Assessment of unsupervised classification techniques for intertidal sediments

E. Ibrahim, S. Adam, J. Monbaliu
Dept. of Civil Engineering, Katholieke Universiteit Leuven, Belgium

Keywords: intertidal sediments, unsupervised classification, imaging spectroscopy, clustering, artificial data

ABSTRACT:

The aim of this study is to explore three techniques for unsupervised classification of airborne hyperspectral imagery of intertidal flats. The unsupervised classification techniques considered are k-means (hard clustering), the Gustafson-Kessel algorithm (Fuzzy clustering), and the mixture of Gaussians model (probabilistic clustering). The behavior and suitability of these techniques is analyzed for sediment classification. Artificial data sets based on real airborne and field spectra are used for this purpose. The sensitivity of the techniques is investigated on two spectral aspects: the effect of within class (intra-class) variability and the effect of spectral dimensionality using feature selection. This sensitivity is expressed as classification accuracy in terms of the Kappa statistic (?) that indicates how better the classification is than chance agreement. The results show that the three techniques are suitable for sediment classification. When there is no feature selection involved, the mixture of Gaussians results in the best classification results. When feature selection is considered, sediment classification accuracy increases for all three techniques applied on the artificial imagery.

1 INTRODUCTION

An intertidal zone in a marine aquatic environment is the area of the foreshore and seabed that is exposed to air at low tide and submerged at high tide generating important geochemical processes (Silva et al., 2005). To describe these processes, extensive field knowledge is required. Since accurate data collection is often costly, inefficient, or unattainable, remote sensing can be a fine and resourceful alternative. Specifically, unsupervised classification techniques are of high interest where an image is classified on the basis of its reflectance values without taking field measurements into account. Field knowledge plays a role only in the analysis and identification of the classified groups. In practice, there is globally no “absolute best” unsupervised classification methodology, where the reliability of a result is based on the aim of the classification and its use (Everitt, 2001). Therefore, various unsupervised classification techniques are explored and assessed in this study for their performance in sediment classification. The techniques selected are K-means, Gustafson-Kessel algorithm (GK), and the mixture of Gaussians model (MG). In order to be in total knowledge and control of the test records, synthetic data sets are utilized as real imagery can include various uncertainties.

2 METHODOLOGY

2.1 *An introduction to clustering*

Clustering is a means of unsupervised classification defined as finding a structure in a collection of unlabeled data. It partitions an $N \times n$ data set X into c clusters (Lillesand and Kiefer, 2000) where X is the studied image, N is the number of pixels, and n is the spectral dimensionality. Each of the N pixels is characterized by a spectrum of the n frequency bands and is represented by a row vector (Balasko et al., 2004). Each type of clustering entails its limitations and specialties based on the concept of dissimilarity or *distance measures* (Everitt, 2001). Two popular distance norms are utilized in this study: the Euclidean and the Mahalanobis distance measures. When these distance norms are used for finding clusters as fixed or non-adaptive norms, limitations arise as the techniques then impose defined geometrical structures. Yet, when an adaptive norm is used for each cluster, more freedom in cluster features such as cluster shape, size, and orientation becomes possible (Kim et al, 2005).

2.1.1 *The clustering techniques*

The investigated techniques are of three clustering types: hard, fuzzy, and probabilistic. From hard clustering techniques, *k-means* is chosen where a pixel is allocated to a cluster minimizing the within cluster sum of squares using a non-adaptive Euclidean distance measure (Everitt, 2001, Balasko et al., 2004). In fuzzy techniques, each pixel can belong to more than one cluster yet in a different degree of belonging. The *Gustafson-Kessel algorithm (GK*) applies an adaptive distance norm of the Mahalanobis distance measure (Balzano and Del Sorbo, 2007, Gustafson and Kessel, 1979). The hard and fuzzy clustering are based on the "Fuzzy Clustering and Data Analysis Toolbox" by Balasko B, Abonyi J., and Feil B. (2004).

As for probabilistic clustering, the *mixture of Gaussians (MG)* model is used. It is a model-based probabilistic approach that constitutes of models describing each cluster. Clusters are considered as various Gaussian distributions according to their covariance structure (Beaven et al., 2000, Banfield et al., 1993). For this study, the popular expectation-maximization algorithm (EM) optimizes the fit between the data and the models (Dempster et al., 1977), and the Bayesian Information Criterion (BIC) selects the most suitable model describing the data (Fraley and Raftery, 1998). MG was performed by means of the Mixture Modeling software, MIXMOD (Biernacki et al, 2006).

2.1.2 *Accuracy assessment*

An artificial image is a combination of various artificial *classes* built on the basis of specific sediment *types* acquired from imagery and field spectra. The clustering techniques applied to the artificial data sets result in *clusters*. The comparison between *classes* and *clusters* indicates the accuracy of the classification. Confusion matrices are used to assess this accuracy as they compare the relationship between the artificial data as the reference data and the "corresponding" results of the unsupervised classification techniques (Lillesand and Kiefer, 2000). Yet with unsupervised classification, a resulting cluster is not automatically labeled nor identified as corresponding to a specific class. So, a class is investigated with respect to all clusters, and the cluster containing most of the pixels closest to the mean of that class is considered as its corresponding cluster. Based on the confusion matrix, the accuracy is then expressed in terms of the kappa statistic (k) where the difference between the clustering accuracy and the chance agreement between the classes and the clusters is calculated (Lillesand and Kiefer, 2000). It results in a value between 0 and 1

for each classification, where 0 indicates that the clustering is no better than grouping the data by chance:

$$k = \left[N \sum_{i=1}^{c} y_{ii'} - \sum_{i=1}^{c} (y_i . y_{i'}) \right] / \left[N^2 - \sum_{i=1}^{c} (y_i . y_{i'}) \right] \quad (1)$$

Where

c is the number of classes or clusters, $1 \le i \le c$.
N is the total number of pixels in the artificial image classified
i' refers to the cluster corresponding to class i
$y_{ii'}$ is the number of observations in row i and column i' in the confusion matrix
y_i is the number of observations in row i in the confusion matrix
$y_{i'}$ is the number of observations in column i' in the confusion matrix

2.2 *Building artificial data*

2.2.1 *Data available*

To build the synthetic images, data extracted from a hyperspectral image is taken into consideration. This image acquired by Airborne Hyperspectral Sensor (AHS) on the 17th of June 2005 covers the IJzermonding estuary located on the Belgian coast. Its spatial resolution is 3.4 meters and 19 non-corrupted bands are available covering the visible part of the spectrum VIS (10 bands, 0.455-0.746µm, 30nm wide), the Near Infra-red NIR (8 bands, 0.774-1.004µm, 30nm wide), and the short wave infra-red SWIR-1 (1 band, 1.622µm, 200nm wide). Furthermore, field sampling was carried out at a timeframe close the image acquisition. The resulting samples were analyzed for relative moisture content (RMC), mud content (MC), and chlorophyll *a* content (chl *a*). The sediments are considered "wet" when RMC> 30%, high chl *a* when the chl *a* content > 40 mg/m^2, and either muddy or sandy based on the clay and silt content of a threshold equal to 30% of particles (Deronde et al., 2006). The spectral measurements were carried out by means of an Analytical Spectral Device (ASD) that records reflectance covering the spectrum from the VIS to SWIR with a resolution varying from 3nm to 10nm.

2.2.2 *Sediment types*

To build the synthetic data, three major sediment *types* identified at the study area are selected. Representative spectra of those *types* are extracted from the AHS image and the basic statistics of each of the types is calculated such that for $type_i$, mean μ_i and standard deviation s_i are calculated per spectral band (μ_{ib} and s_{ib}). *Type1* is wet, clayey, and with high chl *a* content; *type2* is wet, clayey, yet with low chl *a* content. *Type3* is dry, sandy, and with low chl *a* content.

2.2.3 *Sediment properties in artificial data*

Synthetic data sets are constructed using all the *n* bands of the AHS image. An artificial image is a combination of a number of *classes* X_i. Each class is of N_i pixels and built on the basis of the statistics of each *type*; a mean spectrum, μ_{ib}, and a standard deviation in terms of s_{ib}. To relate the artificial data to the reality of spectral distribution in the field, three properties, RMC, chl *a* content, and mud content, that distinguish the three sediments types play a role in building the data. The first step in building theses images consists of applying the effect of moisture content to the different types. This is explained in 2.2.4. In the second step, the effect of chl *a* content is applied as in 2.2.5 to the spectra of *type1* and *type2* generated in the first step. These types show variability in chl *a* content where a chl *a* absorption feature appears in the red region at around 673nm.

The effect of clay content is not included directly as the clay content absorption features are not visible with field or airborne spectra (Adam et al., 2008). However, correlations between mud con-

tent and chl *a* content have been demonstrated (Van Engeland, 2008). Therefore, by considering the effect of chl *a* and RMC in the data, the effect of clay content is considered indirectly.

2.2.4 *Effect of moisture content*

For changes in RMC, an increase or decrease of the whole spectrum is considered (Weidong et al., 2001; Muller & Décamps, 2000). Reflectance normally decreases as water content increases. Yet, at a value of RMC referred to as "the cut-off thickness" that is soil dependent, reflectance increases (Neema et. al, 1987). Assuming at each spectral band of a $type_i$ the spectra are normally distributed, a spectrum of $class_i$ represented by the N_i *by n* matrix X_i is generated by m_i and a standard deviation represented in terms of s_i. So, a spectrum *k* in $class_i$ is calculated as follows:

$$\mathbf{X_{ik}} = \boldsymbol{\mu}_\mathbf{i} + r.\mathbf{s}_\mathbf{i}.s \tag{2}$$

Where

$1 \le i \le c'$, c' is the number of classes
$1 \le k \le N_i$
m_i is a matrix containing the mean values per band for $type_i$
s_i is a matrix of standard deviations per band of $type_i$
s is a factor multiplied by s_i and determines the standard deviation of a class ($s > 0$)
N_i is the number data points in $class_i$
r is a random value following a normal distribution: *mean=0* and *standard deviation=1*

In addition, random noise factors are added to each band in the order of ±1% of the reflectance values.

2.2.5 *Effect of chl a content*

In the field samples acquired from the IJzermonding, RMC and chl *a* show a significant correlation with an $r^2 = 0.56$. Due to the cut-off thickness in spectra affected by RMC, the positive correlation between moisture content and chl *a* can be demonstrated in the artificial data by allowing the presence of a random chl *a* dip. This is introduced to the spectra generated in 2.2.4 for *type1* and *type2*. The ratio between the minimal reflectance in the absorption feature and the reflectance outside the absorption feature has a positive correlation with chl *a* content (Adam et al., 2008). With airborne spectra, Pearson's correlation resulted in $r^2 = 0.57$ between chl *a* content and the ratio between the AHS bands 9 and 8 of 718nm and 689nm respectively.

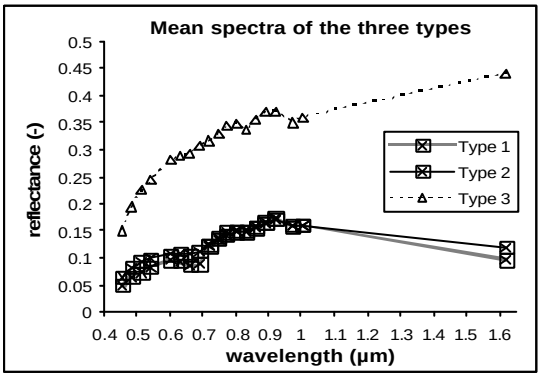

Figure 1: Mean spectra μ_i of each *type*

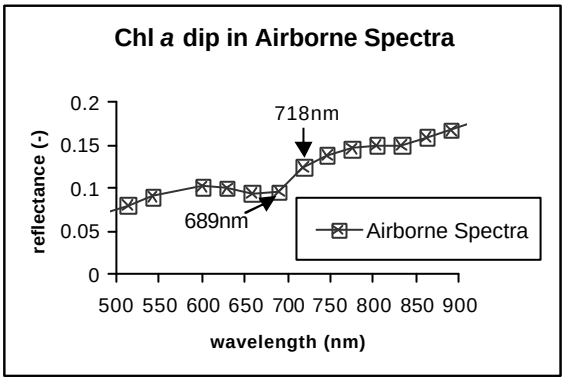

Figure 2: Chl *a* dip in AHS spectra

Considering that the chl *a* dip is bounded by band 5 and band 9 of 0.601µm and 0.718µm respectively, the effect of chl *a* is first introduced at band 8 of 0.689µm in the following manner:

$$X_{ik8'} = min_{i8}.s + r'.min_{i8}.s \tag{3}$$

Where

$X_{ik8'}$ is the new value at band 8

min_{i8} is the minimum reflectance value of $band_8$ in $type_i$

r' is a random value / $0 < r' = m$, where m varies fulfilling the condition: $X_{ik8'} < X_{ik5}$ as band 5 notes the start of the dip

Finally, bands 6 and band 7 of 0.689 and 0.630µm respectively are altered according to the new values of band 8.

3 RESULTS

3.1 *Clustering and inter/intra class variability*

The dependence of the clustering techniques on spectral variance in the data is analyzed while spatial properties are preserved. So, the classes are set to have the same number of pixels $N_i = 500$. First, the artificial data is built containing the three *types* and of various standard deviations represented by *s*. With *c=c'*=3, *figure 3(a)* shows the results obtained by the three techniques. They all resulted in acceptable *?* values through the range of intra-class variability. Yet, referring to the confusion matrices, a good classification by the three techniques for *type3* is noticed while *type1* and *type2* are more difficult to discriminate. Therefore, the second step is to reduce the inter-class variability. So, the images are built using 2 *types* at a time, *c=c'=2*. The clustering results show a better performance in distinguishing *type1* and *type2* by k-means; MG still resulted in *?* =1, while GK resulted in a poor classification of the spectra *(figure 3)*. As for distinguishing *type3* from the other two types, all methods lead to high values of kappa. Therefore, in real imagery, since *type3* can be detected easily, one could mask it out of the image and then apply the clustering techniques on the other two types for better classification accuracy.

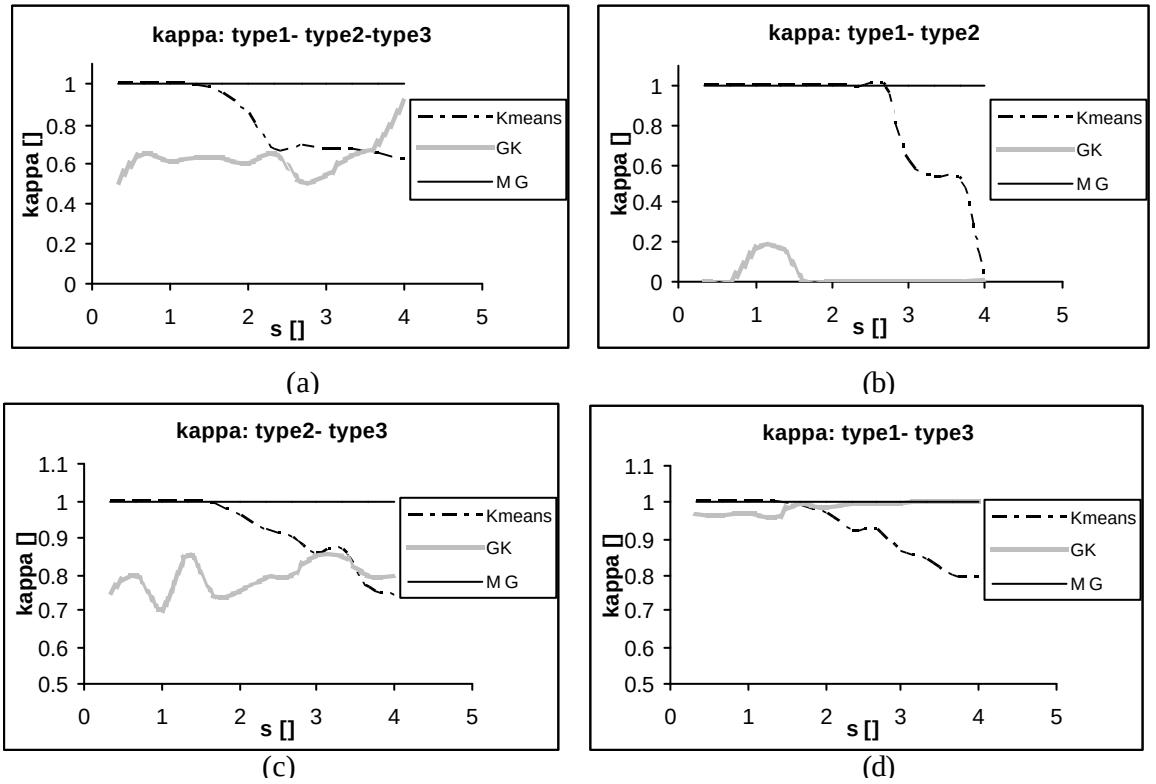

Figure 3: Accuracy in terms of Kappa with respect to standard deviation of (a) the three types (b) *type1* and *type2* (c) *type2* and *type3* (d) *type1* and *type3*

3.2 *Feature selection*

3.2.1 *Based on supervised classification*

According to Deronde et al. (2006), spectral dimensionality can be decreased resulting in not only computational efficiency, but also an increase in classification accuracy. A band selection by the sequential floating forward selection algorithm (SFFS) based on the methodology of Deronde et al (2006) is carried out for the AHS image to test the clustering techniques. The bands selected for RMC are 455nm, 513nm, 918nm, and 1622nm, for mud content 689nm, 833nm, 1004nm, and for Chl *a* content 718nm, 689nm, 774nm.

After building the artificial data, these bands are selected and a comparison is carried out between the results obtained by using these few bands and all the *n* bands (*figure 4).* Using bands chosen for any of the three properties, similar or better results to using the *n* bands are obtained for k-means, better results for GK, and the same result of *?*=1 for MG. Therefore, using only three or four bands, the computational efficiency increases, and the results are generally better than using all the 19 bands of the AHS image.

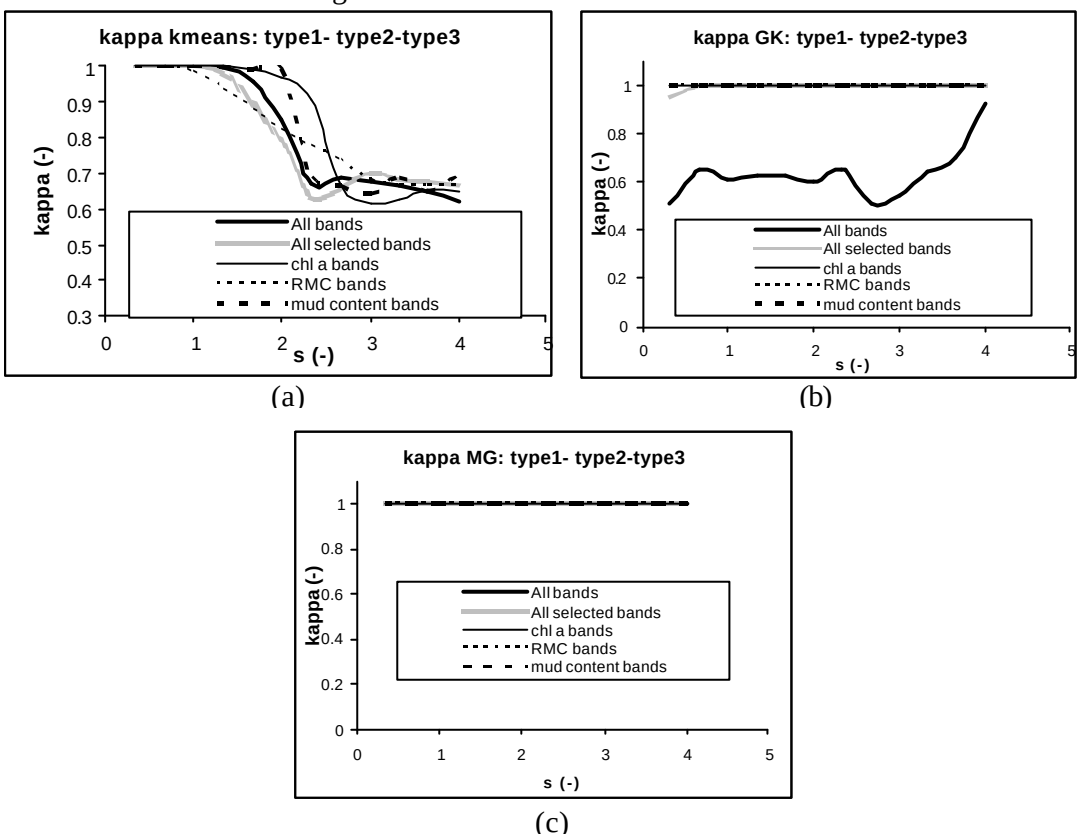

Figure 4: the effect of feature selection on kappa accuracy for (a) k-means (b) GK (c) MG

3.2.2 *Based on random band selection*

Hyperspectral data within adjacent bands are usually highly correlated. As a random attempt to reduce these correlations, 4 bands are selected from the blue, green, red, and NIR parts of the spectrum repetitively. With a range of standard deviations considered, $0 < s = 4$, the three techniques result in high classification accuracy. MG and GK resulted in kappa values of minimum 0.95. The k-means results varied between $0.6 < ? < 0.8$ for higher and lower standard deviations respectively.

4 CONCLUSIONS

Unsupervised classification is essential for intertidal sediment characterization due to the many difficulties incorporated in field work on such regions. Three techniques are investigated in this work, namely k-means, the Gustafson-Kessel algorithm (GK), and the mixture of Gaussians model (MG). Artificial data based on real imagery and sediment properties are built to test the data with varying intra-class variability and spectral resolutions.

From this study, it can be concluded that the unsupervised techniques are capable of discriminating sediment types successfully. With varying inta-class variability, the three techniques lead to acceptable results with superiority to MG.

Furthermore, it is revealed that the techniques are dependent on inter-class variability represented by the occurrence of different sediment types in an image. Therefore, a good approach can be to cluster first the easily distinguished types and mask them out from the image. This can lead to an increase in the classification accuracy of the remaining types.

Feature selection also increases the classification accuracy. By referring to the work of Deronde et al., 2006, the extracted bands from the AHS image lead to generally better classification accuracy for the three techniques. GK and MG show a perfect classification of the three types on a range of intra-class variability $0 < s = 4$.

Furthermore, if four bands are selected randomly, one from each of the blue, green, red, and NIR regions of the spectrum, the classification accuracy is superior to using all the 19 bands. MG and GK result in very high kappa values ($? > 0.95$) for any combination of the four bands on the whole range of standard deviations. The accuracy of k-means also increases resulting in $? > 0.6$.

In conclusion, all three techniques are able to distinguish the three sediment types considered for this study. Yet, if there is no feature selection involved, MG can be considered the most robust. If there is feature selection involved, the three techniques lead to good classification accuracies, yet with superiority to MG and GK.

5 FURTHER WORK

Further study is being done to interpret the results obtained using artificial imagery for the real imagery. There are three major points addressed. First, the same class sizes are considered for the artificial images. Therefore, the effect of having various class sizes in an image on classification accuracy is investigated. Furthermore, the spatial distribution of the spectra and the role of neighboring pixels are examined. Although these clustering techniques do not take spatial information into account, the spatial distribution can affect the choice of suitable seed pixels initiating the different algorithms. Finally, the issue of the number of clusters to be retrieved from an image is considered. In a real image, the number of clusters in the image is not known from beforehand as it is in this study. Therefore, this aspect is investigated in order to be able to apply the clustering techniques successfully on real imagery.

ACKNOWLEDGEMENTS

The research presented in this paper is funded by the Belgian Science Policy Office in the frame of the STEREO II programme – project SR/00/109 - ALGASED (*remote sensing for characterization of intertidal sediments and microphytobenthic algae*)

For this reseach, there was also financial contribution from the CISS-project funded by the Research Foundation-Flanders (FWO Vlaanderen) under contract no. G0480.05.

The field campaigns are supported by the Belgian Science Policy in the framework of the STEREO program – project 043, 072 and 109.

The authors are grateful to the ALGASED team members

Finally, the authors express their gratitude to the contribution of Ms. Sindy Sterckx (Flemish institute for technological research - VITO). The work of Ms. Sterckx work was in the framework of the SEDOPTICS project funded by the Belgian Science Policy Office in the frame of the STEREO programme – project SR/00/72

REFERENCES

Adam, S., Monbaliu, J., Toorman, E. 2008. Quantification of bio-physical intertidal sediment properties using hyperspectral laboratory and in situ measurements. *Presented at the 28th EARSeL Symposium and Workshops, Remote Sensing for a Changing Europe*, 2-7 June, 2008, Istanbul.

Balasko, Abonyi, B., J. and B. Feil. 2004, Fuzzy Clustering and Data Analysis Toolbox: For Use with Mat lab, University of Veszprem, Hungary

Balzano W. and Del Sorbo M.R. 2007. Genomic comparison using data mining techniques based on a possibilistic fuzzy sets model. *Biosystems* 88, Issue 3: 343-349.

Banfield, J and Raftery. 1993. A. Model-based Gaussian and non-Gaussian clustering, *Biometrics* 49: 803–821

Beaven, S.G., Stein, D., Hoff, L.E. 2000. Comparison of Gaussian mixture and linear mixture models for classification of hyperspectral data. *Proceedings of the IEEE Geoscience and Remote Sensing Society. Honolulu, Hawaii, USA*: 1597-1599

Biernacki, C., Celeux, G., Govaert, G., Langrognet, F. 2006. Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics & Data Analysis* 51: 587-600

Dempster, AP, Laird N. M. and Rubin D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.* 39: 1–38

Deronde, P. Kempeneers and R.M. Forster. 2006. Imaging Spectroscopy as a Tool to Study Sediment Characteristics on a Tidal Sandbank in the Westerschelde. *Estuarine, Coastal and Shelf Science* 69: 580-590

Everitt, B. S., 2001, Cluster Analysis , Edward Arnold, London.

Gustafson D. E. and W.C. Kessel, 1979, Fuzzy clustering with fuzzy covariance matrix. *In Proceedings of the IEEE CDC, San Diego*: 761-766.

Kim D., Kwang H. Lee, and Doheon Lee, 2005, Detecting clusters of different geometrical shapes in microarray gene expression data. *Bioinformatics Advance Access, Bioinformatics* 21: 1927-1934

Lillesand, T.M. and Kiefer, R.W. 2000. Remote sensing and image interpretation. John Wiley & Sons, Inc., New York, NY, USA.

Muller E. and Décamps H. 2000. Modeling soil moisture-reflectance. *Remote sensing of environment* 76: 173-180

Neema, D., Shah, A, & Patel, A. N. 1987. A statistical model for light reflection and penetration through sand. *International Journal of Remote sensing* 8: 1209-1217

Silva J., Santos R., Calleja M., and Duarte C. 2005. Submerged versus air-exposed intertidal macrophyte productivity: from physiological to community-level assessments. *Journal of Experimental Marine Biology and Ecology* 317, Issue 1: 87-95

Van Engeland, T. 2005. Using field data and hypsespectral remote sensing to model microalgae distribution and primary production on an intertidal mudflat. Unpublished MSc thesis. Universiteit Gent, Belgium.

Weidong, L., Baret F., Xingfa G., Qingxi T., Lanfen Z., Bing Z. 2002. Relating soil surface moisture to reflectance. *Remote sensing of environment* 81: 238-246