Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática

Dissertação de Mestrado

Fuzzy clustering não supervisionado na detecção automática de regiões de *upwelling* a partir de mapas de temperatura da superfície oceânica

Pedro Guerra de Almeida Franco

Orientadora: Prof. Doutora Susana Nascimento

Trabalho apresentado no âmbito do Mestrado em Engenharia Informática, como requisito parcial para obtenção do grau de Mestre em Engenharia Informática.

Monte de Caparica
2009

Agradecimentos

À orientadora, Professora Doutora Susana Maria Santos Nascimento Martins Almeida, pela árdua tarefa de me acompanhar neste ano e meio, e pela capacidade de exposição da natureza de um trabalho de investigação científica e escrita de uma dissertação. Sabendo que todas as críticas tiveram o único propósito de melhorar esta dissertação e as minhas capacidades como investigador, reconhece-se também a paciência e compreensão demonstradas em todas as etapas, de maior ou menor dificuldade, que compuseram o trabalho desenvolvido nesta tese.

A Fátima Sousa, Joaquim Dias e Filipe Neves, oceanógrafos do Instituto de Oceanografia - Universidade de Lisboa, pelo fundamental trabalho de obtenção, pré-processamento e anotação de regiões de upwelling para as 61 imagens que compuseram o estudo feito, bem como pelas proveitosas reuniões que serviram para organizar muito do trabalho elaborado.

Aos amigos e colegas de curso, que permitiram uma melhor experiência durante estes últimos anos.

Principalmente, aos que me têm acompanhado desde sempre: Lena, Manela, Carmo, Ana, Guida e António, João e Paula, e, de um modo verdadeiramente geral, a toda a família, por tudo o que me deram e continuam, em todos os momentos, a dar. Serão sempre a minha preferência.

Esta dissertação foi desenvolvida no CENTRIA, Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa (FCT-UNL) e suportada por uma bolsa concedida pela Fundação para a Ciência e Tecnologia (FCT/MCTES) no âmbito do projecto *Learning Spatio-Temporal Oceanographic Patterns* - LSTOP (PTDC/EIA/68183/2006).

Resumo

O afloramento costeiro (*upwelling*) ao largo da costa de Portugal Continental é um fenómeno bem estudado na literatura oceanográfica. No entanto, existem poucos trabalhos na literatura científica sobre a sua detecção automática, em particular utilizando técnicas de *clustering*. Algoritmos de agrupamento difuso (*fuzzy clustering*) têm sido bastante explorados na área de detecção remota e segmentação de imagem, e investigação recente mostrou que essas técnicas conseguem resultados promissores na detecção do *upwelling* a partir de mapas de temperatura da superfície do oceano, obtidos por imagens de satélite. No trabalho a desenvolver nesta dissertação, propõe-se definir um método que consiga identificar automaticamente a região que define o fenómeno. Como objecto de estudo, foram analisados dois conjuntos independentes de mapas de temperatura, num total de 61 mapas, cobrindo a diversidade de cenários em que o upwelling ocorre.

Focando o domínio do problema, foi desenvolvido trabalho de pesquisa bibliográfica ao nível de literatura de referência e estudos mais recentes, principalmente sobre os temas de técnicas de agrupamento, agrupamento difuso e a sua aplicação à segmentação de imagem. Com base num dos algoritmos com mais influência na literatura, o Fuzzy c-means (FCM), foi desenvolvida uma nova abordagem, utilizando o método de inicialização *'Anomalous Pattern'*, que tenta resolver dois problemas base do FCM: a validação do melhor número de clusters e a dependência da inicialização aleatória. Após um estudo das condições de paragem do novo algoritmo, AP-FCM, estabeleceu-se uma parametrização que determina automaticamente um bom número de *clusters*. Análise aos resultados obtidos mostra que as segmentações geradas são de qualidade elevada, reproduzindo fidedignamente as estruturas presentes nos mapas originais, e que, computacionalmente, o AP-FCM é mais eficiente que o FCM. Foi ainda implementado um outro algoritmo, com base numa técnica de *Histogram Thresholding*, que, obtendo também boas segmentações, não permite uma parametrização para a definição automática do número de grupos. A partir das segmentações obtidas, foi desenvolvido um módulo de definição de *features*, a partir das quais se criou um critério composto que permite a identificação automática do *cluster* que delimita a região de upwelling.

Palavras-chave: Agrupamento difuso, número de *clusters*, detecção automática de afloramento costeiro, segmentação de imagem.

Abstract

Coastal upwelling by the shore of Continental Portugal is a well studied phenomena in oceanographic literature. However, there exists few work developed in the scientific literature about its automatic recognition, in particular by applying clustering techniques. Fuzzy clustering algorithms have been widely used in areas as remote sensing and image segmentation, and recent investigation has shown that those techniques achieve promising results in the identification of the upwelling regions, with the use of maps of sea surface temperatures obtained via satellite images. In the work to be developed in this dissertation, it is proposed to define a method that is able to detect automatically the areas where the event occurs. It were studied two independent sets of sea surface maps, in a total of 61, covering the whole range of situations where the upwelling is present.

Focusing on the problem's domain, it was developed work on bibliographic research on reference literature and recent work, mainly the topics of clustering techniques, fuzzy clustering and their application to image segmentation. Based on Fuzzy c-means (FCM), one of the most important reference algorithms, it was applied a new initialization, using the 'Anomalous Pattern' algorithm, that addresses two of the basic FCM issues: validating the best number of clusters and its dependance of the random initialization. After a study on the stop conditions of the new algorithm, AP-FCM, it was established a parametrization that determines automatically a good number of clusters. Analysis to the results showed that the achieved segmentations were of high quality, closely reproducing the original maps, and that, computational-wise, AP-FCM performed more efficiently than FCM. It was implemented yet another algorithm, based on a Histogram Thresholding technique, that, while obtaining good results, it supplied no method to automatically detect a good number of clusters. With the resulting segmentations, it was developed a module of feature definition, that allowed to define a criterion to identify the cluster that represents the border of the upwelling region.

Keywords: Fuzzy clustering, number of clusters, automatic detection of upwelling, image segmentation.

Tabela de símbolos

$\Re^p$	Conjunto dos números reais num espaço com p dimensões
∇f	Gradiente num campo escalar f
$\|\|a-b\|\|$	Distância euclideana entre dois pontos, a e b
C	Conjunto de clusters numa partição
c	Número de *clusters* numa partição
n	Número de entidades de um conjunto de dados
n_k	Cardinalidade de um cluster
p	Dimensionalidade de um conjunto de dados
T_k	Temperatura média do cluster k
$T(Y)$	Dispersão total de dados de um conjunto de dados normalizado Y
U	Matriz $n \times c$ de pertenças
u_{ik}	Valor de pertença na matriz U
V	Conjunto de protótipos de clusters
v_k	Protótipo de um cluster k
X	Conjunto de dados sob a forma de matriz de atributos
x_i	Entidade i de um conjunto de dados X
$\overline{x}$	Ponto médio de um conjunto de dados X
Y	Conjunto de dados normalizado

Conteúdo

1. Introdução

1.1 Descrição e contexto da Dissertação

Na Sociedade da Informação, as imagens digitais têm tomado um papel de cada vez maior importância. Desde máquinas fotográficas digitais de uso pessoal, a monitorização de eventos na superfície terrestre por satélite, passando por câmaras de vigilância ou identificação de cenários, os mecanismos de aquisição deste tipo de dados estão presentes nas mais variadas situações. Consequentemente, métodos de análise e processamento de imagem também têm evoluído bastante. A segmentação de imagem é um dos métodos de maior importância no processamento de imagem e, englobando-se numa área mais abrangente, de identificação e reconhecimento de padrões, tem aplicabilidade em diversas áreas, como a identificação de retinas ou a detecção de alterações no cérebro, a partir de ressonâncias magnéticas.

As técnicas de *clustering*, ou agrupamento [1], são das mais utilizadas para segmentar imagens e dedicam-se a dividir um determinado conjunto de dados em vários grupos, baseando-se em características ou atributos das entidades do conjunto. As suas versões *fuzzy* [2], também criam vários sub-conjuntos mas têm a particularidade de permitirem que uma determinada entidade seja associada a vários grupos, atribuindo um grau de pertença de uma entidade a cada grupo encontrado. No caso específico da segmentação de imagens, os grupos criados pelos algoritmos de *clustering* retratam conjuntos de píxeis, que são agrupados em regiões contíguas da imagem.

A partir de mapas de temperatura da superfície oceânica, a problemática da identificação automática de regiões de upwelling pode ser definida como um problema de segmentação de imagem e reconhecimento automático de padrões, ou seja, dado um mapa SST, pretende-se segmentar e identificar as regiões de upwelling das restantes águas, que podem ser classificadas como *background*. Para tal, serão utilizadas técnicas de *fuzzy clustering*. Destaque-se que se entende por regiões de upwelling como a união de todas as estruturas de águas presentes num mapa de temperaturas onde o fenómeno está presente.

Por outro lado, hoje em dia o conhecimento sobre a Natureza e o seu funcionamento toma um papel de destaque. Nesse âmbito, uma das áreas de maior relevo é a oceanografia, ciência que estuda a compreensão, descrição e previsão de fenómenos que se dão nos oceanos. O *coastal upwelling*, ou afloramento costeiro [3], é um desses eventos e traduz-se no surgimento de águas mais frias junto a uma costa terrestre.

Sem o auxílio de ferramentas computacionais, os oceanógrafos conseguem detectar a existência ou ausência do upwelling a partir de imagens de satélite, transformadas em mapas de

[1] Ao longo do texto será utilizado o termo *clustering* e *clusters*, já que a sua utilização é consensual, mesmo entre especialistas portugueses da área de aprendizagem automática e *machine learning*.

[2] Também para os vários tipos de *clustering* se seguirão as designações em inglês: *fuzzy clustering*, para agrupamento difuso, e *hard* ou *crisp clustering*, para agrupamento rígido.

[3] Será utilizado o termo upwelling, já que a sua utilização é consensual, mesmo entre oceanógrafos portugueses. Sendo o único tipo de upwelling tratado nesta dissertação, sempre que se referir o termo está sub-entendido que se refere ao upwelling costeiro.

temperatura da superfície oceânica, também designados por mapas SST (*Sea Surface Temperature*). Porém, não há um método que permita a detecção automática da sua região de ocorrência, quando o fenómeno está presente. Neste trabalho, um dos objectivos principais será definir um método que o permita fazer, dando a possibilidade de fazer um estudo mais profundo sobre a sua duração temporal, abrangência espacial e frequência.

Esta dissertação enquadra-se num projecto de investigação do Centro de Inteligência Artificial (CENTRIA) da Universidade Nova de Lisboa, em parceria com o Instituto de Oceanografia da Universidade de Lisboa, de onde são disponibilizados os mapas de temperatura do largo da costa portuguesa.

1.2 Reconhecimento de upwelling em imagens SST e sua detecção automática

Nas últimas décadas, os satélites tornaram-se elementos fundamentais para a monitorização de fenómenos à superfície da Terra e a medição da temperatura das águas é agora feita com recurso a imagens aéreas, ao contrário do passado, onde eram necessárias medições *in loco*. Estas medições por satélite e uma posterior transposição dos valores lidos para um mapa de temperaturas das águas dão origem a imagens onde é possível verificar a existência, ou não, do upwelling. A partir dos mapas de temperatura, a detecção a olho-nu é relativamente fácil de fazer por parte de oceanógrafos, com recurso a um ajuste da escala de cores para evidenciar o fenómeno, porém, devido ao elevado número de dados, é um trabalho que consome muito tempo e não existe ainda um método que consiga detectar automaticamente a existência e extensão das zonas de upwelling.

O upwelling costeiro é um fenómeno marítimo que se define como o aparecimento de águas vindas do fundo do oceano junto a uma costa terrestre. A sua existência surge como consequência do movimento de rotação natural da Terra e correntes aéreas (vento) que têm direcção paralela à costa, que conjuntamente criam a chamada Força de Coriolis. No caso específico da costa atlântica portuguesa, o upwelling acontece durante os meses de Verão (de Junho a Setembro, sensivelmente) e surge devido aos ventos que sopram de Norte para Sul. Esta combinação de factores leva a que a água que normalmente se encontraria à superfície seja submetida à Força de Coriolis e "empurrada" para longe da costa, surgindo no seu lugar águas vindas do fundo do oceano. Estas águas que surgem junto à costa são caracterizadas por terem uma temperatura mais reduzida que as que se encontravam anteriormente junto à costa e possuirem maior concentração de nutrientes (nitratos, fosfatos e silicatos), que associados à crescente exposição solar, criam condições ideais para o desenvolvimento de fitoplâncton [1, 2]. Esta situação leva a um aumento e proliferação de toda a cadeia alimentar marítima nas zonas onde o upwelling está presente.

Como é natural num país costeiro e onde a indústria piscatória tem muito impacto, o upwelling ao largo da costa portuguesa é um fenómeno já estudado e documentado, por exemplo, ao

nível da sua influência na fauna marítima [2, 3, 4]. A Figura 1.1 apresenta um mapa de temperaturas SST onde o upwelling está bem definido, ou seja, tem contornos nítidos e é de fácil visualização. A região que se identifica como pertencente ao upwelling corresponde aos píxeis mais frios, desde os que possuem tom azul até aos de tom verde, inclusivé.

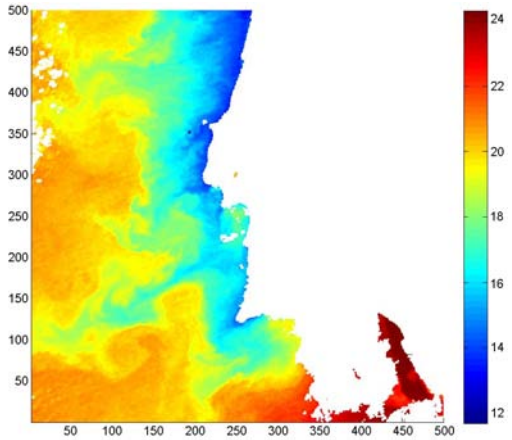


Figura 1.1 Mapa de temperaturas SST (2 de Agosto de 1998), com upwelling bem definido.

Por outro lado, devido ao aquecimento global e eventos associados, como o famoso El Niño e uma maior frequência condições meteorológicas extremas, nos últimos anos tem-se assistido a uma maior consciência e estudo de fenómenos climáticos à escala global, sendo os oceanos um dos elementos fulcrais nesses eventos. Cada vez mais, procuram-se desenvolver métodos que permitam maior conhecimento da Terra, onde se incluem modelos de detecção e previsão de temperaturas terrestres e oceânicas, precipitação, correntes marítimas, etc. O estudo do upwelling enquadra-se nesse âmbito.

Desde logo, uma das características do problema que dificulta o reconhecimento do fenómeno é o cenário de detecção remota sem a presença de objectos físicos ou modelo analítico que defina as regiões do upwelling. Assim, tendo em conta a imprecisão natural dos dados obtidos por detecção remota e não havendo um método padronize o conhecimento oceanográfico e o fenómeno, o reconhecimento automático das suas regiões é um problema inerentemente difuso.

Em [5], Marcello *et al.* descrevem as principais dificuldades da detecção de fenómeno a partir de imagens de satélite: i) haver frequentemente ruído nas imagens, como em situações com nuvens sobre o oceano; ii) a taxa de variação de temperaturas, mesmo quando há ocorrência de upwelling, ser reduzida, dificultando a tarefa de delinear fronteiras; iii) haver uma grande variação da morfologia do fenómeno, impedindo a definição de um modelo geométrico para a sua detecção e a criação de um modelo analítico para as estruturas encontradas.

A abordagem seguida em [5] para a detecção automática de regiões de upwelling começa com uma etapa de processamento da imagem (tratamento de zonas terrestres ou nebulosas), após a qual é definida uma região de interesse que, sabendo que o upwelling "cresce" a partir da costa na direcção do oceano, é composta pelos píxeis que se encontram junto à costa, e é criada

uma imagem segmentada, a partir do histograma da zona de interesse. Posteriormente, numa segunda fase, são definidos alguns pontos de interesse, como filamentos da região de upwelling que se propagam para o largo do oceano, e, nesses pontos, é aplicado um passo de crescimento iterativo, com técnicas de segmentação por histograma ou *watershed*.

Em [6], é proposto um modelo de previsão do upwelling, baseado em redes neuronais, a partir de mapas de temperatura oceânica e registos da velocidade do vento. Também com recurso a mapas de temperatura obtidos por satélite e informação sobre o vento obtida em bóias colocadas no oceano, em [7], Plattner *et al.* fazem a detecção de upwelling no Lago Michigan, mas a abordagem é fundamentalmente vocacionada para o estudo de várias características do fenómeno (frequência, localização, duração, extensão, magnitude e condições do vento), não se dedicando à sua detecção automática.

1.3 Motivação da Dissertação

Desde logo, uma das grandes motivações do trabalho desenvolvido está ligada ao estudo de dois dos problemas de fundo do algoritmo *Fuzzy c-means* (FCM), relacionados com o número de *clusters* em que se segmenta um determinado conjunto de dados e com a dependência da inicialização de protótipos para a obtenção de bons resultados. Aplicando uma técnica já proposta para algoritmos de *crisp clustering*, criar-se-à uma extensão ao FCM, o algoritmo de referência para *fuzzy clustering*, que se pretende que resolva ambas as questões e, ao mesmo tempo, seja computacionalmente mais eficiente, sem perder qualidade de resultados.

Após a resolução do problema da obtenção automática de boas segmentações, sem a necessidade de introdução do número de grupos, também se pretende modelar o conhecimento de oceanógrafos para definir um método que, a partir de uma boa segmentação consiga identificar automaticamente os *clusters* correspondentes às regiões de upwelling. Este facto é de fundamental importância, já que o processo de identificação visual do fenómeno consome bastante tempo. Hoje em dia, conseguindo-se adquirir grandes volumes de dados em pouco tempo, ter um método que automaticamente consiga identificar automaticamente as regiões de upwelling em centenas de mapas de temperatura seria um grande avanço.

O objectivo de aplicar técnicas de *fuzzy clustering* a um vasto conjunto de imagem será o de criar um grupo de bons resultados que possam constituir um conjunto de dados de referência para um estudo mais profundo do upwelling, extraindo informação específica do domínio do problema. Assim, o estudo terá como base duas amostras independentes de mapas de temperatura, uma relativa ao ano de 1998 e outra ao ano de 1999, que retratam uma diversidade de situações onde o fenómeno do upwelling ocorre. A utilização de duas amostras distintas permite também a aplicação de uma metodologia que utiliza uma amostra como conjunto de treino, possibilitando a validação dos resultados obtidos pela análise à outra amostra, funcionando como conjunto de teste.

Outro ponto importante prende-se com a crescente importância que os sistemas de detecção

automática vêm tomando no dia-a-dia. Como consequência do aumento dos métodos de aquisição de dados digitais, estes sistemas estão presentes em cada vez maior quantidade, tomando diversos papéis e funcionalidades, como, por exemplo, sistemas de segurança (reconhecimento de faces [8] ou leitura da íris), prevenção de acidentes (detecção de incêndios florestais [9] ou *tsunamis*), cuidados de saúde (análise de electro-cardiogramas [10] ou segmentação de ressonâncias magnéticas [11]), entre outros. O problema que se propõe resolver, com recurso a técnicas de segmentação de imagem, enquadra-se numa perspectiva de detecção automática de fenómenos naturais, sendo útil, por exemplo, para o estudo e construção de modelos climáticos e oceanográficos.

1.4 Contribuições da Dissertação

As principais contribuições do trabalho elaborado neste dissertação são as seguintes:

1. AP-FCM - Inicialização determinística do FCM com o algoritmo *Anomalous Pattern* [12]. Esta abordagem resolve duas questões problemáticas no algoritmo original, não sendo necessário introduzir o número de *clusters* como parâmetro de entrada e eliminando a geração aleatória de protótipos iniciais e consequente necessidade de múltiplas execuções para cada aplicação.

2. Estudo do AP-FCM para obtenção de uma boa segmentação de mapas SST - Para o problema em causa, nenhum dos índices de validação testados para o algoritmo FCM conseguiu obter bons resultados no que toca à determinação de um número de *clusters* que origine uma boa segmentação. Foi feito um estudo às diferentes condições de paragem do algoritmo AP-FCM e concluíu-se que, com a aplicação de uma condição que estuda a contribuição para a dispersão total de dados, as segmentações obtidas automaticamente, ou seja, sem indicar o número de *clusters*, são de boa qualidade. Destacaram-se ainda as vantagens de eficiência computacional resultantes da aplicação do AP-FCM, por comparação com o FCM.

3. Visualização de resultados de segmentação - Para uma boa análise dos resultados obtidos, foi desenvolvido um método que, a partir dos resultados de *fuzzy clustering*, permite visualizar as fronteiras dos *clusters* sobre os mapas de temperatura originais. Este método, juntamente com a visualização de mapas de pertença difusa, possibilita uma análise qualitativa dos resultados de segmentação.

4. Definição de *features* e critério para identificação do *cluster* de interesse que contém a frente de upwelling e sua anotação automática - Com recurso a uma análise de *features* definidas a partir de uma boa segmentação, foi construído um critério que consegue identificar com precisão o *cluster* que define a frente de upwelling, possibilitando a sua anotação. O método desenvolvido apresentou bons resultados e pode ser de grande utilidade para futuras análise do fenómeno do upwelling ao largo da Costa Portuguesa.

5. Definição de fronteiras difusas das regiões de upwelling - Para além da definição de uma fronteira do upwelling como uma simples linha a separar duas regiões no oceano, foi implementado um módulo que permite a definição e visualização de fronteiras difusas. Indo ao encontro da natureza do upwelling, as fronteiras difusas caracterizam-se por serem definidas também por uma região, entre *clusters* distintos, na qual se considera existir uma transição, mais ou menos suave, dependendo dos parâmetros utilizados para a criação da fronteira, nomeadamente da medida de *fuzziness* e α-cut.

1.5 Organização da Dissertação

A dissertação está organizada sob a seguinte forma: no Capítulo 2 é feito um estado-da-arte das técnicas de *clustering*, com especial ênfase para o *clustering* por partição e difuso, e abordada a problemática da segmentação de imagem. O Capítulo 3 apresenta o algoritmo proposto nesta dissertação e a composição, com base em *features* definidas, do critério criado para identificação automática das regiões de upwelling. No Capítulo 4 são apresentados e analisados os resultados do estudo experimental feito nesta dissertação. Para terminar, o Capítulo 5 contém a conclusão, resumindo os objectivos atingidos, contribuições efectuadas e trabalho futuro.

No Anexo A é feito um estudo experimental comparativo entre os algoritmos AP-FCM e FCM. No Anexo B são apresentados os 61 mapas que foram utilizados nesta dissertação. Os Anexos C a G apresentam os resultados (mapas de pertença e visualização de fronteiras de *clusters*) para os algoritmos estudados (AP_{C1}-FCM, AP_{C3}-FCM, AP_{C4}-FCM, FCM e *Iterative Thresholding*, respectivamente) e o Anexo H contém os resultados do critério definido para identificação automática das regiões de upwelling.

2. Algoritmos de *clustering* e aplicação em segmentação de imagem

Clustering é uma técnica de análise de dados cujo objectivo é separar um conjunto de entidades em vários grupos ou categorias distintas, com a particularidade de se pretender que entidades pertencentes a um determinado grupo sejam o mais possível semelhantes entre si e diferentes relativamente a entidades de grupos diferentes. Cada um dos grupos encontrados tem a designação de *cluster*. Na literatura científica, a designação de "entidade" também pode ser referida como "indíviduo" ou "ponto" de um determinado conjunto de dados. Jain e Dubes [13] definem um *cluster* como sendo formado por um conjunto de pontos:

> *"Clusters may be described as connected regions of a multi-dimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points." - [13]*

Por ser um método que não utiliza informação prévia sobre os dados ou grupos existentes, diz-se que o *clustering* é um tipo de "aprendizagem não supervisionada", por contraste com a "aprendizagem supervisionada" onde há conhecimento prévio sobre as classes existentes. Assim, pode-se dizer que é uma técnica que pesquisa os dados por alguma estrutura intrinsecamente presente nestes.

Em [14], os autores fazem uma distinção na aplicação desta técnica entre *clustering* para compreensão e *clustering* para utilidade. No primeiro caso, o objectivo é segmentar um determinado conjunto de dados em classes distintas, com base em atributos ou características das entidades do conjunto, permitindo a classificação de cada entidade com uma determinada etiqueta e uma análise posterior aos resultados, possibilitando o estudo e interpretação destes. Para este tipo de *clustering*, os autores dão como exemplos aplicações em áreas tão distintas como a biologia, pesquisa de dados na Internet, análise climática, entre outros. O *clustering* para a utilidade é descrito como um tipo de *clustering* que é utilizado quando se pretende obter alguma aplicação ou resultado a partir de um conjunto de dados. São dados os exemplos de sumarização, compressão de dados e busca eficiente de "vizinhos mais próximos", casos onde as entidades individuais podem ser abstraídas do *cluster* a que são associadas. Como será de fácil percepção, o objectivo do trabalho a desenvolver na dissertação enquadra-se no *clustering* para compreensão, procurando-se identificar num mapa de temperaturas oceânicas uma região que se possa etiquetar como sendo pertencente à região de upwelling.

Xu e Wunsch II [15] definiram as várias fases que tipicamente compõem o processo de *clustering* (ver Figura 2.1):

- Selecção e extracção de atributos: No primeiro passo, são escolhidos os atributos que mais interessam ou influenciam o problema em causa e/ou a estrutura de *clustering* presente nos dados. Uma boa escolha de atributos pode permitir uma computação mais eficaz, tanto espacial como temporalmente, bem como uma melhor representação da estrutura

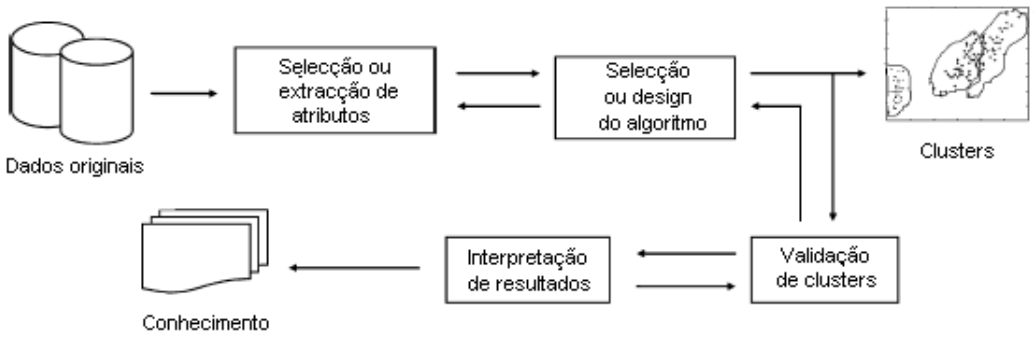


Figura 2.1 Fases de um processo de *clustering* [15]

real dos *clusters* existentes. Destaque-se a diferença entre selecção e extracção: no primeiro caso, utilizam-se atributos já existentes no conjunto de dados, sendo desejável que se seleccionem atributos que consigam separar, entre si, as entidades em estudo, enquanto que na extracção de atributos, estes são gerados a partir de atributos já existentes, criando um novo conjunto de medidas com um interesse especial, num determinado domínio de dados.

- Selecção ou design do algoritmo: A escolha do algoritmo a utilizar é um ponto fulcral nas técnicas de *clustering*. Uma vez que não se conhece um algoritmo que consiga bons resultados em todos os cenários, deve ser estudado o problema, os atributos escolhidos e os resultados esperados antes de fazer a selecção do algoritmo. Nesta etapa, também devem ser escolhidas a medida de similaridade e função objectivo que define o critério de *clustering*.

- Validação de *clusters*: Após a aplicação do algoritmo escolhido, é normalmente aplicado um passo para verificar a qualidade dos *clusters* obtidos. Mais informação sobre este tópico é apresentada na Secção 2.4.

- Interpretação de resultados: O último passo das técnicas de *clustering* é a análise de resultados, que, idealmente, permite retirar a informação pretendida dos dados, conforme a natureza do problema em causa.

2.1 Tipos de algoritmos de *clustering*

Dependendo do ponto de análise, encontram-se na literatura várias taxonomias distintas quanto à classificação do tipo de algoritmos de *clustering*. Esta questão prende-se com a quantidade de algoritmos e técnicas existentes, bem como com o facto de estes serem aplicados nos mais diversos cenários/problemas. Assim, é difícil conseguir estruturar toda a família de algoritmos

de uma mesma forma, uma vez não diferem em apenas uma característica do seu funcionamento, podendo tomar abordagens distintas ao problema ou sobrepôr-se em alguns tópicos e diferenciarem-se noutros. Nesta secção abordar-se-ão as principais técnicas em que se baseiam os algoritmos.

Em [16] são definidas algumas propriedades importantes que podem servir de diferenciação entre algoritmos distintos, como, por exemplo, o tipo de atributos que são aceites, a escalabilidade, a capacidade de trabalhar em conjuntos de dados de alta-dimensionalidade, a capacidade de descobrir *clusters* de formas irregulares, o tratamento de pontos desviados (*outliers*), a complexidade temporal e a dependência da ordem pela qual os dados são apresentados ao sistema.

Como dito anteriormente, existem diversas classificações distintas dos algoritmos existentes, contudo a divisão mais consensual será entre *clustering* hierárquico e *clustering* por partição, sendo referida, por exemplo, em [13, 14, 17].

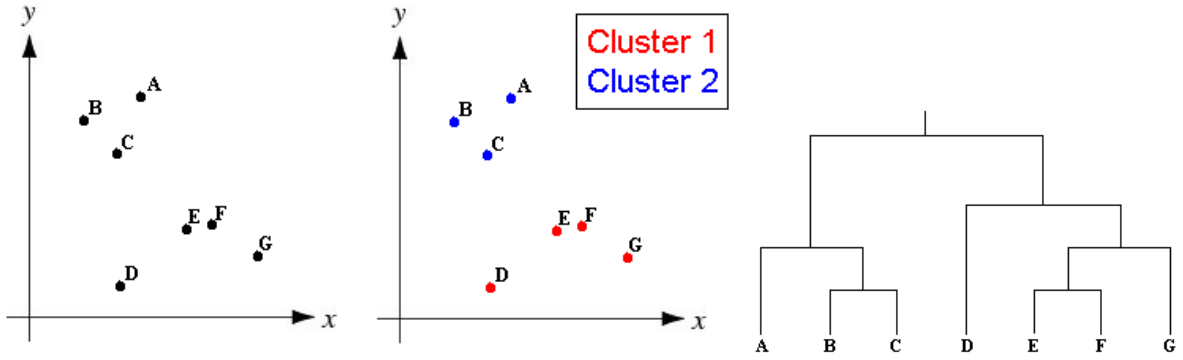
Em *clustering* por partição, o resultado final é uma divisão do grupo de dados em vários sub-grupos, que correspondem aos *clusters* descobertos (ver Figura 2.2(b)). Na Secção 2.2 este tipo de *clustering* é abordado em maior detalhe.

O *clustering* hierárquico organiza o conjunto de dados em forma de árvore binária, permitindo que cada *cluster* contenha sub-*clusters*. A árvore é visualizada sob a forma de um dendograma, cuja raíz é um *cluster* aglomerando todos os elementos do conjunto de dados e as folhas são cada elemento separadamente (ver Figura 2.2(c)). Este tipo de *clustering* pode ser também classificado em *clustering* aglomerativo e *clustering* divisivo. O *clustering* aglomerativo tem como ponto de partida todas as entidades isoladas e agrupa iterativamente aquelas que se encontrem mais próximas umas das outras, até se atingir um *cluster* com todas as entidades. No *clustering* divisivo o procedimento é inverso, partindo de um *cluster* único que engloba todas as entidades e separando até se atingir um estado com todas as entidades a representar um *cluster*. Por questões de eficiência, o *clustering* hierárquico divisivo não é habitualmente usado [15]. Para uma descrição mais detalhada sobre este método, há bastante informação disponível na bibliografia e poder-se-à consultar, por exemplo, [13, 14].

A Figura 2.2 ilustra a diferença entre os resultados obtidos com *clustering* por partição, onde cada entidade do conjunto é atribuída a um *cluster*, e com *clustering* hierárquico, onde é gerado um dendograma que agrupa as entidades em vários níveis sequenciais.

A abordagem a tomar durante a elaboração da dissertação será baseada em *clustering* por partição. Este método já obteve bons resultados em [1, 18], onde foi introduzida a aplicação de técnicas de *fuzzy clustering* para a detecção de regiões de upwelling. As técnicas de *fuzzy clustering*, descritas em maior detalhe na Secção 2.3, permitem associar uma entidade a vários grupos, mediante um grau de pertença. Em [19], comparam-se algoritmos de *clustering* hierárquicos e particionais para classificação de imagens multi-espectrais obtidas por satélite e o *clustering* hierárquico obtém piores resultados.

No entanto e como já referido, não se pode reduzir as diferenças entre técnicas de *clustering* com base na distinção de *cluster* particional e hierárquico, existindo uma grande variedade de abordagens e procedimentos como, por exemplo, baseado em redes neuronais, em densidade dos dados, em grafos, entre outros. Em [15, 16, 17] podem-se encontrar estados-da-arte



(a) (b) (c)

Figura 2.2 (a) Estrutura original de um conjunto de dados com 2 dimensões; (b) resultado da aplicação de um algoritmo de *clustering* por partição; (c) dendograma resultante da aplicação de um algoritmo de *clustering* hierárquico.

interessantes onde são abordados diversos tipos de algoritmos de *clustering*.

2.2 Clustering por partição crisp

A técnica de *clustering* particional associa um grupo de dados a *c clusters*, sendo que na versão original deste método (conhecida por *crisp clustering*), cada entidade é associada a um, e um só, *cluster*. Contrariamente, nas versões *fuzzy*, poderá ser associada a mais do que um *cluster*, com um grau de pertença a cada um. Esta última versão de *clustering* particional será abordada em maior detalhe na Secção 2.3. As associações são feitas com base nas características do grupo de dados e o resultado final pretendido é que entidades que possuam maiores semelhanças entre si sejam agrupadas num mesmo *cluster* e separadas de entidades que possuam um menor nível de semelhança.

Formulando matematicamente, dado um grupo de dados $X = \{x_1, ..., x_i, ..., x_n\}$, onde $x_i = (x_{i1}, x_{i2}, ..., x_{ip}) \in \Re^p$, sendo cada medida x_{ij} o j-ésimo atributo da entidade i, encontrar um grupo de c *clusters*, $C = \{C_1, ..., C_k, ..., C_c\} (c \leq n)$, tal que:

$$C_k \neq \emptyset, k = 1, ..., c, \tag{2.1}$$

$$\cup_{k=1}^{c} C_k = X, \tag{2.2}$$

$$C_k \cap C_q = \emptyset, k, q = 1, ..., c, k \neq q. \tag{2.3}$$

A primeira condição garante que nenhum *cluster* é vazio, a segunda impõe que todas as entidades sejam associadas a um *cluster* e a última refere-se à propriedade de que a intercepção entre quaisquer dois conjuntos seja nula, ou seja, uma entidade não pode ser associada a dois *clusters*. Note-se que esta última condição é relaxada nas técnicas de *fuzzy clustering* (ver Secção 2.3).

Tipicamente, o conjunto X é um parâmetro de entrada sob a forma de uma tabela $n \times p$, com n linhas, correspondentes às entidades, e p colunas, correspondentes aos atributos. Neste caso, diz-se que o modelo de dados de *input* é de matriz de atributos. Jain e Dubes [13] definem os vários modelos de dados de entrada, indicando dois grandes grupos de formatação de dados: matriz de atributos e matriz de proximidade ou similaridade. Neste último caso, o valor real de cada entidade, ou dos seus atributos, não é conhecido, conhecendo-se apenas uma medida de similaridade, ou dissimilaridade, calculada entre cada par de atributos. Os vários tipos de representação de dados podem ainda ser discriminados pela sua escala (quantitativa ou qualitativa) ou tipo de atributos (binário, discreto ou contínuo).

2.2.1 k-means

O algoritmo de *clustering* por partição mais reconhecido na literatura científica é o *k-means*, introduzido em 1967, por MacQueen (*cf.* [20, 21]). Muito do seu sucesso deve-se ao facto de ter uma implementação simples, bem como de ser computacionalmente rápido e eficiente [12]. Este método tem como objectivo encontrar uma partição que minimize as distâncias de cada entidade ao protótipo do *cluster* ao qual se encontra associada. O protótipo, ou centróide, de um *cluster* é definido pelo seu ponto médio. Formulando, para um *cluster* C_k, sendo n_k a sua cardinalidade, ou seja, o número de entidades a si atribuídas, o seu protótipo v_k é dado pela equação:

$$v_k = (1/n_k) \sum_{k=1}^{n_k} x_i : x_i \in C_k. \tag{2.4}$$

A associação de uma entidade a um *cluster* é feita pela regra do protótipo mais próximo [21]:

$$u_{ik} = \begin{cases} 1, d_{ik} < d_{ij}, \forall j \neq k \\ 0, \text{caso contrário,} \end{cases} \tag{2.5}$$

onde

$$d_{ik} = ||x_i - v_k|| \tag{2.6}$$

é uma distância definida entre a entidade x_i e o protótipo do *cluster* k, v_k. $U = [u_{ik}]$ é uma matriz de pertença, com dimensão $n \times c$, onde cada posição tem o valor 1, se a entidade i pertencer ao *cluster* k, ou 0, caso não pertença.

A noção de distância não é universal e é uma questão inerente ao espaço em que é utilizada ou ao tipo de problema. Dependendo do tipo de algoritmo e do problema em causa, o desejável é ter uma medida que se adapte o melhor possível de modo a obter bons resultados. Na versão original do *k-means* é utilizada a distância Euclideana:

$$d(x,y) = ||x-y|| = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}. \tag{2.7}$$

Esta medida, que corresponde à Distância de Minkowski de ordem 2, para além de ser a mais intuitiva (visualizando, corresponde à distância em linha recta entre dois pontos num espaço $\Re^2$ ou $\Re^3$), é das mais utilizadas em técnicas de *clustering*. Em [17, 21], apresentam-se descrições mais detalhadas desta e de outras medidas de dissimilaridade ou correlação entre entidades que podem ser aplicadas. O erro quadrático para um *cluster* C_k é a soma de todas as distâncias Euclideanas entre as entidades a si atribuídas e o seu centróide, medida também chamada de variação intra-*cluster* e definida por:

$$e_k^2 = \sum_{i=1}^{n_k} d_E(x_i, v_k), x_i \in C_k. \tag{2.8}$$

Facilmente se deduz que a fórmula dos erros totais é dada por:

$$E_c^2 = \sum_{k=1}^{c} e_k^2. \tag{2.9}$$

Utilizando a distância Euclideana ($||.||$), conclui-se que a função objectivo do algoritmo *k-means*, que se pretende minimizar, é dada por:

$$J(X, V, ||.||) = \sum_{k=1}^{c} \sum_{i=1}^{n} ||x_i - v_k||^2. \tag{2.10}$$

A diminuição desta medida pretende que se atinja um resultado de *clustering* onde todas as entidades estejam o mais próximo possível dos centróides dos grupos a que são associadas, favorecendo a compactação intra-*cluster*. Os passos do algoritmo *k-means* são descritos na Tabela 2.1. Refira-se há situações em que não é possível calcular os centróides, devido à própria natureza dos dados - quando o conjunto de dados possui atributos categóricos (verbais) e as distâncias entre entidades estão disponíveis numa matriz de similaridade, por exemplo -, pelo que há versões alternativas, como o PAM (*Partitioning Around Medoids*), onde como representantes de cada *cluster* são utilizados medóides, que são definidos como as entidades mais representativas de cada *cluster*.

As principais vantagens do *k-means* são a sua fácil implementação, ter uma convergência relativamente rápida e, para *clusters* esféricos, obter resultados bons, tanto em qualidade das partições obtidas como na rapidez de execução. Como desvantagens, apresenta os seguintes problemas: o número de *clusters* a serem encontrados é um parâmetro de entrada e afecta de sobremaneira os resultados obtidos, os resultados são dependentes da posição dos centróides iniciais e o facto de não funcionar bem em universos com *clusters* reais não esféricos. A questão do número de grupos é abordada na Secção 2.4.1. A dependência dos resultados em relação à posição inicial dos protótipos evidencia-se especialmente com a convergência do algoritmo para mínimos locais, ou seja, uma má inicialização dos protótipos pode levar a uma convergência para um mínimo local, não se obtendo a melhor estrutura de *clustering*. Normalmente, a solução passa por executar o algoritmo várias vezes, partindo de protótipos iniciais distintos e, posteriormente, verificar qual a melhor partição encontrada, em termos de minimização do

1. Escolher o número de *clusters* (c) e medida de distância a utilizar.
2. Seleccionar c pontos no espaço dos atributos das entidades existentes. Estes pontos são os centróides iniciais.
3. Associar cada entidade ao centróide que lhe é mais próximo, pela Equação (2.5).
4. Actualizar os centróides, pela Equação (2.4).
5. Repetir os passos 2 e 3 até se violar alguma condição de paragem (número de iterações, nenhuma entidade mudar de *cluster*, percentagem de entidades que alteram de *cluster* ser menor que um determinado valor limiar previamente estabelecido, etc.).

Tabela 2.1 Passos de execução do algoritmo *k-means*

critério de *clustering*. Esta situação introduz outro tópico muito importante nas técnicas de *clustering*: a validação de resultados (ver Secção 2.4).

A inicialização aleatória dos protótipos iniciais é a mais comum e em [22] é mostrado, comparando com outros três métodos, que os resultados por si obtidos não são de má qualidade, sendo ainda de destacar a sua fácil implementação. Em [12], Mirkin propõe uma nova inicialização para o algoritmo *k-means*, onde os protótipos iniciais, em vez de serem gerados aleatoriamente, são calculados pela iteração de um método denominado por *Anomalous Pattern*. No trabalho desenvolvido nesta dissertação, este algoritmo será introduzido como método de inicialização do algoritmo Fuzzy c-means (Sub-secção 2.3.1).

2.3 Clustering por partição *fuzzy*

Como referido na secção anterior, os algoritmos de *clustering* particional dedicam-se a encontrar divisões num dado conjunto e o seu resultado final é uma partição, onde cada entidade fica associada a um *cluster*. Assim, pode-se definir uma função de pertença das entidades aos *clusters* derivando a Equação (2.5), obtendo:

$$u_{ik} = \begin{cases} 1, \text{ entidade i pertence ao } \textit{cluster} \text{ k} \\ 0, \text{ caso contrário.} \end{cases} \tag{2.11}$$

Esta função retorna valores pertencentes ao conjunto $\{0,1\}$, assumindo o valor unitário nos casos em que a entidade i pertence ao *cluster* k. A função faz as associações de uma maneira

rígida, seguindo sempre a regra do protótipo mais próximo, e não tem em conta eventuais graduações de pertença como forma de graduar a tomada de decisão de pertencer a um grupo. Por exemplo, no caso extremo, quando uma entidade se encontra à mesma distância de dois protótipos, os algoritmos de *crisp clustering* associam-na somente a um dos *clusters*, normalmente dependente da ordem pela qual são tratados os dados. Note-se que no resultado final e, consequentemente, numa análise futura dos dados por parte de um utilizador, é ignorado o facto de a entidade estar tão próxima do *cluster* a que pertence como a um ao qual não fica associada e ainda, o valor de pertença nulo é igual para uma entidade que esteja significativamente mais próxima a um protótipo do que outra, que se encontre mais distante.

Há problemas onde pode ser interessante, ou até necessário, lidar com informação mais detalhada sobre as pertenças aos *clusters*. Os algoritmos de *fuzzy clustering* surgiram como aplicação da definição de conjuntos difusos, introduzidos em 1965 por Zadeh (*cf.* [17, 21, 23]). Estes conjuntos permitem lidar com as incertezas que surgem muitas vezes em diversas situações. As partições resultantes destes algoritmos fazem corresponder uma entidade a um *cluster* com um grau de pertença, no intervalo [0,1], onde um maior valor indica uma maior pertença. Quanto mais próximo (ou distante) estiver uma entidade de um protótipo de um *cluster*, maior (ou menor) será o seu grau de pertença a esse *cluster*. Assim, considere-se um conjunto $X = \{x_1, ..., x_i, ..., x_n\}$, de n entidades. Uma c-partição difusa associa n entidades a c *clusters*, com os valores de pertença a serem definidos por uma matriz $n \times c$, $U = [u_{ik}]$, sujeita às seguintes restrições [21]:

$$0 < u_{ik} < 1, \forall i = 1, ..., n, k = 1, ..., c, \tag{2.12}$$

$$\sum_{k=1}^{c} u_{ik} = 1, \forall i = 1, ..., n, \tag{2.13}$$

$$0 < \sum_{i=1}^{n} u_{ik} < n, \forall k = 1, ..., c. \tag{2.14}$$

A primeira condição estabelece os limites aos valores de pertença, a segunda assegura que os valores de pertença de uma entidade ficam associados exaustivamente em relação aos c *clusters*, ou seja, a soma das pertenças de uma entidade aos c grupos existentes tem o valor 1, e a última garante a não existência de *clusters* vazios. Facilmente se verifica que as partições resultantes dos algoritmos de *crisp clustering* são um subconjunto das partições dos algoritmos de *fuzzy clustering*, onde todos os valores de pertença têm o valor 0 ou 1.

Após a execução de um algoritmo de *fuzzy clustering*, normalmente aplica-se um passo *desfuzzificador*, onde a natureza difusa da partição é eliminada com a associação de cada entidade ao *cluster* cujo seu valor de pertença é máximo.

Para exemplificar a natureza difusa de um conjunto de dados, muitos autores recorrem ao conjunto da borboleta, introduzido por Ruspini (*cf.* [21]). Este exemplo é composto por 15 pontos dispersos num espaço cartesiano a duas dimensões, numa forma que se assemelha a uma borboleta (ver Figura 2.3(a)), com o número ideal de *clusters* igual a 2, correspondendo a cada uma das asas.

O resultado da aplicação de um algortimo de *crisp clustering* particional a este conjunto pode ser verificado na Figura 2.3(b). Analisando a estrutura do conjunto, facilmente se percebe que cada um dos dois *clusters* encontrados corresponde a uma das asas da "borboleta", porém não há nenhuma razão específica pela qual o ponto central seja associado a qualquer um dos grupos. Esse ponto tem tanta afinidade com a "asa esquerda", como com a "asa direita", no entanto, pela natureza dos algoritmos de *crisp clustering*, ele acaba por ficar agrupado com uma das asas, numa associação dependente da localização dos protótipos iniciais do algoritmo ou da ordem pelas quais as distâncias aos protótipos são tratadas. Por não lidar com estas ambiguidades, a matriz de pertenças de um algoritmo de *crisp clustering* reflecte essa associação aleatória (ver entidade 8 da Tabela 2.2). Por seu lado, os algoritmos de *fuzzy clustering* já conseguem lidar com essas ambiguidades, permitindo que o ponto central tenha um grau de pertença igual para ambos os *clusters* encontrados (Figura 2.3(c)). Analisando a Tabela 2.3, verifica-se que enquanto os pontos relativos a cada uma das asas possuem uma pertença muito elevada, acima de 0.85, ao *cluster* que engloba as entidades da asa a que pertence, a entidade 8, correspondente ao ponto central, tem uma pertença de 0.5 em ambos os *clusters*, indicando um elevado grau de ambiguidade. Note-se que, mesmo aplicando o passo *desfuzzificador* referido anteriormente, onde o ponto central poderia ser associado a um dos grupos aleatoriamente, a informação sobre a incerteza encontrada pode ficar sempre guardada na matriz de pertenças, possibilitando uma análise *a posteriori*.

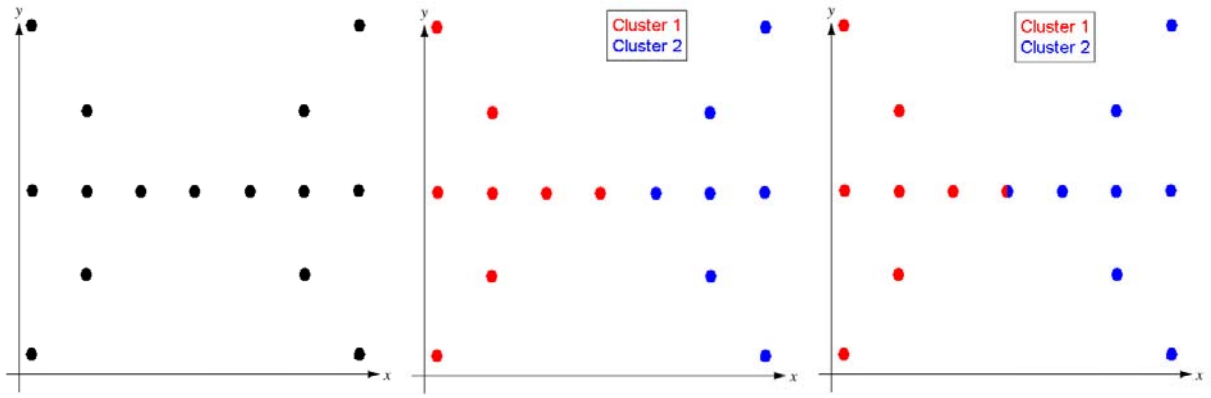


Figura 2.3 (a) Estrutura original do conjunto Borboleta; (b) associação resultante da aplicação de um algoritmo de *crisp clustering*; (c) associação resultante da aplicação de um algoritmo de *fuzzy clustering*, sem execução do passo *desfuzzificador*.

2.3.1 Fuzzy c-means (FCM)

O *Fuzzy c-means* (FCM), proposto por Dunn (*cf.* [21]) e generalizado por Bezdek [24], é um dos algoritmos difusos mais utilizados. Também há referências ao FCM com o nome *Fuzzy*

u_{ik}	*Cluster* 1	*Cluster* 2
1	1	0
2	1	0
3	1	0
4	1	0
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1
13	0	1
14	0	1
15	0	1

Tabela 2.2 Matriz de pertenças da aplicação de um algoritmo de *crisp clustering* ao conjunto Borboleta.

u_{ik}	*Cluster* 1	*Cluster* 2
1	0.8656	0.1344
2	0.9731	0.0269
3	0.8656	0.1344
4	0.9468	0.0532
5	0.9988	0.0012
6	0.9468	0.0532
7	0.8829	0.1171
8	0.5	0.5
9	0.1170	0.8830
10	0.0532	0.9468
11	0.0012	0.9988
12	0.0532	0.9468
13	0.1344	0.8656
14	0.0269	0.9731
15	0.1344	0.8656

Tabela 2.3 Matriz de pertenças da aplicação de um algoritmo de *fuzzy clustering* ao conjunto Borboleta.

k-means, por ser uma extensão do algoritmo *k-means*. A função objectivo deste algoritmo é extrapolada a partir da Equação (2.10), continuando a utilizar os erros quadráticos das distâncias de cada entidade aos protótipos a que pertence, neste caso tem em conta os graus de pertença, entre 0 e 1, a múltiplos *clusters*, resultando na função:

$$J_m(X,U,V,||.||) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^m ||x_i - v_k||^2. \tag{2.15}$$

Tal como no algoritmo *k-means*, valores inferiores da Equação (2.15) indicam melhores estruturas de *clustering*. O parâmetro m é designado por elemento *fuzzificador* e afecta o grau de sobreposição dos grupos encontrados. Normalmente, o valor utilizado para m é 2 [21]. Este parâmetro pode variar entre $[1,+\infty[$, sendo que para $m = 1$, o FCM comporta-se semelhantemente ao *k-means* e quando $m \to +\infty$, tende-se para um estado completamente difuso com os valores de pertença a tenderem todos para $\frac{1}{C}$ [21].

O algoritmo FCM utiliza uma optimização alternante [21, 24] entre os valores de pertença e os protótipos dos *clusters* para iterativamente chegar ao melhor resultado possível. Sendo U uma matriz $n \times c$ de pertenças, com $u_{ik} \in [0,1]$ a identificar a pertença da entidade i ao *cluster* k, e V uma matriz $c \times p$ com os protótipos dos *clusters*, onde $v_k \in \Re^p$ identifica o protótipo do *cluster* k e U^t, V^t, as matrizes de pertenças e com protótipos na iteração t do algoritmo, tem-se por optimização alternante do FCM a iteração dum ciclo onde $V^{t-1} \to U^t \to V^t$, iterando até $\left\|V^t - V^{t-1}\right\| < \varepsilon$, com ε a ser um valor reduzido e pré-definido ou se atingir alguma outra condição de paragem (ver Tabela 2.4).

Este algoritmo sofre das mesmas desvantagens que o *k-means*, nomeadamente ao nível da necessidade de introduzir como parâmetros de entrada o número de *clusters*, a dependência da inicialização dos protótipos e da forma dos *clusters* obtidos ser hiper-esférica. Tal como no *k-means*, para contornar a questão da dependência dos protótipos iniciais, ou seja, para garantir que não se obteve um mau resultado devido a uma má inicialização, normalmente executa-se o algoritmo várias vezes, partindo de diferentes pontos de inicialização, e para resultado final escolhe-se a melhor execução, tipicamente medida em termos da função objectivo do algoritmo. Note-se que o número de vezes que será necessário executar o algoritmo depende bastante do conjunto de dados a que este está a ser aplicado, variando conforme a sua complexidade (número de entidades e dimensionalidade), e do número de protótipos existentes. Para além da inerente natureza difusa do algoritmo, em [21] refere-se ainda a vantagem acrescida do FCM de convergir mais rapidamente que o *k-means*. Os passos do algoritmo estão descritos na Tabela 2.4.

1. Escolher o número de *clusters* (c), a medida de distância, o *fuzzificador* m e uma condição de paragem (ε reduzido, T iterações).
2. Inicializar a matriz de pertenças U^0 e $t = 0$.
3. Incrementar t.
4. $V^t = \Delta(U^t)$
5. $U^t = \Gamma(V^{t-1})$
6. Repetir os passos 3, 4 e 5 até se violar alguma condição de paragem ($\left\|V^t - V^{t-1}\right\| < \varepsilon$, $t = T$)

Tabela 2.4 Passos de execução do algoritmo FCM

As equações $U^t = \Gamma(V^{t-1})$ e $V^t = \Delta(U^t)$ são definidas respectivamente por:

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d^2(x_i, v_k)}{d^2(x_i, v_j)} \right)^{\frac{2}{m-1}} \right]^{-1}, \forall i,k \tag{2.16}$$

e

$$v_k = \frac{\sum_{i=1}^{n} (u_{ik})^m {x_i}^m}{\sum_{i=1}^{n} (u_{ik})}, \forall i. \tag{2.17}$$

Em [25], Gath e Geva propõem um algoritmo UFP-ONC ("*Unsupervised Fuzzy Partition - Optimal Number of Classes*") baseado em dois passos: (i) aplicação do algoritmo FCM com a

distância euclideana como métrica utilizada e, (ii) aplicação do FCM com uma distância exponencial, baseada numa estimativa de máxima verossimilhança. O UFP-ONC vai incrementando o número de *clusters*, introduzindo um novo protótipo em regiões onde as entidades existentes possuiem um valor de pertença baixo, em relação aos *clusters* já existentes. Destaque-se a eliminação do factor aleatório da inicialização original dos protótipos do FCM. A introdução do passo (ii) do algoritmo faz com que o UFP-ONC funcione bem em *clusters* de várias formas e densidades.

2.4 O problema da inicialização e validação no FCM

Nesta secção abordam-se questões de relevo na aplicação do algoritmo FCM, nomeadamente a problemática de definição do número de *clusters* numa determinada partição (Secção 2.4.1) e a validação de resultados (Secção 2.4.2).

2.4.1 Número de *clusters*

Uma das questões de fundo dos algoritmos de *clustering* particional é a definição do número de *clusters*. A grande maioria dos algoritmos exige algum parâmetro de entrada que afecta, directa ou indirectamente, o número de *clusters* a procurar. No entanto, percebe-se facilmente que a qualidade dos resultados obtidos está intimamente ligada ao número de grupos que se procuram e, se em alguns casos, por experiência ou conhecimento prévio, se sabem quantos grupos "naturais" existem num conjunto de dados, noutros isso não acontece, incluindo-se como parte do problema a sua descoberta. É neste último cenário que se inclui o problema a ser tratado nesta dissertação, com a indefinição sobre o "melhor" número de grupos que possibilitam a identificação automática das regiões de upwelling em cada mapa SST. Destaque-se o facto de que para o problema em causa, considera-se um bom número de *clusters* como aquele que gera uma segmentação que permita definir com o máximo de exactidão as regiões de upwelling.

A Figura 2.4 exibe uma hipotética situação em que o número real de *clusters* não é claramente bem definido, possibilitando interpretações ambíguas. A Figura 2.4(a) mostra a disposição das entidades do conjunto num eixo bidimensional, enquanto as Figuras 2.4(b) e (c) possuem a associação de cada entidade a um *cluster* resultante da aplicação de um algoritmo de *clustering* particional, procurando 2 e 6 grupos, respectivamente. Intuitivamente, ambos os resultados são aceitáveis. Há casos em que esta ambiguidade não necessita de ser tratada, sendo considerável como resultado correcto mais do que uma das partições resultantes. No entanto, noutras situações pode ser necessário haver um estudo mais profundo dos resultados obtidos, de modo a facultar mais informação que possibilite uma melhor decisão quanto ao número de *clusters* a utilizar.

As tentativas de descoberta do número ideal de *clusters* num dado grupo de dados, baseiam-se frequentemente na análise de resultados com o auxílio de índices de validação (Secção 2.4). A aplicação mais comum destas técnicas segue o prodecimento descrito na Tabela 2.5.

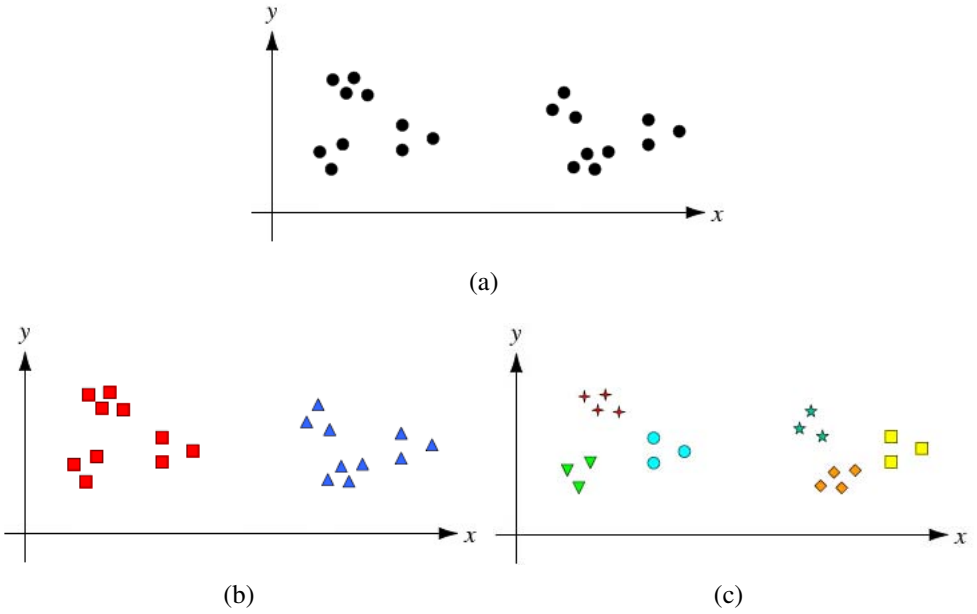


Figura 2.4 (a) Conjunto de dados original; (b) Resultado de aplicação de um algoritmo de *clustering* para 2 *clusters*; (c) Resultado de aplicação de um algoritmo de *clustering* para 6 *clusters*.

Há na literatura vários autores que propõem alterações a este modelo, como, por exemplo, em [26], Wang *et al.* derivam-no para um modelo que também abordando o problema da inicialização dos *clusters*, a cada alteração do número de *clusters* (de c_{min} a c_{max}) computam uma função que calcula o pior *cluster* e dividem-no em dois, fazendo a incrementação do número de *clusters* sem inicializações aleatórias.

Outro tipo de abordagem ao problema baseia-se em técnicas de amostragem. O príncipio base destes métodos é que se se pegar em duas (ou mais) amostras de um conjunto de dados, em que a percentagem f de entidades na amostra não seja muito reduzida ($f > 50\%$), se existir realmente uma estrutura de *clustering* nos dados originais, essa mesma estrutura se reflectirá nas amostras do conjunto, ou seja, o melhor número de *clusters* é aquele em que os resultados da aplicação de um algoritmo a várias amostras forem mais estáveis [27, 28]. Para verificar a estabilidade de resultados, há várias medidas que analisam a similaridade entre duas partições, como o *'matching coefficient'*, coeficiente de Jaccard [13] ou a partir de uma matriz de consenso [27] ou matriz de confusão.

2.4.2 Procedimento de validação

Nas técnicas de aprendizagem não supervisionada, em que o objectivo é a análise e exploração de dados, a validação de resultados é um dos pontos fundamentais. Note-se que mesmo não havendo conhecimento prévio sobre a existência, ou não, de grupos naturais nos dados, todos

1. Dado um conjunto de dados X, definir um algoritmo A, o número mínimo de *clusters* c_{min}, o número máximo de *clusters* c_{max} e algoritmo de validação V.
2. Inicializar MelhorCusto, MelhorPartição, MelhorNúmeroClusters.
3. De c_k = c_{min} até c_{max} :
   - Aplicar A(c_k) ao conjunto de dados e obter um resultado R.
   - Se custo V(R) melhor que MelhorCusto: MelhorCusto = custo V(R), MelhorPartição = R, MelhorNúmeroClusters = c_k.
4. O melhor número de *clusters* é MelhorNúmeroClusters, com a partição MelhorPartição.

Tabela 2.5 Procedimento genérico para fixar o melhor número de *clusters*, a partir de um índice de validação

os algoritmos de *clustering* particional conseguem descobrir uma estrutura de *clustering* num qualquer grupo de dados, mesmo que ela seja fictícia. Este facto ilustra a necessidade de encontrar um método que consiga analisar se a partição encontrada é de boa ou má qualidade. A validação de resultados é a funcionalidade que, após a aplicação do algoritmo de *clustering*, trata de trabalhar sobre os resultados de modo a obter mais informação sobre estes, dando ao utilizador mais dados para a interpretação (ver processo de *clustering* na Figura 2.1). Algumas das utilidades da validação de resultados são descritas em [14]:

- Definir a tendência de *clustering*, ou seja, verificar a existência ou ausência de padrões ou estruturas nos dados.
- Determinar o número correcto de *clusters* (ver Secção 2.4.1).
- Avaliar como é que os resultados da análise de *clustering* se enquadram nos dados, sem informação externa.
- Comparar os resultados da análise de *clustering* com resultados exteriores conhecidos, como informação sobre "*class labels*", ou seja, atribuições prévias de algumas entidades a determinados grupos.
- Comparar resultados de duas execuções diferentes para verificar qual a melhor.

Segundo Jain e Dubes [13], as técnicas de validação, também designadas de índices de validação, têm que possuir as seguintes características: i) sentido intuitivo; ii) teoria matemática fundamentada; iii) ser computável. Estes índices estão consensualmente divididos em três categorias:

- Validação interna: estas medidas avaliam a qualidade dos resultados obtidos com base em informação obtida exclusivamente dos dados e resultados. As medidas mais frequentemente utilizadas são a coesão de cada *cluster* e a separação entre *clusters* distintos. Relembre-se que estas são duas características que definem idealmente uma boa estrutura de *clustering*.
- Validação externa: as medidas de validação externa são utilizadas comparando os resultados obtidos com alguma informação extra, como dados relativos à classe de cada entidade.
- Validação relativa: a validação relativa é utilizada para comparar diferentes resultados de *clustering* ou *clusters* individuais.

Na literatura sobre o tema, facilmente se encontram inúmeros índices de validação. Na Secção 2.4.3 estão descritos os índices que foram aplicados no estudo feito no âmbito desta dissertação. Halkidi *et al.* [20] disponibilizam uma boa análise a diferentes técnicas e índices de validação.

2.4.3 Índices de Validação

Nesta sub-secção serão apresentados e descritos os índices de validação aplicados à validação do algoritmo FCM, no trabalho desenvolvido. Teve-se o objectivo de aplicar índices de referência e bem estabelecidos na literatura (Xie-Beni [29] e Fukuyama-Sugeno *cf.* [30]), bem como um índice relativamente recente, do ano de 2003, proposto por Pakhira *et al.* [31].

2.4.3.1 Índice de Xie-Beni

O índice de Xie-Beni, proposto em [29], é um dos índice de validação interna para *fuzzy clustering* com mais reconhecimento e utilização na área. Este método utiliza os protótipos dos *clusters* e os valores de pertença de cada entidade aos *clusters* resultantes da execução de um algoritmo e, através das medidas de separação entre *clusters* e compactação de cada *cluster*, avalia a qualidade dos resultados. O índice é definido por:

$$V_{XB} = \frac{\sum_{k=1}^{c}\sum_{i=1}^{n} u_{ik}^2 \|v_k - x_i\|^2}{n \min_{k,j} \|v_k - v_j\|}. \tag{2.18}$$

Note-se que o valor referente a compactação é constituído pela expressão

$$\frac{\sum_{k=1}^{c}\sum_{i=1}^{n} u_{ik}^2 \|v_k - x_i\|^2}{n} \tag{2.19}$$

e refere-se ao valor médio das distâncias de cada entidade aos protótipos a que pertence, com o respectivo peso do valor de pertença. A separação entre *clusters* é definida por

$$\min_{k,j} \|v_k - v_j\|, \tag{2.20}$$

ou seja, a distância entre os dois protótipos mais próximos um do outro. Com o objectivo de ter uma estrutura de *clustering* bem definida, com compactação intra-*cluster* e separação inter-*cluster*, quanto mais reduzido for o valor do índice de Xie-Beni, melhor é qualidade de uma partição.

Os maiores problemas deste método são o facto de ser monotonicamente decrescente com o aumento do número de *clusters*, quando este tende para o número de entidades - note-se que esta situação é um cenário limite teórico e não é uma questão a ter em conta no trabalho desenvolvido, já que o número de grupos com que se trabalhou nunca esteve perto do número total de entidades -, ser dependende do factor de *fuzzificação m* e não ter em conta a geometria natural dos dados [20].

2.4.3.2 Índice de Fukuyama-Sugeno

Outro índice de validação que utiliza a matriz de pertenças e os protótipos dos *clusters* foi proposto por Fukuyama e Sugeno (*cf.* [30]). Este índice baseia-se na diferença entre dois termos: $J_m(u,v)$ que representa a compactação intra-cluster do conjunto difuso e $K_m(u,v)$, que funciona como medida de separação inter-*cluster*:

$$\begin{aligned} V_{FS} &= J_m(u,v) - K_m(u,v) \\ &= \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ik}^m ||x_i - v_k||^2 - \sum_{k=1}^{c}\sum_{i=1}^{n} u_{ik}^m ||v_k - \overline{v}||^2, \end{aligned} \tag{2.21}$$

onde $\overline{v}$ é a média de todos os protótipos, ou seja, $\overline{v} = \sum_{k=1}^{c} \frac{v_k}{c}$. Pretendendo atingir um resultado com *clusters* compactos e separados uns dos outros, as partições óptimas são obtidas pelo valor de c que minimize o índice.

2.4.3.3 Índice de Pakhira-Bandyopadhyay-Maulik

O índice PBMF, proposto por Pakhira *et al.* [31], é a versão *fuzzy* do índice PBM, proposto também em [31], e favorece a criação de um grupo reduzido de *clusters* compactos, com uma grande separação entre, pelo menos, dois *clusters*. Sendo c o número de *clusters*, $D_c = max_{k,j=1}^{c} \left\|v_k - v_j\right\|$ identifica a distância máxima entre pares de protótipos de *clusters*:

$$V_{PBMF} = (\frac{1}{c} \times \frac{E_1}{J_m} \times D_c)^2. \tag{2.22}$$

Este índice é composto por três factores que "competem" entre si à medida que o número de *clusters* aumenta, na medida em que o primeiro factor diminui com o aumento de c, ao contrário

dos restantes. O primeiro factor, $\frac{1}{c}$, indica a divisibilidade de um sistema de *clustering*. O segundo factor tem em conta a separação intra-cluster, ou compactação, de uma partição, sendo que o termo $E_1 = \sum_{i=1}^{n} u_{i1} \|x_i - v_1\|$ é fixo para um determinado conjunto de dados e serve para evitar que este factor fique muito reduzido. O melhor número de *clusters* é indicado pela maximização do índice.

2.4.3.4 Outros índices

Outros dos importantes e reconhecidos índices de validação interna são o *Partition Coefficient* (V_{pc}) e *Partition Entropy* (V_{pe}) [**?**, 32]. Estes dois índices são calculados apenas com base nas matrizes de pertença resultantes da aplicação de um algoritmo. Ambos os índices têm como base teórica a ideia de que uma partição menos difusa, ou seja, com valores de pertença mais elevados, corresponde a uma melhor partição. A sua definição é a seguinte:

$$V_{pc} = \frac{\sum_i^n \sum_k^c u_{ik}^2}{n} \tag{2.23}$$

e

$$V_{pe} = \frac{-\sum_i^n \sum_k^c (u_{ik} \log u_{ik})}{n}. \tag{2.24}$$

Os seus resultados óptimos são um valor máximo para (V_{pc}) e mínimo para (V_{pe}). A maior desvantagem para ambos os índices prende-se com o facto de, por utilizar apenas os valores de pertença, ignorar a estrutura patente nos dados [26].

2.5 Segmentação de imagem por *clustering*

A segmentação de imagem é uma técnica, incluída numa área mais global de processamento e análise de imagem, que se dedica à divisão de uma imagem em vários segmentos que se pretendem distintos entre si e homogéneos no seu interior. Esta técnica tem uma grande utilidade em áreas muito diversas que vão desde a visão de robots, reconhecimento de objectos, processamento de imagens aéreas e médicas, havendo na literatura diversas técnicas e algoritmos com aplicações nestas área. Em [33], as técnicas de segmentação de imagem são divididas em: a) baseadas em espaço de características, que utilizam propriedades das imagens ou dos píxeis que a compõem (*feature-space based*); b) técnicas que também utilizam informação espacial dos píxeis para formar os segmentos (*image-domain based*); c) técnicas que abordam questões físicas que afectem a imagem, como sombras e luminosidades diferentes (*physics based*). Em [34], é introduzido ainda uma nova definição para técnicas baseadas em detecção de diferenças entre píxeis (*edge based*). As técnicas que utilizam *clustering*, tanto versões *crisp* como *fuzzy*, enquadram-se na segmentação *feature-space based*. A própria definição do problema feita por Jain e Dubes em [13], permite ver os pontos de contacto entre a segmentação de imagem e as técnicas de *clustering* particional:

> *"The problem of image segmentation can be stated as follows: Partition a given image into regions or segments such that pixels belonging to a region are more similar to each other than pixels belonging to different regions. [...] A clustering method can group the pixels in the feature space into clusters. These clusters are then mapped to the spatial domain to display the segmented image." - [13]*

Facilmente se faz a transposição das Equações (2.1)-(2.3), referentes às técnicas de *clustering* particional em geral, para a segmentação de imagem: tendo uma imagem com largura m e altura n, o objectivo é agrupar o conjunto de $m \times n$ pontos, composto por todos os píxeis, criando c zonas adjacentes e homogéneas relativamente a alguma característica da imagem. Essas características podem ser a cor da imagem (válido para vários espaços de cores como, por exemplo, RGB, LUV, escala de cinzentos, etc.), onde cada píxel é representado por uma entidade que possui tantos atributos quantos sejam necessários para representar a cor (três para RGB ou LUV, um para uma escala de tons de cinzento de 0 a 255, por exemplo) ou alguma outra qualidade presente na imagem ou nos seus píxeis (como temperatura, textura, intensidade, etc.).

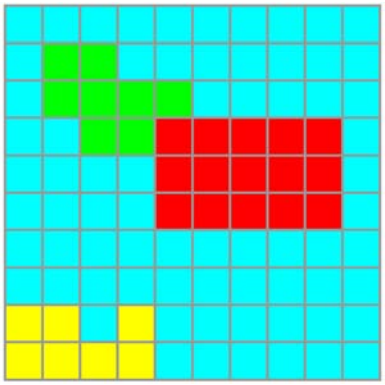
No trabalho a desenvolver, a segmentação será feita exclusivamente através dos valores de temperatura da superfície oceânica presentes em cada píxel dos mapas disponíveis. Note-se que apesar de ser possível visualizar os mapas de temperatura como uma imagem a cores, essa informação é obtida mapeando os valores de uma matriz de temperaturas num espaço de cores e as segmentações são feitas tratando cada píxel como entidades distintas e que se agrupam em *clusters* e não fazendo uso da cor visualizada.

Na Figura 2.5 exemplifica-se a segmentação de uma imagem em vários grupos distintos. No exemplo, a figura da alínea (a) é segmentada em 4 *clusters*, com base na cor dos píxeis. A alínea (b) simula o resultado da aplicação de um algoritmo de segmentação à figura da alínea (a). Destaque-se que cada um dos *clusters* ficou a agrupar grupos de píxeis de cores diferentes: verdes, vermelhos, amarelos e azuis. Este exemplo é também um bom caso para voltar a destacar a importância que o número de *clusters* tem na obtenção de uma boa segmentação. Se uma segmentação fosse feita para dividir a imagem em 3 grupos, o resultado não seria o ideal, podendo-se obter um *cluster* que agrupasse conjuntamente píxeis verdes e vermelhos, ou azuis e amarelos, por exemplo.

Tal como nas técnicas de *clustering*, a segmentação de imagem é um tema de grande importância e, como tal, há uma grande quantidade de técnicas que abordam este problema sobre uma perspectiva diferente. Em [33] pode-se encontrar uma descrição de várias abordagens ao problema de segmentação de imagem.

2.5.1 Segmentação de imagem por *fuzzy clustering*

No trabalho a desenvolver, o foco estará sobre técnicas que abordam a segmentação de imagem com *fuzzy clustering*. O algoritmo FCM é dos mais utilizados com sucesso nessa abordagem, sendo de destacar a importância da detecção de alterações do cérebro com imagens de ressonâncias magnéticas tem na literatura [35, 36, 37]. No âmbito da aplicação desta dissertação,



(a)

1	1	1	1	1	1	1	1	1	1
1	4	4	1	1	1	1	1	1	1
1	4	4	4	4	1	1	1	1	1
1	1	4	4	2	2	2	2	2	1
1	1	1	1	2	2	2	2	2	1
1	1	1	1	2	2	2	2	2	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
3	3	1	3	1	1	1	1	1	1
3	3	3	3	1	1	1	1	1	1

(b)

Figura 2.5 (a) Imagem a ser segmentada; (b) Informação em cada píxel sobre o *cluster* a que foi associado.

segmentação de imagens de satélite da superfície terrestre, também se encontram alguns estudos como em [1, 38, 39].

Uma das principais desvantagens da utilização do algoritmo FCM na segmentação de imagem prende-se com o facto de este utilizar apenas os valores dos píxeis, seja cor ou uma outra característica, para fazer a segmentação, ignorando a informação espacial dos píxeis [35, 37, 40]. Note-se que esta questão é particularmente importante devido à própria definição de segmentação de imagem, que refere que, para além de se pretender agrupar vários píxeis com um valor semelhante, estes devem ser contíguos. Esse factor faz com que o FCM seja particularmente sensível a ruído e *outliers* (píxeis com valores anómalos relativamente aos restantes).

Comparativamente às técnicas de *crisp clustering*, as de *fuzzy clustering* são mais utilizadas na segmentação de imagem principalmente por dois factores: a própria natureza difusa de muitas imagens, onde a associação de um píxel a um determinado *cluster* pode não ser algo trivial, principalmente em imagens reais, não criadas artificalmente, e também por o resultado incluir os valores de pertença, preservando mais informação sobre a imagem e os seus píxeis [35].

Uma das técnicas recentes com maior relevância trata-se do chamado *spectral clustering* [41, 42]. Tratando uma imagem como um conjunto de píxeis, o objectivo passa por criar um grafo, onde as arestas são pesadas com o valor de uma determinada medida de similaridade, e o objectivo passa por segmentar sub-grafos, que se pretendem que sejam o mais distintos possíveis entre si. Esta abordagem é feita com o cálculo de vectores-próprios a partir de uma matriz de similaridade e o passo de *clustering* é feito com a aplicação de um algoritmo de *clustering* sobre uma composição dos vectores-próprios obtidos. Em [41], é proposta um tipo de segmentação de grafos que aborda não só a dissimilaridade entre sub-grafos distintos mas também utiliza uma similaridade intra-cluster e é utilizado o *k-means* como algoritmo de *clustering*. Esta técnica de *clustering* tem, as vantagens de conseguir descobrir *clusters* de formas para além de

hiper-esféricas, como espirais e de ser bastante eficiente, mesmo em grandes conjuntos de dados, desde que se defina uma boa matriz de similaridade. Em oposição ao *k-means* ou FCM, destaca-se ainda a vantagem de eliminar a necessidade de executar várias vezes o algoritmo para evitar mínimos locais [42]. Este tipo de *clustering* não foi explorado no estudo elaborado devido à complexidade inerente ao cálculo de matrizes de similaridade, para os píxeis dos mapas de temperatura do problema em causa.

Como referido na Secção 2.1, não existe um algoritmo de *clustering* que satisfaça todos os problemas e, da mesma forma, mesmo sendo considerado um bom algoritmo para a segmentação de imagem, a qualidade dos resultados obtidos pelo FCM varia conforme o domínio do problema ou tipo de imagem em questão. Por esta razão, muitos dos trabalhos desenvolvidos na literatura sobre a aplicação de *fuzzy clustering* para segmentação de imagem tratam de introduzir alterações ao FCM original para que este se adapte melhor a um determinado problema.

Em [40], para reduzir o efeito do ruído é introduzida uma função espacial, com recurso aos valores de pertença, definida por:

$$h_{ki} = \sum_{j \in NB(x_k)} u_{ji}, \tag{2.25}$$

onde $NB(x_k)$ representa a vizinhança do píxel x_k, podendo ser definida como uma janela quadrada de raio variante. Uma nova matriz de pertenças é calculada pela fórmula:

$$u_{ki}^{l} = \frac{u_{ki}^{p} h_{ki}^{q}}{\sum_{i=1}^{c} u_{ki}^{p} h_{ki}^{q}}, \tag{2.26}$$

em que p e q são parâmetros que permitem controlar o peso relativo que a vizinhança tem no cálculo da nova matriz U. O algoritmo é designado de $sFCM_{p,q}$ e para o caso em que $p = 1, q = 0$ é idêntico ao FCM original. Também para tratar imagens com ruído, Ahmed *et al.* [43] propuseram uma alteração à função de minimização J_m (Equação (2.15)), com o acrescentar de um termo que utiliza a vizinhança dos píxeis, enquanto em [35], é introduzido um filtro que minimiza as diferenças entre píxeis vizinhos. Em [36], aplicam-se funções kernel na função objectivo para introduzir restrições espaciais.

As alterações referidas até aqui baseiam-se todas em transformações da função objectivo do FCM. Alternativamente, em [23, 44], também para a utilização de informação espacial, nomeadamente a relação entre cada píxel e os seus vizinhos, propõe-se que, em cada píxel, se incorpore nos atributos os valores dos píxeis vizinhos. Por exemplo, se um píxel for representado por um atributo $x_{i,j}$, se se utilizar como vizinhança uma janela de tamanho 3×3, passaria a ser representado pelos valores $(x_{i-1,j-1}, x_{i-1,j}, x_{i-1,j+1}, x_{i,j-1}, x_{i,j}, x_{i,j+1}, x_{i+1,j-1}, x_{i+1,j}, x_{i+1,j+1})$. Esta solução, mesmo variando o tipo de vizinhança, tem o problema do aumento da dimensionalidade do problema. No exemplo dado, passa-se de representar um píxel com uma para nove dimensões, sendo que esse aumento é pode aumentar com a subida da dimensionalidade original ou da dimensão da janela de vizinhança utilizada. Em [23], são propostos dois métodos para combater esse inconveniente: a utilização de uma média de valores da vizinhança,

passando um píxel a ser representado, no caso com uma vizinhança 3x3, por $(x_{i,j}\ \overline{V_{i,j}})$, em que $\overline{V_{i,j}} = 1/9\sum_{r=-1}^{1}\sum_{c=-1}^{1} x_{i+r,j+c}$ e o aumento de dimensionalidade passa a ser "apenas" para o dobro, ou utilizar apenas valores limite da vizinhança, como os cantos da vizinhança, sendo que neste caso o aumento depende dos valores utilizados.

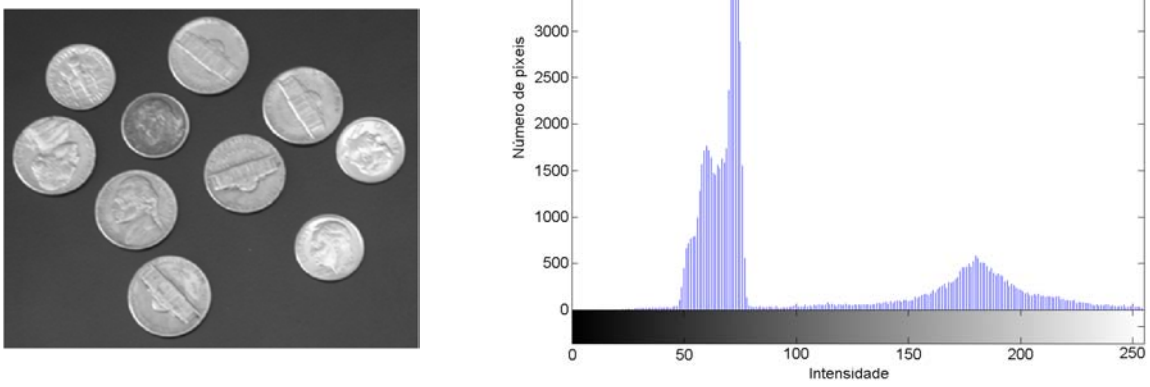
Em [45], é proposta uma versão do FCM parcialmente supervisionada, designada por ssFCM. Segundo os autores, o ssFCM resolve o problema do número de *clusters* numa segmentação bem como faz a associação entre os *clusters* e cada classe física distinta. No entanto, é ressalvado que se assume que todas as classes físicas contém previamente entidades a si associadas. Para o problema em causa, apenas se poderia indicar um conjunto das águas mais frias e mais quentes de cada mapa de temperatura como pertencentes ao *cluster* mais frio e ao *cluster* mais quente de uma segmentação. Porém, para resultados de *clustering* com um número relativamente elevado de grupos (5, 6 ou 7, por exemplo) não é possível classificar píxeis para todos os *clusters*, já que não se está na presença de objectos físicos distintos.

2.5.2 Segmentação de imagem por Histogram Thresholding

Uma outra técnica de segmentação de imagem muito reconhecida na literatura é denominada por *Histogram Thresholding*. Como o nome indica, estas técnicas utilizam histogramas da imagem alvo para atingirem uma determinada segmentação. Um histograma é uma representação da distribuição de frequências de um determinado atributo, sendo visualizado normalmente sob a forma de um gráfico de barras. No caso específico da segmentação de imagem esse atributo é obtido a partir de uma medida ou intensidade de cada píxel existente. Por exemplo, a Figura 2.6(b) contém um histograma da imagem representada na Figura 2.6(a) [46]. Destaque-se a grande frequência de píxeis com uma intensidade reduzida, entre 50 e 80, de tons mais escuros que correspondem ao *background* da imagem e a frequência elevada de píxeis de tons mais claros, representativos das várias moedas, com vários níveis de brilho.

Nas técnicas de *Histogram Thresholding* a determinação de um bom número de grupos é feita através da análise à morfologia do histograma, em termos de "picos" e "vales". Entende-se por "picos" como regiões contíguas da medida utilizada para a construção do histograma, com uma elevada frequência e "vales" como as regiões contíguas entre os "picos". Intuitivamente, pretende-se encontrar conjuntos semelhantes de píxeis, agrupados num "pico" do histograma, que estão separados da restante imagem por regiões de frequência inferior. Um bom número de *clusters* poderá ser encontrado pelo cálculo do número de "picos" do histograma. Na Figura 2.6(b), consegue-se distinguir claramente o *background* da imagem dos restantes objectos, através da identificação de um "vale" entre os dois picos de frequência superior, um correspondente ao fundo e outro relativo ao conjunto de moedas.

As versões mais comuns das técnicas de *Histogram Thresholding* são aplicadas a imagens em tons de cinzento, sendo que o atributo medido é precisamente o tom de cada píxel (variando entre 0 e 255, tipicamente). Para o problema da detecção do upwelling a partir de mapas de temperatura SST, o histograma é calculado com base na frequência dos valores de temperatura dos píxeis.



(a) (b)

Figura 2.6 (a) Imagem exemplificativa com um conjunto de moedas; (b) Histograma da imagem da alínea (a).

Uma das técnicas de "Histogram Thresholding" com mais reconhecimento na literatura foi proposta por Ridler e Calvard [47] e define um *threshold* inicial para segmentar a imagem e adapta-o iterativamente à medida dos dados do conjunto. A adaptação do *threshold* é feita através do cálculo de uma média ponderada entre os píxeis de valor superior e os de valor inferior ao *threshold* actual, garantindo assim que o *threshold* final separa os píxeis do histograma com base na sua estrutura real. Os passos do algoritmo estão descritos na Tabela 2.6.

Este método de segmentação gera resultados apenas com dois grupos, correspondentes aos píxeis de valor superior ao *threshold* final num grupo e os píxeis de valor inferior noutro grupo. Na Secção 4.5, referente ao estudo experimental por aplicação deste algoritmo, é apresentado uma versão iterativa que permite a obtenção de segmentações com um número de *clusters* superior.

1. Construir o histograma do conjunto de dados e um threshold τ inicial (por exemplo, a média do conjunto de dados).
2. Segmentar dois sub-conjuntos: R_1, conjunto de pontos com valor inferior a τ e, R_2, conjunto de pontos com valor igual ou superior a τ.
3. Sendo τ_1 a média de R_1 e τ_2 a média de R_2, re-calcular $\tau = \frac{\tau_1+\tau_2}{2}$. Se τ alterar o valor, voltar ao passo 2, caso contrário termina com a segmentação em dois grupos R_1 e R_2.

Tabela 2.6 Passos de execução do algoritmo iterativo de segmentação por Histogram Thresholding [47].

3. Anomalous-Pattern FCM para segmentação e anotação de regiões de upwelling em mapas SST

Neste capítulo são descritos os métodos implementados e estudados para a resolução do problema em causa nesta dissertação. Na Secção 3.1 é apresentado o algoritmo *Anomalous Pattern*-FCM e as suas condições de paragem. A Secção 3.2 contém a definição das *features* propostas no contexto do problema e a composição do critério que permite a anotação automática de regiões de upwelling. Na Secção 3.3 introduz-se a identificação de fronteiras difusas de *clusters*, abordando também a sua visualização. A Secção 3.4 apresenta e descreve a arquitectura do sistema implementado.

3.1 O Algoritmo AP-FCM

Uma das contribuições desta dissertação é a introdução de um método de inicialização alternativo para o *Fuzzy c-means*. A nova inicialização foi proposta por Mirkin [12] e aplicada ao algoritmo *k-means*, como referido na Sub-secção 2.2.1. Baseada num método, também introduzido por Mirkin, denominado *Anomalous Pattern*, a inicialização pretende seleccionar iterativamente, para protótipos iniciais, pontos que definam grupos o mais possível afastados da média do conjunto de dados. Os passos do método *Anomalous Pattern* estão descritos na Tabela 3.1.

Como é de fácil percepção, os resultados obtidos pela aplicação iterativa do *Anomalous Pattern* é afectada pela condição de paragem utilizada. Dado o príncipio base de extrair os *clusters* mais desviantes do conjunto de dados, a esse passo iterativo foi dado o nome de "Divisão & Conquista", para o qual Mirkin define as seguintes condições de paragem:

1. AP-C1: Todas as entidades serem associadas a um dos protótipos resultantes de uma iteração do *Anomalous Pattern*.

2. AP-C2: A contribuição dos primeiros *c clusters* encontrados para a dispersão de dados ser superior a um determinado valor limiar.

3. AP-C3: A contribuição do último *cluster* encontrado para a dispersão de dados ser inferior a um determinado valor.

4. AP-C4: O número de grupos extraídos tenha atingido um valor pré-determinado.

Note-se que a última condição de paragem indica directamente, tal como no FCM, com o parâmetro de entrada, o número de *clusters*, enquanto as restantes condições apenas afectam esse valor de forma indirecta, dependendo do conjunto de dados.

O cálculo da dispersão de dados de cada *cluster* será baseado na definição apresentada em [12]. Para um conjunto de dados normalizado de dimensão $n \times p$, $Y = [y_{ij}]$, com n entidades

1. Seleccionar como ponto de referência, c_0, a média de todo o conjunto de dados.
2. Escolher como protótipo tentativo, c, a entidade mais afastada do ponto de referência.
3. Aplicar a regra do protótipo mais próximo (Equação (2.5)) a todo o conjunto de dados, entre c e c_0.
4. Calcular novo protótipo tentativo, c', como a média de todos os pontos associados a c no passo anterior. Se c' e c forem diferentes, colocar $c = c'$ e voltar ao passo anterior. Se forem iguais, continua para o passo seguinte.
5. Retorna o protótipo tentativo c como o centróide relativo ao *cluster* mais desviante do conjunto de dados e as entidades a si associadas.

Tabela 3.1 Passos do algoritmo *'Anomalous Pattern'*.

e p atributos, e uma partição S desse conjunto, $S = S_1 \cup S_2 \cup ... \cup S_c$, Mirkin propõe que a dispersão total do conjunto, $T(Y) = \sum_{i=1}^{n} \sum_{j=1}^{d} y_{ij}^2$, seja composta por dois termos:

$$T(Y) = B(S,v) + W(S,v). \tag{3.1}$$

$B(S,v)$ é a componente responsável pela dispersão de dados relativamente aos protótipos dos *clusters*:

$$B(S,v) = \sum_{k=1}^{c} \sum_{j=1}^{p} n_k v_{kj}^2, \tag{3.2}$$

onde n_k é a cardinalidade do *cluster* c_k. $W(S,v)$ refere-se à soma das dispersões de cada protótipo às entidades a si associadas. Este valor corresponde aos erros quadráticos intra-*cluster* (ver Equação (2.9)). Para uma partição de *crisp clustering*, a contribuição relativa de um *cluster* para a dispersão de dados é dada por:

$$W(k) = \frac{B_{k+}}{T(Y)} = \frac{\sum_{j=1}^{p} n_k v_{kj}^2}{\sum_{l=1}^{n} \sum_{j=1}^{p} y_{lj}^2}, \tag{3.3}$$

onde B_{k+} representa a dispersão total de um *cluster* k. A normalização aplicada para obter o conjunto Y, a partir de um conjunto original X, é feita colocando a origem do referencial

no ponto médio de X e transformando a escala de modo a limitar os valores dos atributos ao intervalo $[-1,1]$.

O método de inicialização do *k-means* proposto em [12] corresponde à aplicação iterativa do algoritmo *'Anomalous Pattern'*, utilizando os protótipos resultantes como protótipos iniciais para a execução do algoritmo *k-means*.

No trabalho incluído na dissertação, utilizar-se-à este mesmo método de inicialização para a inicialização do FCM. Assim, comparativamente à aplicação do FCM original, elimina-se a necessidade de escolha do número de *clusters*, utilizando para protótipos iniciais os resultantes da aplicação do *Anomalous Pattern* (ver Tabela 3.2). Ao algoritmo FCM com uma inicialização baseada neste método foi dado o nome de '*Anomalous Pattern - Fuzzy c-means*' (AP-FCM). Para a aplicação do AP-FCM com uma condição de paragem AP-Cx, o algoritmo toma a designação de AP$_{Cx}$-FCM.

1. Inicializar t com o valor 1 e I_t como todo o conjunto de dados. Seleccionar condição de paragem.
2. Aplicar o '*Anomalous Pattern*' a I_t e receber um novo protótipo, c_t, e sub-conjunto de entidades a si associadas, S_t.
3. Testar condição de paragem. Se não se verificar, recalcular $I_{t+1} = I_t - S_t$ e $t = t+1$, voltar ao passo 2. Caso contrário, prossegue para o passo 4.
4. Remoção de *clusters* cuja cardinalidade seja inferior a um valor pré-determinado.
5. Executar o *FCM* com protótipos iniciais $c_1,...,c_t$.

Tabela 3.2 Passos da aplicação do algoritmo *Anomalous Pattern* como método de inicialização do FCM (AP-FCM).

Note-se que a aplicação do passo 4 serve para eliminar grupos compostos por entidades demasiadamente afastadas para serem relevantes para o problema, sendo passíveis de se referirem a ruído nos dados. Pode, no entanto, haver problemas em que seja interessante, ou mesmo necessária, a sua análise. O valor pré-determinado pode ser introduzido como parâmetro de entrada pelo utilizador.

No trabalho experimental a efectuar nesta dissertação, será feito um estudo para analisar as diferentes condições de paragem e ajustar os seus parâmetros para se obter uma segmentação efectiva do problema a tratar. Destaque-se que se pretende que as segmentações obtidas possibilitem a definição, de um modo o mais exacto possível, da região de upwelling. No Anexo A

é apresentado um estudo comparativo entre o FCM e o AP-FCM, com três conjuntos de dados de referência, onde se verificam as vantagens computacionais do AP-FCM e se analisa o comportamento das suas condições de paragem para a detecção de um bom número de grupos.

3.2 Definição de *features* e critério para identificação e anotação de regiões de upwelling

A problemática da identificação de regiões de upwelling está definida em dois grandes módulos. Inicialmente, a partir de um mapa de temperatura, pretende-se obter uma boa segmentação, que possibilite a identificação da região pretendida. A aplicação do algoritmo AP-FCM, introduzido na secção anterior, tem como objectivo a obtenção de uma segmentação sem recorrer à necessidade de introduzir o número de *clusters* como parâmetro de entrada. Posteriormente, é necessário encontrar o sub-conjunto de *clusters* que definem a região de upwelling. Este sub-conjunto representa a região de interesse do estudo elaborado, sendo que o objectivo passa por separar essa região do *background*, ou seja, separar os *clusters* correspondentes à região de upwelling dos restantes. O *cluster* de temperatura média superior ainda pertencente à região de upwelling é considerado o *cluster* de interesse, já que é o *cluster* que separa a região de interesse do *background*.

Nesta secção introduzem-se os padrões das regiões de upwelling em mapas de temperatura e define-se um conjunto de *features* que, a partir de uma segmentação previamente obtida, possibilitam a criação de um critério composto para identificação do *cluster* de interesse e, consequentemente, da região de upwelling.

3.2.1 Caracterização de padrões de upwelling em mapas SST

Em situações ideais, a região de upwelling pode ser caracterizada como uma região que cresce perpendicularmente à Costa Oeste da Península Ibérica, e que é separada das restantes águas, não pertencentes a essa região, por uma zona de transição com um gradiente térmico elevado, ou seja, variação relativamente brusca da temperatura entre as duas regiões. No entanto, nem sempre ocorrem essas situações ideiais, havendo situações onde há uma existência excessiva de ruído nos mapas de temperatura ou a diferença entre a região de upwelling e o oceano ao largo não é tão evidente. O grupo de mapas SST disponibilizado pelo Instituto de Oceanografia da Universidade de Lisboa foi produzido com o objectivo de cobrir as várias situações em que o upwelling ocorre. Podemos caracterizar três situações distintas: mapa SST sem ruído e com região de upwelling bem definida em termos de gradientes térmicos (Figura 3.1), mapa SST sem ruído e região de upwelling mal definida (Figura 3.2) e mapa SST com ruído, devido fundamentalmente à presença excessiva de regiões com nuvens (Figura 3.3).

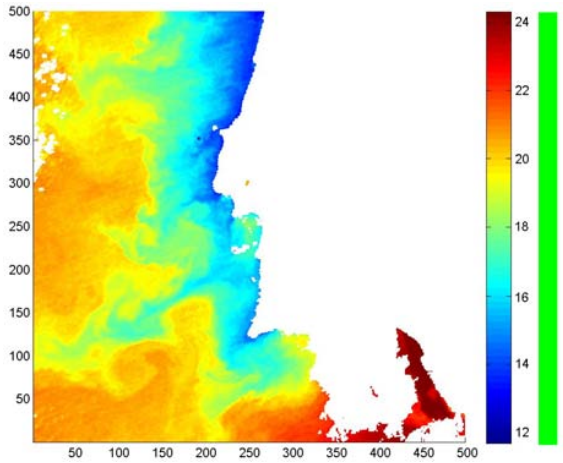


Figura 3.1 Mapa de temperaturas SST, (2 de Agosto de 1998), com upwelling bem definido, com anotação visual da região de upwelling.

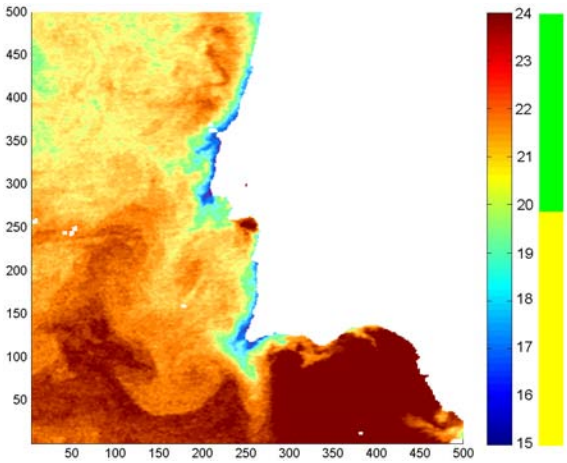


Figura 3.2 Mapa de temperaturas SST (12 de Junho de 1998), com upwelling mal definido, com anotação visual da região de upwelling.

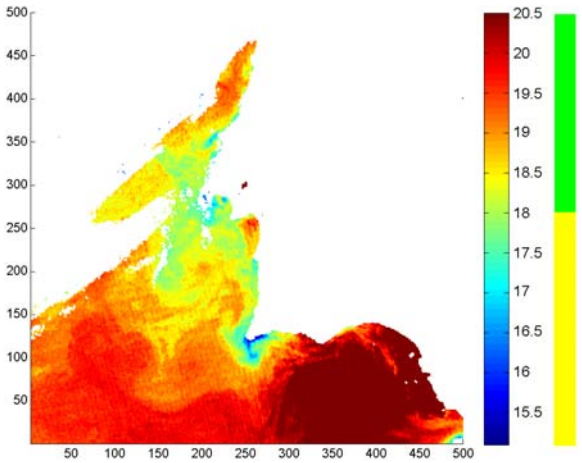


Figura 3.3 Mapa de temperaturas SST (09 de Junho de 1998), com upwelling mal definido e presença excessiva de ruído sob a forma de nuvens, com anotação visual da região de upwelling.

3.2.2 Definição de fronteiras *crisp* de *clusters*

Para permitir a análise das segmentações obtidas foi desenvolvido um método para anotar automaticamente as fronteiras obtidas. No trabalho desenvolvido em [18], essa análise foi feita somente com base na observação dos mapas de pertença (como o da Figura 3.4(a)), no entanto nesse resultado acaba-se por perder informação relativa aos valores de temperatura dos píxeis, sendo esse elemento um factor de cariz fundamental para a análise do upwelling.

Independentemente do algoritmo de segmentação utilizado, após a sua aplicação, os *clusters* foram ordenados por temperaturas médias, ficando o *cluster* com temperatura média mais fria identificado com a *label* 1 e o *cluster* com a temperatura média mais elevada identificado com a *label* c, sendo c o número total de *clusters*. Ou seja, independentemente da inicialização dos protótipos, seja esta feita aleatoriamente (FCM) ou com um método alternativo (AP-FCM), o resultado final tem sempre os *clusters* ordenados por temperatura média. Esta ordenação, para além de regularizar as *labels* dos *clusters* para todas as imagens, ajuda ao próprio reconhecimento do upwelling, já que sabendo que o fenómeno é obrigatoriamente representado pelos *clusters* com temperatura média inferior, pode-se definir o upwelling como o sub-conjunto dos primeiros k *clusters*. Na Sub-secção 3.2.6 apresenta-se o critério com *thresholds* para encontrar esse valor de k. A Figura 3.4(a) exemplifica uma segmentação com identificação ordenada por temperatura média. Nos mapas de temperatura SST em que o upwelling está presente, os *clusters* mais frios encontram-se sempre mais próximos da linha costeira.

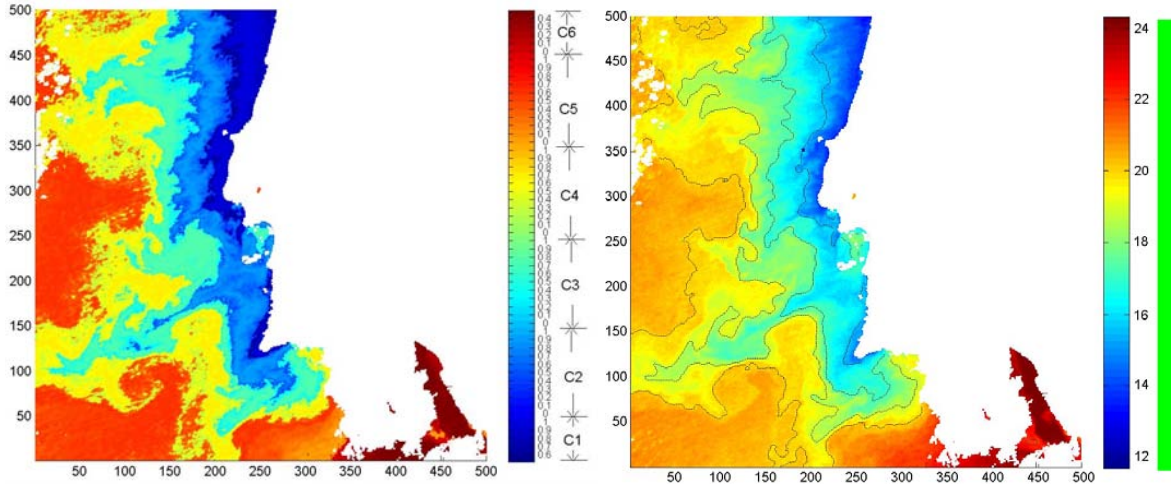
Define-se como fronteira de um *cluster* k, o conjunto de píxeis a si pertencentes, em que algum dos seus vizinhos pertença a um *cluster* que não seja k. Os vizinhos de um píxel (i, j) são definidos como uma "4-vizinhança", ou seja, o píxel superior (i+1,j), píxel inferior (i-1,j), píxel à esquerda (i,j-1) e píxel à direita (i,j+1). Contudo, para não coexistirem duas fronteiras entre o mesmo par de *clusters* k e $k+1$ (i.e., a fronteira de C_k e a fronteira de C_{k+1}), a fronteira de cada *cluster* k é apenas definida pela sua fronteira exterior, ou seja, os píxeis pertencentes ao *cluster* C_k cujos vizinhos pertençam a um *cluster* identificado com uma *label* superior. Assim, a visualização das fronteiras nos mapas SST é feita com a marcação dos píxeis relativos à fronteira exterior de cada *cluster*.

Note-se que esta é uma definição de fronteiras *crisp*, feitas com base na classificação de píxeis após o passo de *desfuzzificação*, ignorando assim a natureza difusa, tanto dos algoritmos de agrupamento difuso como do próprio upwelling. Na Secção 3.3 são apresentadas medidas de *fuzziness* para estudar a identificação de fronteiras difusas.

Na Figura 3.4(b) visualizam-se as fronteiras obtidas a partir do mapa de segmentação visualizado na Figura 3.4(a). Uma boa análise aos resultados de segmentação deve ter em conta ambos os resultados, já que a visualização de fronteiras sobre o mapa de temperaturas permite ter a noção de variações de temperaturas entre *clusters*, estando no entanto sempre dependende da escala de cores aplicada. A visualização do mapa de pertenças apresenta os *clusters* com a sua estrutura natural, permitindo a análise das estruturas da regiões de upwelling e restantes águas, sem depender da escala de cores.

Refira-se que, para efeitos de visualização, aos resultados de segmentação é aplicado um

passo de pós-processamento que elimina regiões não contíguas de *clusters*. Um dos objectivos da segmentação de imagem é a geração de regiões contíguas de píxeis e mesmo no domínio do problema em causa, pequenas regiões de píxeis pertencentes a um *cluster* que estão envolvidas em regiões de píxeis de um outro *cluster* são consideradas ruído. Neste passo de pós-processamento também é melhorada a definição das fronteiras *crisp* pela aplicação de um passo de *smooth*.



(a) (b)

Figura 3.4 Visualização, em (b), das fronteiras crisp dos *clusters* obtidos na segmentação visualizada na figura (a), com identificação de *clusters* ordenada por temperatura média. Temperaturas médias de cada *cluster*: $T_1 = 16°C, T_2 = 18°C, T_3 = 18.6°C, T_4 = 19.2°C, T_5 = 19.6°C, T_6 = 20.7°C$.

3.2.3 Definição de diferença relativa de temperatura entre *clusters*

Com o objectivo de resolver o problema da definição de uma frente da região de upwelling, foi feita uma análise às temperaturas médias dos *clusters* obtidos nas segmentações. A base dessa análise teve como princípio as indicações dadas por oceanógrafos relativamente à própria natureza do fenómeno, indicando que a frente das regiões de upwelling poderia ser definida como a região onde se verifica uma maior, ou mais acentuada, diferença de temperatura. Assim, definiu-se uma nova *feature*, com base nas diferenças de temperaturas entre pares adjacentes de *clusters*, mas com a introdução do factor n_k, a cardinalidade do *cluster* k, que permite estabelecer maiores valores para a feature $TDiff$ quando o *cluster* k tem menor cardinalidade, ou seja, é de extensão mais reduzida.

A introdução do factor n_k vai de encontro à definição de diferença acentuada de temperatura, que se encontra na frente da região de upwelling, já que penaliza os *clusters* que possuem uma

cardinalidade muito elevada e que, por essa razão, ocupam uma área muito vasta, geograficamente falando. Assim, a feature $TDiff$, definida para cada *cluster* k, é calculada pela seguinte fórmula:

$$TDiff(k) = \frac{T_{k+1} - T_k}{n_k}, (1 \leq k \leq c-1). \tag{3.4}$$

3.2.4 Detecção de extensão cumulativa de *clusters*

De acordo com as anotações fornecidas por oceanógrafos, um dos factores que afectam a definição das regiões de upwelling prende-se com a região geográfica que esta ocupa, ou seja, verifica-se que a extensão da região anotada nunca ultrapassa um certo limite. Esta situação está ligada à relação que o fenómeno tem com a morfologia do fundo do oceano. A Figura 3.5 contém um mapa de batimetria (profundidade do oceano) e, por análise dos mapas SST e informação fornecida por oceanógrafos, verifica-se que os limites da região de upwelling estão relacionados com a profundidade existente, não atingindo, normalmente, regiões com mais de 4000 metros de profundidade. Assim, sabemos que as águas frias, relativas a uma região de upwelling, começam por surgir junto à costa e se vão expandido costa fora. No entanto, essa expansão não acontece infinitamente, ficando a região de upwelling limitada a uma determinada porção dos mapas de temperatura disponíveis. Com base nessa análise, outra das *features* definidas para complemento à detecção das frentes de upwelling, estuda a extensão dos primeiros k *clusters*, em termos de percentagem de píxeis por eles ocupados.

A análise a esta *feature* permite-nos afirmar, com um elevado grau de certeza, que certos *clusters* não podem pertencer à região de upwelling. Se, por exemplo, os primeiros 3 *clusters* de uma segmentação ocuparem 95% dos píxeis disponíveis, ou seja, todos aqueles píxeis que não contém o valor NaN, sabemos que, pelo menos, o *cluster* 3 não faz parte da região de upwelling, já que, nos mapas em análise, é impossível que a região pretendida ocupe tal percentagem de píxeis. Para um *cluster* k, a sua cardinalidade relativa cumulativa é definida pela fórmula:

$$CCard(k) = \frac{\sum_{j=1}^{k} n_j}{n} \tag{3.5}$$

3.2.5 Detecção de ruído excessivo causado por extensões nebulosas

Outra questão que pode afectar a definição da região de upwelling é a presença excessiva de extensões nebulosas. Segundo oceanógrafos, estas regiões, que acabam por ser ruído nos mapas SST, são problemáticas já que a sua existência pode afectar os valores dos píxeis seus vizinhos, fazendo com que os valores de SST lidos não sejam fiáveis. Consequentemente, a definição da região de upwelling pode ser afectada em regiões com grandes extensões nebulosas, facto que se pode confirmar pelas anotações feitas por oceanógrafos, para identificação das regiões de upwelling dos anos de 1998 e 1999.

Para fazer um estudo da presença de extensões nebulosas, para cada píxel com valor NaN,

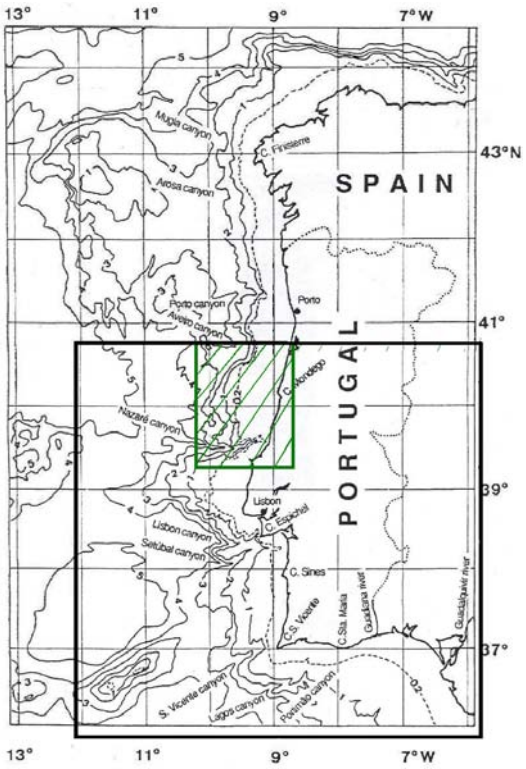


Figura 3.5 Mapa de batimetria ao largo da Península Ibérica. O quadrado marcado com cor preta corresponde área geográfica de cada mapa de temperaturas e, a verde, está marcada a região onde é testada a vizinhança de extensões nebulosas aos *clusters*.

definiu-se uma janela de dimensão 8×8 píxeis à sua volta como sendo a sua vizinhança e, para cada *cluster* em análise, contam-se quantos píxeis a si associados estão na vizinhança de algum píxel *NaN*, com excepção dos píxeis relativos à massa terrestre continental (presentes em todas as imagens). Por análise empírica, verificou-se que nas marcações feitas por oceanógrafos os casos onde a indicação é alterada pela presença de nuvens acontece numa região restrita, indicada a verde na Figura 3.5. Assim, é nessa região que se testa a intersecção dos *clusters* com nuvens. Note-se que situações em que as extensões nebulosas estejam muito afastadas da costa, e por isso excluídas da região onde se faz a análise, já são tratadas na análise da feature relativa à extensão dos *clusters*. Sendo *NaNNeighbour* a região que incorpora todos os píxeis na vizinhança 8×8 de um píxel de valor *NaN* e $|x|$ a cardinalidade de um conjunto de píxeis x, tem-se como cálculo do ruído causado por extensões nebulosas para um *cluster* k, a seguinte medida:

$$CloudNoise(k) = |\{x : x \in c_k \cup x \in NaNNeighbour\}|. \tag{3.6}$$

3.2.6 Definição de critério de decisão para anotação da fronteira de upwelling

A definição do critério para identificação do cluster de interesse, que define a região de upwelling, teve como princípio a obtenção de fronteiras o mais semelhantes possíveis às anotações, feitas por oceanógrafos, pretendendo-se que este modelasse o critério que os próprios especialistas usam para definir essas regiões. Uma das primeiras conclusões retiradas a partir da análise das anotações de regiões de upwelling feitas por oceanógrafos, foi que em várias anotações a fronteira de upwelling indicada variava em termos de gama de temperaturas, ao longo da imagem. Por exemplo, a anotação feita para o mapa de temperaturas visualizado nas Figuras 3.2 e 3.3, a anotação indica que a Norte do Cabo Espichel a fronteira de upwelling é indicada pelos píxeis de cor verde, e a Sul a fronteira já é definida pelos píxeis de cor amarela. Assim, numa única imagem, os píxeis com a temperatura representada pela cor amarela, estão incluídos na região de upwelling a Sul do Cabo Espichel, mas a Norte essa situação já não se verifica. Sendo óbvio que as segmentações obtidas pelos algoritmos em estudo não conseguem tratar situações destas e obter bons resultados, ou seja, píxeis de temperaturas iguais ficarem classificados em *clusters* diferentes, estabeleceu-se que a análise para o critério de definição da fronteira de upwelling seria feita separadamente na Região Norte e na Região Sul, sendo essa separação feita na latitude onde se encontra o Cabo Espichel (38.4º N), já que se trata do ponto onde mais frequentemente há diferenciação nas anotações disponibilizadas. Esta posição corresponde à linha 260 nas matrizes dos mapas de temperatura SST disponíveis. Assim, para a análise à Região Norte, no mapa de pertenças *desfuzzificado* resultante de uma segmentação, todos os píxeis de latitude inferior ao Cabo Espichel são colocados a *NaN*, acontecendo a situação inversa para a análise à Região Sul. Por esta razão, em alguns mapas, onde o critério definido indica *clusters* diferentes como limites da região de upwelling, na visualização de frente da região, a fronteira tem uma mudança brusca na latitude correspondente ao Cabo Espichel. Destaque-se que esta divisão dos mapas de temperatura é algo natural para os oceanógrafos, já que houve indicações de que as duas regiões têm características relativamente diferentes, nomeadamente em termos de temperaturas médias e salinidade.

Por análise dos resultados de segmentação da aplicação dos algoritmos FCM e AP-FCM, e comparação com as marcações das regiões de upwelling, conseguiu-se reduzir o espaço de análise nas segmentações obtidas. Para todas as segmentações obtidas com a aplicação do algoritmo AP_{C3}-FCM, com a análise da dispersão de dados como condição de paragem, verificou-se que a fronteira de upwelling, de acordo com as anotações, correspondia sempre à fronteira exterior de um dos primeiros 3 *clusters*. Assim, a análise de *features* foi feita unicamente nesses *clusters*, considerados críticos e que definem uma região de interesse para este problema.

O critério que foi criado para identificar o cluster de interesse, para as duas regiões analisadas distintamente, é composto pela análise das várias *features* definidas. Esta situação é natural já que há várias situações que um oceanógrafo analisa quando define uma região de upwelling, pelo que também há lugar à análise dessas mesmas situações quando se implementa um método automático para essa definição. A composição do critério pode ser definida como uma sequência de passos, que tem como resultado uma segmentação binária, definindo a região

de upwelling e as restantes águas presentes no mapa. Os principais passos são os seguintes: (i) separação do mapa de segmentação em Região Norte e Região Sul; (ii) análise à feature $TDiff$; (iii) análise à extensão dos *clusters*; (iv) análise de existência de extensões nebulosas; e (v) construção do mapa de segmentação binário, com base nos resultados obtidos após (ii), (iii) e (iv). A Figura 3.6 contém um esquema com a sequência dos passos que compõe o módulo de definição do cluster de transição da região de upwelling.

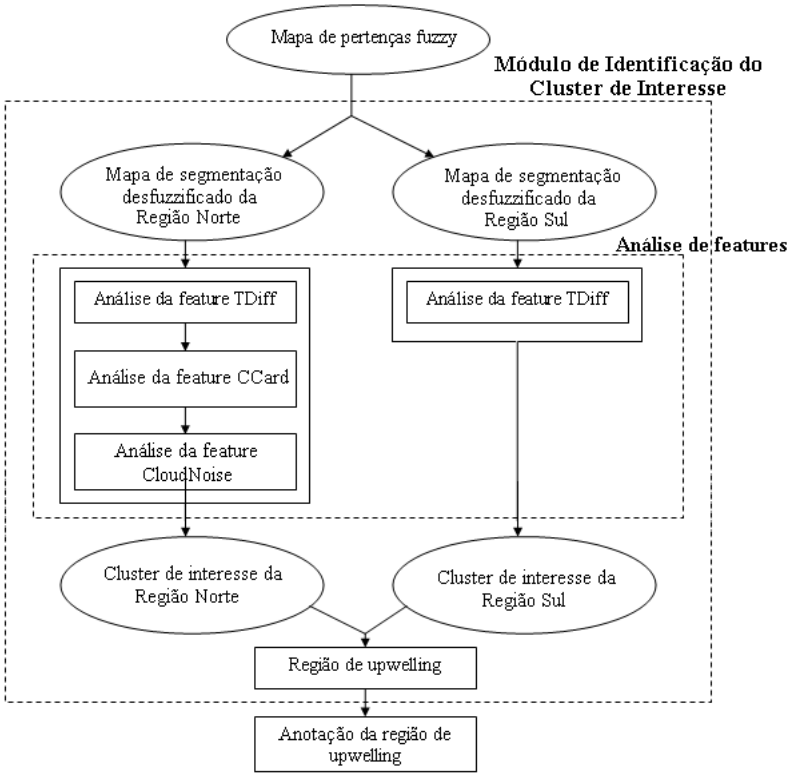


Figura 3.6 Esquema da composição do critério para identificação do cluster de interesse, para a Região Norte e Região Sul.

A análise à feature $TDiff$ (Equação 3.4) é a que fornece melhor informação relativamente à decisão de que sub-conjunto de *clusters* define o upwelling, sendo que as análises posteriores, efectuadas sobre a Região Norte, melhoram a qualidade dos seus resultados. Destaque-se que na Região Sul, não se efectuam as análises às *features* todas, uma vez que nessa região não ocorrem as situações que cada *feature* analisa. Assim, na origem do critério composto está a análise de $TDiff(3)$, que começa por identificar o *cluster* de interesse (*Border_Cluster*). O método de análise de $TDiff$ é composto pelos passos 2-6, para a Região Norte, e 19-23, para Região Sul, do critério composto apresentado no Algoritmo 1. Sendo τ_T um valor limiar experimentalmente estabelecido, a análise de $TDiff(3)$ resulta na indicação de um *cluster*,

Border_Cluster, que representa o limiar da região de upwelling, ou seja, define essa região como a união dos primeiros *Border_Cluster clusters*. Havendo duas regiões distintas onde a *feature* é analisada, τ_T tem a designação de τ_{TN} para a Região Norte e de τ_{TS} para o Sul. Note-se que nesta análise nunca se indica que apenas o primeiro *cluster* faz parte da região de upwelling, uma vez que nos resultados obtidos, essa situação só acontece devido a situações estudadas em posteriores análises de *features*: extensão de *clusters* e extensões nebulosas.

De seguida, faz-se a análise da *feature CCard* (Equação 3.5), que analisa os *clusters* em termos de percentagens de píxeis ocupados. Nesta fase, cuja aplicação é feita somente na Região Norte, apenas se tem em conta os primeiros *Border_Cluster clusters*, que resultaram do passo prévio, pelo que, o objectivo passa por diminuir o número de *clusters* que definem a região de upwelling, enquanto a percentagem de píxeis por si ocupado for excessiva. Sendo $CCard(k)$ a cardinalidade relativa cumulativa do *cluster* k e τ_C um valor limiar, a partir do qual se considera como excessiva a percentagem de píxeis ocupados, a análise é feita nos passos 7-9 do Algoritmo 1.

A última análise feita estuda a perturbação na definição da fronteira de upwelling causada pela presença de extensões nebulosas, através da *feature CloudNoise* (Equação 3.6). Tal como no passo anterior, esta análise "recebe" um valor de *Border_Cluster*, como indicação dos *clusters* que definem a região de upwelling e é sobre esse resultado que estuda a existência de extensões nebulosas que possam condicionar a qualidade da região obtida. Assim, define-se um *threshold* τ_N como um valor limiar a partir do qual se considera que um *cluster* contém demasiado ruído na sua vizinhança para se poder afirmar que pertence à região de upwelling. Note-se que a análise não é feita no primeiro *cluster*, uma vez que nos 4 mapas de temperatura onde não há anotação feita para a totalidade da Região Norte, os resultados não são possíveis de melhorar. O resultado final para *Border_Cluster* é calculado nos passos 10-12 do Algoritmo 1.

Concluída a análise das *features* que compõem o critério para identificação do cluster de transição da região de upwelling, o resultado final são dois valores de *Border_Cluster*: um para a Região Norte (*Border_Cluster_North*) e outro para a Região Sul (*Border_Cluster_South*). Relembre-se que a separação Norte/Sul é feita na linha correspondente à latitude do Cabo Espichel, pelo que na visualização do resultado final, acima dessa linha tem-se a fronteira exterior do *Border_Cluster_North* e nas linhas inferiores a fronteira exterior do *Border_Cluster_South*. O critério composto, agrupando a análise de todas as *feature* é definido pelo Algoritmo 1.

Na aplicação do critério composto para identificação automática do cluster de interesse, os resultados obtidos são dependentes dos *thresholds* aplicados em cada passo. Inicialmente, os valores utilizados foram encontrados com base na observação empírica de resultados e na sua comparação com indicações das regiões de upwelling feitas por oceanógrafos. Contudo, para contornar a observação estritamente empírica, foi desenvolvido um método de obtenção de *thresholds* com base no estudo do ganho de informação, apresentado em [48].

Em Teoria da Informação, o ganho de informação de um qualquer atributo a, respectivamente a uma classe V, indica a redução de incerteza sobre o valor de V quando se sabe o valor de a, ou seja, quão útil é o atributo a para descobrir V. Aplicando ao cálculo automático de

Algoritmo 1 Composição do critério para identificação do *cluster* de interesse

{Análise Região Norte, a partir do mapa de pertenças *desfuzzificado* da Região Norte.}
if $(TDiff(3) >= \tau_{TN})$ **then**
    $Border\_Cluster \leftarrow 3$;
else
    $Border\_Cluster \leftarrow 2$;
end if
while $((CCard(Border\_Cluster) > \tau_C)\ \&\&\ (Border\_Cluster > 0))$ **do**
    $Border\_Cluster \leftarrow Border\_Cluster - 1$;
end while
while $((CloudNoise(Border\_Cluster) > \tau_N)\ \&\&\ (Border\_Cluster > 0))$ **do**
    $Border\_Cluster \leftarrow Border\_Cluster - 1$;
end while
if $(Border\_Cluster == 0)$ **then**
    $Border\_Cluster \leftarrow 1$;
end if
$Border\_Cluster\_North \leftarrow Border\_Cluster$;

{Análise Região Sul, a partir do mapa de pertenças *desfuzzificado* da Região Sul.}
if $(TDiff(3) >= \tau_{TS})$ **then**
    $Border\_Cluster \leftarrow 3$;
else
    $Border\_Cluster \leftarrow 2$;
end if
$Border\_Cluster\_South \leftarrow Border\_Cluster$;

thresholds, podemos definir duas classes V, sendo que uma classe indica o atributo como pertencente à região de upwelling e a outra como não pertencente, e ter como atributos os valores das *features* que são estudadas. Por exemplo, para a análise à *feature* $TDiff(3)$, os atributos possuem o valor da *feature* e a classe V, para cada imagem, indica se o terceiro *cluster* pertence, ou não, à região de upwelling. Para cada mapa de temperatura, a atribuição do valor para V foi feito por comparação com os mapas "ground-truth", ou seja, se o *cluster* em análise se encontrar incluído na região de upwelling anotada por oceanógrafos, então V tem o valor de 1, indicando essa pertença. Caso contrário, é atribuído o valor de 0 a V, indicando que o *cluster* não pertence à região de upwelling.

Assim, tendo um conjunto S composto por tuplos (atributo, classe), onde S_i é o conjunto dos atributos de S pertencentes à classe $V = i$. Para o problema corrente, com duas classes que definem a pertença, ou não, à região de upwelling, a informação esperada para calcular o valor de uma amostra é dada pela fórmula:

$$I(S_1,S_2) = -\sum_{i=1}^{2} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}. \tag{3.7}$$

Cada atributo, a_j, de S pode ser usado para criar uma partição em S, com conjuntos (S_{i1}, S_{i2}), onde S_{i1} contém as amostras de S, pertencentes à classe $V = i$, que são inferiores a a_j e S_{i2} as que são superiores ou iguais. Esta técnica de segmentação do conjunto de atributos é conhecida por *Entropy Based Discretization* [48], onde as partições geradas permitem estudar a aplicação de um atributo como valor limiar para definir a classe V. A informação esperada com base nas partições criadas por a_j é conhecida como a sua entropia:

$$E(a_j) = -\sum_{j=1}^{2} \frac{|S_{1j}| + |S_{2j}|}{|S|} I(S_{1j},S_{2j}). \tag{3.8}$$

Pretende-se utilizar como threshold o valor de a_j que melhor discrimina a pertença à região de upwelling, ou seja, o atributo cuja partição gerada obtém um maior ganho de informação. O cálculo do ganho de informação é definido por:

$$Gain(a_j) = I(S_1,S_2) - E(a_j). \tag{3.9}$$

Refira-se que a utilização deste método para cálculo automático de *thresholds* tem como vantagem o facto de poder ser aplicado outros conjuntos de mapas de temperatura. Ou seja, em vez de se utilizarem os *thresholds* definidos empiricamente a partir da análise exclusiva dos mapas de 1998 a qualquer conjunto de mapas que estejam disponíveis, a análise do ganho de informação pode ser aplicada a novos conjuntos de mapas para obtenção automática dos *thresholds*.

3.3 Identificação e visualização de fronteiras difusas de upwelling

3.3.1 Definição de medidas de caracterização de fronteiras difusas

Como explicado anteriormente, a definição de uma fronteira de upwelling não pode ser considerada como uma técnica exacta. A mistura de águas com temperaturas distintas leva a uma situação onde cada fronteira deverá ser melhor caracterizada como uma região, ou conjunto de pixeis, que se situam no limiar de, pelo menos, duas massas de água distintas e, com o passo *desfuzzificador* aplicado às técnicas de *fuzzy clustering*, a fronteira que se obtém vai contra esse mesmo princípio. Como tal, desenvolveu-se um estudo exploratório, com recurso a medidas de *fuzziness*, para permitir a definição de cada fronteira como um objecto de fronteira difusa.

Foram implementadas duas medidas de *fuzziness* e cada medida é aplicada ao nível do píxel, para todos os píxeis das imagens, utilizando os resultados da execução dos algoritmos difusos, nomeadamente a matriz de pertenças U. O resultado final da aplicação de uma destas medidas a uma segmentação difusa é um mapa onde cada píxel contém o seu valor de difusividade,

conforme a medida utilizada. Assim, para cada píxel (entidade) i de uma imagem, as medidas são definidas por:

- *Ignorance Uncertainty* [49] - Medida que calcula a incerteza associada ao desprezar dos valores de pertença, com excepção do valor máximo, aquando o passo *desfuzzificador*. Por exemplo, duas entidades com valores de pertença $u_1 = [0.1, 0.9]$ e $u_2 = [0.45, 0.55]$ ficam associadas ao mesmo *cluster*, com valores de pertença máximos bem distintos. Calculada através de uma medida de entropia, que estuda a concentração dos valores de pertença num *cluster* ou a sua dispersão por vários *clusters*, a medida tem a seguinte fórmula:
$$IU(i) = -\frac{1}{log_e\ c} \sum_{k=1}^{c} u_{ik}\ log_e(u_{ik}). \tag{3.10}$$

- *Exageration Uncertainty* [49] - Medida que calcula a incerteza causada pela associação de uma entidade a um *cluster*, mesmo sendo possível que o valor de pertença seja relativamente reduzido. Com os valores de pertença exemplificativos da medida anterior, a entidade 2 fica associada ao segundo *cluster* apesar de possuir um valor de pertença de "apenas" 0.55. A medida é calculada medindo as diferenças entre o valor unitário (pertença máxima possível) e o valor de pertença de cada entidade ao *cluster* a que é associado, após o passo *desfuzzificador*:
$$EU(i) = 1 - max(u_{ik}), \forall\ k = 1, .., c \tag{3.11}$$

As duas medidas variam no intervalo entre 0, quando o grau de difusividade é nulo, ou seja, o valor de pertença a um dos *clusters* é máximo (1) e 0 para os restantes, e 1, quando as pertenças são muito difusas.

3.3.2 Visualização de fronteiras difusas

A visualização de fronteiras difusas é um elemento de relevância para o estudo do upwelling por parte dos oceanógrafos, pelos motivos anteriormente explicados. Para a visualização efectiva é necessário recorrer à definição de um α-cut. Dado um conjunto de dados X e uma matriz de pertenças U, associando n entidades a c *clusters*, um α-cut define o conjunto de entidades com um valor de pertença a um *cluster*, igual ou superior a α, ou seja:

$$\alpha - cut(U, X, \alpha) = x \in X : u_{xk} \geq \alpha, \forall\ k = 1, .., c. \tag{3.12}$$

Esta definição pode ser transposta de modo a se poder obter α-cuts sobre imagens. Dada uma medida M, definida para todos os pontos existentes de uma imagem I, um α-cut permite-nos obter um conjunto de píxeis de uma imagem cuja medida seja superior a um determinado α:

$$\alpha - cut(M, I, \alpha) = x \in I : M(x) \geq \alpha \tag{3.13}$$

A visualização é feita com recurso à aplicação de uma medida de *fuzziness* sobre os resultados de uma segmentação difusa, obtida a partir da execução de um algoritmo de *fuzzy clustering* sobre uma matriz SST. Essa aplicação resulta numa matriz com o mesmo formato que o mapa SST mas onde cada píxel, em vez de um valor de temperatura oceânica, tem o valor da medida de *fuzziness* aplicada. Assim, instanciando-se a medida M da Equação 3.13 com a matriz de *fuzziness*, aplica-se o α-cut sobre M e obtêm-se os píxeis do mapa de temperaturas original (I) que definem fronteiras difusas, ou seja, aqueles píxeis cujo valor de *fuzziness* seja superior ao α estabelecido. Os píxeis obtidos pela aplicação do α-cut são os píxeis que definem a fronteira difusa.

Como é esperado, os píxeis "no interior" de cada *cluster* têm um maior valor de pertença a esse *cluster* e menor *fuzziness*, pelo que as frontteiras difusas tendem a ser definidas pelo conjunto de píxeis junto às fronteiras de cada *cluster*. A parametrização do valor de α verifica-se de importância fundamental, uma vez que é esse valor que afecta o formato da fronteira.

A visualização efectiva é feita com a sobreposição sombreada dos píxeis que definem a fronteira difusa sobre a imagem SST original. Por exemplo, na Figura 3.7(a) representa-se um exemplo dum mapa de temperaturas, sobre uma matriz de píxeis. Na Figura 3.7(b), cada píxel contém um valor simulado de uma medida de *fuzziness* e a fronteira difusa que se pretende visualizar é definida pelos píxeis a sombreado, obtidos com a aplicação de um $\alpha - cut$, com $\alpha = 0.6$. A visualização da fronteira difusa sobre os valores de temperatura fica, exemplificativamente, como na Figura 3.7(c), ao contrário de uma fronteira *crisp*, que seria apenas representada por uma linha de píxeis (Figura 3.7(d)).

3.4 Arquitectura do sistema FuzzyUpwell

Como uma das contribuições desta dissertação, o trabalho desenvolvido resultou numa proposta de uma arquitectura de um sistema que engloba todos os métodos estudados. A Figura 3.8 contém um esquema com os vários módulos e secções que compuseram o estudo feito com os algoritmos de *fuzzy clustering*. A componente com a aplicação do algoritmo de *Histogram Thresholding* enquadra-se no mesmo sistema, com a exclusão das componentes que possuem pertenças difusas, ou seja, o Módulo de Segmentação não resulta num mapa de pertenças difusas e não pode ser aplicado o Módulo de Definição de Fronteiras Difusas.

Para solucionar o problema da detecção automática do upwelling, o primeiro módulo (Módulo de Segmentação) aplicado a cada uma dos mapas de temperatura tratou de obter boas segmentações, ou seja, segmentações que permitam definir com rigor a região de upwelling. Neste módulo são aplicados os algoritmos estudados (FCM, AP-FCM e Histogram Thresholding) aos mapas de temperatura e tem como resultado um mapa de pertenças, associando cada píxel a um cluster com um determinado valor de pertença. Como já referido, exceptuando na aplicação do algoritmo *Histogram Thresholding*, as pertenças são *fuzzy*. Estes mapas podem ser visualizados, como o resultado dos algoritmos de *clustering* ou permitir a marcação de fronteiras de *clusters* sobre o mapa SST original.

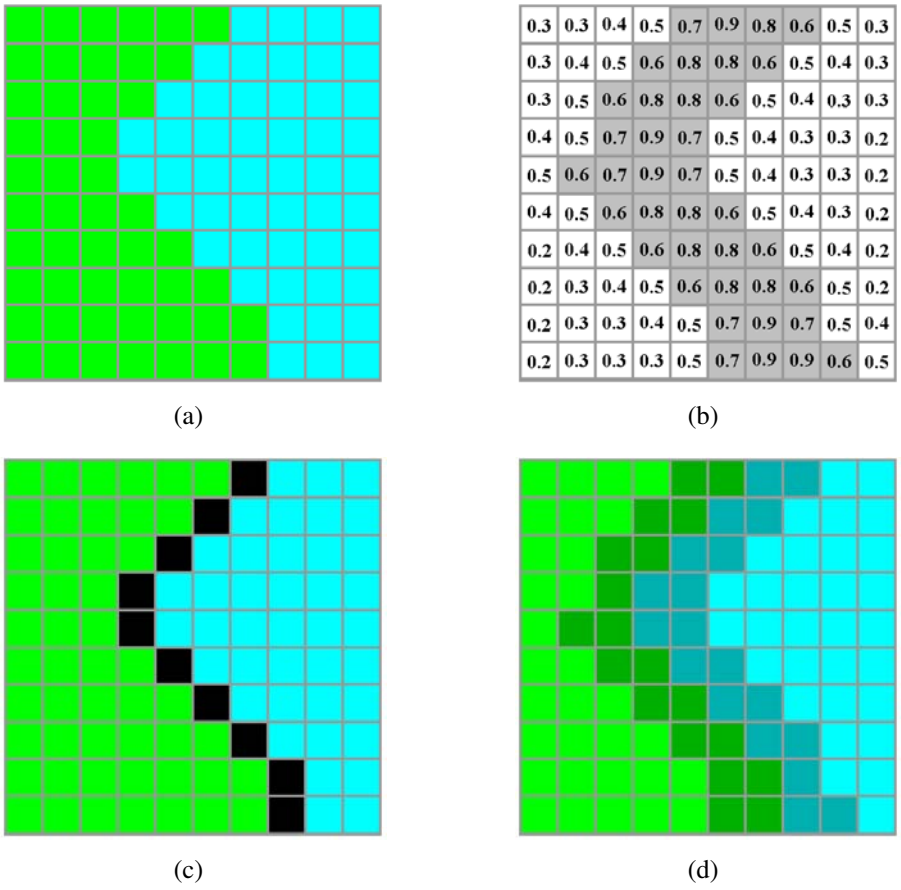


Figura 3.7 (a) Visualização exemplificativa de valores de temperatura num mapa de píxeis; (b) Representação de uma fronteira difusa num mapa de píxeis, definida com um $\alpha - cut$ de 0.6, sobre uma medida simulada; (c) Visualização de fronteira *crisp* sobre valores de temperatura da alínea (a); (d) Visualização de fronteira difusa da alínea (b) sobre valores de temperatura da alínea (a).

Figura 3.8 Arquitectura dos módulos desenvolvidos, para algoritmos de *fuzzy clustering*.

Posteriormente, com base nos resultados obtidos no primeiro módulo (mapa de pertenças), desenvolveu-se um método para definir a região de upwelling (Módulo de Identificação da Região de Upwelling, descrito em maior detalhe na Secção 3.2). Este módulo faz uma análise à segmentação que recebe como *input* e, com recurso a uma definição de *features*, identifica uma fronteira para a região de upwelling. O Módulo de Definição de Fronteiras Difusas (Secção 3.3) engloba um trabalho exploratório com o objectivo de identificar e visualizar fronteiras difusas, com base em medidas de *fuzziness*.

As anotações fornecidas por oceanógrafos tomaram um papel fundamental nos módulos de segmentação e de definição de fronteira da região de upwelling, já que foram a única informação utilizada para validar e melhorar cada passo e método testado no estudo feito, tendo também sido utilizadas nas etapas de avaliação de resultados obtidos. Os passos de avaliação são aplicados para avaliar qualitativamente as segmentações obtidas no Módulo de Segmentação e as regiões de upwelling identificadas no Módulo de Identificação da Região de Upwelling.

Refira-se que no âmbito do projecto no qual esta dissertação se enquadra, todos os módulos foram integrados num *interface*, no qual, dado um novo mapa SST, é possível aplicar automaticamente todos os passos necessários à anotação automática do upwelling.

4. Estudo Experimental

Neste capítulo são apresentados e analisados os resultados obtidos na elaboração desta dissertação. Após a definição dos objectivos do estudo (Secção 4.1) e dos conjuntos de dados utilizados (Secção 4.2), são apresentadas secções relativas à segmentação de mapas de temperatura com cada um dos algoritmos estudados: FCM, AP-FCM e *Histogram Thresholding*. Nas Secções 4.6 e 4.7 são comparadas as segmentações obtidas em termos de qualidade para uma boa identificação de regiões de upwelling e eficiência computacional dos algoritmos, respectivamente. Na Secção 4.8 apresenta-se um estudo com base no cálculo de gradientes máximos nas fronteiras dos *clusters* e na Secção 4.9 são analisados os resultados da aplicação do critério para reconhecimento automático da fronteira de upwelling.

Todos os resultados obtidos no trabalho desta dissertação foram executados com o *software* MATLAB [46], num computador portátil Acer, com processador AMD Turion 64 X2 (1.60 GHz) e 768 MB de RAM, com o sistema operativo Windows XP - Media Center Edition (Service Pack 2).

4.1 Objectivos

O primeiro passo do sistema para permitir a detecção de regiões de upwelling a partir de mapas de temperatura SST é obter uma segmentação desse mapa. O objectivo passa por tentar separar os píxeis que não contêm valores *NaN* em vários grupos "naturais" que representem significativamente as estruturas distintas de massas de água existentes da imagem. Com o intuito primário de detectar o upwelling, uma boa segmentação será aquela em que um sub-conjunto dos *clusters* encontrados representa fidedignamente a região do upwelling. Por exemplo, para o mapa de temperaturas representado na Figura 4.1 a região de upwelling estende-se até aos píxeis de cor verde, inclusivé. Assim, verifica-se que a segmentação da Figura 4.1(a), obtida por aplicação do algoritmo AP_{C4}-FCM (número total de *clusters* como condição de paragem) com 5 *clusters*, não permite definir um sub-conjunto de *clusters* que representem fidedignamente a região de upwelling, já que os píxeis esverdeados mais distantes da costa encontram-se num mesmo *cluster* que píxeis amarelos, que já não pertence à região de upwelling pretendida. Com a segmentação obtida na Figura 4.1(b) (AP_{C4}-FCM com 6 *clusters*), os píxeis verdes encontram-se praticamente todos num mesmo *cluster*, sendo que a região de upwelling pode ser correctamente identificada por esse *cluster*, juntamente com os *clusters* de temperatura média inferior.

Analisando os resultados obtidos por cada um dos algoritmos estudados (FCM, AP-FCM, *Histogram Thresholding*), o primeiro objectivo do estudo experimental é conseguir um algoritmo, e sua parametrização, que possibilite a obtenção de boas segmentações para o maior número de imagens. Através de métodos de comparação de segmentações, é feita uma análise qualitativa às segmentações obtidas. Paralelamente, é feita uma comparação em termos de custos computacionais entre os algoritmos, possibilitando a análise em termos de eficiência

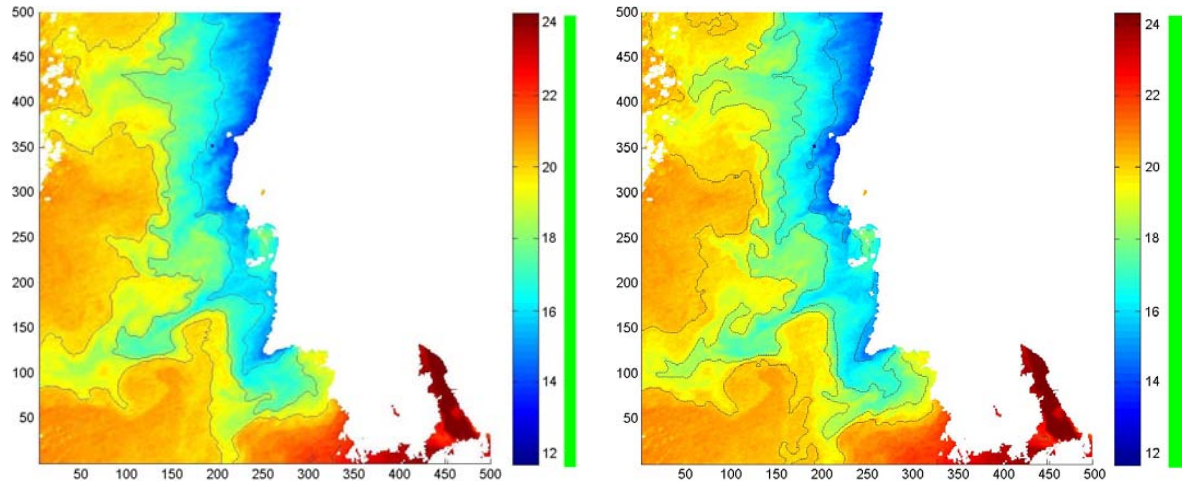


Figura 4.1 Para um mapa de temperaturas onde a região de upwelling se prolonga até aos píxeis verdes: (a) Segmentação que não permite representar correctamente o upwelling; (b) Segmentação que permite representar correctamente o upwelling.

computacional.

Posteriormente, é feito um estudo sobre a identificação das regiões de upwelling, com base no critério composto definido na Secção 3.2. Pretendendo-se identificar automaticamente regiões de upwelling o mais semelhantes possíveis às anotações feitas por oceanógrafos, o objectivo passa por avaliar qualitativamente os resultados da aplicação do critério composto e comparar a sua aplicação com *thresholds* definidos empiricamente e por análise ao ganho de informação.

4.2 Imagens SST e mapas binários ground-truth

O trabalho desenvolvido teve como objecto de estudo um conjunto de 61 mapas de temperatura, correspondentes a duas épocas completas de upwelling, referentes aos anos de 1998 (30 mapas) e 1999 (31 mapas). Na definição de *thresholds* para parametrização dos algoritmos utilizados e aplicação no critério composto para identificação da região de upwelling, os mapas relativos ao ano de 1998 foram utilizados como conjunto de treino, servindo os mapas de 1999 para conjunto de teste, possibilitando a confirmação dos resultados apurados com o conjunto de treino. Refira-se que as duas amostras são independentes e possuem uma diversidade de situações onde o fenómeno de upwelling se encontra presente.

Todos os mapas de temperatura estão sob a forma de uma matriz píxeis de dimensão 500×500, onde cada píxel possui o valor da temperatura de superfície oceânica, em graus Celsius, na sua localização geográfica. Os píxeis correspondentes a massa terrestre, extensões nebulosas

ou leituras incorrectas por parte do satélite, possuem o valor NaN. Todos os mapas representam a mesma região, entre as latitudes 36.8º Norte e 40.8º Norte, e longitudes 5.8º Oeste e 12.1 º Oeste, e cada píxel tem uma resolução espacial de $1.1Km \times 1.1Km$.

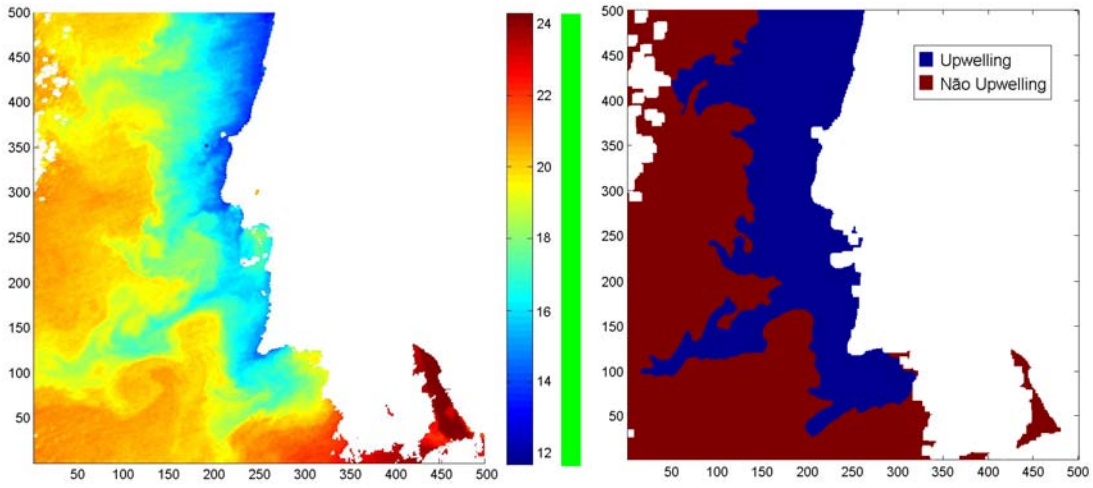
Para todos os mapas de temperatura do conjunto de dados sobre o qual se trabalhou foi também disponibilizada uma anotação, feita por oceanógrafos, com informação verbal sobre a escala de cores que define os limites de upwelling. A Tabela 4.1 exibe as três anotações feitas paras os mapas de temperatura visualizados nas Figuras 3.1, 3.2 e 3.3 (pág. 45), onde ao lado de cada mapa SST está colocada uma barra vertical, com a cor dos píxeis limítrofes da região de upwelling, segundo a anotação fornecida. Destaque-se o facto de que existem situações, tipicamente em casos em que o upwelling não está bem definido, onde a anotação varia consoante a região do mapa. Os limites indicados em cada anotação indicam, em termos de gama de temperaturas, os píxeis mais quentes que ainda estão incluídos na região de upwelling. Por exemplo, para o mapa SST da Figura 3.1 a indicação dada foi de que o upwelling ocorre desde as águas mais frias, de tom azul junto à costa, até às águas representadas por píxeis de cor verde, inclusivé. Note-se que estas anotações contrariam a natureza difusa do fenómeno, já que a transição de píxeis de uma cor para outra não se faz de uma forma brusca, havendo como é natural uma região de transição na gama de cores. Esta situação faz com que muitas vezes seja difícil marcar com exactidão uma região de upwelling e, consequentemente, dois oceanógrafos possam definir duas regiões diferentes para uma mesma imagem. O Anexo B disponibiliza todos os mapas SST estudados, onde se pode verificar a grande variedade de situações onde o fenómeno do upwelling ocorre. São também apresentadas as anotações feitas por oceanógrafos para cada mapa.

Mapa SST	Anotação
19980612	Limite verde até C. Espichel; para Sul deste cabo, limite amarelo.
19980802	Limite verde para toda a costa Oeste.
19980924	Limite verde até C. Espichel; para Sul, limite amarelo.

Tabela 4.1 Anotações dadas por oceanógrafos para a definição da região de upwelling para três mapas SST.

Assim, o estudo elaborado teve como princípio a tentativa de obter segmentações que permitam definir as mesmas regiões que as anotações de oceanógrafos referem. Para permitir a avaliação automática da qualidade das segmentações, criaram-se mapas binários "ground-truth" a partir das anotações para cada imagem. Nestes mapas, aos píxeis pertencentes à região de upwelling foi atribuída uma mesma *label* e aos restantes píxeis uma outra *label*. Por exemplo, a Figura 4.2(b) representa o mapa "ground-truth", onde os píxeis a azul correspondem à região de upwelling, para o mapa de temperaturas visualizado na Figura 4.2(a).

Para trabalhar os dados dos mapas de temperatura da superfície oceânica que foram disponibilizados, é feita uma transformação da matriz $a \times l$ original, onde cada posição se refere a um ponto no mapa, para uma matriz de atributos de dimensão $(a * l) \times 1$, ou seja, um vector coluna em que o único atributo considerado é a temperatura. O conjunto de dados recebido pelos algoritmos de *clustering* é o vector coluna criado a partir de cada mapa SST.



(a) (b)

Figura 4.2 (a) Mapa de temperaturas SST (2 de Agosto de 1998); (b) Mapa binário "ground-truth" do mapa SST da alínea (a).

4.3 Segmentação de imagens SST com FCM e sua validação

Para a aplicação do algoritmo FCM (Secção 2.3.1) a cada um dos mapas SST, é feita uma transformação do mapa original para uma matriz de atributos. Assim, o *input* para o algoritmo será uma matriz-coluna contendo todos os píxeis com valores de temperatura legíveis, ou seja, removendo os píxeis relativos a extensões nebulosas ou terrestres, identificados com o valor *NaN*, e o número total de *clusters*.

A obtenção de segmentações com a execução do algoritmo FCM foi feita para os valores de c entre 2 e 10, inclusivé. Estes valores foram estabelecidos pretendendo gerar segmentações que vão desde o mais redutor possível (2 *clusters*), até um estado em que se atingem resultados com sobre-segmentação, ou seja, o número de regiões encontradas nos mapas é mais do que o pretendido, havendo um nível de detalhe superior ao desejável. O parâmetro de *fuzzificação m* foi estabelecido em 2, seguindo o valor mais utilizado na literatura.

Sabendo que a qualidade dos resultados obtidos pela aplicação do FCM é dependente da geração aleatória de protótipos iniciais, uma das soluções para contornar esta questão mais reconhecida na literatura [13, 21] passa por aplicar o algoritmo várias vezes, sempre ao mesmo conjunto de dados mas com protótipos iniciais distintos e, dentro das várias aplicações escolher aquela que melhore a função objectivo J_m. No trabalho desenvolvido, cada execução do algoritmo FCM conteve 10 computações, ou seja, o resultado de uma aplicação é o melhor resultado de 10 computações do algoritmo, partindo de protótipos iniciais distintos. Assim, para cada execução, entende-se por número de iterações total e tempo total de execução do algoritmo como a soma das iterações e tempo gasto no total das 10 computações.

Mesmo sendo possível reduzir o valor em algumas computações, um número bastante reduzido de computações é insuficiente para garantir os melhores resultados, levando à possibilidade de convergência da função objectivo para mínimos locais.

No Anexo F estão disponíveis os resultados de segmentação (mapas de pertença e fronteiras *crisp* de *clusters*), entre $c = 2$ e $c = 10$ *clusters*, por aplicação do FCM.

4.3.1 Validação do melhor número de *clusters*

A tentativa de resolução do problema em causa com o algoritmo FCM carece evidentemente de resolução para a questão transversal aos algoritmos de agrupamento relativa ao número de grupos. Mesmo sabendo que num mapa de temperaturas oceânicas não existem, *per se*, grupos naturais distintos, é necessário encontrar boas segmentações. Para resolver essa situação o método utilizado baseou-se na análise de índices de validação aplicados aos resultados obtidos, seguindo o procedimento de validação descrito na Tabela 2.5.

Por análise experimental, verificou-se que as segmentações feitas com base em números de *clusters* reduzidos, 2 ou 3, dão origem a piores resultados para a identificação da região de upwelling. Uma das razões para essa situação prende-se desde logo com a localização geográfica dos mapas SST e as características da região. Por exemplo, independentemente da ocorrência de upwelling ou não, os píxeis que estão mais a Sul nos mapas, nomeadamente os que se encontram numa latitude inferior ao Algarve, possuem normalmente uma temperatura média superior aos restantes. Assim, quando se obtém uma segmentação com 2 *clusters*, independentemente do algoritmo, é natural que os grupos encontrados reflictam essa morfologia, segmentando águas da Região Norte, de temperatura mais reduzida, de águas da Região Sul, de temperatura superior. Devido ao comportamento típico do upwelling na região em estudo, surgindo junto à costa Oeste da Península Ibérica e "crescendo" na direcção do oceano, numa segmentação com 2 *clusters* é improvável que se consiga identificar correctamente a região de upwelling. Em aplicações com 3 *clusters* verifica-se que, tipicamente, as segmentações obtidas ainda são muito redutoras em relação às massas de água que se visualizam no mapa SST original, não conseguindo assim uma segmentação ideal, ou seja, aquela que permite distinguir os píxeis identificados como upwelling nas marcações feitas por oceanógrafos.

Com base nessa análise, constata-se nos resultados das Figuras 4.3-4.7 que a definição do número de *clusters* com base nos índices PC (Equação 2.23), PE (Equação 2.24), XB (Equação 2.18) e PBMF (Equação 2.22) não resulta, regra geral, em boas segmentações para os 61 mapas. O índice que obtém melhores resultados é o FS (Equação 2.21), já que análise empírica aos resultados obtidos indicou que as segmentações obtidas com um número de *clusters* entre 5 e 7 dão origem a bons resultados, em termos de comparação com as anotações feitas por oceanógrafos. Análise aos resultados obtidas com 9 e 10 *clusters* mostrou que, embora possibilitando o reconhecimento de regiões de upwelling, as segmentações obtidas estavam sobre-segmentadas. No problema em questão, não havendo uma definição exacta, entende-se por sobre-segmentação como um estado em que a segmentação obtida contém informação excessiva para a definição do upwelling, nomeadamente através de um excessivo número de *clusters*.

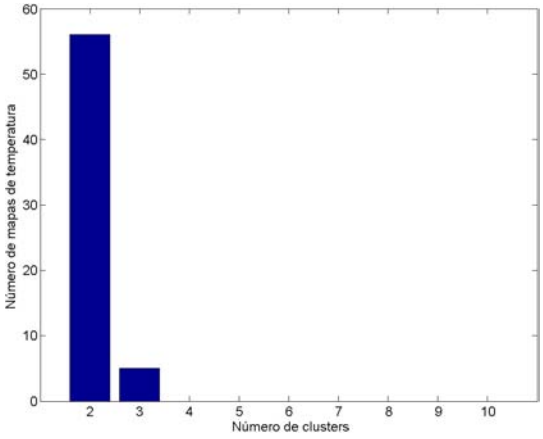


Figura 4.3 Frequência do número final de *clusters*, de acordo com o índice *Partition Coefficient*.

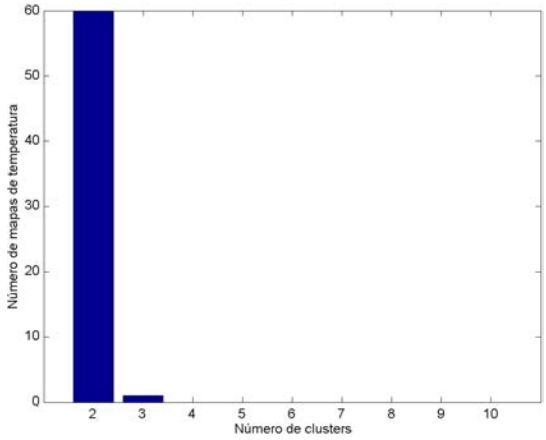


Figura 4.4 Frequência do número final de *clusters*, de acordo com o índice *Partition Entropy*.

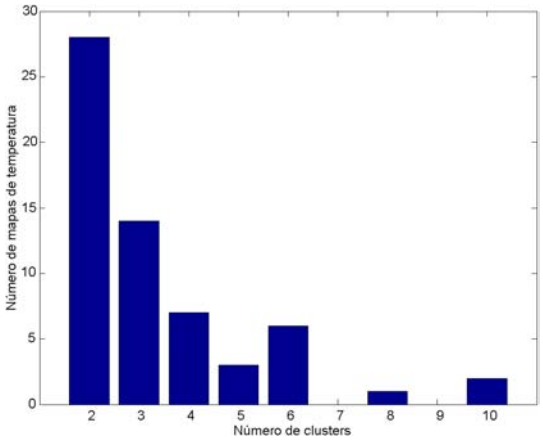


Figura 4.5 Frequência do número final de *clusters*, de acordo com o índice Xie-Beni.

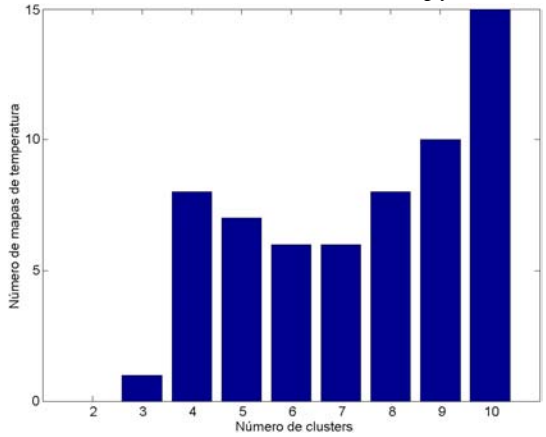


Figura 4.6 Frequência do número final de *clusters*, de acordo com o índice Fukuyama-Sugeno.

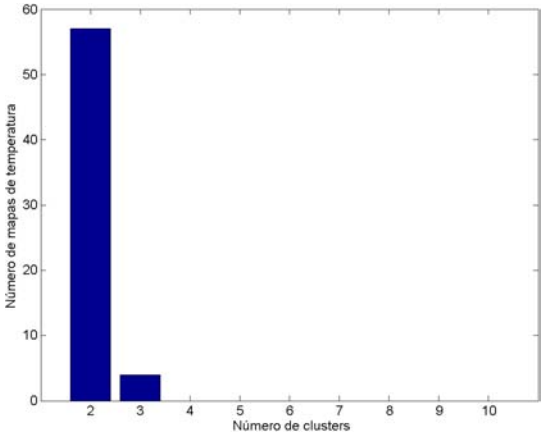


Figura 4.7 Frequência do número final de *clusters*, de acordo com o índice PBMF.

4.4 Segmentação de imagens SST com AP-FCM

Como referido aquando a introdução do algoritmo, a aplicação do AP-FCM (Secção 3.1) aos mapas SST em estudo pode-se dividir em duas fases distintas: a primeira fase, denominada por "Divisão & Conquista" (D&C), é composta pela iteração do algoritmo *Anomalous Pattern* e tem como objectivo a resolução de dois factores que são apontados como problemas do FCM: a necessidade de introduzir como parâmetro de entrada o número de *clusters* e a susceptibilidade que os resultados têm relativamente à inicialização de protótipos iniciais; a segunda fase é composta pela execusão do algoritmo FCM, onde os protótipos iniciais utilizados são os resultantes da fase anterior. Assim, o estudo em termos de número de iterações gastas pela aplicação do AP-FCM é feito tendo em conta as iterações utilizadas em cada uma das fases do algoritmo.

Uma das condições de paragem para o algoritmo de Divisão & Conquista estudada foi a iteração do *Anomalous Pattern* até todas as entidades terem sido afectadas a um dos protótipos gerados (AP-C1). A Figura 4.8 apresenta as frequências para os valores de número de *clusters* com esta condição de paragem e verifica-se que em 59 de 61 mapas de temperatura o número de grupos resultante é superior a 7. Por análise visual aos resultados obtidos, constata-se que as segmentações obtidas se encontram com sobre-segmentação para definir o upwelling. Por exemplo, na segmentação visualizada na Figura 4.10(a), a condição de paragem AP-C1 resultou em 9 *clusters* e, apesar de se poder a definir a região de upwelling, como a união dos primeiros 4 *clusters*, constata-se que há um excessivo número de grupos, verificando-se, por exemplo a existência, em algumas regiões, de fronteiras de *clusters* a distâncias muito próximas entre si. Contrariamente, a Figura 4.10(b) apresenta uma segmentação onde se consegue fazer uma relação entre os grupos obtidos e a cor dos píxeis visualizados na imagem, permitindo uma boa definição da região de upwelling. O Anexo C contém as segmentações (mapas de pertença e fronteiras *crisp* de *clusters*) obtidas pela aplicação do algoritmo AP_{C1}-FCM, sendo possível confirmar a sobre-segmentação de vários resultados.

Com base nas segmentações obtidas, verificou-se que na gama de valores possíveis para o número de *clusters*, que com a aplicação do AP-FCM vai desde 2 *clusters* até ao número indicado pela afectação de todas as entidades a um dos protótipos gerados pelo *Anomalous Pattern*, os melhores resultados para a detecção do upwelling são obtidos tipicamente por segmentações com um número médio de *clusters*, entre 5 e 8. Esta análise vai ao encontro de um dos objectivos de qualquer problema de segmentação de imagem: a pretenção de obter segmentações que não sejam muito redutoras (poucos *clusters*), nem segmentações com excessivo detalhe ou sobre-segmentação (muitos *clusters*).

Com o estudo da contribuição para a dispersão total de dados (Equação 3.3), em [12] propõe-se duas condições de paragem: a soma das contribuições dos primeiros *k clusters* gerados pelo *Anomalous Pattern* ser superior a um determinado *threshold* (condição AP-C2) ou a contribuição do último *cluster* ser inferior a um determinado *threshold* (AP-C3). O algoritmo AP_{C2}-FCM não foi alvo do estudo feito porque verificou-se experimentalmente a impossibilidade de definir um *threshold* que obtenha segmentações com um número de *clusters* considerado bom para

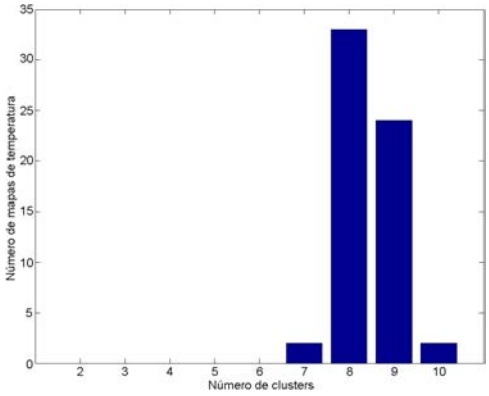


Figura 4.8 Frequência do número final de *clusters* por aplicação do algoritmo AP_{C1}-FCM, que tem o tratamento de todas as entidades como condição de paragem para o *Anomalous Pattern*.

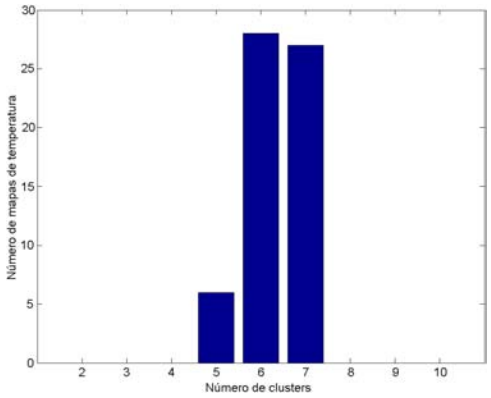


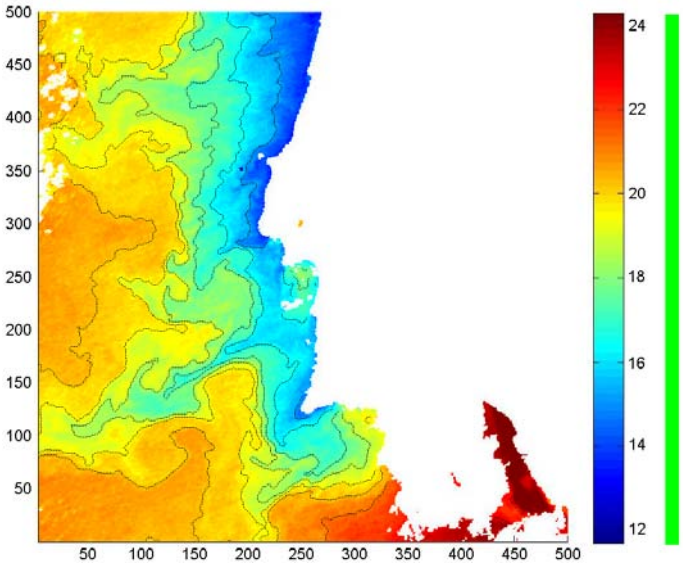
Figura 4.9 Frequência de número final de *clusters* por aplicação do algoritmo AP_{C3}-FCM, com condição de paragem estabelecida por análise à dispersão de dados do último *cluster* gerado pelo *Anomalous Pattern*, com *threshold* definido em 0.1%

.

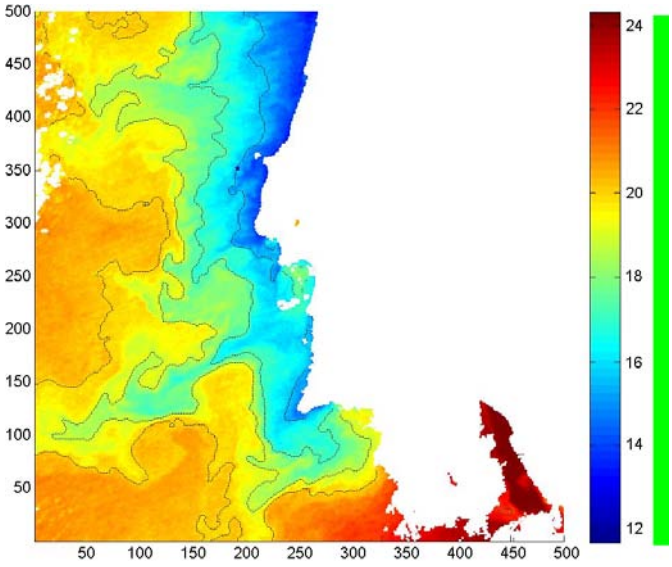
todos os mapas de temperatura. Esta impossibilidade deve-se à grande variabilidade na contribuição acumulada para a dispersão de dados entre os vários mapas de temperatura, entre o quinto e sétimo *clusters* extraídos pelo algorimo *Anomalous Pattern*.

Contrariamente, verificou-se que a condição AP-C3 é bastante estável para o conjunto de imagens disponíveis. O valor para o *threshold* foi estabelecido com base na análise empírica das contribuições obtidas para cada *cluster* e fixado em 1×10^{-3}. Destaque-se que este valor foi estabelecido inicialmente apenas com análise às segmentações obtidas para ano de 1998, resultando nas 30 imagens em partições com 5, 6 ou 7 *clusters* e que quando aplicado ao ano de 1999, a gama de valores resultante para o número de *clusters* não se alterou. Esta situação e os valores da contribuição para a dispersão total de dados de cada *cluster* extraído pelo *Anomalous Pattern* podem ser verificados no Anexo C, nas Tabelas C.1 e C.2 (pág. 133, 134).

Para o *threshold* definido para a condição AP-C3, a Figura 4.9 mostra a frequência das segmentações em termos de número total de *clusters*, quando aplicado o AP-FCM com a análise da contribuição para a dispersão de dados como condição de paragem. Refira-se que o objectivo de determinar o número total de *clusters* com este método pretende com que esse valor não seja calculado estaticamente (definindo 6 como uma boa segmentação para todas as imagens, por exemplo), mas que o seja com base na própria estrutura dos dados presentes em cada mapa de temperaturas. Nas Figuras 4.10(b), 4.11(a) e 4.11(b) podem-se visualizar as fronteiras dos clusters obtidas pelo algoritmo AP_{C3}-FCM, com *threshold* em 1×10^{-3}, resultando em boas segmentações para três cenários onde há ocorrência do upwelling, definindo automaticamente o número de *clusters* em 6, 7 e 6, respectivamente. Para os conjuntos de mapas relativos aos anos de 1998 e 1999, apresentam-se as segmentações resultantes pela aplicação do algoritmo AP_{C3}-FCM no Anexo D, onde se pode verificar que as segmentações obtidas não se encontram em



(a)



(b)

Figura 4.10 (a) Segmentação do mapa de temperatura de 2 de Agosto de 1998, com a condição de paragem AP-C1 para o algoritmo AP-FCM, resultando em 9 *clusters*; (b) Segmentação para o mesmo mapa SST, com condição de paragem AP-C3 (com threshold 1×10^{-3}), resultando em 6 *clusters*.

sub ou sobre-segmentadas, sendo de boa qualidade para a identificação da região de upwelling.

Outra condição de paragem apresentada em [12], necessita da introdução do número total de *clusters*, como parâmetro de entrada. No entanto, esta solução vai contra o objectivo do estudo do AP-FCM para solucionar o problema da necessidade de introduzir o número de grupos como parâmetro. Por motivos de completude e para comparação com a versão original do FCM, também se aplicou esta condição de paragem ao conjunto de mapas em estudo.

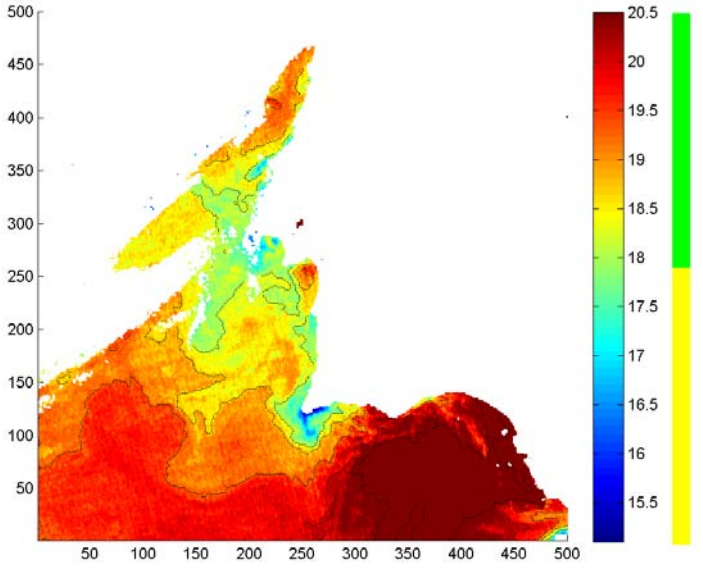
4.5 Segmentação de imagens SST por Histogram Thresholding

Como alternativa às técnicas de segmentação de imagem com recurso a algoritmos de *clustering*, implementou-se também uma segmentação com base numa técnica com *Histogram Thresholding*, descrito na Tabela 2.6 da Secção 2.5.2. Uma vez que as segmentações obtidas por esse método são estritamente binárias, e que executando apenas uma única aplicação não se obtêm bons resultados, já que em 31.15% dos 61 mapas de temperatura acaba por segmentar apenas as massas de água a Sul do Algarve, devido à sua temperatura média elevada, foi implementada uma versão iterativa do algoritmo, de modo a obter um maior número de *clusters*, cujos passos estão descritos na Tabela 4.2.

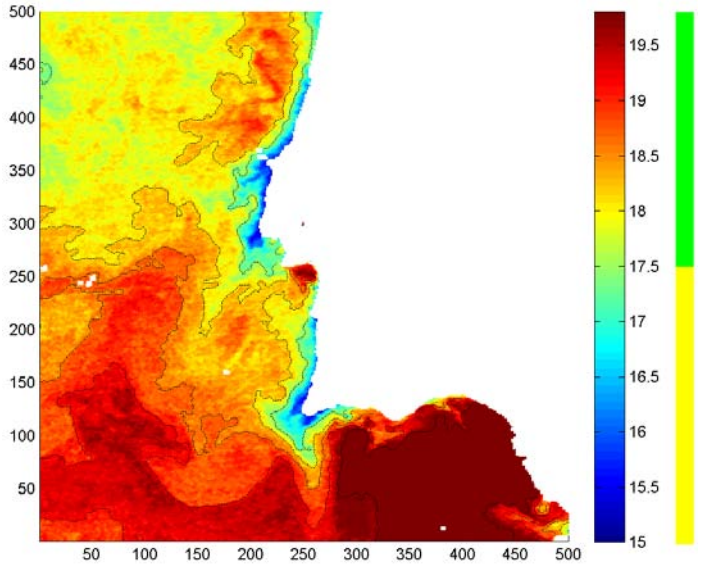
1. Segmentar o mapa I, em duas regiões pelo algoritmo *Histogram Thresholding*. $c = 2$.
2. Testar condição de paragem. Se quebrar condição, termina, com o mapa I segmentado em c grupos, caso contrário, prossegue.
3. Tendo o mapa I dividido em c segmentos, selecciona o *cluster* com maior cardinalidade: $|S| = max(|S_1|, ..., |S_c|)\}$.
4. Segmentar S_j em duas regiões pelo algoritmo *Histogram Thresholding*. Se não se tiver atingido um valor de c pré-estabelecido, ou outra condição de paragem, $c = c+1$ e volta-se ao passo 2.

Tabela 4.2 Passos da iteração do algoritmo *Iterative Thresholding*.

Uma das vantagens do método aplicado prende-se com o seu bom comportamento em imagens onde os histogramas apresentam uma grande variabilidade, já que por análise empírica, verificou-se que os histogramas obtidos para o conjunto de mapas disponível não seguem sempre a mesma distribuição, podendo ser unimodais (Figura 4.12(a)), bimodais (Figura 4.12(b)) ou multimodais (Figura 4.12(b)). Idealmente, os histogramas das imagens com uma região de



(a)



(b)

Figura 4.11 Segmentações por aplicação do algoritmo AP_{C3}-FCM, com threshold 1×10^{-3}, aos mapas de temperatura de: (a) 9 de Junho de 1998 e (b) 12 de Junho de 1998, resultando em 7 e 6 *clusters*, respectivamente.

upwelling bem definida seriam bimodais, com uma frequência elevada dos píxeis (de temperatura inferior) correspondentes à região de upwelling, frequência elevada dos píxeis (de temperatura superior) não pertencentes à região de upwelling, e uma frequência reduzida dos píxeis situados na fronteira da região de upwelling, indicando uma diferença brusca de temperatura. Contudo, devido à falta de uma boa definição do gradiente térmico, tal situação não se verifica, havendo uma maior variedade de situações em que o upwelling ocorre.

Na aplicação desta técnica ao problema de segmentação de mapas de temperatura, o eixo das abcissas do histograma contém os valores de temperatura, divido em intervalos de temperatura, e o eixo das ordenadas a frequência de píxeis para cada um dos intervalos. O número de intervalos aplicado em cada imagem foi obtido por aplicação da medida de Freedman-Diaconis (*cf.* [50]), que calcula a amplitude de cada intervalo, com base na distribuição do conjunto de dados, segundo a seguinte fórmula:

$$h = 2\frac{IQR(X)}{n^{\frac{1}{3}}}, \tag{4.1}$$

onde $IQR(X)$ representa a amplitude interquartil do conjunto X, ou seja, a diferença entre o terceiro e primeiro quartis, e n o número de entidades do conjunto X. O número de intervalos é calculado dividindo a amplitude térmica de cada imagem por intervalos de amplitude h. A Figura 4.13 contém a representação do número de intervalos resultante por aplicação deste método, para as imagens do ano de 1998 e 1999. Note-se que o número de intervalos resultante nunca é muito reduzido (o valor mínimo é 63) e que, se tal acontecesse, a qualidade dos resultados poderia ser afectada negativamente, já que quanto maior for cada intervalo de temperatura, mais informação se perde, em termos dos valores de temperatura dos píxeis que acabam por ficar todos no mesmo intervalo. A existência de mapas de temperatura com mais de 300 intervalos não é excessiva, já que, para os dois anos estudados e excluindo os valores *NaN*, o número de píxeis é sempre superior a 70000, ou seja, um valor muito superior ao número total de intervalos.

Desde logo, se verifica que uma das grandes vantagens desta técnica está relacionada com a redução do espaço de pesquisa, sendo este diminuído desde o número total de píxeis, que podem ultrapassar os 150000 nos mapas disponíveis, para o número de intervalos do histograma vezes 2 (valor médio do atributo para cada intervalo e frequência para cada intervalo). No Anexo G pode-se verificar a boa qualidade das segmentações obtidas pela aplicação do algoritmo *Iterative Thresholding*, nomeadamente nas segmentações com um número de *clusters* considerado ideal (entre 5 e 7).

4.6 Comparação da qualidade das segmentações resultantes dos algoritmos FCM, AP-FCM e Histogram Thresholding

Para a comparação dos resultados obtidos pelos algoritmos testados, o primeiro estudo feito tratou de comparar as segmentações obtidas com a versão original do algoritmo FCM (Tabela 2.4)

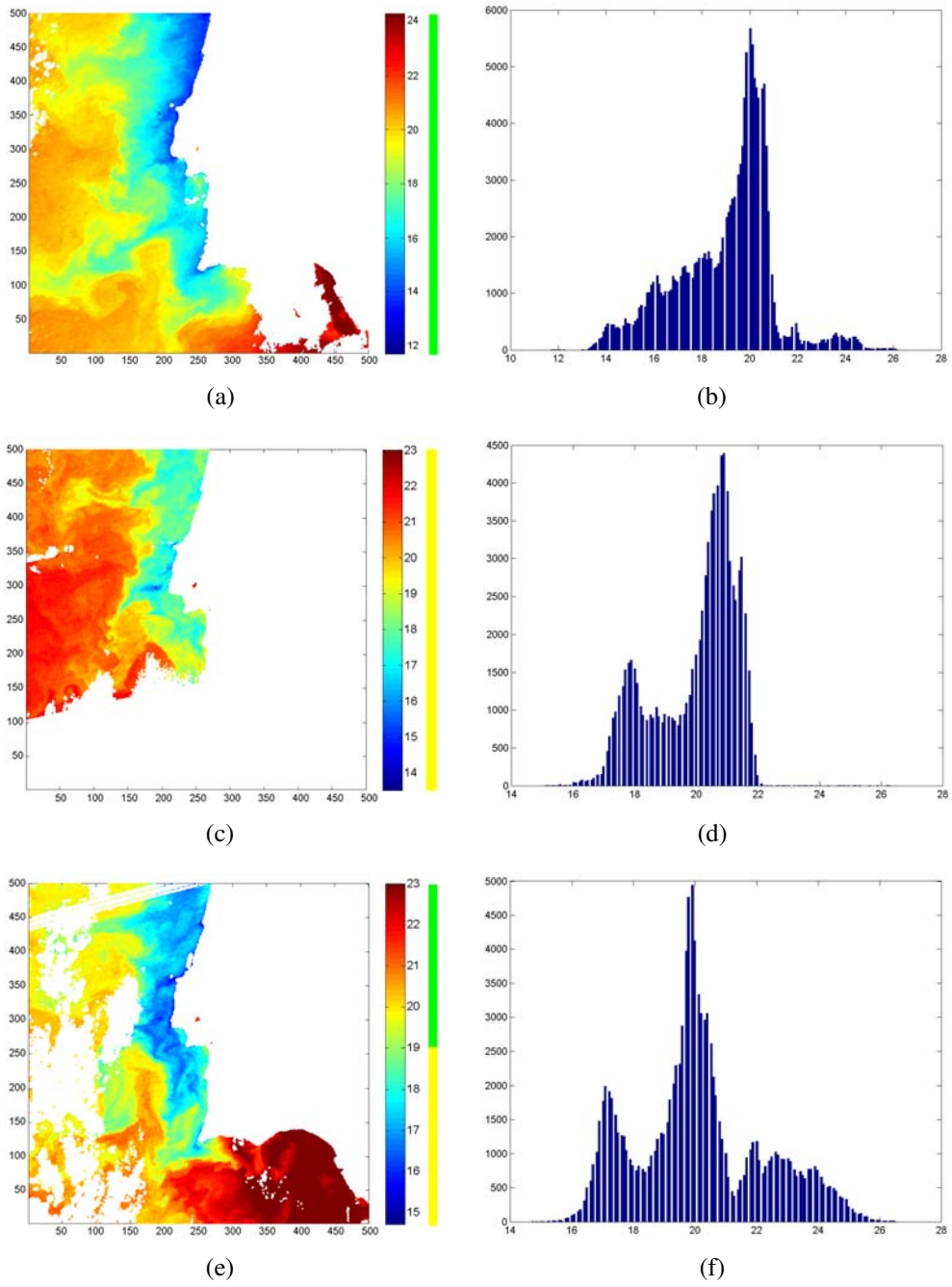


Figura 4.12 (a) Histograma (unimodal) do mapa SST 19980802; (b) Histograma (bimodal) do mapa SST 19980812; (c) Histograma (multimodal) do mapa SST 19980819.

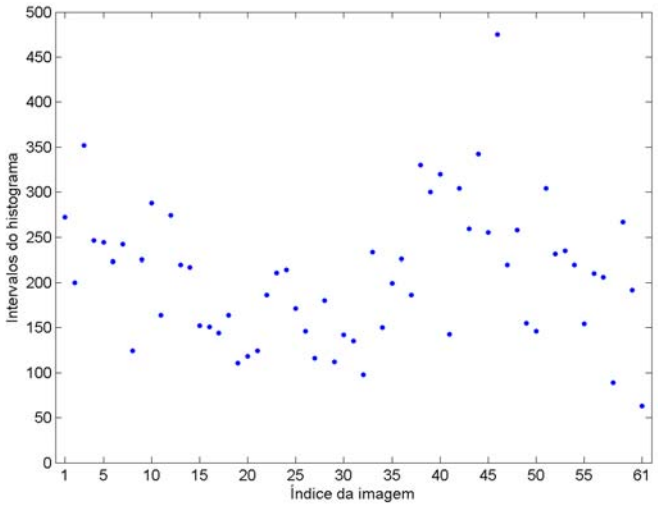


Figura 4.13 Número de intervalos para os histogramas dos mapas de temperatura de 1998 e 1999, calculados com base na medida de Freedman-Diaconis.

com as do algoritmo AP-FCM com a condição de paragem AP-C4 (número total de *clusters*), garantindo-se que se comparam segmentações com o mesmo número de grupos. Nesta análise, não se teve em conta se as segmentações obtidas permitiam uma boa identificação das regiões de upwelling, sendo que o objectivo foi verificar o efeito que a inicialização de protótipos pelo método *Anomalous Pattern* tem sobre a execução do algoritmo FCM. Esta comparação foi feita com base na associação dos píxeis aos *clusters*, após o passo de *desfuzzificação*. Relembre-se que, em qualquer um dos casos, a identificação dos *clusters* está ordenada pela temperatura do seu protótipo, pelo que as segmentações obtidas por ambos os algoritmos são passíveis de serem comparadas *cluster* a *cluster*.

A Tabela 4.3 contém essa comparação feita em termos de percentagens de píxeis classificados em *clusters*, ou seja, a razão entre os píxeis afectados a *clusters* distintos, entre segmentações dos dois algoritmos (FCM e AP_{C4}-FCM), e o número total de píxeis da imagem, excluindo os valores *NaN*. Assim, quanto mais semelhantes forem as segmentações obtidas pelos dois algoritmos, menor será a percentagem. Desde logo, nota-se que a percentagem de píxeis associados a clusters diferentes acompanha o aumento do número de *clusters*, notando-se também o aumento do desvio padrão da percentagem de diferenças, sendo esse facto indicativo de que mesmo continuando a haver segmentações semelhantes - para 8 *clusters*, 25 das 61 segmentações diferem em menos de 3% dos píxeis -, as variações atingem valores cada vez mais elevados. Verificou-se ainda que para o total de 427 segmentações comparadas, relativamente a 61 mapas de temperatura, com o número de *clusters* a variar entre 2 e 8, em 79.86% das segmentações a diferença entre os resultados dos dois algoritmos é inferior a 8% dos píxeis do respectivo mapa. Assim, tendo em costa os valores reduzidos da percentagem de píxeis classificados em clusters distintos, a Tabela 4.3 serve como indicativo da efectiva semelhança entre as

segmentações de ambos os algoritmos, ou seja, com uma execução do AP_{C4}-FCM conseguem-se atingir resultados bastante semelhantes aos do FCM.

	2c	3c	4c	5c	6c	7c	8c
Média (%)	0,02	0,24	1,86	4,30	7,08	8,42	15,38
σ(%)	0,15	1,02	9,64	8,98	11,21	12,67	18,25

Tabela 4.3 Comparação entre FCM e AP-FCM com número de *clusters* como condição de paragem (AP-C4) para o *Anomalous Pattern* em termos de píxeis classificados em *clusters* distintos (%).

Outro ponto de análise para avaliar as segmentações obtidas pelos algoritmos FCM e AP-FCM teve como base a comparação com os mapas binários "ground-truth" criados para cada mapa de temperatura, como descritos na Secção 4.2. Procurando uma segmentação onde uma das fronteiras dos *clusters* obtidos seja o mais semelhante possível à fronteira da região de upwelling no mapa "ground-truth", aplicou-se um passo de transformação de uma segmentação com $c > 2$ para uma segmentação binária com $c = 2$. Nesse passo, formulado no Algoritmo 2, a cada *cluster* presente na segmentação é associada a label do *cluster* do mapa "ground-truth" com uma temperatura média mais próxima, ou seja, a cada *cluster* na segmentação associa-se uma *label* indicando a pertença, ou não, à região de upwelling pretendida. Com este passo, torna-se uma segmentação com um qualquer número de *clusters* numa segmentação binária, onde um grupo indica a região de upwelling e o outro o resto do mapa de temperaturas.

Algoritmo 2 Transformação de uma segmentação (S1) com $c > 2$ *clusters* numa segmentação binária, por comparação com mapa "ground-truth" (GT).

{$T_{k,S}$ indica a temperatura média de um *cluster* k numa segmentação S.}

for $k = 1 : c$ **do** {Para todos os *clusters* da segmentação S1}
    {Associação do cluster k de $S1$ ao *cluster* do mapa "ground-truth" mais próximo.}
    if ($|T_{k,S1}$ - $T_{1,GT}|$) < ($|T_{k,S1}$ - $T_{2,GT}|$) **then**
        $BinarySegment(k) = 1$;
    else
        $BinarySegment(k) = 2$;
    end if
end for
{Como *output*, tem-se a segmentação binária *BinarySegment*, onde a cada cluster k de S1, está associada a *label* 1, caso se considere esse *cluster* como pertencente ao upwelling, ou 2, caso contrário.}

A avaliação por comparação com mapas "ground-truth" foi feita com base em matrizes de confusão binárias, criadas a partir da *label* associando, ou não, cada píxel à região de upwelling. Assim, a matriz de confusão da Tabela 4.4 é definida por:

- *uu* - Número de píxeis classificados como pertencentes à região de upwelling no mapa resultado e, igualmente, no mapa "ground-truth".

		Mapa "Ground-Truth"	
		Upwelling	Não Upwelling
Mapa Segmentação	Upwelling	*uu*	*un*
	Não Upwelling	*nu*	*nn*

Tabela 4.4 Matriz de confusão genérica para comparação entre mapas binários.

- *un* - Número de píxeis classificados como pertencentes à região de upwelling no mapa resultado e como não pertencentes à região upwelling no mapa "ground-truth".
- *nu* - Número de píxeis classificados como não pertencentes à região de upwelling no mapa resultado e como pertencentes à região de upwelling no mapa "ground-truth".
- *nn* - Número de píxeis classificados como não pertencentes à região de upwelling no mapa resultado e, igualmente, no mapa "ground-truth".

Com base na matriz de confusão, foram adaptadas medidas de referência [48] para analisar a qualidade das segmentações binárias geradas, quando comparadas com os mapas "ground-truth":

- $Accuracy = (uu+nn)/(uu+un+nu+nn)$. Calcula a percentagem de píxeis classificados correctamente.
- $Recall = uu/(uu+nu)$. Calcula a percentagem de píxeis pertencentes à região de upwelling, de acordo com o mapa "ground-truth", correctamente classificados.
- $Precision = uu/(uu+un)$. Calcula a percentagem de píxeis classificados como região de upwelling, no mapa de segmentação binária, correctamente classificados.
- $F_\beta = (1+\beta^2)\frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$. Medida baseada na definição de *effectiveness* em [51]. Também conhecida por $F-measure$, foi utilizada a versão mais comum, com $\beta = 1$, que resulta na média harmónica entre as medidas Recall e Precision: $F_1 = \frac{2}{\frac{1}{Recall}+\frac{1}{Precision}}$.

Refira-se que as medidas *Precision* e *Recall* não conseguem quantificar, por si só, a qualidade de uma segmentação. Veja-se que uma alta percentagem de *Recall* pode ser conseguida à custa de uma baixa percentagem de *Precision*, bastando para tal classificar todos os píxeis como pertencentes à região de upwelling. Contrariamente, também se pode elevar a taxa de *Precision* e reduzir o valor de *Recall*, por exemplo, classificando como upwelling apenas o primeiro *cluster*, quando a região pretendida é ocupada pelos primeiros 2 ou 3 *clusters*. Por esta razão, os melhores indicadores da qualidade de uma segmentação são a *Accuracy* e a $F-Measure$, que é calculada com base nas medidas *Recall* e *Precision*. Ambas as medidas variam no intervalo $[0,1]$ e o resultado ideal é para o valor máximo de 1.

Em alternativa às medidas calculadas com base na matriz de confusão, também se fez um estudo analisando as distâncias entre protótipos de duas segmentações. Em [12], é proposta a

medida *Average distance between centroids* (*adc*) para analisar a estabilidade de uma validação cruzada, que tem como princípio a comparação das distâncias entre protótipos de segmentações distintas. Para a comparação de segmentações binárias, onde $c'_l (l = 1,2)$ são os protótipos do mapa "ground-truth" e $c_k (k = 1,2)$ os protótipos da segmentação binária, *adc* é a média das distâncias de cada um dos protótipos de c'_l ao protótipo mais próximo de c_k, ou seja:

$$adc = \frac{1}{2} \sum_{l=1}^{2} min_{k=1,2}(|c'_l - c_k|). \tag{4.2}$$

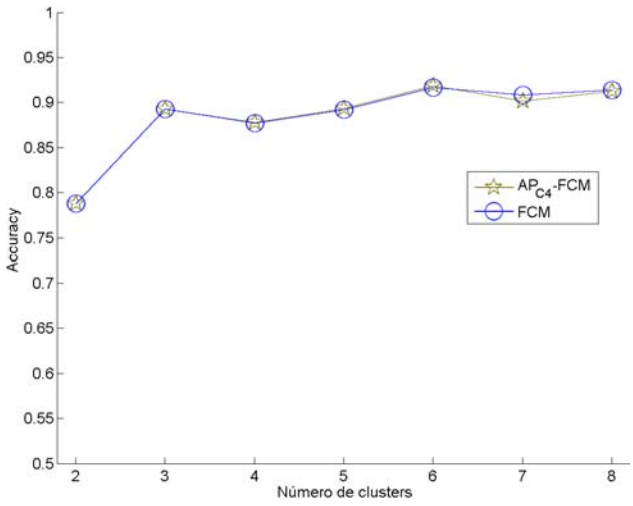
A medida *adc* representa uma distância média entre protótipos, pelo que pretendendo-se que as segmentações sejam o mais semelhantes possíveis entre si, as melhores segmentações são aquelas com menor valor.

A análise aos gráficos da Figura 4.14 indicam-nos claramente que o comportamento de ambos os algoritmos, com o número de *clusters* pré-fixado, é muito semelhante, independentemente da medida utilizada. Conhecendo a convergência iterativa do algoritmo FCM, independentemente dos protótipos iniciais, este facto não é de estranhar. Contudo, este estudo tem fundamentalmente o propósito de comparar as segmentações obtidas pelos dois algoritmos com o mesmo número de *clusters*, não resolvendo a problemática da determinação automática de uma segmentação com um número de *clusters* que permita uma boa definição da região de upwelling.

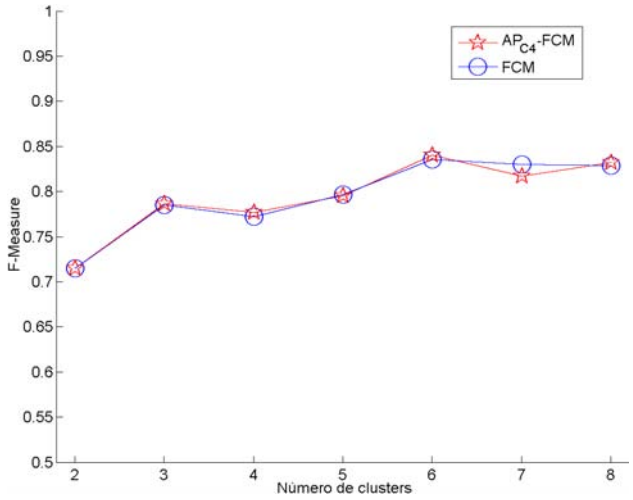
Refira-se que no gráfico da Figura 4.14, o número máximo de grupos para o algoritmo AP_{C4}-FCM é de 8, uma vez que, a partir desse valor o algoritmo já não obtém segmentações em muitos dos mapas de temperatura. Relembre-se o funcionamento do algoritmo AP-FCM (Tabela 3.1, pág. 42), que extrai os *clusters* iterativamente, associando-os a protótipos tentativos, tendo como limite a situação onde todas as entidadas já foram associadas a um protótipo tentativo (condição de paragem AP-C1). Assim, como se vê no gráfico de barras da Figura 4.8, há 33 mapas de temperatura cujo limite de *clusters* é 8, pelo que, se estabeleceu esse valor como limite máximo para as comparações entre os algoritmos AP_{C4}-FCM e FCM.

Para estudar segmentações que determinam automaticamente o número de *clusters*, comparam-se os resultados dos algoritmos AP_{C3}-FCM (condição de paragem definida com base na contribuição para a dispersão de dados do último *cluster* extraído pelo *Anomalous Pattern*) e FCM validado com os índices FS (Equação 2.21) e XB (Equação 2.18). Os restantes índices apresentados na Secção 2.4.3 (PBMF, *Partition Entropy* e *Partition Coefficient*) não foram explorados neste estudo devido ao facto de que as segmentações resultantes pela validação do FCM com esses indices serem sub-segmentadas para todos os mapas de temperatura de 1998 e 1999, como se confirma nos gráficos das Figuras 4.3, 4.4 e 4.7 (pág. 66).

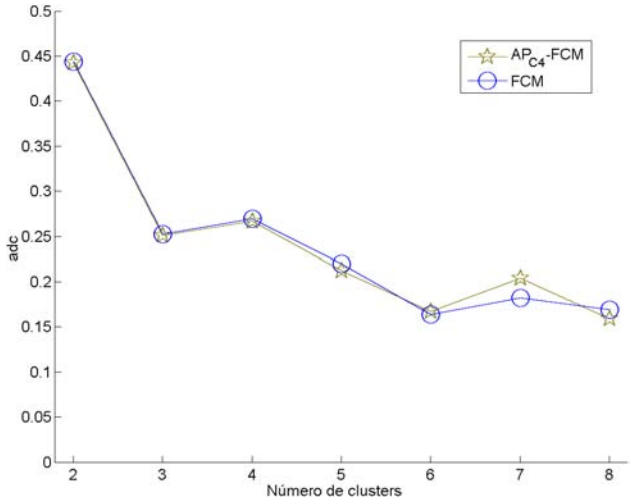
Os resultados da Figura 4.15(a) mostram que as segmentações obtidas com o algoritmo FCM, validado com o índice de XB, são as de pior qualidade, independentemente da medida utilizada. Este facto já era esperado já que, como explicado na Secção 4.3.1, a validação com este índice origina frequentemente resultados sub-segmentados, nomeadamente com 2 e 3 *clusters*, como se verifica no gráfico da Figura 4.5 (pág. 66). Comparando as segmentações obtidas com



(a)



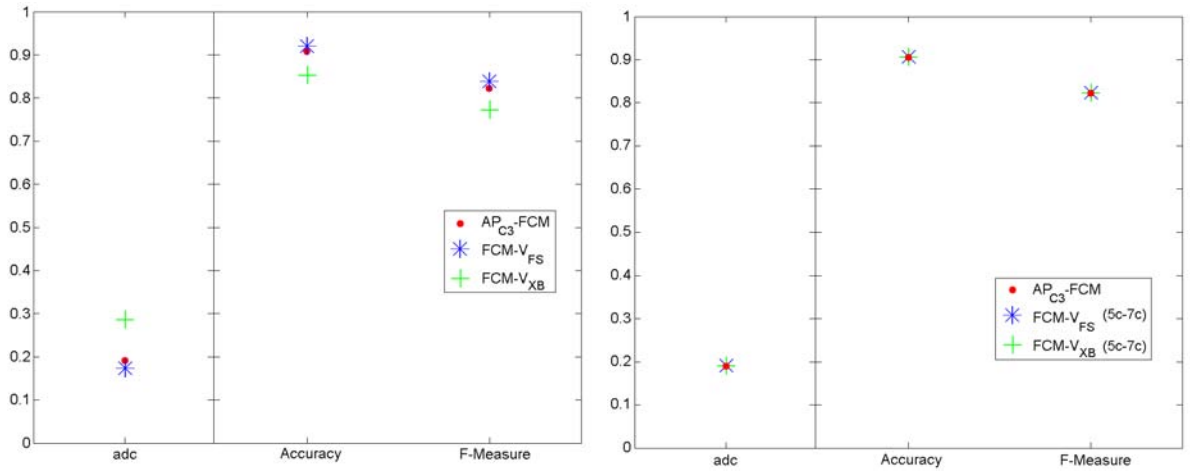
(b)



(c)

Figura 4.14 Comparação entre as segmentações obtidas pelos algoritmos AP$_{C4}$-FCM e *FCM* com as medidas: (a) *Accuracy*; (b) *F-Measure*; (c) *adc*.

os algoritmos AP$_{C3}$-FCM e FCM, validado pelo índice FS, verifica-se que, apesar de bastante semelhantes, as que apresentam melhor qualidade são obtidas pelo FCM. No entanto, como já referido, essas segmentações resultam frequentemente em resultados com sobre-segmentação, como se verifica no histograma da Figura 4.6, enquanto as segmentações obtidas por aplicação do AP$_{C3}$-FCM resultam sempre num número de *clusters* considerado bom para a identificação do upwelling (5, 6 ou 7 *clusters*). Para contornar esta questão, repetiu-se a análise às segmentações obtidas pelo FCM, validadas pelos mesmos índices, mas analisando apenas os resultados com um número de *clusters* a variar entre $c_{min} = 5$ e $c_{max} = 7$. O número de *clusters* resultante pela aplicação do algoritmo AP$_{C3}$-FCM, que dá origem a segmentações consideradas boas, serviu de indicador para o estabelecimento deste intervalo limitado de validação para o FCM. A frequência relativamente ao número de *clusters* resultante pela validação do FCM com os índices FS e XB, entre o referido intervalo, pode ser consultada nas Figuras 4.16 e 4.17, respectivamente. Desta forma, a qualidade dos resultados obtidos pelas três versões testadas é praticamente igual, como se visualiza na Figura 4.15(b), não sendo possível indicar um dos algoritmos como gerador de melhores segmentações.



(a) (b)

Figura 4.15 Comparação com as medidas *adc*, *Accuracy* e *F-Measure* entre os algoritmos AP$_{C3}$-FCM e: (a) *FCM* validado pelos índices de validação (de $c_{min} = 2$ a $c_{max} = 10$); (b) *FCM* validado pelos índices de validação (de $c_{min} = 5$ a $c_{max} = 7$)

Para testar a qualidade das segmentações obtidas, também foi aplicado um método baseado numa análise no espaço ROC (*Receiver Operating Characteristic*) [48]. O espaço ROC é utilizado para visualizar graficamente a qualidade de classificadores binários, utilizando valores calculados a partir da matriz de confusão binária apresentada na Tabela 4.4, nomeadamente a taxa de verdadeiros positivos, $TVP = uu/(uu + un)$, e a taxa de falsos positivos, $TFP = un/(un + nn)$. No espaço ROC, o eixo dos xx representa a medida TVP (ou sensitividade) e o eixo dos yy, a medida TFP (ou $1 - especificidade$). Para analisar a qualidade de

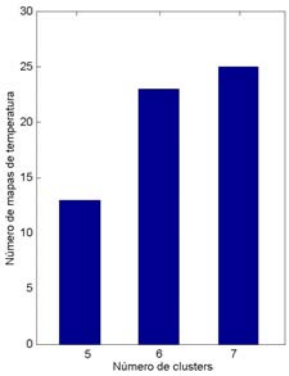


Figura 4.16 Frequência do número final de *clusters* com o algoritmo FCM validade com o índice FS, (de $c_{min} = 5$ a $c_{max} = 7$).

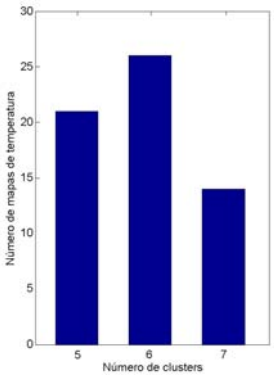


Figura 4.17 Frequência do número final de *clusters* com o algoritmo FCM validade com o índice XB, (de $c_{min} = 5$ a $c_{max} = 7$).

classificações binárias, as segmentações com um número de *clusters* superior a 2, são transformadas em segmentações binárias, pelo método descrito no Algoritmo 2, e comparadas com os mapas "ground-truth".

Uma classificação a partir de uma matriz de confusão resulta num ponto no espaço ROC, pelo que a análise das segmentações relativas a 1998 e 1999 gera um ponto para cada um dos mapas de temperatura. Uma classificação perfeita fica colocada no ponto $(0,1)$, correspondendo a $(TFP, TVP) = (0,1)$, ou seja, a segmentação binária é igual ao mapa "ground-truth". A linha que divide o espaço ROC, unindo os pontos $(0,0)$ e $(1,1)$ é considerada como uma classificação aleatória, pelo que quanto mais próximos os pontos estiverem da classificação perfeita, melhor é a qualidade das segmentações obtidas. A Figura 4.18 apresenta os resultados da comparação das segmentações binárias obtidas pelos algoritmos AP_{C3}-FCM e FCM validado, entre $c = 5$ e $c = 7$, pelos índices FS (Equação 2.21) e XB (Equação 2.21), para os dois conjuntos estudados, relativos aos anos de 1998 e 1999. Para facilitar a análise, a Tabela 4.5 sumariza os resultados, com o número de segmentações em cada intervalo 0.1×0.1 do espaço ROC. Verifica-se que, conforme o algoritmo utilizado, entre 33% a 43% das classificações para o ano de 1998 se encontram no intervalo mais próximo da classificação perfeita, melhorando ligeiramente no ano de 1999 (42%-48%). Mesmo as classificações que se encontram fora desse intervalo (TFP entre 0 e 0.1, TVP entre 0.9 e 0.1), estão situadas mais próximas à classificação perfeita do que à linha de classificação aleatória, pelo que se pode confirmar a qualidade efectiva das segmentações obtidas. Entre os três métodos de obtenção automática de segmentações, não se consegue destacar nenhum dos algoritmos como gerador de melhores resultados.

Por outro lado, comparando as segmentações obtidas pelos algoritmos *Iterative Thresholding* e AP_{C4}-FCM, verifica-se com base na análise aos gráficos da Figura 4.19 que as segmentações deste último são de ligeiramente melhor qualidade. Outro factor contra a utilização do *Iterative Thresholding*, é a inexistência de um método de determinação do número total de *clusters*, com base nos próprios dados, que consiga boas segmentações, impossibilitando a aplicação

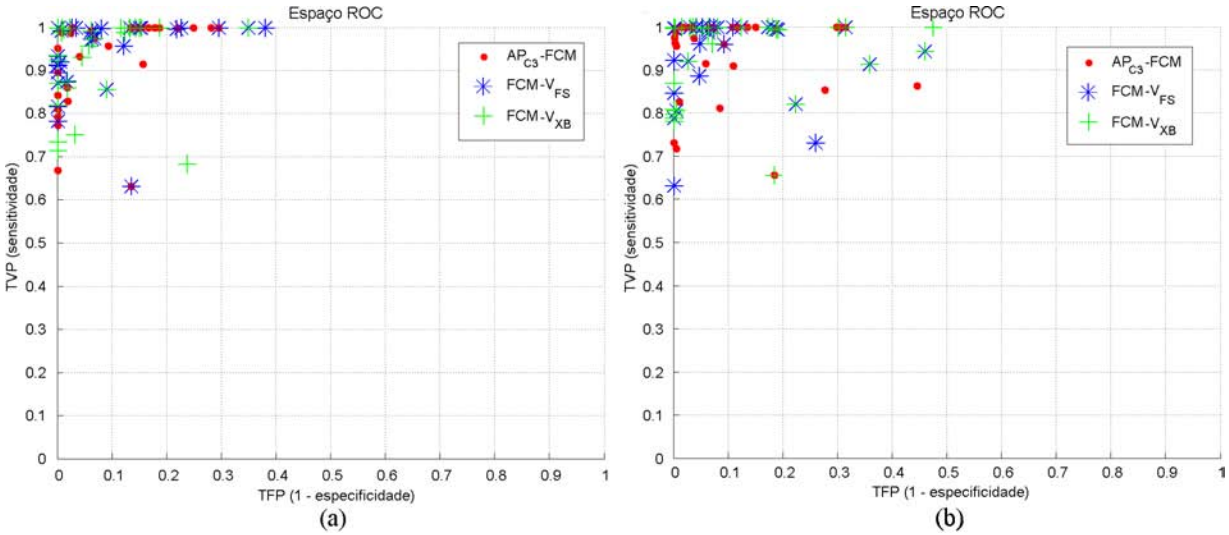


Figura 4.18 Classificação das segmentações binárias obtidas pelos algoritmos AP$_{C3}$-FCM e FCM validado, entre $c = 5$ e $c = 7$, pelos índices FS (FCM$_{V_{FS}}$) e XB (FCM$_{V_{XB}}$), comparadas com os mapas "ground-truth", para os anos de: (a) 1998; (b) 1999.

(a)

TVP (sensitividade)	AP$_{C3}$-FCM	FCM - V_{FS}	FCM - V_{XB}	AP$_{C3}$-FCM	FCM - V_{FS}	FCM - V_{XB}	AP$_{C3}$-FCM	FCM - V_{FS}	FCM - V_{XB}	AP$_{C3}$-FCM	FCM - V_{FS}	FCM - V_{XB}
0.9-1.0	10	13	12	7	5	8	4	3			2	1
0.8-0.9	5	5	5									
0.7-0.8	2	1	3									
0.6-0.7	1			1	1				1			
	0.0-0.1			0.1-0.2			0.2-0.3			0.3-0.4		

TFP (1 - especificidade)

(b)

TVP (sensitividade)	AP$_{C3}$-FCM	FCM - V_{FS}	FCM - V_{XB}	AP$_{C3}$-FCM	FCM - V_{FS}	FCM - V_{XB}	AP$_{C3}$-FCM	FCM - V_{FS}	FCM - V_{XB}	AP$_{C3}$-FCM	FCM - V_{FS}	FCM - V_{XB}	AP$_{C3}$-FCM	FCM - V_{FS}	FCM - V_{XB}
0.9-1.0	15	15	13	6	6	5	1			2	2	3		1	2
0.8-0.9	2	3	3				1	1	1	1					
0.7-0.8	2	1	3					1							
0.6-0.7		1		1		1									
	0.0-0.1			0.1-0.2			0.2-0.3			0.3-0.4			0.4-0.5		

TFP (1 - especificidade)

Tabela 4.5 Número de classificações resultantes pelos algoritmos AP$_{C3}$-FCM, FCM validado, entre $c = 5$ e $c = 7$, pelos índices FS, FCM$_{V_{FS}}$, e XB, FCM$_{V_{XB}}$, em cada intervalo 0.1×0.1 no espaço ROC, para os anos de: (a) 1998; (b) 1999.

deste algoritmo para um reconhecimento automático de regiões de upwelling.

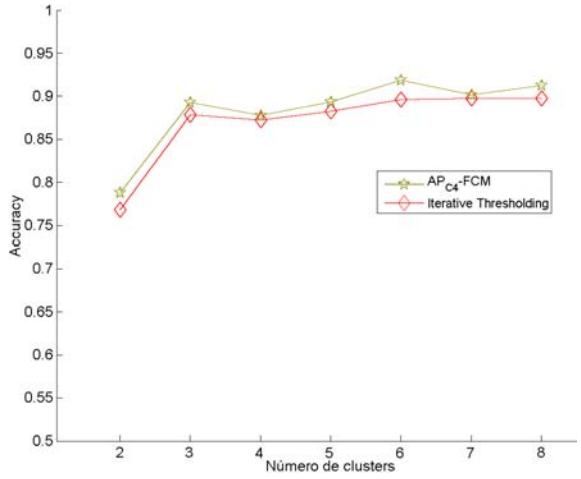
As condições de paragem normalmente definidas para o algoritmo de *Histogram Thresholding* [47] (apresentado na Tabela 2.6, pág. 40), como número máximo de iterações, por exemplo, têm como princípio apenas uma aplicação do algoritmo e não uma iteração sequencial deste, como é feito pela utilização do *Iterative Thresholding*. Como referido anteriormente, a inexistência de um modelo analítico que permita formular o que é uma região de upwelling é um dos factores que dificulta a descoberta de um bom número de *clusters* e, consequentemente, a sua identificação automática.

Pelo facto de que as segmentações resultantes da aplicação do algoritmo AP_{C3}-FCM serem de boa qualidade e obterem automaticamente um bom número de grupos para a identificação da região de upwelling, considera-se esse algoritmo como o melhor para a resolução da problemática desta dissertação. Por esse motivo, nas Secções que tratam de identificar o *cluster* de transição da região de upwelling, as análises são feitas sobre os seus resultados de segmentação.

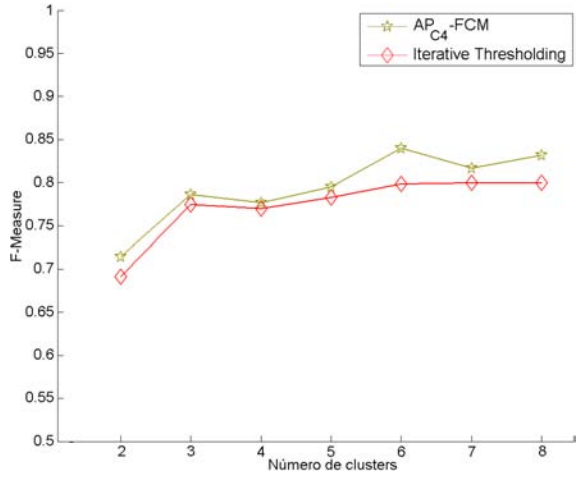
4.7 Comparação computacional entre FCM, AP-FCM e Histogram Thresholding

Com base na análise da secção anterior e tendo em conta o referido facto de que, entre os algoritmos FCM e AP_{C4}-FCM, nenhum resulta, como regra, em melhores segmentações, podem-se comparar os resultados em termos computacionais. Assim, analisando o parâmetro que melhores indicações nos dá, o número de iterações utilizadas por cada algoritmo, verifica-se que quando comparadas as iterações totais de 10 computações do FCM com uma execução do AP_{C4}-FCM (Figura 4.20(a)), este último obtém resultados claramente melhores. Se se comparar apenas a média das iterações ao longo das 10 computações do FCM (Figura 4.20(b)), destacam-se os seguintes factos: (i) a utilização dos protótipos gerados pelo algoritmo Divisão & Conquista na inicialização do FCM, melhora a sua performance, como se verifica pela redução do número de iterações do passo 'AP_{C4}-FCM (FCM)' quando comparado com as do algoritmo FCM original; e (ii) para números de *clusters* reduzidos (entre 2 e 4), o *overhead* das iterações do algoritmo Divisão & Conquista piora a performance total do AP_{C4}-FCM, no entanto, para números de *clusters* superiores, a situação inverte-se, passando o algoritmo a apresentar um número de iterações inferior às do FCM, mesmo comparando unicamente com 1 computação média. Para o intervalo de referência entre $c = 5$ e $c = 7$, uma computação do FCM gasta, em média, mais 7 iterações que o algoritmo AP_{C4}-FCM. Ressalve-se mais uma vez que, na prática, nunca se aplica o FCM apenas com uma computação, pelo que as comparações feitas com essa metodologia servem apenas o propósito deste estudo experimental.

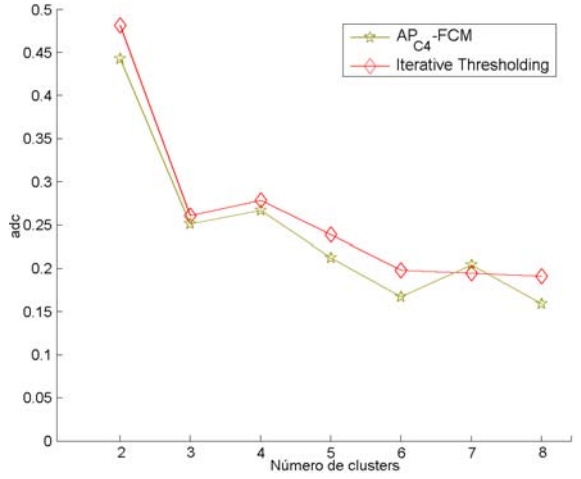
Em termos de tempo gasto na aplicação de cada algoritmo, o gráfico da Figura 4.20(c) demonstra que uma computação do algoritmo AP-FCM demora mais tempo a concluir do que uma computação do FCM. No entanto, se se comparar com o tempo total de 10 computações do FCM, verifica-se que o AP-FCM é de execução mais rápida. Para o número de *clusters* considerados como originários de uma boa segmentação, poderia-se reduzir o número de computações



(a)

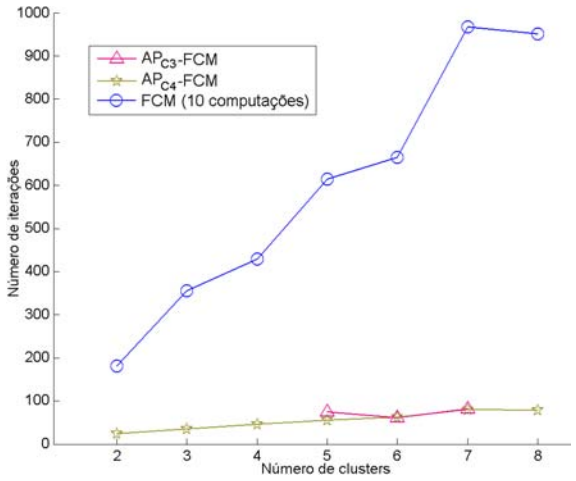


(b)

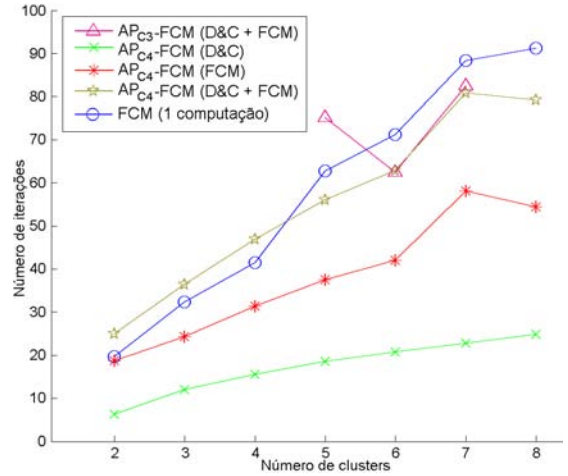


(c)

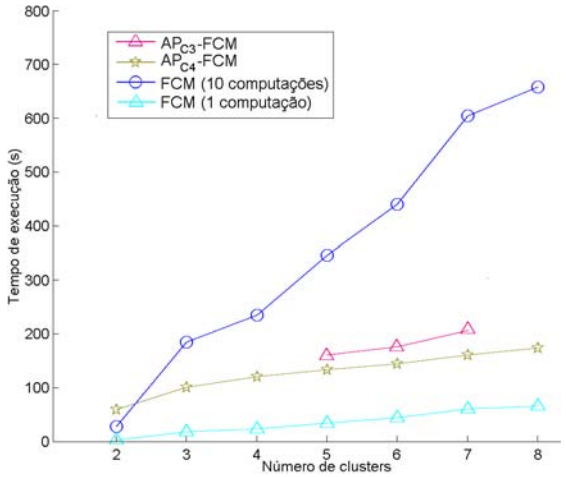
Figura 4.19 Comparação entre as segmentações obtidas pelos algoritmos AP$_{C4}$-FCM e *Iterative Thresholding* com as medidas: (a) *Accuracy*; (b) *F-Measure*; (c) *adc*.



(a)



(b)



(c)

Figura 4.20 Comparação entre as segmentações obtidas pelos algoritmos AP$_{C3}$-FCM, AP$_{C4}$-FCM e *FCM* em termos de: (a) iterações totais; (b) uma computação média do algortimo FCM e iterações totais de AP$_{C3}$-FCM e AP$_{C4}$-FCM (decomposto nos passos d& Conquista e FCM) ; (c) tempo de execução.

do FCM para metade, que a performance do AP-FCM permaneceria superior. Relativamente à comparação em termos de tempo de execução, relembre-se que a performance dos algoritmos é sempre dependente da eficiência das suas implementações.

Como referido anteriormente, a aplicação do algoritmo AP$_{C4}$-FCM não resolve um dos problemas fundamentais que se propõe estudar nesta dissertação, já que continua a estar dependente do parâmetro de entrada que define o número de grupos da segmentação. Verifica-se que o valor médio de iterações totais para o algoritmo que melhor resolve essa questão, AP$_{C3}$-FCM, tem o mesmo comportamento computacional que o algoritmo AP$_{C4}$-FCM, pelo que mantém-se a utilização de menos iterações do que apenas uma computação (média) do FCM (Figura 4.20(b)).

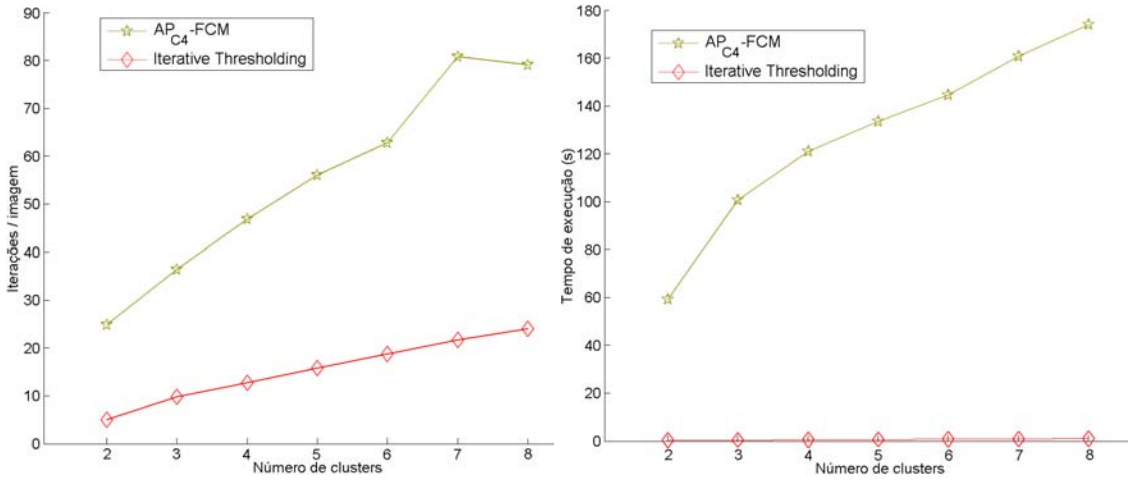
A validação do FCM (de $c_{min} = 2$ a $c_{max} = 10$) do número de *clusters* através de índices de validação não resultou em boas segmentações para a identificação do upwelling, como visto na Secção 4.3.1. Mesmo aplicando a validação apenas a partições com números de *clusters* que resultam em segmentações que permitem identificar a região de upwelling, como por exemplo com a validação de $c_{min} = 5$ a $c_{max} = 7$, na própria aplicação do FCM é necessária executar de c_{min} a c_{max} *clusters*, com r computações por execução, resultado em $(c_{max} - c_{min} + 1) * r$ computações. Diminuindo ao máximo o número de iterações necessárias numa aplicação do algoritmo FCM validado por um índice de validação, temos o cenário em que se varia o número de *clusters* entre $c_{min} = 5$ e $c_{max} = 7$) e apenas se executa uma computação do FCM por cada aplicação. Nesse caso, o número de iterações necessárias, em termos médios, será a soma das iterações utilizadas em $c = 5$ (62.7), $c = 6$ (71.2) e $c = 7$ (88.4), resultando em 222.3 iterações, mais a computação necessária para o cálculo do índice de validação para cada um dos três resultados.

Pela aplicação do algoritmo AP$_{C3}$-FCM conseguem-se obter boas segmentações ($Accuracy = 0.91$, $F - Measure = 0.82$), de qualidade semelhante às obtidas pelo FCM (Figura 4.15) e com utilização, em média, de 72.54 iterações por cada mapa SST, ou seja, cerca de $1/3$ das iterações mínimas por aplicação do FCM com um índice de validação.

Analisando os resultados da Figura 4.21, onde se compara em termos computacionais as aplicações dos algoritmos *Iterative Thresholding* e AP$_{C4}$-FCM, verifica-se que o *Iterative Thresholding* obtém melhores resultados, tanto em termos de número de iterações (Figura 4.21(a)) como em tempo de execução (Figura 4.21(a)). A aplicação do algoritmo *Iterative Thresholding*, quando comparado com o AP$_{C4}$-FCM, acaba por ter como grande vantagem a redução do custo computacional, não resolvendo no entanto a problemática da obtenção automática de um bom número de grupos que conduza a uma segmentação efectiva.

4.8 Estudo do cálculo de gradientes máximos

Com base nos resultados de segmentação obtidos (AP$_{C3}$-FCM), estudou-se também o gradiente dos pontos-fronteira de cada *cluster*. Esta análise está relacionada com a definição de região de upwelling, que indica que, em condições ideais, essa região termina quando há uma diferença brusca de temperatura, ou seja, o gradiente tem um valor mais elevado. O objectivo desta



(a) (b)

Figura 4.21 Comparação em termos computacionais dos algoritmos AP$_{C4}$-FCM e *Iterative Thresholding* em termos de: (a) número de iterações por imagem; (b) tempo de execução.

análise é verificar se existe alguma ligação entre a fronteira da região de upwelling e a fronteira de gradiente máximo.

O gradiente de um campo escalar bidimensional $f(x,y)$ é um vector calculado pelas derivadas parciais de f, em ordem a x e a y:

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right). \tag{4.3}$$

A direcção, sentido e norma do vector gradiente num determinado ponto de f são calculados através de cálculo vectorial simples: $\nabla f(x,y) = \sqrt{\frac{\partial f}{\partial x}(x,y)^2 + \frac{\partial f}{\partial y}(x,y)^2}$. No problema em causa, o campo escalar f representa os mapas de temperatura oceânica disponíveis e os gradientes são calculados nas fronteiras dos *clusters* obtidos.

Assim, no estudo efectuado nesta secção, o cálculo dos gradientes foi feito sobre as fronteiras *crisp* dos *clusters* encontrados. Com base nas informações dadas por oceanógrafos, a identificação da fronteira das regiões de upwelling poderia ser feita com base na procura da fronteira com um gradiente médio mais elevado, ou seja, com a variação mais brusca de temperatura. Contudo, verificou-se que essas fronteiras não correspondem necessariamente à fronteira do cluster de interesse, havendo um conjunto de situações que afecta os gradientes.

A Figura 4.22 apresentam-se as três variações existentes na análise ao gradiente máximo, com a fronteira cujo gradiente tem o maior valor a estar representada pela linha azul. A Figura 4.22(a) contém a situação onde a fronteira de gradiente máximo indica correctamente a fronteira da região de upwelling. Na Figura 4.22(c), a fronteira com gradiente máximo encontra-se na região do Golfo de Cádiz. Ao longo das duas épocas de upwelling disponíveis,

é facilmente perceptível que essa região, tal como toda a região numa latitude inferior ao Cabo de Sagres, tem uma temperatura média bastante superior ao resto do mapa, pelo que a fronteira com a variação mais brusca de temperatura acaba por ser nesta região. Refira-se que não se consegue visualizar uma grande diferença entre os píxeis dos *clusters* separados pela fronteira de máximo gradiente devido ao mapeamento temperatura-cor ter sido ajustado para facilitar uma boa identificação da região de upwelling. Por outro lado, na Figura 4.22(b), tem-se um caso onde a fronteira de gradiente máximo se encontra no interior da região de upwelling. Nesta situação a fronteira indicada representa, por vezes, o que se designa por "upwelling activo", ou seja, as águas que mais recentemente surgiram junto à costa e que ainda não se misturaram com águas que já se encontravam à superfície.

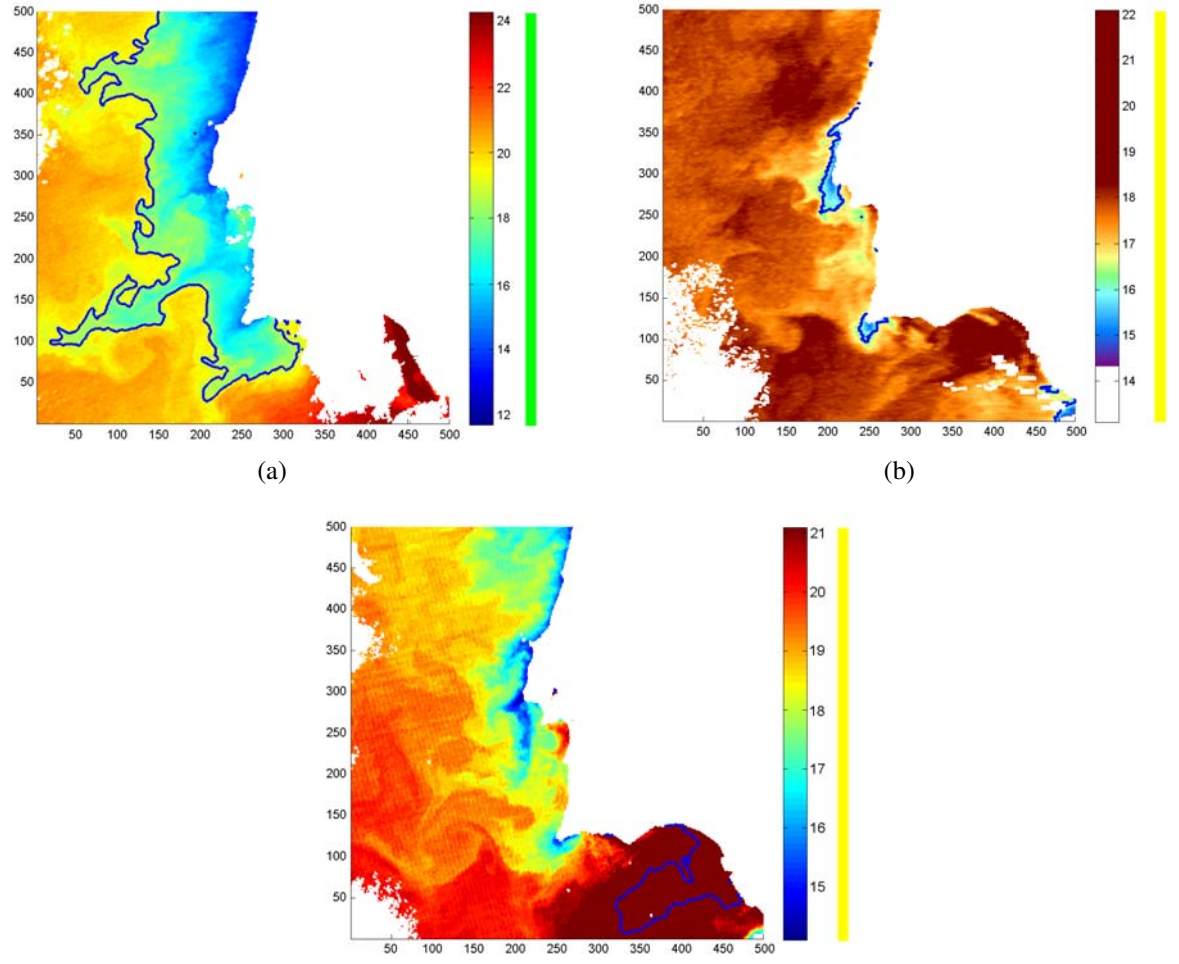
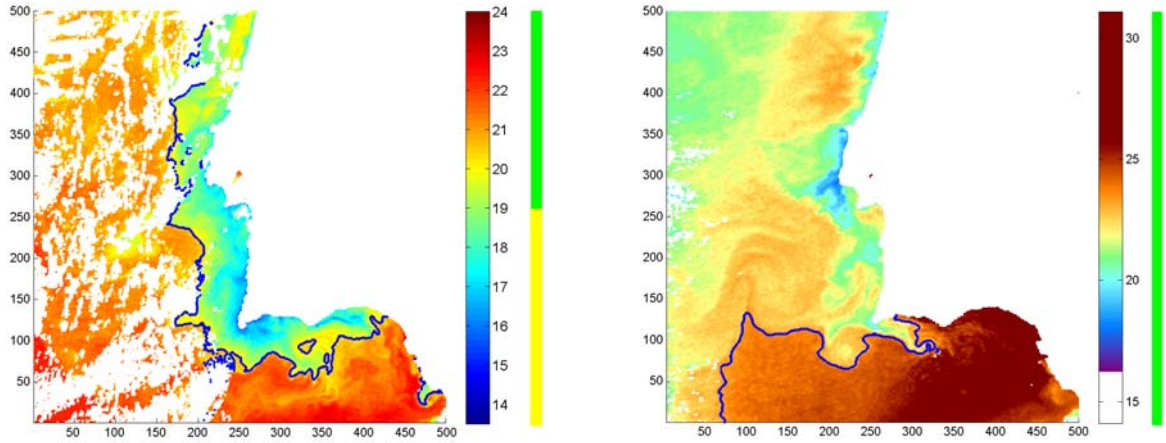


Figura 4.22 Visualização da fronteira de maior gradiente para os mapas de temperatura de: a) 2 de Agosto de 1998; b) 30 de Junho de 1999 e; c) 11 de Julho de 1998.

Dada a situação particular da Figura 4.22(c), optou-se por excluir da análise os *clusters* que ao longo dos dois anos de mapas disponíveis, nunca pertencem à região de upwelling. Nomeadamente, a partir do quarto *cluster*, inclusive, até ao *cluster* de temperatura mais quente. Assim, tendo em conta apenas os três primeiros *clusters* encontrados, verificou-se que no conjunto de 61 mapas SST, a fronteira de gradiente máximo pertence quase sempre a um *cluster* pertencente à região de upwelling. Em 5 imagens, a excessiva presença de nuvens no Norte da imagem afecta a região de upwelling e, consequentemente, o gradiente máximo ultrapassa a fronteira de upwelling definida por oceanógrafos (por exemplo, Figura 4.23(a)). Para essas 5 imagens, na Região Sul a fronteira de gradiente máximo não ultrapassa a fronteira da região de upwelling. A outra situação em que a fronteira de gradiente máximo não está contida na região de upwelling acontece quando o quarto *cluster* se encontra já numa latitude muito a Sul e, por isso, a fronteira do terceiro *cluster* separa as águas mais quentes dessa região (por exemplo, Figura 4.23(b)). Esta situação ocorre em 3 das 61 imagens. Resumidamente, em 53 de 61 imagens a fronteira de máximo gradiente pertence a um *cluster* que faz parte da região de upwelling.



(a) (b)

Figura 4.23 Visualização da fronteira de maior gradiente para os mapas de temperatura de: a) 11 de Setembro de 1998 e; b) 20 de Junho de 1999.

Excluindo os casos problemáticos, verifica-se que em 15 imagens a fronteira de gradiente máximo corresponde à fronteira da região de upwelling, como definida pela anotação textual de oceanógrafos, em cerca de 20 imagens a fronteira definida representa o que se designa "upwelling activo" e nas restantes, o máximo gradiente encontra-se no interior da região de upwelling mas sem definir particularmente, ou em toda a latitude do mapa, o upwelling ou "upwelling activo". Refira-se que, tal como no reconhecimento da região de upwelling, diferentes oceanógrafos podem discordar quanto à identificação da região de "upwelling activo". Pode-se concluir que a análise da fronteira de gradiente máximo não se revelou eficaz para a detecção

do *Border_Cluster*, ou seja, o *cluster* que limita a região de upwelling.

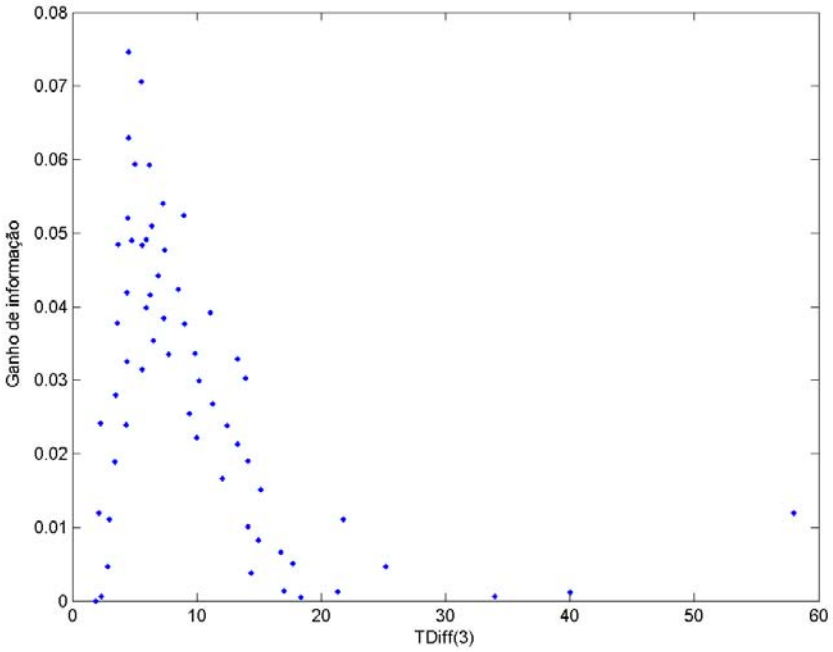
4.9 Detecção e anotação de fronteiras de upwelling

O estudo experimental realizado nesta secção tratou de analisar os resultados do critério composto definido na Secção 3.2.6, tendo como objectivo a obtenção de regiões de upwelling que representem as anotações feitas por oceanógrafos o mais fidedignamente possível. Para esse propósito foi definido, para as regiões Norte e Sul de cada mapa de temperatura, o valor ideal de *Border_Cluster*, ou seja, o *cluster* que melhor define a fronteira da região de upwelling, em cada uma das segmentações obtidas pelo algoritmo AP$_{C3}$-FCM. Esta definição permite também medir quantitativamente a qualidade dos resultados obtidos, comparando os valores de *Border_Cluster* para cada imagem com os valores ideais. A qualidade dos resultados do critério composto foi medida através da percentagem de valores de *Border_Cluster* resultantes iguais aos valores considerados ideais.

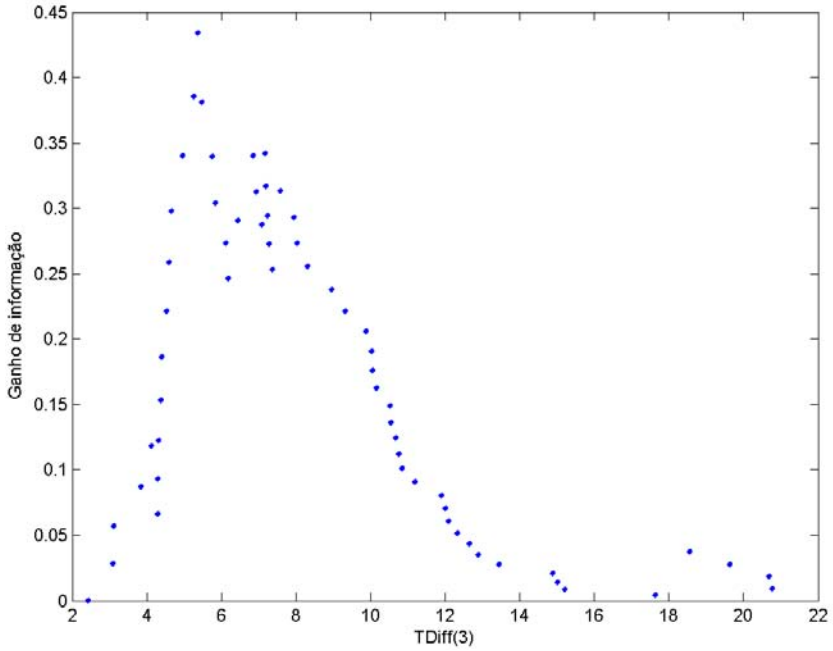
Numa análise empírica aos resultados obtidos, verificou-se que na Região Sul dos mapas de temperatura, a aplicação da análise às *features CCard* e *CloudNoise* não melhorava os resultados obtidos, já que as situações abordadas na definição dessas duas *features* ocorrem somente na Região Norte. Assim, na Região Sul o resultado de *Border_Cluster* é obtido exclusivamente pela aplicação do primeiro passo do critério composto, ou seja, análise da *feature* $TDiff$. O estudo realizado tratou de definir valores utilizados nos únicos parâmetros que o Algoritmo 1 (pág. 53) necessita: os *thresholds* τ_{TN} (passo 2), τ_C (passo 7) e τ_N (passo 10), para a Região Norte, e τ_{TS} (passo 19), para a Região Sul.

O primeiro passo feito neste estudo foi uma análise às segmentações resultantes da aplicação do algoritmo AP$_{C3}$-FCM e, empiricamente, estabelecer os *thresholds* que permitissem atingir bons resultados. Esta análise foi feita apenas sobre os mapas de temperatura do ano de 1998. Alternativamente, foram também calculados *thresholds* com base na análise ao ganho de informação. Como referido na Secção 3.2.6, para cada uma das *features*, o atributo que melhor funciona como *threshold* é aquele que melhor consegue discriminar a pertença, ou não, à região de upwelling, ou seja, aquele que maximiza o ganho de informação, $Gain(a_j)$. Na Figura 4.24 são apresentados os valores de ganho de informação em função do valor da *feature* estudada em cada mapa SST, para o cálculo de *thresholds* para a análise de $TDiff$, na Região Norte e na Região Sul. A *feature* é estudada no terceiro *cluster* uma vez que se considera ser o cluster crítico para a identificação da região de upwelling. Verifica-se o bom comportamento do método para cálculo de *thresholds* pelo facto de que os atributos com os limites superior e inferior de cada *feature* estudada obterem um ganho de informação muito reduzido, já que não conseguem segmentar correctamente a informação de pertença ou não dos *clusters* à região de upwelling, e que quanto mais próximo o valor da *feature* está ao atributo com valor máximo de ganho de informação, maior é o próprio ganho.

No Anexo H são apresentadas os gráficos relativos ao cálculo de *thresholds* para a análise



(a)



(b)

Figura 4.24 Valores do ganho de informação de cada atributo em função do valor da *feature* estudada ($TDiff$) para: (a) Região Norte; (b) Região Sul.

às *features CCard* e *CloudNoise*, verificando-se que o bom comportamento do ganho de informação se mantém. A Tabela 4.6 apresenta os valores dos *thresholds* utilizados na definição do critério fronteira com base na análise empírica e na análise pelo ganho de informação.

A diferença entre a utilização das duas abordagens de detecção de *thresholds* (análise empírica e estudo do ganho de informação) resulta na alteração da fronteira de upwelling na Região Norte em 10 das 61 imagens, sendo que são todas situações com ocorrências de nuvens, pelo que a própria definição da região de upwelling é susceptível de dúvidas. Na Região Sul, não se registam quaisquer diferenças. Refira-se que a diferença, em 12.6%, entre o *threshold* definido por análise ao ganho de informação e *threshold* definido experimentalmente para o passo de análise à extensão cumulativa de *clusters* (estudo da *feature CCard*), apesar de parecer elevada, acaba por alterar apenas 1 dos 61 mapas SST dos conjuntos testados.

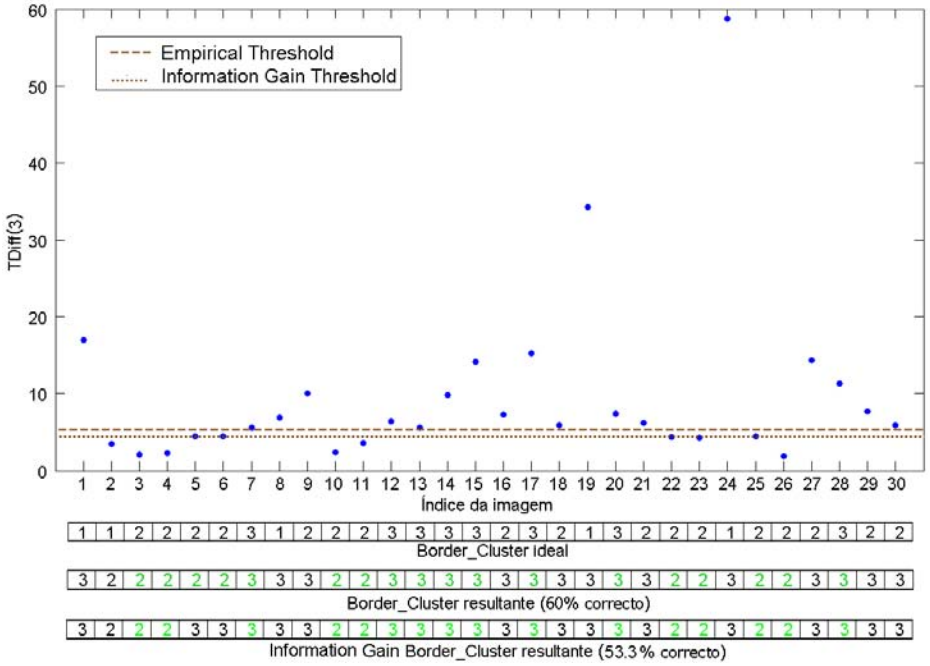
Como referido anteriormente, a análise à feature $TDiff$ é a que fornece a melhor informação quanto à definição do cluster de interesse da região de upwelling. Nas Figuras 4.25 e 4.26 visualiza-se graficamente a análise feita para as duas regiões analisadas, para os mapas dos anos de 1998 (conjunto de treino) e 1999 (conjunto de teste), sendo que para cada imagem se visualiza o valor da *feature* em cada mapa SST. Se a feature $TDiff(3)$ estiver acima do *threshold* utilizado, $Border\_Cluster$ indica o terceiro *cluster* dessa imagem como último *cluster* pertencente à região de upwelling, caso contrário indica o segundo *cluster*. Independentemente do método de obtenção de *thresholds*, os valores de $Border\_Cluster$ resultantes encontram-se a verde se forem iguais ao valor de $Border\_Cluster$ considerado ideal. A taxa de sucesso é calculada através do número de $Border\_Cluster$ correctamente identificados. Verifica-se que na Região Sul (Figure 4.26), esta análise sucede em 90% dos mapas de temperatura, sendo esse um resultado de qualidade alta. Na Região Norte, mais susceptível a situações que afectam a definição da região de upwelling, nomeadamente a presença de ruído sob a forma de extensões nebulosas, a análise à feature $TDiff(3)$ obtém entre 51.6% e 60% de resultados correctos.

	Empírico	Ganho de Informação
τ_{TN}	5.3×10^{-5}	4.47×10^{-5}
τ_{TS}	5.3×10^{-5}	5.39×10^{-5}
τ_C	65%	52,40%
τ_N	1250	532

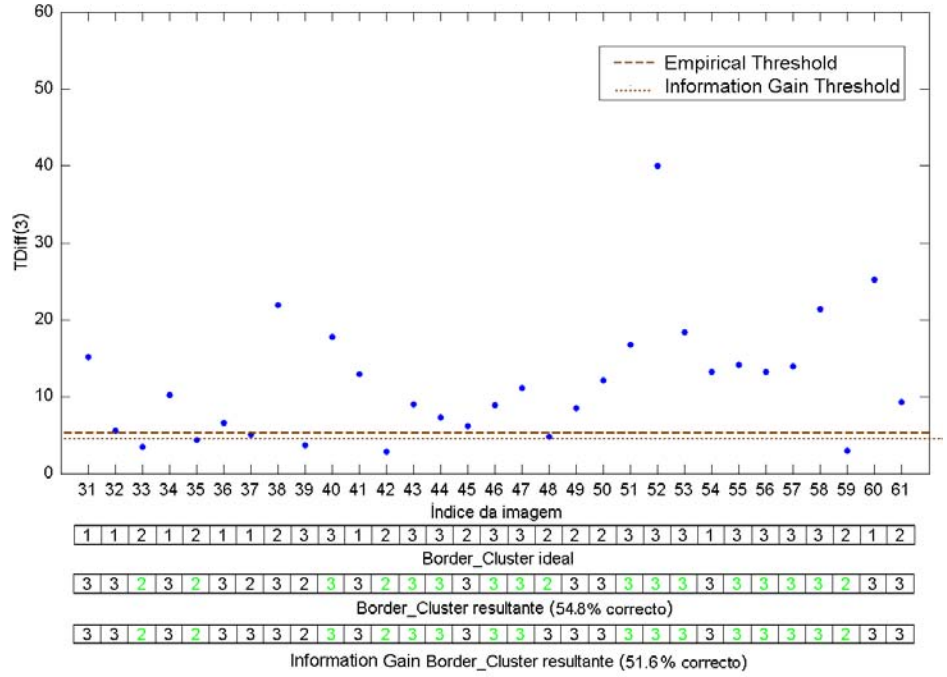
Tabela 4.6 Valores de *thresholds* com base em estudo empírico e estudo por ganho de informação.

Na Figura 4.26 não há distinção entre *threshold* definido empiricamente e por análise do ganho de informação devido ao facto de que ambos possuem praticamente o mesmo valor (ver Tabela 4.6), razão pela qual os resultados da aplicação de ambos os *thresholds* serem iguais.

Na Figura 4.27 apresenta-se a análise feita a extensão cumulativa dos *clusters* indicados como pertencentes à região de upwelling, após a análise da feature $TDiff(3)$. As barras empilhadas de cada mapa de temperaturas indicam a soma cumulativa das percentagens de píxeis da Região Norte ocupados pelos primeiros dois ou três *clusters*, dependendo do valor de $Border\_Cluster$ cada mapa. As sub-barras estão ordenadas pela ordem dos *clusters*, sendo que a primeira sub-barra, correspondente à percentagem de píxeis associados ao primeiro *cluster*,

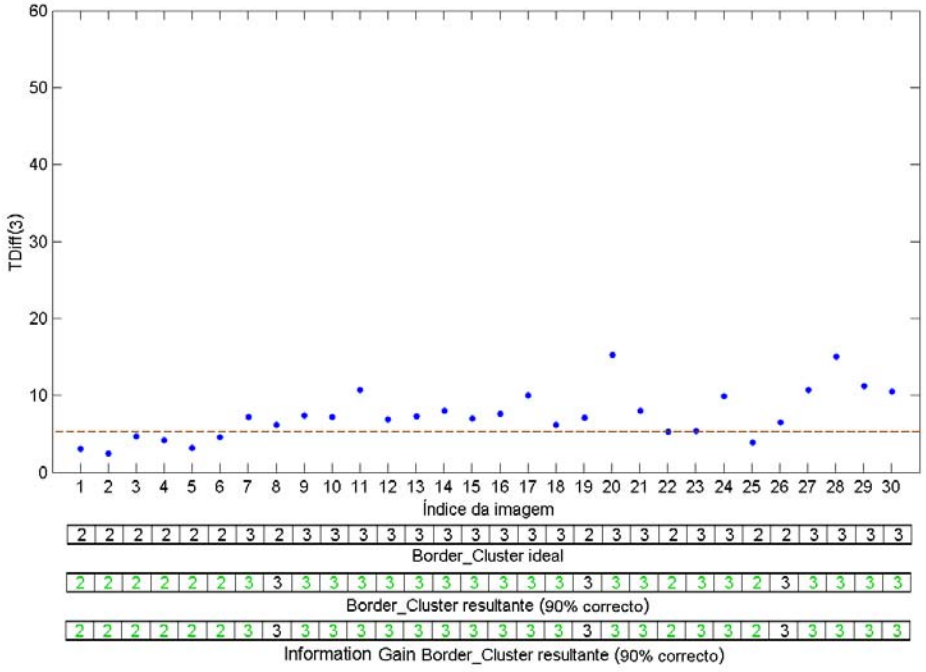


(a)

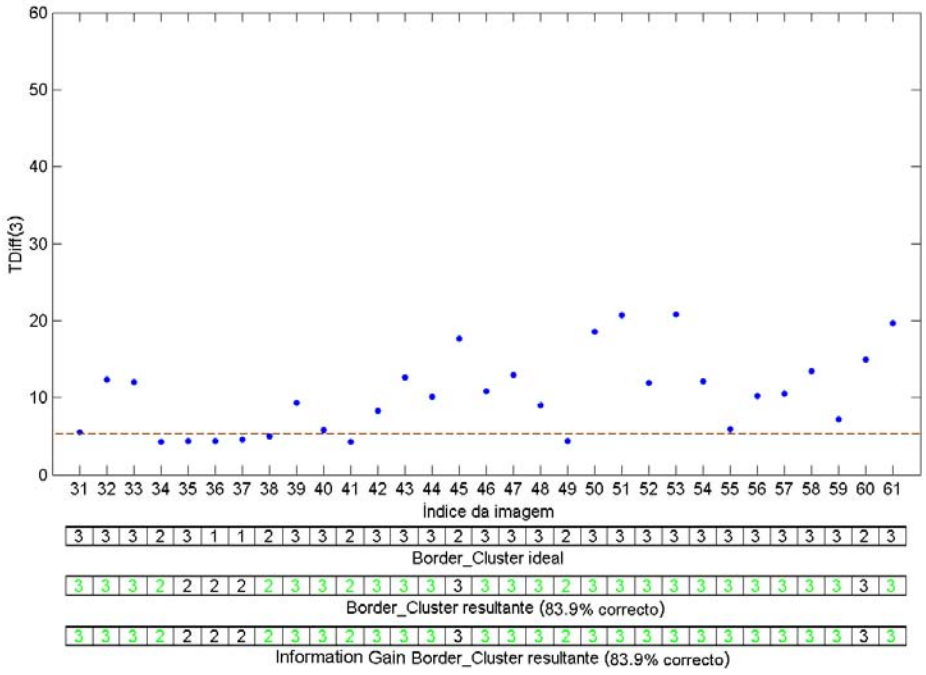


(b)

Figura 4.25 Análise à feature $TDiff(3)$, na Região Norte, para os anos de: (a) 1998; (b) 1999.



(a)



(b)

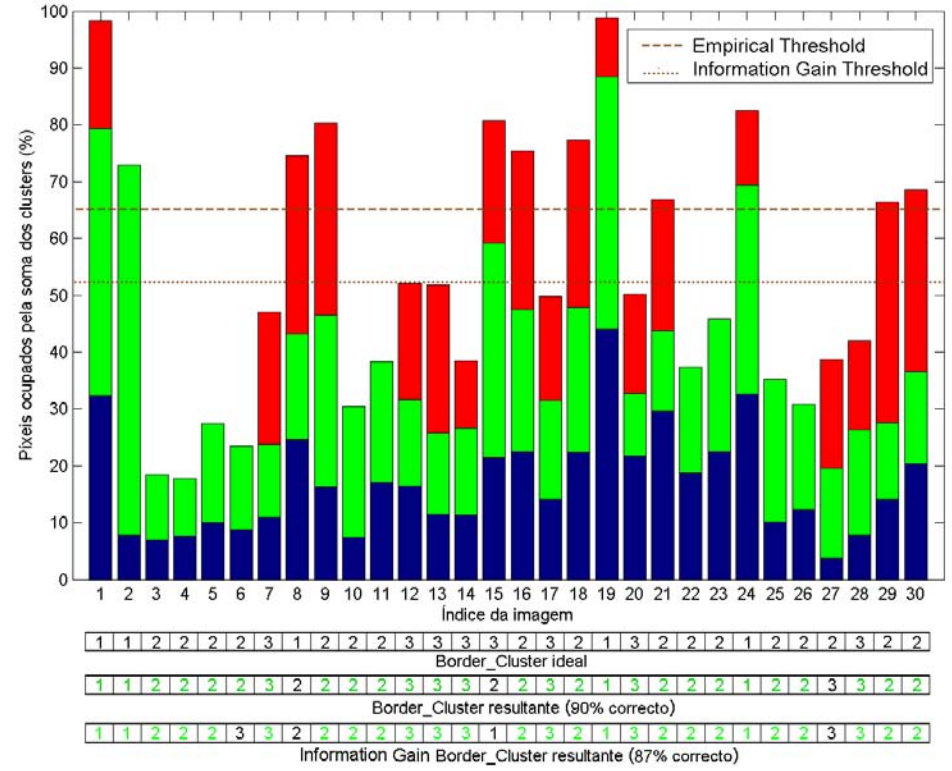
Figura 4.26 Análise à feature $TDiff(3)$, na Região Sul, para os anos de: (a) 1998; (b) 1999.

é a que começa na ordenada 0% e visualizam-se apenas as sub-barras referentes aos primeiros *Border_Cluster clusters*, onde *Border_Cluster* é definido pela análise prévia da *feature TDiff*. Como referido anteriormente, o objectivo desta análise é refinar a detecção do *Border_Cluster*, diminuindo o número de *clusters* que se indicam como pertencentes à região de upwelling, se estes ocuparem uma porção demasiado elevada dos píxeis existentes na imagem. Assim, para as imagens onde as percentagens cumulativas dos primeiros *Border_Cluster clusters* ultrapassa o valor limiar estabelecido, eliminam-se os *clusters* necessários, até que a soma das percentagens não ultrapasse o valor pretendido. Com a aplicação deste segundo passo do critério composto, a qualidade dos resultados aumenta para $87\% - 90\%$, para o ano de 1998, sendo que para o ano de 1999 a subida não é tão acentuada, ficando-se pelos $71\% - 74\%$. Uma das razões que pode explicar a diferença entre a qualidade de resultados para o conjunto treino e para o conjunto teste prende-se com o simples facto de que o ano de 1999 possui uma maior quantidade de mapas SST considerados problemáticos. Informações dadas por oceanógrafos referem inclusivé que o ano de 1999 é considerado pelos próprios como um mau ano para o estudo do fenómeno. Esta situação pode-se verificar logo nas primeiras imagens, até ao dia 27 de Junho de 1999 (mapa com o índice 37), onde a presença de nuvens ou os gradientes térmicos não muito acentuados são factores que afectam uma boa definição de regiões de upwelling. Mesmo na Figura 4.27(b), verifica-se que há um padrão de percentagens elevadas nos mapas até ao índice 37, que não se mantém no resto do ano.

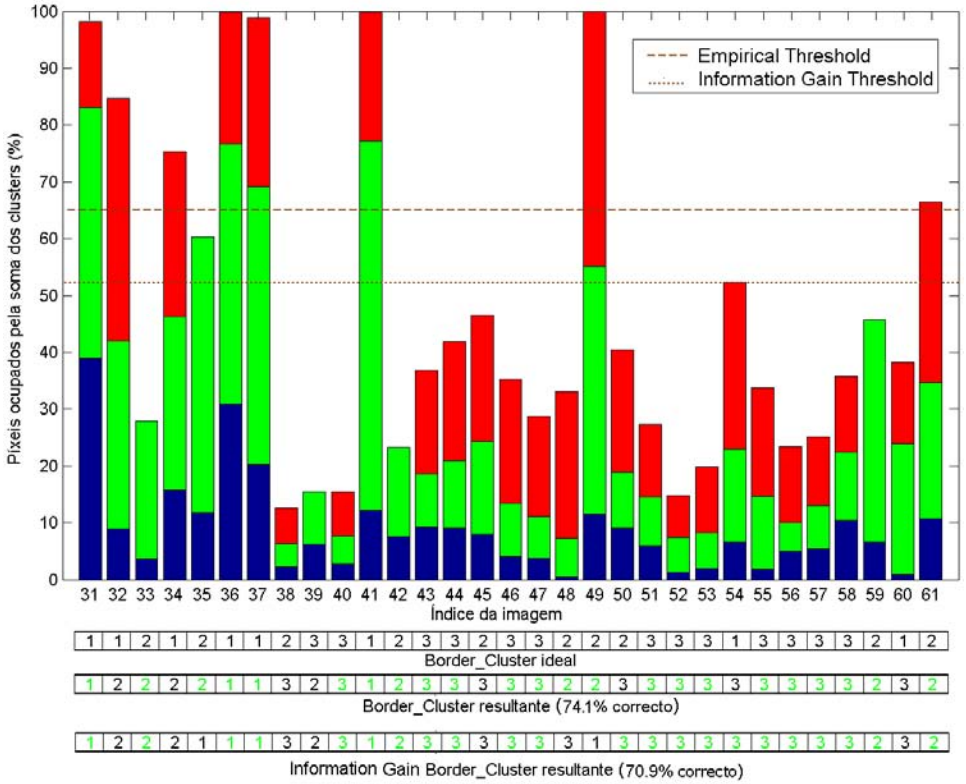
O último passo do critério composto trata de analisar perturbações causadas pela excessiva presença de extensões nebulosas (estudo da *feature CCard*). Na Figura 4.28 está apresentada a análise feita para os dois anos estudados. Para cada mapa de temperaturas estão visualizadas as barras correspondentes aos *clusters* que ainda se consideram como pertencentes à região de upwelling, ordenados decrescentemente pela *label* dos *clusters*. A barra mais à esquerda (vermelha) corresponde ao terceiro *cluster* e a barra mais à direita (azul) ao primeiro *cluster*, ou seja, só aparece a barra do segundo ou terceiro *cluster*, se estes não tiverem sido excluídos na aplicação de um dos passos anteriores do critério composto. Assim, quando se considera que um *cluster* está demasiadamente próximo de uma extensão nebulosa, ou seja, contém muitos píxeis vizinhos a essas extensões e ultrapassa o *threshold* estabelecido, exclui-se esse *cluster* da região que define o upwelling. Com este passo, a qualidade dos resultados obtidos não se alterou positivamente, mas refira-se que o passo anterior, que analisa a extensão cumulativa dos *clusters*, já resolve situações de excesso de nuvens. Por exemplo, o mapa de temperatura de 9 de Agosto de 1998 fica com o resultado correcto na aplicação do passo anterior, embora a região de upwelling seja afectada pela presença excessiva de nuvens.

A qualidade dos resultados também está afectada pela própria incerteza inerente às anotações existentes para cada mapa de temperaturas e utilizadas como *"ground-truth"*. Por exemplo, os próprios oceanógrafos referem ter dúvidas nas anotações feitas quando não há um gradiente elevado nas regiões fronteiriças do upwelling ou quando há ruído causado por extensões nebulosas.

As Tabelas 4.7 e 4.8 apresentam a evolução da qualidade dos resultados após cada passo do critério com posto, com *thresholds* definidos por análise ao ganho de informação. Verifica-se

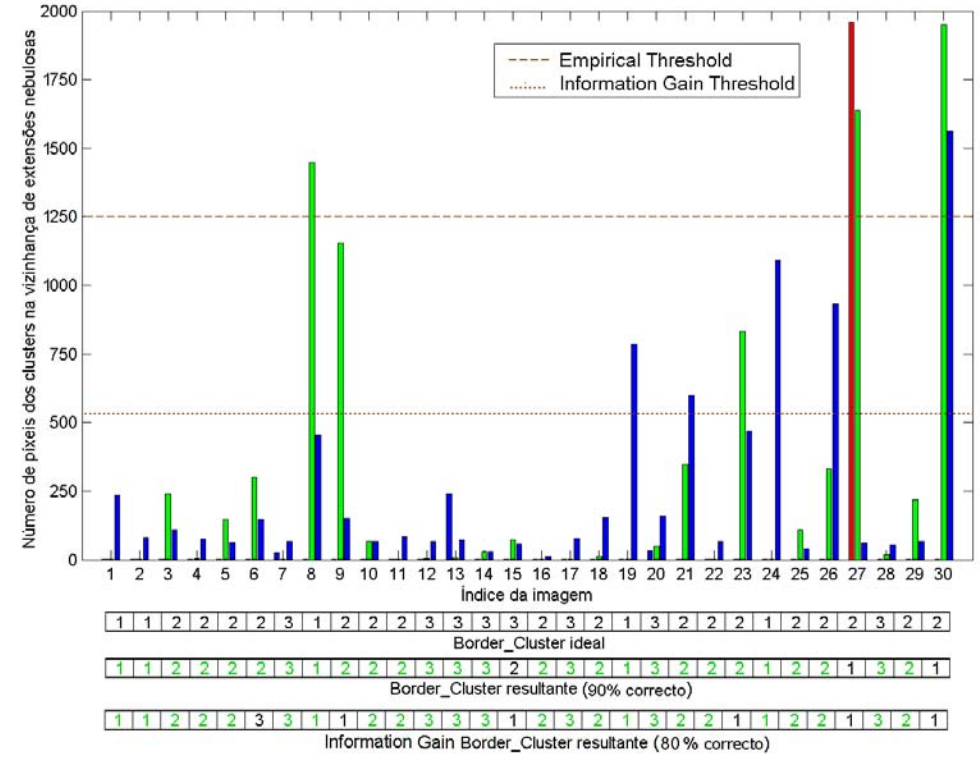


(a)

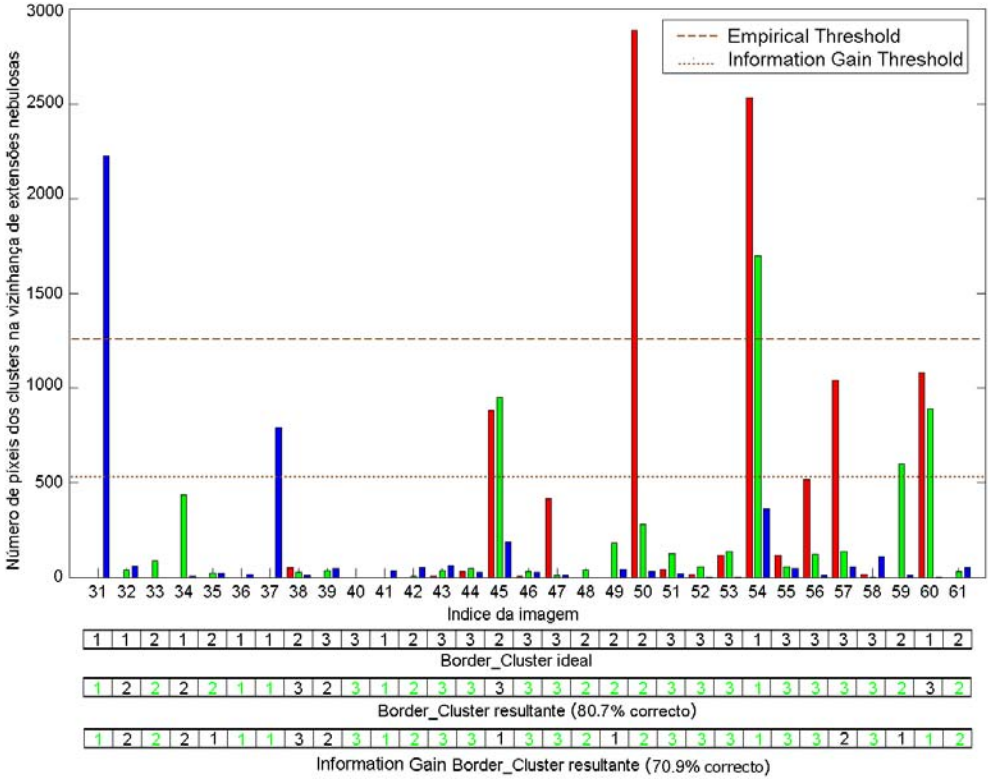


(b)

Figura 4.27 Análise à feature *CCard*, para os anos de: (a) 1998; (b) 1999.



(a)



(b)

Figura 4.28 Análise à feature *CloudNoise*, para os anos de: (a) 1998; (b) 1999.

	# Resultados correctos (Norte)		
	1998	1999	1998+1999
Após análise $TDiff$	16	16	32
Após análise *CCard*	26	22	48
Após análise *CloudNoise*	25	22	47

Tabela 4.7 Evolução do número de resultados certos para os 61 mapas SST relativos aos anos de 1998 e 1999, para a Região Norte, com *thresholds* definidos com base no ganho de informação.

	# Resultados correctos (Sul)		
	1998	1999	1998+1999
Após análise $TDiff$	27	26	53

Tabela 4.8 Número de resultados certos para os 61 mapas SST relativos aos de 1998 e 1999, para a Região Sul, com *threshold* definido com base no ganho de informação

que, apesar de uma ligeira redução, a qualidade dos resultados obtidos é alta para os dois anos de mapas SST utilizados neste estudo. Refira-se ainda que, como referido na Secção 3.2.6, a utilização de *thresholds* definidos automaticamente por análise ao ganho de informação tem a vantagem de poder ser aplicada com maior rigor a outros conjuntos de mapas de temperatura, ao contrário da aplicação de *thresholds* definidos empiricamente por análise aos mapas de temperatura de 1998.

Com base nos resultados obtidos, a aplicação do terceiro passo do critério composto para identificação da região de upwelling, que analisa a influência de extensões nebulosas, não introduz melhorias aos resultados finais, mesmo sendo um factor de importância significativa para os oceanógrafos. Esta situação deve-se à imprevisibilidade das situações onde há maiores extensões nebulosas, sendo que não se conseguiu definir um método linear para modelar a relação entre a presença de nuvens e a região de upwelling. Assim, este terceiro passo deverá ser retirado do critério composto, sem afectar a qualidade de resultados.

Conclui-se que ambos os modos de obtenção de *thresholds* conseguem obter boas taxas de sucesso na detecção do $Border\_Cluster$ ideal e que, a qualidade dos resultados com o conjunto de treino (ano de 1998) desceu cerca de 15% quando aplicado ao conjunto de teste (ano de 1999). No entanto, mesmo para o ano de 1999, a taxa de sucesso manteve-se acima dos 70%, resultado que se pode considerar de boa qualidade. Relembre-se novamente a noção de que o conjunto de dados de 1999 retrata o que oceanógrafos consideram um mau ano de upwelling, dando origem a uma maior percentagem de mapas de temperatura onde há muito ruído devido a extensões nebulosas ou os gradientes térmicos não são muito acentuados, facto que permite ter mais confiança na boa qualidade dos resultados obtidos para esse ano.

A validação de resultados do critério para definição da fronteira da região de upwelling também pode ser feita com as medidas de análise à qualidade de segmentações introduzidas

	Accuracy	Recall	Precision	F-Measure
Thresholds definidos por análise ao ganho de informação)	0,95	0,92	0,88	0,89
Thresholds definidos empiricamente)	0,96	0,93	0,89	0,90

Tabela 4.9 Qualidade das regiões identificadas pelo critério composto, com base em medidas calculadas a partir de matrizes de confusão.

na Secção 4.6, com a diferença de que agora a segmentação que se comparou com os mapas "ground-truth" binários é resultante da aplicação do critério composto, que já é binária por natureza, indicando os píxeis como pertencentes, ou não, à região de upwelling.

Os resultados apresentados na Tabela 4.9 indicam-nos que os resultados obtidos na aplicação dos passos de obtenção de uma segmentação e escolha da fronteira do upwelling, de acordo com o critério definido, são de boa qualidade, principalmente tendo em conta a natureza difusa do fenómeno. Como já referido, a definição da região de upwelling não é algo linear e a classificação de cada píxel com uma classe binária de pertença ou não à região de upwelling exclui completamente essa incerteza, existente maioritariamente nas regiões limítrofes do upwelling. Confirma-se ainda que a diferença entre a qualidade dos resultados obtidos com os dois métodos de obtenção de *thresholds* é desprezável.

No Anexo H, onde apresentados os resultados da aplicação do critério composto para a identificação de regiões de upwelling, pode-se confirmar a qualidade elevada das regiões anotadas, por comparação com a anotação feita por oceanógrafos.

4.10 Estudo da identificação de fronteiras difusas

Numa análise empírica aos resultados obtidos por cada uma das medidas de *fuzziness* introduzidas na Secção 3.3, na aplicação a vários mapas de temperatura, verifica-se que as duas medidas conseguem identificar fronteiras difusas. Contudo, uma análise à qualidade de cada fronteira obtida é dificultada pela não existência de uma definição do que é uma boa fronteira difusa.

Tratando-se de um trabalho exploratório, o estudo feito teve como base apenas um mapa de temperatura, onde o upwelling está bem definido e não há ruído por ocorrência de nuvens, apresentado na Figura 4.29(a), com as fronteiras dos clusters obtidos automaticamente pelo algoritmo AP_{C3}-FCM. A partir do mapa de pertenças difusas resultante (Figura 4.29(b)), aplicou-se a medida *Ignorance Uncertainty* a todos os píxeis do mapa, obtendo-se o mapa de *fuzziness* visualizado na Figura 4.29(c). Neste mapa de *fuzziness* consegue-se comprovar que as regiões de maior *fuzziness* correspondem às fronteiras dos *clusters*, sendo ainda possível verificar-se que a transição do quarto (píxeis amarelos) para o quinto cluster (píxeis cor-de-laranja) é onde há um gradiente de temperatura mais reduzido, levando à identificação de uma fronteira difusa de maior extensão.

Com base na definição de α-cut (Secção 3.3), é intuitivo que a parametrização de α afecta a extensão das fronteiras identificadas, ou seja, quando maior for α, menor vai ser a fronteira difusa, já que há menos píxeis com o valor da medida de *fuzziness* utilizada superior a α.

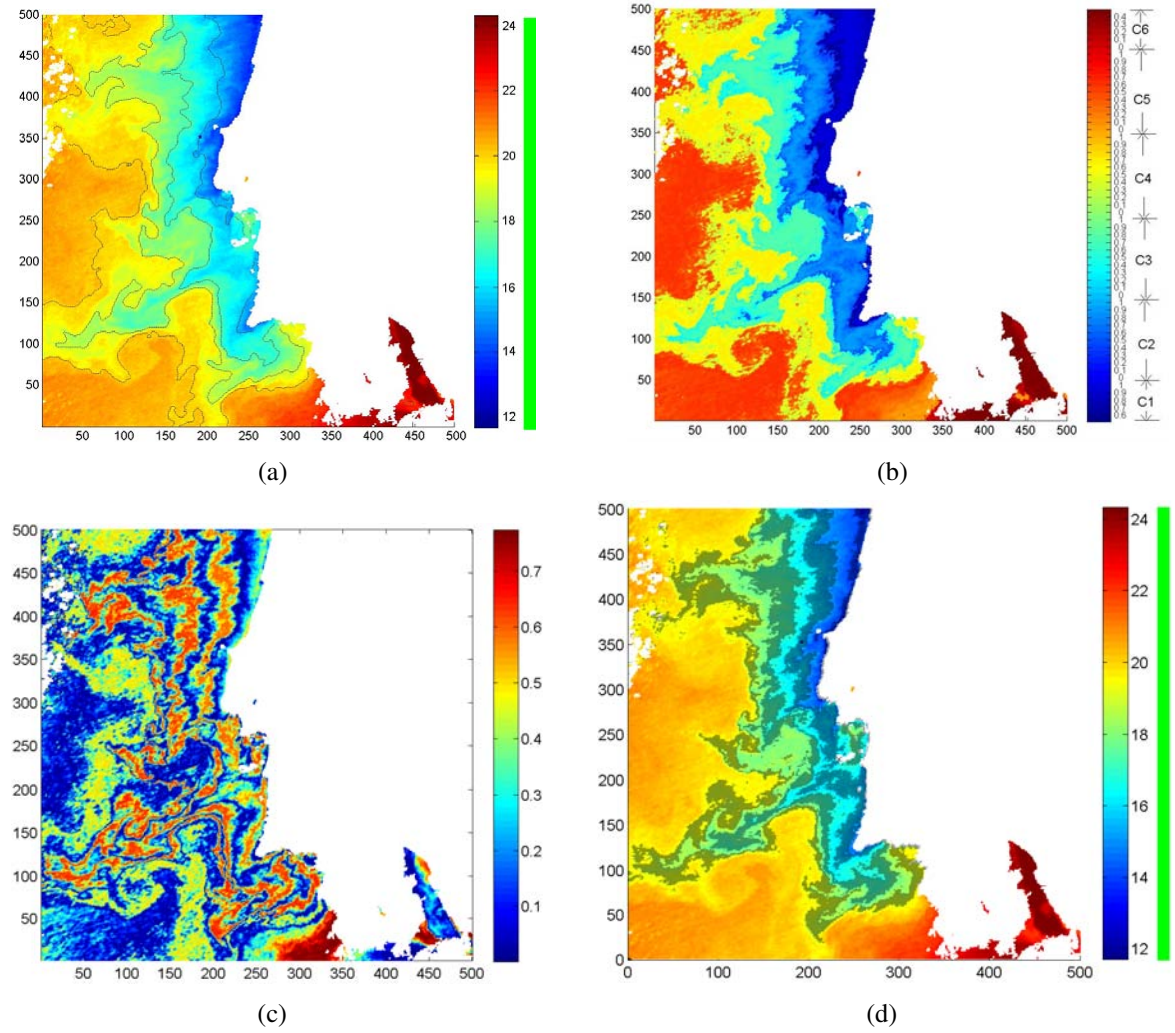
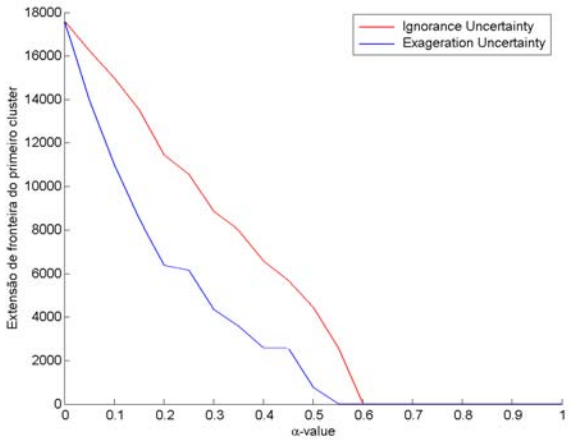


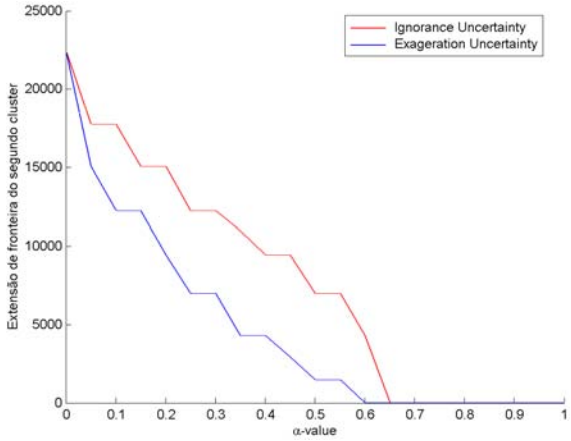
Figura 4.29 (a) Mapa de temperaturas de 02 de Agosto de 1998; (b) Mapa de pertenças difusas resultante da aplicação do algoritmo AP$_{C3}$-FCM; (c) Mapa de medida de *fuzziness Ignorance Uncertainty* para o mapa de temperaturas da alínea (a), obtido com base em segmentação com 6 *clusters* (AP$_{C3}$-FCM); (c) Visualização de fronteiras difusas para os primeiros três *clusters*, com base no mapa de *fuzziness* da alínea (b) e com $\alpha - cut = 0.3$.

O estudo das fronteiras difusas tem maior utilidade para os *clusters* pertencentes ao objecto-de-interesse, ou seja, à região de upwelling, evitando assim "ruído" desnecessário criado pela visualização fronteiras difusas nos *clusters* exteriores. Por essa razão, apenas se estudaram os *clusters* identificados com uma *label* inferior ou igual ao resultado da aplicação do critério composto para identificação do cluster de interesse, ou seja, desde o *cluster* 1 até ao *cluster* $Border\_Cluster$ determinado.

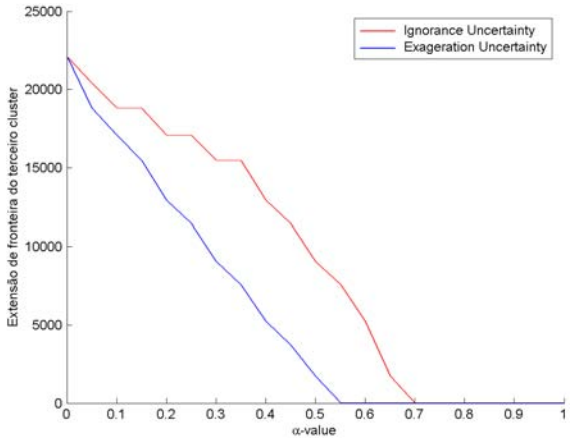
Os gráficos da Figura 4.30 apresentam, para os três *clusters* correspondentes à região de upwelling, o valor da extensão das fronteiras difusas obtidas, ou seja, o número de píxeis da região identificada como sendo fronteira, em função do parâmetro α (calculado em intervalos de 0.05). Verifica-se que o comportamento de ambas as medidas é o esperado, com a redução da extensão dos *clusters* em função do aumento de α, e bastante similar nas três fronteiras estudadas, com a medida *Ignorance Uncertainty* a obter, para os mesmos valores de α, extensões superiores à medida *Exageration Uncertainty*, excluindo os intervalos em que as extensões atingem o valor 0.



(a)



(b)



(c)

Figura 4.30 Efeito da parametrização do valor de α na extensão da fronteira difusa identifcada para: (a) o primeiro cluster; (b) o segundo cluster; e, (c) o terceiro cluster.

4.11 Análise de imagens SST sem upwelling

O conjunto de 61 mapas de temperatura, relativos aos anos de 1998 e 1999, não conteve nenhum mapa onde o fenómeno do upwelling não estivesse presente. No entanto, foi possível fazer um estudo em dois mapas que representam situações sem upwelling, com uma ligeira variação da região geográfica representada mas continuando a representar a costa Oeste da Península Ibérica.

Na região em estudo, sem a ocorrência de upwelling, o gradiente térmico presente no mapa de temperaturas deixa de ter uma direcção perpendicular à costa e passa a ser paralelo, ficando com os píxeis/*clusters* de temperatura inferior na Região Norte e com os píxeis/*clusters* de temperatura superior na Região Sul. Na Figura 4.31 visualiza-se um mapa de temperaturas sem upwelling e as fronteiras resultantes da sua segmentação por aplicação do algoritmo AP_{C3}-FCM (*threshold* = 1×10^{-3}). Desde logo, verifica-se que a disposição dos *clusters* encontrados se altera, comparativamente a situações onde o upwelling está presente, deixando de ter o *cluster* de temperatura média inferior a surgir junto à costa e os *clusters* seguintes a "sobreporem-se" na direcção do oceano, para passar a ter o *cluster* mais frio a Norte e o mais quente a Sul.

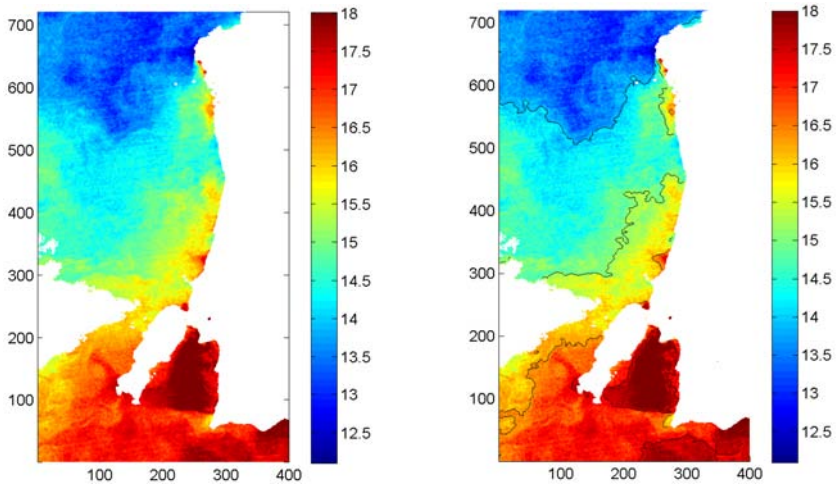


Figura 4.31 (a) Mapa de temperaturas de 22 de Abril de 2002; (c) Segmentação obtida pelo algoritmo AP_{C3}-FCM, com *threshold* = 1×10^{-3}, resultando em 5 *clusters*.

No estudo feito para detectar a ocorrência, ou não, do upwelling em mapas de temperatura, analisou-se a amplitude térmica (diferença entre temperatura máxima e temperatura mínima) média das linhas de uma mapa de temperaturas. Sendo a a altura de um mapa de temperaturas, ou seja, o número de linhas de uma matriz X de temperaturas, e $X(r,:)$ a linha r da matriz X, a medida (Amplitude Térmica Média sobre Linhas) estudada é formulada da seguinte forma:

$$ATML(X) = \frac{\sum_{r=1}^{a} max(X(r,:)) - min(X(r,:))}{a}. \tag{4.4}$$

Intuitivamente, se o upwelling estiver presente, haverá uma maior amplitude do que nos casos em que não ocorre, já que o gradiente de temperaturas tende a variar na horizontal. Com base no gráfico da Figura 4.32, onse se representa o valor da medida $ATML$ para os anos de 1998 e 1999, juntamente com os dois mapas de temperatura sem ocorrência de upwelling, verifica-se que os mapas sem upwelling possuem uma amplitude térmica média menor que todos os 61 mapas de 1998 e 1999. O intervalo que separa os mapas com e sem upwelling é relativamente reduzido (0.08 ºC), mas note-se que o mapa com menor $ATML$ para o conjunto de 61 mapas, relativo ao dia 02 de Junho de 1999, representa um caso onde há muito ruído causado por excesso de nuvens e que, excluindo esse mapa, o intervalo de segurança entre a ocorrência, ou não, de upwelling aumenta para 0.65 ºC.

Refira-se que o estudo feito nesta secção tem um cariz fundamental na criação de um método robusto para detecção automática de regiões de upwelling ao longo de vários anos, sendo que, a aplicação de um algoritmo de segmentação e detecção do cluster de interesse, só é útil caso se confirme a existência do fenómeno.

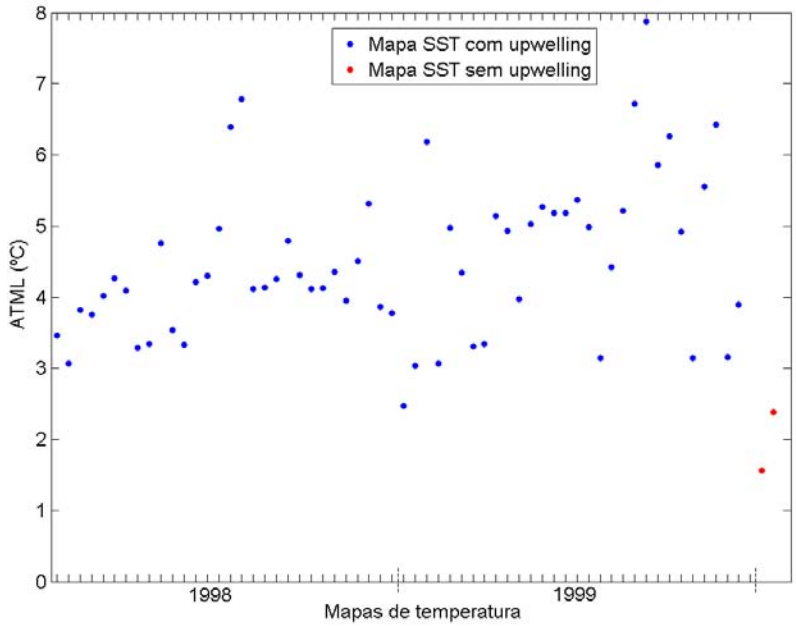


Figura 4.32 Valores médios da amplitude térmica das linhas dos mapas de temperaturas do conjunto de mapas relativos ao ano de 1998 e 1999, e de dois mapas de temperatura sem upwelling.

4.12 Sumário

No estudo experimental elaborado neste capítulo, demonstrou-se que o algoritmo AP-FCM obtém segmentações de qualidade semelhante às do algoritmo FCM, com duas grandes vantagens: uma maior eficiência computacional, já que elimina a necessidade de múltiplas computações do algoritmo FCM, e a utilização da condição de paragem que estuda a contribuição para a dispersão total de dados (AP-C3) e que possibilita a detecção automática de um número de *clusters* que gera boas segmentações. O AP$_{C3}$-FCM evita resultados sob ou sobre-segmentados, que são os resultados obtidos com a validação do FCM com os índices de validação estudados, e serve ainda como indicador para um bom número de *clusters*, possibilitando a limitação do intervalo de validação para o FCM.

A aplicação do algoritmo de *Iterative Thresholding*, quando comparado com o algoritmo AP-FCM, verificou-se resultar em segmentações de ligeiramente pior qualidade mas, apesar da grande vantagem em termos computacionais, não foi definido um método que possibilitasse a detecção automática de um bom número de *clusters*, pelo que o algoritmo $AP_{C3} - FCM$ é o que melhor performance obtém na aplicação ao problema da segmentação de mapas de temperatura. Num estudo à qualidade das segmentações, por comparação com mapas "ground-truth", criados a partir da anotação feita por oceanógrafos, verificou-se que, num espaço ROC, cerca de 40% das classificações se encontram próximas do ponto de classificação perfeita, e as restantes encontram-se, em média, bem mais perto desse ponto do que da linha de classificação aleatória. Foram também utilizadas outras medidas de qualidade, que confirmaram a boa qualidade das segmentações obtidas, nomeadamente *Accuracy* (0.9) e F-Measure (0.82). Relembre-se que ambas as medidas variam no intervalo $[0,1]$, tendo 1 como o valor ideal.

O *threshold* para o algoritmo $AP_{C3} - FCM$, definido por análise empírica dos resultados para o ano de 1998, ficou estabelecido em 1×10^{-3} e resultou em partições com um número de grupos entre 5 e 7, inclusivé, para os 61 mapas estudados. As segmentações geradas com um número de *clusters* nesse intervalo verificaram-se ser de boa qualidade e, quando aplicado o mesmo *threshold* ao ano de 1999, o intervalo de número de grupos resultantes manteve-se igual.

Na análise feita sobre o critério composto para a identificação de regiões de upwelling, destaca-se a elevada qualidade das segmentações obtidas. Analisando os resultados, sem a aplicação do terceiro passo do critério (análise à *feature CloudNoise*), que se verificou não ser útil para melhorar os resultados obtidos, para o conjunto de 61 mapas de temperatura, o critério sucede em 53 (86.9%), para a Região Sul e em 48 (78.7%), com *thresholds* definidos por análise automática ao ganho de informação.

5 . Conclusão e Trabalho Futuro

Em termos de objectivos atingidos, destaca-se principalmente a qualidade dos resultados obtidos na obtenção automática de segmentações e na identificação dos *clusters* correspondentes às regiões de upwelling. O algoritmo *Anomalous Pattern* verificou-se ser uma muito boa inicialização para o FCM, nomeadamente devido à detecção automática do número de clusters e no cálculo determinístico dos protótipos iniciais, através do algoritmo $AP_{C3} - FCM$, ou seja, aplicação do *Anomalous Pattern* com o estudo da contribuição para a dispersão de dados como condição de paragem. A utilização do $AP_{C3} - FCM$ elimina de uma só vez a necessidade de executar o FCM várias vezes, para evitar que fique preso em mínimos locais devido a uma má inicialização de protótipos, e de fazer uma aplicação de c_{min} a c_{max} *clusters*, utilizando índices de validação para detectar um bom número de clusters, conseguindo assim ter uma melhor performance em termos computacionais (número de ciclos e tempo de execução do algoritmo). Refira-se ainda que nenhum dos índices de validação estudados para o FCM conseguiu obter bons resultados para a detecção das regiões de upwelling, originando, tipicamente, resultados sub ou sobre-segmentados. A parametrização obtida do algoritmo $AP_{C3} - FCM$ permite gerar segmentações que retratam de forma fidedigna as estruturas das regiões de upwelling presentes nos mapas SST originais, eliminando assim a necessidade de haver um oceanógrafo que ajuste manualmente, para todos os mapas de temperatura, as escalas de cores para evidenciar as regiões de upwelling. Destaque-se ainda a metodologia usada para a definição do *threshold* aplicado ao algoritmo $AP_{C3} - FCM$, com a utilização de um conjunto de treino, relativo ao ano de 1998, a estabelecer valores para o número de *clusters* entre 5 e 7, e a validação destes resultados com a aplicação ao conjunto de teste, relativo a 1999, onde o intervalo para o número de *clusters* resultante não se alterou.

A partir de mapas "ground-truth" construídos com base nas anotações textuais feitas por oceanógrafos, a qualidade elevada dos resultados obtidos, tanto ao nível de segmentação como de fronteira da região de upwelling, pôde ser comprovada através de medidas calculadas com base numa matriz de confusão (análise no espaço ROC, *Accuracy* e *F-measure*) ou na comparação entre protótipos das segmentações e dos mapas "ground-truth" (*Average Distance between Centroids*).

Destacam-se também os bons resultados obtidos com a aplicação do critério composto para identificação automática do cluster de interesse, criado com o objectivo de modelar num algoritmo o conhecimento do domínio de aplicação de oceanógrafos. O módulo de definição de *features* ($TDiff$, $CCard$ e $CloudNoise$) permitiu a construção do critério composto, utilizando informação específica do domínio. A utilização do critério composto para identificação do cluster de interesse permite a criação de um método sistemático para identificação das frentes de upwelling, eliminando a subjectividade das inspecções visuais feitas por oceanógrafos. É de ressalvar que a *feature CloudNoise* não melhorou a qualidade dos resultados, pelo que o critério composto a aplicar não deverá fazer o seu estudo. A taxa de sucesso na identificação do cluster de interesse para os 61 mapas de temperatura ficou em cerca de 79% para a Região Norte

e 87% para a Região Sul.

A utilização da obtenção de *thresholds*, do critério composto para identificação do *cluster* de interesse da região de upwelling, com base na análise ao ganho de informação das *features* de cada mapa de temperatura, apresenta-se como uma grande vantagem, já que possibilita que, quando se utilizarem novos conjuntos de dados, se apliquem *thresholds* com base na informação exclusiva a esses conjuntos. Também neste caso, se aplicou a metodologia com um conjunto de treino (1998) e conjunto de teste (1999), e verificou-se que a qualidade alta dos resultados se manteve.

Outra contribuição desta dissertação é a visualização de fronteiras dos clusters obtidos numa segmentação sobre os mapas de temperatura oceânica. Em [1], tinha-se a introduzido a visualização dos mapas de pertenças difusas, mas tendo em conta a natureza do fenómeno e o facto de se encontrar ligado à temperatura das águas, a visualização de fronteiras sobre as próprias temperaturas oceânicas permite retirar mais informação, em termos oceanográficos, sobre a região de upwelling. Num estudo exploratório, introduziram-se ainda medidas de *fuzziness* para a identificação, e visualização, de fronteiras difusas. Devido à natureza do fenómeno, as fronteiras difusas representam de uma forma mais fidedigna as fronteiras que separam as regiões de upwelling do restante mapa, já que duas massas de água dificilmente se podem separar com uma linha com um pixel de largura.

Tendo em conta a boa qualidade dos resultados obtidos, tanto em termos de segmentação como na definição da fronteira da região de upwelling, futuramente deverá ser feito um estudo relativamente às regiões encontradas. Esse estudo poderá ter um âmbito alargado e, definindo novas *features* a partir dos resultados obtidos, pode-se estudar várias características físicas do fenómeno, como temperaturas que ocorrem tipicamente na região de upwelling (média, máximo, mínimo, amplitude térmica, por exemplo) e morfologia (análise sobre os limites físicos da região de upwelling, possivelmente tendo em conta a região do ano). Com os resultados obtidos nesse estudo, poder-se-à introduzir uma nova condição de paragem para o algoritmo *Anomalous Pattern*, que tenha em conta precisamente informação relacionada com o fenómeno ou que utilize dados dos clusters extraídos em cada iteração, com relevância em termos físicos (variação de temperatura ou localização geográfica, por exemplo). Seguindo o mesmo princípio, se os resultados desse estudo forem bons, poder-se-à aplicar uma condição de paragem ao algoritmo *Iterative Thresholding* que possibilite a obtenção automática de segmentações que permitam uma boa definição do upwelling. A elevada eficiência computacional do algoritmo é um factor a ter em conta.

No âmbito desta dissertação, o trabalho desenvolvido foi integrado numa *interface* interactiva, de acordo com a arquitectura proposta, FuzzyUpwell (Figura 3.8, pág. 58), com o objectivo de estruturar os métodos e algoritmos estudados. Este sistema possibilita a introdução de um novo mapa de temperatura e, por exemplo, aplicar a detecção automática da região de upwelling, bem como determinar se o mapa de temperaturas tem ou não presente o padrão de upwelling. A médio prazo, o objectivo será fazer um estudo espaço-temporal às regiões de upwelling, com recurso a um maior conjunto de dados, englobando várias épocas de upwelling.

A . Estudo experimental comparativo entre AP-FCM e FCM

Nesta secção estudou-se o comportamento dos algoritmos FCM e AP-FCM na sua aplicação a alguns conjuntos de dados mais utilizados na literatura. Um dos objectivos do estudo feito é estudar a validação do algoritmo FCM e o efeito das condições de paragem do algoritmo *Anomalous Pattern* na detecção automática de um bom número de grupos para os conjuntos de dados testados. Os resultados do algoritmo FCM foram validados pelos índices apresentados na Secção 2.4.3, nomeadamente, os índices XB (Equação 2.18), FS (Equação 2.21) e PBMF (Equação 2.22). Paralelamente, podem ser comparados os algoritmos relativamente a dois aspectos computacionais: número de ciclos/iterações e tempo de execução de cada um dos algoritmos. Note-se que o tempo de execução é dependente da eficiência da implementação, pelo que o melhor indicador para comparação entre a eficiência de ambos os algoritmos é o número de iterações, que se pretende que seja o mais reduzido possível.

Foram utilizados os seguintes conjuntos (disponíveis no repositório de *Machine Learning* da Universidade da Califórnia - Irvine [52]):

- Iris Data Set: Conjunto de dados que contém três classes, correspondentes a três tipos de lírios, com 50 amostras para cada classe, sendo que cada amostra contém quatro atributos. Uma das classes é linearmente separada das restantes, que não são linearmente separáveis entre si. O número ideal de *clusters* é 2.

- Wisconsin Breast Cancer Database (WBCD): Contém 699 amostras, com 10 atributos, sobre casos diagnosticados com cancro, existindo duas classes: cancros benignos e malignos. O conjunto de dados foi disponibilizado pelo Dr. William H. Wolberg da University of Wisconsin Hospitals. O número ideal de *clusters* é 2.

- Wine Data Set: Contém 178 amostras de três tipos de vinho, onde cada amostra contém 13 atributos relativos a análises químicas ao vinho. O número ideal de *clusters* é 3.

Refira-se que os valores obtidos com os índices de validação XB e PBMF foram validados com os resultados dos conjuntos Iris e Cancer, para partições com um número de *clusters* reduzido, onde há menos espaço para variação, de [31] (em execuções com o parâmetro fuzzificador $m = 1.5$). Os resultados obtidos por aplicação do índice FS foram validados com os mesmos conjuntos, com os resultados fornecidos em [53].

No estudo feito, para contornar o problema da dependência dos protótipos iniciais referido na Secção 2.3.1, em cada aplicação do algoritmo FCM, foram computadas 10 execuções e os índices de validação foram calculados com base na melhor execução, relativamente à função objectivo J_m. Seguindo a prática mais comum na literatura de referência, para o parâmetro m foi utilizado o valor 2 para ambos os algoritmos [21].

Relativamente à detecção automática do número de grupos para cada conjunto, a partir da validação do algoritmo FCM por aplicação de índices de validação, nas Tabelas A.1, A.2, A.3 o valor a *bold* indica o melhor número de *clusters* para o respectivo índice. Destaca-se que o

índice de Xie-Beni foi o que apresentou melhores resultados, falhando apenas para o conjunto Wine, onde indica 2 como sendo o número ideal de *clusters*. O valor de $c_{max} = 12$ foi estabelecido tendo em conta o número de entidades dos conjuntos e o número de grupos considerado ideal. O cálculo dos índices de validação para partições com um número de *clusters* muito elevado é desprezável tendo em conta que, para os três conjuntos, o número ideal de grupos é sempre inferior a 4.

As partições obtidas pelo algoritmo AP$_{C1}$-FCM resultam todas em resultados com excesso de *clusters*, quando comparados com os valores ideais (Tabela A.4). As condições de paragem AP-C2 e AP-C3, relativas à análise da contribuição para a dispersão de dados dos *clusters* extraídos pelo *Anomalous Pattern* , devem ser analisadas consoante o problema em causa e são mais úteis em situações onde há uma maior amostra de dados disponíveis. Para os conjuntos de dados em causa, facilmente se faz uma análise experimental e se descobrem os valores para os *thresholds* a utilizar em cada uma das condições de paragem para se obter o número considerado ideal. Contudo, seriam necessárias mais amostras significativas de cada um dos conjuntos de dados, de modo a validar o estudo feito com essas condições de paragem. Já para o problema da detecção de regiões de upwelling, há uma grande amostra de dados e tomando-se cada mapa de temperatura como um conjunto de dados, foi possível validar o estudo feito com a análise à dispersão de dados, ou seja, estabeleceu-se um *threshold* com base na análise a um determinado conjunto de mapas de temperatura (ano de 1998) e, posteriormente, aplicou-se o mesmo *threshold* aos mapas de temperatura de 1999 e verificou-se que o comportamento do algoritmo permaneceu o mesmo, validando os resultados. Por motivos de completude, refira-se que os seguintes *thresholds* permitem atingir um número ideal de *clusters* são os seguintes: conjunto Iris (2 *clusters*), 79.34% (AP-C2) e 47.03% (AP-C3); conjunto Wine (3 *clusters*), 49.24% (AP-C2) e 18.69% (AP-C3); conjunto WBCD (2 *clusters*), 42.64% (AP-C2) e 41.17% (AP-C3).

FCM - # *clusters*	XB	FS	PBMF
2	**0,05**	-401,80	28,39
3	0,14	-450,49	43,90
4	0,20	-475,98	45,37
5	0,23	**-544,92**	46,30
6	0,31	-389,10	51,42
7	0,37	-395,71	51,18
8	0,43	-335,53	50,00
9	0,32	-384,01	45,77
10	0,32	-330,52	48,39
11	0,55	-294,44	51,75
12	0,49	-291,32	**51,89**

Tabela A.1 Validação do algoritmo FCM para o conjunto de dados Iris, com os índices Xie-Beni (XB), Fukuyama-Sugeno (FS) e Pakhira *et al.* (PBMF).

Nos resultados obtidos destaca-se a vantagem que a aplicação do algoritmo AP-FCM representa relativamente ao eliminar da susceptibilidade do FCM aos protótipos iniciais. Comparando os valores dos índices de validação nos resultados obtidos na aplicação do FCM (Tabelas A.1, A.2, A.3), com a versão do AP-FCM que também recebe o número de *clusters* como condição de paragem, AP$_{C4}$-FCM (Tabelas A.5, A.6, A.7), verifica-se que as partições obtidas

FCM - # *clusters*	XB	FS	PBMF
2	**0,06**	-10285531,82	508273,62
3	0,13	-11461000,17	931951,65
4	0,09	-14389328,75	1363968,06
5	0,10	-13398692,43	1396354,04
6	0,11	-16455277,84	2127941,74
7	0,08	-16109988,15	2574077,81
8	0,26	-11302023,99	1626030,83
9	0,11	**-18154235,38**	2768389,73
10	0,53	-11065413,84	1546676,75
11	0,10	-16690077,63	**2950285,24**
12	0,20	-13222239,51	2800912,62

Tabela A.2 Validação do algoritmo FCM para o conjunto de dados Wine, com os índices Xie-Beni (XB), Fukuyama-Sugeno (FS) e Pakhira *et al.* (PBMF).

FCM - # *clusters*	XB	FS	PBMF
2	**0,11**	-13508,07	252,01
3	1,21	-25728,44	249,97
4	8,78	**-31024,80**	209,56
5	169980,68	-16542,22	**268,38**
6	325076,38 $\times 10^7$	-20136,56	248,93
7	112452,86 $\times 10^7$	-22573,82	224,39
8	17514,95 $\times 10^7$	-14041,79	267,94
9	187059,80 $\times 10^7$	-15818,86	246,31
10	18126,59 $\times 10^6$	-17205,57	237,32
11	2544293,08 $\times 10^8$	-12494,62	256,92
12	28892,64 $\times 10^6$	-13257,09	251,11

Tabela A.3 Validação do algoritmo FCM para o conjunto de dados WBCD, com os índices Xie-Beni (XB), Fukuyama-Sugeno (FS) e Pakhira *et al.* (PBMF).

Conj. Dados	# *clusters* (AP_{C1}-FCM)	XB	FS	PBMF	Tempo Execução (s)	Iter. D&C	Iter. FCM	Total iter.
Iris	5	0,23	-544,92	46,30	0,13	22	35	57
Wine	5	0,10	-13400061,72	1396293,20	0,29	36	63	99
WBCD	9	1,80 $\times 10^1$1	-15818,86	2463,72	3,30	52	423	475

Tabela A.4 Parâmetros de análise do algoritmo AP-FCM para os três conjuntos de dados, com a condição de paragem AP-C1 (todas as entidades associadas a um cluster extraído pelo *Anomalous Pattern*).

AP_{C4}-FCM - # c	XB	FS	PBMF
2	**0,05**	-401,80	28,39
3	0,17	-450,49	43,90
4	0,20	-475,97	45,37
5	0,23	**-544,93**	**46,30**

Tabela A.5 Valores dos índices de validação Xie-Beni (XB), Fukuyama-Sugeno (FS) e Pakhira *et al.* (PBMF), aplicados ao algoritmo AP_{C4}-FCM para o conjunto de dados Iris, com o número de *clusters* como condição de paragem.

AP_{C4}-FCM - # c	XB	FS	PBMF
2	**0,06**	-10285531,75	508273,62
3	0,13	-11460976,98	931951,05
4	0,09	**-14389329,17**	1363968,05
5	0,10	-13400059,28	**1396293,31**

Tabela A.6 Valores dos índices de validação Xie-Beni (XB), Fukuyama-Sugeno (FS) e Pakhira *et al.* (PBMF), aplicados ao algoritmo AP_{C4}-FCM para o conjunto de dados Wine, com o número de *clusters* como condição de paragem.

AP_{C4}-FCM - # c	XB	FS	PBMF
2	**0,11**	-13510,21	252,02
3	1,21	-25728,45	249,97
4	7,09	**-31024,05**	210,93
5	27161,1	-16542,03	**268,52**
6	$2{,}54 \times 10^{8}$	-20136,55	248,93
7	$2{,}35 \times 10^{10}$	-22573,82	224,67
8	$1{,}92 \times 10^{16}$	-14041,79	267,94
9	$2{,}17 \times 10^{11}$	-15818,87	246,39

Tabela A.7 Valores dos índices de validação, Xie-Beni (XB), Fukuyama-Sugeno (FS) e Pakhira *et al.* (PBMF), aplicados ao algoritmo AP_{C4}-FCM para o conjunto de dados WBCD, com o número de *clusters* como condição de paragem.

são muito semelhantes, particularmente quando o número de *clusters* é reduzido (2, 3 ou 4), sendo que o algoritmo AP_{C4}-FCM não necessita de múltiplas computações. Note-se que os resultados que se apresentam para o algoritmo AP_{C4}-FCM estão limitados ao número máximo de *clusters* que o *Anomalous Pattern* consegue extrair, ou seja, o mesmo número de *clusters* que o algoritmo apresenta para a condição de paragem AP-C1.

Por outro lado, pode-se comparar os dois algoritmos (FCM e AP-FCM) em termos computacionais. Uma vez que um dos propósitos do algoritmo AP-FCM é eliminar a geração aleatória dos protótipos iniciais, e consequente necessidade de múltiplas computações para evitar maus resultados devido a más inicializações, o principal factor a ter em conta é o número de iterações que cada um dos algoritmos utiliza. Nas Figuras A.1, A.2, A.3, podem-se comparar os dois algoritmos em termos de: (a) número total de iterações, ou seja, as iterações de 10 computações do FCM e de uma execução do AP_{C4}-FCM; (b) uma computação média do FCM e uma execução do AP_{C4}-FCM (decomposto nos dois passos que o compõem: *Anomalous Pattern*, seguido de FCM); e (c) comparação do tempo de execução de 10 computações do FCM e uma execução do AP_{C4}-FCM. Os resultados visualizados nas alíneas a) e b) das referidas Figuras, indicam-nos que o número de iterações gastas pelo AP_{C4}-FCM está na mesma gama de valores do que uma computação média do FCM, sendo que para 10 computações do FCM, o AP_{C4}-FCM necessita de menos iterações, na ordem de dez vezes menos.

Assim, verifica-se que para obter uma partição praticamente igual, o algoritmo AP-FCM precisa de menos iterações. Podendo-se contestar que 10 computações do FCM são em demasia para os conjuntos de dados estudados, não deixa de ser verdade que tendo em conta a ordem

de diferença entre iterações totais dos dois algoritmos, o AP-FCM consegue efectivamente eliminar o factor aleatório da geração de protótipos iniciais a que o FCM está sujeito.

A comparação apresentada nas Figuras A.1(c), A.2(c), A.3(c) em termos de tempo gasto por cada um dos algoritmos não difere da análise ao número de iterações, já que verifica-se que o AP-FCM necessita de menos tempo que 10 computações do FCM, sendo, no entanto, de ressalvar que esta medida está inerentemente ligada à máquina onde o algoritmo é executado e à própria implementação do algoritmo.

Resumindo, a partir da análise feita no estudo experimental elaborado com conjuntos de dados de referência, conclui-se que o AP-FCM consegue efectivamente eliminar a necessidade de múltiplas computações do FCM e a sua dependência dos protótipos iniciais, conseguindo obter deterministicamente partições muito semelhantes às obtidas pelo FCM. A validação do FCM através de índices de validação revelou que o índice que melhores resultados consegue na detecção de um bom número de grupos é o índice de Xie-Beni. Por seu lado, verificou-se que a condição de paragem AP-C1 resulta sempre num número excessivo de *clusters*, enquanto as condições AP-C2 e AP-C3, após um estudo dos *thresholds* utilizados, permitem encontrar o número correcto de grupos, não sendo, no entanto, o método mais exacto para o fazer, devido à inexistência de novos conjuntos que permitam validar os *thresholds* estabelecidos.

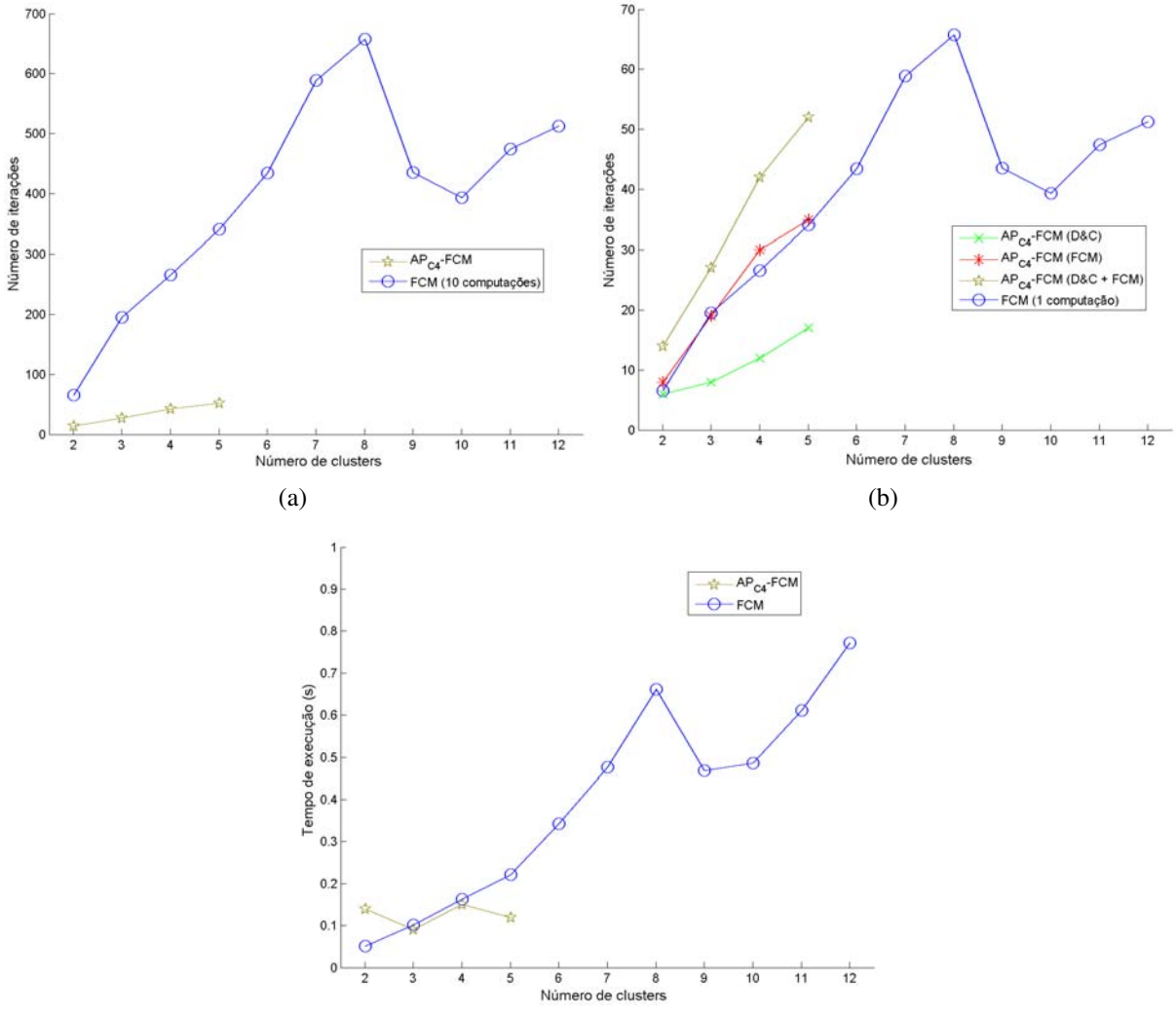


Figura A.1 Para o conjunto Iris, comparação entre: (a) número de iterações utilizadas pelos algoritmos AP$_{C4}$-FCM e FCM, com 10 computações; (b) número de iterações utilizadas pelo algoritmo AP$_{C4}$-FCM e uma computação média do FCM. ; (c) tempo de execução dos algoritmos AP$_{C4}$-FCM e FCM, com 10 computações.

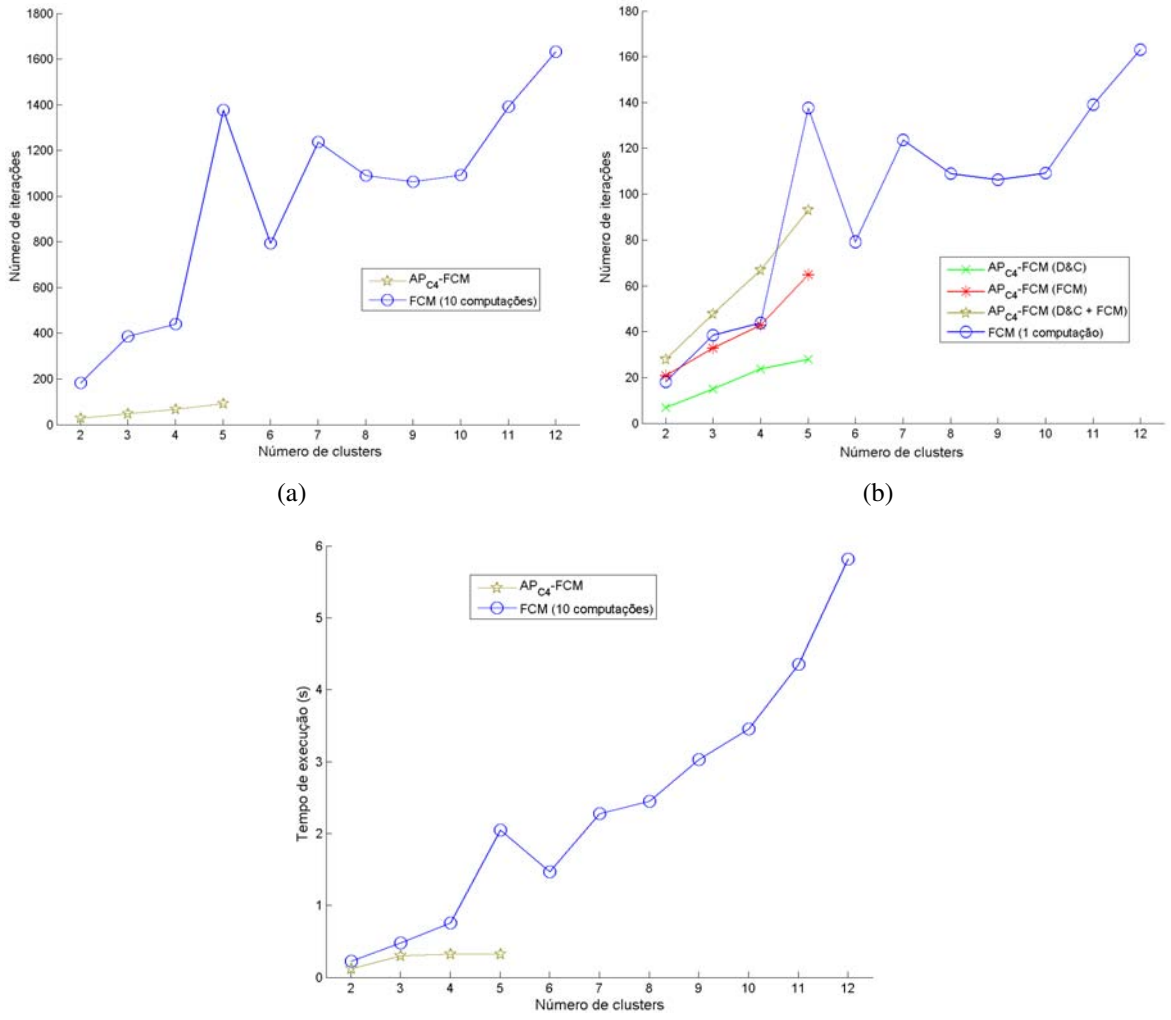


Figura A.2 Para o conjunto Wine, comparação entre: (a) número de iterações utilizadas pelos algoritmos AP$_{C4}$-FCM e FCM, com 10 computações; (b) número de iterações utilizadas pelo algoritmo AP$_{C4}$-FCM e uma computação média do FCM. ; (c) tempo de execução dos algoritmos AP$_{C4}$-FCM e FCM, com 10 computações.

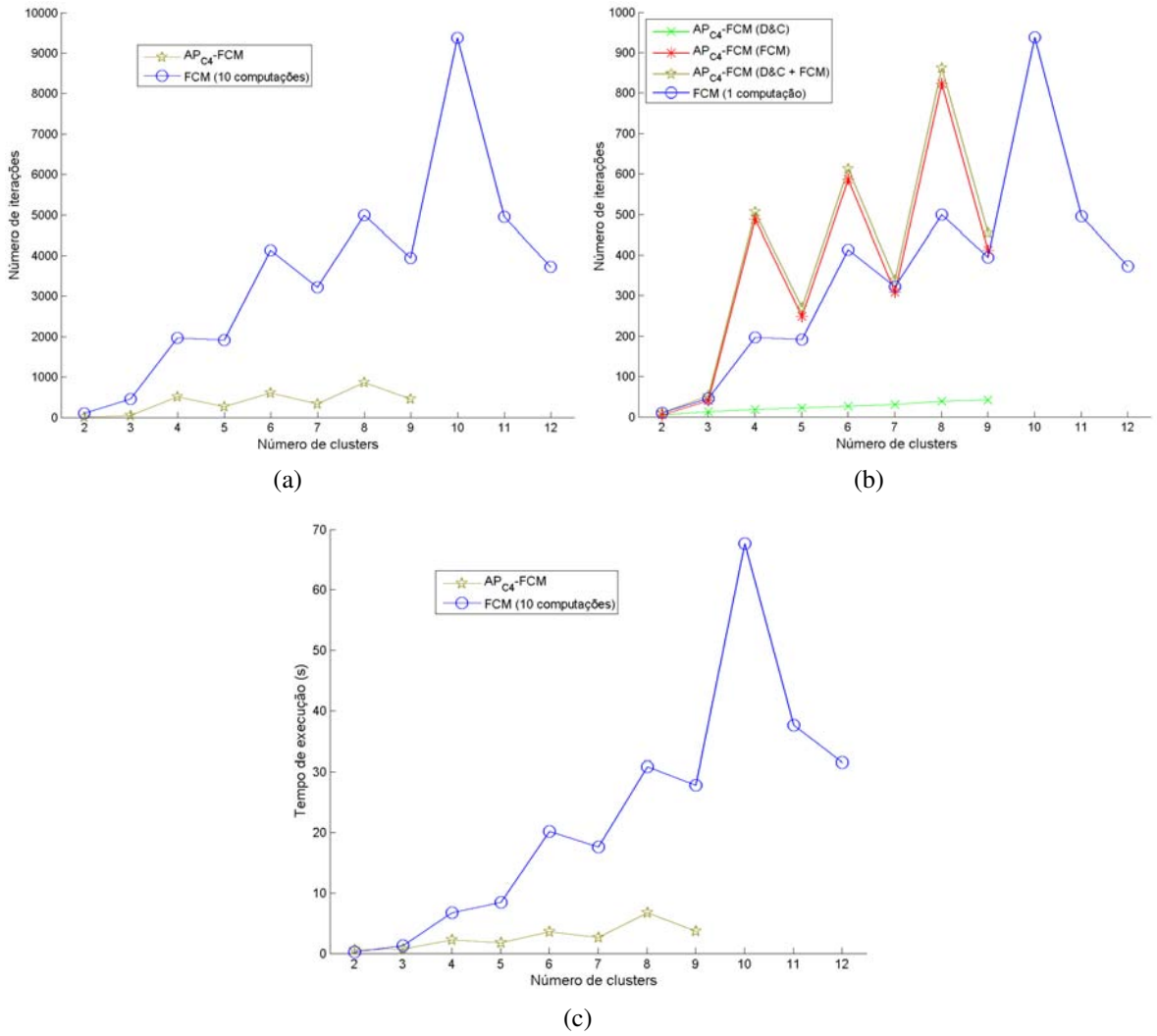


Figura A.3 Para o conjunto WBCD, comparação entre: (a) número de iterações utilizadas pelos algoritmos AP$_{C4}$-FCM e FCM, com 10 computações; (b) número de iterações utilizadas pelo algoritmo AP$_{C4}$-FCM e uma computação média do FCM. ; (c) tempo de execução dos algoritmos AP$_{C4}$-FCM e FCM, com 10 computações.

B . Mapas de temperatura SST [1]

B.1 Ano 1998

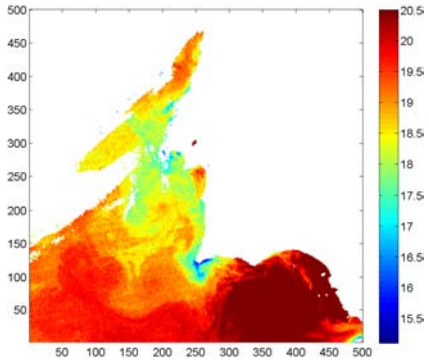


Figura B.1 19980609

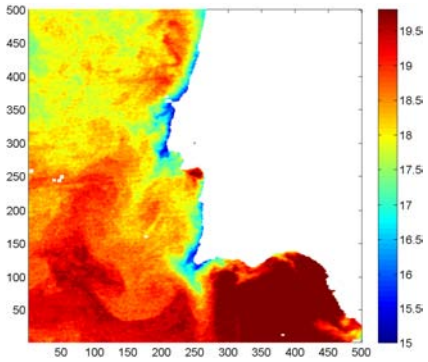


Figura B.2 19980612

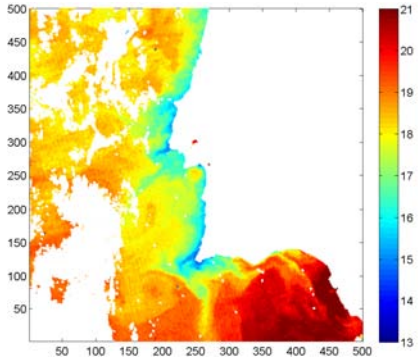


Figura B.3 19980614

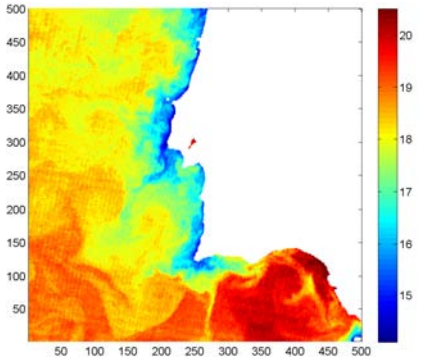


Figura B.4 19980618

[1] A identificação de cada mapa de temperatura está sob o formato 'yyyymmdd', onde 'yyyy' indica o ano, 'mm' o mês e 'dd' o dia correspondentes. Em anexos futuros, o mesmo formato é utilizado, bem como 'yyyymmdd_xc', indicando uma segmentação com 'x' *clusters* para o respectivo mapa. A barra com cores ao lado de cada mapa de temperatura representa a anotação textual feita por oceanógrafos.

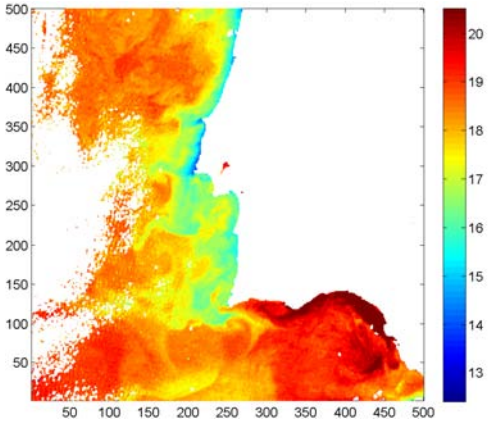


Figura B.5 19980623

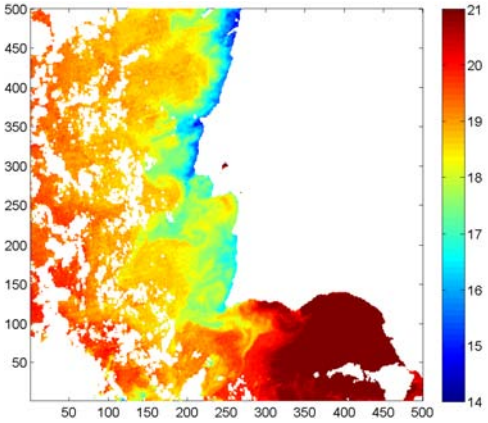


Figura B.6 19980625

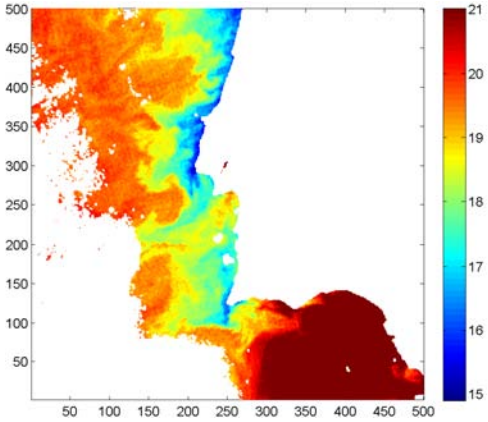


Figura B.7 19980628

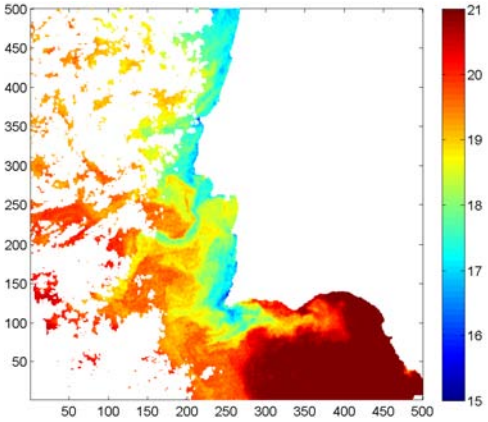


Figura B.8 19980703

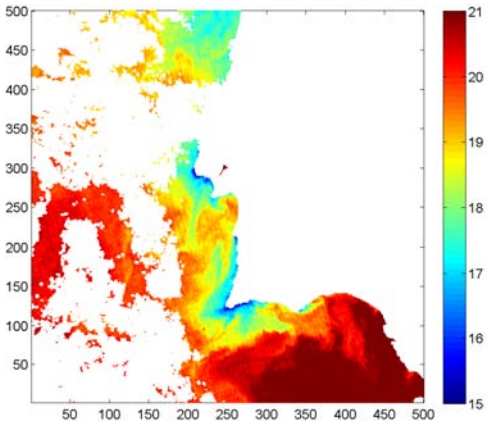


Figura B.9 19980707

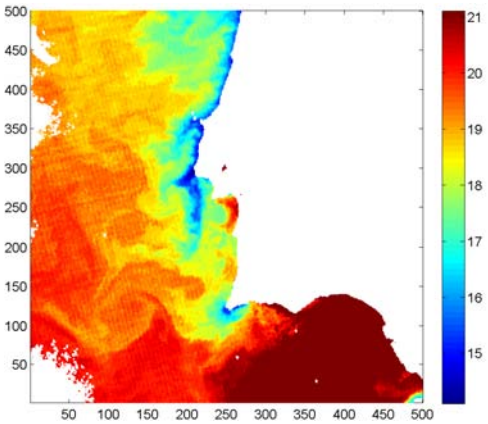


Figura B.10 19980711

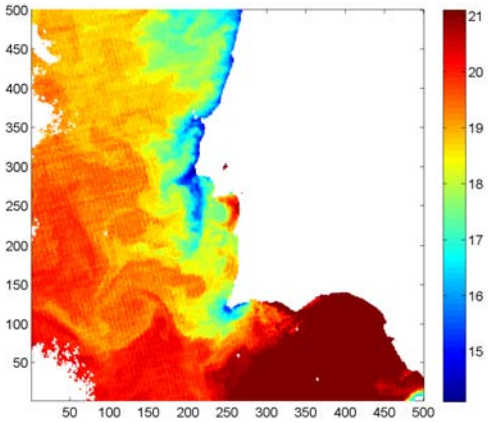


Figura B.11 19980715

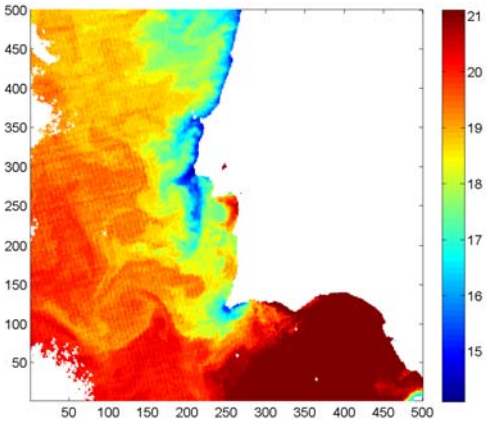


Figura B.12 19980718

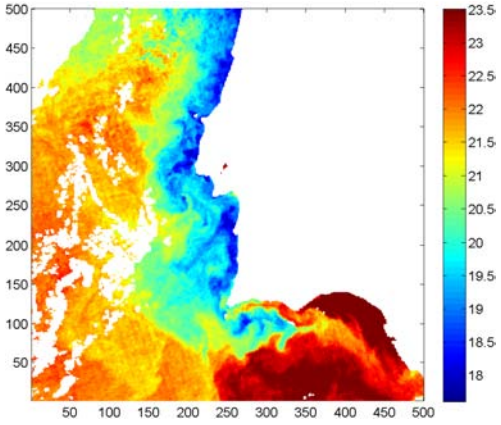


Figura B.13 19980721

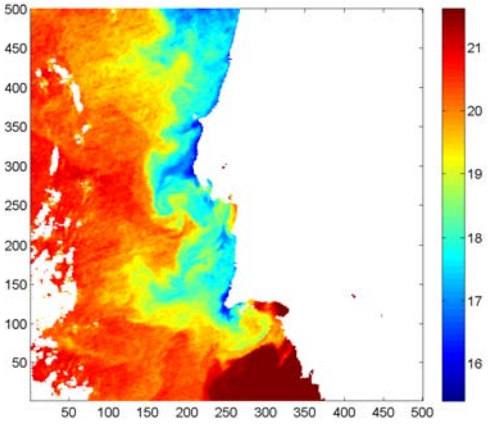


Figura B.14 19980724

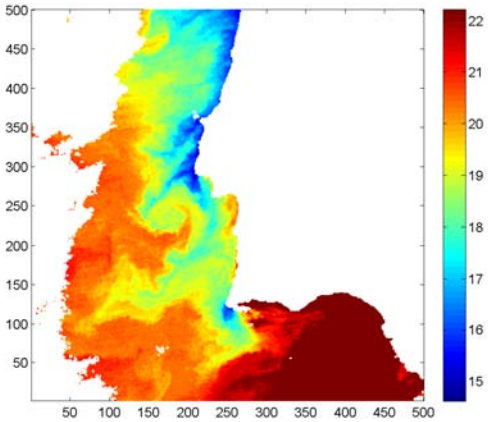


Figura B.15 19980728

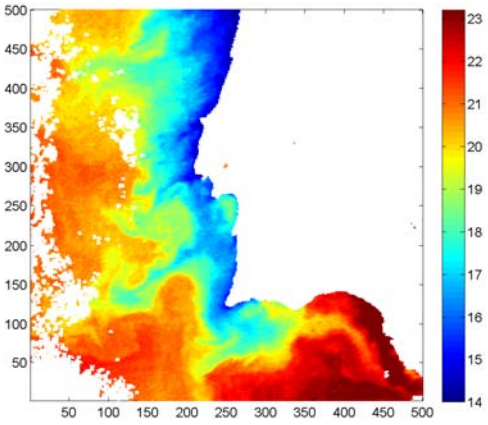


Figura B.16 19980801

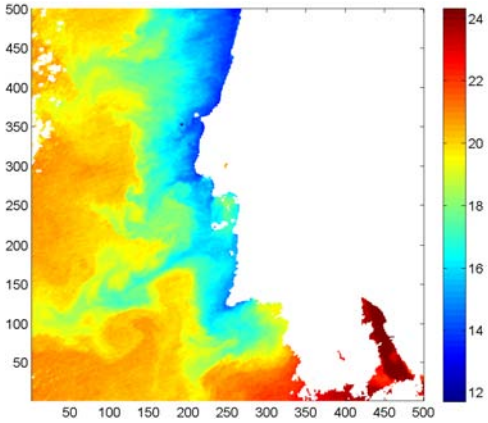


Figura B.17 19980802

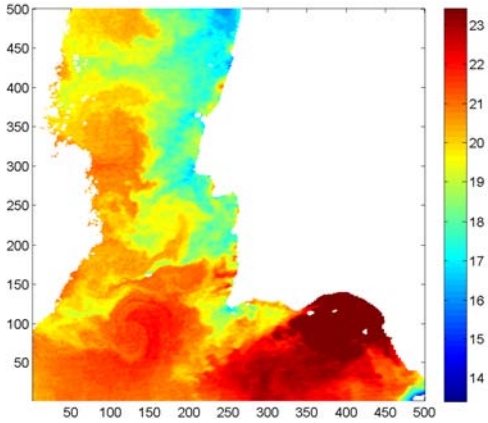


Figura B.18 19980805

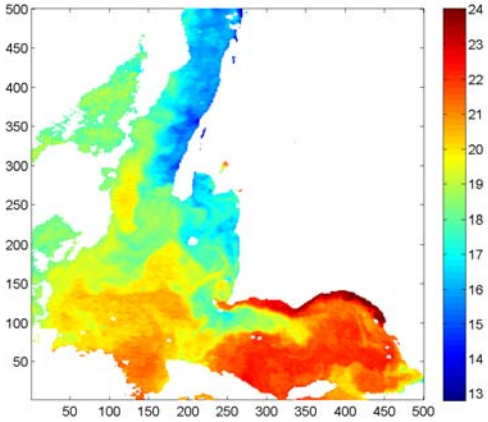


Figura B.19 19980810

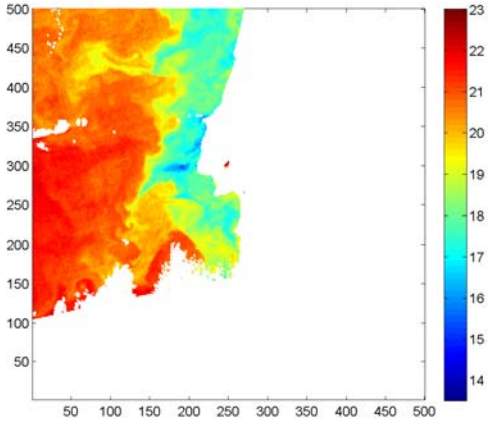


Figura B.20 19980812

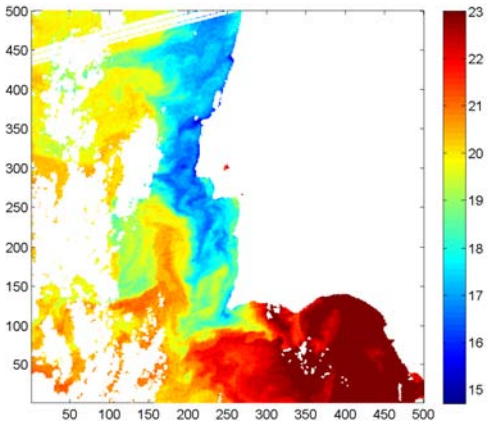


Figura B.21 19980819

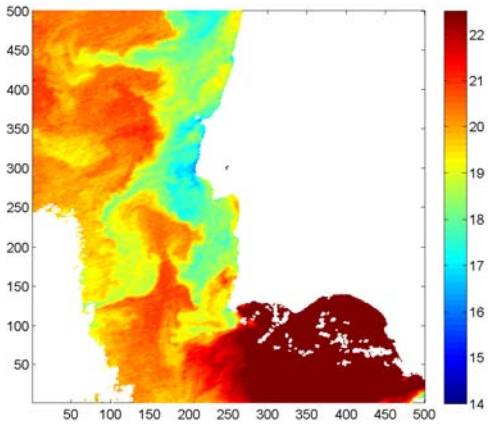


Figura B.22 19980821

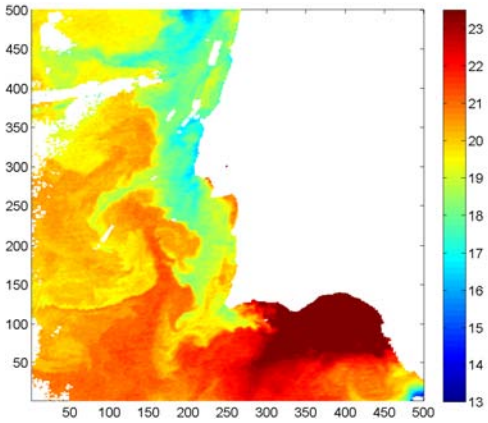


Figura B.23 19980823

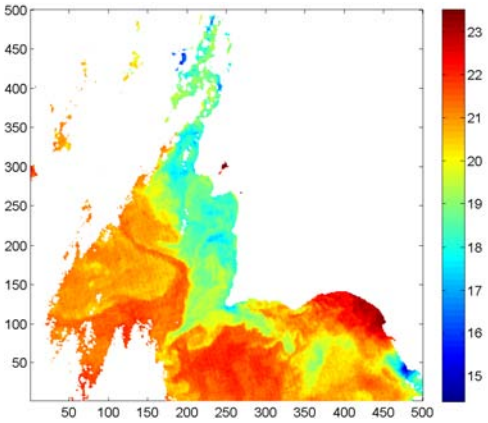


Figura B.24 19980830

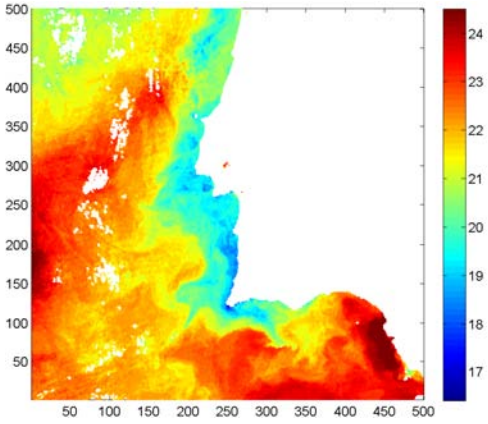


Figura B.25 19980905

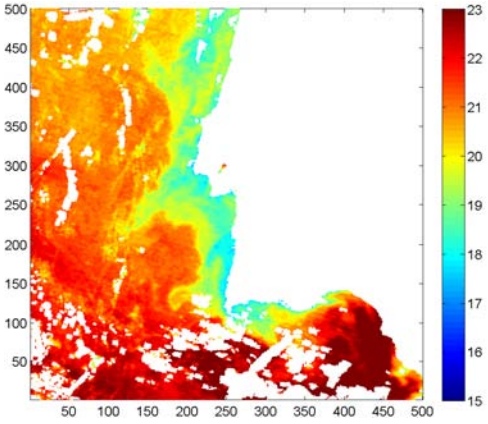


Figura B.26 19980908

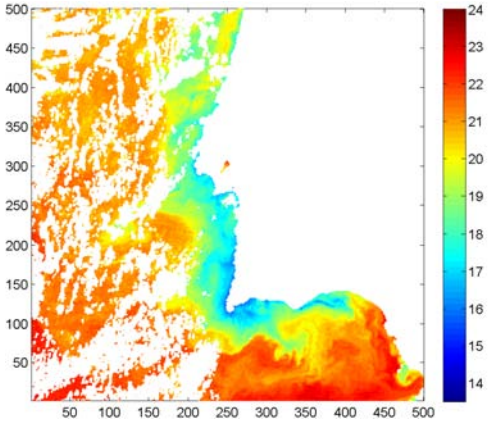


Figura B.27 19980911

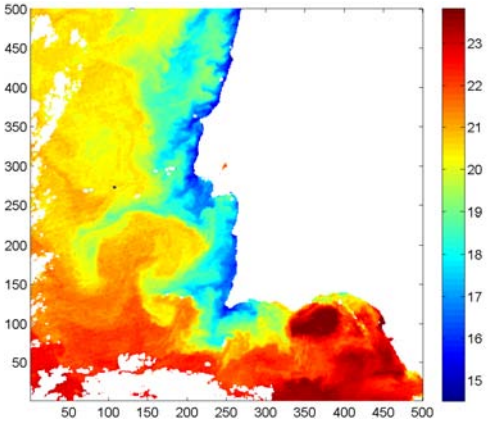


Figura B.28 19980915

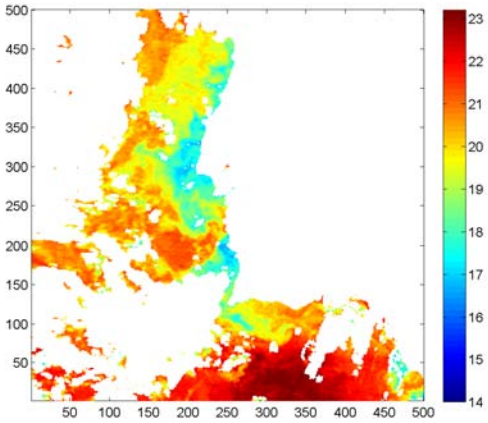


Figura B.29 19980924

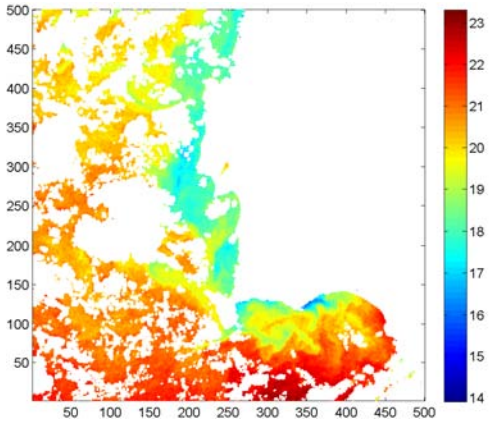


Figura B.30 19980930

B.2 Ano 1999

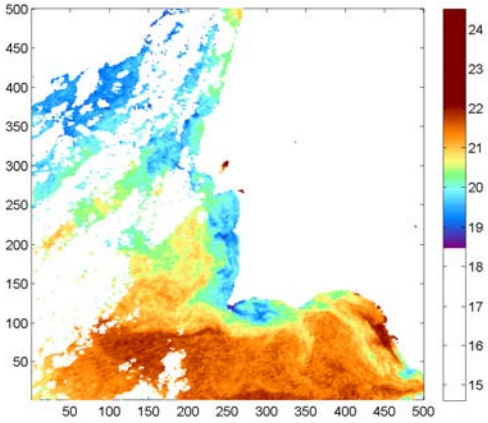


Figura B.31 19990602

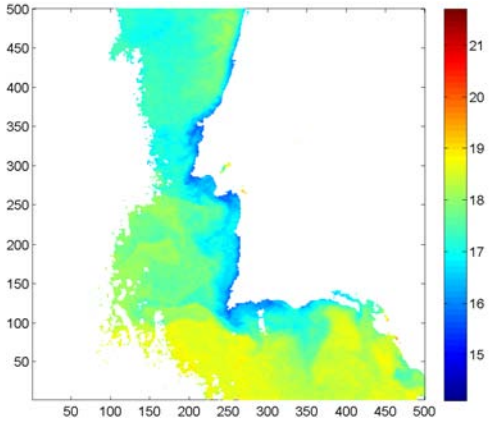


Figura B.32 19990608

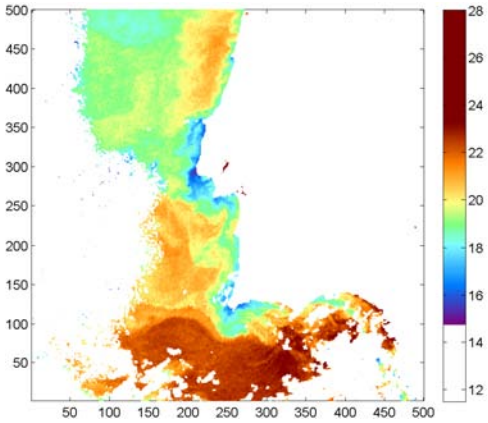


Figura B.33 19990610

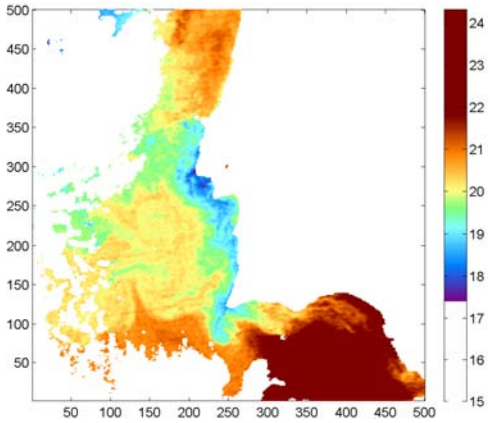


Figura B.34 19990614

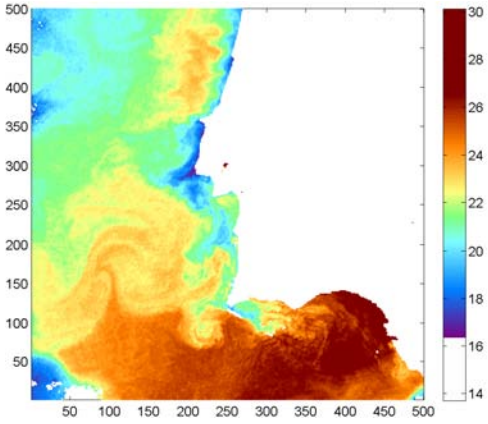


Figura B.35 19990619

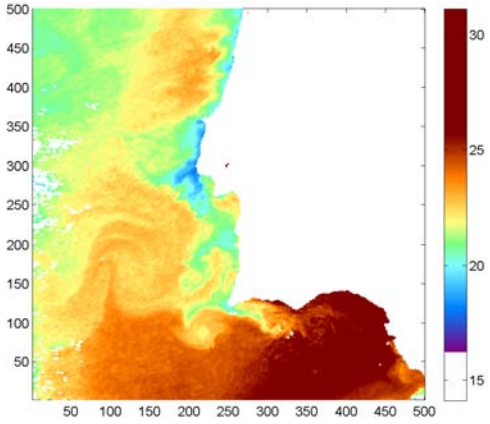


Figura B.36 19990620

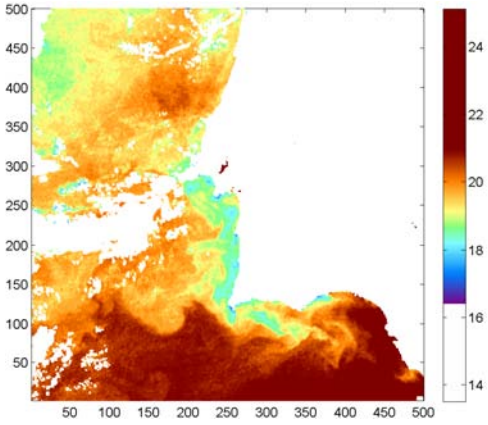


Figura B.37 19990627

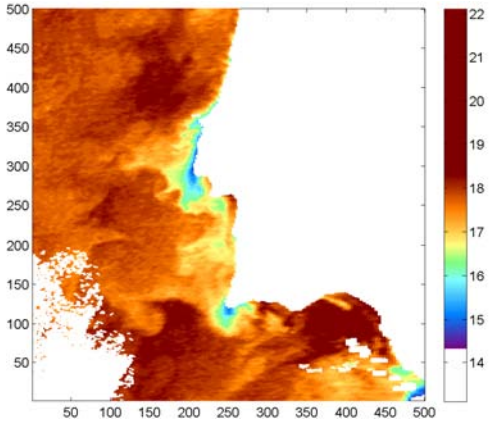


Figura B.38 19990630

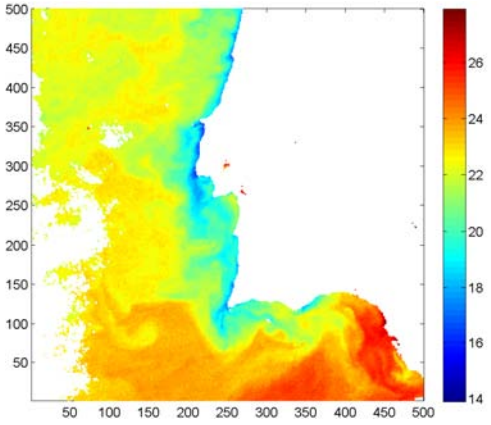


Figura B.39 19990706

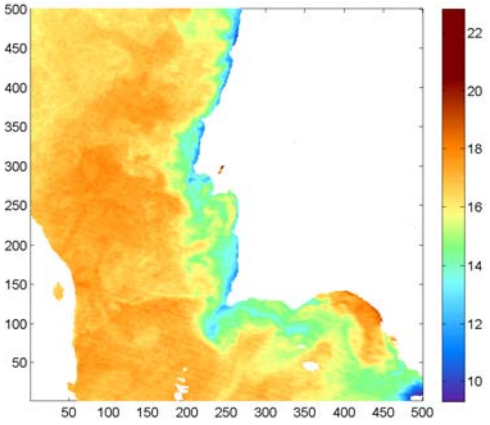


Figura B.40 19990708

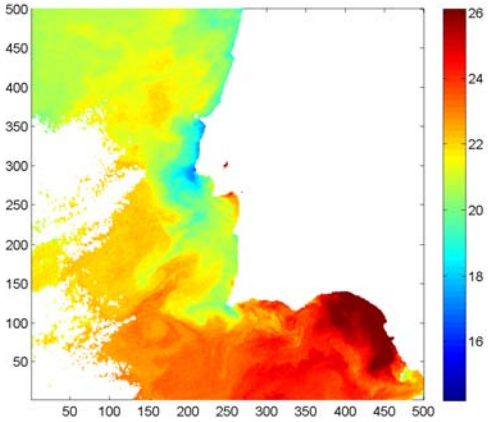


Figura B.41 19990714

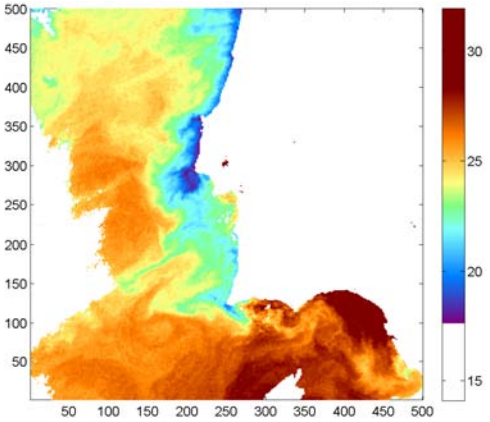


Figura B.42 19990715

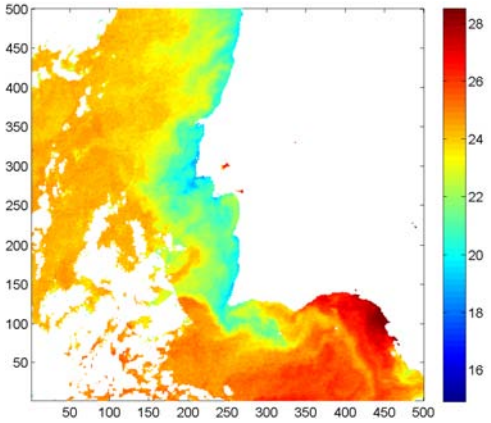


Figura B.43 19990719

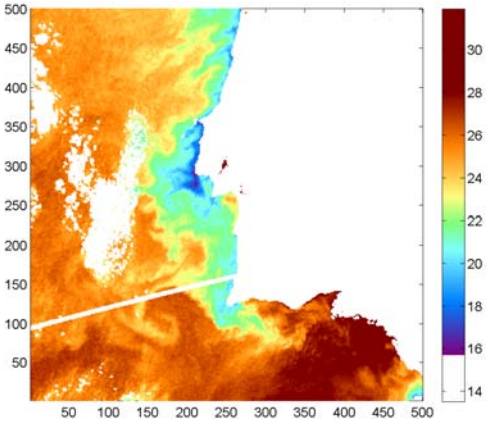


Figura B.44 19990721

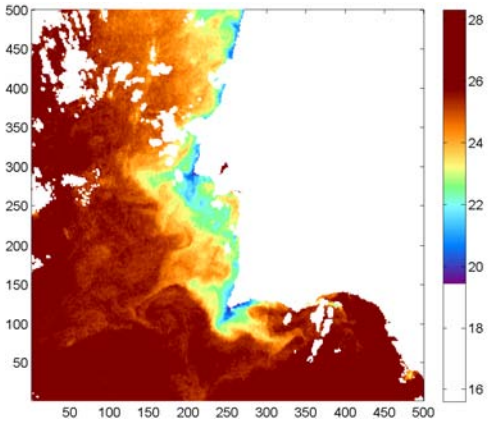


Figura B.45 19990729

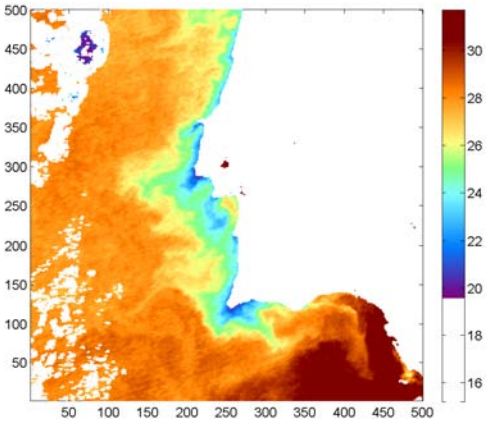


Figura B.46 19990731

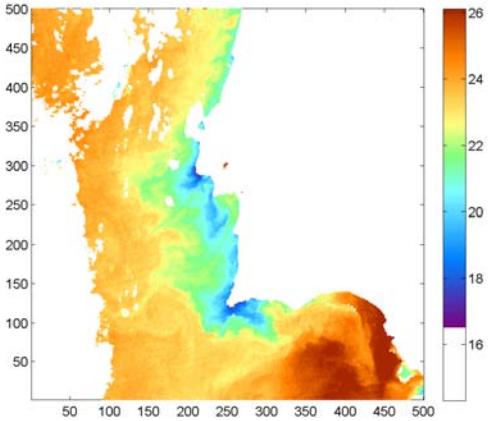


Figura B.47 19990801

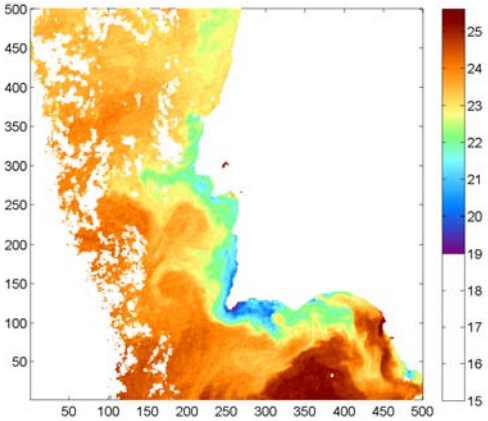


Figura B.48 19990810

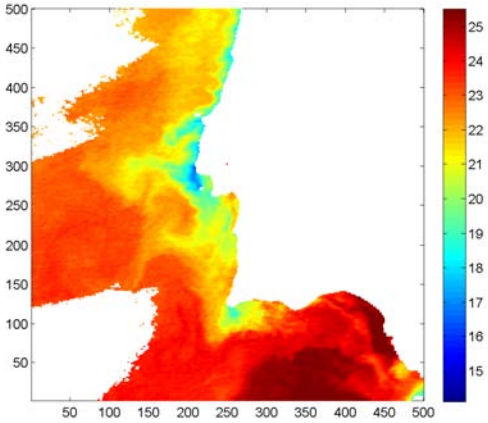


Figura B.49 19990814

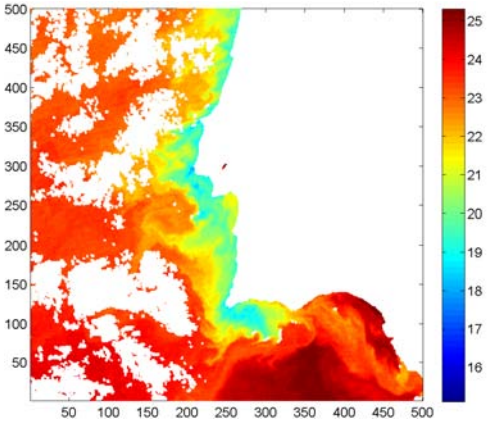


Figura B.50 19990817

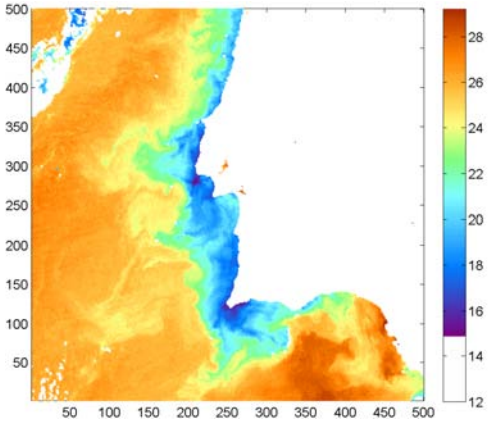


Figura B.51 19990821

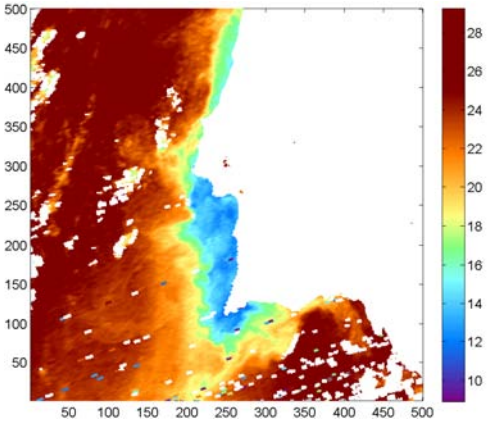


Figura B.52 19990823

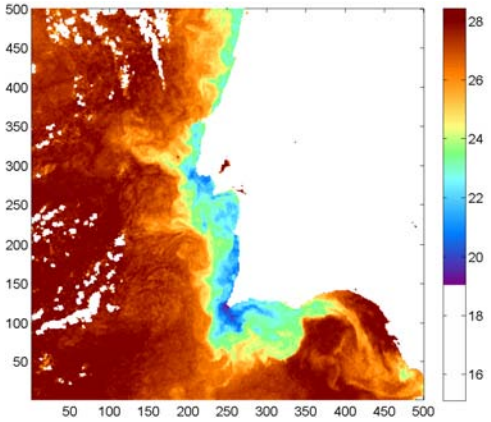


Figura B.53 19990826

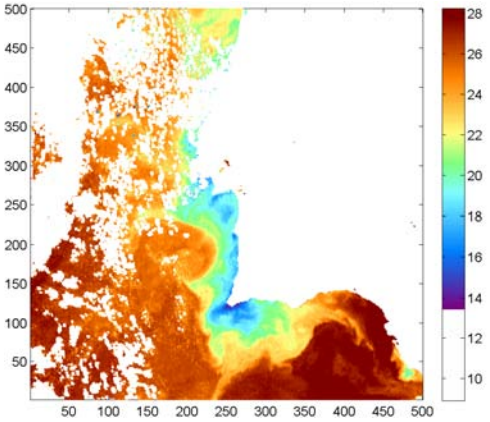


Figura B.54 19990830

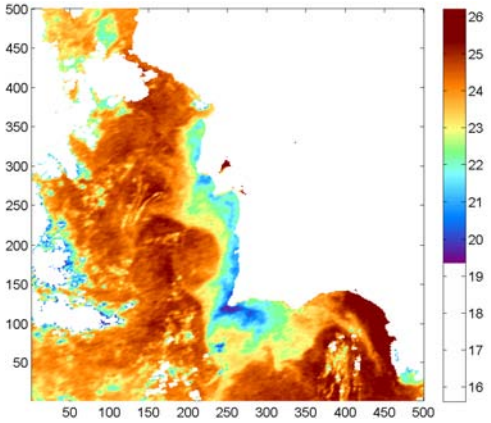


Figura B.55 19990901

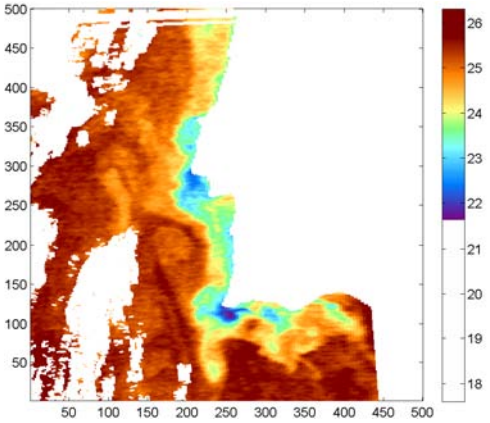


Figura B.56 19990908

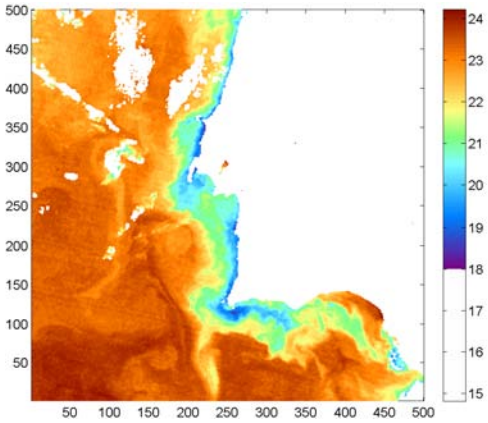


Figura B.57 19990910

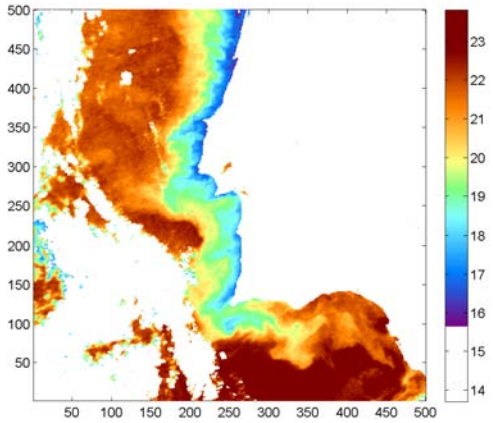


Figura B.58 19990914

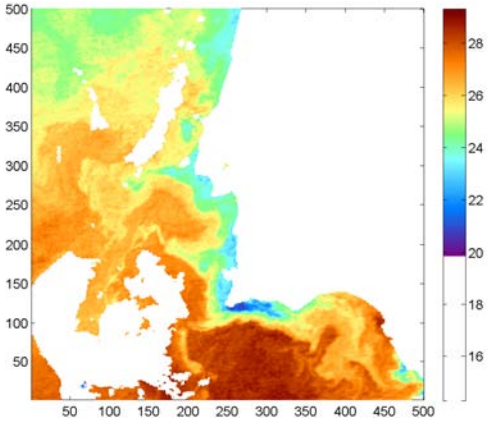


Figura B.59 19990928

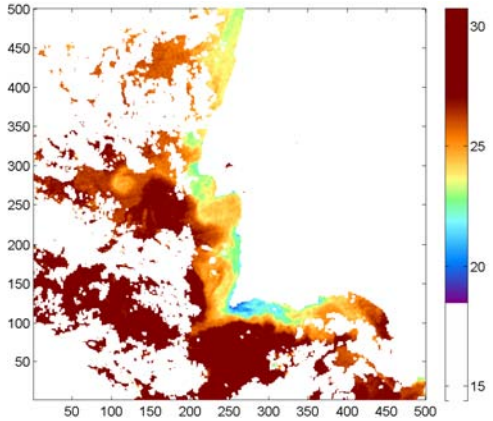


Figura B.60 19990930

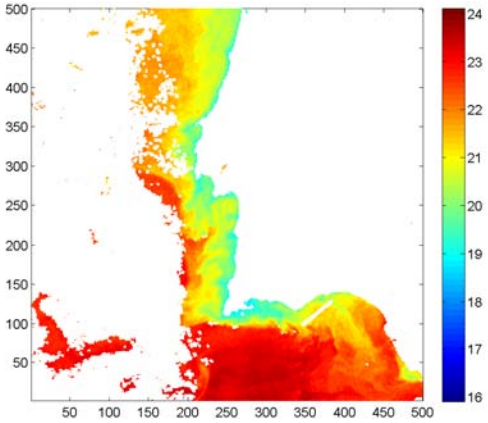


Figura B.61 19991003

B.3 Anotações textuais

B.3.1 Ano 1998

Índice	Mapa SST	Anotação
1	980609	Limite verde até C. Espichel; para Sul deste cabo, limite amarelo
2	980612	Limite verde até C. Espichel; para Sul deste cabo, limite amarelo
3	980614	Limite verde até C. S. Vicente; para Sul e Leste limite amarelo
4	980618	Limite verde até C. S. Vicente; para Sul e Leste limite amarelo
5	980623	Limite amarelo para toda a costa Oeste
6	980625	Limite amarelo para toda a costa Oeste
7	980628	Limite amarelo para toda a costa Oeste
8	980703	Limite verde até C. Espichel; para Sul, limite amarelo
9	980707	Limite verde até C. Espichel; para Sul, limite amarelo
10	980711	Limite amarelo para toda a costa Oeste
11	980715	Limite verde até Peniche; para Sul limite amarelo
12	980718	Limite amarelo para toda a costa Oeste
13	980721	Limite amarelo do C. Mondego para Sul (a Norte deste cabo muito confuso)
14	980724	Limite amarelo para toda a costa Oeste
15	980728	Limite amarelo para toda a costa Oeste
16	980801	Limite amarelo para toda a costa Oeste
17	980802	Limite verde para toda a costa Oeste
18	980805	Limite amarelo para toda a costa Oeste
19	980810	De Peniche até Lisboa, limite azul; de Lisboa até Sines limite verde, depois limite amarelo
20	980812	Limite amarelo para toda a costa
21	980819	Limite verde até C. Espichel; para Sul, limite amarelo
22	980821	Limite amarelo para toda a costa Oeste
23	980823	Limite verde até Peniche; para Sul, limite amarelo
24	980830	Limite amarelo de Lisboa para Sul
25	980905	Limite verde até C. S. Vicente; para Sul e Leste limite amarelo
26	980908	Limite amarelo para toda a costa Oeste
27	980911	Limite verde até C. Espichel; para Sul, limite amarelo
28	980915	Limite verde até C. Espichel; para Sul, limite amarelo
29	980924	Limite verde até C. Espichel; para Sul, limite amarelo
30	980930	Limite verde até C. S. Vicente; para Leste limite amarelo

Tabela B.1 Anotações textuais feitas por oceanógrafos para a definição da região de upwelling para as imagens relativas ao ano de 1998.

B.3.2 Ano 1999

Índice	Mapa SST	Anotação
31	990602	Limite verde até C. S. Vicente; para Sul e Leste limite amarelo
32	990608	Limites azul escuro até C. Espichel, azul claro até C.S. Vicente e verde para Sul e Leste
33	990610	Limite azul claro até C. Espichel; limite amarelo para Sul
34	990614	Limite azul claro até C. Espichel; limite verde para Sul
35	990619	Limite verde até C. Sines; limite amarelo para Sul
36	990620	Limite verde para toda a costa Oeste
37	990627	Limite verde do C. Roca até C. S. Vicente; limite amarelo para Sul e Leste
38	990630	Limite amarelo para toda a costa Oeste
39	990706	Limite verde para toda a costa Oeste
40	990708	Limite amarelo para toda a costa Oeste
41	990714	Limite verde até C. Roca; limite amarelo para Sul
42	990715	Limite verde até C. Roca; limite amarelo para Sul
43	990719	Limite verde até C. Roca; limite amarelo para Sul
44	990721	Limite amarelo para toda a costa Oeste
45	990729	Limite amarelo para toda a costa Oeste
46	990731	Limite amarelo para toda a costa Oeste
47	990801	Limite verde até C. Roca; limite amarelo para Sul
48	990810	Limite verde até C. Roca; limite amarelo para Sul
49	990814	Limite amarelo para toda a costa Oeste
50	990817	Limite amarelo para toda a costa Oeste
51	990821	Limite verde até C. Roca; limite amarelo para Sul
52	990823	Limite amarelo para toda a costa Oeste
53	990826	Limite amarelo para toda a costa Oeste
54	990830	Só interessa para Sul do C. Roca - limite amarelo
55	990901	Limite amarelo para toda a costa Oeste
56	990908	Limite amarelo para Sul de Peniche
57	990910	Limite amarelo para toda a costa Oeste
58	990914	Limite amarelo para toda a costa Oeste
59	990928	Limite verde até C. Roca; limite amarelo para Sul
60	990930	Só interessa para Sul do C. Roca - limite amarelo
61	991003	Limite verde até C. Roca; limite amarelo para Sul

Tabela B.2 Anotações textuais feitas por oceanógrafos para a definição da região de upwelling para as imagens relativas ao ano de 1999.

C . Resultados AP_{C1}-FCM [1]

C.1 Mapas de Segmentação

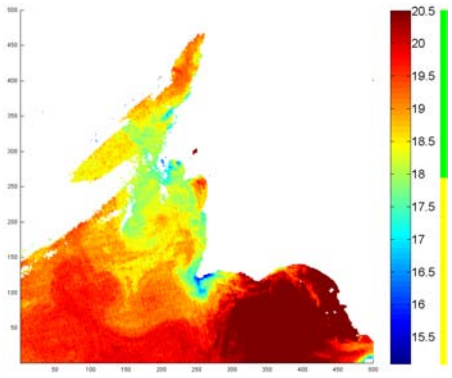


Figura C.1 19980609

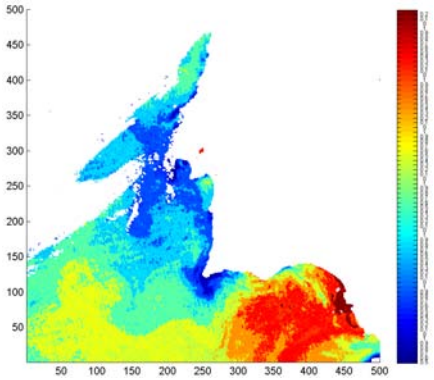


Figura C.2 19980609_8c

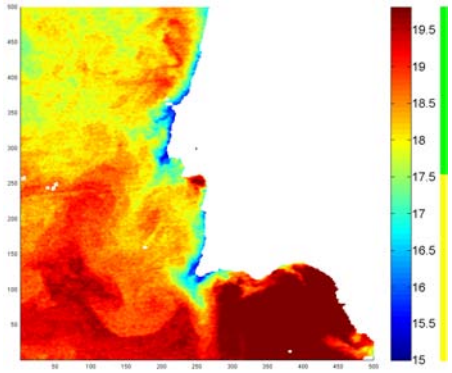


Figura C.3 19980612

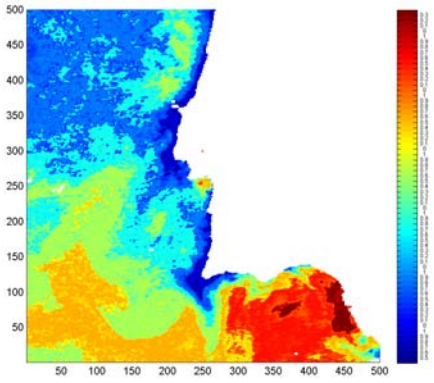


Figura C.4 19980612_7c

[1] Devido à grande quantidade de resultados disponíveis, optou-se escolher uma pequena amostra (5 mapas) representativa do conjunto de 61 mapas existentes para apresentar nos anexos à versão impressa da dissertação. Os resultados referentes a todas os mapas de temperatura utilizados estão em anexo digital (formato DVD).

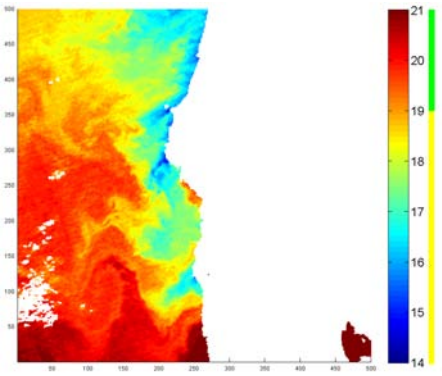


Figura C.5 19980715

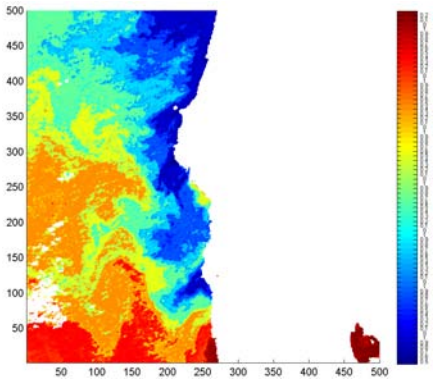


Figura C.6 19980715_8c

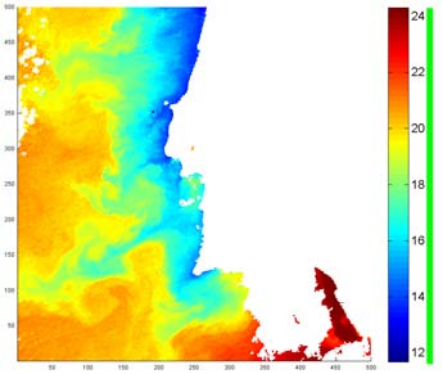


Figura C.7 19980802

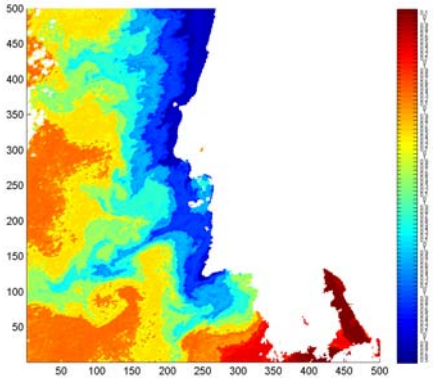


Figura C.8 19980802_9c

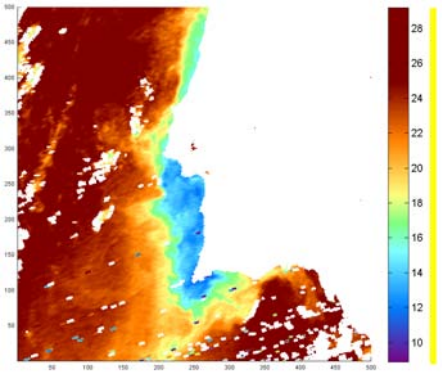


Figura C.9 19990823

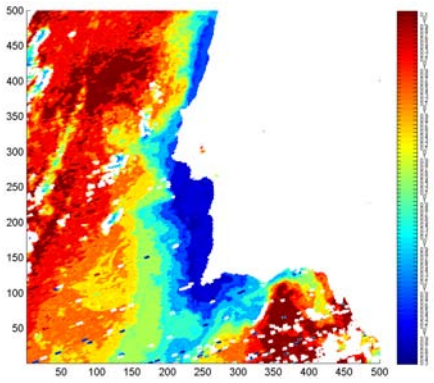


Figura C.10 19990823_9c

C.2 Visualização de Fronteiras

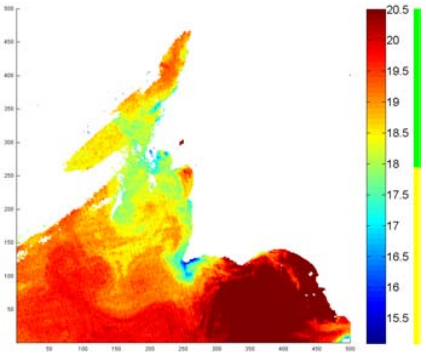


Figura C.11 19980609

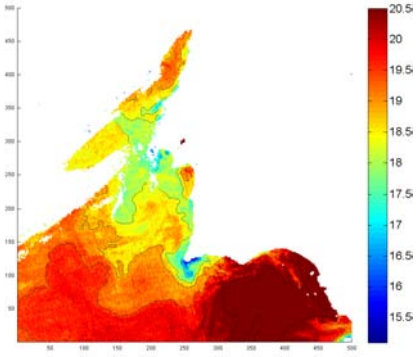


Figura C.12 19980609_8c

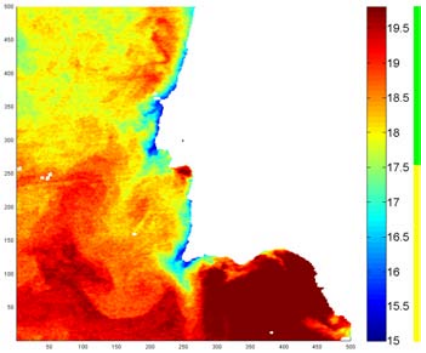


Figura C.13 19980612

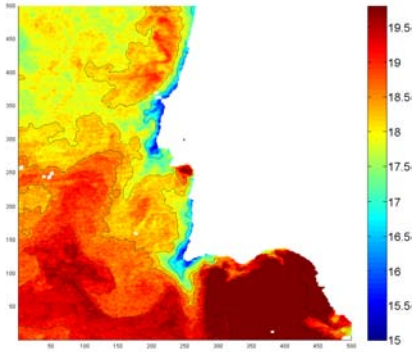


Figura C.14 19980612_7c

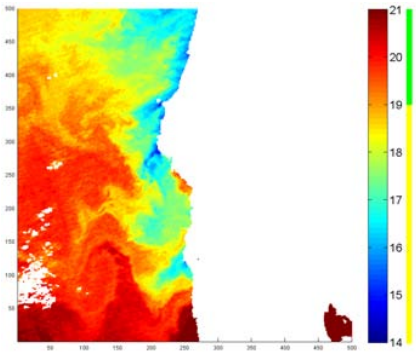


Figura C.15 19980715

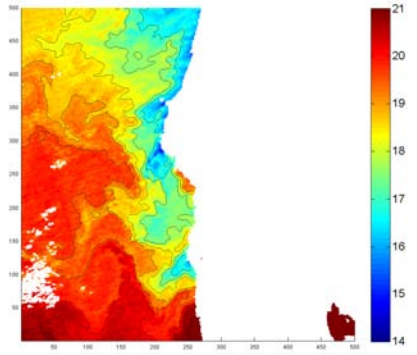


Figura C.16 19980715_8c

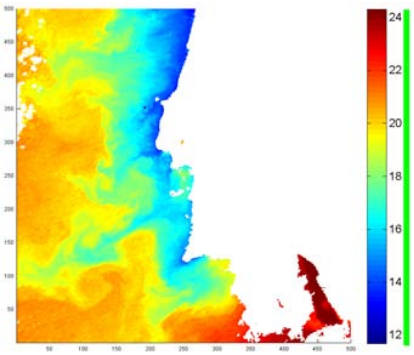


Figura C.17 19980802

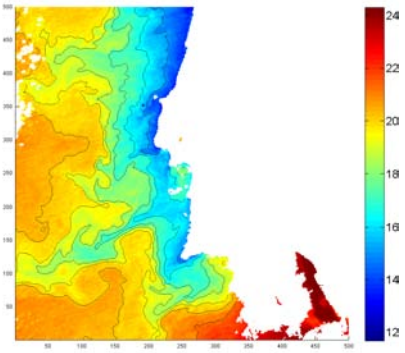


Figura C.18 19980802_9c

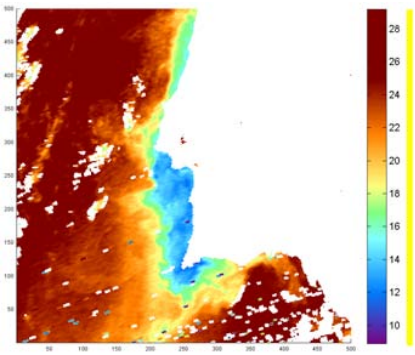


Figura C.19 19990823

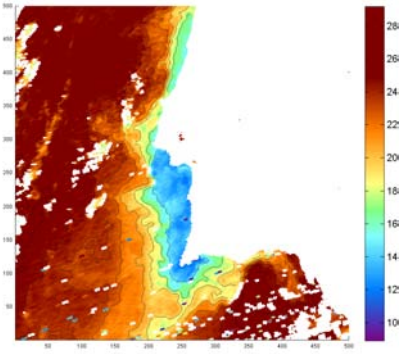


Figura C.20 19990823_9c

C.3 Tabela contribuição para a dispersão de dados

C.3.1 Ano 1998

	1º Cluster	2º Cluster	3º Cluster	4º Cluster	5º Cluster	6º Cluster	7º Cluster	8º Cluster	9º Cluster	10º Cluster
19980609	46,68	34,34	3,25	2,21	0,23	0,11	**0,02**	0,0011		
19980612	45,11	31,50	5,25	0,97	0,32	**0,01**	0,01			
19980614	46,30	34,40	3,80	4,79	0,30	0,53	**0,05**	0,0267	0,0028	
19980618	35,12	44,36	4,97	1,56	0,48	0,18	**0,04**	0,0037		
19980623	44,15	33,03	3,15	2,88	0,14	**0,08**	0,03	0,0016		
19980625	31,78	50,68	2,65	2,98	0,20	0,26	**0,01**	0,0040		
19980628	50,25	36,80	1,19	2,59	0,16	**0,08**	0,01	0,0037		
19980703	49,46	37,67	1,74	2,71	0,24	0,14	**0,02**	0,0023		
19980707	39,35	43,05	3,29	2,96	0,22	0,11	**0,02**	0,000001		
19980711	51,88	30,43	3,41	3,57	0,18	0,15	**0,02**	0,0075	0,0015	
19980715	34,88	45,67	2,98	1,02	**0,07**	0,02	0,01	0,0002		
19980718	33,19	44,39	2,76	4,88	0,11	0,12	**0,02**	0,0009		
19980721	37,82	44,00	5,21	2,69	0,11	0,23	**0,01**	0,0035		
19980724	26,52	48,19	12,41	2,68	0,20	0,34	**0,02**	0,0220	0,000011	
19980728	51,61	33,11	3,04	1,32	0,39	**0,06**	0,06	0,0067	0,000003	
19980801	32,97	49,11	2,31	2,48	**0,07**	0,08	0,01	0,0026	0,0001	
19980802	48,64	29,94	2,82	1,30	0,11	**0,03**	0,01	0,0019	0,0003	
19980805	42,85	37,02	4,81	2,26	0,27	0,11	**0,01**	0,0081	0,0001	
19980810	42,48	40,90	3,82	2,77	**0,09**	0,07	0,01	0,0030	0,0002	
19980812	31,46	57,38	1,81	1,16	**0,07**	0,02	0,01	0,0007		
19980819	48,89	38,75	1,48	2,12	0,13	0,15	**0,01**	0,0035	0,0004	
19980821	57,83	27,66	1,33	1,80	0,20	0,14	**0,02**	0,0156	0,0068	0,0002
19980823	48,31	30,71	6,28	2,66	0,33	0,11	**0,02**	0,0086	0,0004	
19980830	31,09	52,72	2,19	1,59	0,12	**0,10**	0,01	0,0179	0,000018	
19980905	45,69	36,51	3,45	1,66	0,17	**0,09**	0,02	0,0022		
19980908	39,93	44,12	2,77	2,63	0,17	**0,08**	0,02	0,0005		
19980911	55,61	28,26	3,27	1,87	0,14	**0,04**	0,02	0,0056	0,0011	
19980915	47,54	37,44	2,15	1,73	0,21	0,11	**0,01**	0,0048		
19980924	42,06	40,04	5,12	3,00	0,19	0,21	**0,02**	0,0098	0,0008	
19980930	50,28	34,33	3,30	1,69	0,13	**0,06**	0,01	0,0037		

Tabela C.1 Tabela da contribuição para a dispersão de dados dos clusters extraídos pelo *Anomalous Pattern* (%). A *bold* marcação do número de clusters por aplicação da condição de paragem AP-C3 com o *threshold* definido em 0.1%.

C.3.2 Ano 1999

	1º Cluster	2º Cluster	3º Cluster	4º Cluster	5º Cluster	6º Cluster	7º Cluster	8º Cluster	9º Cluster	10º Cluster
19990602	40,22	48,34	1,49	1,47	0,10	**0,04**	0,02	0,0023		
19990608	41,35	42,65	3,59	1,94	0,15	**0,09**	0,02	0,0005		
19990610	40,27	34,17	6,24	2,60	0,12	**0,09**	0,0039	0,0027		
19990614	58,01	26,20	2,22	2,30	0,29	**0,05**	0,0041	0,0010		
19990619	49,20	35,31	1,24	1,53	0,14	**0,08**	0,01	0,0004		
19990620	53,55	29,39	3,66	1,51	0,10	**0,04**	0,01	0,0044	0,0019	
19990627	48,21	35,46	1,46	2,46	**0,06**	0,08	0,01	0,01		
19990630	46,70	25,67	5,11	2,68	0,34	0,11	**0,0003**			
19990706	35,13	43,76	4,18	1,65	0,44	0,13	**0,05**	0,03	0,0009	
19990708	54,24	26,08	4,05	1,17	0,30	**0,05**	0,04	0,0005		
19990714	47,09	31,94	4,10	1,88	0,10	**0,08**	0,01	0,0019		
19990715	36,47	39,27	5,35	2,92	0,18	**0,08**	0,01	0,0009		
19990719	30,20	49,44	3,07	3,89	0,21	0,25	**0,005**	0,005		
19990721	31,47	47,53	3,62	2,87	0,20	0,12	**0,02**	0,00005		
19990729	27,69	49,77	3,13	4,52	0,17	0,18	**0,01**	0,00003		
19990731	32,35	44,91	3,95	4,77	0,20	0,45	**0,04**	0,01	0,0046	
19990801	45,88	32,72	6,28	3,14	0,36	0,16	**0,03**	0,02	0,0002	
19990810	43,41	32,30	5,00	3,31	0,43	0,11	**0,02**	0,01		
19990814	41,10	36,13	3,20	2,42	**0,09**	0,11	0,0033	0,0015		
19990817	54,18	28,84	3,76	2,86	0,19	0,18	**0,01**	0,01	0,0005	
19990821	62,24	23,82	2,67	0,69	0,18	**0,02**	0,02	0,0031	0,0014	
19990823	58,48	25,68	3,70	0,85	0,22	**0,02**	0,01	0,0022	0,00003	
19990826	63,48	23,35	3,50	0,66	0,24	**0,03**	0,02	0,0031	0,0016	
19990830	51,18	31,48	3,49	1,56	0,15	**0,05**	0,0050	0,0004		
19990901	50,98	30,44	4,60	1,52	0,27	**0,04**	0,0045	0,0042		
19990908	57,52	25,41	4,13	1,36	0,22	0,11	**0,01**	0,01		
19990910	58,94	25,14	3,92	0,35	0,65	**0,02**	0,02	0,0002		
19990914	52,45	31,29	2,85	0,92	0,13	**0,02**	0,01	0,0025	0,00004	
19990928	0,28	38,86	42,62	2,87	3,16	**0,07**	0,06	0,01	0,01	
19990930	0,46	49,04	35,97	2,74	1,31	0,24	**0,05**	0,02	0,02	0,0000001
19991003	45,91	41,35	2,46	0,69	0,12	**0,04**	0,05	0,003	0,0007	

Tabela C.2 Tabela da contribuição para a dispersão de dados dos clusters extraídos pelo *Anomalous Pattern* (%). A *bold* marcação do número de clusters por aplicação da condição de paragem AP-C3 com o *threshold* definido em 0.1%.

D . Resultados AP_{C3}-FCM

D.1 Mapas de Segmentação

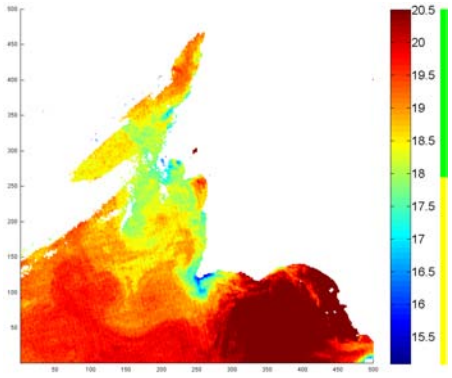


Figura D.1 19980609

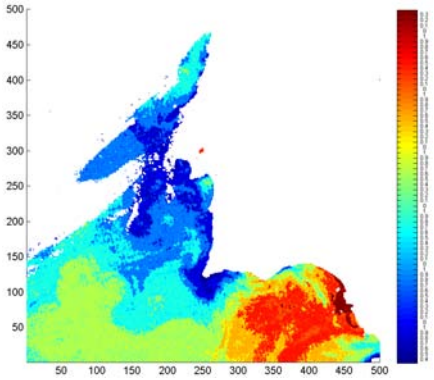


Figura D.2 19980609_7c

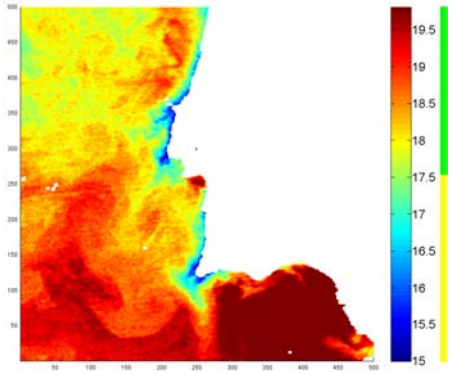


Figura D.3 19980612

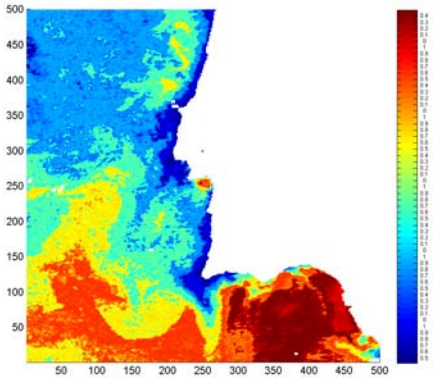


Figura D.4 19980612_6c

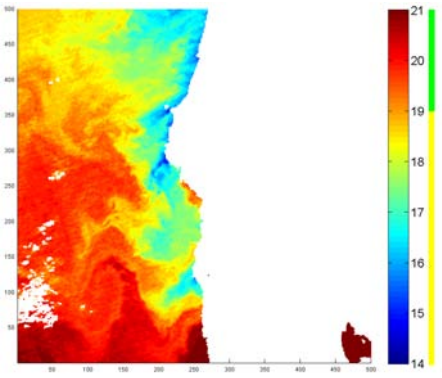


Figura D.5 19980715

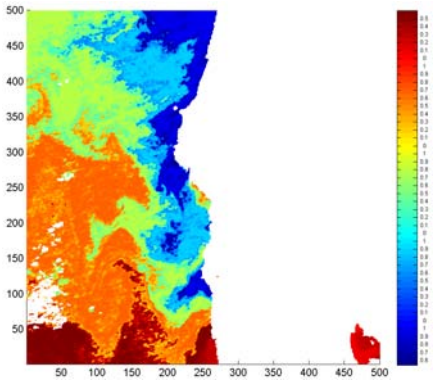


Figura D.6 19980715_5c

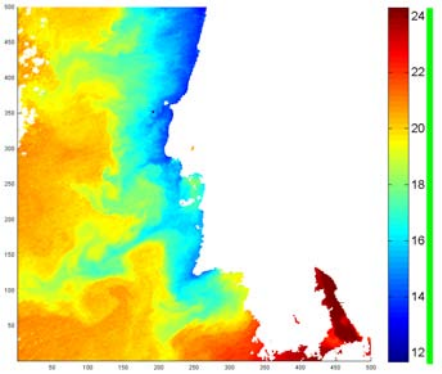


Figura D.7 19980802

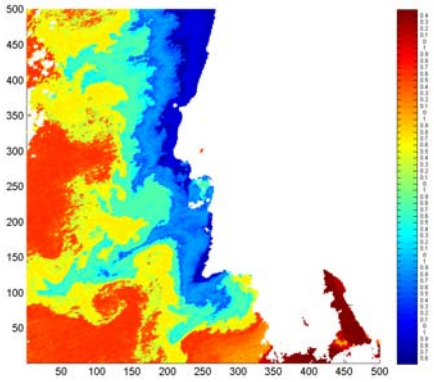


Figura D.8 19980802_6c

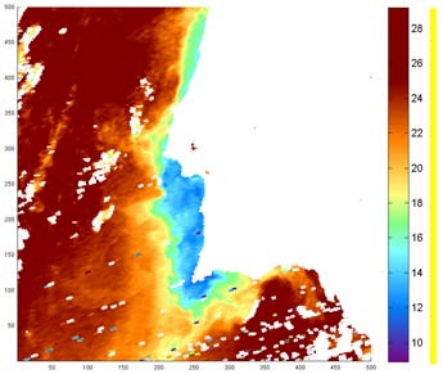


Figura D.9 19990823

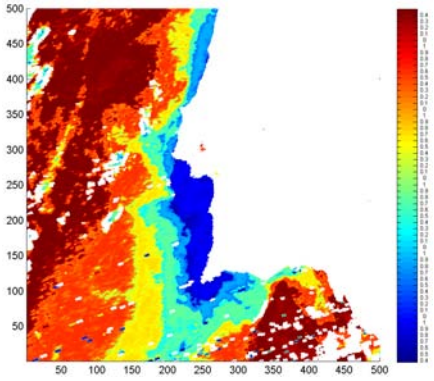


Figura D.10 19990823_6c

D.2 Visualização de Fronteiras

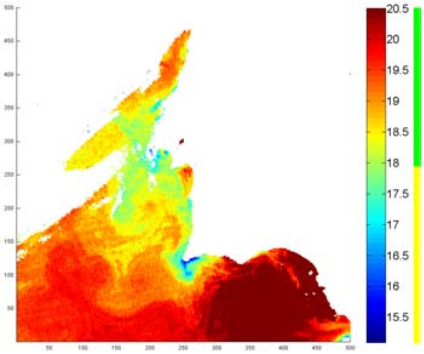


Figura D.11 19980609

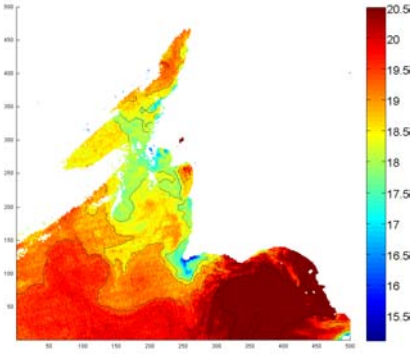


Figura D.12 19980609_7c

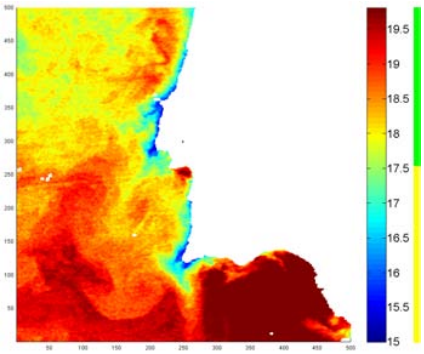


Figura D.13 19980612

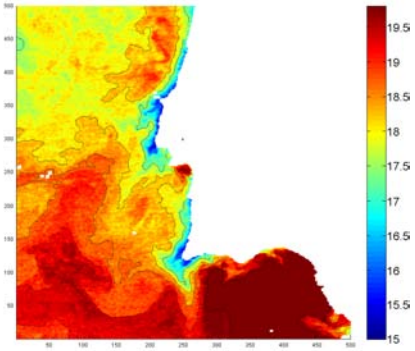


Figura D.14 19980612_6c

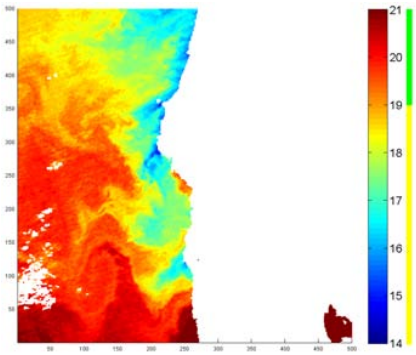


Figura D.15 19980715

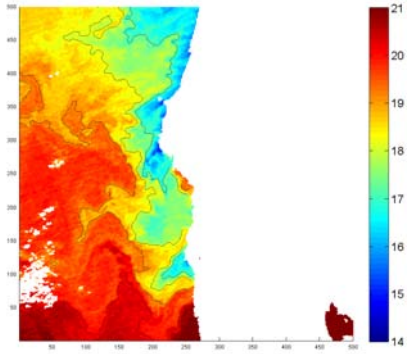


Figura D.16 19980715_5c

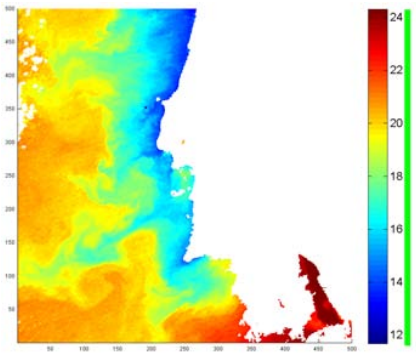


Figura D.17 19980802

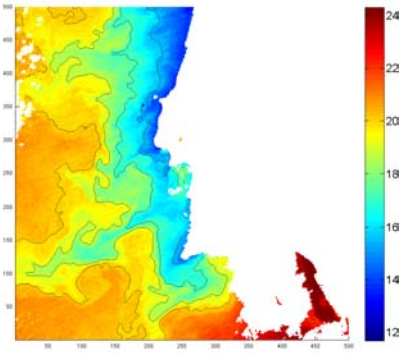


Figura D.18 19980802_6c

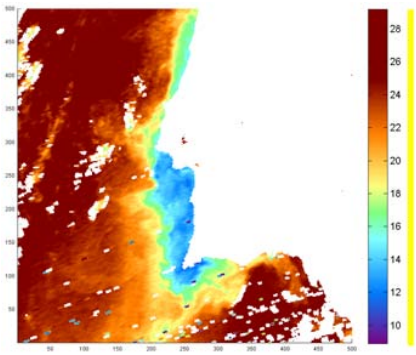


Figura D.19 19990823

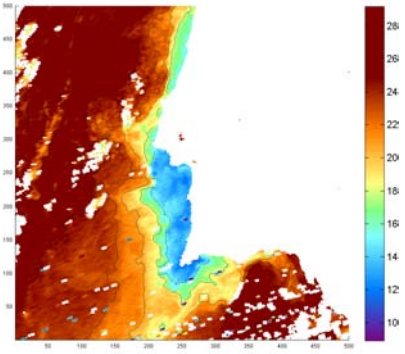


Figura D.20 19990823_6c

E . Resultados AP$_{C4}$-FCM

Devido à elevada quantidade de anexos e à semelhança entre os resultados dos algoritmos FCM e AP$_{C4}$-FCM, optou-se por anexar na versão impressa apenas os resultados do FCM. Em anexo digital, estão disponíveis todos os resultados (mapas de segmentação e fronteiras sobre mapas SST) para os dois algoritmos.

F . Resultados FCM

F.1 Mapas de Segmentação

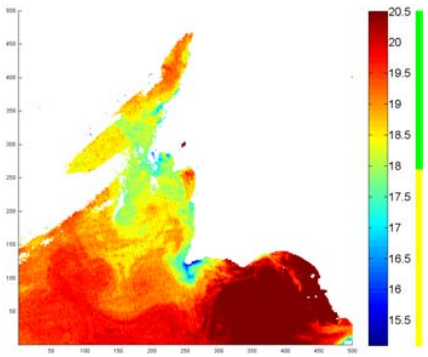


Figura F.1 19980609

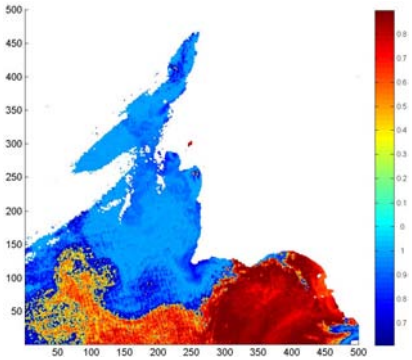


Figura F.2 19980609_2c

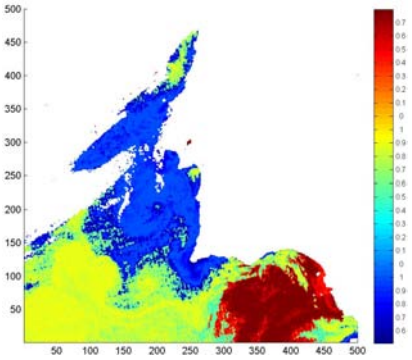


Figura F.3 19980609_3c

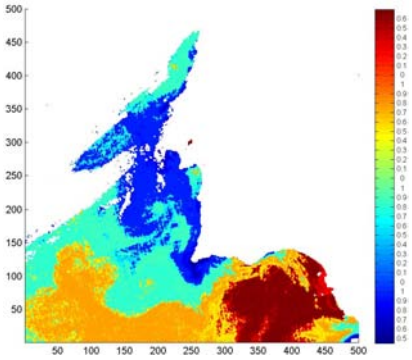


Figura F.4 19980609_4c

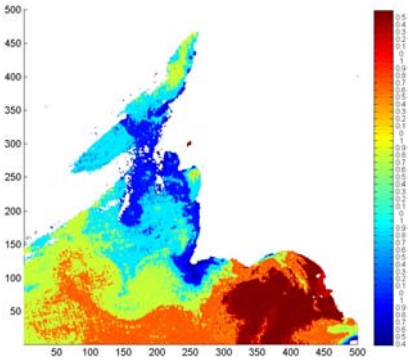


Figura F.5 19980609_5c

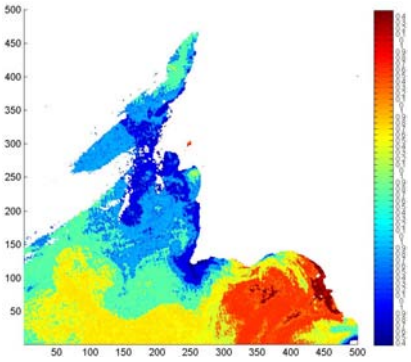


Figura F.6 19980609_6c

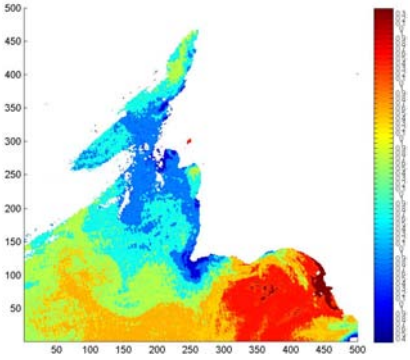


Figura F.7 19980609_7c

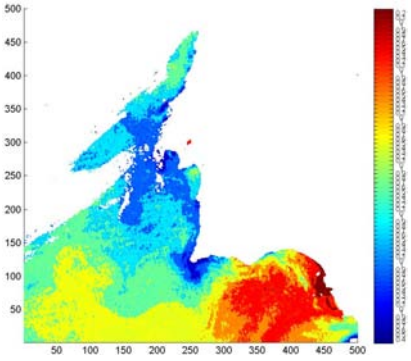


Figura F.8 19980609_8c

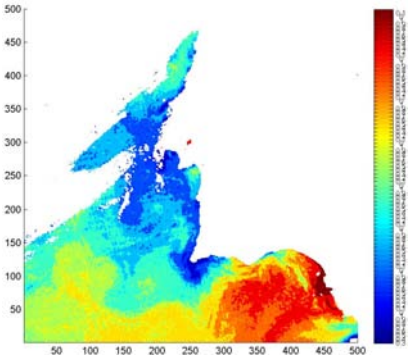


Figura F.9 19980609_9c

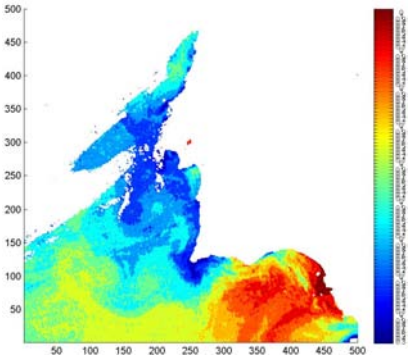


Figura F.10 19980609_10c

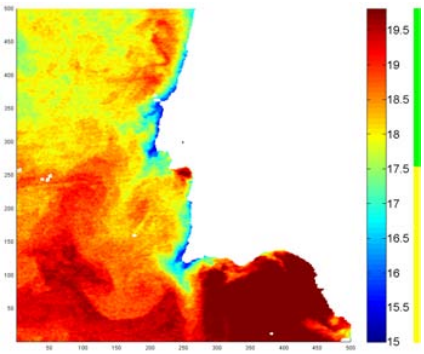


Figura F.11 19980612

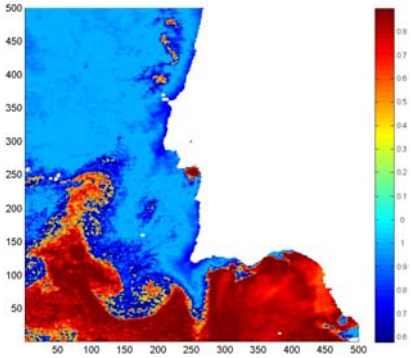


Figura F.12 19980612_2c

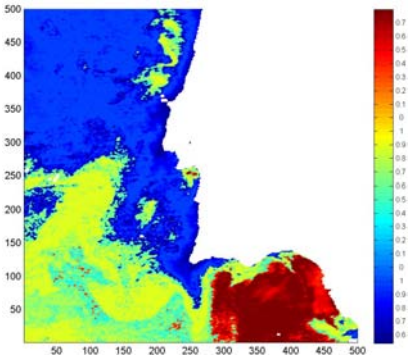


Figura F.13 19980612_3c

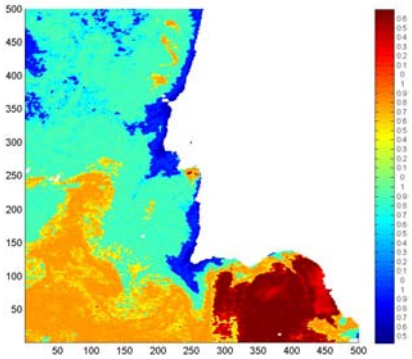


Figura F.14 19980612_4c

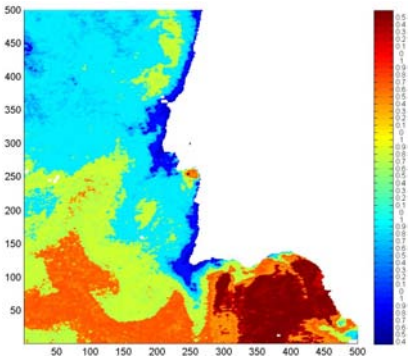


Figura F.15 19980612_5c

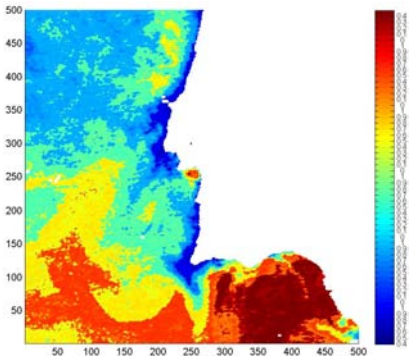


Figura F.16 19980612_6c

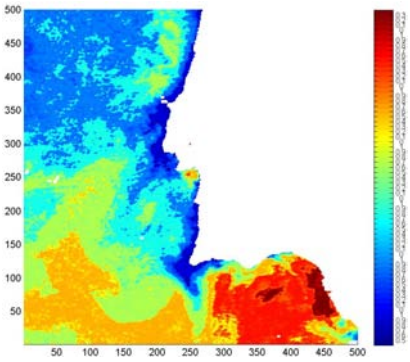


Figura F.17 19980612_7c

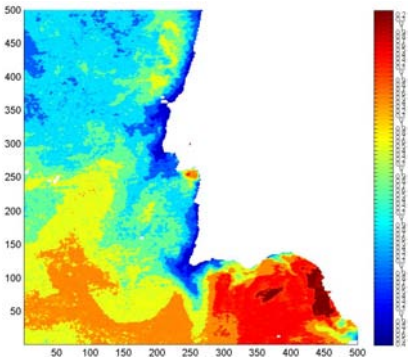


Figura F.18 19980612_8c

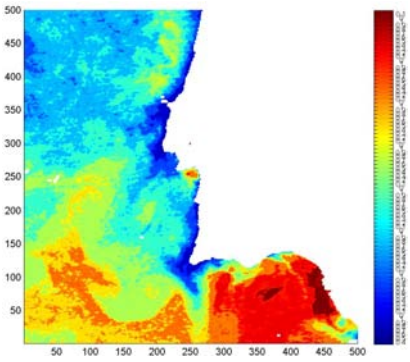


Figura F.19 19980612_9c

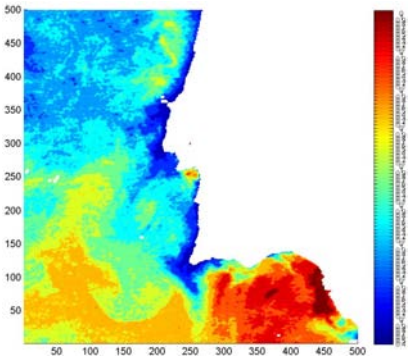


Figura F.20 19980612_10c

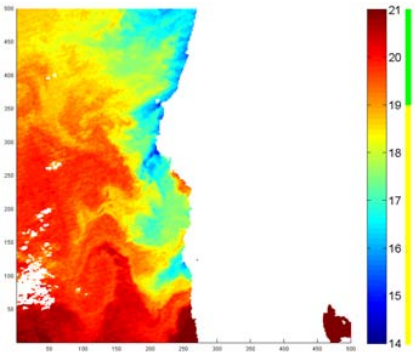


Figura F.21 19980715

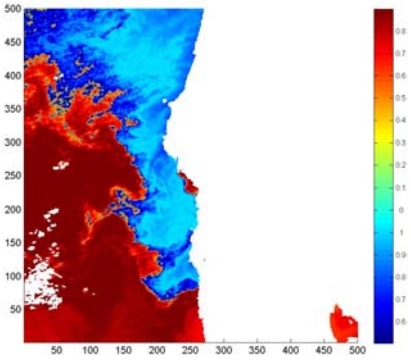


Figura F.22 19980715_2c

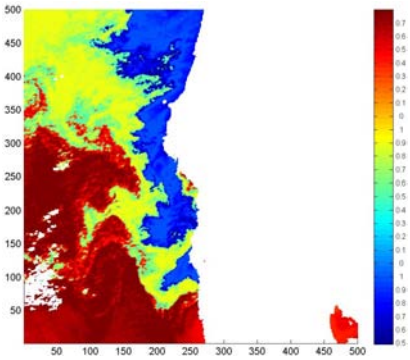


Figura F.23 19980715_3c

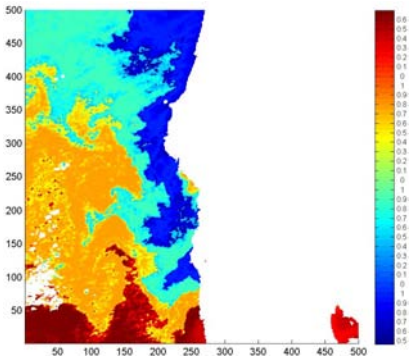


Figura F.24 19980715_4c

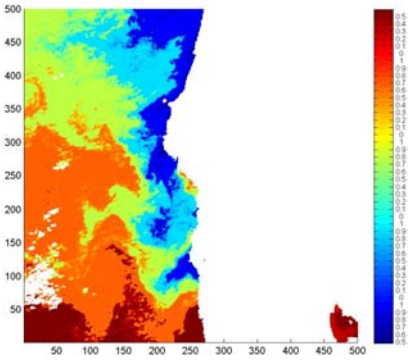


Figura F.25 19980715_5c

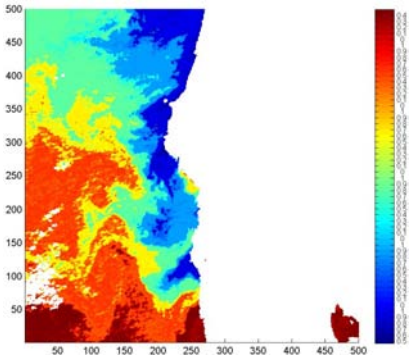


Figura F.26 19980715_6c

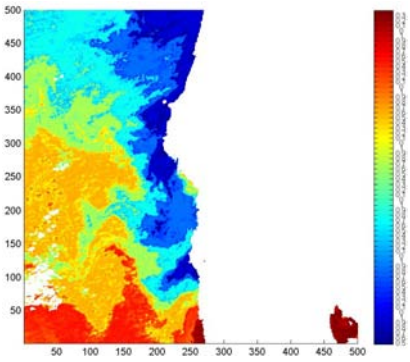


Figura F.27 19980715_7c

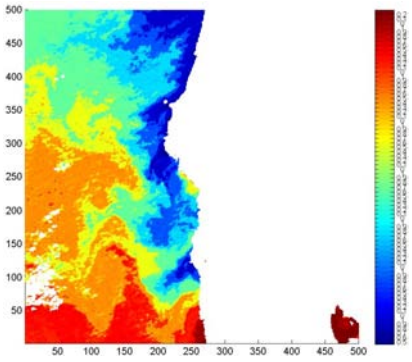


Figura F.28 19980715_8c

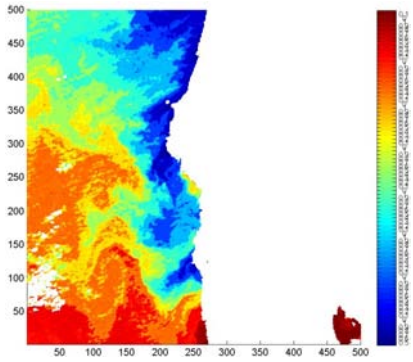


Figura F.29 19980715_9c

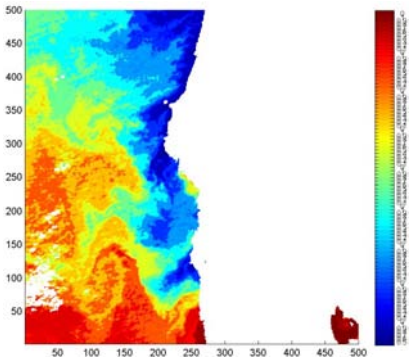


Figura F.30 19980715_10c

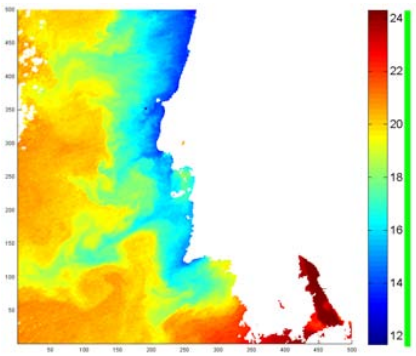


Figura F.31 19980802

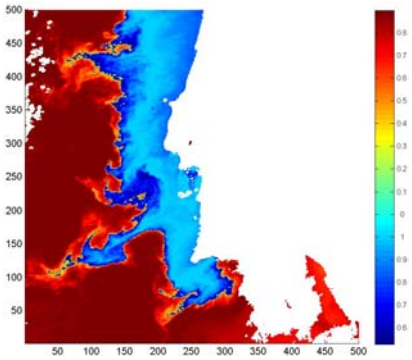


Figura F.32 19980802_2c

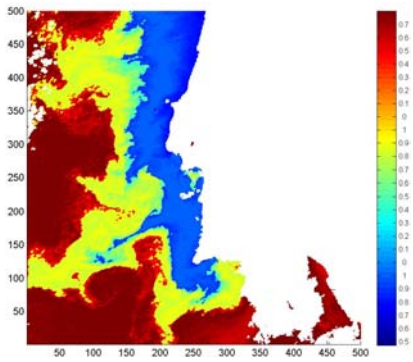


Figura F.33 19980802_3c

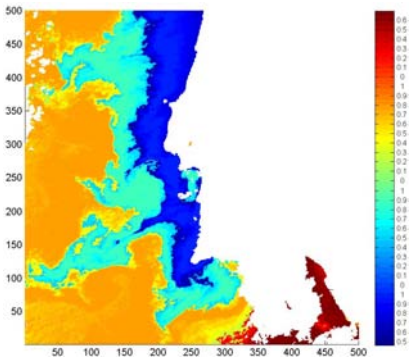


Figura F.34 19980802_4c

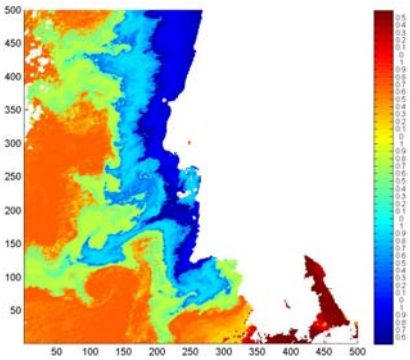


Figura F.35 19980802_5c

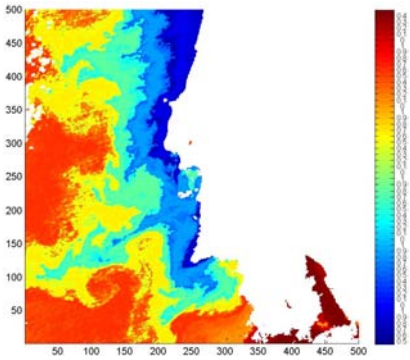


Figura F.36 19980802_6c

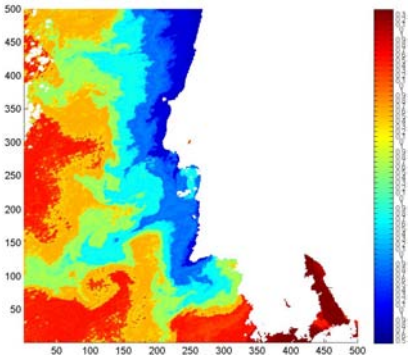


Figura F.37 19980802_7c

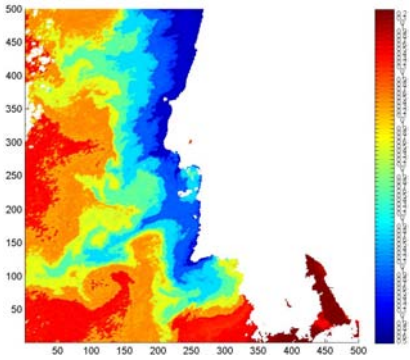


Figura F.38 19980802_8c

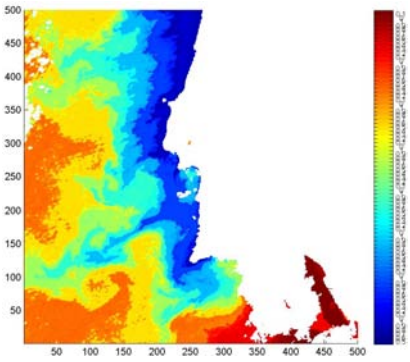


Figura F.39 19980802_9c

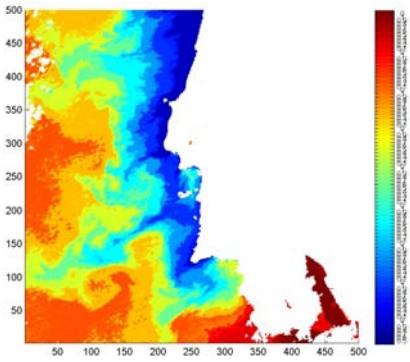


Figura F.40 19980802_10c

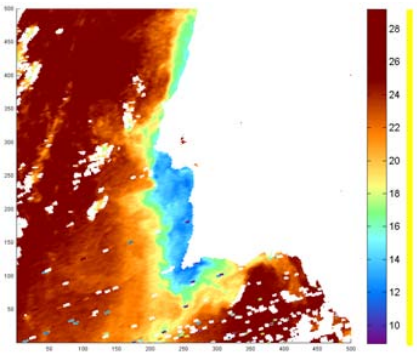


Figura F.41 19990823

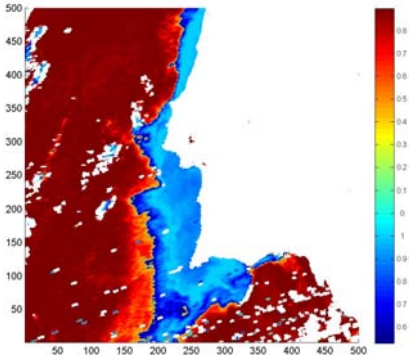


Figura F.42 19990823_2c

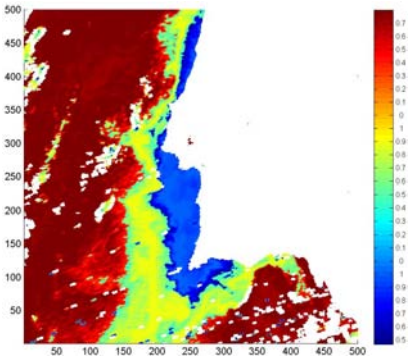


Figura F.43 19990823_3c

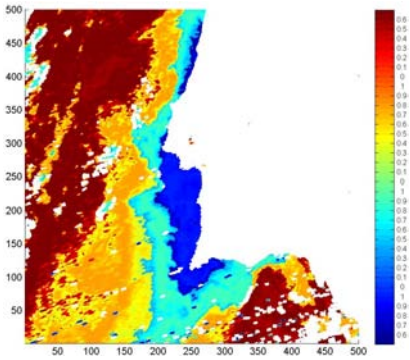


Figura F.44 19990823_4c

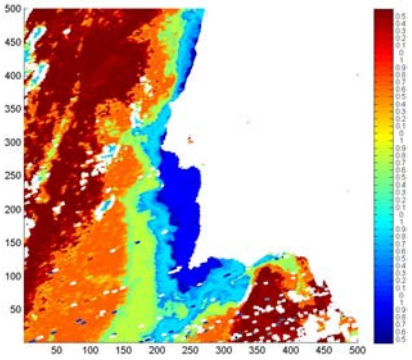


Figura F.45 19990823_5c

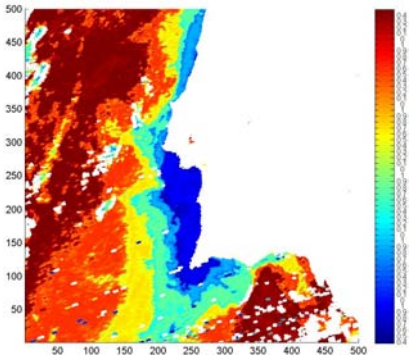


Figura F.46 19990823_6c

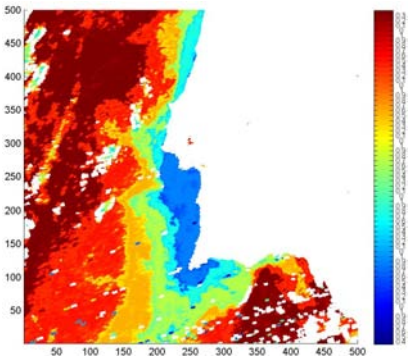


Figura F.47 19990823_7c

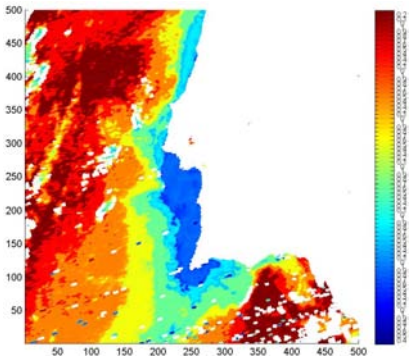


Figura F.48 19990823_8c

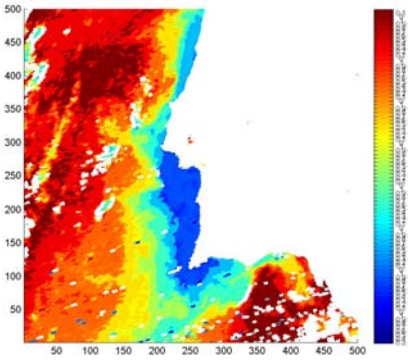


Figura F.49 19990823_9c

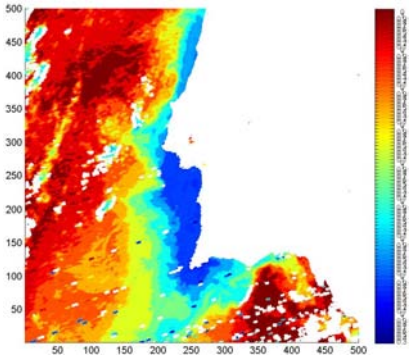


Figura F.50 19990823_10c

F.2 Visualização de fronteiras

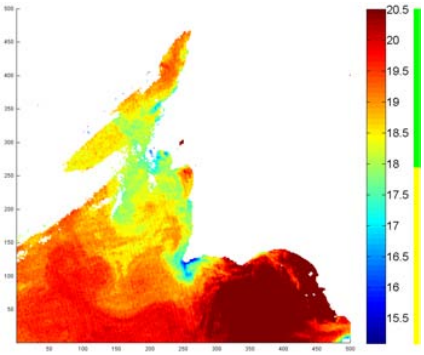


Figura F.51 19980609

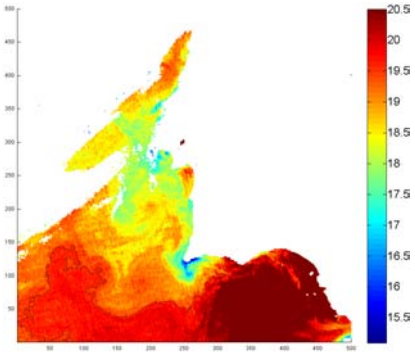


Figura F.52 19980609_2c

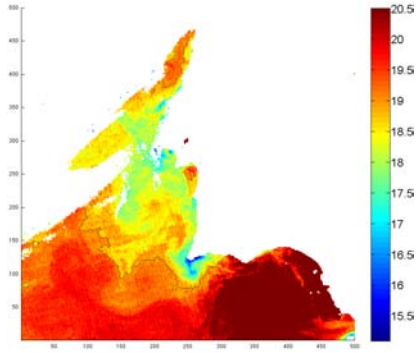


Figura F.53 19980609_3c

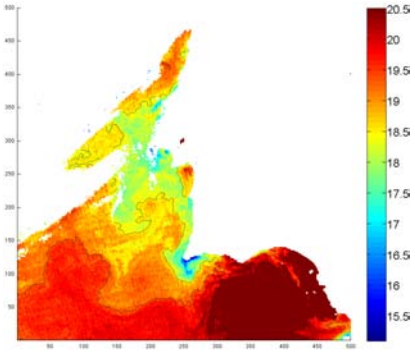


Figura F.54 19980609_4c

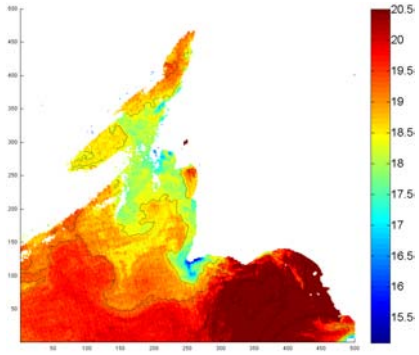


Figura F.55 19980609_5c

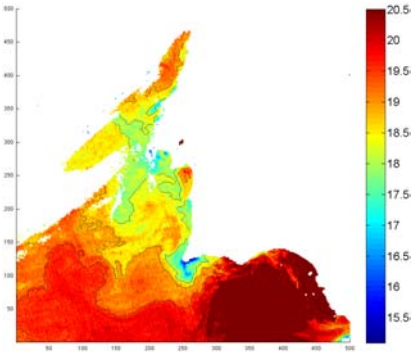


Figura F.56 19980609_6c

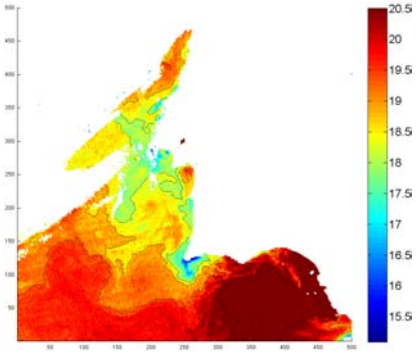


Figura F.57 19980609_7c

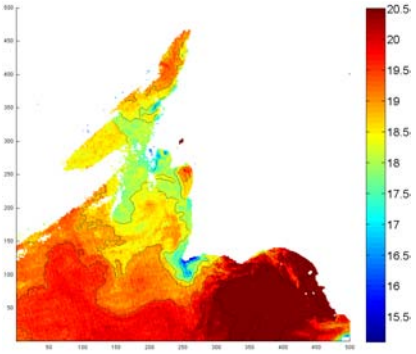


Figura F.58 19980609_8c

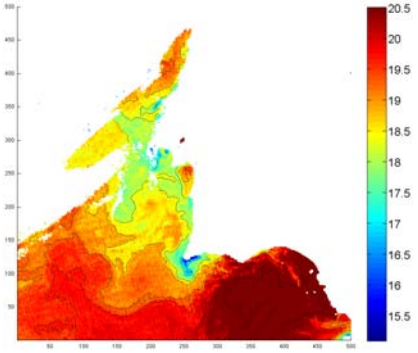


Figura F.59 19980609_9c

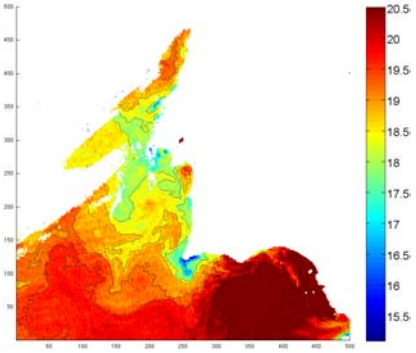


Figura F.60 19980609_10c

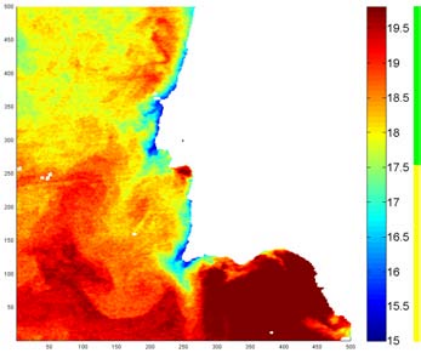


Figura F.61 19980612

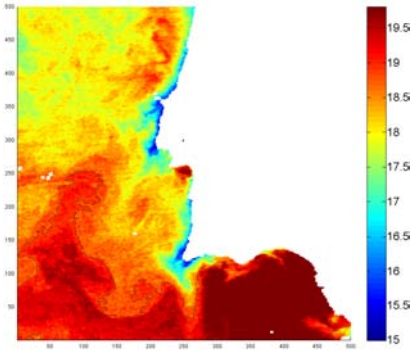


Figura F.62 19980612_2c

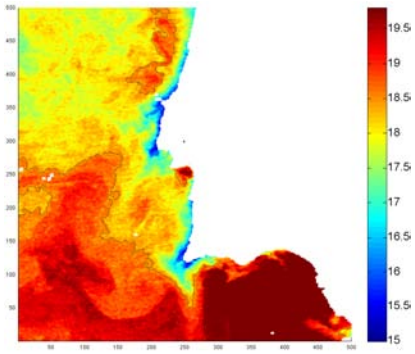


Figura F.63 19980612_3c

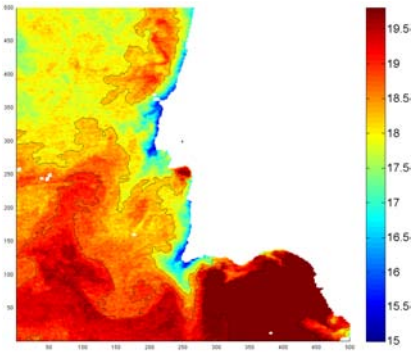


Figura F.64 19980612_4c

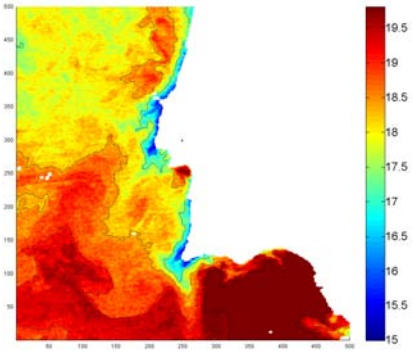


Figura F.65 19980612_5c

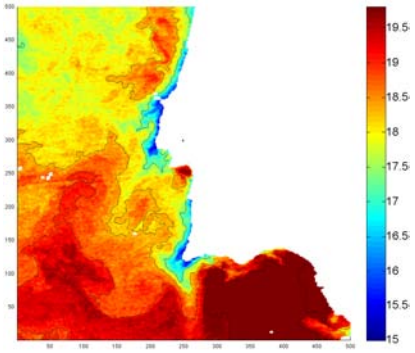


Figura F.66 19980612_6c

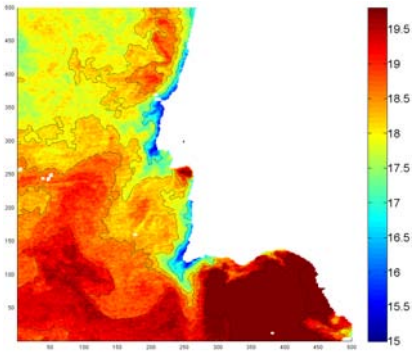


Figura F.67 19980612_7c

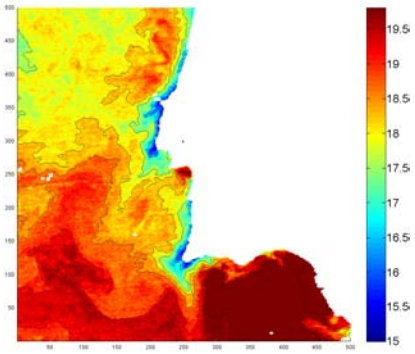


Figura F.68 19980612_8c

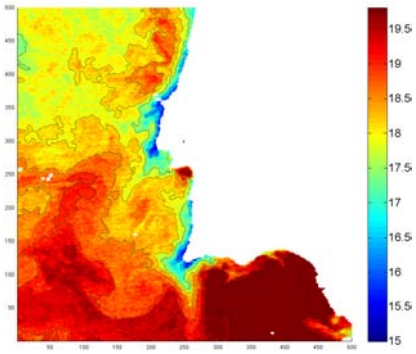


Figura F.69 19980612_9c

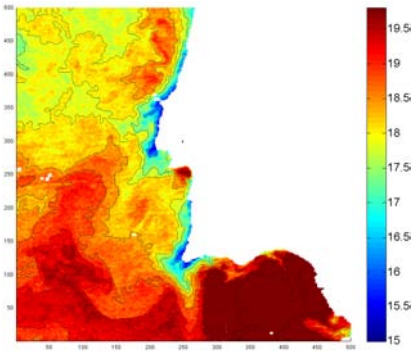


Figura F.70 19980612_10c

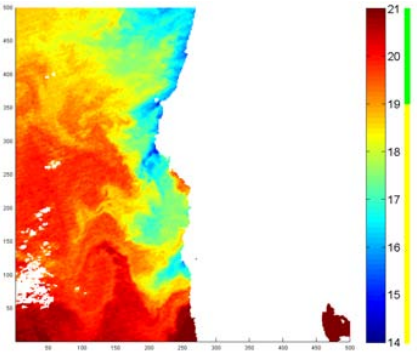


Figura F.71 19980715

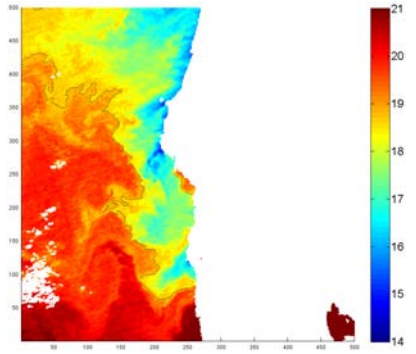


Figura F.72 19980715_2c

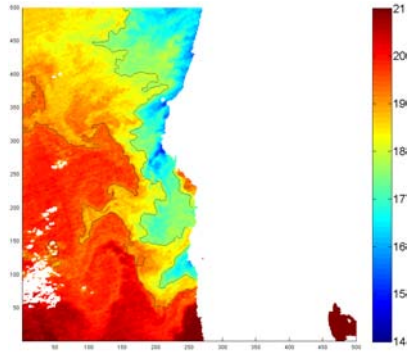


Figura F.73 19980715_3c

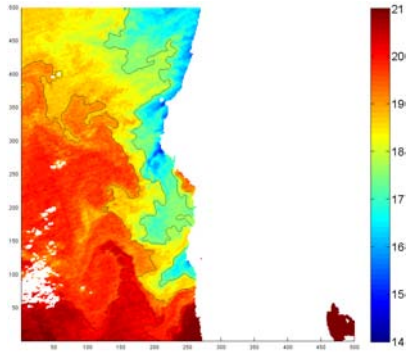


Figura F.74 19980715_4c

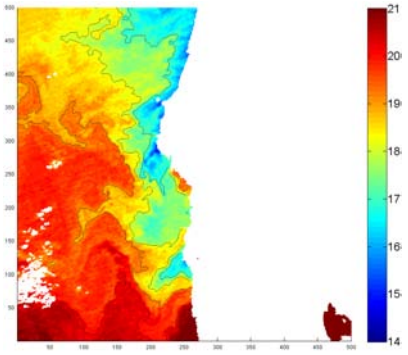


Figura F.75 19980715_5c

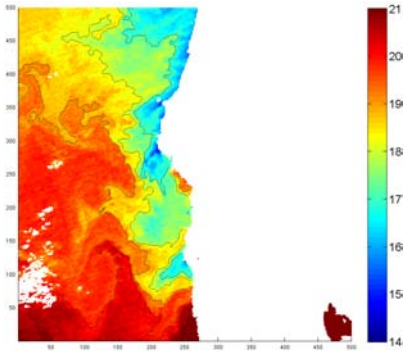


Figura F.76 19980715_6c

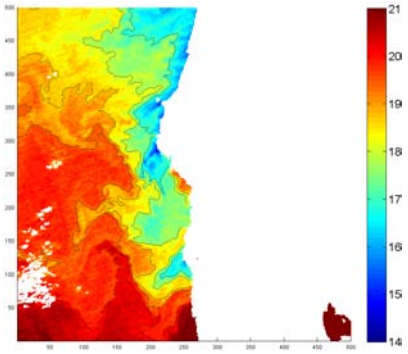


Figura F.77 19980715_7c

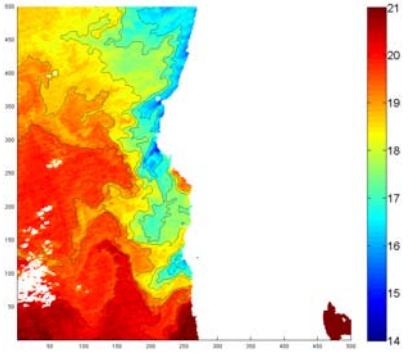


Figura F.78 19980715_8c

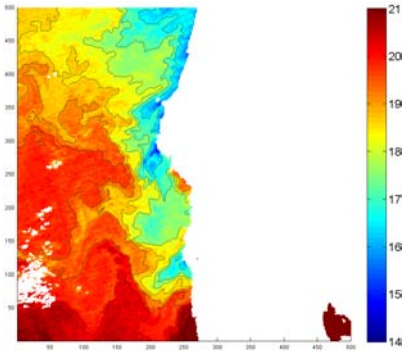


Figura F.79 19980715_9c

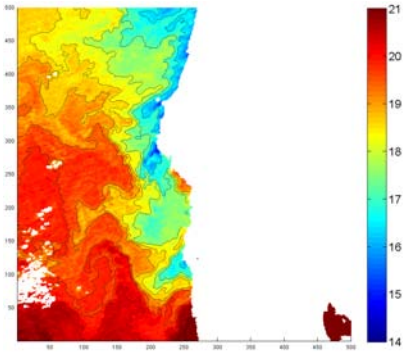


Figura F.80 19980715_10c

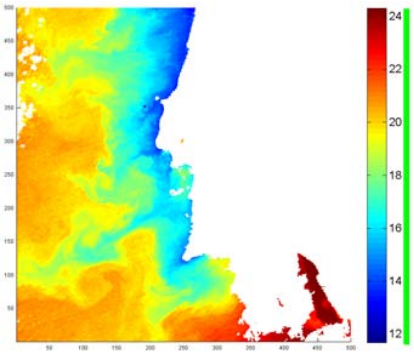


Figura F.81 19980802

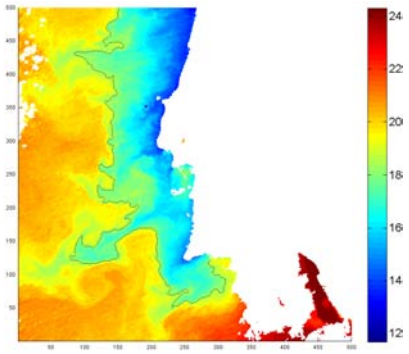


Figura F.82 19980802_2c

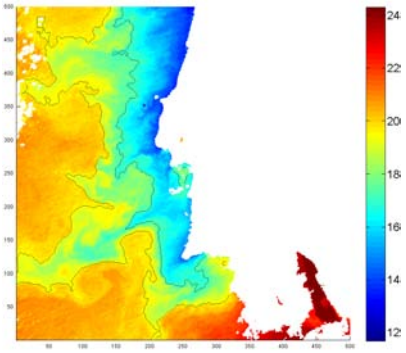


Figura F.83 19980802_3c

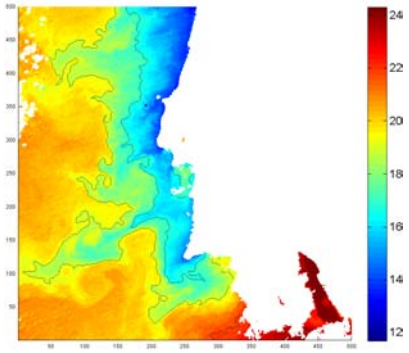


Figura F.84 19980802_4c

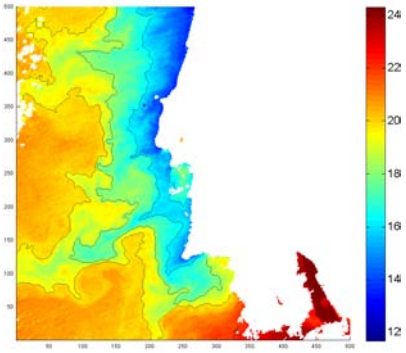


Figura F.85 19980802_5c

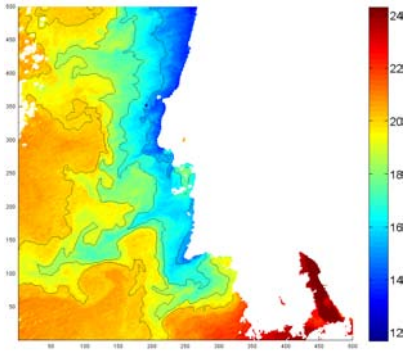


Figura F.86 19980802_6c

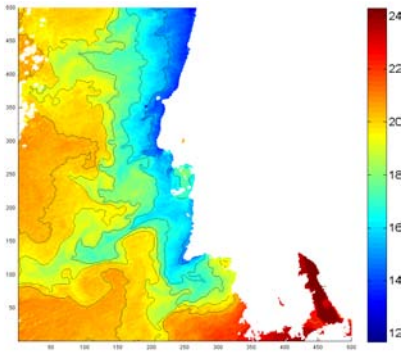


Figura F.87 19980802_7c

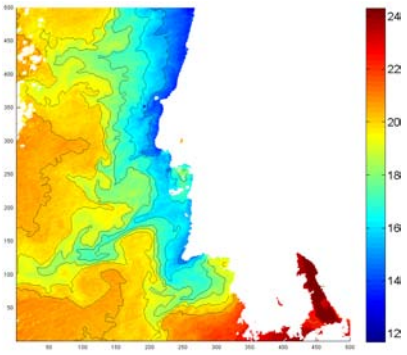


Figura F.88 19980802_8c

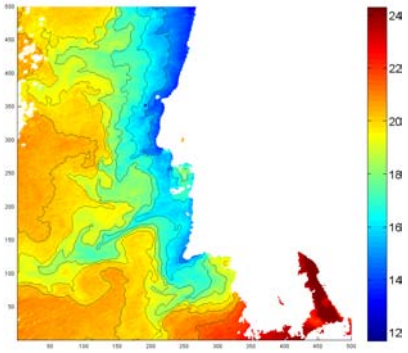


Figura F.89 19980802_9c

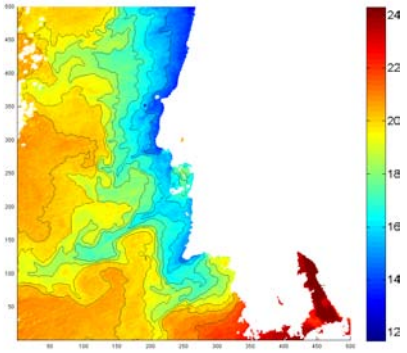


Figura F.90 19980802_10c

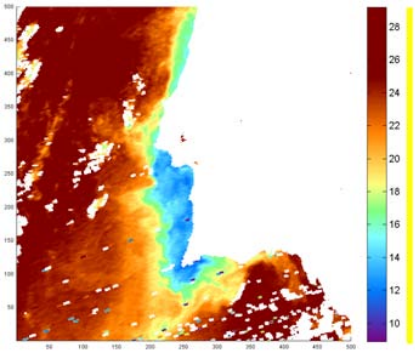


Figura F.91 19990823

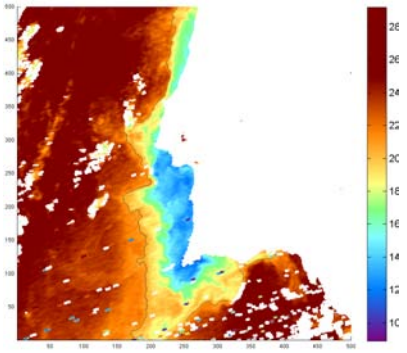


Figura F.92 19990823_2c

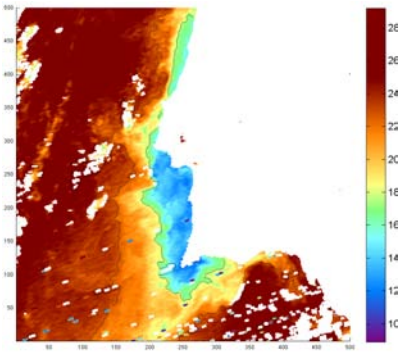


Figura F.93 19990823_3c

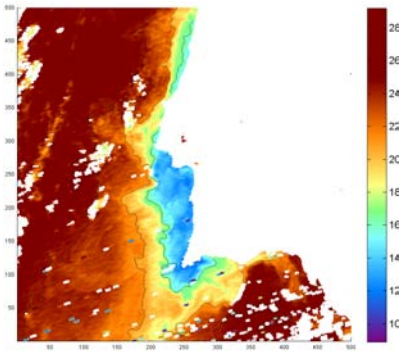


Figura F.94 19990823_4c

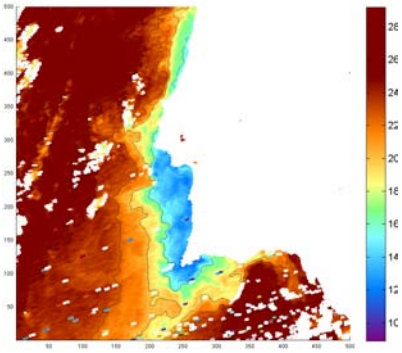


Figura F.95 19990823_5c

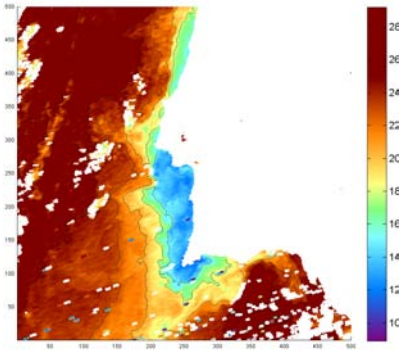


Figura F.96 19990823_6c

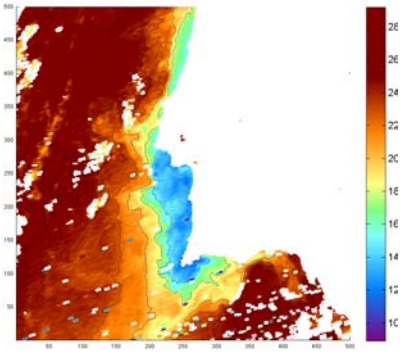


Figura F.97 19990823_7c

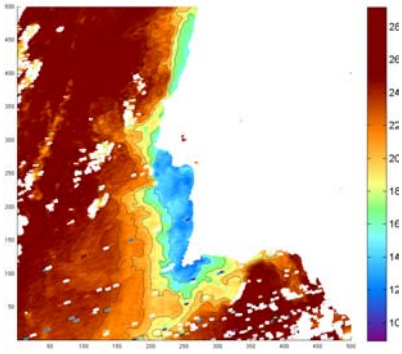


Figura F.98 19990823_8c

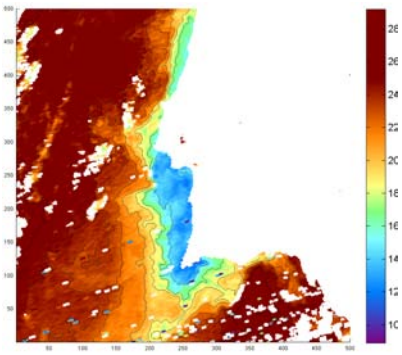


Figura F.99 19990823_9c

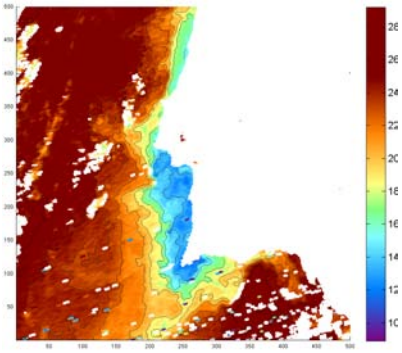


Figura F.100 19990823_10c

G . Resultados Iterative Thresholding

G.1 Visualização de fronteiras

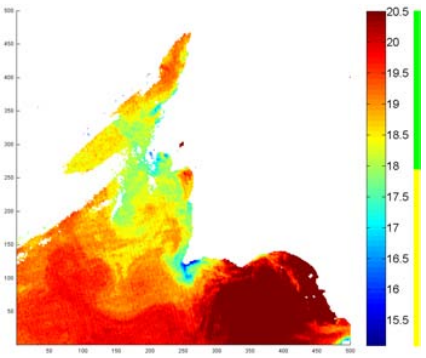


Figura G.1 19980609

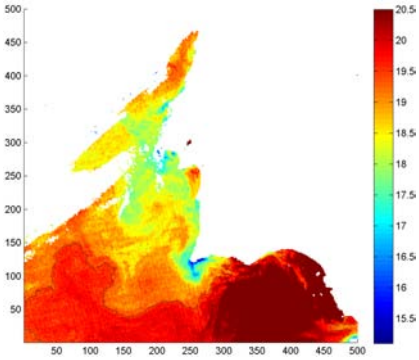


Figura G.2 19980609_2c

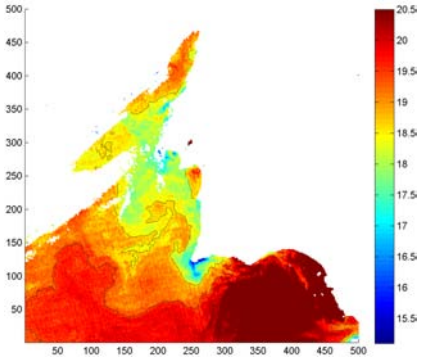


Figura G.3 19980609_3c

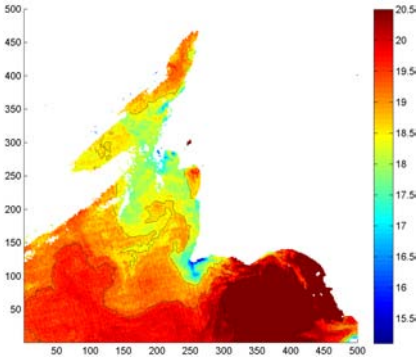


Figura G.4 19980609_4c

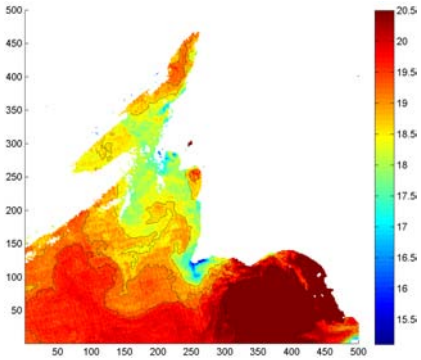


Figura G.5 19980609_5c

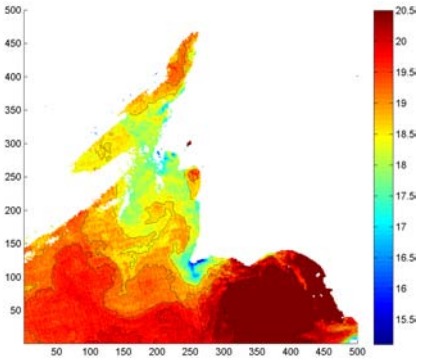


Figura G.6 19980609_6c

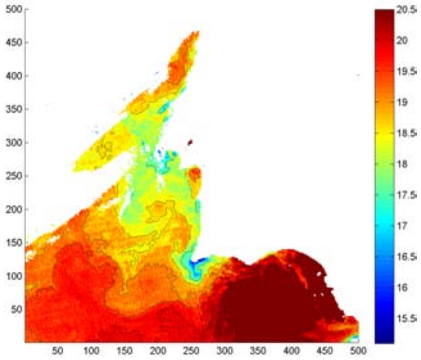


Figura G.7 19980609_7c

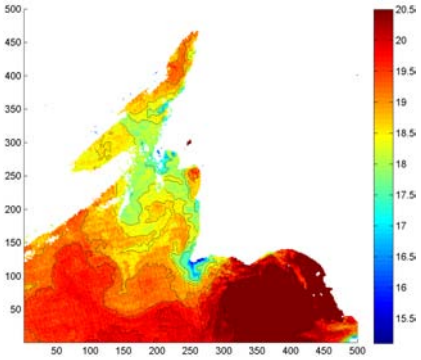


Figura G.8 19980609_8c

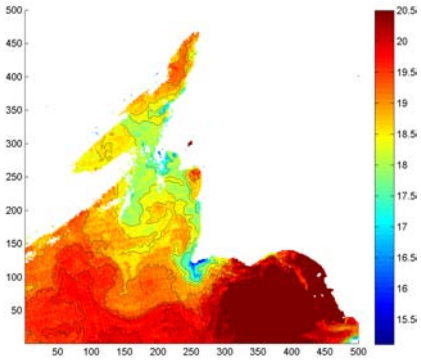


Figura G.9 19980609_9c

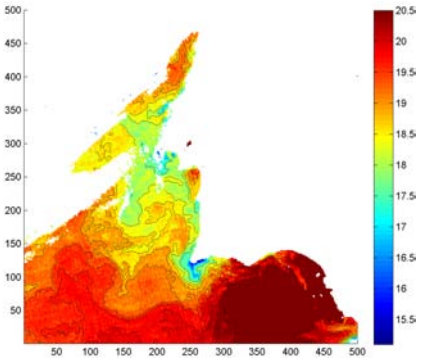


Figura G.10 19980609_10c

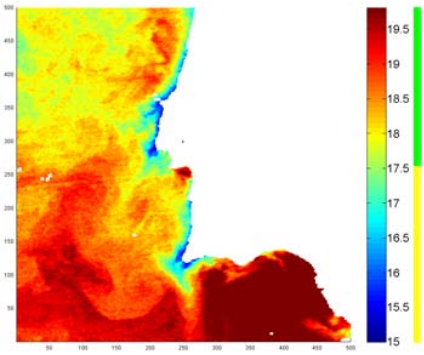


Figura G.11 19980612

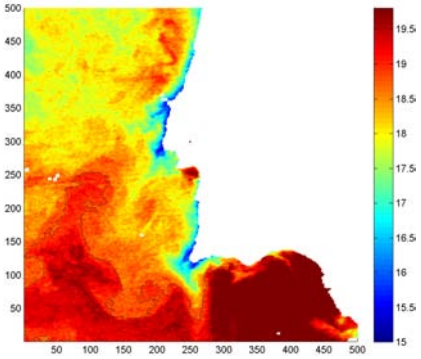


Figura G.12 19980612_2c

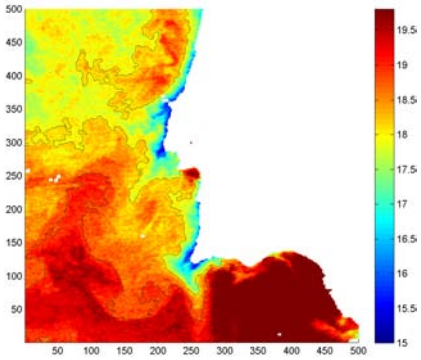


Figura G.13 19980612_3c

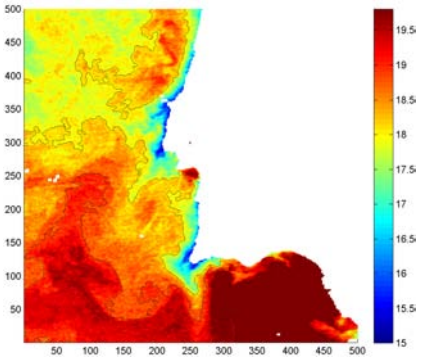


Figura G.14 19980612_4c

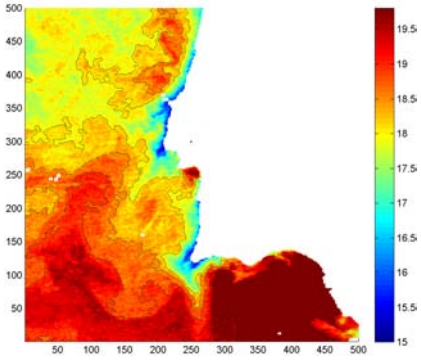


Figura G.15 19980612_5c

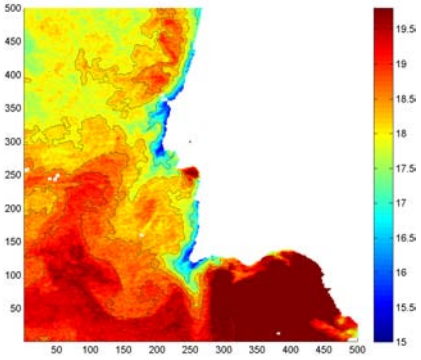


Figura G.16 19980612_6c

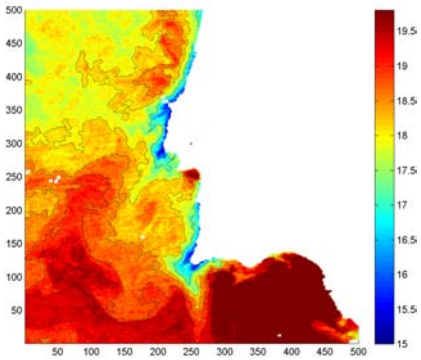


Figura G.17 19980612_7c

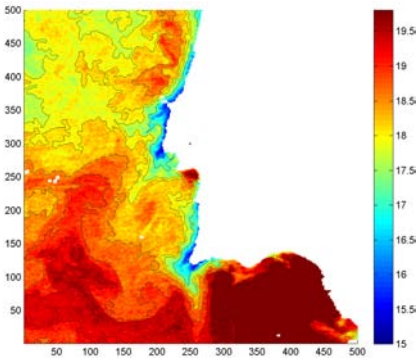


Figura G.18 19980612_8c

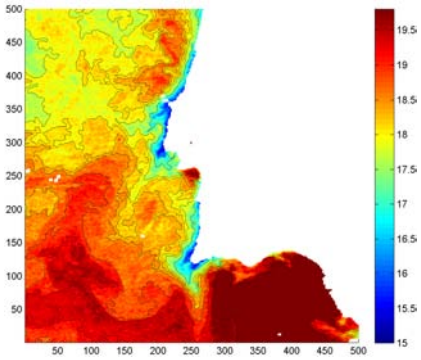


Figura G.19 19980612_9c

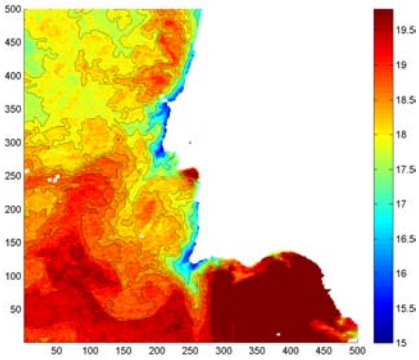


Figura G.20 19980612_10c

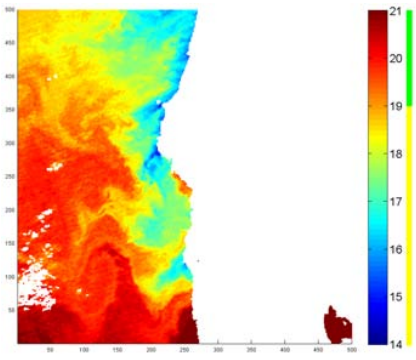


Figura G.21 19980715

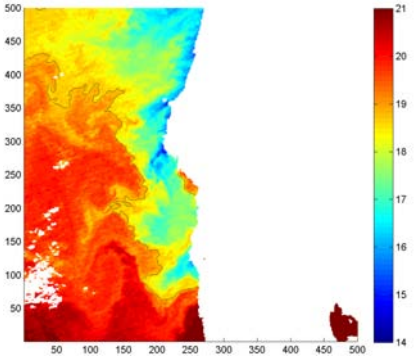


Figura G.22 19980715_2c

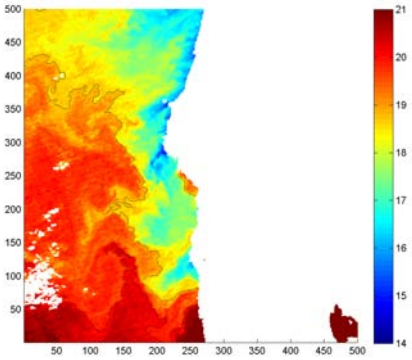


Figura G.23 19980715_3c

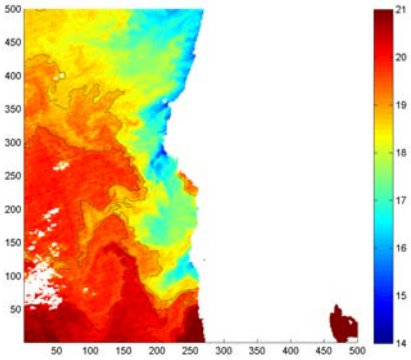


Figura G.24 19980715_4c

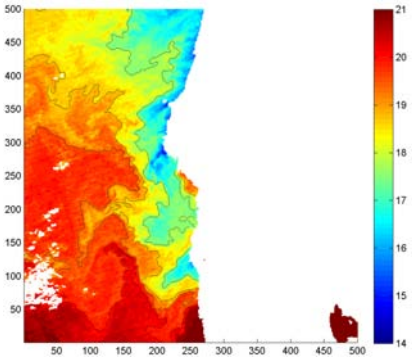


Figura G.25 19980715_5c

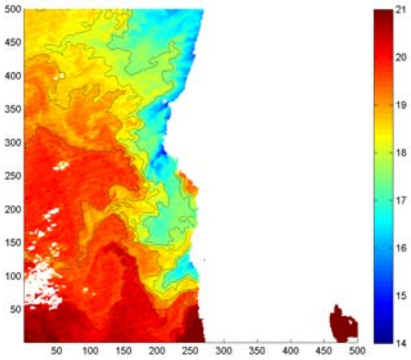


Figura G.26 19980715_6c

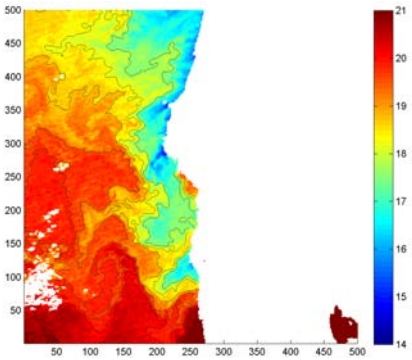


Figura G.27 19980715_7c

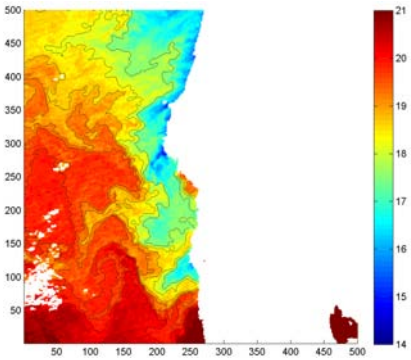


Figura G.28 19980715_8c

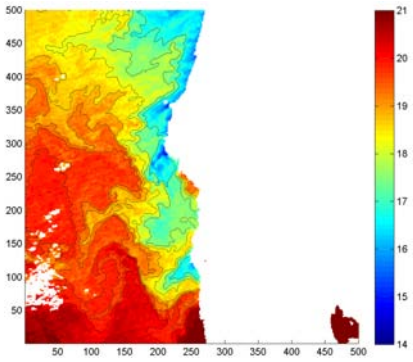


Figura G.29 19980715_9c

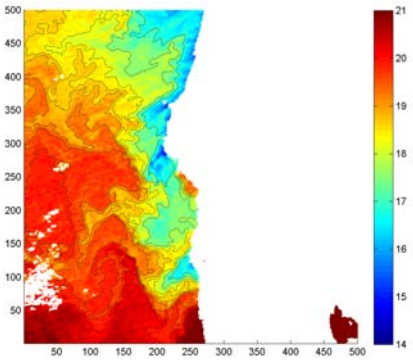


Figura G.30 19980715_10c

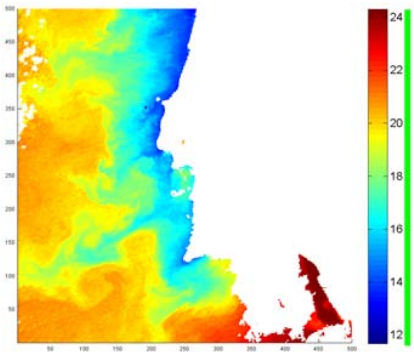


Figura G.31 19980802

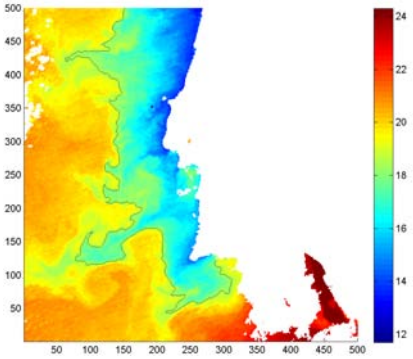


Figura G.32 19980802_2c

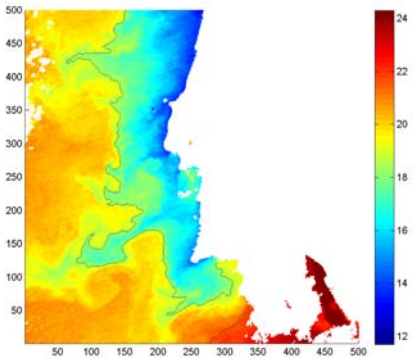


Figura G.33 19980802_3c

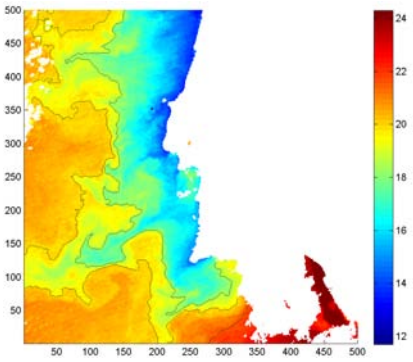


Figura G.34 19980802_4c

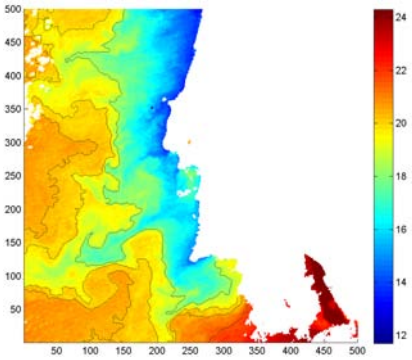


Figura G.35 19980802_5c

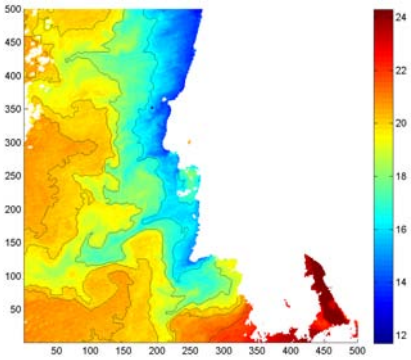


Figura G.36 19980802_6c

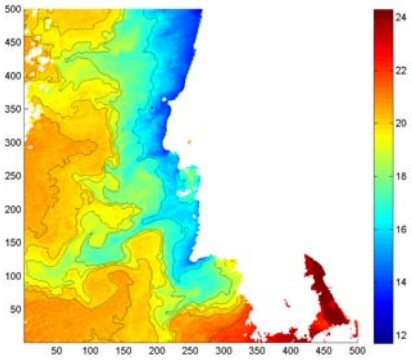


Figura G.37 19980802_7c

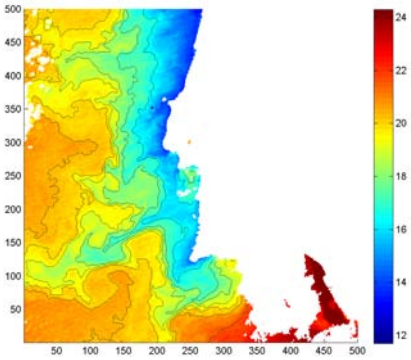


Figura G.38 19980802_8c

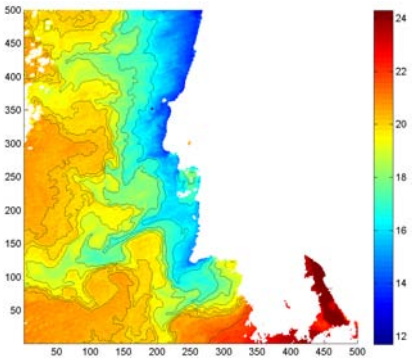


Figura G.39 19980802_9c

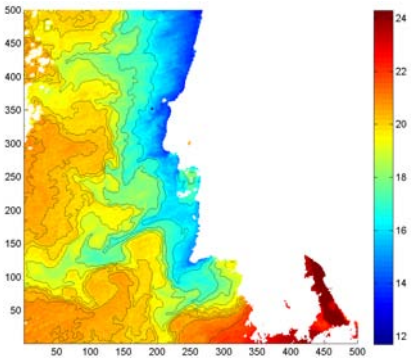


Figura G.40 19980802_10c

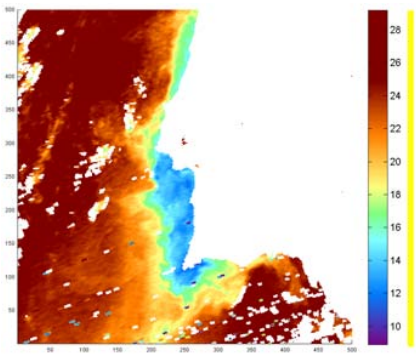


Figura G.41 19990823

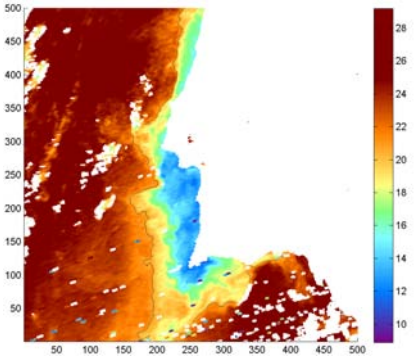


Figura G.42 19990823_2c

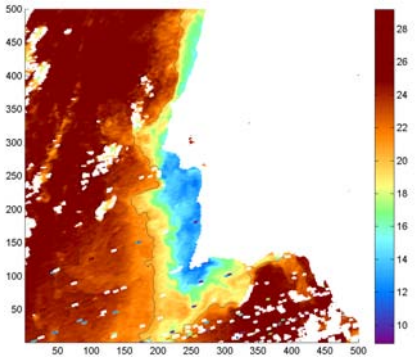


Figura G.43 19990823_3c

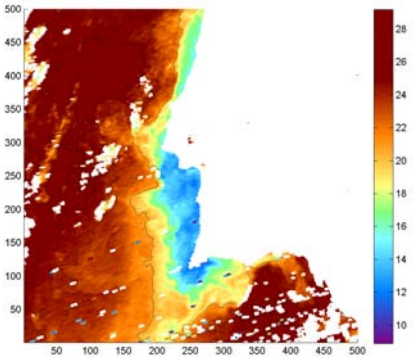


Figura G.44 19990823_4c

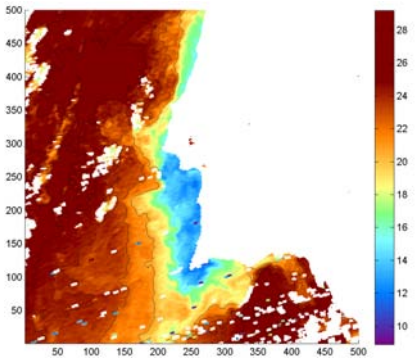


Figura G.45 19990823_5c

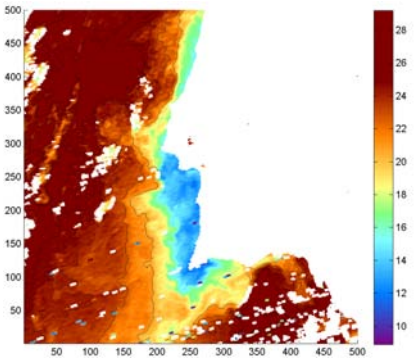


Figura G.46 19990823_6c

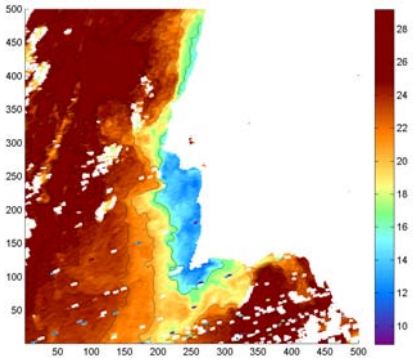


Figura G.47 19990823_7c

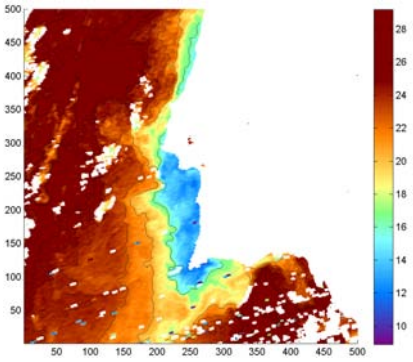


Figura G.48 19990823_8c

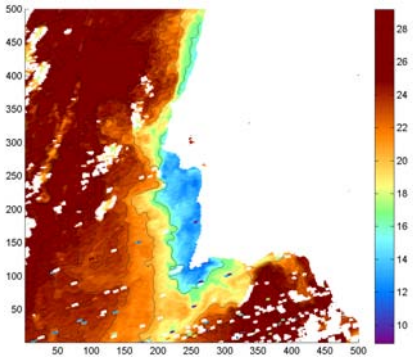


Figura G.49 19990823_9c

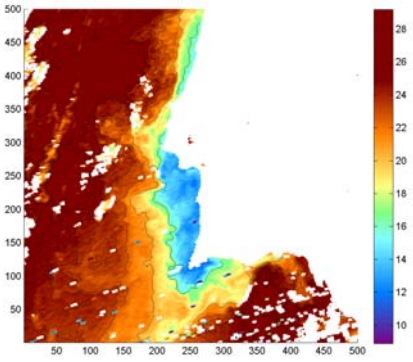


Figura G.50 19990823_10c

H . Critério de definição de fronteira

H.1 Definição de *thresholds* com base no ganho de informação

H.1.1 Análise da *feature* *CCard*

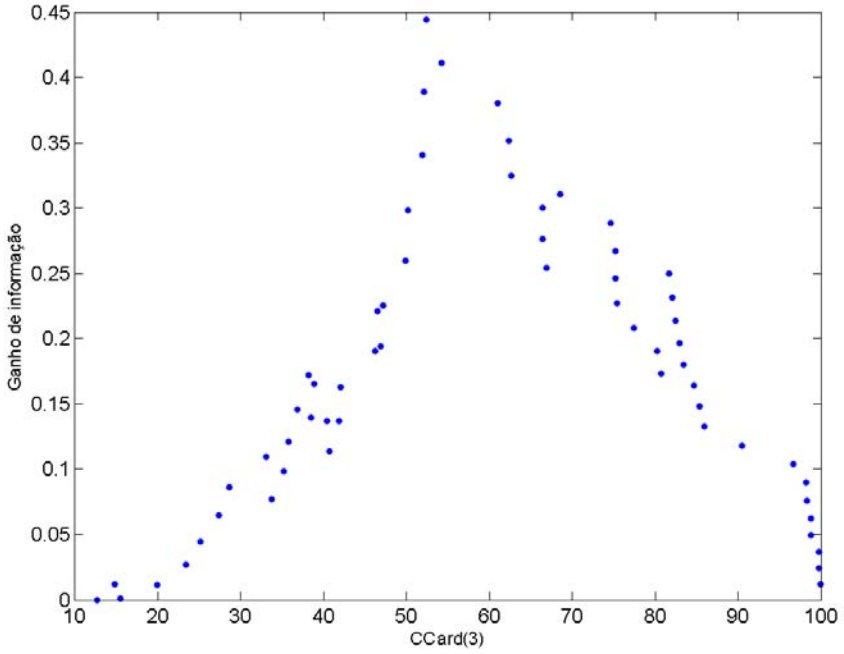


Figura H.1 Análise do ganho de informação em função dos atributos da *feature* *CCard*(3), para a Região Norte.

H.1.2 Análise da *feature* *CloudNoise*

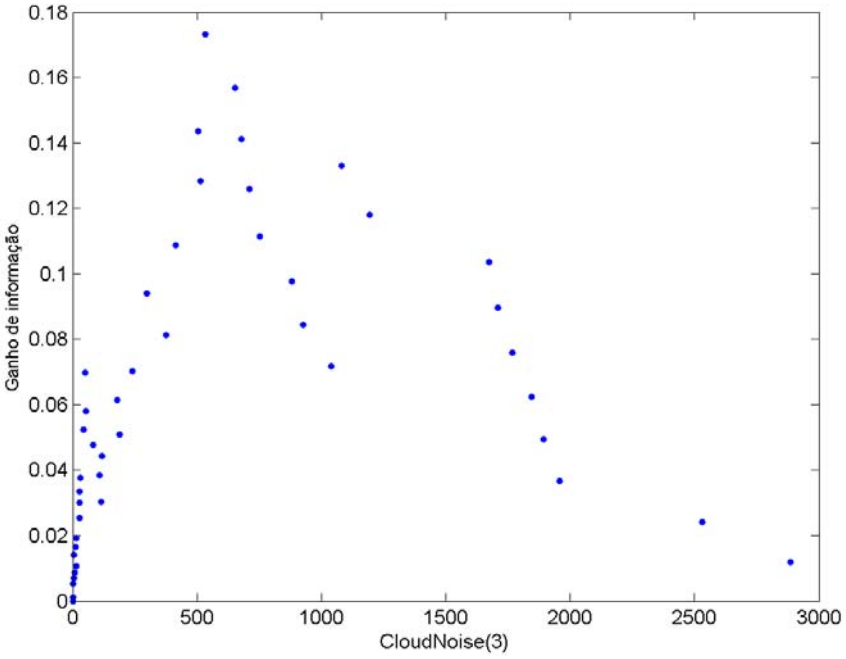


Figura H.2 Análise do ganho de informação em função dos atributos da *feature* *CloudNoise*(3), para a Região Norte.

H.2 Visualização de resultados de aplicação do critério composto [1]

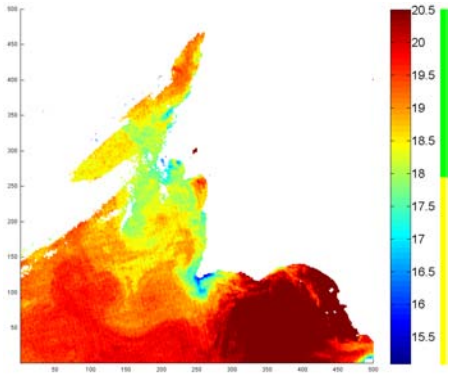


Figura H.3 19980609

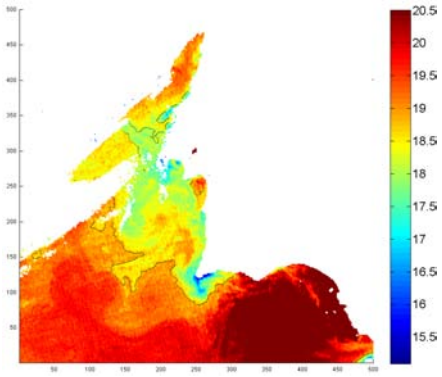


Figura H.4 19980609_7c_1N_2S

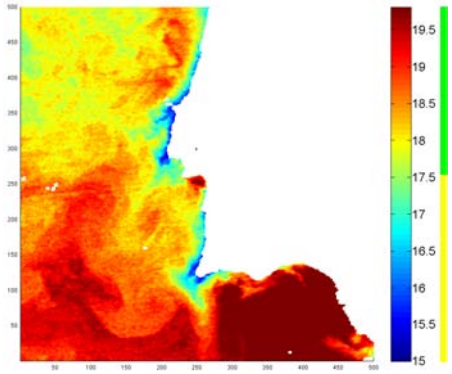


Figura H.5 19980612

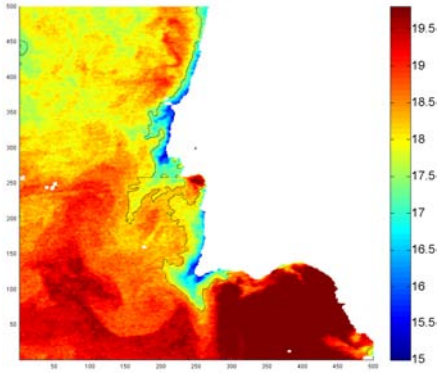


Figura H.6 19980612_6c_1N_2S

[1] A identificação de cada resultado da aplicação do critério está sob o formato 'yyyymmdd_xc_nN_sS', onde 'x' indica o número de *clusters*, e 'n', 's' indicam o número de clusters identificados como pertencentes à região de upwelling, na Região Norte e Sul, respectivamente. A fronteira visualizada corresponde à fronteira exterior do cluster 'n', para a Região Norte, e 's', para a Região Sul. Nas imagens escolhidas para a versão impressa deste relatório, não há diferença no número de clusters resultante entre a definição de *thresholds* por ganho de informação e experimentalmente, pelo que os resultados visualizados são válidos para ambos os métodos.

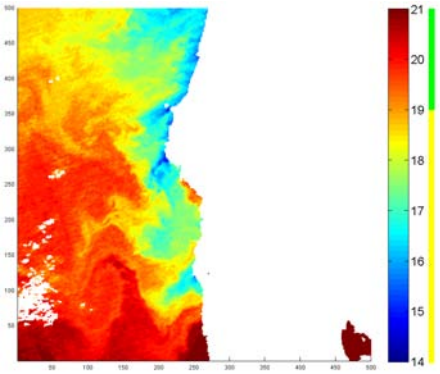


Figura H.7 19980715

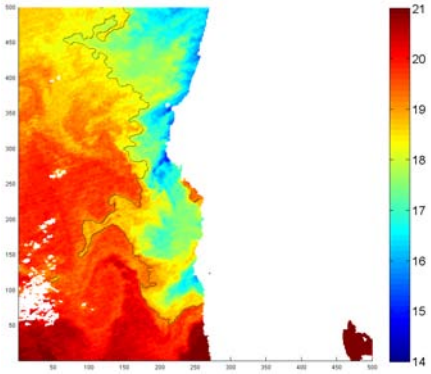


Figura H.8 19980715_5c_2N_3S

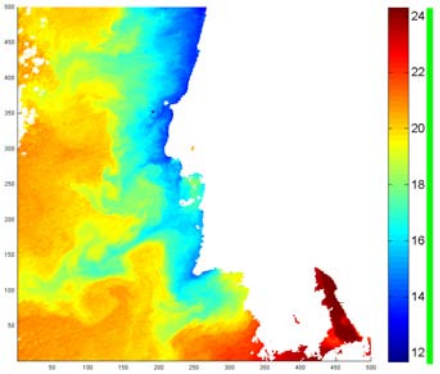


Figura H.9 19980802

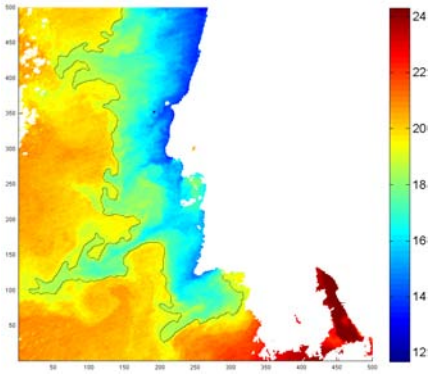


Figura H.10 19980802_6c_3N_3S

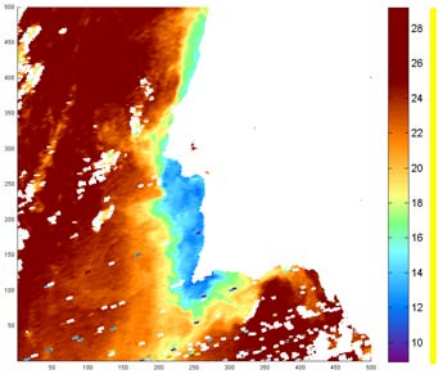


Figura H.11 19990823

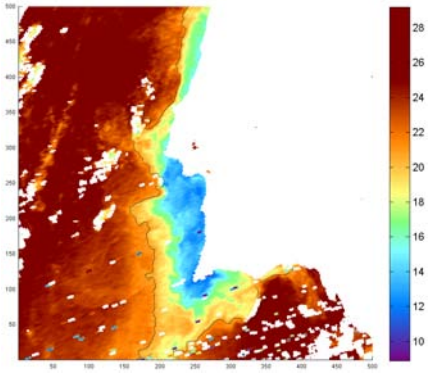


Figura H.12 19990823_6c_3N_3S

Bibliografia

[1] S. Nascimento, F.M. Sousa, H. Casimiro e D. Boutov. Applicability of fuzzy clustering for the identification of upwelling areas on sea surface temperature images. *Proceedings of the 2005 UK Workshop on Computational Intelligence*, pág. 143–148, 5-7 Setembro 2005.

[2] I. Ambar e J. Dias. Remote Sensing of Coastal Upwelling in the North-Eastern Atlantic Ocean. V. Barale e M. Gade (editores), *Remote Sensing of the European Seas (Cap. 9)*, pág. 141-152. Springer, 2008.

[3] A.M.P. Santos, M.d.F. Borges e S. Groom. Sardine and horse mackerel recruitment and upwelling off Portugal. *ICES Journal of Marine Science*, 58:589–596(8), Junho 2001. doi:10.1006/jmsc.2001.1060.

[4] M.J. Almeida e H. Queiroga. Physical forcing of onshore transport of crab megalopae in the northern portuguese upwelling system. *Estuarine, Coastal and Shelf Science*, 57:1091–1102(12), Agosto 2003. doi:10.1016/S0272-7714(03)00012-X.

[5] J. Marcello, F. Marques e F. Eugenio. Automatic tool for the precise detection of upwelling and filaments in remote sensing imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 43(7):1605–1616, 2005. doi:10.1109/TGRS.2005.848409.

[6] S.K.T. Kriebel, W. Brauer e W. Eifler. Coastal upwelling prediction with a mixture of neural networks. *Geoscience and Remote Sensing, IEEE Transactions on*, 36(5):1508–1518, 1998. doi:10.1109/36.718854.

[7] S. Plattner, D. Mason, G. Leshkevich, D. Schwab e E. Rutherford. Classifying and forecasting coastal upwellings in lake michigan using satellite derived temperature images and buoy data. *Journal of Great Lakes Research*, 32(1):63–76, 2006.

[8] J.-C. Terrillon, M. David e S. Akamatsu. Automatic detection of human faces in natural scene images by use of a skin color model and of invariant moments. *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pág. 112–117, Abril 1998. doi:10.1109/AFGR.1998.670934.

[9] Zhanqing Li, A. Khananian, R.H. Fraser e J. Cihlar. Automatic detection of fire smoke using artificial neural networks and threshold approaches applied to AVHRR imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 39(9):1859–1870, Setembro 2001. doi:10.1109/36.951076.

[10] R. Silipo e C. Marchesi. Artificial neural networks for automatic ECG analysis. *Signal Processing, IEEE Transactions on*, 46(5):1417–1425, Maio 1998. doi:10.1109/78.668803.

[11] M.S. Atkins e B.T. Mackiewich. Fully automatic segmentation of the brain in MRI. *Medical Imaging, IEEE Transactions on*, 17(1):98–107, Fevereiro 1998. doi:10.1109/42.668699.

[12] B. Mirkin. *Clustering For Data Mining: A Data Recovery Approach*. Chapman & Hall/CRC, 2005.

[13] A.K. Jain e R.C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[14] P. Tan, M. Steinbach e V. Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 12-14 Fevereiro 2005.

[15] R. Xu e D. Wunsch II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[16] P. Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[17] A.K. Jain, M.N. Murty e P.J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999. doi:10.1145/331499.331504.

[18] H. Casimiro. Relatório do projecto de final de curso de Engenharia Informática - Departamento de Informática FCT/UNL, 2005.

[19] H.G. Wilson, B. Boots e A.A. Millward. A comparison of hierarchical and partitional clustering techniques for multispectral image classification. *Geoscience and Remote Sensing Symposium, 2002. IGARSS '02. 2002 IEEE International*, 3:1624–1626 vol.3, 24-28 Junho 2002.

[20] M. Halkidi, Y. Batistakis e M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001.

[21] S. Nascimento. *Modeling Proportional Membership in Fuzzy Clustering*. IOS PRESS, 2005.

[22] J.M. Pena, J.A. Lozano e P. Larranaga. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*, 20:1027–1040(14), Outubro 1999. doi:10.1016/S0167-8655(99)00069-0.

[23] S. Bandyopadhyay. Satellite image classification using genetically guided fuzzy clustering with spatial information. *International Journal of Remote Sensing*, 26:579–593, 2005. doi:10.1080/01431160512331316432.

[24] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[25] I. Gath e A.B. Geva. Unsupervised optimal fuzzy clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):773–780, 1989. doi:10.1109/34.192473.

[26] H.J. Sun, S.R. Wang e Q.S. Jiang. FCM-based model selection algorithms for determining the number of clusters. *Pattern Recognition*, 37(10):2027–2037, 2004.

[27] S. Monti, P. Tamayo, J. Mesirov e T. Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52:91–118(28), Julho 2003.

[28] A. Ben-Hur, A. Elisseeff e I. Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pág. 6–17, 2002.

[29] X.L. Xie e G. Beni. A validity measure for fuzzy clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13(8):841–847, 1991. doi:10.1109/34.85677.

[30] W. Wang e Y. Zhang. On fuzzy cluster validity indices. *Fuzzy Sets and Systems*, 158(19):2095 – 2117, 2007. doi:10.1016/j.fss.2007.03.004.

[31] M. Pakhira, S. Bandyopadhyay e U. Maulik. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37(3):487 – 501, 2004. doi:10.1016/j.patcog.2003.06.005.

[32] N.R. Pal e J.C. Bezdek. On cluster validity for the Fuzzy c-means model. *Fuzzy Systems, IEEE Transactions on*, 3(3):370–379, 1995. doi:10.1109/91.413225.

[33] L. Lucchese e S.K. Mitra. Color image segmentation: a state-of-the-art survey. *Proceedings of the Indian National Science Academy (INSA-A)*, 67:207–221, 2001.

[34] W. Skarbek e A. Koschan. Colour image segmentation — a survey. Technical report, Institute for Technical Informatics, Technical University of Berlin, 1994.

[35] J. Kang, L. Min, Q. Luan, X. Li e J. Liu. Novel modified Fuzzy c-means algorithm with applications. *Digital Signal Process*, 2008. doi:10.1016/j.dsp.2007.11.005.

[36] S. Chen e D. Zhang. Robust image segmentation using fcm with spatial constraints based on new kernel-induced distance measure. *Systems, Man, and Cybernetics, Part B, IEEE Transactions on*, 34(4):1907–1916, 2004. doi:10.1109/TSMCB.2004.831165.

[37] Z. Yang, F. Chung e W. Shitong. Robust fuzzy clustering-based image segmentation. *Applied Soft Computing*, 2008. doi:10.1016/j.asoc.2008.03.009.

[38] U. Sangthongpraow, P. Thitimajshima e Y. Rangsanseri. Modified Fuzzy c-means for satellite image segmentation. *Proceedings in ACRS 1999*, 1999.

[39] A. Turca, E. Ocelíková e L. Madarász. Fuzzy c-means algorithms in remote sensing. *Proceedings SAMI 2003: 1st Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence*, pág. 10, 2003.

[40] K. Chuang, H. Tzeng, S. Chen, J. Wu e T Chen. Fuzzy c-means clustering with spatial information for image segmentation. *Computerized Medical Imaging and Graphics*, 30:9–15, 2006. doi:10.1016/j.compmedimag.2005.10.001.

[41] J. Shi e J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, Agosto 2000. doi:10.1109/34.868688.

[42] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. doi:10.1007/s11222-007-9033-z.

[43] M.N. Ahmed, S.M. Yamany, N. Mohamed, A.A. Farag e T. Moriarty. A modified Fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. *Medical Imaging, IEEE Transactions on*, 21(3):193–199, 2002.

[44] J. Horváth. Image segmentation using Fuzzy c-means. *Proceedings SAMI 2006: 4th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence*, pág. 144–151, 20-21 Janeiro 2006.

[45] A. Bensaid, L. Hall, J. Bezdek e L. Clarke. Partially supervised clustering for image segmentation. *Pattern Recognition*, 29(5):859 – 871, 1996. doi:10.1016/0031-3203(95)00120-4.

[46] MATLAB, ver. 7.3.0.267 (r2006b), The MathWorks Inc., 2006.

[47] T. Ridler e S. Calvard. Picture thresholding using an iterative selection method. *Systems, Man and Cybernetics, IEEE Transactions on*, 8(8):630–632, Agosto 1978. doi:10.1109/TSMC.1978.4310039.

[48] J. Han e M. Kamber. *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems, 1ª Edição)*. Morgan Kaufmann, Setembro 2000.

[49] A.-X. Zhu. Measuring uncertainty in class assignment for natural resource maps under fuzzy logic. *Photogrammetric engineering and remote sensing*, 63:1195–1202, 1997.

[50] J. Engel. The multiresolution histogram. *Metrika*, 46(1):41–57, Janeiro 1997.

[51] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

[52] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/, 2009.

[53] C. Duo, L. Xue e C. Du-Wu. An adaptive cluster validity index for the Fuzzy c-means. *International Journey of Computer Science and Network Security*, 7(2):146–156, 2007.