Vol.2016-MPS-108 No.23
Vol.2016-BIO-46 No.23
2016/7/5

クラスタリングによる海洋データの構造視覚化

林 勝悟[1,a)]　細田 滋毅[2]　小野 智司[3]　沼尾 正行[4]　福井 健一[4]

概要：世界中の海洋で水温や塩分といった海洋データが測定されている．時にエラー測定値が測定されてしまうこともあるが，様々な要因に由来する海洋データの非線形な自然変動とエラー測定値を識別することは難しい．海洋データのエラー測定値の高精度な自動検知を最終的な目的として，本研究では初期的な検討として，クラスタリングによって海洋データの視覚化を行う．得られた知見から，エラー測定値検知のモデル構築に組み込むべき要素の検討を行う．

キーワード：クラスタリング，系列データ，時空間データ

Structure Visualization of Oceanic Data using Clustering

Hayashi Shogo[1,a)]　Shigeki Hosoda[2]　Ono Satoshi[3]　Numao Masayuki[4]　Fukui Ken-ichi[4]

Abstract: In the ocean around the world, oceanic data such as temperature and salinity is being measured. Error data is also sometime measured, but discrimination between error data and normal data that vary non-linearly because of natural factors is difficult. For the final goal to realize automated high-quality error detection, in this research, we visualize the structure of oceanic data using clustering. Utilizing the knowledge of the result, we consider what the appropriate model for error detection is like.

Keywords: clustering, sequential data, spatio-temporal data

1. はじめに

近年，干ばつや洪水，大型の台風といった異常気象が世界各地で発生しているが，その一つの大きな要因として海洋が挙げられる．海水は大気の約 1,000 倍の比熱を持つため，地球の約 7 割の表面積を持つ海洋が，面した大気に熱を供給して空気の流動を引き起こすためである．従って，海洋の継続的な観測が，長期スケールで起こる地球規模の気候変動のメカニズムの解明やその予測に繋がる．

このような背景から，地球全体の海洋内部の常時観測を目的とした国際プロジェクト「アルゴ (Argo) 計画」が 2000 年に始動され，現在 30 カ国以上の国と気象・海洋機関によって進められている．3,500 以上の海洋観測センサ「アルゴフロート」が世界中の海洋に展開され（図 1），水深約 2,000m まで下降した後，上昇しながら水温・塩分・圧力（深度と同義として扱える）を測定する（図 2）．一回の上昇によって測定された深度方向の系列データは「プロファイル」（図 3）と呼ばれ，測定場所（緯度・経度）や測定日時と共に衛生を介してデータセンターへと送られて，品質管理を受けた後，インターネット上にデータが公開される．

品質管理は，漂流物がアルゴフロートのセンサに付着したり，センサの一時的な故障によって誤って測定されたエラー測定値を取り除くために行われる．長期的には僅かにしか変化しない海洋データの研究において，でたらめな値を持つエラー測定値はその研究結果に大きく影響する恐

[1] 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
[2] 国立研究開発法人海洋研究開発機構
Japan Agency for Marine-Earth Science and Technology
[3] 鹿児島大学大学院理工学研究科
Graduate School of Science and Engineering, Kagoshima University
[4] 大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University
[a)] hayashi@ai.sanken.osaka-u.ac.jp

Vol.2016-MPS-108 No.23
Vol.2016-BIO-46 No.23
2016/7/5

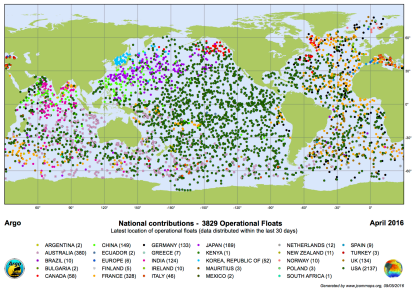


図 1　アルゴフロートの分布
Fig. 1　Distribution of Argo floats.

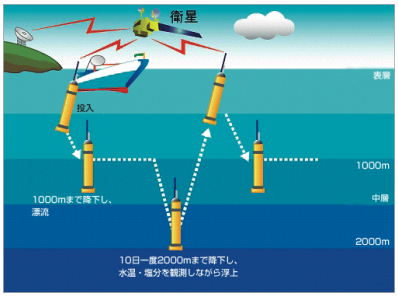


図 2　アルゴフロートによる測定
Fig. 2　Measurement of Argo float.

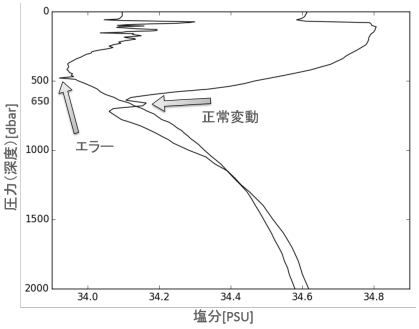


図 3　プロファイルとエラーの例
Fig. 3　Example of profile and fault value.

れがある．実際に，エラー測定値を含んだデータの分析によって，大西洋の温暖化に関して誤った研究結果が報告された例がある [5]．気候変動に関する研究結果は政策決定にさえ影響を及ぼしうるため，アルゴデータのエラー測定値を除去する品質管理は社会的にも非常に重要な役割を持つ．

アルゴプロジェクトでは現在二段階の品質管理手法 [4] を行なっている．一つは即時品質管理であり，ある海域の正常測定値の上限や下限を定めるルールなど，海洋技術者の専門知識と長年の経験値に基づいたヒューリスティックルールがアルゴデータに適用される．しかし，時間や海域によって動的に変化する海洋データから単純なルールのみでエラー測定値のみを全て検出することは困難である．もう一つは遅延品質管理であり，海洋技術者が目視でプロファイルを確認してエラー測定値を検出する．専門知識や長年の経験を活かして臨機応変にエラー測定値を検出することができるが，日々刻々と測定される大量のデータに人手で全て対応しないといけない．そのような人的リソースへの依存性から，技術者の検出基準の違いによるデータ品質の均一性や，そもそも遅延品質管理を行えない機関が存在するといった問題も存在する．これらはアルゴ計画の長年の大きな課題 [6] であり，全海洋の観測の精度や信頼性に大きく関わる．そのため，アルゴ計画では自動で高精度なエラー測定値検出手法が切望されている．

本研究では，海洋データの深度系列のエラー測定値検知を行うために，著者らはその系列性に着目した．海洋データは，海域, 季節, 大気場の変動や潮の満ち引きなど多様な自然要因に由来して値が動的に変化するため，測定値自体の予測が難しいが，深度方向に密に並んでいるため，検知を行いたい測定値をその前後層の値と比較することによって，エラー測定値を効率的に識別できる可能性がある．しかし，深度系列の相関性も，海域や季節に強く依存する．その例として，図 3 に二つの海域における塩分濃度の深度系列を示している．両者とも圧力値 500,650 付近で前後層と比較して急激な変動を起こしているため，エラーであると判断されそうだが，右側矢印で示した点は正常の自然変動である．このような差異を見分けるためには，深度・緯度・経度・季節によって，どのような海洋データの変動構造が存在するのかを知る必要がある．海域毎の変動構造を理解できれば，その特性に合わせて適切な検知モデルを構築することができる．そのため，本研究では，海洋データの適切なエラー測定値検知モデルを構築するための初期検討として，海洋データの深度系列に対してクラスタリングを行うことによって，深度・緯度・経度・季節による変動構造の調査を行った．

2.　関連研究

小野ら [2] は一プロファイルを深度による系列データとみなして，条件付き確率場によるエラーラベルと正常ラベルの系列ラベリングとして問題を定式化した．複雑に変動する海洋データの深度系列の測定値を一点ずつ独立として扱わずに，前後層のラベルの依存関係を無向グラフによって表現することで，系列情報を活用した．北太平洋域の特定のエラー種に限定した実験結果として，プロファイル単位の分類において 95 %を超える精度と再現率を出している．また，条件付確率場の素性関数には即時品質管理のヒューリスティックルールが活用されているため，即時品質管理手法を条件付き確率場によって系列情報を扱えるように拡張したモデルとも解釈することができる．

しかし，条件付き確率場は教師あり学習であるため，教師データに含まれないエラーの種類への対応に疑問が残る．実際に小野らは，教師データ数が少ない，エラーラベルが深度系列に連続して出現するエラーの検出精度を課題として挙げている．

3.　クラスタリングによる海洋データの構造視覚化

本研究では，海洋データの深度系列に対して，クラスタ併合を打ち切る閾値を変化させることによって，局所的な変動や全域的な変動を捉える階層的クラスタリングを適用した．得られたクラスタにおける測定位置や季節を調べることによって，海域毎の変動構造について調査を行なった．

Vol.2016-MPS-108 No.23
Vol.2016-BIO-46 No.23
2016/7/5

3.1 対象データ

本研究では，緯度・経度・深度・季節の測定値との関連性を調べるために，浅い層（0-500[dbar]），中間層（500-1,000[dbar]），深い層（1,000-2,000[dbar]）に分けられた，水温の深度系列計 3 種のデータに対してそれぞれ階層的クラスタリングを行なった．使用したアルゴデータは，日本でアルゴ計画を担っている国立研究開発法人海洋研究開発機構の海洋技術者により遅延品質管理が施された北太平洋（東経 140 度-西経 140 度，北緯 10 度-北緯 50 度）の正常水温データ 10,000 プロファイルである．また，測定深度を揃えるために，各プロファイルに線形補間を行い，圧力 10[dbar] 毎の水温データの深度系列を求めた．

3.2 階層的クラスタリング

あるデータ i の深度と水温の系列を $(< p_i^{(1)}, t_i^{(1)} >, < p_i^{(2)}, t_i^{(2)} >, ..., < p_i^{(n_i)}, t_i^{(n_i)} >)$，データ i と j の共通深度集合を $L = \{< p_i^{(l)}, p_j^{(m)} > | p_i^{(l)} = p_j^{(m)}\}$ と表す．このとき階層的クラスタリングにおける i と j の距離 $d(i,j)$ は，系列長による違いを考慮するために，以下のように系列長 $|L|$ で規格化されたユークリッド距離を用いた．

$$d(i,j) = \frac{\sqrt{\sum_{l,m \in L} (t_i^{(l)} - t_j^{(m)})^2}}{|L|} \quad (1)$$

クラスタ間距離測定には，最長距離法を採用した．最長距離法の場合，本来同一のクラスタに所属すべきクラスタが離れてしまう可能性があるが，局所的な変動を見ることができる．クラスタ併合を打ち切る閾値は，図 4 のようにクラスタ併合を行なった時にクラスタ間距離が急激に大きくなる点を選んだ．

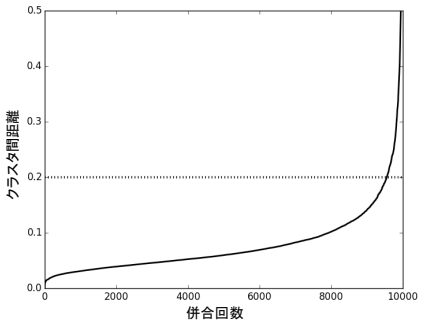


図 4 併合時のクラスタ間距離
Fig. 4 Distance among clusters of hierarchical clustering.

4. 実験結果

それぞれの層に階層的クラスタリングを適用した結果，浅い層では **449** クラスタ，中間層では **301** クラスタ，深い層では **301** クラスタ得られた．色分けされた全クラスタの空間配置を図 5 に図示する．全ての層において，経度方向に長いクラスタが多数形成されていることから，北太平洋では深度に関わらず測定値は緯度方向より経度方向に相関があることが示唆される．

次にクラスタ内のデータ点数のヒストグラムを図 6 に図示する．浅い層において最大のクラスタでもデータ点数は 180 に満たず，データ点数の分散が比較的小さいヒストグラム形状になっているが，深度が深くなるにつれて少数のクラスタが非常に多くのデータ点数を持つようになっている．これは深度が深いほど，水温の変動が小さくなり，大きなクラスタが形成されやすくなっているからである．しかし，中間層・深い層では 2，3 点から成るクラスタが多数形成されていることから，一部の海域では特異な変動を起こしていることが確認できる．

それぞれの層のクラスタリング結果の内，6 クラスタの測定値・空間配置・測定月を色と記号で対応させて図 7 に図示する．図 7 中段のクラスタの空間配置において，どの層でも主にクラスタが近隣の連続する海域のみのデータから形成されていることから，空間的な相関が強いことが確認できる．中緯度帯では，緯度方向にも多少相関が見られるが，高緯度及び低緯度では，経度方向に細長いクラスタが形成されている．これらの位置には，北太平洋海流と北赤道海流が流れていることから，経度方向の相関が特に強いことに海洋学的な妥当性があると言える．

また，図 7 上段のクラスタの測定値と空間配置から，概して高緯度帯ほど水温が冷たくなっていることを確認できる．しかし，浅い層の低緯度に位置するクラスタ 3・5 は 200[dbar] 以上では中緯度に位置する 2・3 よりも水温が低くなっている．中間層において 600[dbar] 以上では中緯度帯のクラスタ 2・6 の方が低緯度の 5 よりも水温が低くなっていることから，この水温の逆転は 200-600[dbar] において発生しているものと思われる．

一方で，小さいクラスタを確認すると，図 7 の浅い層のクラスタ番号 1，中間層の 1，深い層の 3 など，日本近海の太平洋沖に多数形成されており，これらのクラスタは総じて変動が激しかった．特に中間層・深い層ではこれら以外のクラスタは非常に安定しているが，中間層 1 と深い層 3 では局所的に大きな変動が確認される．この海域は温暖な黒潮と寒冷な親潮が衝突する海域であるため，いずれの層においても水温の変動が激しく，大きなクラスタが形成されにかったことは，海洋学的に妥当である．

形成されたクラスタの測定月を確認すると，浅い層の中緯度以北のクラスタ 2・3・4 では季節に偏りがある．この特徴は図示していない他のクラスタにも共通し，季節依存性が示唆される．クラスタ 3 と 4 は空間的に近く，測定月はあまり重複せず，圧力 100[dbar] 以上では測定値もほぼ同じであることから，この海域での季節依存性は 100[dbar] 以下で特に顕著になると予想される．クラスタ 6 は高緯度に位置するが，季節の偏りは見受けられない．これは上述の変動が激しい海域に位置しているため，季節

による変動が海域特有の変動の激しさによって打ち消されているものと思われる．中間層及び深い層では測定月の偏りはほとんど見られなかった．

中間層と深い層の挙動はおおよそ似た傾向を示した．高緯度帯の東北東方面に細長い少数の巨大なクラスタが形成されている．一方で，中緯度の海域では，やや大きいクラスタが多数形成されていたため，これらのクラスタを目視で確認したところ，クラスタ間の測定値の違いは僅かであった．そのため，中緯度の海域も高緯度帯同様に変動が穏やかであるが，高緯度帯より少し変動が激しいと言える．

5. エラー測定値検知のための議論

一般的に北太平洋海域は穏やかであると言われているが，実験の結果，季節・深度・緯度・経度によって様々な変動構造が存在することを確認できた．そのため，単純に海域を区切って異常検知技術を適用しても望ましい結果は得られないと想定される．例えば，親潮と黒潮がぶつかる変動が激しい海域と北太平洋海流が流れる比較的変動が穏やかな海域に対して，深度系列のスライド窓をとり，One Class SVM[3] や Local Outlier Finder[1] を適用した場合を考える．このとき，異常検知を行う閾値を敏感に設定すると変動が激しい海域のエラー測定値の誤検出が相次ぎ，閾値を鈍感に設定すると穏やかな海域のエラー測定値の見逃しが多発する可能性がある．しかし，変動構造が異なる海域毎に異常検知のモデルを変えることによって，効率的な異常検知が可能になる．例えば，高緯度の浅い層では季節依存性が強いことが期待されるので，テストデータと同じ海域で同じ季節のデータのみを異常検知に使用すればよい．また海域の変動の激しさに対応して，異常検知の閾値を適切に決定することができる．今後は，クラスタリング結果から得られた知見を基に，海域を変動構造毎に分割して，深度系列のスライド窓に対する One Class SVM や Local Outlier Finder の適用を検討する．

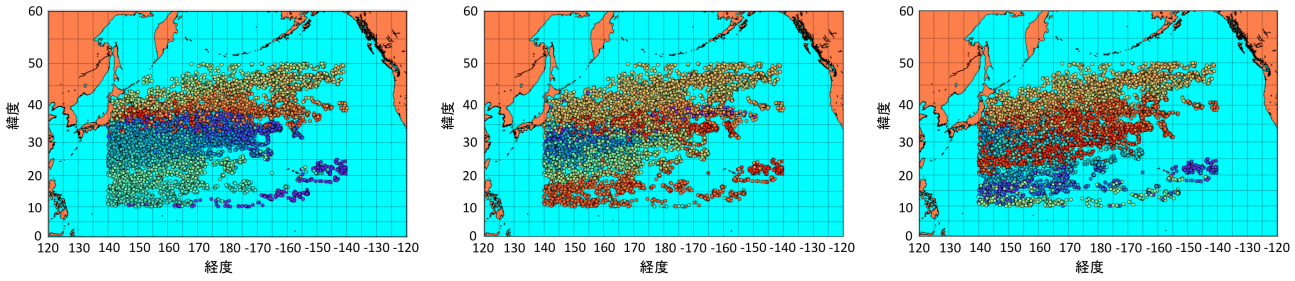
6. まとめ

本研究では，海洋データの深度系列のエラー測定値検知モデルを構築するための初期的な検討として，浅い層・中間層・深い層に分けた深度系列に対して，階層的クラスタリングの適用を行なった．得られたクラスタの測定値・空間配置・測定季節について調べることによって，空間的な相関構造，高緯度かつ浅い層での季節による変動構造，変動が激しい一部の海域などを発見することができた．今後は，これらで得られた海域による変動構造の違いをもとに，深度系列を活用した異常検知を試みる．

参考文献

[1] Breunig, M. M., Kriegel, H.-P., Ng, R. T. and Sander, J.: LOF: identifying density-based local outliers, *ACM sigmod record*, Vol. 29, No. 2, ACM, pp. 93–104 (2000).
[2] Ono, S., Matsuyama, H., Fukui, K. and Hosoda, S.: Error detection of oceanic observation data using sequential labeling, *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, IEEE, pp. 1–8 (2015).
[3] Tax, D. M. and Duin, R. P.: Support vector data description, *Machine learning*, Vol. 54, No. 1, pp. 45–66 (2004).
[4] Team, A. D. M.: Report of the Argo Data Management Meeting, *Proc. Argo Data Management Third Meeting, Marine Environmental Data*, p. 42 (2002).
[5] Willis, J. K., Lyman, J. M., Johnson, G. C. and Gilson, J.: In Situ Data Biases and Recent Ocean Heat Content Variability, *Journal of Atmospheric and Oceanic Technology*, Vol. 26, No. 4, pp. 846–852 (2009).
[6] 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 (2013).

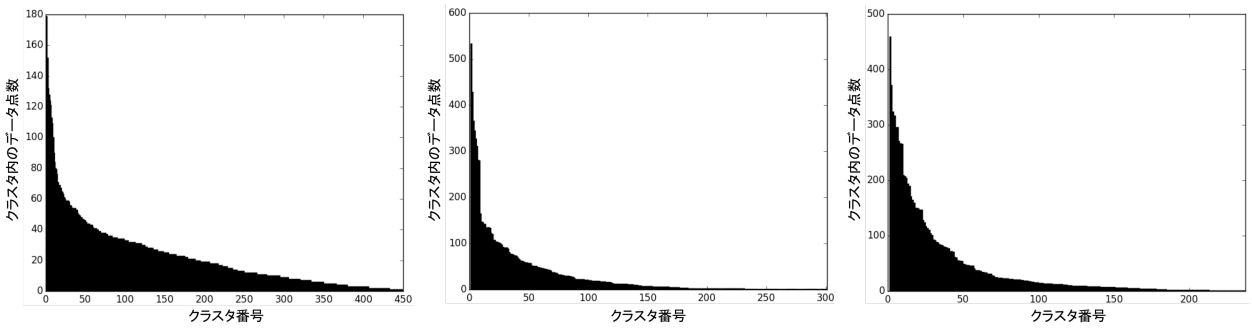
Vol.2016-MPS-108 No.23
Vol.2016-BIO-46 No.23
2016/7/5



(a) 浅い層　(b) 中間層　(c) 深い層

図 5　色分けされた全クラスタの空間配置

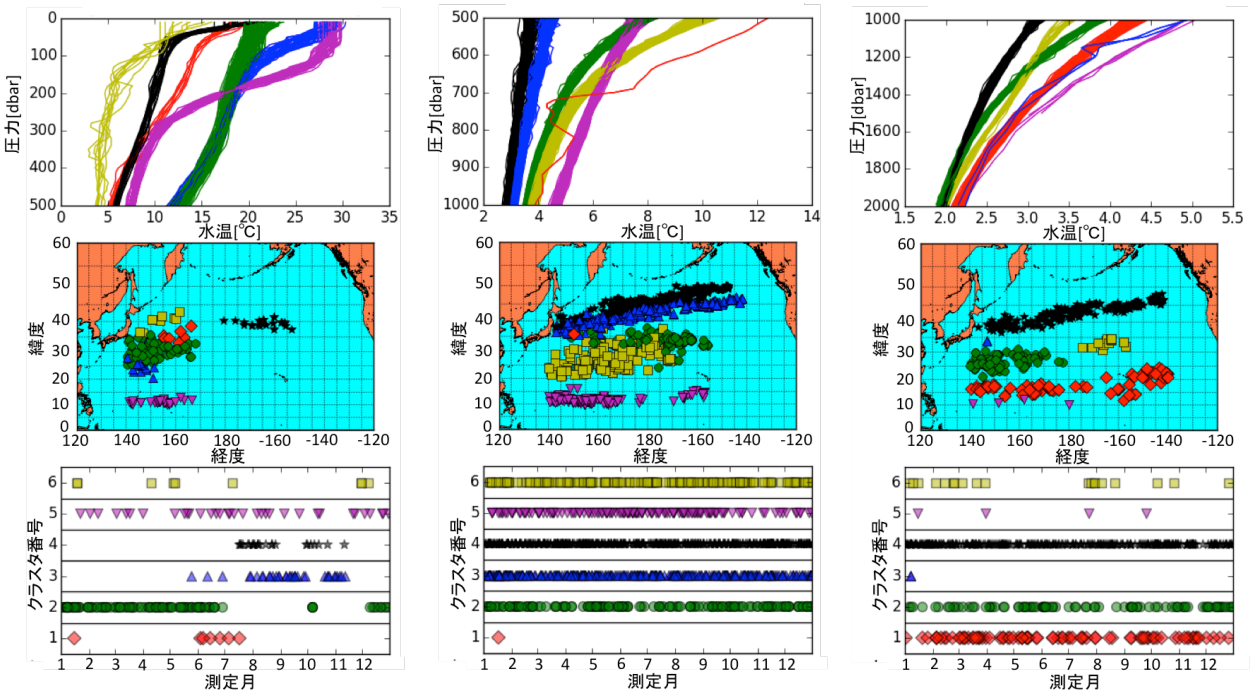
Fig. 5　Spatial distribution of all colored clusters.



(a) 浅い層　(b) 中間層　(c) 深い層

図 6　クラスタ内のデータ点数の降順ヒストグラム

Fig. 6　Sorted histogram of number of data in clusters.



(a) 浅い層　(b) 中間層　(c) 深い層

図 7　6 クラスタの測定値 (上段)，空間配置 (中段)，測定月 (下段)

Fig. 7　Tempreture(upper row), spatial distribution(middle row), measured month(lower row) of six clusters.