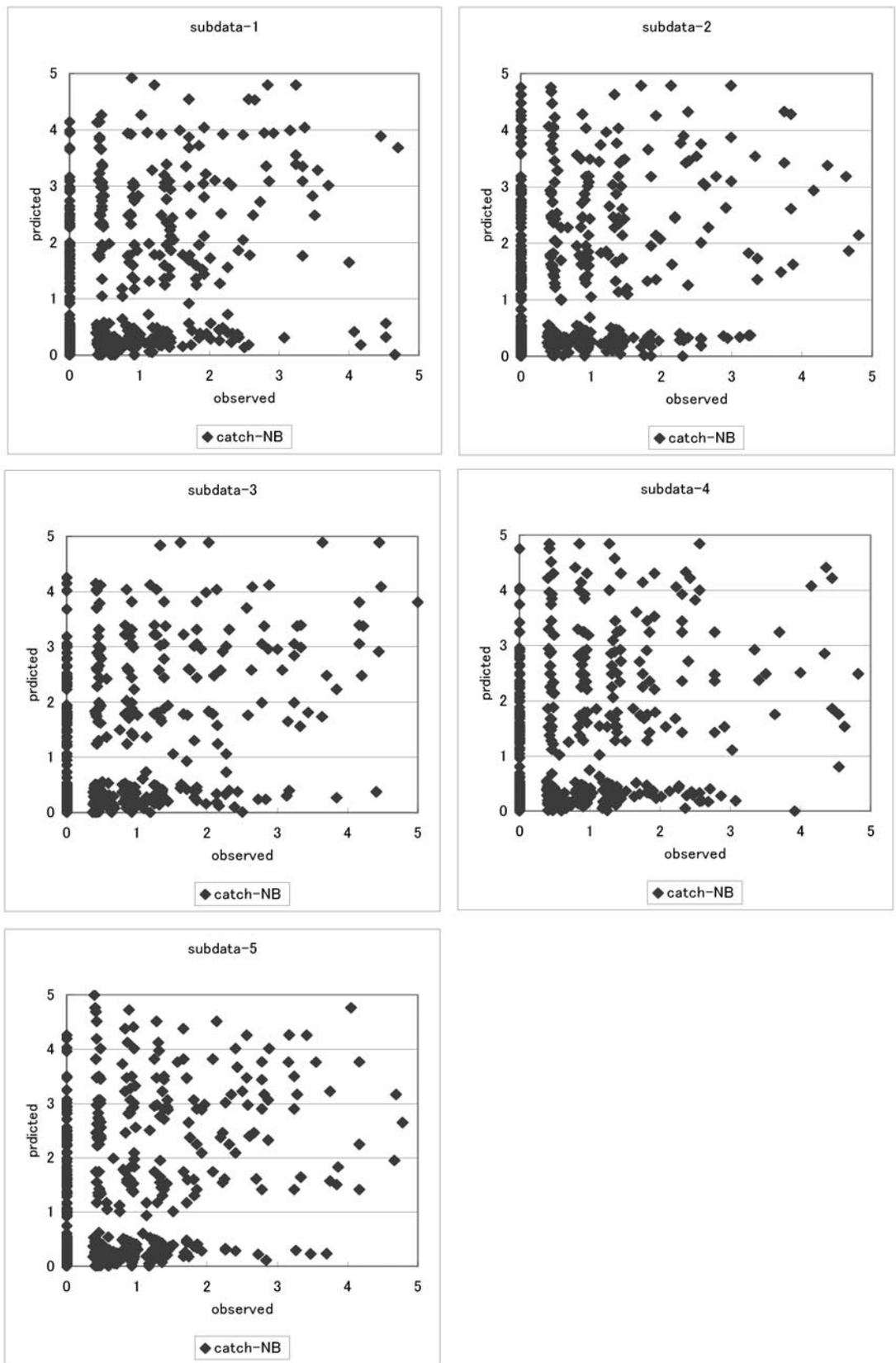

Fig. 5–17. Correlation plots of the observed and the predicted CPUE in the ad hoc method for silky shark in each sub-dataset used for 5-fold cross-validation.

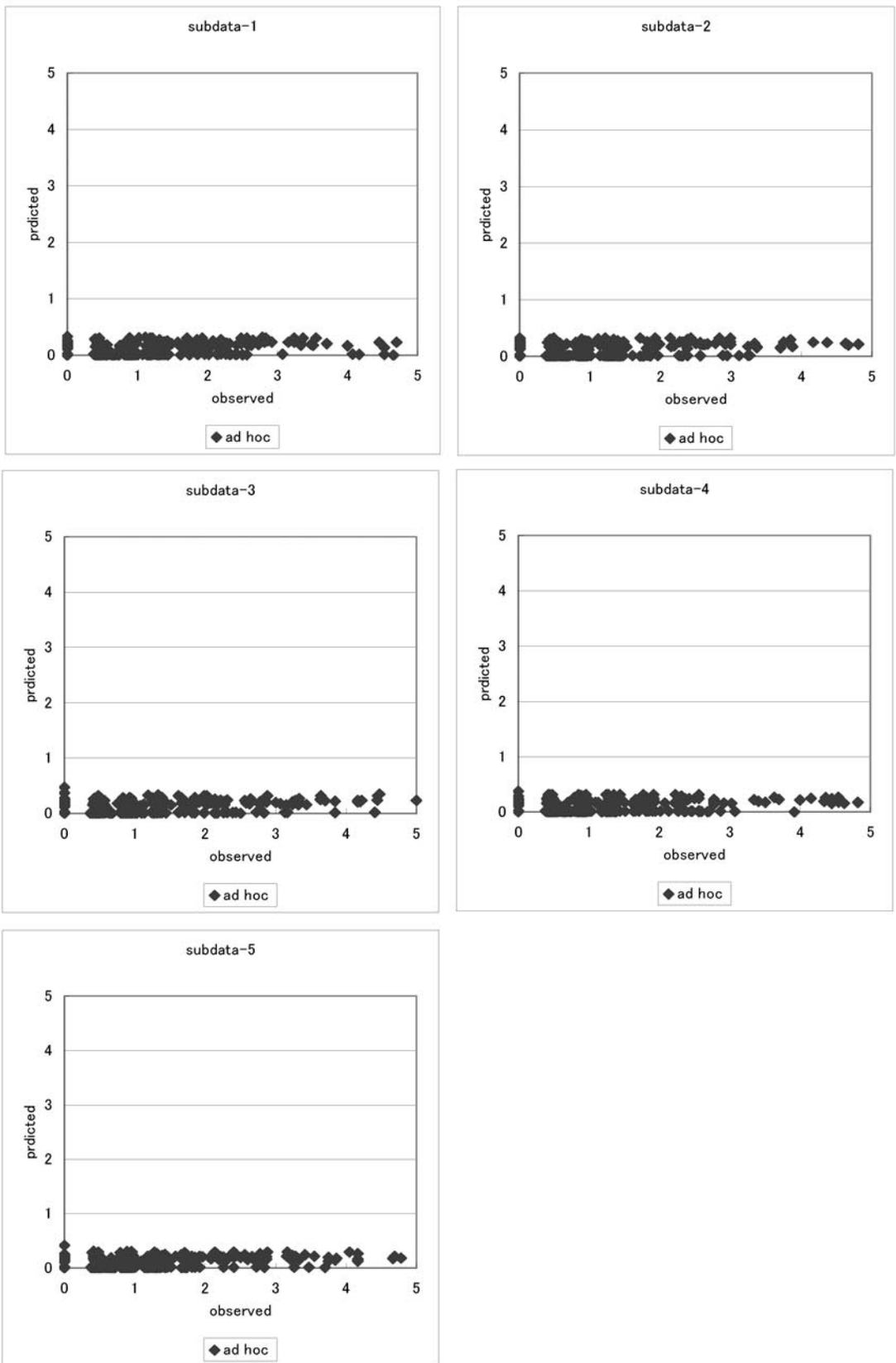

Fig. 5–18. Correlation plots of the observed and the predicted CPUE in the ad hoc method for silky shark in each sub-dataset used for 5-fold cross-validation.

モデルにおける標準化されたDeviance残差はFig. 5-11のようになる。これを見る限り，わずかではあるがマイナスの方向に偏っており，これはゼロ・キャッチをゼロでなく正の微少量と予測してしまうケースが大半のために起こった現象であると考えられる。

そして，Type Ⅲの平方和に基づくLSMEANS（least squared means）によるCPUE年トレンドを計算したところ，Fig. 5-12のようになった（平均CPUEを1と仮定した相対値を表す）。3つのモデル（Tweedieモデル，ad hocな方法，Catch-NBモデル）におけるCPUE年トレンドには明らかな違いが認められる。特に，全ての応答変数に微少量を足し込むad hocな方法では全体的にフラットな感じであり，CPUEが減少傾向にあるTweedieモデルやCatch-NBモデル（この2つのモデルの年トレンドは良く似ている）と多少異なっている。

さらに，5-fold cross-validationの概要について述べる。前節と同様，ランダムに分割されたデータ・セットはTable 5-5の通りであるが，表中でBase以外の5つのケース（Ⅰ－Ⅴ）に関して“Rule”はTweedieモデルもしくはad hocな方法でのパラメーター推定を行うために用いたデータを表し，“C.V.”は正解を故意に隠してcross-validationに使用するための教師無しデータを表している。

全体としてみた場合，すなわち上のⅠからⅤまでのcross-validationの結果を繋ぎ合わせた場合の観測値と予測値のPearson's相関係数および平均二乗誤差（MSE: mean squared error）はTable 5-6の通りである。また，5つのケース（Ⅰ－Ⅴ）毎，すなわち正解を隠した5-fold cross-validationにおけるサブセット毎の相関係数およびMSEの値は，Table 5-7およびTable 5-8のようになる。いずれも，Tweedieモデルの方が他の方法に比べて良くなっており，相関係数は以下Catch-NBモデル，ad hocな方法の順（サブセット毎に見た場合にはCatch-NBモデルとad hocな方法との逆転有り），MSEは以下ad hocな方法，Catch-NBモデルの順になっている。なお，3つの候補モデルにおける観測値と予測値の相関プロットは，全体としてはFig. 5-13― Fig. 5-15のようになり，それぞれのサブセット毎のプロットはFig. 5-16― Fig. 5-18のようになる。

これらの相関プロットおよび相関係数ないしMSEの表から判断すると，Tweedieモデルは全体的に多少予測値が観測値より小さくなる傾向があるが，相関係数が高くMSEの値も小さく，3つの中では一番バランスが取れている。

Catch-NBモデルは，残差（観測値―予測値）の符号の偏りはあまり見られないが，観測値と予測値の差が比較的大きい。その結果MSEの値が非常に大きくなっており，相関係数の値はad hocな方法に比べると少し高くなっている。

Ad hocな方法では，観測値の大きさにかかわらずほとんど全てのCPUEを0.5以下と推定しており，大きなバイアスを持つ。すなわち，ゼロ・キャッチの割合が多い場合に，実用上はほとんど使用出来ないと考えるべきであろう。

MSEの値を見るとCatch-NBモデルより小さくなっているが，これは全体の85%近くを占めるゼロとなる観測データのほとんどを小さな正の値に予測しているゆえの現象であり，モデルの性能はCatch-NBと比較しても極端に悪い。

5-5. まとめ

第5章では，ゼロ・キャッチ問題と呼ばれる漁獲がゼロであるデータが含まれる場合に，CPUEの自然対数を取ったものを応答変数とする共分散分析モデルが使用出来ない問題について，詳細に議論した。Tweedie分布と呼ばれる，確率過程の考え方を利用した，ゼロ・データを統一的に取り扱えるモデルを解析に使用したが，まぐろ漁業に関係する2つの実際例を通してTweedieモデルとゼロ・キャッチ問題に対して利用されている既存の回避策との比較を行った。n-fold cross-validationと呼ばれるデータをランダムにn-分割し，そのうちの一つのサブセットを順番にわざと隠して予測する方法により精度評価を行ったところ，評価の指標としては観測値と5-fold cross-validationから得られた予測値のPearson's相関係数およびMSE（mean squared error: 平均二乗誤差）を用いたが，いずれの例においても両方の指標についてTweedieモデルのパフォーマンスが良くなることが示された。なお，以下に各節毎の内容を要約する。

5-1節では，ゼロ・キャッチ問題について記述し，現状で使用されている回避策である全ての応答変数に微小量（定数項）を足し込んだ上で共分散分析モデル（CPUE-LogNormalモデル）を利用するad hocな方法，離散変量であるCatchを応答変数に設定してPoisson誤差あるいは負の二項（NB, negative binomial）誤差を用いるCatchモデル（Catch-Poisson or catch-NB），ゼロ・キャッチ率を二項分布によるlogitモデルや正規分布によるprobitモデル等を用いて推定し，非ゼロ部分にCPUE-LogNormalモデルもしくはCatchモデルを適用するDelta型2段階モデル，2段階モデルの2つの尤度関数を数珠繋ぎにしてパラメーターを同時推定するZero-Inflatedモデルについて紹介し，そ

れぞれのモデルの適用の現状，特徴，長所と短所などについて記述した。

5-2節では，ゼロ・データが統一的に取り扱い可能な，複合 Poisson 分布の概念の拡張でもある Tweedie 分布について，モデルの特徴や適用上の注意，モデルの持つ長所や短所，パラメーター推定の手順などについて簡潔に説明した。

5-3節では，この Tweedie モデルを日本のはえ縄商業船によるインド洋におけるキハダ資源の CPUE 標準化（漁獲量・努力量データ解析）に適用し，CPUE の年トレンド推定（要因分析）を行うとともに，ad hoc な方法（共分散分析モデル）との比較を行った。ゼロ・キャッチ率は10% 強とそれほど高くはないが，まぐろ類は漁船がターゲットとして狙っていることもあり，極端な割合ではない。相関係数および MSE を指標とした5-fold cross-validation の結果，Tweedie モデルの精度は ad hoc なモデルのそれに比べて多少は高かったが，その差異はさほど大きくはなっていない。また両者の CPUE 年トレンドは多少の違いは見受けられるが，非常に良く似ている。これらの原因はゼロ・キャッチ率にあると考えられ，この例のようにゼロ・キャッチの割合がさほど高くない場合には Tweedie モデルの有意性が顕著に表れないとも考えられる。

このことから，ゼロ・キャッチ率がさほど高くない場合には，全ての応答変数に微小量（定数項）を足し込む ad hoc な方法を適用しても，実用上大きな問題は生じない，と考えたこともあり，第 3 章（3-3節）のインド洋キハダの実際例（3-3-1節）においても，この ad hoc な共分散分析モデルを使用している。

5-4節では，この Tweedie モデルを日本のはえ縄公庁船による北太平洋におけるクロトガリザメ資源の CPUE 標準化（漁獲量・努力量データ解析）に適用し，ad hoc な方法（共分散分析モデル），および Catch-NegativeBinomial モデルとの比較を行った。ゼロ・キャッチ率は80% を超え非常に高くなっているが，サメ類は漁船がターゲットとしては狙っていない混獲種であり，漁獲回避が要請されている。相関係数および MSE を指標とした5-fold cross-validation の結果，Tweedie モデルの精度は ad hoc なモデルや Catch モデルのそれに比べてかなり高くなっている。２つの従来法の比較では，相関係数は Catch モデルの方が高く，MSE は ad hoc な方法が一見したところ優れている。しかし，この ad hoc な方法では殆ど全ての CPUE 予測値が0.5以下とかなり低くなっており，観測値と予測値の相関プロットから判断しても，バイアスが大きく精度も極端に低いと考えられる。CPUE 年トレンドは，Tweedie モデルと Catch-NB モデルとはかなり良く似ており全体的に減少傾向にあるのに対し，ad hoc モデルではフラットな感じになっており，違いが見受けられる。

この混獲種の例のように，ゼロ・キャッチの割合が高い場合には，Tweedie モデルのパフォーマンスが Catch モデルや ad hoc な方法のそれと比べて非常に優れており，今後の使用を推奨していきたい。一方で，全ての CPUE に定数項を足し込む ad hoc な共分散分析モデルは，予測値に極端なバイアスが生ずるため，計算が簡便なこともあり現状では非常に多く利用されてはいるが，望ましいことではない，と考えている。

結論として，実用上は，ゼロ・キャッチ率が少ない場合には全ての応答変数に定数項を加えた ad hoc な方法でもやむを得ないが，ゼロ・キャッチ率が半分以上など，その割合が高い場合には Tweedie モデル（それが不可能な場合には代用としての Delta 型２段階モデルもしくは Zero-inflated モデル）を使用すべき，と考えている。その中間に位置する場合には Catch モデルを利用することも一案であるが，今後シミュレーションを含めた更なる検討が必要である。

第 6 章　結論：まとめと今後の課題

第 6 章は，本論文全体の結論部である。6-1節では，本研究で得られた知見について，まぐろ類を対象にした水産資源解析（CPUE 標準化）における問題解決の視点から，また（理論的な検討をも含めた上で）統計モデルやデータマイニング的なアプローチの CPUE 解析（社会問題・自然現象）への応用の観点からに大別して，系統的かつ具体的に整理する。6-2節では，本論文での研究成果を踏まえた今後の課題について，記述する。

6-1. 本研究で得られた知見

まず，CPUE 解析を水産資源解析における１つの重要な問題と捉えた場合，その主要な目的は，要因分析（年（Year）に関係する要因効果推定を通じた CPUE 年トレンドの抽出）であり，主に相対資源量の増減傾向の把握のため，及び資源の絶対量推定のためのモデルへのチューニング・インデックスとして，標準化された CPUE を利用することが多い。特に，後者の場合には使用する CPUE シリーズの年トレンドの微妙な差異が，資源評価モデルによる絶対量推定値の違いとなって表れることも多く（付録 B），CPUE の解釈に当たっては，細心の注意が必要である。

本研究では，主にCPUE年トレンド抽出に影響を与える原因と考えられる，以下の3つの問題について取り上げる。これらの問題は全て，CPUE年トレンドの推定及びチューニング・インデックスとしての指標の信頼性のみならず，CPUEに影響を与えていると考えられる様々な説明要因の統計的なチェックの意味においても，極めて重要である。

1) 同一の統計モデルにおける，情報量規準やstepwise検定を用いた説明要因（主効果および積の形で表現される2要因交互作用）の取捨選択（第3章）
2) 標準化されたCPUEに相対的なエリア・サイズを掛け合わせて得られる，資源量指数の計算過程でのCPUE解釈の問題（第4章）
3) ゼロ・キャッチを含むモデル（ad hocな方法，Catchモデル，Delta型2段階モデ，Zero-Inflatedモデル，Tweedieモデル）を比較する問題（第5章）

1)では，CPUE標準化に多く用いられる共分散分析モデルを中心に，小標本の場合やパラメーター数の標本数に占める割合が大きい場合，大標本の場合，ネスト構造を持つモデルの場合などを取り上げて，実際の漁業データや仮想データを用いた様々な情報量規準（ネスト構造を持つモデルではstepwise検定も合わせて使用）によるモデル選択，すなわち説明変数の取捨選択を行い，異なった指標を利用することによって違ったモデルが選択され，ひいては抽出されたCPUE年トレンドの差異を生じることを説明した。ここでは，特に水産資源解析分野で広く用いられている情報量規準であるAICと，他の規準を使用した場合のCPUE年トレンドの違いについて強調しておきたい。なお，このCPUE年トレンド推定の問題は，この変数選択のみならず，2)，3)でも起こりうる。

2)では，ミナミマグロ資源におけるCPUE解釈の問題，すなわち操業がない時空間のCPUE予測の問題を取り上げて，ニューラルネットワークを利用して解析した。具体的には，過去から現在にかけて漁場が縮小しているミナミマグロ資源において，過去に漁獲があり現在操業がない時空間のCPUEの予測，つまり操業がないセルのCPUEと操業が行われたセルのCPUEの比（0から1までの間の値を取る）の推定は，CPUEに相対面積指数を掛け合わせた資源量指数の計算にダイレクトに影響しており，抽出されたCPUE年トレンドの違いの一番の原因である欠測セルのCPUEを予測することは，極めて重要な問題である。ニューラルネットワークによるCPUE予測値を元にして計算されたこのCPUE比は経年変化がさほど大きくなく，その値は0.8～1前後を推移している。これは，1998年から2000年にかけて局所的に行われた日本の調査漁獲におけるCPUE比（年，季節，エリアは非常に局所的であるが0.7前後を記録）と比べてみても，極端な矛盾は見られない。以上の点から，ニューラルネットワークなどにより操業がない時空間における精度の高いCPUE予測値を得る解析は，漁海況予測なども含め極めて広い適用範囲が想定され，調査コストの軽減や効率的な船の運行なども含めて，水産資源解析学やまぐろ漁業への実用上の貢献度が高いと考えられる。

3)では，まぐろはえ縄漁業で混獲されるサメ類などを想定し，ゼロ・キャッチを含む場合，特にその割合が高い場合の解析方法について詳細に議論した。具体的には，ゼロ・データを統一的に取り扱える確率過程の考え方を利用したTweedieモデルを用いて，ゼロ・キャッチ率が10%程度と低い日本のはえ縄商業船によるインド洋キハダ資源のCPUE解析，及びその割合が80%以上と高い日本のはえ縄公庁船による北太平洋クロトガリザメ資源のCPUE標準化を行い，従来法であるad hocな共分散分析モデルやCatch Negative-BinomialモデルとのCPUE年トレンドを比較した。その結果，前者のゼロ・キャッチ率が低いターゲット種のケースでは，年トレンドに極端な違いが見られなかったのに対し，後者のゼロ・キャッチ率が高い混獲種の場合には，Tweedieモデルから得られたCPUEの年トレンドは，他の方法，特に旧来多く使用されてきたad hocな方法とかなり異なる結果が得られたため，適用に当たっては注意が必要である。

次に，統計モデルやデータマイニング的なアプローチのCPUEへの適用の視点から，そして統計学や情報科学の理論的な側面から，本研究で得られた成果について，整理・検証する。主要な3つの課題は，以下のa)～c)であり，前述の1)～3)を言い換えたものと考えることも可能であるが，水産資源解析における実際問題への適用の観点からは，得られたCPUE年トレンドの違いや結果の与える影響に重点を置くのに対し，この応用統計学的な観点からは，仮想データによる計算機シミュレーションおよび漁業データによるクロス・バリデーションを通じたモデルの精度評価やモデルの性能チェックに焦点を合わせている。

a) 共分散分析モデルにおける，様々な情報量規準やstepwise検定による計算機実験を通じたモデル（説明要因（変数）の組合せ）の性能チェック（第3章）
b) ニューラルネットワークを利用したミナミマグロ資源の操業がないセルのCPUE予測と簡

便な要因分析法（CPUE トレンド抽出法）の提案（第 4 章）

c） ゼロ・キャッチを含むモデルにおける，Tweedie 分布モデルの精度評価，および従来の手法（ad hoc な方法・Catch モデル）との性能の比較（第 5 章）

a）では，CPUE 標準化に多く用いられる CPUE-LogNormal モデル（共分散分析モデル）を取り上げて，小標本の場合やパラメーター数の標本数に占める割合が大きい場合，大標本の場合，ネスト構造を持つモデルの場合，正規混合分布モデルの場合など幾つかのケースにおいて，複数の候補となるモデル（説明要因（変数）の組合せ）の中から定めた真のモデルから乱数を発生させ，正しいモデルを選ぶという選択パフォーマンスを通じて，AIC をはじめとする様々な情報量規準，および F 検定に代表される stepwise 検定の性能評価を行った。

第 3 章で得られた主な結果は，理論的な検証も含めると，以下のようになる。

ⅰ）小標本の場合やパラメーター数の標本数に占める割合が大きい場合には，水産資源解析で広く利用されている AIC（Akaike's information criterion）はパラメーター数の多い複雑なモデルを選択しがちであり，AIC に有限修正を施した情報量規準である c-AIC（finite correction of Akaike's information criterion）の選択パフォーマンスが，AIC のそれに比べて優れていることを 2 元配置分散分析型の計算機シミュレーションにより示した。

ⅱ）大標本の場合にも，AIC が複雑なモデルを選びがちであることを例示し，一致性と呼ばれる漸近的に望ましい性質を持つ情報量規準である BIC, HQ, CAIC の真のモデルを選ぶ選択パフォーマンスが，AIC のそれと比較して全体的に高くなることを回帰分析型シミュレーションにより実証した。ただし，説明変数間の相関が高い場合，すなわち多重共線性を持つ場合には AIC の性能が一致性を持つ情報量規準のそれよりも良くなるケースが存在する。

ⅲ）大標本の場合に一致性を持つ情報量規準である HQ において，2 より大きい任意定数と定め方の自由度が高い定数項 c を2.01ないし2.71（または3.59）とおくことの妥当性を情報量規準間のペナルティ項の大きさの grid search などによる評価を通じて示し，$c = 2^{*}k\,[\log(n)/\{\log(\log(n))\}-2]$（$0<k<1$）の形で k=0.05ないしは0.1程度の微少量に設定する方法の性能が良いことを，回帰分析型のシミュレーション実験により示した。

ⅳ）ネスト構造を持つモデルにおいて，真のモデルが候補となるモデルを含まない場合に AIC が持つ偏りを修正した，AIC の精密評価でもあり性能が良いと言われてきた情報量規準 TIC（Takeuchi's information criterion: 竹内情報量規準）が，実はある種の一般化線形モデル（連結関数が恒等写像かつ正規誤差を持つ場合）には AIC と同等になることを，分散分析モデルにおける TIC の導出および理論的な検証から導き，TIC の選択パフォーマンスが AIC のそれとほぼ同等であるという分散分析型シミュレーションの結果と合わせて，このような場合に計算の複雑な TIC を敢えて使用する必要がないことを示した。

ⅴ）ネスト構造を持つモデルにおいて情報量規準と並んで使用可能な stepwise 検定について，検定のパスによって最終的なモデル選択結果が異なることを示し，その場合の判断基準（検定結果の解釈）について，2 通りの方法を提示した。また，分散分析型のシミュレーションを通じて F 検定に代表される stepwise 検定と AIC や BIC などの代表的な情報量規準との比較を行い，全体としては BIC の選択パフォーマンスがわずかながら優れていること，及び検定の有意水準を小さく設定した場合にはパラメーター数の少ない単純なモデルが選択されがちであることを示した。なお，共分散分析モデルにおける情報量規準 AIC と stepwise な F 検定との理論的な比較検討も行った。

ⅵ）体長組成データの年齢分解などに使用される正規混合分布モデルにおいて，カイ二乗検定が理論的に使用出来ないこと，および AIC や BIC などを含む罰金付き最尤法が実際上適用可能なこと，具体的には，コンポーネント数を過大推定する可能性はあるが，漸近的に過小推定することはないという片方からの一致性が成立することを説明し，コンポーネント数および各正規分布のパラメーター推定のためのシミュレーションを通じて，AIC や BIC の情報量規準，特に Dilichlet 事前分布を仮定した Bayes 型情報量規準の選択パフォーマンスが，コンポーネント数を正しく推定するという意味において高くなることを実証した。

第 4 章および第 5 章では，ニューラルネットワークと Tweedie モデルを使用して，それぞれミナミマ

グロ資源の操業がない時空間の CPUE 予測と簡便な要因分析法の提案，およびゼロ・キャッチを含む場合の従来法である ad hoc な方法や Catch モデルとの比較も含めたモデルの精度評価を行ったが，ここでは計算機シミュレーション実験ではなく実際の漁業データのみを利用しているため，モデルの性能チェックに n-fold cross-validation を使用し，その尺度としては観測値と予測値の相関係数，相関プロット，絶対誤差，平均二乗誤差（MSE: mean squared error）などを利用した。

b）では，ミナミマグロ資源を題材とした教師付きニューラルネットワークによる CPUE の予測性能が，全く同じ条件での解析である MCMC 法に基づく EM algorithm によるそれよりも格段に高いことを示し，この CPUE 予測値を元に算出された CPUE 年トレンドが，一般化線形モデル（共分散分析モデル）によるそれと比較的似ていることを例示した。この過程では，今回提示したニューラルネットワークによる予測値に基づく簡便な要因分析手法を使用しており，その妥当性の 1 つの傍証になっていると考えられる。

c）では，ゼロ・キャッチの割合が少ない場合に対応する日本のはえ縄商業船によるインド洋キハダ資源の CPUE 解析，およびゼロ・キャッチ率が高い場合に該当する日本のはえ縄公庁船による北太平洋クロトガリザメ資源の CPUE 標準化を通じて，Tweedie モデルと全ての CPUE に定数項を加える ad hoc な方法，離散変量である漁獲量を応答変数にした Catch Negative-Binomial モデルとの比較をクロス・バリデーションにより行った。その結果，ゼロ・キャッチ率が低い場合には Tweedie モデルの有意性が顕著に表れず，ad hoc な方法を使用しても実用上さほど問題が生じないと考えられる一方，ゼロ・キャッチの割合が高い場合には，Tweedie モデルの精度が他に比べて非常に高くなり使用が推奨されるが，ad hoc な方法では観測値の大きさにかかわらず予測値が極端に小さくなることもあり，バイアスの大きさを鑑みると，このような事例に適用すべきでないと思われる。

6-2. 今後の研究課題

最後に，これらの研究成果を踏まえた今後の課題や問題点について，水産資源学，応用統計学といった視点を分けずに整理し，列挙する。

まず，情報量規準と stepwise 検定を中心にした，同一の統計モデルにおける説明変数の取捨選択については，使用した仮定や条件に関する一般性・普遍性の担保が最大の課題である。Catch-Poisson モデルや Catch Negative-Binomial モデルなどの一般化線形モデル（GLM: generalized linear model）によるシミュレーションも含め，今回取り上げられなかった予測誤差最小の観点からの計算機実験にも取り組んで行きたい。CPUE データの特性として，観測誤差が一定ではなく絶対値に比例していると考えられるために，共分散分析モデルにおいては正規誤差ではなく対数正規誤差の仮定を出発点としたが，その妥当性を検証する意味も含めて，Box-Cox 変換などによる尤度の比較も行っていきたい。そして，Kalman-filter などに代表される状態空間モデルを用いた CPUE の時系列解析の適用も，合わせて視野に入れていきたい。

また，ニューラルネットワークや Tweedie 分布モデルの漁業データへの適用を通じたクロス・バリデーションによるモデルの精度評価，性能チェックでは，検証に用いる指標の検討，すなわち観測値と予測値の Pearson's 相関係数や平均二乗誤差（MSE：mean squared error），絶対誤差などの妥当性に関する議論や，これらに代わる指標の開発が急務である。さらに，n-fold cross-validation においては，多く用いられるランダム分割ではなく，系統的な誤差を持つ場合のシミュレーション実験も必要であろう。

なお，各々の手法に目を向けた場合には，ニューラルネットワークでは要因分析手法の検討，特に連続変量の取り扱い方法やカテゴリカル変数における添字の加工方法（総和を取るべきか，平均を取るべきか等）に議論の余地があり，Tweedie モデルのゼロ・キャッチデータへの適用に関しては，Delta 型 2 段階モデルもしくは Zero-inflated モデルを含めた比較や計算機シミュレーションを通じた検証が必要である，と思われる。

以上のような問題点が，本研究を通じて得られた今後の研究課題となる。

謝　辞

本研究をとりまとめるにあたり，終始懇切丁寧な御指導御鞭撻を賜りました筑波大学社会工学系教授椿広計博士に心から厚く御礼申し上げる。また，本論文の御校閲を頂いた同教授吉田健一博士，同講師佐藤忠彦博士，匿名の予備審査委員の先生方に深く感謝の意を表する。さらに，本研究を遂行するにあたり数々の有益なご助言を頂いた東京大学海洋研究所助教授平松一彦博士，筑波大学数学系教授赤平昌文博士，元東京大学海洋研究所教授故松宮義晴博士，東京海洋大学海洋科学部教授山田作太郎博士，同準教授田中栄次博士，同助教北門利英博士，統計数理研究所教授江口真透博士，同準教授南美穂子博士，東京医科歯科大学準教授

吉岡耕一博士（現国士舘大学教授），日本鯨類研究所袴田高志氏，遠洋水産研究所元浮魚資源部長鈴木治郎博士，同温帯性まぐろ資源部長宮部尚純氏，同数理解析研究室長余川浩太郎氏，同数理解析研究室竹内幸夫氏，市野川桃子博士，小国奈津子氏，同鯨類管理研究室岡村寛博士，同国際海洋資源研究員西田勤博士，同熱帯性まぐろ研究室長岡本浩明博士，同温帯性まぐろ研究室高橋紀夫博士，伊藤智幸博士，黒田啓行博士，同混獲生物研究室長清田雅史博士，同混獲生物研究室松永浩昌氏，同熱帯性まぐろ資源部長本多仁博士，同図書室近藤禮子氏に厚く御礼申し上げる。

参考文献

Akaike H., 1973: Information theory and an extention of the maximum likelihood principle. *2nd International Symposium on Information Theory*. (ed. by Petrov, B. N., and Csaki, F.), Akadimiai Kiado, Budapest, 267–281.

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

麻生英樹，1988: ニューラルネットワーク情報処理 産業図書，198 pp.

麻生英樹，津田宏治，村田　昇，2003: パターン認識と学習の統計学，岩波書店，225 pp.

Badcook E. A. and Mcallister M. M., 2001: Bayesian generalized linear models for standardization catch rate indices of abundance, ICCAT/SCRS/01/43, 29 pp.

Bayler P. B., 1993: Quasi-likelihood estimation of marked fish recapture. *Canadian Journal of Fish and Aquatic Science*. **50**, 2077–2085.

Bedrick E. J. and Tsai C. L., 1994: Model selection for multivariate regression in small samples. *Biometrics,* **50**, 226–231

Berry M. J. A. and Linoff G., 1997: Data mining techniques: for marketing, sales, and customer support, John Wiley and Sons, New York, 454 pp.

Bollen K. A., 1989: Structural equations with latent variables, John Wiley and Sons, New York. 528 pp.

Bozdogan H., 1987: Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika,* **52** (3), 345–370.

Breiman L. J. Friedman R. A. Olshen R. A. and Stone C. J., 1983: Classification and regression trees, Wadsworth International Group, Belmont, California, 368 pp.

Broomhead D. S. and Lowe D., 1988: Multivariable functional interpolation and adaptive networks. *Complex Systems,* **2**, 321–355.

Burnham K. P. and Anderson D. R., 1998: Model selection and inference: A practical information -Theoretic approach-, Springer, New York, 353 pp.

Chernoff H. 1954: On the distribution of the likelihood ratio. *Annals of Mathimatical Statistics*., **25**, 573–578.

CCSBT, 1998: Report of the first meeting of the scientific assessment group (SAG), 41 pp.

Chen D. G. and Ware D. M.,1999: A neural network model for forecasting fish stock recruitment. *Canadian Journal of Fish and Aquatic Science,* **56**, 2385–2396.

Cleveland W. S. and Grosse E., 1991: Computational methods for local regression. *Statistics and Computing*, **1**, 47–62.

Dobson A. J., 1990: An introduction to generalized linear models, Chapman and Hall, London, 174 pp.

Edwards D., 2000: Introduction to graphical modeling (2^{nd} edition), Springer, New York, 333 pp.

Eguchi S., and Yoshioka K., 2001: Maximum penalized likelihood estimation of finite mixture with a structural model. Research memorandum, *The Institute of Statisical Mathematics,* **809**, 30pp.

Fahrmeir L. Tutz G. and Fahrmeir L., 2001: Multivariate statistical modeling based on generalized linear models (2^{nd} edition), Springer, New York, 517 pp.

Gavaris S., 1980: Use of a multiplicative model to estimated catch rate and effort from commercial data. *Canadian Journal of Fish and Aquatic Science,* **37**, 2272–2275.

Gavaris S., 1988: An adaptive framework for the estimation of population size, CAFSAC Research Document, 88/12, 12 pp.

Hannan E. J. and Quinn B. G., 1979: The determination of the order of autoregression. *J. Royal Statist, Soc. Ser. B,* **41**, 190–195.

Haralabous J. and Georgakarakos S., 1996: Artificial neural networks as a tool for species identification of fish schools. *ICES Journal of Marine Science*, **53**, 173–180.

Hartigan J. A., 1975: Clustering algorithms, John Wiley and Sons, New York, 352 pp.

Hastie T. and Tibshirani R., 1990: Generalized Additive Models, Chapman and Hall, London, 335 pp.

Hastie T. Tibshirani R. and Friedman J., 2001: The elements of statistical learning: Data mining, inference and prediction, Springer, New York, 533 pp.

Haykin S., 1994: Neural networks: A comprehensive foundation, Macmillan, New York, 750 pp.

Hilborn R. and Walters C. L., 1992: Quantitative fisheries stock assessment. Chapman and Hall, 570pp.

平松一彦，1995: 統計モデルによる CPUE 標準化．漁業資源研究会議北日本底魚部会報，**28**，87–97.

Hinton M. G. and Nakano H., 1996: Standardizing catch and effort statistics using physiological, ecological, or behavioral constraints and environmental data, with an application to blue marlin (*Makaira Nigricans*) catch and effort data from Japanese longline fisheries in the Pacific. *Inter-American Tropical Tuna Commission Bulletin,* **21** (4), 171–200.

Hurvich C. M. and Tsai C. L., 1989: Regression and time series model selection in small samples. *Biometrika,* **76**, 297–307.

Hurvich C. M. and Tsai C. L., 1991: Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika,* **78**, 499–509

ICCAT, 1997: Report of the bluefin tuna methodology session. International Commission for the Conservation of Atlantic Tunas, Coll. Vo. Sci. Pap. Vol. XL **VI** (1), 187–201.

Jensen F. V., 2001: Baysian networks and decision graphs, Springer, New York, 268 pp.

Jorgensen B, 1997: *The theory of dispersion models,* Chapman and Hall, London, 237 pp.

狩野　裕，三浦麻子，2002: AMOS, EQS, CALIS によるグラフィカル多変量解析（増補版): 目で見る共分散構造分析，現代数学社，293 pp.

Kass G. V., 1980: An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* **29**, 119–127.

Kiso K., Akamine T., Ohnishi S., and Matsumiya Y, 1992 : Mathematical examinations of the growth of sea-run and fluviatile forms of female masu salmon *Oncorhynchus masou* in rivers of the southern Sanriku district, Honshu, Japan. *Nippon Suisan Gakkaishi,* **58**, 1779–1784.

Kohonen, T., 1989: Self-Organization and Associative Memory (3rd edition), Springer, Berlin, 368 pp.

Lambert D., 1992: Zero-Inflated Poisson regression, with an application to defects in manufacturing. *Technometrics,* **34** (1), 1–14.

Leroux B., 1992: Consistent estimation of a mixture distribution. *Annals of statistics,* **20**, 1350–1360.

Large P A., 1992: Use of a multiplicative model to estimate relative abundance from commercial CPUE data. *ICES Journal of Marine Science,* **49**, 253–261

Little R. C. Milliken G. A. Stroup W. W. and Wolfinger R. D., 1996: SAS system for mixed models, SAS Institute Inc., USA, 633 pp.

Little R. J. A. and Rubin D. B., 2002: Statistical analysis with missing data (2nd edition), John Wiley and Sons, New York, 381pp.

Lo N. C. L. D. Jacobson L. D. and Squire J. L., 1992: Indices of relative abundance from fish spotter data based on Delta-Lognormal models. *Canadian Journal of Fish and Aquatic Science*. **49**, 2515–2526.

Matsunaga H., Shono H., and Yokawa K., 2002: Standardization of CPUE of Pacific bluefin tuna caught by Japanese distant-water and offshore longliners in the spawning ground from 1953–2000. ISC-BFT-WG/02/Doc.11, 7 pp.

McCullagh P. and Nelder J. A., 1989: Generalized linear models (2nd edition), Chapman and Hall, London, 511 pp.

宮川雅巳，1997: グラフィカルモデリング，朝倉書店，177 pp.

Miyashita T., Shono H., and Okamura H., 2001: GLM analysis of the JSV data for the Antarctic minke whale. IWC-SC/53/IA15, 11 pp.

Neath A. A., and Cavanaugh J. E., 1997: Regression and time series model selection using variants of the Schwarz information criterion. *Commun. Statist.-Theory. Meth.,* **26** (3), 559–580.

能勢幸雄，石井丈夫，清水　誠，1988：水産資源学，東京大学出版会，217 pp.

O'Brien C. M. B. Kell L. T. Santiago J. and V. Ortiz de Zarate., 1997: The use of generalized linear models for the modeling of catch-effort series. II: Application to north Atlantic albacore surface fishery. ICCAT/SCRS/97/49, 14 pp.

Okamoto H., Satoh K., Shono H., and Miyabe N., 2003: Standardized Japanese longline CPUE for yellowfin tuna in the Atlantic Ocean up to 2001. ICCAT/SCRS/2003/056, 24 pp.

大滝　厚，堀江宥治，Dan Steinberg，1998: 応用2進木解析法，日科技連，273 pp.

Pella J. J. and Tomlinson P. K., 1969: A generalize stock production model. *Inter-American Tropical Tuna Commission Bulletin,* **13**, 421–496.

Quinlan J. R., 1993: C4.5: Programs for machine learning, Morgan Kaufmann, San Mateo, 302 pp.

Quinn T. J. and Deriso R. B., 1999: Quantitative fish dynamics, Oxford, New York, 542 pp.

Reed W. J., 1996: Analyzing catch-effort data allowing for randomness in the catching process. *Canadian Journal of Fish and Aquatic Science,* **43**, 174–186.

Repley B. D., 1994: Neural networks and flexible regression and discrimination. *Advances in Applied Statistics,* Supplement to Journal of Applied Statistics, **21**, 39–57.

Ridout M. Hinde J. and Demetrio G. G. B., 2001: A score test for testing a zero-inflated Poisson regression model against zero inflated negative binomial alternatives. *Biomaterics,* **57**, 219–223.

Rissanen J., 1983: A universal prior for integers and estimation by minimum description length. *Ann. of Statist.,* **11**, 416–431.

Robson D. S., 1966: Estimation of the relative fishing power of individual ships. *Research Bulletin, International Commission for the North-west Atlantic Fisheries*, **3**, 5–14.

Rumelhart D. E. McClelland J. L. and PDP Research Group, 1986: Learning internal representation by error propagation. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, **1**, MIT Press, Cambridge, 318–362.

Russell E. S., 1931: Some theoretical consideration on the "overfishing" problem. *Journal du conseil / Conseil permanent international pour l' exploration de la mer.* **6** (1), 3–20.

坂本慶行，石黒真木夫，北川源四郎，1983: 情報量統計学，共立出版，236pp.

櫻井茂明，1998：ファジィ帰納学習による数値予測規則の学習．電気学会論文誌，**C118** (9), 1369–1375.

櫻田玲子，2005：VPA を用いたインド洋キハダの資源量推定に関する研究，
東京海洋大学海洋科学部資源管理学科平成17年度卒業論文，66pp.

佐藤整尚，1996: 統計モデルとしてのニューラルネットワーク，統計数理，**44** (1)，85–98.

Schwarz G., 1978: Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Searle S. R. Casella G. and McCulloch C. E., 1992: *Variance components*. John Wiley and Sons, New York, 501 pp.

Shono H., 2000: Efficiency of finite correction of Akaike's information criteria. *Fisheries Science,* **66** (3), 608–610.

Shono H., 2001: Comparison of statistical models for CPUE standardization by information criteria − Poisson model vs. Log-normal model −, IOTC-WPM/01/1, 12 pp.

Shono H., 2002: Attempts for estimation of standardized CPUE by tree-regression models and neural networks. CCSBT-ESC/0209/38, 18 pp.

Shono H., 2004: Attempts for multiple imputation of SBT CPUE by new statistical method. CCSBT-ESC/0409/43, 10 pp.

Shono H., 2005: Is model selection using Akaike's information criterion appropriate for CPUE standardization in large samples? *Fisheries Science,* **71** (5), 976–984.

Shono, H., 2008?: Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research,* (To appear)

Shono H., and Ogura M., 1999: The standardized skipjack CPUE including the effect of searching devices, of the Japanese distant water pole and line fishery in the Western Central Pacific Ocean. ICCAT-SCRS/99/59, 18 pp.

Shono H., and Ogura M., 2000: The standardized albacore CPUE of the Japanese distant water pole and line fishery, including the effect of searching devices. NPALB/00/8, 9 pp.

Shono H., Matsumoto T., Ogura M., and Miyabe N., 2000: Preliminary analysis of effect of fishing gears on catch rate for the Japanese purse seine fishery. SPC-SCTB13/WP/RG-3, 13 pp.

Shono H., Tsuji S., Takahashi N., and Itoh T., 2001: Preliminary analysis for CPUE standardization and area stratification by tree-regression models. CCSBT-SC/0108/30, 17 pp.

Shono H., Okamoto H., and Nishida T., 2002: Standardized CPUE for yellowfin tuna (*Thunnus albacares*) resources in the Indian Ocean by generalized linear models (GLM) (1960–2000). IOTC/WPTT/02/12, 12 pp.

Shono H., Okamoto H., and Nishida T., 2005: Standardized CPUE for yellowfin tuna (*Thunnus albacares*) resources in the Indian Ocean up to 2003 by Generalized Linear Models (GLM) (1960–2003). IOTC-2005-WPTT-15, 16pp.

庄野　宏，2000: 情報量規準とステップワイズ検定の比較と水産資源解析への応用．遠洋水産研究所報告，**37**，1–8.

庄野　宏，2001：情報量規準 TIC と c-AIC によるモデル選択の有効性．遠洋水産研究所報告，**38**，21–28.

庄野　宏，2004：CPUE 標準化に用いられる統計学的アプローチに関する総説．水産海洋研究．**68**（2），106–120.

庄野　宏，2006：モデル選択手法の水産資源解析への応用，―情報量規準とステップワイズ検定の取り扱い―．計量生物学，**27**（1），55–67.

庄野　宏，椿　広計，2006：ニューラルネットワークによる水産資源解析．―CPUE 予測と要因分析の試み―．計量生物学，**27**（1），35–53.

Simonoff J. S., 1998: Smoothing methods in statistics, Springer, New York, 338 pp.

Smith M., 1996: Neural networks for statistical modeling, International Thomson Computer Press, Boston, 256 pp.

Soto M., Moron J., and Pallares P., 2000: Standardized catch rates for yellowfin (*Thunnus albacares*) from the Spanish purse seine fleet (1984–1995). IOTC-WPTT/00/04, 15 pp.

Soto M., Pallares P., Gaertner D., Delgado deMolina, Fonteneau A., and Ariz Y. J., 2002: Standardization of tropical purse seine fishing effort by generalized linear model (GLM). IOTC-WPTT/02/26, 15 pp.

Stefansson G., 1996: Analysis of groundfish survey abundance data: combining the GLM and delta approaches. *ICES Journal of Marine Science*. **53**, 577–588.

Sugiura N., 1978: Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist.-Theory. Meth.,* **7** (1)，13–26.

Takahashi N., Tsuji S., Itoh T., and Shono H., 2001: Abundance indices of southern bluefin tuna based on the Japanese longline fisheries data, 1969-2000, along the interim approach agreed for the 2001 stock assessment. CCSBT-SC/0108/28, 39 pp.

高橋行雄，大橋靖雄，芳賀敏郎，1989：SAS による実験データの解析，367 pp.

竹内　啓，1976：情報統計量の分布とモデルの適切さの規準．数理科学，**153**，12–18.

竹澤邦夫，1999：バギング樹形モデルを用いたモデル合成．システム農学，**15**（1），1–8.

竹澤邦夫，2001：みんなのためのノンパラメトリック回帰，吉岡書店，560 pp.

寺野隆雄，2002：データマイニングの展望．計測と制御，**41**（5），315–324.

Toscas P. and Thomas M., 1998: Spatial analysis of southern bluefin tuna catch per unit effort data: A best linear unbiased predictor approach. CCSBT-SC/9807/10, 33 pp.

豊田秀樹，1998a：共分散構造分析［入門編］，朝倉書店，319 pp.

豊田秀樹，1998b：非線形多変量解析，朝倉書店，325 pp.

豊田秀樹，2000：共分散構造分析［応用編］，朝倉書店，303 pp.

豊田秀樹，2003：共分散構造分析［技術編］，朝倉書店，238 pp.

椿　広計，1988：一般化線形模型の問題点と疑似尤度の一般化．応用統計学，**17**（1），1–12.

Tsukimoto H, 2000：Extracting rules from trained neural networks. *IEEE Transactions on Neural networks*, **11** (2)，377–389.

月本　洋，森田千絵，2000：予測モデルからのルール抽出，「発見科学とデータマイニング」，共立出版，24–33.

Tukey J. W., 1977: Exploratory data analysis. Addison Wesly, Reading, Mass.

Tweedie M. C. K., 1984: An index which distinguishes

between some important exponential families. In Statistics: Applications and New Directions. *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*. (ed. by Ghosh J. K. and Roy J.), Calcutta: Indian Statistical Institute, 579–604.

内田　治，2002：例解データマイニング入門，日本経済新聞社，214 pp.

Vapnik V., 1998: Statistical learning theory, John Wiley & Sons, New York, 736 pp.

Verbeke G. and Molenberghs G., 1997: Linear mixed models in practice — A SAS oriented approach —, Springer, New York, 306 pp.

Venables W. N. and Toscas P. J., 2002: Estimating a CPUE series for southern bluefin tuna using enhanced tree-based modeling methods. CCSBT-ESC/0209/31, 24 pp.

Watters G. and Deriso R., 2000: Catch per unit of effort of bigeye tuna: a new analysis with regression tree and simulated annealing. *Inter-American Tropical Tuna Commission Bulletin,* **21** (8), 531–571.

Wedderburn R. W. M., 1974: Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biomatrika*, **61**, 439–447.

Whittaker E., 1923: On a new method of graduation. *Proceedings of the Edinburgh Mathematics Society,* **41**, 63–75.

Wise B., Bugg A., Shono H., Barry S., Nishida T., Barratt D., and Kalish J., 2002: Standardization of Japanese longline catch rates for yellowfin tuna in the Indian Ocean using GAM analyses. IOTC-WPTT/02/11, 15 pp.

Witten I. H. and Frank E., 2000: Data mining: Practical machine learning tools and techniques with JAVA implementations. Morgan Kaufman, 371 pp.

山田作太郎，田中栄次，1999：水産資源解析学，成山堂書店，151 pp.

Yokawa K., and Shono H., 2000: Preliminary stock assessment of swordfish (*Xiphias gladius*) in the Indian Ocean. IOTC-WPB-00-02, 5 pp.

Yokawa K., Okazaki M., Okamura H., Matsumoto T., Uozumi Y., and Saito H., 2002: An estimation of effective fishing effort of Japanese longliners on the Atlantic blue marlin, *Makaira Nigricans,* in the Atlantic Ocean. *Handbook and Abstract of Third International Billfish Symposium,* 25.

吉富康成，2002：ニューラルネットワーク，朝倉書店，174 pp.

付録 A.　分散分析モデルにおける TIC の導出ならびに AIC との比較

AIC の有限修正の場合と比較すると TIC のペナルティ項は複雑な形をしており，理論上は計算可能であるが実際には厄介なことが多い。そこで，付録 A では論文中の仮想例やシミュレーション実験で仮定したような正規誤差を持つ 2 元配置分散分析モデル（分散分析モデルは CPUE 標準化に良く用いられる）を用いて TIC の導出および理論的な側面から AIC との比較を行った（庄野，2001）。

まず，Table A-1のような 2 元配置分散分析モデルを考える。標本数 n，パラメーター数 p（n>p），及び n/p の全てが整数であるとし，式（A.1）のような要因を含むモデルを仮定する。実際には，

$$\text{Log (CPUE)} = \text{(Intercept)} + \text{(Year)} + \text{(Area)} + \text{(Error)},\ \text{Error} \sim N(0, \sigma^2) \qquad \text{(A.0)}$$

の形の分散分析型モデルで（Area）の主効果が認められないものと仮定することとする。

$$\text{Log (CPUE)} = \text{(Intercept)} + \text{(Year)} + \text{(Error)},\ \text{Error} \sim N(0, \sigma^2) \qquad \text{(A.1)}$$

このとき，未知パラメーターベクトル $\Theta := (\mu, \sigma^2)$，$\mu = (\mu_1, \cdots, \mu_p)$ に対する対数尤度関数 l はその確率密度関数 f を書き下すことによって（A.2）式のように表現される。

$$\begin{aligned} l(\Theta \mid Y) &= \sum_{i=1}^{n} \log f(y_i \mid \mu_i, \sigma^2) \\ &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 \\ &\quad - \frac{1}{2\sigma^2}\{(y_1-\mu_1)^2 + \cdots + (y_{\frac{n}{p}}-\mu_1)^2 \\ &\qquad + (y_{\frac{n}{p}+1}-\mu_2)^2 + \cdots + (y_{2\frac{n}{p}}-\mu_2)^2 \\ &\qquad + \cdots \\ &\qquad + (y_{(p-1)\frac{n}{p}+1}-\mu_p)^2 + \cdots + (y_n-\mu_p)^2\} \end{aligned} \qquad \text{(A.2)}$$

そして，以後は計算の簡略化のために（A.2）式の { } の中身を☆とおく。すなわち，

$$\begin{aligned} ☆ &= (y_1-\mu_1)^2 + \cdots + (y_{\frac{n}{p}}-\mu_1)^2 \\ &\quad + (y_{\frac{n}{p}+1}-\mu_2)^2 + \cdots + (y_{2\frac{n}{p}}-\mu_2)^2 \\ &\quad + \\ &\quad + (y_{(p-1)\frac{n}{p}+1}-\mu_p)^2 + \cdots + (y_n-\mu_p)^2 \end{aligned} \qquad \text{(A.3)}$$

である。

このとき，j=1,…,p に対して

$$\begin{aligned} &\frac{\partial l}{\partial \mu_j} = \frac{1}{\sigma^2}\sum_{k=(j-1)n/p+1}^{j*n/p}(y_k - \mu_j), \\ &\frac{\partial^2 l}{\partial \mu_j^2} = -\frac{1}{\sigma^2}\frac{n}{p}, \\ &\frac{\partial^2 l}{\partial \mu_i \partial \mu_j} = 0 \ (i \neq j), \\ &\frac{\partial^2 l}{\partial \mu_j \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{k=(j-1)n/p+1}^{j*n/p}(y_k - \mu_j), \end{aligned} \qquad \text{(A.4)}$$

となることと

$$\begin{aligned} &\frac{\partial l}{\partial \sigma^2} = \frac{1}{2\sigma^4}\{☆\} - \frac{n}{2\sigma^2}, \\ &\frac{\partial^2 l}{\partial (\sigma^2)^2} = -\frac{1}{\sigma^6}\{☆\} + \frac{n}{2\sigma^4} \end{aligned} \qquad \text{(A.5)}$$

より，また（A.6）式の仮定

$$E[(y_k - \mu_j)] = 0,\ E[(y_k - \mu_j) = \sigma\ (k = (j-1)*n/p+1, \cdots, j*n/p) \qquad \text{(A.6)}$$

を用いると

$$\begin{aligned} &E\left[\frac{\partial^2 l}{\partial \mu_j^2}\right] = -\frac{1}{\sigma^2}\frac{n}{p}, \\ &E\left[\frac{\partial^2 l}{\partial \mu_i \partial \mu_j}\right] = 0 \ (i \neq j), \\ &E\left[\frac{\partial^2 l}{\partial (\sigma^2)^2}\right] = -\frac{1}{\sigma^6}E[\{☆\}] + \frac{n}{2\sigma^4} = -\frac{n}{2\sigma^4} \end{aligned} \qquad \text{(A.7)}$$

Table A-1. ANOVA model (corresponding to CPUE standardization) for TIC derivation

Area	1	2	・	・	・	n/p	Parameter
Year-1	Y(1)	Y(2)	・	・	・	Y(n/p)	u(1)
Year-2	Y(n/p+1)	Y(n/p+2)	・	・	・	Y(2n/p)	u(2)
・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・
・	・	・	・	・	・	・	・
Year-p	Y((p-1)n/p+1)	Y((p-1)n/p+2)	・	・	・	Y(n)	u(p)

Remark) Y(・), u(・) show the corresponding samples and unknown parameters.

となる。よって，

$$J(\Theta)=J(\mu,\sigma^2)=-E\left[\frac{\partial}{\partial\Theta\partial\Theta^t}l(\Theta\mid Y)\right]$$
$$=\begin{bmatrix}\frac{1}{\sigma^2}\frac{n}{p} & & & \\ & \ddots & & \\ & & \frac{1}{\sigma^2}\frac{n}{p} & \\ & & & \frac{n}{2\sigma^4}\end{bmatrix} \quad (A.8)$$

$$J(\Theta)^{-1}=J(\mu,\sigma^2)^{-1}=\begin{bmatrix}\sigma^2\frac{p}{n} & & & \\ & \ddots & & \\ & & \sigma^2\frac{p}{n} & \\ & & & \frac{2\sigma^4}{n}\end{bmatrix} \quad (A.9)$$

と表される。

また，

$$E\left[\frac{\partial l}{\partial\mu_j}\cdot\frac{\partial l}{\partial\mu_j}\right]=\frac{1}{\sigma^2}\frac{n}{p}\ (\because\text{標本の独立性の仮定より}) \quad (A.10)$$

$$E\left[\frac{\partial l}{\partial\sigma^2}\cdot\frac{\partial l}{\partial\sigma^2}\right]=E\left[\left(\frac{1}{2\sigma^4}\{\ \star\}-\frac{n}{2\sigma^2}\right)\left(\frac{1}{2\sigma^4}\{\ \star\}-\frac{n}{2\sigma^2}\right)\right]$$
$$=\frac{1}{4\sigma^8}E\left[\{\ \star\}^2\right]-\frac{n}{2\sigma^6}E\left[\{\ \star\}\right]+\frac{n^2}{4\sigma^4}$$
$$=\frac{1}{4\sigma^8}E\left[\{\ \star\}^2\right]-\frac{n^2}{4\sigma^4}=\frac{1}{4\sigma^8}\left(\frac{\kappa}{\sigma^4}-n^2\right) \quad (A.11)$$

とおくと（上のようにκを定める），

$$\{\star\}^2=\sum_{j=1}^{p}\sum_{k=(j-1)n/p+1}^{j*n/p}(y_k-\mu_j)^4+\sum_{i,j,k\neq l}(y_k-\mu_i)^2(y_k-\mu_j)^2 \quad (A.12)$$

$$E\left[\sum_{j=1}^{p}\sum_{k=(j-1)n/p+1}^{j*n/p}(y_k-\mu_j)^4\right]=E\left[\sum_{i=1}^{n}(y_i-\mu_i)^4\right]=nE\left[(y_i-\mu_i)\right.$$
（∵添字の付替え）(A.13)

$$E\left[\sum_{i,j,k\neq l}(y_k-\mu_i)^2(y_k-\mu_j)^2\right]=n(n-1)\sigma^4 \quad (A.14)$$

となることから

$$\frac{\kappa}{\sigma^4}-n^2=\frac{n}{(\sigma^2)^2}E\left[(y_i-\mu_i)^4\right]+\frac{n(n-1)\sigma^4}{\sigma^4}-n^2$$
$$=\frac{n}{(\sigma^2)^2}E\left[(y_i-\mu_i)^4\right]-n$$
$$=n\left\{\frac{1}{(\sigma^2)^2}E\left[(y_i-\mu_i)^4\right]-1\right\}=(*) \quad (A.15)$$

と変形出来る。

ここで，μ_i，σ^2をそれぞれ最尤推定量$\hat{\mu}_i$，$\hat{\sigma}^2$=RSS/nで置き換える。また，平均まわりの4次の積率を，標本における平均まわりの4次の積率$\hat{\mu}(4)$を用いて表現すると，

$$(*)=n\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^4\Big/(\hat{\sigma}^2)^2-1\right\}$$
$$=n\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^4}{\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2\right\}^2}-1\right]=n\left\{n^2\cdot\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^4\Big/(RSS)^2-1\right\}$$
$$=n\left\{\frac{n^2\cdot\hat{\mu}(4)}{(RSS)^2}-1\right\} \quad (A.16)$$

となる。但し，

$$RSS:=\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2 \quad \text{(residual sum of squares)},$$
$$\hat{\mu}(4):=\frac{1}{n}\sum_{i}^{n}(y_i-\hat{\mu}_i)^4 \quad (A.17)$$

とおく。

以上により

$$I(\hat{\Theta})=I(\hat{\mu},\hat{\sigma}^2)=E_g\left[\frac{\partial}{\partial\Theta}l(\hat{\Theta}|Y)\cdot\frac{\partial}{\partial\Theta^t}l(\hat{\Theta}|Y)\right]$$
$$=\begin{bmatrix}\frac{1}{\hat{\sigma}^2}\frac{n}{p} & & & * \\ & \ddots & & \\ & & \frac{1}{\hat{\sigma}^2}\frac{n}{p} & \\ * & & & \frac{n}{4\hat{\sigma}^4}\left\{\frac{n^2\cdot\hat{\mu}(4)}{(RSS)^2}-1\right\}\end{bmatrix} \quad (A.18)$$

$$J(\hat{\Theta})^{-1}I(\hat{\Theta})=\begin{bmatrix}1 & & & * \\ & \ddots & & \\ & & 1 & \\ * & & & \frac{1}{2}\left\{\frac{n^2\cdot\hat{\mu}(4)}{(RSS)^2}-1\right\}\end{bmatrix} \quad (A.19)$$

となり（上の行列の対角成分以外（*印の部分）については最終的な結果に影響しないなめに計算を省略する），

$$\mathrm{trace}\left\{J(\hat{\Theta})^{-1}I(\hat{\Theta})\right\} = p + \frac{1}{2}\left\{\frac{n^2\cdot\hat{\mu}(4)}{(RSS)^2} - 1\right\}$$
$$= \frac{1}{2}\left\{2p - 1 + \frac{n^2\cdot\hat{\mu}(4)}{(RSS)^2}\right\} \quad (A.20)$$

と表せることから，この場合の TIC は

$$TIC(M) = -2 * l(\hat{\Theta}|Y) + 2\hat{t}$$
$$= -2 * l(\hat{\Theta}|Y) + 2p - 1 + \frac{n^2\cdot\hat{\mu}(4)}{(RSS)^2} \quad (A.21)$$

となり，ペナルティ項の評価が出来たことになる。

なお，式（A.21）は式（A.1）のような分散分析型のモデルに適用可能なことに注意する必要がある。

また，これと AIC の形

$$AIC(M) = -2 * l(\hat{\Theta} \mid Y) + 2(p+1) \quad (A.22)$$

を比較することにより，

$$TIC(M) > AIC(M) \Rightarrow \frac{n^2\cdot\hat{\mu}(4)}{(RSS)^2} > 3,$$
$$TIC(M) = AIC(M) \Rightarrow \frac{n^2\cdot\hat{\mu}(4)}{(RSS)^2} = 3,$$
$$TIC(M) < AIC(M) \Rightarrow \frac{n^2\cdot\hat{\mu}(4)}{(RSS)^2} < 3 \quad (A.23)$$

という関係式が成立していることが分かる。

この分散分析モデルでは，両者（AIC と TIC）のペナルティ項の値を比較することにより，精密評価の度合いを実際の標本を用いて数値的に検討することが可能である。

次に，理論的な側面から TIC と AIC を比較することを考える。

すなわち，TIC のペナルティ項の（真の分布 g に対する）期待値を考えると，

$$\mathrm{trace}\left\{J(\Theta)^{-1}I(\Theta)\right\} = p + \frac{1}{2}\left\{\frac{1}{(\sigma^2)^2}E\left[(y_i-\mu_i)^4\right] - 1\right\}$$
$$= \frac{1}{2}\left\{2p - 1 + \frac{1}{(\sigma^2)^2}E\left[(y_i-\mu_i)^4\right]\right\} \quad (A.24)$$

より

$$-2 * l(\hat{\Theta}|Y) + 2 * E_g[\hat{t}] = -2 * l(\hat{\Theta}|Y) + 2p - 1 + \frac{E\left[(y_i-\mu_i)^4\right]}{(\sigma^2)^2} \quad (A.25)$$

と表現出来る。これと AIC のペナルティ項の期待値を取ったもの

$$-2 * l(\hat{\Theta}|Y) + 2(p+1) \ (= AIC(M)) \quad (A.26)$$

との差を考えると，

$$E_g\,[TIC\,(M) - AIC\,(M)] = \frac{E\left[(y_i-\mu_i)^4\right]}{(\sigma^2)^2} - 3 = \beta - 3 \quad (A.27)$$

となる。このとき

$$\beta = \frac{E\left[(y_i-\mu_i)^4\right]}{(\sigma^2)^2} \quad (A.28)$$

は確率分布の中央での尖り具合や裾の重さ・長さを測る尖度（kurtosis）という指標になっている。よく知られているように，正規分布の尖度は 3 であることから，今回の分散分析モデルにおける TIC と AIC の差の理論値は 0 になることが証明される。よって，この場合には TIC と AIC は実用上もほとんど差がないと考えられ，本論文の計算機実験においてもそのような結果が示された（本文中の Table3-13, p.32）。

また，p 個の回帰係数を持つ（A.29）式のような重回帰分析モデル

$Y = X\,\Theta + \varepsilon$，　$\varepsilon \sim N\,(0,\ \sigma^2 I_n)$

（但し $\Theta = (\theta_1, \cdots, \theta_p)$ とする）　(A.29)

における TIC も今回の分散分析の場合とほぼ同じ手順にて計算することが可能であり

$$TIC(M) = -2 * l(\hat{\Theta}|Y) + 2p - 1 + \frac{n^2\cdot\hat{\mu}(4)}{(RSS)^2} \quad (A.30)$$

と表される。

従って，上と同様の尖度 β を用いた理論的な検討により，

$$E_g\,[TIC\,(M) - AIC\,(M)] = \beta - 3 = 0 \quad (A.31)$$

となり，AIC と TIC の理論値に完全に一致することが示された。

今回の付録 A で取り上げた統計モデルは分散分析と重回帰分析のみであるが，TIC のペナルティ項は上の意味での統計モデルには依存しないため，この論理を正規分布に従う他の統計モデル（時系列モデルなど）に拡張することは比較的容易である。ゆえに，正規誤差モデルにおける TIC の理論値は AIC のそれと同じであり，標本に基づく実測値に

ついてもほとんど差がないと考えられる。

なお，Poisson 分布，二項分布など他の確率分布に従う場合には

$$E_g\,[\mathrm{TIC}\,(\mathrm{M})\ -\ \mathrm{AIC}\,(\mathrm{M})] = \frac{E\left[(y_i-\mu_i)^4\right]}{(\sigma^2)^2} - 3 = \beta - 3 \tag{A.32}$$

となる保証がないため（実際には TIC の具体的な形を求めること自体困難な場合が多いと思われる），ここでの推論は適用出来ないことに注意する必要がある。

付録 B. 代表的な資源評価モデルの概略および標準化された CPUE が資源量推定結果に与える影響の例

本研究のメインテーマである CPUE 解析を利用して得られた標準化された CPUE 年トレンドは，資源の絶対量を推定するためのプロダクションモデル（Pella and Tomlinson，1969） や VPA（virtual population analysis）（Gavaris，1988）などにチューニングインデックスとして使用されている。ここでは，まぐろ類の解析に用いられている代表的な資源評価モデルの概略および標準化された CPUE が資源量推定結果に与える影響の事例について，簡単に記述する。

最初に，まぐろ類の資源評価モデルについて，使用可能な年別の漁獲データが年齢別に分かれていないか，分かれているかによって，２つに大別出来る。前者は logistic 曲線などに基づくプロダクションモデルが，後者は人工動態のコホート解析などにも使用されることがあるチューニング VPA が広く利用されている。なお，漁獲物の年齢査定は資源量推定に与える影響からして非常に重要であり，耳石などの生物情報を用いることが一般的であるが，そのような情報が利用出来ない場合には体長組成データにより年齢分解を行うことが必要であり，正規混合分布モデルなどが使用されている。本論文の3.6節では，体長組成の年齢分解を想定し，正規混合分布モデルの成分数推定を取り上げた。

まず，年別漁獲量によるプロダクションモデルは（B.1）式のように表される。

$$\frac{dB_t}{dt} = rB_t\left(1-\frac{B_t}{K}\right) - Y_t = rB_t\left(1-\frac{B_t}{K}\right) - qE_tB_t \tag{B.1}$$

但し，B・Y・r・K・q・E・t（添字）はそれぞれ資源重量・漁獲重量・内的増加率・環境収容力・漁具能率・漁獲努力量・時間（年など）を表す。

（B.1）式は，離散形の差分方程式（B.2）で表現されることも多い。

$$B_{t+1} - B_t = rB_t\left(1-\frac{B_t}{K}\right) - Y_t = rB_t\left(1-\frac{B_t}{K}\right) - qE_tB_t \tag{B.2}$$

漁獲重量や CPUE（=Y/E=qB）などに誤差構造を仮定し，最小二乗法や最尤法による目的関数最適化を行い，パラメーター（r, K, q など）を推定する。

以下，最小二乗法および最尤法（正規誤差を仮定）による目的関数例を示す。

$$SSQ = \sum_t \left(\log\frac{Y_t}{E_t} - \log(q\hat{B}_t)\right)^2 \searrow$$

$$LL = -\sum_t \left\{\frac{1}{2}\log(2\pi\sigma^2) + \frac{(Y_t - q\hat{B}_tE_t)^2}{2\sigma^2}\right\} \nearrow \tag{B.3}$$

次に，年別年齢別漁獲量によるチューニング VPA の定式化について述べる。

$$N_{a+1,y+1} = N_{a,y}\exp\{-(F_{a,y}+M_a)\}(a=0,1,\ldots,A-1)$$

$$N_{A+,y+1} = N_{A,y}\exp\{-(F_{A,y}+M_A)\} + N_{A+,y}\exp\{-(F_{A+,y}+M_{A+})\}$$

$$C_{a,y}(F_{a,y}) = N_{a,y}\frac{F_{a,y}}{F_{a,y}+M_a}\left[1-\exp\{-(F_{a,y}+M_a)\}\right]$$

（漁獲方程式） (B.4)

但し，C・N・F・M はそれぞれ漁獲尾数・資源尾数・漁獲死亡係数・自然死亡係数（漁獲以外の死亡）を表し，添字の a, y はそれぞれ age, year を表す。ここでは，漁獲量と資源量の関係が上の漁獲方程式でモデル化され，資源尾数の定式化では各コホート（年級群）の動きを Forward もしくは Backward に追っていることになる。このチューニング VPA の考え方を模式化すると，表 B.1のようになる。なお，年齢の添字 A+（プラスグループ）は，これ以上年齢査定が出来ない高齢部分をまとめた年級群を表し，プラスグループの全体に占める割合が高い場合には資源推定結果が頑健でなくなる傾向にあるため，このことからも年齢査定，体

長組成の年齢分解（3-6節）の重要性がうかがえる。

チューニングVPAでも，プロダクションモデルの場合と同様にして，観測されたCPUEとモデルから推定されたそれとの差異をモデル化し，最小二乗法や最尤法によりパラメーター推定することが多い（式B.5）。

$$LL=-\sum_{a}\sum_{y}\left\{\frac{1}{2}\log(2\pi\sigma^2)+\frac{(\log(CPUE_{a,y})-\log(q_aN_{a,y}))^2}{2\sigma^2}\right\}\nearrow \quad \text{(B.5)}$$

（B.5）式は，CPUE観測値に対して正規誤差を仮定した最尤法による目的関数を表し，（B.4）で漁獲尾数と資源尾数の関係を表している漁獲方程式における漁獲死亡係数F（Table B-1の濃色部分など）を未知パラメーターと設定することが多い。この場合には，コホートは高齢部分からBackwardに追っていくことになり（Table B-1の斜めに塗りつぶされた部分），プラスグループより1歳若い年齢Aとプラスグループ（Table B-1の薄色部分）の比α（（B.6）式）を推定，もしくは固定することにより，モデルの整合性が保たれる。

$$F_{A+,y}=\alpha_{(y)}F_{A,y} \quad \text{(B.6)}$$

（αは年に依存すると仮定することもあり，依存しないと設定することもある）

なお，プロダクションモデルとチューニングVPAの中間に位置する，年齢別漁獲尾数が存在しない場合に，選択率と呼ばれる年齢別の獲られやすさを仮定して年齢毎に漁獲を振り分ける，コホート解析モデルも考案されている。

例B.1. 標準化されたCPUEの資源量推定結果に与える影響（櫻田，2005）

最後に，本論文でも何回か取り上げた，インド洋キハダ資源のチューニングVPA解析例を通じて，標準化されたCPUEが資源量推定結果に与える影響について論じる。Fig. B-1はチューニングインデックスとして用いた2つの標準化されたCPUE年トレンド（日本のはえ縄船および台湾のはえ縄船による指標）を表し，Fig. B-2はこれらのCPUEシリーズと年別年齢別漁獲量に基づいて推定された資源尾数の絶対値を表す。2つの推定された資源量のトレンドは良く似ているが，その絶対値は多少違いがあり，解釈には細心の注意が必要である。

Table B-1. Pattern diagrams for tuned VPA（年齢 and 年 shows the age and year in the table, respectively）

年齢/年	1		y		m-1	m
0						
a						
A						
A+						

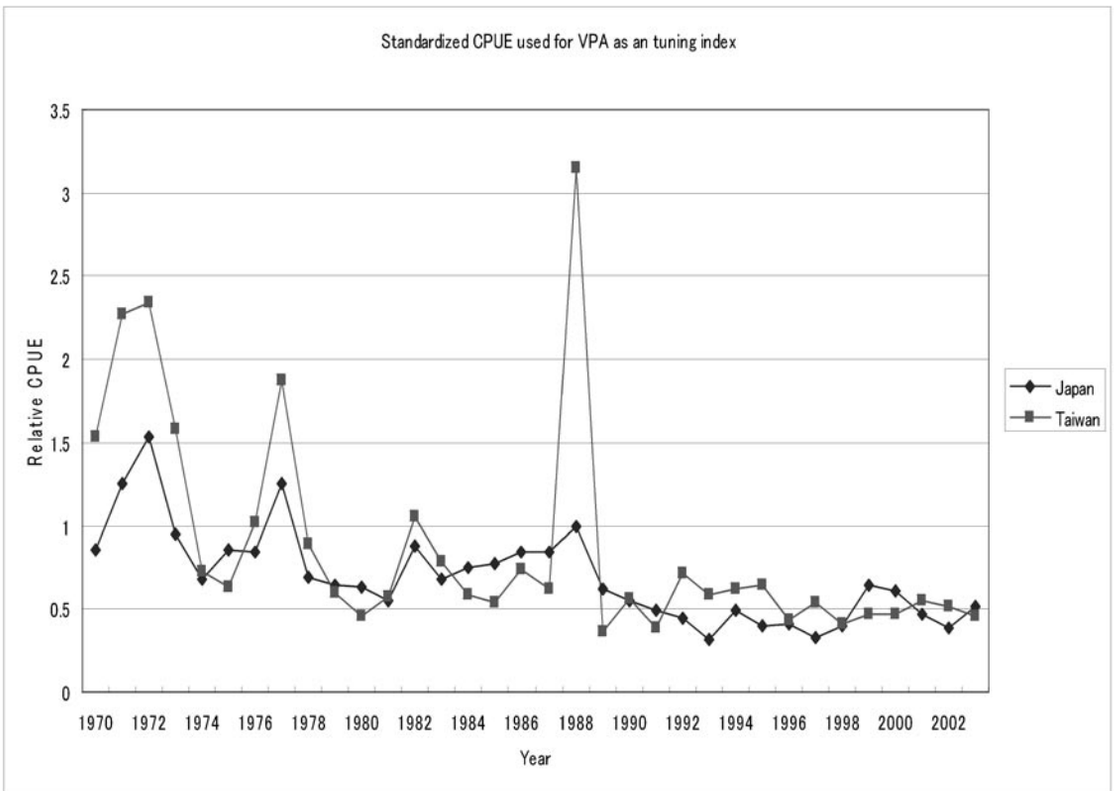

Fig. B-1. Year trends of standardized CPUE (for yellowfin tuna in the Indian Ocean caught by the Japanese and Taiwanese longline commercial vessels) as a tuning indices used for VPA.

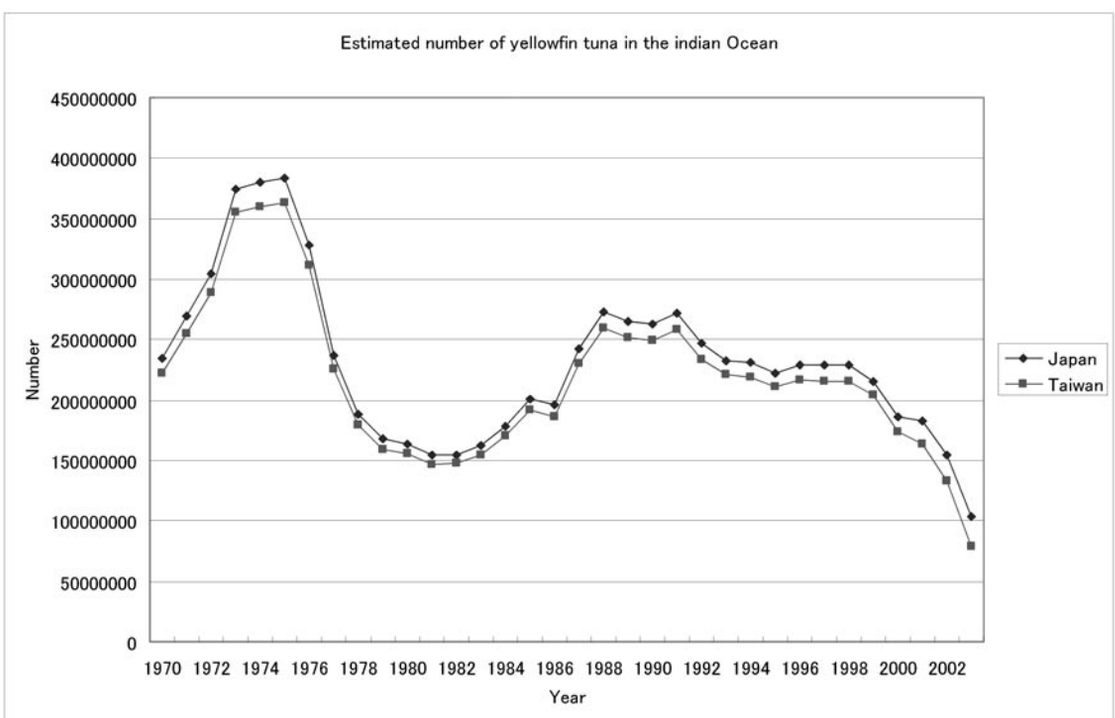

Fig. B-2. Year trends of estimated stock number (for yellowin tuna in the Indian Ocean obtained from the VPA) by the standardized CPUE in Fig. B-1.

統計モデルとデータマイニング手法の水産資源解析への応用

庄野　宏　（遠洋水産研究所）

本論文では，水産資源解析学における様々な問題，特に魚の資源密度に対応し，相対的な資源量を表すCPUE（catch per unit effort：単位努力当たり漁獲量）の解析に関する様々な問題について，遠洋域に生息するまぐろ類・関連種の漁業データや計算機によるシミュレーション実験を利用し，統計モデルおよびデータマイニング的なアプローチにより問題解決するための手法を提案した。CPUEは漁獲量を投下した努力量で割ることによって定義される，漁獲効率を表し，資源密度に比例することから相対資源量に対応する重要な概念である。しかし，漁船などの加工されていないCPUEは，季節・海区・漁具など資源密度以外の様々な時空間的な要因や環境要因などを含んでおり，資源の年変動を知るためにはこれらの影響を取り除く必要がある。そこで，CPUEの自然対数を応答変数に設定し，正規誤差の下で考えられる要因効果を説明変数に組み込んだ共分散分析モデル（CPUE Log-Normalモデル）や，離散変数であるCatchを応答変数と設定し，Poisson分布や負の二項分布などを仮定した一般化線形モデル（Catch Poissonモデル，Catch Negative-Binomialモデルなど）を用いて年の要因効果を推定することが伝統的に行われてきた。この作業をCPUE標準化と呼び，統計モデルに加えて近年では樹形モデルやニューラルネットワーク等のデータマイニング的なアプローチも用いられるようになってきている。本研究では，この水産資源解析における主要な問題であるCPUE標準化を論文のメインテーマとし，以下の3つの課題について取り上げて詳しく検討した。

1）CPUE標準化を想定した分散分析型モデルにおける，様々な情報量規準やstepwise検定を通じた要因効果の取捨選択，モデルの性能評価（第3章）
2）ニューラルネットワークによるミナミマグロの操業がない時空間のCPUE予測および簡便な要因分析法（CPUE年トレンド抽出法）の提案（第4章）
3）ゼロ・キャッチを多く含む場合の，Tweedieモデルの性能評価，および従来の手法（ad hocな共分散分析モデル・Catchモデル）との比較検討（第5章）

本論文の第1章は序論であり，研究の背景と目的，論文の構成を記述した。第2章では，CPUE標準化の現状について，統計モデル・データマイニング手法・漁業資源特有の問題に分けて整理し，レビューを通じて明らかになった問題点について，特に本研究で取り上げる3つの主要課題に関して，概説した。

第3章では，CPUE標準化に対応する一般化線形モデルを用いて，小標本の場合，大標本の場合など様々なケースを取り上げ，水産分野で広く知られている情報量規準AICの他に，BIC, CAIC, c-AIC, HQ, TICなどを使用し，実際の漁業データを用いて利用する情報量規準によってモデル選択結果が異なること，および複数の候補モデルの中から定めた真のモデルから乱数を発生させて正しいモデルを選ぶという選択パフォーマンスをシミュレーションにより計算し，情報量規準の良さを評価した。なお，ネスト構造モデルでは，カイ二乗検定やF検定などのstepwise検定も使用可能であり，計算機実験を通じて情報量規準とstepwise検定の性能を比較した。この変数選択の問題は，CPUEに影響を与えている要因効果を統計的に取捨選択するという意味において重要であるが，使用する情報量規準やstepwise検定によるモデル選択結果が，推定されたCPUE年トレンドという要因分析結果の違いを引き起こし，これらをチューニング指標として組み込んだモデルでの資源の絶対量推定結果の大きな差異となることもあり，極めて本質的な問題であると考えられている。なお，本章の具体的な研究成果は，次の通りである。

- 小標本の場合や未知パラメーター数の標本数に占める割合が高い場合に，AICに有限修正を施した規準であるc-AICによるモデル選択結果がAICなどによるそれと異なることを例示し，さらに分散分析型のシミュレーションを通じて，c-AICの選択パフォーマンスがAICのそれに比べて高くなることを証明した。
- 大標本の場合にAICが偏りを持つ可能性があることを示し，使用する規準により選択結果に差が生じること，および漸近的に望ましい性質である一致性を持つ情報量規準（BIC, HQ and CAIC）がAICに比べて全体として優れていることを，それぞれ漁業データによる実例および回帰分析型の実験により示した。合わせて，HQにおける定数項cの検討を行い，推奨値と推奨式を提案した。
- ネスト構造を持つモデルにおいて，従来性能が良いと言われてきたAICの精密評価であるTICが正規誤差を持ちかつ連結関数が恒等写像であるような一般化線形モデルではAICと同等になることを理論的に証明し，合わせてTICとAICの選択パフォーマンスにはほとんど差がないことを，

計算機実験により示した。

- ネストモデルにおいて，計算機実験により情報量規準と stepwise 検定の比較を行い，一般に前者が多少優れていること，後者で有意水準を小さく設定した場合にパラメーター数が少ない単純なモデルが選ばれがちであることを示した。

第4章では，ミナミマグロ資源における CPUE 解釈の問題，すなわち操業がない時空間の CPUE 予測の問題を取り上げて，ニューラルネットワークを利用した解析を行った。CPUE を相対資源量の観点から捉えた場合，標準化された CPUE に相対的な面積指数を掛け合わせたものとして考えることが自然であり，これを資源量指数（AI: abundance index）と呼んでいる。ミナミマグロ資源では過去から現在にかけて漁場が縮小しており，このような過去に漁獲があり現在操業がないセルの CPUE をどのように設定するか，極論すれば周囲と同じと考えるかそれとも 0 と仮定するかが資源量指数の計算に影響してくる。ひいては，資源量指数から得られた CPUE 年トレンドの違いとなって表れる。そこで，本論文では，このような欠測セルの CPUE を教師付きニューラルネットワークの代表的なアルゴリズムである誤差逆伝播法を用いて予測を行い，合わせて得られた予測値から CPUE 年トレンド抽出を行うための簡便な要因分析手法を提案した。ニューラルネットワークの精度評価のために，クロス・バリデーションにより同じ条件での MCMC 法に基づく EM algorithm との比較を行った。n-fold cross-validation により観測値と予測値の相関係数および MSE（平均二乗誤差）に基づき，モデルの性能評価および比較検討を行った。結果として，ニューラルネットワークによる CPUE 予測値に基づく，操業がないセルの CPUE と操業が行われたセルの CPUE 比は，0.8～1 前後を推移しており，1998年から2000年にかけて局所的に行われた日本の調査漁獲における CPUE 比（年，季節，エリアは非常に局所的であるが0.7前後を記録）と比べ極端な矛盾は見られない。また，ニューラルネットワークによる CPUE の予測性能は，全く同じ条件での解析である MCMC 法に基づく EM algorithm によるそれよりも格段に高く，CPUE 予測値を元に算出された CPUE 年トレンドは一般化線形モデル（共分散分析）によるそれと比較的良く似ていた。このことから，ニューラルネットワークの予測性能の良さ，および提案した簡便な要因分析法の妥当性が言える。

第5章では，まぐろはえ縄漁業で混獲されるサメ類などを想定し，ゼロ・キャッチ問題と呼ばれる漁獲がゼロであるデータが含まれる場合に，CPUE の自然対数を取ったものを応答変数とする共分散分析モデルが使用出来ない問題について，詳細に議論した。Tweedie 分布と呼ばれる，複合 Poisson 分布の拡張であるゼロ・データを統一的に取り扱えるモデル使用し，ゼロ・キャッチ率が10% 程度と低い日本のはえ縄商業船によるインド洋キハダ資源の CPUE 解析，およびその割合が80% 以上と高い日本のはえ縄公庁船による北太平洋クロトガリザメ資源の CPUE 標準化を行った。実際には，Tweedie モデルと全ての CPUE に定数項を加える ad hoc な共分散分析モデル，Catch Negative-Binomial モデルに基づく CPUE 年トレンドを比較した。その結果，ゼロ・キャッチ率が低いターゲット種のインド洋キハダ資源では Tweedie モデルと ad hoc な方法で年トレンドに極端な違いが見られなかったのに対し，ゼロ・キャッチ率が高い混獲種の北太平洋クロトガリザメ資源では，Tweedie モデルからの CPUE 年トレンドが，Catch モデルや ad hoc な方法からのトレンドと異なっていた。また，ニューラルネットワーク解析と同様に，n-fold validation を利用した観測値と予測値の相関係数や MSE に基づくモデルの性能評価を行ったところ，いずれの例においても，両方の指標に関して Tweedie モデルの精度が良かった。クロス・バリデーション結果から判断すると，ゼロ・キャッチ率が低い場合には Tweedie モデルの有意性が顕著に表れず，ad hoc な方法を使用しても実用上さほど問題が生じないと考えられる一方，ゼロ・キャッチの割合が高い場合には，Tweedie モデルの精度が他に比べて非常に高くなり使用が推奨される。なお，ゼロ・キャッチ率が高い場合には，相関係数が Tweedie モデル，Catch モデル，ad hoc な方法の順，MSE は Tweeide モデル，ad hoc な方法，Catch モデルの順に優れていたが，ad hoc な方法では観測値の大きさにかかわらず予測値が極端に小さくなることもあり，バイアスの大きさを考慮すると，サメ類などのゼロ・キャッチ率が高い場合には適用すべきでない，と結論付けられる。

最後の第6章は，本論文の結論部であり，今回取り上げた3つの課題に関する研究成果について，水産資源学の観点から，および応用統計学の視点からに分類して再度系統的に整理し，合わせて，今後の研究課題について記述した。