1B2-OS-02a-3in

機械学習による海洋観測データの良否分類に向けた初期検討

A Preliminary Study on Error Detection of Oceanic Observation Data by Machine Learning

松山 開[*1] 小野 智司[*1] 福井 健一[*2] 細田 滋毅[*3]
Haruki Matsuyama Satoshi Ono Ken-ichi Fukui Shigeki Hosoda

[*1]鹿児島大学大学院 理工学研究科 情報生体システム工学専攻
Department of Information Science and Biomedical Engineering, Graduate School of Science and Engineering, Kagoshima University

[*2]大阪大学 産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

[*3]独立行政法人 海洋研究開発機構
Japan Agency for Marine-Earth Science and Technology

Argo, a global ocean monitoring system for climate change, consists of more than 3,000 floats located in the global oceans and is operated by over 30 countries. Every 10 days, the Argo floats produce temperature and salinity data at a depth from 2,000m to the surface of the sea. However, it was inevitable to observe the ocean without any errors due to substance adhesion, sensor failure and other reasons. The goal of this study is to propose a method for error detection of the observation data by the floats, which has been performed by a human expert. Before designing the error detection method, this paper surveys the cases corrected by the expert and comprehensively understands the property of the observation data so that appropriate machine learning models and features are revealed for error detection.

1. はじめに

異常気象の一因とされる気候変動のメカニズムは十分に理解されていないが，海が変動の駆動源と考えられている．これは，地球上のおよそ 7 割を占める海水が大気の 1,000 倍以上の比熱を持ち，大気の状態を大きく変化させるためである．海洋の変動を把握するためには，全世界の海洋内部をくまなく継続的に観測する必要があるが，これまでの船舶観測で実現することは難しかった．

これを受け，2000 年より海洋観測システム「アルゴ」が稼動している．これは，「アルゴ計画」のもとで運営される，全球観測データをリアルタイムで取得することを目的とした国際プロジェクトであり，全世界で 30 カ国以上が参加している [Argo Science Team 01]．全球アルゴ観測網を実現するために，アルゴフロートと呼ばれる水温・塩分を計測可能な自動昇降型海洋観測ロボットを海へ投入し，衛星経由でデータを取得することにより実現している．アルゴで得られた観測データは，品質管理を施した後にインターネットを通じて公開される．現在，3,500 台以上のアルゴフロートが常に稼動しており，大量の海洋データの蓄積に成功している．このプロジェクトにより，従来知り得なかった地球規模の変動が捉えられ，気候変動などのメカニズム解明に向けて研究が進められている．

アルゴフロートの観測サイクルを図 1 に示す．アルゴフロートは 10 日間隔で，水深 2,000m 付近から水温と塩分を観測しながら浮上する．1 回分の浮上によって生成される観測データはプロファイルと呼ばれ，プロファイルには各観測層における水温および塩分が記録される．

自動観測されたデータは，予期しないエラーを含むことがある．アルゴフロートの場合，その要因を特定することは一般的に困難であるため，観測値の信頼性を示すラベルが導入されている．このラベルの割り当ては，国際アルゴ計画で決められた品質管理手法に則り，各国のデータ管理チームによって行わ

連絡先: 松山開，鹿児島大学大学院 理工学研究科 情報生体システム工学専攻，〒 890-0065 鹿児島市郡元 1-21-40，k4391399@kadai.jp

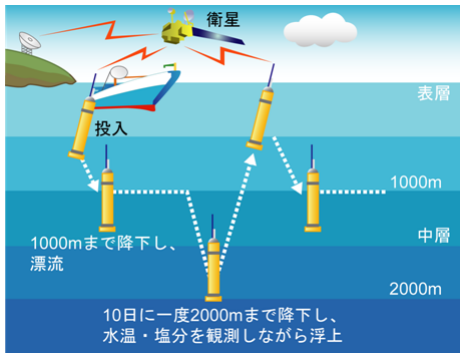

図 1: アルゴフロートの観測サイクル

れている [Argo Data Management Team 02]．しかし，海面付近で水温や塩分といった観測量は，時間や天候などの影響により激しく変動するため，観測量に対するデータ品質管理の十分な自動化手法が確立されておらず，自動的な品質管理ではエラーの見落としや誤検出が発生している．このため，現在の品質管理では最終的に専門技術者が目視で確認を行い，手動で補正を行わなければならい．また，技術者による補正が困難な国もあり，全球データの品質の均一性が担保されない問題も生じる [細田 13]．これらは国際アルゴ計画における長年の大きな課題であり，全球海洋環境モニタリングの精度・信頼性に関わるほど重大である．

本研究では，データ品質管理の専門技術者が目視および手動で行っているエラーの検出および補正を自動的に行える方式の実現を目指す．アルゴフロートの水温・塩分センサ値（以下，アルゴデータ）の誤差検出・指標決定ならびに補正に対し，機械学習を応用する．上記のエラー検出方式の実現に向け，本稿では対象となるアルゴデータに関する基礎的な検討を行う．まず，専門技術者によって補正が行われたデータの事例を観察し，密度逆転とオフセットの 2 種類のエラーについて特徴を調査する．次に，自己組織化マップ（Self-Organizing Map: SOM）[Kohonen 95] を用いてエラーを含むプロファイルのクラスタ

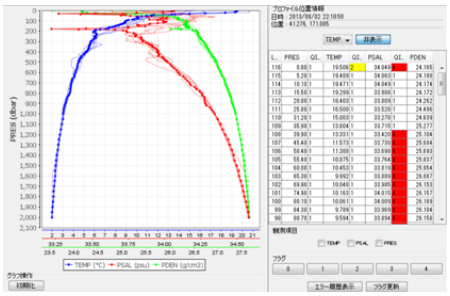

図 2: 密度逆転の例 1

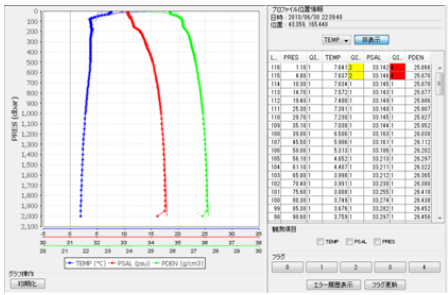

図 3: 密度逆転の例 2

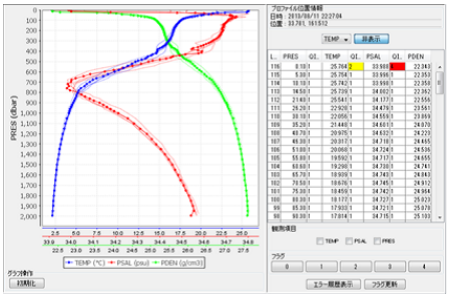

図 4: 密度逆転の例 3

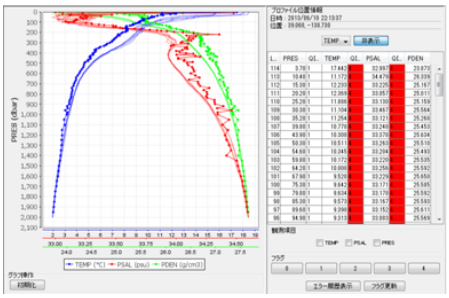

図 5: 密度逆転の例 4

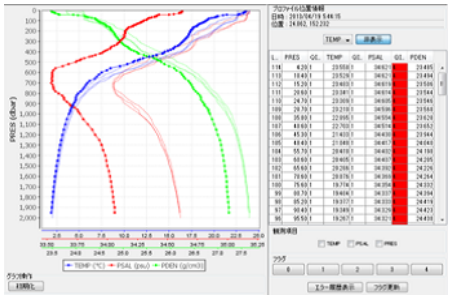

図 6: 塩分センサ異常例 1

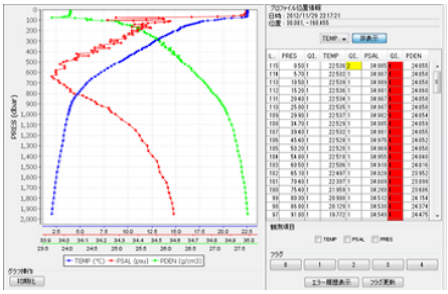

図 7: 塩分センサ異常例 2

リングを行うことで，アルゴデータ全体の特性を俯瞰し，エラー検出方式で利用する機械学習モデルや属性，素性の設計に資する知見を得る．最後に，対象問題の難しさ，および，エラー検出方式における要求事項について考察する．

2. 問題の概要と特徴

2.1 品質管理の現状

アルゴデータに含まれるエラーは，ハードウェアやソフトウェアに起因するもの，外的要因によって発生するセンサ汚濁やデータ受信の不具合に起因するものなどがある．エラーパターンはそれぞれ異なるものの，傾向が見える事象もいくらか存在するため，自動的に品質を管理する手法が定められている．しかし，水温や塩分といった観測量は，海面付近で天候などの影響により激しく変動するため，詳細な管理手法を記述することは難しく， 既存の自動品質管理手法で全てのエラーに対応することは技術的に困難である． また，目視確認を行う技術者のスキルも各国でばらつきがあり，人的資源も限られている．このような理由により，全球アルゴ観測網にとって重要なデータの均一性を担保することが難しい．日本のアルゴデータの高精度品質管理を受け持っている海洋研究開発機構では，アルゴ計画開始以降 130,000 プロファイル以上を取得し，このうち約 90,000 プロファイルについて品質管理を実施している．年間約 10,000 プロファイルの品質管理を施しているが，実際にエラーを検出する割合は 10〜20 %に相当する．エラーが検出されたプロファイルについては，具体的な状況とエラーの起こっている観測層を特定するため，目視確認を行われる．

2.2 エラーの分類と特徴

エラーの出現パターンは多様であり，できる限り多くエラーに対応できるよう，自動品質管理には多数のエラー検出規準が含まれる．これらの検出過程には，過去のプロファイルや周辺のアルゴフロートのアルゴデータを参考にして判断するような複雑な条件もある．ここではエラーの頻度が高い密度逆転と，プロファイル全体に影響しデータの品質を大きく左右する塩分センサ異常について着目する． 本稿では，解析の第一歩として，比較的自然変動の影響が小さくデータが十分揃っている 400m〜1400m 間の観測層について取り扱う．

2.2.1 密度逆転

密度は，水深，水温，塩分によって決定され，海域によらず深度とともに単調増加する．そこで，海面から 2,000m まで 100 層以上の観測層のうち，ある閾値より大きな鉛直密度の重軽の関係が逆転する場所を検知することでエラーを検知し，その深度を特定する．この逆転が 1 層のみであれば自動検知は可能であるが，複数層にわたる逆転の場合，エラーが起こっている状況を特定出来ないため，すべての可能性を網羅した自動検知は困難である．密度逆転は，水温または塩分のどちらか一方の観測不良によって引き起こされることが殆どであり， 上下層の値の関係で決まる．実際にどの層のどちらの観測値がエラーを含むか特定するためには，専門技術者による目視確認が必要となる．一方で，自然現象に伴う変動が含まれるため，観測不良と自然現象を切り分ける必要がある．

密度逆転によりエラーが生じた例を図 2〜5 に示す．青色のグラフは水温 [℃]，赤色は塩分 [PSS-78]，緑色は密度を表す．図 2 では，自動検出により密度逆転が生じていると判断され，水温，塩分ともにエラーとしてラベルが割り当てられたが，水温は過去のプロファイルから見て正常であることが確認できたため，目視補正により正常値を表すラベルに補正されている．

また，密度逆転は検知された観測層によってアルゴデータの信頼性が異なる．図 3, 4 では，ともに水深 2,000m 付近で密度逆転が発生しているが，図 4 は閾値以下の範囲であるため自動検出されない．しかし，海流や天候などの影響を受けやすい海面付近では，閾値を超えても正常なラベルに補正されることもあり，反対に 2,000m 付近の安定した観測層において密度逆転が閾値内であったとしても，補正によりエラーに変更される．2,000m 付近の観測層で密度逆転が検出された場合，全観測層がエラーに補正されることがある．

2.2.2 塩分センサの異常によるエラー

プロファイルのなかには，密度逆転は生じていないものの，過去のデータや近傍データと比較したとき，プロファイル全体が平行移動したようなエラー（オフセット）が見られることがある． エラーの例を図 6, 7 に示す．これは，観測層全体に現れる場合や，深層のみに現れる場合，投入当初から持続しているアルゴフロートに突発的に起こる場合など多岐にわたる．このエラーの補正は比較的困難であるが，過去や近傍のプロファイルとの比較により解決できる．

3. SOM を用いた解析

3.1 SOM の概要

SOM[Kohonen 95] は，Kohonen により提案された教師なし学習を行うニューラルネットワークであり，多次元データの分類・解析に効果的な手法として知られている．入力となる多次元データを 2 次元空間に非線形写像することにより，多次元データの分布を 2 次元平面上で可視化できる．類似した特徴を持つパターンはマップの近い位置に配置され，類似しないパターンは遠い位置に配置される．これにより，入力データを類似度に応じて自動的に分類するクラスタリングの分野で注目されている．

本稿では，アルゴデータに SOM を適用することでデータ全体を俯瞰し，今後採用する機械学習モデルや属性・素性の検討を行うための手がかりを得ることを目的とする．

3.2 実験条件

アルゴデータは海域ごとに傾向が異なり，また，国ごとによって品質管理の精度が異なる．このため，世界最高の水準で品質管理を行っている日本で，専門技術者による目視確認および補正が行われたアルゴデータを対象とした．より具体的には，北太平洋のうち以下の海域で観測されたプロファイル（約 500 個）を使用した．

- 10N-30N,140E-120W(240E)
- 30N-40N,150E-130W(230E)
- 40N-50N,155E-135W(225E)

プロファイルには，水深 0〜2,000m の水温・塩分値が保存されているが，プロファイルごとに観測層が異なる．データを均一化するために，観測層間で水温・塩分値を線形的に補完し，水深 5m 間隔での観測データとなるように加工した．また，1,500m 以深で一部のプロファイルに欠損が生じており，海面付近では大きな変動が含まれることから，水深 400〜1,400m を対象として解析を行った．プロファイル全体の俯瞰には，標準的な六角格子を持つバッチ学習型 SOM を用いた．近傍関数はガウス関数，近傍半径は減少戦略とし，学習変数である参照ベクトルはランダムに初期化した．SOM の学習結果の可視化には，標準的な U-matrix 表示 [Ultsch 93] を用いた．

3.3 実験結果

SOM の出力層を 10×10 として実行したときの，水温における補正前後の SOM の出力を図 8，9 に，塩分における補正前後の SOM の出力を図 10，11 にそれぞれ示す．上記の図において，マップ上のグレースケールの濃淡により，データ空間上の近さを表現している（淡い：近い，濃い：遠い）．また，エラーを扱うことから，代表ラベルは 1：良, 2：おそらく良, 3：おそらく否, 4：否として，SOM の各マイクロクラスタに属するデータのラベルのうち，最も悪いラベルで代表させた．赤色の円は補正によりラベル値が変更された箇所を示す．

いくつかのクラスタに分かれたものの，自然現象による変動成分が大きく，エラーの種類を反映した分類とは言い難い．しかしながら，マップの赤丸で示した部分では，補正前後でラベル値が変更されたクラスタが得られている．

図 8 および 9 において，補正前後で良否が反転された 2 つのクラスタはいずれも同様の傾向を示した．上記クラスタに含まれるデータ例を図 12 に示す．図 12 において，橙色のグラフは補正前のラベル値，紫色のグラフは補正後のラベル値を表す．750〜800m 付近および 950〜1,050m 付近の観測層において，塩分の観測値が異常に低いものの水温は正常であり，目視品質管理により水温の品質監視ラベル値が良に変更されたことがわかる．上記のクラスタ内の他のプロファイルにおけるラベル値補正は，密度逆転例 1 （図 2）に示した種類であり，塩分の観測不良にのみ起因していたため，すべてのプロファイルにおいて水温のラベルが 4(否) → 1(良) へ変更されていた．また，水温マップ上に，1(良) → 4(否) への補正が行われているプロファイルの存在を確認できるが，これらは密度逆転とは別の要因で補正が行われていた．

図 10，11 に示される塩分のクラスタリング結果においても，3 つのクラスタは同様の傾向を示した．上記クラスタに含まれるデータ例を図 13 に示す．1,300〜1,400m の観測層において塩分が異常値を示しているため，目視品質管理によりラベル値が 3（おそらく否）→ 4（否）に変更されたことがわかる．いずれのプロファイルも密度逆転例 1 に示した種類のエラーであり，自動品質管理で使用された閾値に近い密度逆転が発生していたため，目視補正によりエラーと判定されていた．

以上のように，SOM によって示されたクラスタにおける補正の主要因とその詳細から，自動品質管理において適切な閾値の設定が困難であることがわかる．

4. 考察

3 章に示した実験結果から，水温および塩分といった直接的な観測データを SOM に入力することで，一部ではエラーに関するクラスタも得られたものの，当初目的であった多様なエラーの全貌を俯瞰することは困難であった．本実験で対象とした 400〜1,400m までの観測層の中腹付近はエラーが比較的少なかったことも，SOM による解析対象として適切ではなかった．以上の理由から，SOM の出力結果には自然変動や海域の依存性が強く反映されてしまったと考える．本実験では，2.2 節に示した理由により，上記の条件で解析を行ったが，水深 1,500〜2,000m の観測データを対象とする，あるいは，過去のプロファイルにおける平均値との差分や勾配などを加味することで，エラーの傾向を可視化できると考える．

一方で，エラーを含むプロファイル全てを目視で確認したところ，水温，塩分ともに 8 割以上の補正が密度逆転に関連しており，残り 2 割がオフセットに関していたことを確認した．このため，それぞれのエラー検出を正確に行う分類器を構築することで，技術者の負担を軽減できると考える．

プロファイルは水深方向および時間方向の時系列データである．また，品質管理ラベルもプロファイルに対し，深さ方向の系列ラベルとして付与されている．よって，アルゴデータの良否判定問題は，機械学習における系列ラベリング問題と捉えることができ， 条件付き確率場 (Conditional Random Field: CRF)[Lafferty 01] などを適用することで，専門技術者による補正の一部を自動化できると考える．

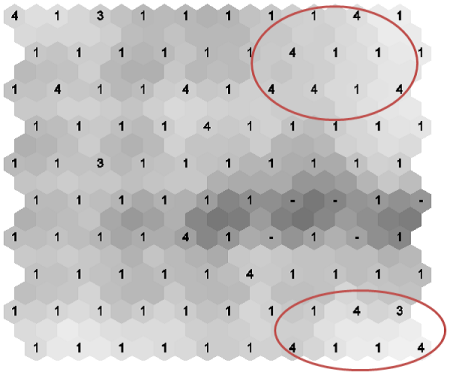

図 8: 水温補正前

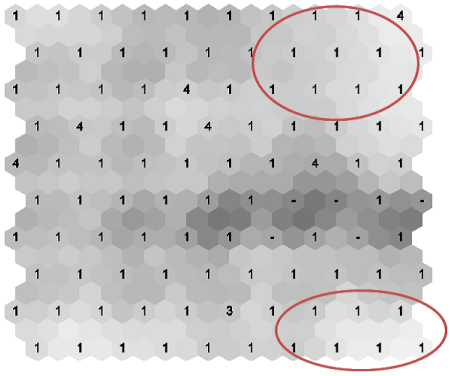

図 9: 水温補正後

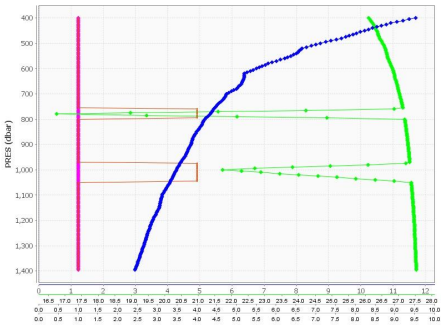

図 12: 水温マップのクラスタの例

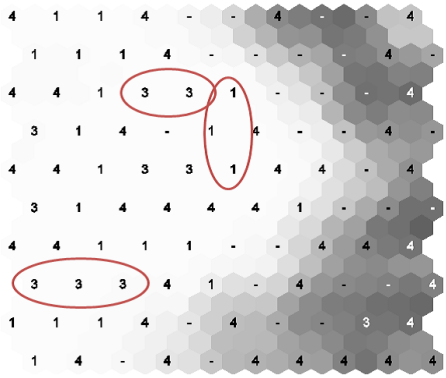

図 10: 塩分補正前

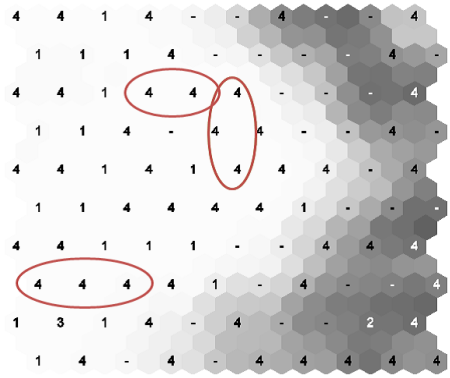

図 11: 塩分補正後

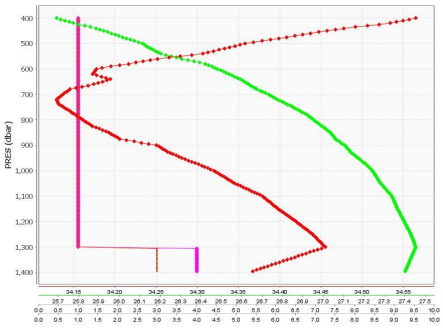

図 13: 塩分マップのクラスタの例

5. おわりに

海洋観測システム「アルゴ」におけるアルゴデータの品質管理の問題点に着目し，機械学習を用いたエラー検出方式の設計の前段階として，機械学習のモデルや素性の設計に関して有用な知見を得るために，対象データの分析を行った．すなわち，エラーの種類を大別し，SOM を用いてエラーパターン全体の俯瞰を試みた．専門技術者による補正は 8 割以上が密度逆転に関して行われていることが確認できたものの，SOM による分類の結果から有用な知見を得ることは困難だった．これは，自然現象の変動が大きいために微小なエラー成分が埋没したことが理由として考えられ，過去のプロファイルを加味した上で，適切な加工を施した入力を属性や素性として利用する必要があることを示している．

今後は，アルゴの品質管理におけるプロファイルの良否判定問題を系列ラベリング問題として捉え，CRF を用いた機械学習の適用を検討する． CRF を適用するにあたり，自然現象による変動に依存せず，エラーの特徴を捉える素性関数を設計する必要がある．今回の実験から，自然現象による変動成分は時空間上で連続性を持っていることと，物理現象の制約を受けていることが特徴として挙げられる．エラーはそれらからの逸脱と考え，素性関数の設計を検討したい．

謝辞

本研究を進めるにあたり，独立行政法人海洋研究開発機構・地球環境変動領域・アルゴデータ班に協力頂いた． また，本研究の一部は，倉田記念日立科学技術財団 倉田奨励金によるものである．ここに記して感謝の意を表する．

参考文献

[Argo Science Team 01] Argo sicence team, Argo: The global array of profiling floats, in Observing the Oceans in the 21st Century, edited by C. J. Koblinsky and N. R. Smith, pp. 248–258, GODAE Project Office, Bureau of Meteorology, 2001.

[Argo Data Management Team 02] Argo Data Management Team, Report of the Argo Data Management Meeting. Proc. Argo Data Management Third Meeting, Marine Environmental Data, 2002.

[細田 13] 細田 滋毅, 全球海洋監視システム「アルゴ」, 人工知能学会第 27 回全国大会, 3K1-OS-08a-1, 2013.

[Kohonen 95] T.Kohonen; Self-Organizing Maps, Springer-Verlag:Berlin, 1995.

[Ultsch 93] Ultsch, A., Self-organizing neural networks for visualization and classification, in: Lausen, O.B., Klar, R. (Eds.), Information and Classification- Concepts, Methods and Applications. Springer Verlag, Berlin, pp. 307-313, 1993.

[Lafferty 01] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proc. Int'l Conf. Machine Learning, 2001.