Classification and Mapping of Seabed Type from Deep Water Multi-Beam Echosounder (MBES) Data

Kazi Ishtiak Ahmed[1,4], Helen Caughey[1,4], Peter Hung[2,4], Paul Harris[1,4], Urška Demšar [1,4], Sean McLoone[2,4], A Stewart Fotheringham[1,4], Xavier Monteys[3], Ronan O'Toole[3]

[1]National Centre for Geocomputation, National University of Ireland Maynooth, Co.Kildare, Ireland
Email: {kazi.ahmed.2009, helen.m.caughey, paul.harris, urska.demsar, stewart.fotheringham}@nuim.ie

[2]Department of Electronic Engineering, National University of Ireland Maynooth, Co.Kildare, Ireland
Email: {peter.hung, sean.mcloone}@eeng.nuim.ie

[3]Geological Survey of Ireland, Beggars Bush, Haddington Road, Dublin 4, Ireland
Email: {xavier.monteys, ronan.o'toole}@gsi.ie

[4]StratAG – Strategic Research in Advanced Geotechnologies, http://www.stratag.ie
Email: stratag@nuim.ie

1. Introduction

Seabed mapping has a variety of applications, such as environmental research, marine and coastal resource management, oil/gas exploration etc. Multi-Beam Echosounder (MBES) is one of the most commonly used instruments for seabed surveys. This is a recent collaboration between StratAG (Strategic Research in Advanced Geotechnologies) and the Geological Survey of Ireland (GSI). GSI and the Irish Marine Institute (MI) have recently conducted extensive surveys in the seaboard of Ireland and the result of the survey is a massive MBES backscatter (scattering of sound waves) database.

The project described in this paper consists of two phases:

- In phase-1, spatial and statistical analysis was conducted to determine the integrability of the datasets (Caughey et al. 2010).
- Phase-2 consists of finding the best classification methodology for the data. This is work in progress, which we describe in this abstract.

The aim of this project is to develop a new methodology that will improve the seabed mapping process from deep water MBES data for GSI and potentially other marine research institutes.

2. Seabed type Classification from MBES Data

2.1 MBES: working principle and current approaches

The MBES systems transmit and receive an array of acoustic beams with individual small widths across the axis of the ship. The intersection of the transmit pulse and the receive-beams results in many simultaneous measurements across a wide swath. These measurements are then interpolated into an acoustic image (Augustin et al. 1994, Lurton 2002, Mayer 2006).

Seabed classification is achieved by the analysis of backscatter amplitudes. This is a complex process due to the inhomogeneity of seafloor as well as the sheer volume of the data (Arescon Ltd. 2001, Xinghua and Yongqi 2004). Automated classification, based on the statistical nature of the acoustic image, is currently the logical and common choice to achieve statistically valid and objective segmentations (Hellequin 1998, Hellequin et al. 2003, Cutter et al. 2003, McGonigle et al. 2009). Our data was generated using Multiview™, which uses a similar approach (Preston et al. 2001, Preston et al. 2004, McGonigle 2009).

A set of features, called Full Feature Vectors (FFVs), are generated using Multiview™ from the backscatter amplitudes (Collins and Preston 2002, Preston 2009). The result is a large matrix containing all the features extracted from the image. This matrix represents the FFV space, where each feature can be regarded as one new dimension/attribute.

2.2 Data Description

The deep sea survey was carried out over a period of three years (2000-2002) in the western seaboard of Ireland using three different pulse lengths (2000ms, 5000ms, and 15000ms) depending on the depth. Majority of the survey data comes from the survey year 2001 using 15000ms pulse length and this data set was used for the analysis and results presented in this paper. The study area extends between 57.4^0N to 46.7^0N and from 24.8^0 W to 9.25^0 W (Figure 1).

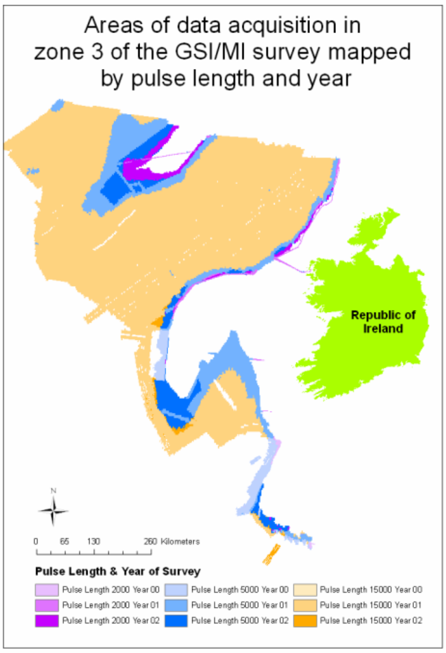

Figure 1. Map of data acquisition in zone-3 of the GSI/MI survey.

Each survey pass included five FFVs: Quantile measure (Q), 'Pace' textural feature (P), 'contrast' feature (C), mean (M), and standard deviation (S). The five features chosen for our analyses were pre-determined by GSI.

3. Seabed type mapping from deep water MBES data

3.1 Data integration based on spatial and statistical analysis

Datasets of pulse length 2000 was removed from the analysis after initial processing. It covered a small proportion of the survey area, consisted of hugely varied data and was considered unreliable in aspects concerning its means and consistency of acquisition.

A range of comparative analyses to assess the feasibility of joining the data were carried out on six overlapping spatially diverse data subsets from pulse lengths 5000 and 15000. These overlaps were not confined to one distinct spatial region (Figure 2).

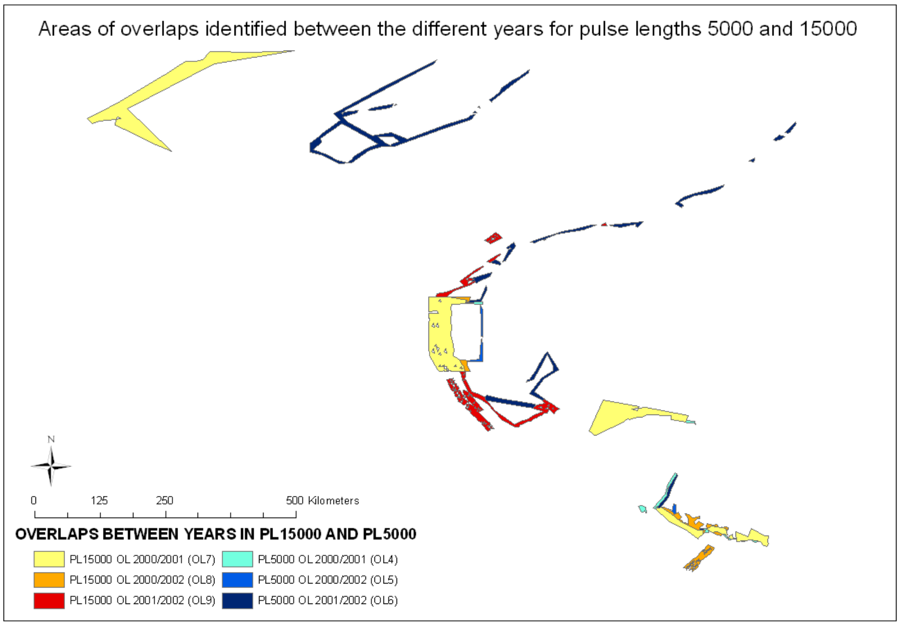

Figure 2. Overlap areas (e.g. OL4, OL5…) between different years and pulse lengths.

We performed a number of statistical tests (summary statistics, hypothesis testing, weighted/un-weighted mean errors, robust and weighted correlations etc.) on these overlaps. It was evident from these results (Caughey et al. 2010) that we cannot reliably join the datasets together and decided to keep them separate for phase-2.

3.2 Correlation analysis

In the next step we examined, through correlation analysis, if there exist any relationships among the FFVs. Global and local correlation analyses were carried out on the five backscatter features: Q, P, C, M, S; and the single bathymetry variable: 'Max_Slope'. From the global correlation matrices (Figure 3), it was evident that there exists a clear relationship between P and C. Using exponential regression, a negative exponent on C was estimated at -0.90. This was used to calculate a new feature variable C1 (i.e. $C1 = C^{-0.90}$), and regenerate a revised set of the global correlation matrices (Figure 4). As expected, P related fairly linearly to C1.

Next, the nature of the relationship between M and Q suggested that there may exist some composite measure for M and Q that may replace or supplement (one or both of) M and Q. In the coming weeks, we intend to investigate this by examining a prediction error variable: MQ1 = M-M*; where M* is predicted using M and Q. This analysis is currently ongoing.

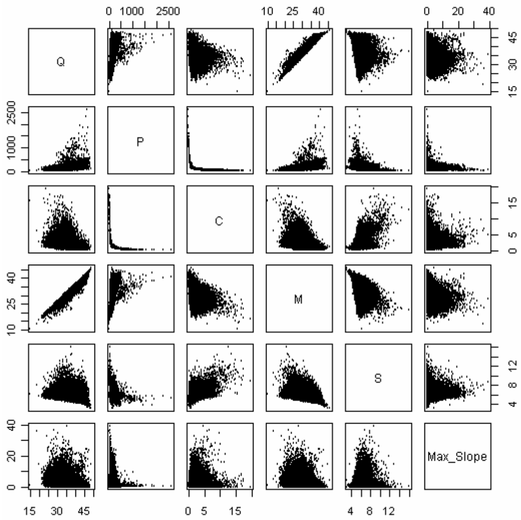

Figure 3. Global correlation matrix.

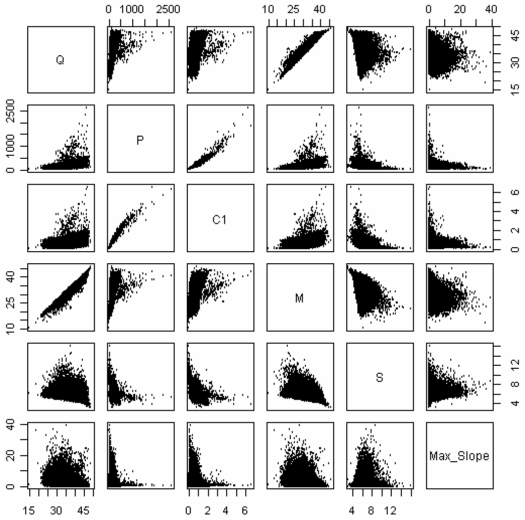

Figure 4. Revised global correlation matrix with C1.

These results will form the basis of selection of an appropriate subset of FFVs for subsequent classification.

3.3 Automatic classification and evaluation of classification quality

The next focus was to develop most suitable techniques for automatic classification of these datasets and combining classification results spatially. We are currently investigating three methods of clustering on the data and look for consistency in the clusters across methods.

A combination of principal component analysis (PCA) and *k*-means clustering, commonly used in acoustic seabed classification, will be attempted (Atallah and Smith 2004, Preston 2009). We will also test two more recent methods, Quality Threshold (QT) clustering (Heyer et al. 1999) and Spectral Clustering (SC) (Shi and Malik 2000, Ng et al. 2002, Luxburg 2007, Chen et al. 2010). This is a work-in progress and this paper focuses on the initial results obtained from SC and QT clustering on the dataset.

With QT clustering, unlike *k*-means, users do not need to specify the number of clusters required. Rather, they choose a maximum diameter for clusters, called QT distance. QT will only return clusters that pass a user-defined quality threshold. With no requirement for random cluster centroid initialisations, QT clustering always returns the same result for the same dataset and QT distance specified. In this paper, Euclidean distance measure is used as it is one of the most commonly employed.

SC method is a relatively new method which exploits pair wise similarities of data instances and is recently shown to be more effective than traditional methods such as *k*-means (Shi and Malik 2000, Dhillon 2001, Xu et al. 2003, Luxburg 2007, Chen et al. 2010).

Spectral clustering is graph-based clustering. It calculates a similarity matrix $S=[s_{ij}]$ between data points in attribute space, which represents the adjacency matrix of the similarity graph $G=(V,E)$, where V is the set of all data points $x_1,\ldots,x_n$. Two vertices in V, x_i and x_j, are connected if similarity s_{ij} between them is larger than a certain threshold – this means that the original similarity matrix linking all possible pairs of vertices is, thereby made sparse by considering most significant similarities. The similarity graph is then partitioned into k groups, so that the edges between groups have low weights and the edges within a group have high weights. This is done by calculating the Laplacian of the sparse similarity matrix and clustering its eigenspace using *k*-means clustering – the name spectral clustering comes from this (Luxburg 2007, Chen et al. 2010). The advantage of this method is that the clusters are based on local similarity in the attribute space and the method is therefore both faster as well as able to detect clusters that simple *k*-means would not recognise (e.g. non-convex clusters).

4. Results

To reduce the computational time during initial investigations, decimation is performed prior to clustering to reduce the size of data by a factor of 10 to around 100,000 samples. The use of both PCA and *k*-means failed to properly produce meaningful clusters, partly due to insufficient between-cluster distance in the principal component subspace. Nonetheless, it appears that the principal subspace can identify a number of outliers that is located outside the main sample 'cloud' (Figure 5).

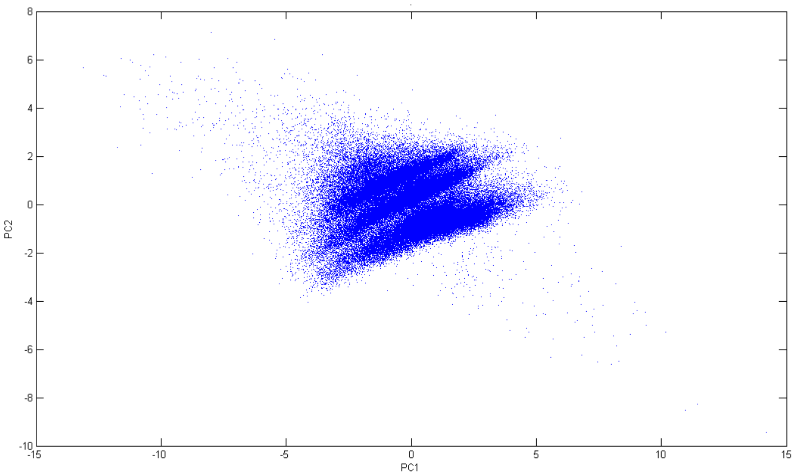

Figure 5. The MBES feature set projected on the first two principal components.

Figure 6 shows the results obtained from QT and SC. Here the attributes Q, P, C, M, and S are considered. For SC, in order to reduce computation time, five random samples of 10,000 records were generated from the dataset, normalized, and then SC was run on each of them. The results were consistent across samples. Figure 6 shows the results for 4, 6, and 8 clusters obtained from the first random sample.

For QT, the computational load is more involved. The dataset is divided into 100 subsets by random permutation. This is followed by 100 separate QT clustering and the results combined. Monte Carlo simulation is carried out 10 times and the final label for each sample is determined by majority voting. Since QT virtually always returns the cluster with decreasing number of samples, the labelling scheme are consistent, both within each data subset and among each Monte Carlo simulation. The QT distances chosen are 3, 2 and 1 which return 12, 20 and 40 clusters, respectively. They were then labelled into 4, 6, and 8 clusters for QT distances 3, 2 and 1 (Figure 6).

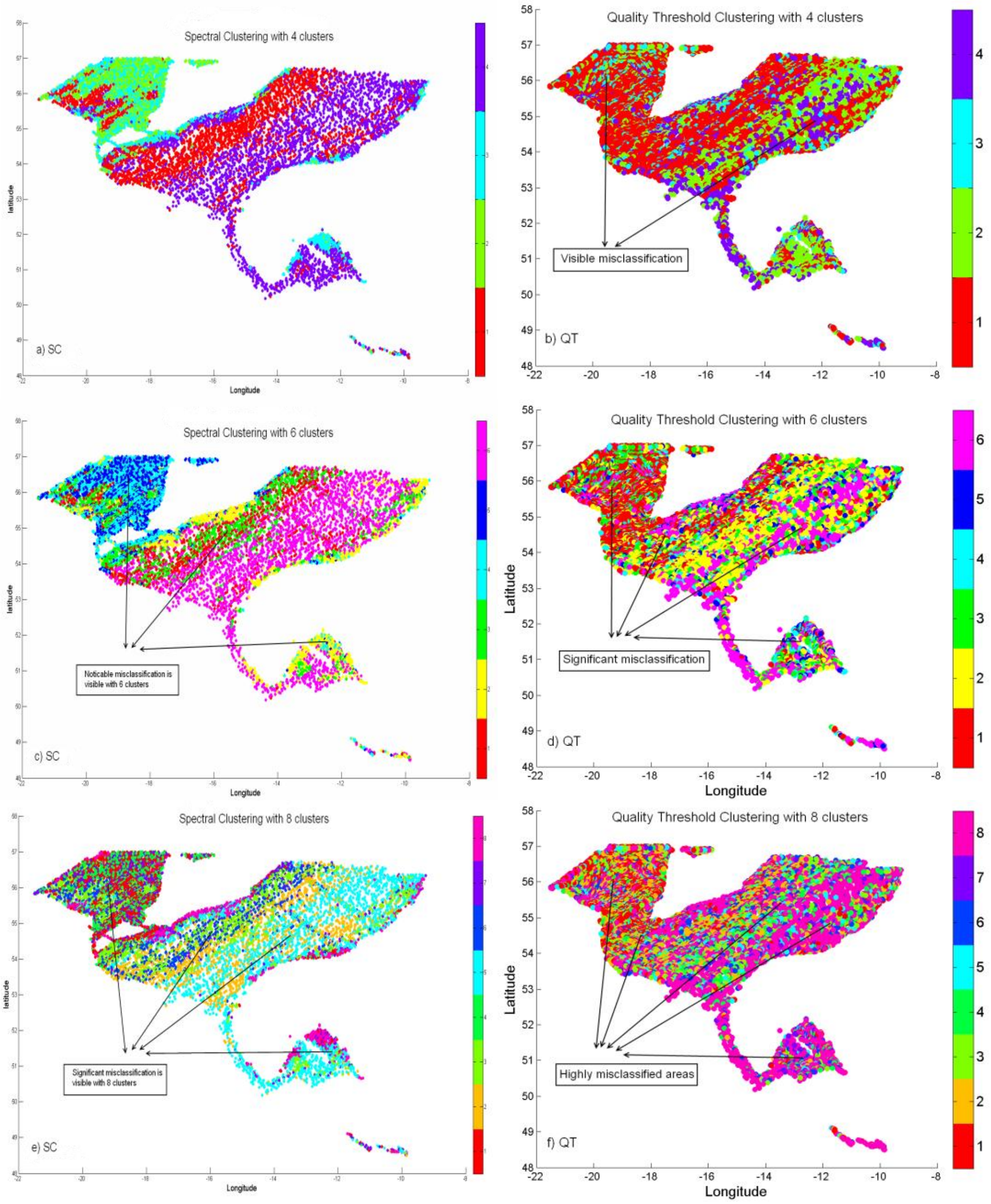

Figure 6. Spectral (SC) (using sparse similarity matrix) and Quality Threshold (QT) (using Euclidean distance) clustering results with 4 (6: a, b), 6 (6: c, d), and 8 (6: e, f) clusters.

From the figures, it seems that misclassification increases with the increased number of classes. In the deep waters, possibilities of variation in the geological features of the sea floor over a short distance are reasonably low. Therefore, it is expected that the datasets would have fewer number of distinct clusters as opposed to the land-cover above sea level. SC methods seemed to have performed well in defining the clusters but this will need to be further verified, while QT allows more robust detection of outliers. These preliminary results also indicate that the number of clusters should be limited to six or below. The labels returned by both methods have clear spatial correlations when compared like-to-like. The final number of classes will be decided upon after expert consultation.

5. Conclusions

In the coming weeks, we plan to use the completed correlation results (section 3.2) to reduce dimensionality. We will run QT and SC on different combinations of FFVs to obtain the best possible clustering as well as test the consistency of the clustering methods. The geological features (rocks, sediments, etc) corresponding to each cluster will then be determined by comparing the results, using expert knowledge, and comparison to results from other surveys.

The combined methodology of feature correlation and classification will undergo expert evaluation against ground truth information or existing seabed maps. Should this new methodology withstand the evaluation tests; it would be a useful new tool available to GSI and other marine institutes.

Acknowledgements

Research presented in this paper is supported by a Strategic Research Cluster grant (07/SRC/I1168) – StratAG, awarded to the National Centre of Geocomputation by Science Foundation Ireland under the National Development Plan. This project is a part of the collaboration of StratAG with the Geological Survey of Ireland (GSI) and INFOMAR, who generously provided the MBES data.

References

Arescon Ltd., 2001, An Approach to seabed classification from multi-beam bathymetric sonar data, (http://arescon.com/pics/mbeam.pdf), last accessed on 16th April 2010.

Atallah L and Smith P, 2004, Automatic seabed classification by the analysis of sidescan sonar and bathymetric imagery, *IEEE proceedings of the 'Radar, Sonar and Navigation'*, 151(5):327-336.

Augustin JM, Edy C, Savoye B and Le Drezen E, 1994, Sonar mosaic computation from multibeam echo sounder, *OCEANS '94, Proceedings of the 'Oceans Engineering for Today's Technology and Tomorrow's Preservation'. conference*, 2:433-438.

Caughey H, Ahmed KI, Harris P, Hung P, Demšar U, McLoone S, Fotheringham AS, Monteys X, O'Toole R, 2010, Developing a statistical methodology to improve classification and mapping of seabed type from deep water multi-beam echosounder (MBES) data, *Proceedings of the 'GIS Research UK' conference*, April, UK.

Chen WY, Song Y, Bai H, Lin CJ and Chang E, 2010, Parallel Spectral Clustering in Distributed Systems, *Accepted by IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010

Collins WT and Preston JM, 2002, Multibeam seabed classification, *International Ocean Systems*, 6(4):12-15.

Cutter JGR, Rzhanov Y, and Mayer LA, 2003, Automated segmentation of seafloor bathymetry from multibeam echosounder data using local Fourier histogram texture features, *Journal of Experimental Marine Biology and Ecology*, 285-286:355-370.

Dhillon IS, 2001, Co-clustering documents and words using bipartite spectral graph partitioning, *Proceedings of the SIGKDD*, California, USA, 269–274.

Heyer LJ, Kruglyak S, and Yooseph S, 1999, Exploring Expression Data: Identification and Analysis of Coexpressed Genes, *Genome Research*, 9(11):1106-1115.

Hellequin L, 1998, Statistical characterization of multibeam echosounder data, *Oceans'98, proceedings of the MTS/IEEE conference*, 1:228-233.

Hellequin L, Boucher JM, and Lurton X, 2003, Processing of high-frequency multibeam echosounder data for seafloor characterization, *IEEE Journal of Oceanographic Engineering*, 28:78-89.

Lurton X, 2002, *An Introduction to Underwater Acoustics: Principles and Applications*. Springer Verlag, Berlin-Heidelberg.

Luxburg UV, 2007, A tutorial on spectral clustering, *Statistics and Computing*, 7(4):395-416.

Mayer LA, 2006, Frontiers in seafloor mapping and visualization, *Marine Geophysical Researches*, 27:7-17.

McGonigle C, Brown C, Quinn R and Grabowski J, 2009, Evaluation of image-based multibeam sonar backscatter classification for benthic habitat discrimination and mapping at Staton Banks, UK, *Estuarine, Coastal and Shelf Science*, 81:423-437.

Ng A, Jordan M, and Weiss Y, 2002, On spectral clustering: analysis and an algorithm. In Dietterich T, Becker S, and Ghahramani Z (eds), *Advances in Neural Information Processing System*, 14:849-856, MIT Press.

Preston JM, Christney AC, Bloomer SF, and Beaudet IL, 2001, Seabed Classification of multibeam sonar Images, *Ocean's 01,proceedings of the MTS/IEEE conference, Honolulu, USA*, 4**:**2616-2623.

Preston JM, Christney AC, Collins WT, and Bloomer S, 2004, Automated acoustic classification of sidescan images, *Oceans'04, proceedings of the MTS/IEEE Techno-Ocean conference, Kobe, Japan*, 4**:**2060-2065.

Preston JM, 2009, Automated acoustic seabed classification of multibeam images of Stanton Banks, *Applied Acoustics*, 70(10): 1277-1287.

Shi J and Malik J, 2000, Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Xinghua Z and Yongqi C, 2004, Seafloor sediment classification based on multibeam sonar data, *Geospatial Information Science*, 7(4):290-296.

Xu W, Liu X and Gong Y, 2003, Document clustering based on non-negative matrix factorization. *Proceedings of the AGM SIGIR*, Toronto, Canada, 267–273.