

Clustering de Dados Oceanográficos no Atlântico Norte – Visualização dos clusters em diferentes espaços, para uma melhor compreensão da sua estrutura

Pires¹, F.J., Bação², F., Lobo^{1,2}, V.

¹Escola Naval, ²ISEGI

Sumário: Neste artigo é descrito um processo de detecção de agrupamentos em dados oceanográficos. A aplicação em si é relevante para uma melhor compreensão dos fenómenos oceanográficos que ocorrem ao largo da costa portuguesa. Esta aplicação também ilustra bem os processos de agrupamento realizados pelas redes SOM (Self Organizing Maps, ou Redes de Kohonen), bem como a importância da visualização na análise de agrupamentos.

Palavras chave: clustering, visualização de dados, oceanografia

1. Introdução

Os SOM têm sido usados com sucesso em muitas aplicações de clustering [1, 2] mas, por vezes, é difícil compreender a razão de ser das escolhas de clusters que são feitas. Esta dificuldade em interpretar SOMs faz com que muitas vezes se opte por técnicas mais tradicionais, como k-médias ou o clustering hierárquico que, apesar de permitirem a partição em agrupamentos, têm o inconveniente de não evidenciarem a estrutura dos dados, dificultando a compreensão desses dados. Neste artigo apresentam-se algumas técnicas de clustering baseadas em SOM que podem levar a uma melhor compreensão da estrutura dos dados.

O estudo clássico dos fenómenos atmosféricos e oceanográficos apoia-se em metodologias de modelação caracterizadas pela forte não linearidade e elevada complexidade dos modelos, que decorrem, em grande medida, da natureza dos dados em que assentam. Estes modelos necessitam de trabalhar uma grande quantidade de dados, cobrindo uma grande extensão geográfica, para capturarem o comportamento macroscópico da atmosfera, do oceano e da interacção atmosfera-oceano.

Uma forma de reduzir a complexidade e, colateralmente, de reduzir também a quantidade de dados necessários para caracterizar as massas (de água e ar) presentes, seria a de identificar grandes zonas de homogeneidade, quando consideradas as variáveis de interesse no seu conjunto (clusters).

Ao caracterizar massas de água ou ar, procurando zonas de homogeneidade nas variáveis de análise, está-se, na verdade a trabalhar com dados georeferenciados a 3 dimensões (3D), o que coloca alguns constrangimentos de vulto visto que a maior parte das aplicações georeferenciadas trabalham a 2 dimensões (2D). É assim necessário procurar abordagens indirectas para se poder trabalhar e principalmente avaliar os resultados obtidos, uma vez que a visualização de resultados sobre dados 3D é um problema não trivial. Este tipo de abordagem (GIS 3D) pode revelar-se importante no estudo da atmosfera (poluição/meteorologia), do oceano (idem e também o estudo de correntes, de inventariação de recursos, eventualmente contribuintes para a extensão da plataforma continental), e também em campos emergentes de aplicação de GIS tais como para aplicações da área da medicina onde se “navega” pelo corpo humano.

Este trabalho propõe-se assim efectuar uma avaliação sucinta da aplicação de algumas técnicas de clustering a dados climatológicos referentes ao Atlântico Norte, fazendo uma abordagem introdutória à problemática do tratamento de dados georeferenciados a 3 dimensões, recorrendo-

se ao repositório de dados do National Oceanographic Data Centre (NODC) [3], para a área do Oceano Atlântico adjacente a Portugal Continental (figura 1). Dada a grande diversidade dos dados disponíveis foram seleccionadas 2 “features” não geográficas: Salinidade (S) e temperatura (T). Os dados seleccionados foram também separados em estações do ano Verão/Inverno.

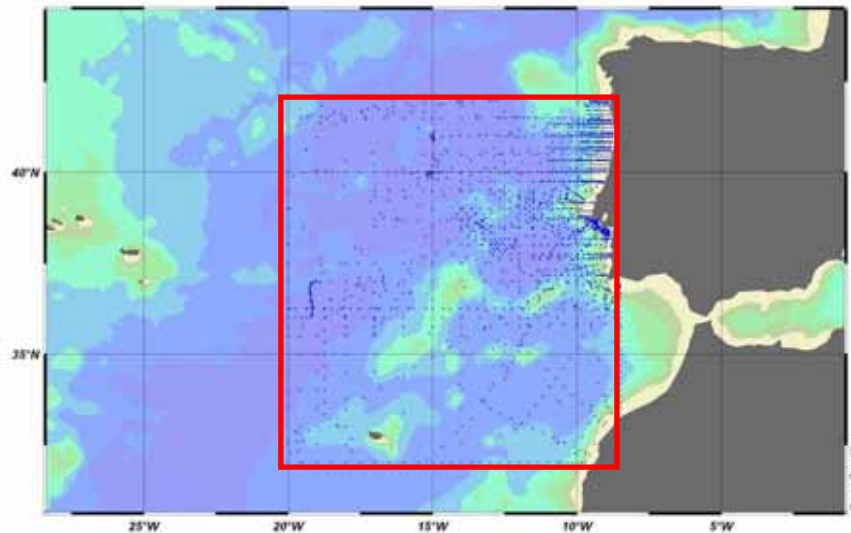


Figura 1 - Distribuição dos dados.

2. Clustering usando como critério apenas temperatura (T) e salinidade (S)

Como os dados são bidimensionais (T e S), é possível visualizarmos a sua distribuição através de um gráfico 2D, apresentado na figura 2 (Inverno e Verão).

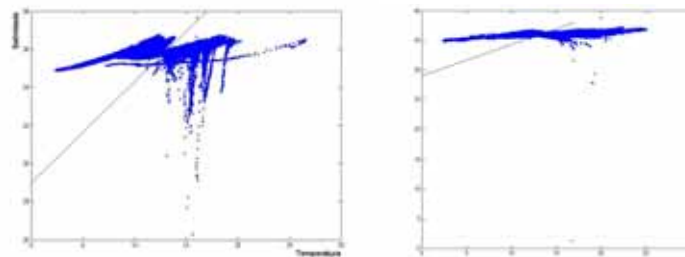


Figura 2 - Distribuição S/T para dados de Inverno (à esquerda) e Verão (à direita).

Deste processo preliminar de “clustering visual” podem-se obter, apenas para os dados de Inverno, alguns resultados razoáveis, tendo em conta o conhecimento à priori já existente: Identificamos dois grandes clusters que, mais tarde veremos, correspondem às camadas de superfície e de profundidade (separados pela linha inclinada a preto) e verificamos que a separação entre as duas é difusa e dificilmente identificável apenas com recurso à observação dos dados no espaço de entrada (S/T). Para os dados de Verão este método não permite identificar nenhum cluster. Fazendo a visualização dos dados a 3D e separando os dois clusters pela linha indicada, podemos confirmar a distribuição apresentada, tal como se observa na figura 3.

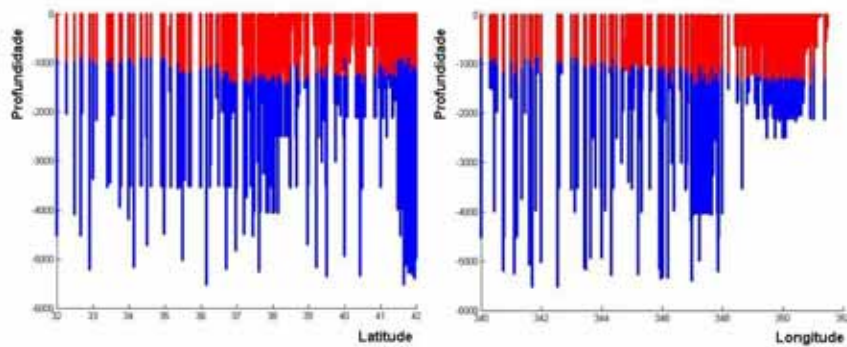


Figura 3 - Projecção dos dados 3D em Latitude e Longitude (dados de Inverno). Grupo à esquerda da linha na figura 2 a azul, grupo à direita a vermelho.

3- SOM com duas variáveis (T e S)

Realizando um SOM com os dados de Verão, que se revelaram mais problemáticos na análise visual, obtemos os resultados da figura 4.

A análise da UMAT [4] aponta para 3 zonas distintas, que correspondem a:

- 1- grandes profundidades (baixas T e S)
- 2- intermédia (valores intermédios e menos correlacionados para T e S)
- 3- superfície (altas T e S)

Verifica-se assim que o SOM permitiu identificar mais um cluster do que a simples observação visual do espaço de entrada.

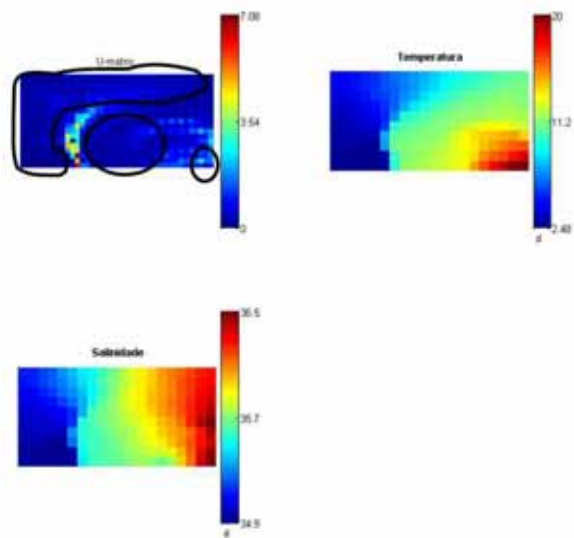


Figura 4 SOM para 2 variáveis (S e T)

4- Projecção dos neurónios e da UMAT sobre o espaço de entrada (2D)

A visualização dos clusters identificados sobre o espaço de entrada permite-nos entender a estrutura dos clusters e obter *insight* sobre o problema em estudo. Cada neurónio corresponde a um ponto no espaço de entrada (S/T) e tem associado a si um valor na UMAT.

Se codificarmos em cores o valor da UMAT para cada neurónio e o representarmos sobre o plano das variáveis de entrada, obtemos um gráfico com uma boa representação dos clusters e da sua relação com as variáveis de entrada, como se pode ver na figura 5.

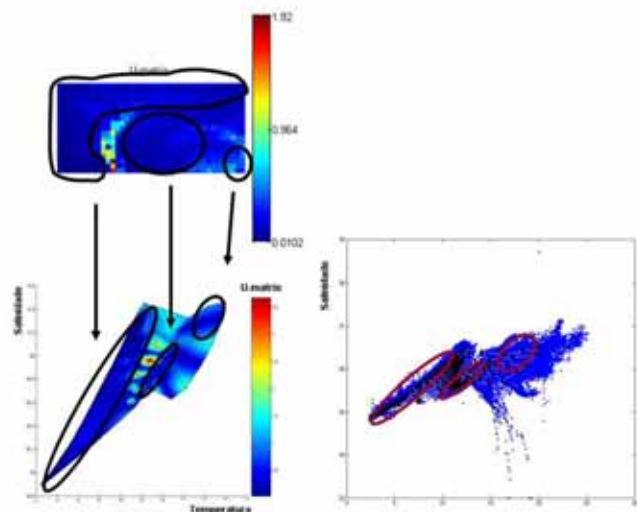


Figura 5 -Visualização dos clusters no espaço de saída vs espaço ST

A projecção acima foi feita interpolando os valores da UMAT que já são, por sua vez, uma interpolação dos verdadeiros valores que correspondem à distância entre neurónios. Daqui resulta um suavizar da distribuição de cores pela superfície que leva a esbater os valores, ou seja, as zonas de separação dos clusters. De qualquer forma, são perfeitamente visíveis os clusters e, a partir daqui, quais os pontos do plano de entrada correspondentes.

Conclui-se portanto que a projecção dos valores da UMAT sobre a distribuição de neurónios no espaço de entrada permitiu localizar os clusters em função dos valores de ST.

5 - Conclusões

Neste artigo mostrou-se a importância, a utilidade e a fragilidade da inspecção visual preliminar dos dados, neste caso possível dada a baixa dimensionalidade do espaço de entrada.

Mostrou-se depois que o SOM, em conjunto com a UMAT, permite uma melhor observação das características presentes nos dados. Sem abdicar da expressividade da inspecção visual o SOM permite identificar grupos que passariam despercebidos e permite fazer a identificação dos clusters mesmo em espaços de dimensão mais elevada.

Mostrou-se ainda que a principal separação entre tipos de características é feita em profundidade, sendo toda a costa portuguesa bastante homogénea em latitude e longitude. Constatou-se que, como esperado, há principalmente duas camadas no oceano, surgindo por vezes uma terceira camada de transição. Os métodos usados permitiram identificar claramente as fronteiras entre essas camadas.

6 - Bibliografia

1. Kohonen, T., *Self-Organizing Maps*. 3rd ed. Information Sciences. 2001, Berlin-Heidelberg: Springer. 501.
2. Bacao, F., V. Lobo, and M. Painho, *Self-organizing Maps as Substitutes for K-Means Clustering*, in *Lecture Notes in Computer Science*, V.S. Sunderam, et al., Editors. 2005, Springer-Verlag: Berlin Heidelberg. p. 476-483.
3. NODC, *National Oceanographic Data Centre*. 2006, <http://www.nodc.noaa.gov>.
4. Ultsch, A., et al. *Knowledge Extraction from Artificial Neural Networks and Applications*. in *Transputer-Anwender-Treffen*. 1993. Aachen: Springer Verlag.