Final Draft
of the original manuscript:

Schiller, H.; Schoenfeld, W.; Krasemann, H.; Schiller, K.:
**Novelty Detection – Recognition and Evaluation of Exceptional Water
Reflectance Spectra**
In: Environmental Monitoring and Assessment (2007) Springer

DOI: 10.1007/s10661-006-9538-5

Novelty Detection: Detection and Evaluation of Exceptional Reflectance Spectra

Helmut Schiller, Wolfgang Schönfeld, Hansjörg L. Krasemann and
Kathrin Schiller

GKSS Forschungszentrum, PF 1160, 21494 Geesthacht, Germany

(author for correspondence, e-mail: schiller@gkss.de)

Abstract. One of the aims of monitoring of coastal waters is the detection of special plankton blooms since they are an important water quality indicator and are sometimes harmful to the ecosystem and maricultures. Such events can be remotely detected by multi-spectral sensors aboard an environmental satellite like ENVISAT since they cause the occurrence of exceptional water colour manifesting in exceptional water reflectance spectra.

To detect such events we first produce a tessellation of ‘normal’ water reflectances from a large set of reflectance data. In the second step exceptional spectra are identified as those spectra which have a large distance from this ‘normal’ reflectance spectra tessellation.

Keywords: novelty search, monitoring, water quality, reflectance, remote sensing

© 2006 Kluwer Academic Publishers. Printed in the Netherlands.

Поди туда – не знаю куда,

принеси то – не знаю что.

Go I know not where,

bring back I know not what.

Russian Fairy Tale

1. Introduction

This order given to the hero of the quoted fairy tale can be understood as a summary of the request to detect exceptional reflectance spectra of coastal water. Here ‘exceptional’ means that from operationally processed multi-spectral satellite data unusual events have to be detected. Of course, neither time and place of such events nor their spectral characteristic are known beforehand. What is sought for are deviations from what is occurring ‘normally’.

Out of range spectra can be caused e.g. by exceptional plankton blooms which may form so called Red Tides (due to the red discolouration of the water) or white water by coccolithophorides. It is of high interest to detect and map the temporal and spatial extensions of these special blooms since they are an important water quality indicator and are sometimes harmful to the ecosystem and maricultures.

Generally speaking the above task is a task of novelty detection: after training a learning system identifies new or unknown data – data different from that presented during the training. Novelty detection is studied in signal processing. For novelty detection from multidimensional data statistical approaches as well as Neural Network based approaches are known. A recent review of available techniques is presented by (Markou 2003), which also shows that there is a great variety of application fields for novelty detection. Examples for novelty detection in the area of analysis of remotely sensed data are (Auguststeijn 2002) and (D’Alimonte 2003).

In (Auguststeijn 2002) two Neural Net (NN) classification schemes are compared w.r.t. their ability to identify novel patterns in Thematic Mapper (TM) imagery of United States Air Force Academy grounds, north of Colorado Springs acquired in the spring of 1993. Even the time and spatial range is rather restricted the variability of the ground is extremely high. In conclusion the Probabilistic Neural Network (PNN) (Washburne 1993) is significantly more successful in marking new ground covers than back-propagation NN.

In (D’Alimonte 2003) novelty detection is used to identify the range of applicability of empirical ocean colour algorithms. Two empirical ocean colour algorithms are constructed on the basis of their respective training dataset. The accuracy of the output of a given algorithm de-

depends on the representativeness of the input in the respective training dataset. To estimate the representativeness it turns out that a multivariate Gaussian function accurately enough models the distribution of the logarithm of the three remote sensing reflectances (wavelengths $\lambda = 490, 555$ and 665 nm) used in the algorithm.

The selection of an appropriate algorithm for novelty detection is intimately connected with the type and structure of the data to be tackled. For the problem at hand this issue is discussed in section 2. In section 3 results are presented which were obtained from North Sea data. Conclusions are presented in section 4.

2. Choice of the algorithm

The data for which we want to develop a novelty detection scheme are produced by remote sensing. The sensor and some important aspects of the reflectance spectra are described in subsection 2.1. The main characteristics of the data are used to construct two-dimensional exemplary data set. The performance of different novelty detection schemes with this exemplary data set is shown in subsection 2.2. It turns out that an appropriate novelty detection scheme can be based on a set of Gaussians.

2.1. MERIS WATER REFLECTANCE SPECTRA

In March 2002, the European Space Agency launched ENVISAT, an advanced polar-orbiting earth observation satellite that provides measurements of the atmosphere, ocean, land, and ice over a five year period. The ENVISAT satellite has an ambitious and innovative payload that will ensure the continuity of the measurements of the ESA ERS satellites. The ENVISAT data support earth science research and allow monitoring of the evolution of environmental and climatic changes.

One of the instruments on board ENVISAT is MERIS (Rast 1999), the MEdium Resolution Imaging Spectrometer. The primary mission of MERIS is the measurement of water colour in the oceans and in coastal areas. Such measurements of the water colour, after atmospheric correction, can be converted into an estimate of concentrations of chlorophyll pigment, suspended matter and *gelbstoff* (coloured dissolved organic material). Main application domains of such data are (1) the ocean carbon cycle, (2) the thermal regime of the upper ocean and (3) the management of fisheries and of coastal zones. MERIS allows global coverage of the earth in three days.

In this paper the spectrum of water leaving radiance reflectance (reflectance, for short) is used to monitor oceanic and coastal environment. The nine spectral bands are centred at wavelengths $\lambda =$

412, 442, 490, 510, 560, 620, 664, 681, 708 *nm*. Examples of water reflectance spectra are shown in fig. 1.

Figure 1. Examples of water reflectance spectra measured by MERIS.

This illustrates the great variability of water reflectance spectra: the sea is not always blue. Randomly chosen pixels from MERIS scenes 2002/2003 in a region¹ of the North Sea are used to produce the scatter-plots of logarithms² of reflectances at different wavelengths in fig. 2.

This shows that there are strong correlations between the reflectances. But obviously the correlation behaviour is not the same for all pixels –

¹ Details of this data are discussed in section 3

² We use the logarithm of the reflectance rather than the reflectance since the latter have distributions strongly skewed to small values, especially in the long wavelength bands. As the reflectance has the unit sr^{-1} formally we divide the reflectance by its unit before taking the logarithm. Any scaling of the unit would not change anything since for the considered algorithms what matters is always the difference of the variables.

Figure 2. Scatter-plots showing examples of water reflectances at different wavelengths measured by the MERIS sensor aboard the environmental satellite ENVISAT.

it changes with the kind of water. This leads to v-like shapes of point clouds in some scatter-plots.

For real applications of novelty detection hundreds of MERIS scenes with ten thousands of pixels each have to be checked. Therefore it is important to devise an algorithm which has a low rate of *false alarms*.

2.2. ALGORITHM PERFORMANCE WITH THE EXEMPLARY DATA

In order to study the pros and cons of different novelty detection schemes in a realistic manner as well as being able to visualise the findings we construct a two-dimensional data set. A randomly chosen subset of 800 points out of 5000 points altogether is shown in fig. 3. The demo data

Figure 3. Subset of the demo data.

set is a superposition of two Gaussians exhibiting the kind of v-like structure seen in the MERIS reflectance data shown above. Like the reflectance data the demo data set shows correlations, its variables are continuous and of same dimension and order of magnitude. Therefore

the demo data set is well suited to exemplify the novelty search in real reflectance data.

In any novelty search a model of the 'known' data $\mathcal{K}$ is built which serves to construct a measure $\rho(\mathbf{x})$ of how certain a given point $\mathbf{x}$ belongs to the kind of data characterised by the proxy $\mathcal{K}$. In the following sample applications we construct different models of the demo data set and discuss for each of the models the isoline of $\rho(\mathbf{x})$ which encloses 95% of the demo data set.

2.2.1. *Simple statistical scheme*

Let $\mathbf{x}^{(m)}$, $m \in [1, \dots, M]$ the M points of the known data $\mathcal{K}$. Then $\mathcal{K}$ is characterised to some extend by its first two moments, the centroid $\mathbf{q}$

$$q_j = \frac{1}{M} \sum_{m=1}^M w^{(m)} x_j^{(m)}$$

and the covariance matrix $\mathbf{C}$

$$C_{ij} = \frac{1}{M} \sum_{m=1}^M w^{(m)} (x_i^{(m)} - q_i)(x_j^{(m)} - q_j)$$

The moments contain information about the position of the points and shape of the point cloud $\mathcal{K}$ which is used in the Mahalanobis distance function. The Mahalanobis distance of a given point $\mathbf{x}$ from the centre of the points $\mathbf{q}$ is

$$\rho_M(\mathbf{x}, \mathbf{q}) = (\mathbf{x} - \mathbf{q})^T (\mathbf{C})^{-1} (\mathbf{x} - \mathbf{q})$$

In fig. 4 the isoline of the Mahalanobis distance from the centre enclosing

Figure 4. Subset of the demo data with isoline of the Mahalanobis distance from the centre which encloses 95% of the demo data.

95% of the demo data is shown. This clearly shows that the ellipsoid of equal Mahalanobis distance accounts for the gross arrangement of the data. But, since it is not able to account for details like the v-like shape the isoline in fig. 4 encloses also such points like *A* which could be representative for 'novelty'. This example shows that such a simple model is not appropriate for our MERIS spectra of water reflectances.

2.2.2. Auto-associative Neural Net

Auto-associative Neural Nets (AANN's) became quite popular for novelty search. AANN's are trained to map the points of $\mathcal{K}$ onto themselves whereby at one stage of the mapping the input information is passed through a manifold of lower dimension than the dimension of $\mathbf{x}$ ('bottleneck'). This can be accomplished only by exploiting correlations existing in $\mathcal{K}$. The AANN's can account for nonlinear correlations and can handle very distinct input sources.

Figure 5. Subset of the demo data with isoline $\rho(\mathbf{x}, AANN(\mathbf{x}))$ which encloses 95% of the demo data.

Since AANN's are trained to map the points $\mathbf{x} \in \mathcal{K}$ onto itself $AANN(\mathbf{x}) \simeq \mathbf{x}$ after successful training $\rho(\mathbf{x}, AANN(\mathbf{x})) = \sqrt{(\mathbf{x} - AANN(\mathbf{x}))^2}$

is small for points from $\mathcal{K}$. Now, to use AANN's for novelty search the idea is that points which are not like the points in the training set are not projected onto themselves and consequently lead to large $\rho(\mathbf{x}, AANN(\mathbf{x}))$. For the demo data the isoline of $\rho(\mathbf{x}, AANN(\mathbf{x}))$ which encloses 95% of the demo data is shown in fig. 5. It clearly demonstrates the AANN's ability to account for nonlinear correlations. The weakness of the AANN scheme comes from the fact that there is no mean to prevent the AANN from mapping also points onto itself which do *not* belong to $\mathcal{K}$. Therefore one can not exclude the possibility of small $\rho(\mathbf{x}, AANN(\mathbf{x}))$ for points like B in fig. 5 which rather are candidates for 'novelty'.

2.2.3. *Neural Net classifier*

To avoid the possibility that regions close to $\mathcal{K}$ are classified as belonging to $\mathcal{K}$ one can generate points close to $\mathcal{K}$ and then train a NN to classify the points. This is illustrated in fig. 6.

In the surrounding of points from $\mathcal{K}$ (dots) points belonging to the class 'non $\mathcal{K}$ ' (crosses) are generated. All the points together then are used as training- and test set to train a NN to classify the points accordingly. In this case the NN has just one output value which we code to represent membership in $\mathcal{K}$ by '0' and non-membership by '1'.

Figure 6. Two class data to train NN classifier.

The isoline of the NN output enclosing 95% of the data is shown in fig.

7.

The isoline now is quite close to $\mathcal{K}$.

For this demo data the generation of neighbouring points is trivial.

For real data the generation of neighbouring points can become a very difficult task in cases of manifolds extending in different directions and having varying intrinsic dimensionalities.

Figure 7. Subset of the demo data with isoline of $NN(\mathbf{x})$ which encloses 95% of the demo data.

2.2.4. Tessellation

Here we first tessellate the known data $\mathcal{K}$ and then define a distance of a given point from the tessellation³.

To start the tessellation we choose by trial and error a radius R (this is discussed below in detail for the tessellation of the real data). Now

- The first point becomes the first centre.

³ This could be the initialisation step of a cluster algorithm (*e.g.* 'K-means clustering') or the construction of a radial basis function (RBF) neural network. But the respective iterations would compromise the description of the border of $\mathcal{K}$ in favour of the better density estimation

- For each of the remaining points the smallest of the distances to all centres is compared with R . If the distance is larger than R the point is included in the set of centres.
- In the second step each point is assigned to that centre to which it is closest (*Voronoi tessellation*).

For each of the patches obtained by the tessellation the mean (centre of gravity) and the covariance matrix is calculated and as distance of a given point to the tessellation the minimum of the Mahalanobis distances is used⁴.

As compared with the two class NN classifier 2.2.3 the advantage of the tessellation approach is its easy and straightforward set up: no need to construct 'no $\mathcal{K}$ ' points, yet still obtaining tight enclosures of $\mathcal{K}$ as seen in fig. 8.

⁴ It should be mentioned that this scheme is a blend of the schemes (Auguststeijn 2002) and (D'Alimonte 2003) mentioned in the introduction: like in the PNN a set of Gaussians is used but here the covariance matrix of the points of each patch is used instead of univariate smoothing parameters.

Figure 8. Subset of the demo data with isoline of distance to the tessellation which encloses 95% of the demo data.

3. North Sea results

As an application for the tessellation novelty detection scheme we used MERIS data of the North Sea. In this section we first describe the construction of the dataset used to describe the 'normal' situation and then the construction of the tessellation and the cleaning of the results. After characterising the data which was used to search for novelties we discuss the different kinds of novelty which were detected by our procedure.

3.1. THE DATASET DEFINING THE 'NORMAL' SITUATION

We used all MERIS scenes from 26th June 2003 to 29th June 2004 in the region (region of interest = roi) shown in fig. 9. We excluded pixels

Figure 9. The roi in the North Sea chosen for the tessellation novelty detection scheme.

from the sun glint part as well as those which had one or more negative reflectances (bad atmosphere correction). Then randomly ~ 1000 pixels were sampled from each scene ending up with 115331 pixels. The logarithms of the remote sensing reflectances in the first nine MERIS channels (centred at wavelengths $\lambda = 413, 443, 490, 510, 560, 620, 665, 681, 708 \text{ nm}$) were used to span the space in which to search for novelties.

The projection of (a tenth of) this data onto the eigenvectors with the two largest eigenvalues is shown in fig. 10. This projection contains 94.9% of the variance of this data. It is obvious that the point cloud has not a simple shape.

Figure 10. The projection of the roi data onto the eigenvectors with the two largest eigenvalues.

3.2. THE TESSELLATION

The algorithm described in 2.2.4 was applied to the above data with varying R and hence varying number of resulting centres. To choose a sensible tessellation one has to compromise between low number of

Figure 11. The shape parameter ϵ as a function of the number of centres.

centres (computing time) and large number of centres (modelling the details of the pdf). For this we study fig. 11 which shows the weighted mean ϵ of the ratios of the two largest eigenvalues *vs.* the number of centres

$$\epsilon = \frac{1}{N} \sum_{c=1}^C n_c \frac{\lambda_2^c}{\lambda_1^c}$$

where n_c , $c = 1, \dots, C$ is the number of points belonging to the c^{th} centre ($N = \sum_{c=1}^C n_c$) and $\lambda_{1(2)}^c$ is the largest (second largest) eigenvalue of the covariance matrix of these points.

The quantity ϵ is a 'shape parameter': if all pixels are forced into one group the ratio $\frac{\lambda_2}{\lambda_1}$ is about a tenth. With an increasing number

of centres this ratio becomes larger and – if many details are modelled – stays constant around 0.6. As a compromise we decided to use the 26-centres tessellation ($R = 2.52$). This decision is supported by the finding that the use of the 31-centres tessellation leads to the same novelty findings.

Figure 12. The distribution of the distances of the training points from the tessellation.

The distribution of the distances of the training points from the tessellation is shown in fig. 12. From the blow-up of this figure, fig. 13, we decided to declare points as representing *novelty* if their distance to the tessellation is above 7.5. The spectra of the 38 pixel of the training set above distance 7.5 is shown in fig. 14.

Figure 13. The distribution of the distances of the training points from the tessellation – blow-up of fig. 12.

3.3. CLEAN UP OF THE FINDINGS

The distance cut 7.5 leaves only 38 pixel from the training points. But if we apply the cut to the 288 complete roi's (s. fig. 9) with 16,575,000 pixels of the time range from which training data set was selected we find 110,718 pixels with distance > 7.5 from the tessellation. Inspection of these pixels show many of them very near to clouds and also many singular. We therefore introduce two conditions which are not related to spectral information:

Figure 14. The spectra of the 38 pixel of the training set above distance 7.5.

1. Pixels have to be at least three pixels away from clouds. This brings down false alarms from 110,718 to 39,860.
2. Novelty has to appear in a minimum patch size. We accept pixels as novel only if in the surrounding 5×5 square at least 19 pixel have a distance above 7.5. This brings down false alarms from 39,860. to 2,670 in altogether 26 scenes.

To get rid of the remaining false alarms we identified four groups of pixels like that shown in fig. 15. For these four groups we calculated the centre of gravity and the covariance matrix and appended this to

Figure 15. One of the four groups which were appended to the tessellation was located near northern Denmark.

the existing tessellation information. After this addition one false alarm shown in fig. 16 remained.

Figure 16. The remaining false alarm in the Skagerrak.

3.4. NOVELTIES

The test data set comprises 223 scenes from 30th June 2004 to 13th October 2005. Application of our novelty detection scheme results in 10 scenes. The first three show a red tide near to the island Helgoland on 30th June, 03rd August and 05th August 2004 (see fig. 17).

Figure 17. Red tide in the German Bight found as novelty.

This red tide has been observed on a measuring campaign between Cuxhaven and Helgoland on 3rd August 2004 and identified as *myrionecta rubra*.

Figure 18. Coccolithophoride found as novelty (example from 23.06.2005).

In the scene at 19th November 2004 novelty was signalled in the Skagerrak area but inspection simply shows a processing failure: a lot of water reflectances are just unity.

The remaining six novelties at 19th, 23rd, 26th, 28th June and 2nd, 18th July 2005 are located in the central and northwestern North Sea. These findings are likely correlated to blooms of coccolithophorides. An example is shown in fig. 18.

4. Conclusions

A novelty detection scheme for MERIS water reflectances has been constructed which has a low false positive rate. It was successful in

finding exceptional algae blooms, one of them could be identified by *in situ* measurements as a red tide.

To reach the low false positive rate it was necessary to use not only spectral information but also to avoid close-by clouds and to require a minimum novelty patch size.

The novelty detection scheme developed is flexible and adaptable to various monitoring needs. It easily allows to exclude explained novelties from subsequent searches.

Acknowledgements

This research was supported by the European Space Agency ESA as a part of the 'MERIS Case2 Regional' project.

The authors like to thank the BEAM/VISAT development team (see URL <http://www.brockmann-consult.de/beam/>). The novelty detection processor was developed on the basis of this open source software package and thus can be freely used.

References

- Markos Markou and Sameer Singh, Novelty detection: a review–part 1: statistical approaches, *Signal Processing*, Volume 83, Issue 12, December 2003, Pages 2481–2497 and
- Markos Markou and Sameer Singh, Novelty detection: a review–part 2:: neural network based approaches, *Signal Processing*, Volume 83, Issue 12, December 2003, Pages 2499–2521.
- M. F. Auguststeijn and B. A. Folkert, (2002) Neural network classification and novelty detection, *International Journal of Remote Sensing*, 23, pp. 2891–2902.
- D. D’Alimonte *et al.* Use of the Novelty Detection Technique to Identify the Range of Applicability of Empirical Ocean Color Algorithms, *IEEE Transactions on Geoscience and Remote Sensing*, 41, 2003, pp. 2833–2843.
- T. P. Washburne, D. F. Specht and R. M. Drake, Identification of unknown categories with probabilistic neural networks, *Proceedings of the 1993 International Conference on Neural Networks*, San Francisco, California, 28 March–1 April, 1993, Volume 1 (Piscataway, NJ: IEEE Service Center), pp. 434–437.
- Rast M., J. L. Bézy and S. Bruzzi (1999) The ESA Medium Resolution Imaging Spectrometer MERIS – a review of the instrument and its mission, *International Journal of Remote Sensing*, 20, pp. 1681–1702; see also URL <http://envisat.esa.int/instruments/meris/descr/character.html>