Universal Journal of Geoscience 2(4): 109-115, 2014
DOI: 10.13189/ujg.2014.020401 http://www.hrpub.org

Regression Tree Analysis of Spm Entropy Groups: Case Study from the Irish Sea

Krivtsov V.[1,2,*], Mikkelsen O.A.[3]

[1]The University of Edinburgh, Crew building,West Mains Road, Edinburgh EH9 3JN, Scotland, UK
[2]Department of Ecology, Kharkov State University, 4 Svobody Square, Kharkov 310077, USSR, Ukraine
[3]MacArtney Underwater Technology, Gl. Guldagervej 48, DK-6710 Esbjerg V, CVR No. 84 16 48 28
*Corresponding Author: e96kri69@netscape.net

Abstract This paper describes our studies of the suspended particulate matter (SPM) in the Liverpool Bay (UK). Monitoring data were analyzed by using entropy analysis. Entropy analysis of in situ particle size spectra revealed 5 basic types, attributable to different sets of environmental conditions. The revealed basic types of in situ particle size spectra were then subjected to the classification trees analysis in order to identify the meteorological and oceanographic variables of importance for the characterisation of the shape of SPM spectra. The results obtained are a step towards a better characterisation of the floc size, and therefore a more precise calculation of the sedimentation and transport rate, and are therefore relevant to the scientific analysis of a wider range of environmental issues.

Keywords Suspended Sediment, Irish Sea, Marine Environment, Particle Size, Entropy Analysis, LISST, Oceanography, Marine Ecosystem

1. Introduction

Suspended particulate matter (SPM) is of fundamental importance in issues of ocean engineering. Its dynamics is indispensable for understanding of corrosion and abrasion of materials, and also the formation of fluid mud, and is therefore relevant to issues of navigation and channel maintenance (Schrottke *et al.*, 2006; Schwartz & Kozerski, 2003). SPM is also important as regards issues of aquatic ecology and environmental management (Hakanson & Eckhell 2005) as it is intimately related to the transport of pollutants and influences water clarity and primary production, and hence also secondary production (Krivtsov *et al.* 2008b). Consequently, SPM characterisation has recently been among the increasingly important topics as regards pollution control, environmental auditing and management (Audry *et al.* 2006; Guo *et al.* 2007; He *et al.* 2006; Karrasch *et al.* 2006; Maldonado *et al.* 1999; Manjunatha *et al.* 2001; Shankar & Manjunatha 1994; Zhou *et al.* 2000), and ecological modelling (Barros & Abril 2005; Ebenhoh *et al.* 2004; Hakanson & Eckhell 2005; Hakanson *et al.* 2004; Hakanson *et al.* 2005; Hakanson *et al.* 2000; Johansson *et al.* 2001; Krivtsov *et al.* 2008a; Lindstrom 2001; Lindstrom *et al.* 1999; Malmaeus & Hakanson 2003, 2004).

The in situ particle size spectrum of suspended particulate matter in aquatic environment influences the feeding pattern of bottom fauna (Cranford *et al.* 2005), affects the transmission and reflectance of light in water (Mikkelsen 2002) and is of importance for numerous sedimentological and a wide range of ecological processes (Krivtsov *et al.* 2008b). It has previously been shown (Sharp & Fan 1963) that such parameters as mean/median particle size and standard or median absolute deviation can be incomplete or even misleading descriptors of the shape of the size spectrum, in particular for multi-modal spectra (Mikkelsen *et al.* 2007; Mikkelsen *et al.* 2005). Here we have applied a combination of entropy modelling with regression tree analysis to deduce 5 basic types of SPM spectra, and describe their relationships with environmental variables.

2. Materials and Methods

2.1. Site Description

The data presented here were collected in Liverpool Bay, an area of the Irish Sea important as regards recreation and shipping (Figure 1). The site is characterised by tidal straining, intertidal regions with exposed banks, high suspended sediment concentration and complex biogeochemical interactions. Tidal currents are strong (up to 1 m/s during springs) and there are occasional large storm surges and waves (in particular associated with westerly winds). The principal freshwater inputs are from rivers Mersey, Dee and Ribble. Further information about the area can be found in our previous publication and references therein (Krivtsov *et al.* 2008b).

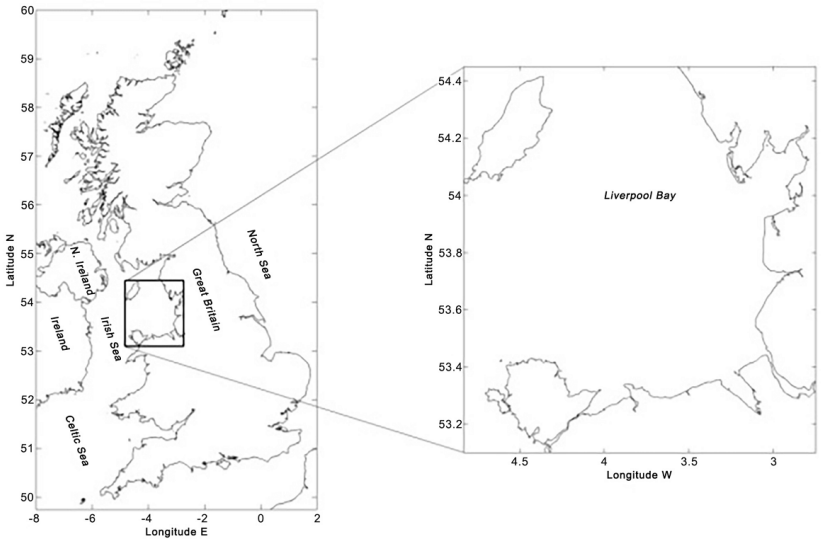

Figure 1. Location of the study site

2.2. Observations

The observational data forming the basis of this paper come from 9 cruises carried out by the Proudman Oceanographic Laboratory and School of Ocean Sciences (Bangor) on the RV Prince Madog, which took place between Sep 2004 and Feb 2006. The oceanographic variables measured during the cruises using a profiling CTD package are standard, and include (among others) temperature, salinity, conductivity, beam attenuation, chlorophyll a fluorescence, photosynthetic active radiation (PAR). Data on wave characteristics are available from the CEFAS wave rider buoy. The principal observational evidence comes from the LISST-100 laser, which provides in situ estimates of volume concentrations in µl/l for 32 size classes corresponding to particle sizes between 2.5 and 500 µm (Agrawal & Pottsmith 2000).

2.3. Entropy Analysis

Entropy analysis has previously (Mikkelsen *et al.* 2007) been used to classify in situ particle (floc) size spectra of suspended particles into groups based on similar distribution characteristics. It was evident that the in situ spectra sorted into groups that reflected different forcing conditions (e.g. variations in turbulence). Importantly, the different forcing conditions were not necessarily reflected in other commonly used distribution measures such as median floc diameter; this suggests that entropy analysis may be an effective approach for investigating the effect of changes in forcing conditions on floc size (Sharp & Fan 1963).

In information theory, the concept of entropy is related to the randomness of an event or a signal. Essentially, entropy links the information content of a signal to its randomness – if a signal has a high entropy (high randomness) the information content is low and vice versa. In particle size terms, this can be illustrated by considering a completely flat size spectrum, i.e. all volume (or mass) in the size spectrum occurs with the same frequency throughout the spectrum. This is essentially a random distribution of matter throughout the size spectrum, so a size spectrum with this shape has maximum entropy. Conversely, in a size spectrum where all particle volume or mass is found in only one bin there is no randomness of the distribution, so the entropy for such a spectrum is at a minimum. Therefore, a particle size spectrum can be characterized in terms of its entropy. For a particle size spectrum with n size bins, the entropy, E, is given as:

$$= -\sum_{i=1}^{n} p_i \log p_i \qquad (1)$$

where p_i is the proportion of particles in size bin I(Shannon 1948). Note that when $p_i = 0$, $p_i \log p_i = 0$ (according to L'Hôspital's rule). The entropy can vary between a maximum value, E_{MAX} of log n when all $p_i = 1/n$ and a minimum value, E_{MIN}, of zero when $p_i = 1$ for exactly one of the bins $i = 1 \ldots n$. The entropy is related to the information gain, I, which is also known as the inequality statistic, by the equation:

$$I = (\log n) - \qquad (2)$$

When E equals E_{MAX}, the proportion of particles is the same in all size bins, and I equals zero. As the value of I increases, the information content of the size spectrum increases.

With an ensemble of size spectra, the inequality statistic can be used to divide the spectra into groups. Optimal grouping maximizes the inequality between the groups and minimizes the inequality within the group, so the spectra in each group all have similar shapes, and the shapes of the spectra differ mainly between groups. The first step in grouping size spectra is to express the proportion of particles in each size bin of each size spectrum as proportions of the grand total(Johnston & Semple 1983). For size distributions expressed as volume concentrations, the volume concentration in each size bin of each spectrum must be divided by the grand total volume concentration, defined as the sum of the volume concentration in all bins in all spectra:

$$Y_{ij} = \frac{\mathrm{VC}_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\mathrm{VC}_{ij}} \quad (3)$$

where N is the number of spectra, J is the number of size bins in each spectrum, VC_{ij} is the volume concentration in spectrum i, bin j and Y_{ij} is the proportion of the total volume concentration in spectrum i, bin j.

Following Johnston and Semple (Johnston & Semple 1983), the total inequality for all spectra is then given as:

$$I = \sum_{j=1}^{J} Y_j \sum_{i=1}^{N} Y_i \log NY_{i,} \quad (4)$$

where $Y_j = \sum_{i=1}^{N} Y_{ij}$ and $Y_i = Y_{ij}/Y_j$.

In case the spectra have been divided into R groups, a measure of the efficiency of the grouping (in terms of maximising between-group inequality) can be obtained from the so-called R_S statistic:

$$\mathrm{RS}=(\mathrm{IB}/\mathrm{I})100 \quad (5)$$

In Eq. (5) I_B is the between-group inequality, which is defined as:

$$I_B = \sum_{j=1}^{J} Y_j \sum_{r=1}^{R} P_{ir} \log \left(\frac{P_{jr}}{N_r/N}\right), \quad (6)$$

where $p_{jr}=(\sum_i \in_r Y_{ij})/Y_j$, and N_r is the number of spectra in group r of R. High R_S values indicate that the inequality is mostly related to differences between the groups and that the inequality within each group is low. In short, the spectra within each group have similar shapes, and the shapes of the spectra differ mainly between groups.

Unfortunately there is no way to predict in advance how the spectra should be grouped or how many groups are desirable. The only way to obtain a best grouping (simply defined as the best R_S statistic) is to perform all possible combinations of N spectra into R groups, compute the R_S statistic for each of the combinations and then choose the combination that yields the largest R_S statistic for that number of groups. This problem is well known from other grouping techniques such as, for example, principal component analysis, where the full set of principal components is as large as the original set of variables, but the vast majority of the variation usually can be explained by the first two to four principal components.

Johnston and Semple (1983) provided a FORTRAN routine that automatically arranges the data into a user-selected number of groups, and then shifts them between groups until an optimal grouping for that number of groups is found. Their routine was later adapted to QBASIC (Woolfe & Michibayashi 1995) for the analysis of sedimentological facies. Entropy analysis was also useful in delineating ecological habitats on the Scotian Shelf off Nova Scotia, Canada (Orpin & Kostylev 2006). Here we have used a Matlab implementation of the Entropy analysis reported previously (Mikkelsen *et al.* 2007).

2.4. Regression Tree Analysis

To investigate whether the entropy group of an SPM spectrum could be predicted using a set of meteorological and oceanographic variables, we applied the data mining method of regression trees using a Matlab function 'treefit' with the 'classification' option. The resulting tree was subsequently pruned using level 4 and displayed using the 'treedisp' operator. The list of variables used in this analysis is given in Table 1.

Regression trees are a representation for piece-wise constant or piece-wise linear functions, and models are given in a form of hierarchical structures of their elements. The models predict the value of a dependent variable (i.e. in our case, the entropy group type of the SPM spectra) from the values of a set of independent variables. The space of examples is partitioned into axis-parallel rectangles and a model is fitted to each of these partitions. A regression tree has an inverse hierarchical structure with a test in each inner node (junction from were two links go to the lower hierarchical levels). Each node tests the value of a certain independent variable, and each leaf (the lowest level of hierarchical tree) displays a linear equation or (in the analysis presented here) just a constant for predicting the value of the dependent variable.

3. Results and Discussion

The 5 groups of spectra resulting from the entropy analysis are displayed in Figure 2, with the groups numbered in the ascending order according to the position of the main modal. To investigate the relationships between the group number and environmental factors, the data were subjected to the regression trees analysis. The variables used for this analysis were the ones known to be important to the SPM characterisation from our previous work and also those showing particularly strong correlations with the spectra grouping (Krivtsov *et al.* 2012)

The classification tree resulting from the regression tree analysis is displayed in Figure 3. It shows that the most important variables for the characterisation of the SPM spectra are temperature, the directions of wind and waves, wave period and orbital velocity. The upper level node is represented by temperature, thus dividing all the spectra into predominantly winter ones (types 1 and 2) and predominantly summer ones (types 4 and 5). The spectra belonging to type 3 constituted rather a small group, and their occurrence was under broadly similar (albeit somewhat more turbulent) conditions as those of type 4 (data not shown).

It should be noted, however, that Type 2 spectra can also be observed during warmer periods, provided there are sufficiently high levels of turbulence. These may e.g. happen either on the tail of a passing depression (when the strong swell comes from the W/ NW) or during sufficiently strong tidal currents - see Figure 3. It should also be noted that, based on the results of these analysis, at the site studied the wave-induced turbulence appears to be more important for the SPM characterisation than tidal currents, which is in line with our previous work (Krivtsov *et al.* 2009).

Table 1. List of variables used in the regression tree analysis

Variable Name	Explanation of the variable
BotSPM	Bottom SPM, mass concentration (gravimetric method)
SurfSPM	Surface SPM, mass concentration (gravimetric method)
BotV	Bottom SPM, volumetric concentration
SurfV	Surface SPM, volumetric concentration
BotBeamAt	Bottom beam attenuation
SurfBeamAt	Surface beam attenuation
BotMedD	Bottom median diameter
SurfMedD	Surface median diameter
BotVoverD	Bottom Volume over Diameter ratio
SurfVoverD	Surface Volume over Diameter ratio
TDav	Water temperature (depth average)
Transmissometer (beam att.)Dav	Beam attenuation of the SeaTech transmissometer
Sal (PSU)Dav	Salinity (depth average)
Density (kg per m^3)Dav	Density (depth average)
Potential Energy Anomaly	Potential Energy Anomaly
Tide level	Tide level
Tide current	Tide current
Tide direction	Tide direction
Water Depth	Water Depth
Epsilon proxy	Estimate of turbulent kinetic energy dissipation
DeltaSPM mg per l	Difference between bottom and surface SPM mass concentration
DeltaTotV	Difference between bottom and surface SPM volumetric concentration
DeltaDiameter	Difference between bottom and surface median diameter
DeltaTransm	Difference between bottom and surface SPM estimated using a SeaTech transmissometer
Kolmogorov Scale	Kolmogorov Scale
DominantWaveDir	Dominant wave direction
DominantWaveT	Dominant wave period
MaxOrbitalU	Maximum orbital velocity due to wave action
WaveEnergy	WaveEnergy
TrueWindSpeed	Wind speed
TrueWindDir	Wind direction
AirTemp	Air temperature
Pressure	Atmospheric pressure
Humidity	Humidity
WindEpsilon	TKE dissipation due to wind
WindWaveAlign	Allignment between wind and waves
WindTideAlign	Allignment between wind and tide
DeltaBotSurfGroups	Difference between the recoded bottom and surface group numbers
GroupBotRecoded	Bottom group number
GrsurfRecoded	Surface group number

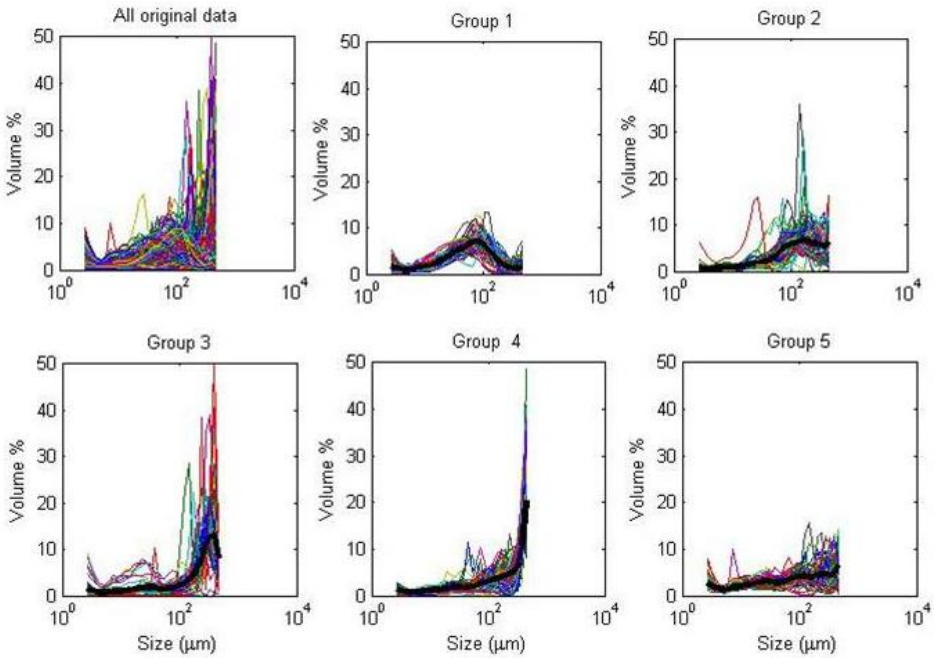

Figure 2. Groups of spectra resulting from entropy analysis and numbered in the order of the increase in the position of the main modal. Note that group 5 appears to have the modal outside the coarse end of the measurements window

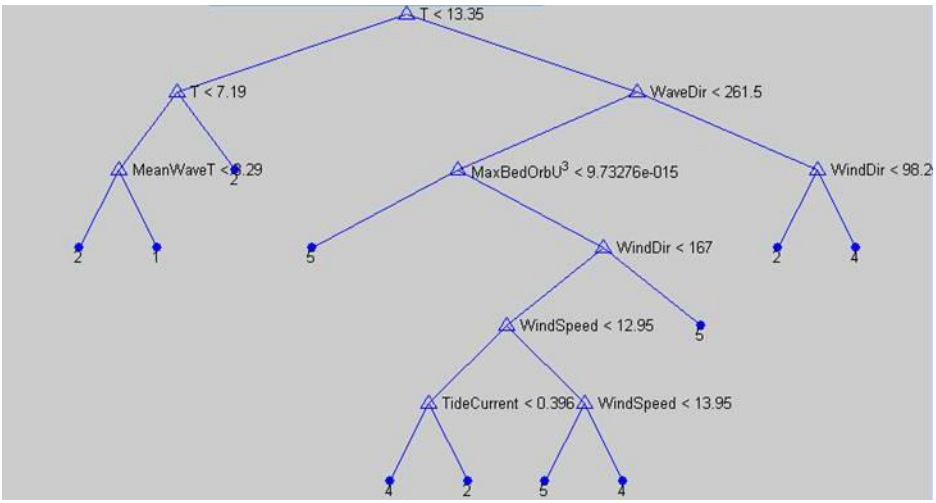

Figure 3. Classification tree for the bottom group of the spectra type. Note that Group 3 only constituted a relatively small proportion of the bottom spectra, and was therefore 'pruned'. Dominant wave period is labelled as 'MeanWaveT'; the other variables as in Table 1

It has previously been argued (Mikkelsen *et al.* 2007) that the shape of the in situ size spectrum must be a function of a limited number of variables, including turbulence, biological 'stickiness' and suspended matter concentration. Therefore, a group of size spectra that all have approximately the same shape should be indicative of a certain set of environmental conditions. Thus the complexity of the in situ size spectra in a particular body of water may be reduced to a few groups, each typical of the forcing conditions varying within a certain limited range (Krivtsov *et al.* 2012). The results presented here not only support these considerations, but also provide an insight how the shape of spm spectra could be estimated from the concurrent environmental conditions. It should be noted, however, that to enable reliable predictions the analysis presented here should be repeated including data on particle sizes larger than the current measurement limit of 500 microns.

4. Conclusion

In this paper, we have shown the evidence that a library of entropy groups could be built for a particular site, and the average shape of the spectrum could subsequently be estimated from measurements of the forcing parameters. Figure 3 shows how a classification tree analysis could be used to deduce the spectra group number from the concurrent values of ambient parameters. Potentially, this strategy may enable computation of average floc effective density, floc settling velocity, and floc fraction, hence providing valuable

information for a good range of engineering, environmental and ecological modelling applications.

Acknowledgements

We are indebted to the technical staff at SOS, and to the crew of RV 'Prince Madog'. Thanks are also due to POL in general, and to M.J. Howarth in particular, for organising and leading the Liverpool Bay monitoring. This paper utilises, with great thanks, materials obtained from CEFAS and POL. We also gratefully acknowledge the contribution of Dr. Sarah Jones (University of Bangor). Sadly, Sarah passed away after the primary data were collected.

REFERENCES

[1] Agrawal Y. C. & Pottsmith H. C. (2000) Instruments for particle size and settling velocity observations in sediment transport. *Marine Geology* 168: 89-114.

[2] Audry S., Blanc G. & Schafer J. (2006) Solid state partitioning of trace metals in suspended particulate matter from a river system affected by smelting-waste drainage. *Science of the Total Environment* 363: 216-236.

[3] Barros H. & Abril J. M. (2005) Constraints in the construction and/or selection of kinetic box models for the uptake of radionuclides and heavy metals by suspended particulate matter. *Ecological Modelling* 185: 371-385.

[4] Cranford P. J., Armsworthy S. L., Mikkelsen O. A. & Milligan T. G. (2005) Food acquisition responses of the suspension-feeding bivalve Placopecten magellanicus to the flocculation and settlement of a phytoplankton bloom. *Journal of Experimental Marine Biology and Ecology* 326: 128-143.

[5] Ebenhoh W., Kohlmeier C., Baretta J. W. & Floser G. (2004) Shallowness may be a major factor generating nutrient gradients in the Wadden Sea. *Ecological Modelling* 174: 241-252.

[6] Guo W., He M. C., Yang Z. F., Lin C. Y., Quan X. C. & Wang H. Z. (2007) Distribution of polycyclic aromatic hydrocarbons in water, suspended particulate matter and sediment from Daliao River watershed, China. *Chemosphere* 68: 93-104.

[7] Hakanson L. & Eckhell J. (2005) Suspended particulate matter (SPM) in the Baltic Sea - New empirical data and models. *Ecological Modelling* 189: 130-150.

[8] Hakanson L., Gyllenhammar A. & Brolin A. (2004) A dynamic compartment model to predict sedimentation and suspended particulate matter in coastal areas. *Ecological Modelling* 175: 353-384.

[9] Hakanson L., Mikrenska M., Petrov K. & Foster I. (2005) Suspended particulate matter (SPM) in rivers: empirical data and models. *Ecological Modelling* 183: 251-267.

[10] Hakanson L., Parparov A. & Hambright K. D. (2000) Modelling the impact of water level fluctuations on water quality (suspended particulate matter) in Lake Kinneret, Israel. *Ecological Modelling* 128: 101-125.

[11] He M. C., Sun Y., Li X. R. & Yang Z. F. (2006) Distribution patterns of nitrobenzenes and polychlorinated biphenyls in water, suspended particulate matter and sediment from mid- and down-stream of the Yellow River (China). *Chemosphere* 65: 365-374.

[12] Johansson H., Lindstrom M. & Hakanson L. (2001) On the modelling of the particulate and dissolved fractions of substances in aquatic ecosystems - sedimentological and ecological interactions. *Ecological Modelling* 137: 225-240.

[13] Johnston R. J. & Semple R. K. (1983) *Classification using information statistics, Concepts and Techniques in Modern Geography No. 37*. GeoBooks, Norwich

[14] Karrasch B., Parra O., Cid H., Mehrens M., Pacheco P., Urrutia R., Valdovinos C. & Zaror C. (2006) Effects of pulp and paper mill effluents on the microplankton and microbial self-purification capabilities of the Biobio River, Chile. *Science of the Total Environment* 359: 194-208.

[15] Krivtsov V., Gascoigne J. & Jones S. E. (2008a) Harmonic analysis of suspended particulate matter in the Menai Strait (UK). *Ecological Modelling* 212: 53-67.

[16] Krivtsov V., Howarth M. J. & Jones S. E. (2009) Characterising observed patterns of suspended particulate matter and relationships with oceanographic and meteorological variables: Studies in Liverpool Bay. *Environmental Modelling & Software* 24: 677-685.

[17] Krivtsov V., Howarth M. J., Jones S. E., Souza A. J. & Jago C. F. (2008b) Monitoring and modelling of the Irish Sea and Liverpool Bay: An overview and an SPM case study. *Ecological Modelling* 212: 37-52.

[18] Krivtsov, V., O. A. Mikkelsen, and S. E. Jones. "Entropy analysis of SPM patterns: case study of Liverpool Bay." *Geo-Marine Letters* 32.3 : 195-204.

[19] Lindstrom M. (2001) Distribution of particulate and reactive mercury in surface waters of Swedish forest lakes - an empirically based predictive model. *Ecological Modelling* 136: 81-93.

[20] Lindstrom M., Hakanson L., Abrahamsson O. & Johansson H. (1999) An empirical model for prediction of lake water suspended particulate matter. *Ecological Modelling* 121: 185-198.

[21] Maldonado C., Dachs J. & Bayona J. M. (1999) Trialkylamines and coprostanol as tracers of urban pollution in waters from enclosed seas: The Mediterranean and Black Sea. *Environmental Science & Technology* 33: 3290-3296.

[22] Malmaeus J. M. & Hakanson L. (2003) A dynamic model to predict suspended particulate matter in lakes. *Ecological Modelling* 167: 247-262.

[23] Malmaeus J. M. & Hakanson L. (2004) Development of a Lake Eutrophication model. *Ecological Modelling* 171: 35-63.

[24] Manjunatha B. R., Balakrishna K., Shankar R. & Mahalingarn T. R. (2001) Geochemistry and assessment of metal pollution in soils and river components of a monsoon-dominated environment near Karwar, southwest coast of India. *Environmental Geology* 40: 1462-1470.

[25] Mikkelsen O. A. (2002) Variation in the projected surface area of suspended particles: Implications for remote sensing assessment of TSM. *Remote Sensing of Environment* 79: 23-29.

[26] Mikkelsen O. A., Curran K. J., Hill P. S. & Milligan T. G. (2007) Entropy analysis of in situ particle size spectra. *Estuarine, Coastal and Shelf Science* 72: 615-625.

[27] Mikkelsen O. A., Hill P. S., Milligan T. G. & Chant R. J. (2005) In situ particle size distributions and volume concentrations from a LISST-100 laser particle sizer and a digital floc camera. *Continental Shelf Research* 25: 1959-1978.

[28] Orpin A. R. & Kostylev V. E. (2006) Towards a statistically valid method of textural sea floor characterization of benthic habitats. *Marine Geology* 225: 209-222.

[29] Shankar R. & Manjunatha B. R. (1994) Elemental Composition and Particulate Metal Fluxes from Netravati and Gurpur Rivers to the Coastal Arabian Sea. *Journal of the Geological Society of India* 43: 255-265.

[30] Shannon C. E. (1948) A mathematical theory of communication. *The Bell System Technical Journal* 27: 379-423 623-656.

[31] Sharp W. E. & Fan P. W. (1963) A sorting index. *Journal of Geology* 71: 76–84.

[32] Schrottke, K., Becker, M., Bartholomä, A., Flemming, B. W., & Hebbeln, D. (2006). Fluid mud dynamics in the Weser estuary turbidity zone tracked by high-resolution side-scan sonar and parametric sub-bottom profiler. *Geo-Marine Letters 26*(3): 185-198.

[33] Schwartz, R. & Kozerski, H. P. (2003) Entry and Deposits of Suspended Particulate Matter in Groyne Fields of the Middle Elbe and its Ecological Relevance. *Acta hydrochimica et hydrobiologica*, 31: 391-399.

[34] Woolfe K. J. & Michibayashi K. (1995) "BASIC" entropy grouping of laser-derived grain-size data: An example from the Great Barrier Reef. *Computers & Geosciences* 21: 447-462.

[35] Zhou J. L., Hong H., Zhang Z., Maskaoui K. & Chen W. (2000) Multi-phase distribution of organic micropollutants in Xiamen Harbour, China. *Water Research* 34: 2132-2150.