Applying Anomalous Cluster Approach to Spatial Clustering

Susana Nascimento and Boris Mirkin

Abstract The concept of anomalous clustering applies to finding individual clusters on a digital geography map supplied with a single feature such as brightness or temperature. An algorithm derived within the individual anomalous cluster framework extends the so-called region growing algorithms. Yet our approach differs in that the algorithm parameter values are not expert-driven but rather derived from the anomalous clustering model. This novel framework successfully applies to the issue of automatically delineating coastal upwelling from Sea Surface Temperature (SST) maps, natural phenomenon seasonally emerging in coastal waters.

Susana Nascimento
Department of Computer Science and NOVA Laboratory for Computer Science and Informatics (NOVA LINCS), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
2829-516 Caparica, Portugal, e-mail: snt@fct.unl.pt

Boris Mirkin
Department of Data Analysis and Machine Intelligence, National Research University Higher School of Economics, Moscow, Russian Federation e-mail: bmirkin@hse.ru and Department of Computer Science, Birkbeck University of London, UK e-mail: mirkin@dcs.bbk.ac.uk

1 Introduction

In a previous work [14, 15], we automated the process of delineation of upwelling regions and boundaries using a fuzzy clustering method supplemented with the anomalous cluster initialization process [13]. Yet that method operates over temperature data only, without any relation to the spatial arrangement of the pixels involved. Therefore, we decided to modify the anomalous cluster framework in such a way that it applies to the pixels spatially located on geographical maps. We apply the view that an upwelling region grows step-by-step by involving nearest cold water pixels. The process is controlled by a function expressing the similarity of pixel temperatures to those already in the region. In a self-tuning version of the algorithm the homogeneity threshold is locally derived from the approximation criterion over a window around the pixels under consideration. This window serves as a boundary regularizer.

The paper is organized in three main sections. Section 2 describes a version of the anomalous cluster model and method relevant to the task. Section 3 describes a modification of this method applicable to clustering pixels according to the sea surface temperature maps. Section 4 gives a glimpse on application of the method to real temperature map data. Conclusion outlines the contents and sketches future work.

2 Anomalous cluster model and alternating method to optimize it

Consider a set of objects P characterized by just one feature x so that, for every $p \in P$, $x(p)$ is a real representing the value of the feature at p. A subset $C \subseteq P$ will be characterized by a binary vector $z = (z_p)$ such that $z_p = 1$ if $p \in C$ and $z_p = 0$ otherwise. We find such a C that approximates the feature x as close as possible. To adjust the quantitative expression of C with z to the measurement scale of x, vector z should be supplied with an adjustable scaling coefficient λ. Also, a preliminary transformation of the x scale can be assumed by shifting the origin of x into a point of standard or norm such as, say, the mean value of x. Therefore, following Mirkin [12, 13], an approximation model can be stated as

$$y_p = \lambda z_p + e_p \tag{1}$$

where y_p are the preprocessed feature values, $z = (z_p)$ is the unknown cluster membership vector and λ is the scaling factor value, also referred to as the cluster intensity value [12, 13]. The items e_p represent errors of the model; they are to emerge because the vector λz_p may have only two different values, 0 and λ, whereas rescaled feature y may have different values. Anyway, the model requires that the errors should be made as small as possible.

Consider the least squares criterion $\Phi^2 = \sum_{p \in P} e_p^2 = \sum_{p \in P} (y_p - \lambda z_p)^2$ for fitting the model (1).

This is a more or less conventional statistics criterion. Yet in the clustering context, Φ^2 bears somewhat unconventional meaning. Indeed, any $z_p = 0$ contributes y_p^2 to Φ^2 independently of the λ value. Therefore, to minimize the criterion, $z_p = 0$ should correspond to those objects at which pre-processed feature values $y(p)$ are zero or near zero. In contrast, those objects $p \in P$ at which maximum or almost maximum absolute values of the feature hold, should be assigned with $z_p = 1$. Moreover, these must be either maximum positive values or minimum negative values but not both. Indeed, as it is well known, the optimum λ at any given C must be the average of $y(p)$ over all $p \in C$, $\lambda(C) = \sum_{p \in C} y_p / |C|$. Substituting this value into criterion Φ^2, one can easily derive the following decomposition

$$\Phi^2 = \sum_{p \in P} y_p^2 - \lambda(C)^2 |C| \tag{2}$$

or, equivalently,

$$\sum_{p \in P} y_p^2 = \lambda(C)^2 |C| + \Phi^2. \tag{3}$$

The latter expression is a Pythagorean decomposition of the data scatter (on the left) in explained and unexplained parts. The smaller the unexplained part, Φ^2, the greater the explained part,

$$g^2(C) = \lambda(C)^2 |C|. \tag{4}$$

Equation (4) gives an equivalent reformulation to the least-squares criterion: an optimal C must maximize it. This cannot be achieved by mixing in C

objects with both high positive and high negative y values because this would make the average $\lambda(C)$ smaller.

Therefore, an optimal C must correspond to either highest positive values of y or lowest negative values of y. Assume, for convenience, the latter case and consider a local search algorithm for finding a suboptimal C by adding objects one by one starting from a singleton $C = \{p\}$. What singleton? Of course that one corresponding to the lowest negative value of y, to make the value of criterion (4) as high as possible.

In publications [12, 13] only local search algorithms were considered. In these algorithms, entities are added (or removed) one by one to warrant a maximum possible local increment of the criterion until that becomes negative. Here, we develop a method which is more suitable for temperature map data. In this new method iterations are carried on in a manner similar to that of the well known clustering method k-means. Specifically, given a central value $c = \lambda(C)$, we add to cluster C all the relevant objects at once, after which the central value is recomputed and another iteration applies. The computations converge to a stable solution that cannot be improved with further iterations.

To arrive at this “batch” clustering method, let us derive a different expression for the criterion Φ^2 by “opening” parentheses in it. Specifically, since $z_p^2 = z_p$ because z_p accepts only 0 and 1 values, we may have

$$\Phi^2(C,\lambda) = \sum_{p\in P}(y_p - \lambda z_p)^2 = \sum_{p\in P} y_p^2 - 2\lambda \sum_{p\in P}(y_p - \lambda/2)z_p \tag{5}$$

As the data scatter $\sum_{p\in P} y_p^2$ is constant, minimizing (5) is equivalent to maximizing the scoring function

$$f(C,\lambda) = \sum_{p\in P} \lambda(y_p - \lambda/2)z_p = \sum_{p\in C} \lambda(y_p - \lambda/2). \tag{6}$$

This can be rewritten as

$$f(C,\lambda,\pi) = \sum_{p\in C} (\lambda y_p - \pi) \tag{7}$$

where π is a parameter that is, optimally, equals $\pi = \lambda^2/2$, and yet can be considered a user-defined threshold in criterion (7). This criterion may be considered as depending on two variables to be determined: λ and C. Therefore, the method of alternating optimization can be applied to maximize it. Each iteration of this method would consist of two steps. First, given λ, find all $p \in P$ such that $\lambda y_p > \pi$ and put them all as C. Of course, these y-values must be negative since $\lambda < 0$ in our setting. Second, given a C, find λ as the within-C average of y_p. Since both steps are optimal with respect to the corresponding variable, this method increases the value of g^2 at each step and, therefore, must converge because there are a finite number of different subsets C. In the follow-up the alternating anomalous clustering algorithm will be referred to as AA-clustering.

3 Adapting AA-clustering to the issue of delineating upwelling areas on sea surface temperature maps

Consider the set of pixels of a Sea Surface Temperature (SST) map of an ocean part as the set P, the feature x being the surface temperatures. Such maps are used in many applications of which we consider the problem of automatic delineation of coastal upwelling. This is a phenomenon that occurs when the combined effect of wind stress over the coastal oceanic waters and the Coriolis force cause these surface waters to move away from the coast. Therefore, deep, cold and nutrient-rich waters move to the surface to compensate for the mass deficiency due to this surface water circulation. As such, it has important implications in ocean dynamics and for the understanding of climate models. The identification and continuing monitoring of upwelling is an important part of oceanography.

Unfortunately the current state is far from satisfactory. Although a number of approaches for segmentation of upwelling approaches have been proposed, they suffer of too complex computational processes to get more or less satisfactory results (see, for instance, [8, 2, 10, 4, 17]).

Therefore, we decided to apply the self-tuning AA-clustering to pixels of the temperature map starting from the coldest pixel which, in fact, corresponds to the nature of upwelling.

Let $P = R \times L$ be a map under consideration where R is the set of rows and L, the columns, so that a pixel p can be presented as $p = (i, j)$ where $i \in L$ is its row coordinate, and $j \in L$, its column coordinate. Then a corresponding sea surface temperature map can be denoted as $x = (x(i, j))$, for all $i \in R$ and $j \in L$. First of all, let us center the temperature, that is, subtract the average temperature $t^* = mean(x)$ of the temperature map x from the temperature values at all pixels in $R \times L$. Let the centered values be denoted as $t(i, j)$, $(i, j) \in R \times L$. The algorithm finds a cluster $C \subseteq R \times L$ in the format of a binary map $Z(R, L)$ with elements z_{ij} defined as $z_{ij} = 1$ if $(i, j) \in C$ and $z_{ij} = 0$, otherwise. Since the pixels are not just arbitrary objects but rather elements of a spatial grid, we need to introduce this property into the AA clustering approach.

For this purpose, let us consider each pixel $p = (i, j)$ as an element of a square window of a pre-specified size $W(i, j)$ centered at p. Based on preliminary experimentation we define the window size as 7×7 (pixels). The usage of a window system appears to be useful not only as a device for maintaining continuity of the cluster C being built, but also that its boundary is of more or less smooth shape. We refer to the pixels in window $W(i, j)$ as the neighborhood of $p = (i, j)$.

The algorithm starts by selecting a seed pixel, $o = (i_o, j_o)$, as a pixel with the lowest temperature value. The cluster C is initialized as the seed $o = (i_o, j_o)$ together with pixels within the window $W(i_o, j_o)$ satisfying the similarity condition

$$c \times t(i, j) \geq \pi, \tag{8}$$

where c is the reference temperature taken at the start as the temperature of the seed pixel o, and π, a similarity threshold, according to the previous section. For convenience, let us refer to pixels in cluster C as those labelled.

Once cluster C is initialized, its boundary set F is defined as the set of such unlabelled pixels, that their neighborhood intersects the cluster. Therefore,

$$F = \{(i', j') \notin C | W(i', j') \cap C \neq \oslash\} \tag{9}$$

Then the algorithm proceeds iteratively expanding the cluster C step by step by dilating its boundary F until it is empty. For each boundary pixel

(i',j') in F we define the boundary expand region as the subset of pixels (i,j) of C that intersect the exploring window centered at pixel (i',j'), that is $(i,j) \in W(i',j') \cap C$ and define c^* as the average temperature of those pixels.

The homogeneity criterion of the algorithm is defined by the following similarity condition (10). This condition involves the reference temperature c^* equal to the mean temperature of the window pixels within the expanding region defined as $c^* = mean\,(T\,(W(i',j') \cap C))$:

$$c^* \times t(i',j') \geq \pi \tag{10}$$

Therefore, in the follow-up we take the self-tuned value for the similarity threshold as half the squared average temperature over the cluster C.

A more or less smooth shape of the growing region is warranted by the averaging nature of the similarity criterion and by involving windows around all pixels under consideration in the frontline.

This method, to which we refer to as the Seed Expanding Cluster (SEC), falls in a rather large subset of finding a segment on an image, the so-called, Seeded Region Growing (SRG) method introduced by Adams and Bischof [1] for region based segmentation (see also [11, 6, 20, 22]. The SRG method tries to grow a region whenever its interior is homogeneous according to a certain feature such as intensity, color or texture, called the *feature of interest.* The algorithm follows the strategy based on the growth of a region, starting from one or several 'seeds' and by adding similar neighboring pixels. The growth is controlled by using a homogeneity criterion so that the merging decision is generally taken based only on the contrast between the evaluated pixel and the region. However, it is not always easy to decide when this difference is small (or large) enough to make a reasonable decision.

The Seeded Region Growing image segmentation approach has been widely used in various medical image applications like magnetic resonance image analysis and unsupervised image retrieval in clinical databases [9, 23, 7, 24]. The approach has been also successfully applied in color image segmentation with applications in medical imaging, content-based image retrieval, and video [5, 21, 22], and yet in remote sensing image analysis [3, 25].

Main challenging issues that arise with SRG methods are:

(i) selection of the initial seed(s) in practical computations to get a good segmentation;

(ii) choosing the homogeneity criterion and specifying its threshold;

(iii) efficiently ordering pixels for testing whether they should be added to the region.

Most approaches of SRG involve homogeneity criteria in the format of difference of the feature of interest between that at the pixel to be labeled and the mean value at the region of interest [1, 6, 20, 22]. A weak point of these algorithms is the definition of the non-homogeneity threshold at which the pixels under consideration are considered as failing the homogeneity test and, therefore, cannot be added to the region. Such a definition is either expert driven or supervised in most of the currently available algorithms [6, 22].

Many SRG algorithms grow the regions using a sequential list sorted according to the dissimilarity of unlabeled pixels to the growth region [1, 7, 22]. The disadvantage is that the segmentation results are very much sensitive to this order.

As can be readily seen, our approach avoids these issues altogether. It utilizes a mathematically derived, though somewhat unusual, homogeneity criterion, in the format of a product rather than the conventional difference between the pixel and the mean of the region of interest. To this end, we first subtract the average temperature value from all the temperature values. This process is moderated via usage of the concept of window of a pre-specified size around the pixels under consideration: only those within the window are involved in the comparison processes. This provides for the spatial homogeneity and smoothness of the growing region. Indeed, only borderline pixels are subject to joining in because the windows around remote pixels just do not overlap the growing region. Therefore, there is no need in specifying the order of testing for labelling among pixels: all those borderline pixels can be considered and decided upon simultaneously. The process starts from a cluster consisting of just one pixel, the coldest one, according to the approximation clustering criterion. The preprocessed temperature of this pixel is negative with a relatively large absolute value. Our region growing process initializes with a fragment of the coldest pixels, which is rather robust. Moreover, the simultaneous borderline labeling considerably speeds up the SRG procedure.

In our experiments, we used also the similarity threshold π derived according to Otsu's thresholding method [18]. This method fine-tunes the similarity threshold by finding the maximum inter-class variance that splits between warm and cold waters, and is considered one of the most popular threshold method in the literature [19].

4 Experimental testing

The method developed and its Otsu's competitor have been applied to a group of 28 500×500 images showing upwelling in Portugal coastal waters; a detailed description can be found in [16]. The selected images cover distinct upwelling situations. Specifically: i) SST images with a well characterized upwelling situation in terms of fairly sharp boundaries between cold and warm surface waters measured by relatively contrasting thermal gradients and continuity along the coast (two topmost images); ii) SST images showing distinct upwelling situations related to thermal transition zones offshore from the North toward the South and with smooth transition zones between upwelling regions; iii) noisy SST images with clouds, where information to define the upwelling front lacks (fourth-line image). Figure 1 (left column) illustrate these types of situations. These images have been manually annotated by expert oceanographers regarding the upwelling regions (binary ground truth maps), which are shown in the right column of Figure 1.

Here we report of experiments on the SEC method at which the value of parameter π has been determined by either as the optimal $\lambda^2/2$ (SelfT-SEC) or by using the Otsu method (Otsu-SEC) applied to ground truth maps.

To compare the performance of various results of seed region growing algorithms, we use the popular precision, recall and their harmonic mean, the F-measure. Precision corresponds to the probability that the detection is valid, and recall is the probability that the ground truth is detected.

Overall, the segmentations are rather good, with 93% of F-scores ranging between 0.768 and 0.985. On analyzing segmentations obtained by the self-tuning threshold version of the algorithm we obtained good results in 75% of the cases. The majority of the lower value scores occur for the images with weak gradients. Figure 2 (left column) illustrates the segmentation results obtained by the self-tuning SEC algorithm for three SST images presented in Figure 1.

By comparing the relative performances of the two unsupervised thresholding versions of SEC algorithm (Otsu-SEC and SelfT-SEC), we analysed the following. The Otsu-SEC wins in 57% of the cases whereas the self-tuning version wins in 43% of images. However, let us remind that the Otsu's method uses information from the ground truth maps, whereas our method is totally unsupervised based on the image under consideration only.

The two versions of the algorithm are implemented in MatLab R2013a. The experiments have been run on a computer with a 2.67 GHz Intel(R) core(TM) i5 processor and 6 Gbytes of RAM. The operating system is Windows 8.1 Pro, 64-bit. The elapsed time of segmentation of an SST image with the Otsu's thresholding version takes 25 seconds, whereas the self-tuning version takes 22 seconds for the task.

5 Conclusion

We have proposed a new method for image segmentation combining ideas of AA-clustering and Seed region growing. This algorithm involves a novel homogeneity criterion (10), no order dependence of the pixel testing, and a version with self-tuning threshold derived from the approximation criterion.

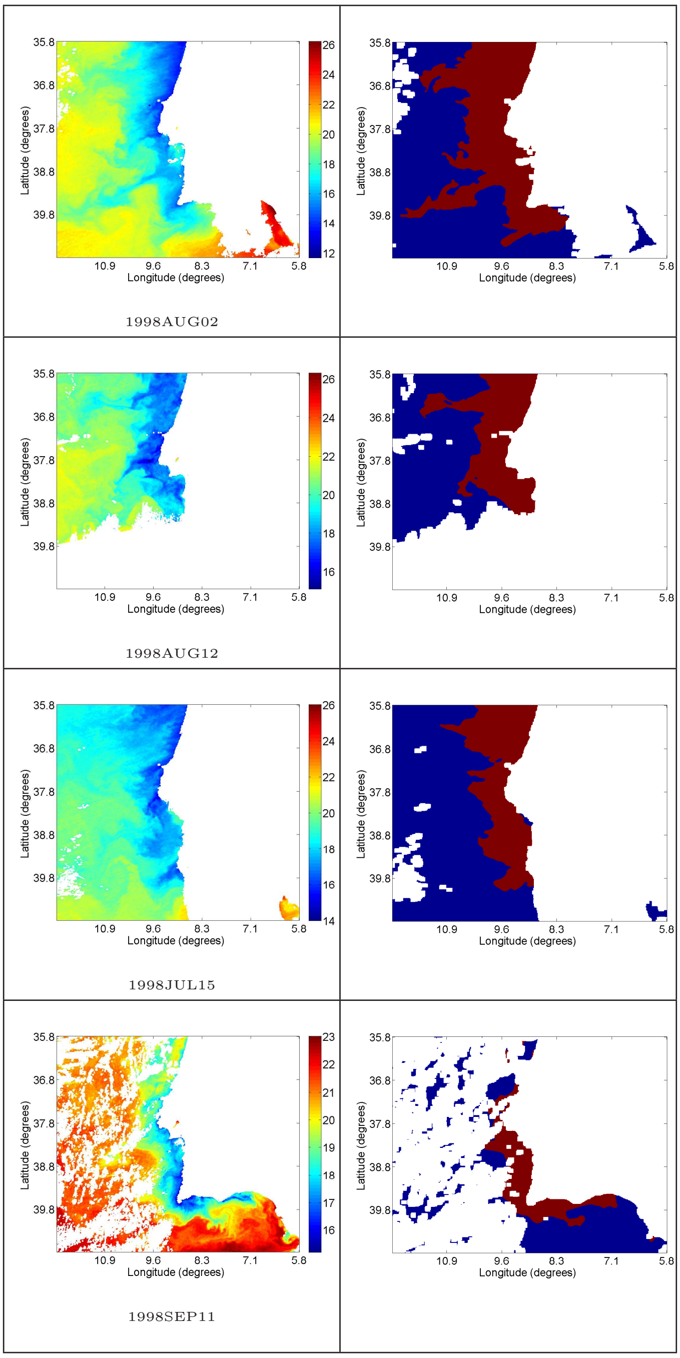
The Otsu's version of the algorithm presents very high F-measure on segmenting SST images showing different upwelling situations. The self-tuning version of the algorithm succeeds for all images presenting contrasting gradients between the coastal cold waters and the warming offshore waters of the upwelling region, and in some images with weak gradients for upwelling.

Further research should be directed toward both extending of the SEC algorithm to the situations with many clusters and applying it to other image segmentation problems.

References

1. Adams, R., Bischof, L.: Seeded region growing. IEEE Transactions on Pattern Analasys and Machine Intelligence **16**, 641-647 (1994)
2. Arriaza, J., Rojas, F., Lopez, M., Canton, M.: Competitive neural-net-based system for the automatic detection of oceanic mesoscalar structures on AVHRR scenes. IEEE Transactions on Geoscience and Remote Sensing **41**(4), 845 - 85 (2003)
3. Byun, Y., Kim, D., Lee, J., Kim, Y.: A framework for the segmentation of high-resolution satellite imagery using modified seeded-region growing and region merging. International Journal of Remote Sensing **32**(16), 4589-4609 (2011)
4. Chaudhari, S., Balasubramanian, R., Gangopadhyay, A.: Upwelling Detection in AVHRR Sea Surface Temperature (SST) Images using Neural-Network Framework. 2008 IEEE International Geoscience & Remote Sensing Symposium II, 926-929 (2008)
5. Fan, J., Yau, D.K.Y., Elmagarmid, A.K., Aref, W.G.: Automatic image segmentation by integrating color-based extraction and seeded region growing. IEEE Transactions on Image Processing **10**(10), 1454-1466 (2001)
6. Fan, J., Zeng, G., Body, M., Hacid, M.-S.: Seeded region growing: an extensive and comparative study. Pattern Recognition Letters **26**(8), 1139-1156 (2005)
7. Harikrishna-Rai, G.N., Gopalakrishnan-Nair, T.R.: Gradient Based Seeded Region Grow method for CT Angiographic Image Segmentation. International Journal of Computer Science and Networking **1**(1), 1-6 (2010)
8. Kriebel, S. T., Brauer, W., Eifler, W.: Coastal upwelling prediction with a mixture of neural networks. IEEE Transactions on Geoscience and Remote Sensing **36**(5), 1508-1518 (1998)
9. Mancas, M., Gosselin, B., Macq, B.: Segmentation Using a Region Growing Thresholding. Proc. SPIE 5672, Image Processing: Algorithms and Systems IV 388 (2005)
10. Marcello, J., Marques, F., Eugenio, F.: Automatic tool for the precise detection of upwelling and filaments in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **43**(7), 1605-1616 (2005)
11. Mehnert, A., Jackway, P.: An improved seeded region growing algorithm. Pattern Recognition Letters **18**(10), 1065-1071 (1997)
12. Mirkin, B.: A sequential fitting procedure for linear data analysis models. Journal of Classification **7**, 167-195 (1990)
13. Mirkin, B.: Clustering: A Data Recovery Approach, 2nd Edition, Chapman and Hall, Boca Raton (2012)
14. Nascimento, S., Franco, P.: Segmentation of upwelling regions in sea surface temperature images via unsupervised fuzzy clustering. In: Corchado, E. and Yin, H. (Eds.), Procs. Intelligent Data Engineering and Automated Learning (IDEAL 2009), LNCS 5788, Springer-Verlag, 543-553 (2009)

15. Nascimento, S., Franco, P., Sousa, F., Dias, J., Neves, F.: Automated computational delimitation of SST upwelling areas using fuzzy clustering, Computers & Geosciences **43**, 207-216 (2012)
16. Nascimento, S., Casca, S., Mirkin, B.: A Seed Expanding Cluster Algorithm for Deriving Upwelling Areas on Sea Surface Temperature Images Computers & Geociences Special issue on "Statistical learning in geoscience modelling: novel algorithms and challenging case studies" (2015) (in press)
17. Nieto, K., Demarcq, H., McClatchie, S.: Mesoscale frontal structures in the Canary Upwelling System: New front and filament detection algorithms applied to spatial and temporal patterns. Remote Sensing of Environment **123**, 339-346 (2012)
18. Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on System, Man, and Cybernetics SMC-**9**(1), 62-66 (1979)
19. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging **13**(1), 146-168 (2004)
20. Shih, F., Cheng, S.: Automatic seeded region growing for color image segmentation. Image and Vision Computing, **23**, 877-886 (2005)
21. Ugarriza, L. G., Saber, E., Vantaram, S.R., Amuso, V., Shaw, M., Bhaskar, R.: Automatic Image Segmentation by Dynamic Region Growth and Multiresolution Merging. IEEE Transactions on Image Processing **18**(10), 2275 - 2288 (2009)
22. Verma, O., Hanmandlu, M., Seba, S., Kulkarni, M., Jain, P.: A Simple Single Seeded Region Growing Algorithm for Color Image Segmentation using Adaptive Thresholding, Procs. of the 2011 International Conference on Communication Systems and Network Technologies, IEEE Computer Society, Washington, DC, USA, pp. 500-503 (2011)
23. Wu, J., Poehlman, S., Noseworthy, M. D., Kamath, M.: Texture feature based automated seeded region growing in abdominal MRI segmentation. Journal of Biomedical Science and Engineering **2**, 1-8 (2009)
24. Zanaty, E. A.: Improved region growing method for magnetic resonance images (MRIs) segmentation. American Journal of Remote Sensing **1**(2), 53-60 (2013)
25. Zhang, T., Yang, X., Hu, S., Su, F.: Extraction of Coastline in Aquaculture Coast from Multispectral Remote Sensing Images: Object-Based Region Growing Integrating Edge Detection. Remote Sensing **5**(9), 44704487 (2013)



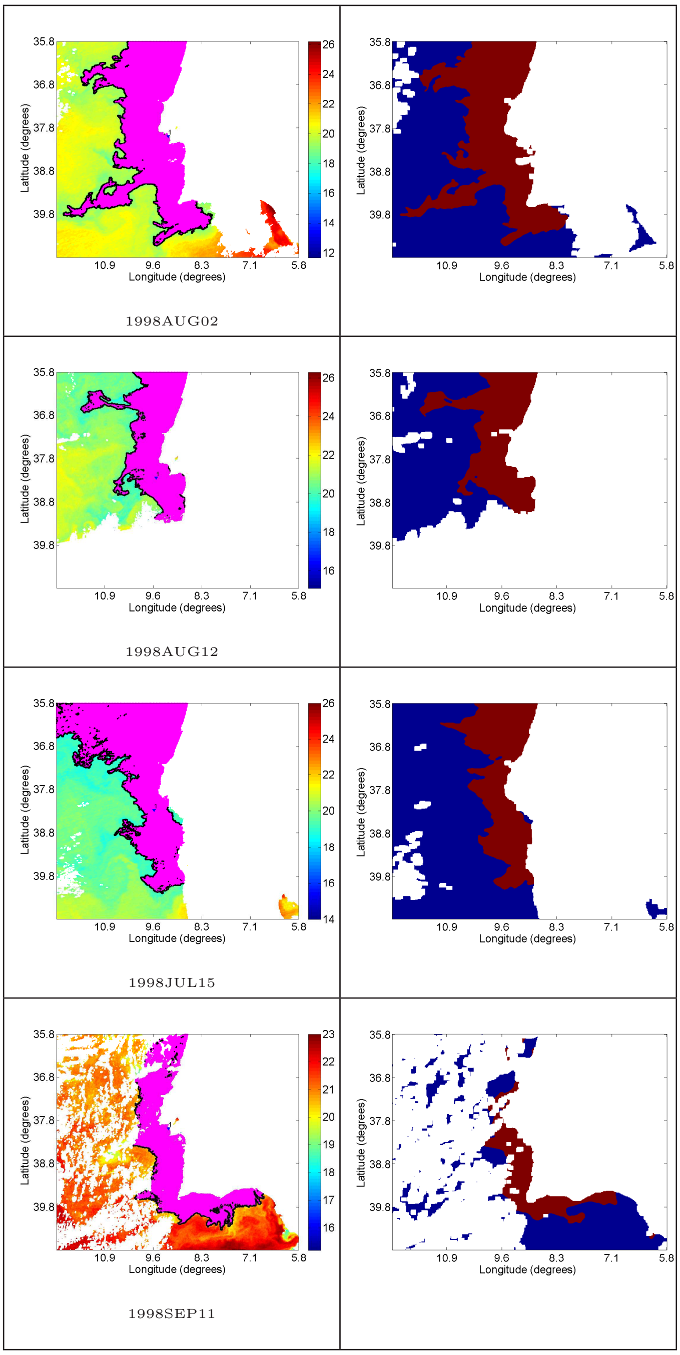
1998AUG02

1998AUG12

1998JUL15

1998SEP11

Fig. 1: Four SST images of Portugal showing different upwelling situations (left column); corresponding binary ground-truth maps (right column).



1998AUG02

1998AUG12

1998JUL15

1998SEP11

Fig. 2: Upwelling areas found by the self-tuning version of SEC algorithm on SST images of Portugal and her coastal waters(left column) versus the binary ground-truth maps (right column).