4P1-OS-38a-4

空間的自己相関を考慮した海洋データのエラー検知

Error Detection in Ocean Data considering Spatial Autocorrelation

林 勝悟 *1 Shogo Hayashi　小野 智司 *2 Satoshi Ono　細田 滋毅 *3 Shigeki Hosoda　沼尾 正行 *4 Masayuki Numao　福井 健一 *4 Ken-ichi Fukui

*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鹿児島大学大学院理工学研究科
Graduate School of Science and Engineering, Kagoshima University

*3国立研究開発法人海洋研究開発機構
Japan Agency for Marine-Earth Science and Technology

*4大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

Ocean data, which is a kind of spatial data, is difficult to detect errors because its characteristics are different among ocean areas. For now, the error detection is conducted with visual checks by ocean data technicians. However, they are time-resource-consuming and difficult to deliver accurately and uniformly quality-controlled ocean data because technicians' skills are not uniform. In this work, we propose a framework for an automated error detection in the ocean data, that is applicable for unknown types of errors, considering the differences among the ocean areas: spatial autocorrelation in latitude, longitude and depth. The framework comprises a training data selection to care the spatial autocorrelation and an error detection. As a result, we found the effective combinations of features, training data selecting methods and anomaly detection methods, regarding the ocean characteristics. Moreover, our proposal training data selecting method worked efficiently, even when training data was little around test data.

1. はじめに

気候変動の予測やメカニズム解明のために，2000 年に 30 カ国以上の海洋・気象機関によって国際アルゴ計画が始動された．アルゴ計画において，世界中の海洋に放流された 3,500 以上の観測フロートが，深度 2,000m から海表まで浮上しながら，水温・塩分といった海洋データの深度系列を測定する（図 1）．この 1 深度系列を 1 プロファイルと呼ぶ（図 2）．種々の理由によりエラーデータが測定されることがある（図 2 中 Error1）．エラーを含むデータによって見積もられた気候変動シグナルは，政策決定にさえ関係する気候変動の研究に悪影響を与える恐れがある．

エラーの除去・補正のために，現在アルゴ計画では二段階の品質管理が行われている [Argo Data Management Team 02]．即時品質管理（Real-time Quality Control：RQC）では，海洋技術者の長年の経験知識を活用したエラー検知ルールが適用される．遅延品質管理（Delayed-mode Quality Control：DQC）では，海洋技術者が知識と経験を活かして目視でエラーの除去・補正を行う．しかし，海洋データには図 2 のように海域で異なる複雑な自然変動が含まれているため，RQC で全エラーに対応することは難しい．また，DQC は海洋技術者の大きな負担を強い，海洋技術者個人の作業に大きく依存しているため，品質管理の均一性に問題が残る．これらの問題はアルゴデータの信頼性に影響を及ぼし，国際アルゴ計画が始動されてから長年解決が望まれてきた．

海洋データのエラー検知問題に対して我々は，海域による変動の違いを考慮するために，クラスタリングを基に海洋の分割を行い，分割された海域毎にエラー検知を行なった [Hayashi 16]．結果として，深層における正常データとの変動の差が僅かなエラーデータに対して高い異常度付与に成功した．しかし，海域分割では海域内のテストデータのトレーニングデータ集合は同一であるため，海域端点付近のデータのエラー検知に問題があった．

本研究では，未知タイプのエラーに適用可能であり，空間的自己相関を考慮した海洋データのエラー検知フレームワークを提案する．提案フレームワークは大きく，空間的自己相関が強い空間的近傍トレーニングデータ集合（以下，近傍データ集合と呼ぶ）の選択と，エラー検知で構成される．テストデータそれぞれに合った近傍データ集合の選択を行うことにより，既存研究 [Hayashi 16] より高いエラー検知精度が期待できる．本研究では，提案フレームワークにおける手法の有効な組み合わせを求めるために，特徴量比較実験，近傍データ集合選択手法比較実験，エラー検知手法比較を行う．

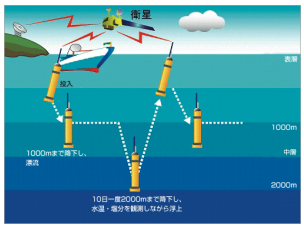


図 1: 観測フロートによる海洋データの測定 *1

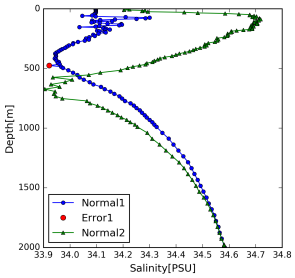


図 2: 変動の性質が異なる 2 海域の塩分プロファイル

連絡先: 林 勝悟, 大阪大学, 産業科学研究所沼尾研究室,
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1,
Tel: 06-6879-8426, Fax: 06-6879-8428,
E-mail: <surname>@ai.sanken.osaka-u.ac.jp

*1 http://www.jamstec.go.jp/J-ARGO/overview/overview_3.html

2. 関連研究

海洋データのエラー検知は，ユークリッド空間で定義される正常な空間的特徴量 $\boldsymbol{s}$（緯度，経度，深度）が与えられたときの異常な非空間的特徴量 $\boldsymbol{x}$（温度，塩分）を検知する，空間的異常検知 [Aggarwal 13] の一種である．空間的異常検知手法 [Kou 06] は，与えられた空間的近傍集合における非空間的特徴量の平均との差を計算するなど，空間的自己相関を考慮した異常検知を行う．

海洋データのエラー検知の既存研究として，上川路ら [Kamikawaji 16] は，深度系列性を考慮するために，Conditional Random Field(CRF) を用いた深度系列ラベリングを行なった．結果として，太平洋の限られたエラー種に対して 95%を超える精度と再現率を達成した．しかし，CRF は教師有り学習であるため，エラーデータと正常データの不均衡問題を受け，未知のエラーを検知できない可能性がある．

3. 空間的自己相関を考慮した海洋データのエラー検知

3.1 提案フレームワーク

提案フレームワーク全体の流れを以下に，そのフローチャートを図 3 に示す．

1. テストデータ（水温もしくは塩分の 1 プロファイル）の近傍データ集合を選択する．
2. テストデータと近傍データ集合を深度系列のスライド窓集合に変換し，同深度毎に分割する．
3. それぞれの深度のテストデータと近傍データ集合に対して，正常トレーニングデータのみを用いて異常検知手法を適用し，テストデータの異常度を得る．
4. 全てのテストデータに対して (1), (2), (3) を行う．

(1), (2) により，緯度，経度，深度における空間的自己相関を考慮することができ，空間的異常検知問題から異常検知問題に近似され，一般的な異常検知手法が適用可能になる．また，スライド窓変換により，深度の系列性考慮が可能になる．(3) により，不均衡問題を受けず，様々な変動が存在する海洋データの未知のエラーに対応可能になる．

3.2 空間的近傍トレーニングデータ集合の選択手法

近傍データ集合の選択手法として，全データを選択する手法（All），テストデータから固定の緯度経度の範囲内のデータを選択する手法（Rect），提案選択手法であるクラスタリングに基づいて選択を行う 2 手法の計 4 手法を使用する．提案選択手法は，階層的クラスタリングで形成されたクラスタに所属するデータを選択するため，空間的自己相関が強い任意形状に空間分布するデータ集合を選択することが期待できる．

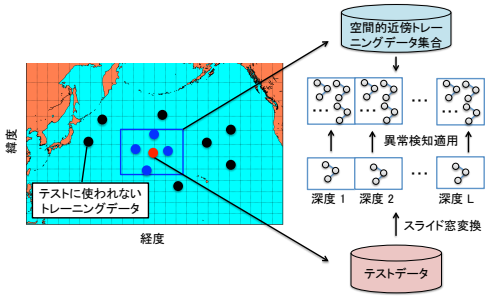


図 3: 提案フレームワーク

3.2.1 階層的クラスタリングに基づく選択手法 1 ：Clust1

階層的クラスタリングに基づく選択方法 1（Clust1）は，階層的クラスタリングで得られたクラスタの重心とテストデータの空間的距離に基いて選択を行う．1 テストデータが得られた際の近傍データ集合の選択手順を以下に記す．

前処理

1. 全トレーニングデータに対して階層的クラスタリングを適用し，クラスタを得る．
2. 各クラスタの緯度経度における重心を保存する．

近傍データ集合選択

3. テストデータと各クラスタ重心との空間的距離を計算し，昇順にソートした候補集合をつくる．
4. 候補集合の中から距離が最短のクラスタに所属するトレーニングデータを近傍データ集合に追加し，そのクラスタを候補集合から取り除く．
5. （4）を繰り返し，近傍データ集合の要素数がある閾値 $MinimumNumber$ を超える，もしくはテストデータとクラスタ重心までの空間的距離がある閾値 $MaxDistance$ を超えれば近傍データ集合への追加を終了する．

なお，本研究では $MinimumNumber$ と $MaxDistance$ を 350 と 1250[m] に設定した．

3.2.2 階層的クラスタリングに基づく選択手法 2 ：Clust2

階層的クラスタリングに基づく選択方法 2（Clust2）は，Clust1 の選択手順とおおよそ同じであるが，前処理で得られたクラスタ重心を計算せずにトレーニングデータ 1 点とテストデータの空間的距離に基いて選択を行う．Clus1 との挙動の違いとして，Clust2 はクラスタの空間的重心が遠くてもクラスタ内の 1 データが近い限り近傍データ集合として追加する．Clust1 と同様に，$MinimumNumber$ と $MaxDistance$ を 350 と 1250[m] に設定した．

3.3 異常検知手法

一般的な異常検知手法である，k-th Nearest Neighbor(k-th NN)，Local Outlier Factor (LOF)，One-class Support Vector Machine (OCSVM)，Isolation Forest (IF) と，空間的異常検知手法である，AvgDiff Algorithm（AvgDiff）[Kou 06] を提案フレームワークにおける異常検知に用いる．また，AvgDiff を拡張して，空間的近傍データ集合の非空間的特徴量の k 近傍のみを異常度評価に考慮する AvgDiff+も使用する．なお，OCSVM のカーネル関数として RBF カーネルを用い，IF では $t = 100$，$s = 256$ とし，AvgDiff 及び AvgDiff+の空間的重み付けとして，指数関数重み付け $c\exp(-\|\boldsymbol{s} - \boldsymbol{s}'\|^2)$ を採用する．ここで，$\boldsymbol{s}$ は空間的特徴量であり，c は正規化定数である．

4. 実験

4.1 データ

2006 年までに北太平洋（北緯 10-50 度，東経 140-西経 140 度）で測定された正常塩分データ 10,000 プロファイルをトレーニングデータとして，2007 年以降に測定されたエラーを含む 2,000 プロファイル（内エラーを含むプロファイル数は 116）をテストデータとしてランダムに抽出した．

前処理として，最小測定深度間隔である 5[m] 毎の線形補間と Z 標準化を行ない，窓幅 M の深度のスライド窓に変換することにより，最終的に実験に用いたデータを得た（表 1）．なお，エラーシグナルを見落とさないために，補間時に補間値の両隣の生データが正常な場合のみ補間値に正常と，それ以外

表 1: データ数（スライド窓幅 10，測定値と勾配の場合）

海域	トレーニング	正常テスト	エラーテスト
浅層	473251	95551	250
中層	773768	157582	467
深層	2264393	478400	4280

にエラーとラベル付けした．同様に，スライド窓変換する際は，スライド窓の補間値が正常ラベルから成る場合のみスライド窓に正常と，それ以外にエラーとラベル付けした．また，深度によって異なる統計的性質を考慮するために，線形補間後のトレーニングデータの深度毎の標準偏差を基に，深度を浅層（0-250[m]），中層（250-650[m]），深層（650-2,000[m]）に分割し，それぞれに別々のモデルを適用した．

4.2 特徴量比較実験

4.2.1 実験設定

特徴量の比較実験として，測定値，深度方向の勾配，スライド窓幅（1,3,5,10）の組み合わせを変えたときの ROC（Receiver Operating Characteristic）曲線の AUC（Area Under the Curve）の評価実験を行なった．なお，近傍データ集合選択手法として All を，異常検知手法として既存研究 [Hayashi 16] で最も結果が良かった k-th NN（k=1）を使用した．なお，実行環境として，Intel(R) Xeon(R) CPU E5-2697 v2 2.70GHz 12-Cores x2，RAM 128GB を用い，言語は Python を使用した．

4.2.2 実験結果と考察

全ての層において窓幅 10 で測定値と勾配を組み合わせた特徴量が最も結果が良く，AUC は浅層から順に 0.864，0.877，0.896 となった．これは，深度系列の変動が海域によって大きく異なるため，窓幅 10 で海洋データに勾配情報を加えることにより，海域の特徴がより正確に表現されたからだと思われる．追加実験として，より大きな窓幅（15,20,30,50）を設定した結果として，窓幅 10 のときより浅層では AUC が減少したが，中深層ではいずれも結果が改善し，窓幅 50 のときに 0.944，0.930 となった．これは，浅層では変動が激しいため，50[m] 以上の空間的自己相関が弱く，逆に中深層では変動が比較的安定しているため，より深い系列性を考慮することにより結果が改善されたと思われる．しかし，窓幅を大きくして次元を増やしすぎると，エラー検知の計算量が膨大になり，深度系列長が短いプロファイルのスライド窓変換が不可能となるため，本研究では窓幅 10 の測定値と勾配を実験で用いる特徴量として決定した．

4.3 空間的近傍トレーニングデータ選択手法比較実験

4.3.1 実験設定

近傍データ集合選択手法比較実験として，All，Rect(緯度, 経度)，Clust1(距離定義, 閾値分位点)，Clust2(距離定義, 閾値分位点) を実行したときの各手法における AUC の評価を行なった．パラメータとして，Rect には (2,4)，(5,10) を，Clust1 と Clust2 の距離定義には complete，single，average を，閾値分位点には 2,4 を設定した．なお，4.2 節の実験結果に基づき特徴量として窓幅 10 で測定値と勾配を組み合わせたものを用い，異常検知手法として k-th NN（k=1）を使用した．

4.3.2 実験結果と考察

浅層では，Clust1(single,2) のときに AUC は最高の 0.890 であった．階層的クラスタリングに基づく選択手法により，変動が激しい浅層においてエラー検知に関係のある，空間的自己相関が強いデータを自動的に選択できたことが分かる．中層では Rect(5,10) のとき AUC が 0.934，深層では All のときに

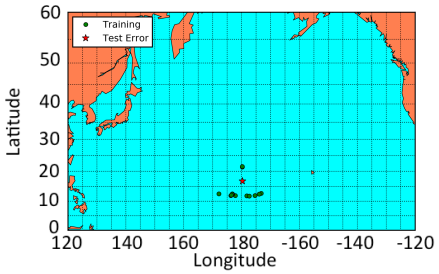


図 4: Rect(5,10) による選択データの空間分布

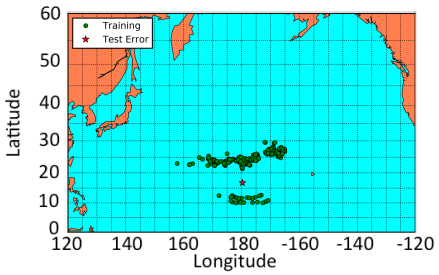


図 5: Clust1(single,2) による選択データの空間分布

0.896 と最も良かった．これは，深度が深いほど海洋データの変動が安定し，緯度経度における空間的自己相関が強くなるため，より広い範囲のデータを選択することにより結果が良くなったと思われる．

Clust1(single,2) の選択性質として，北太平洋中央のように変動が穏やかな海域では，巨大なクラスタが形成されるため広範囲のトレーニングデータが選択され，逆に日本近海のように変動が激しい海域では，小さいクラスタが形成されるため円形範囲のトレーニングデータが選択された．

空間的近傍にトレーニングデータが少ないテストデータ（北緯 17 度，東経 180 度）において，各手法の選択の違いを確認した．Rect(2,4) は範囲にトレーニングデータが無かったため，エラー検知を行うことができなかった．Rect(5,10) は少数のデータを集めることができたが（図 4），トレーニングデータの深度が浅かったため，深度 1960[m] に存在するエラーの検知に失敗した．一方で，Clust1(single,2) は，クラスタリングを通じて変動が似たデータを近隣海域以外から補完したことにより（図 5），エラー検知に成功した．

4.4 異常検知手法比較実験

4.4.1 実験設定

異常検知手法比較実験として，k-th NN，LOF，AvgDiff，AvgDiff+，OCSVM，IF の 5 種類の異常検知手法による AUC の評価実験を行なった．なお，4.2 節と 4.3 節の実験結果に基づき，特徴量には窓幅 10 の測定値と勾配を使用し，近傍データ集合選択手法として，浅層では Clust1(single,2) を，中層では Rect(5,10) を，深層では All を用いた．

4.4.2 実験結果と考察

浅層では，k-th NN（k=1）が最も良い結果を残した（表 2）．これは浅層では局所的な海域依存変動が強いため，非空間的特徴量のデータ空間における近傍データのみを考慮する k-th NN（k=1）が最も適していたことが理由であると考えられる．中層でも k-th NN（k=1）が最も良い結果を出したが，浅層ほど局所的な海域依存変動が強くないため，20 以下の k をパラメータとして持つ手法の結果と大きく変わらなかった．深層では，全データを空間的距離で重みづけた AvgDiff が最高の AUC0.922 を達成した．これは，空間的距離に基づいた重み付けにより，変動が穏やかである深層の僅かな海域依存変動をう

表 2: 各異常検知手法における AUC.
OCSVM は AUC が最大と最小のもののみを示している.

手法	k	ν	γ	AUC 浅層	AUC 中層	AUC 深層
k-th NN	1			**0.890**	**0.934**	0.896
k-th NN	5			0.866	0.928	0.896
k-th NN	10			0.853	0.922	0.896
k-th NN	20			0.835	0.917	0.895
k-th NN	50			0.801	0.908	0.895
LOF	1			0.822	0.844	0.885
LOF	5			0.816	0.907	0.898
LOF	10			0.813	0.933	0.903
LOF	20			0.810	0.933	0.906
LOF	50			0.763	0.926	0.907
AvgDiff				0.771	0.790	**0.922**
AvgDiff+	5			0.877	0.930	0.896
AvgDiff+	10			0.867	0.927	0.896
AvgDiff+	20			0.855	0.924	0.896
AvgDiff+	50			0.829	0.919	0.895
OCSVM		0.05	0.1	0.517		0.429
OCSVM		0.2	0.001		0.809	0.348
OCSVM		0.2	0.1		0.860	
OCSVM		0.5	0.001	0.701		0.348
IF				0.766	0.898	0.500

表 3: 提案手法と RQC の F 値

手法	浅層	中層	深層
提案手法	**0.796**	**0.865**	**0.864**
RQC	0.603	0.630	0.812

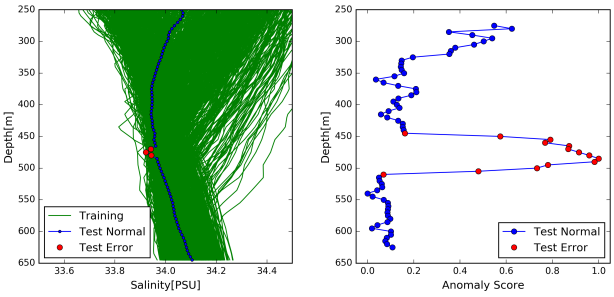
まく捉えることができたからであると思われる.

浅層, 中層, 深層で最良結果を出したそれぞれの組み合わせにおける, 異常標本精度 (≡(異常データのうち異常データであると正しく判定できた数) / (実際の異常データの数)) と正常標本精度 (≡ (正常データのうち正常データであると正しく判定できた数) / (実際の正常データの数)) の F 値 (調和平均) の最大値と, RQC の F 値の比較を行なった. 結果として, いずれの層においても提案手法は, 海洋技術者の長年の経験知識により設計された RQC よりも高い F 値を獲得した (表 3). 特に海洋データの複雑な自然変動が現れやすい浅中層において, 提案手法の F 値は約 0.2 高く, 提案手法の有効性が確認された.

中層で最良結果を出した手法の組み合わせにおいて, 図 2 の識別が難しいエラーを含む塩分データ (Normal1 と Error1) に対する異常度付与結果として, テストデータとその近傍データ集合の補間塩分値とテストデータの異常度を図 6 に示す. エラーは近傍データ集合の正常範囲に含まれているが (図 6a), 提案手法は深度系列性を考慮してエラーシグナルを強調することができたため, 他の正常データと比べて高い異常度付与に成功した (図 6b). しかし, 前処理としてデータに対して補間を行っているため, エラーの端点に対しては正常データと変わらない低い異常度を与えた. また, 浅い方がより海洋データの変動が大きく複雑であるため, 中層の浅い位置のデータに対して深い位置のデータよりも高い異常度を付与した.

最終的に中深層においては, 異常検知手法比較実験結果よりも, 特徴量比較実験において窓幅 50 で測定値と勾配を用いて, All と k-th NN (k=1) を適用した実験結果の方が良かった. これは, 北太平洋の中深層における海洋データの変動は穏やかであり, 深度系列性が強いためであると考えられる. 逆に浅層では, 変動が中深層よりも激しいため, 適切な近傍データ集合の選択を行うことにより, 結果を改善することができた. よって, 提案フレームワークにおける近傍データ集合選択は, 大西洋や浅層のように海洋データの変動が激しい海域で特に有効に働くことが期待できる.

以上の議論をまとめると, 海洋データの変動が穏やかな海域では, 深度系列性を十分に反映することにより, また激しい海域では, 強すぎない深度系列考慮と空間的近傍トレーニングデータ選択方法を適切に組み合わせることで, エラー検知の結果を改善できると言える.



(a) 補間塩分値　(b) 異常度

図 6: 中層における補間塩分値と k-th NN (k=1) が付与した [0,1] に規格化された異常度.
テストエラーはトレーニングデータの正常範囲に入っているが, 提案手法は高い異常度付与に成功した.

5. まとめ

本研究では, 未知のエラーに対応可能であり, 空間的自己相関を考慮した海洋データのエラー検知フレームワーク及びクラスタリングに基づいた空間的近傍トレーニングデータ集合の選択手法を提案した. 実験結果として, 海域の変動の性質に応じて有効な特徴量, スライド窓幅, 近傍データ集合選択手法, 異常検知手法の組み合わせを発見することができ, RQC よりも高い F 値を達成できた. また, 提案データ選択手法は, トレーニングデータが少ない海域でも他海域からデータを補完してエラー検知に成功した. 一方で, 提案フレームワークは非線形な空間的自己相関や時間的自己相関を考慮することができない. 海洋データの構造に即したガウス過程をエラー検知に適用することを今後の展望として検討している.

謝辞

本研究の一部は, 科学研究費補助金挑戦的萌芽研究 (16K12490), 高橋産業経済研究財団及び物質・デバイス領域共同研究拠点：人・環境と物質をつなぐイノベーション創出ダイナミック・アライアンスの補助を受けて行われた.

参考文献

[Aggarwal 13] Aggarwal, C. C.: *Outlier Analysis*, Springer Publishing Company, Incorporated (2013)

[Argo Data Management Team 02] Argo Data Management Team, : *Report of the Argo Data Management Meeting* (2002)

[Hayashi 16] Hayashi, S., Ono, S., Hosoda, S., Numao, M., and Fukui, K.: Error Detection of Ocean Depth Series Data with Area Partitioning and Using Sliding Window, in *The 15th IEEE International Conference on Machine Learning and Applications*, pp. 1029–1033 (2016)

[Kamikawaji 16] Kamikawaji, Y., Matsuyama, H., Fukui, K., Hosoda, S., and Ono, S.: Feature Function Design in Conditional Random Field Using Decision Tree Learning Applied to Error Detection of Ocean Observation Data, in *Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence* (2016)

[Kou 06] Kou, Y., Lu, C., and Chen, D.: Spatial Weighted Outlier Detection., in *SIAM International Conference on Data Mining*, pp. 614–618 (2006)