Statistical Clustering Of Curves In The Geosciences – The Answer To Everything?

L.J. Hamilton
Defence Science & Technology Organisation (DSTO)
13 Garden Street
Eveleigh, NSW 2015, Australia

Abstract. **Classification and exploration of large geodata sets consisting of tens to hundreds of thousands of single valued curves, e.g. oceanic wave spectra and grain size distributions, is often made through use of features or proxy parameters extracted from the curves. Feature extraction is typically enabled through curve fitting, and hence implicit or explicit application of a statistical model. Principal Components Analysis (PCA) is commonly applied to the proxies or to the curves as a dimensional reduction measure, and the first few principal components are used as the final classification features. Statistical clustering is then applied to the selected proxies or principal components to produce groups or classes which best describe the properties of the proxy data set, usually in a least squares sense. A far simpler and model free technique is to directly cluster the curves themselves. The curves are essentially treated as geometric entities, and calculation of features or proxies is unnecessary. The methodology for this concept is outlined, and is demonstrated for several geodata sets, ranging in size from a hundred objects to several tens of thousands, and for spatial scales of several tens of metres to the South Pacific ocean basin.**

I. INTRODUCTION

Many data sets consist of measurements which can be regarded as single valued curves. Curves for the present application are defined as sets of connected points for which order is important (abscissa values increase monotonically), and for which the ordinate values all have the same units. Examples are wind wave energy spectra, grain size distributions, and seabed acoustic backscatter response with angle to the output signal of multibeam sonars. Data sets may consist of hundreds of curves to hundreds of thousands of curves, and each curve may have from tens to hundreds of values.

The "curse of dimensionality" is often invoked for these curves, and dimensional reduction is achieved through feature extraction, which is the calculation of proxy parameters for the curves such as means, slopes, skewness, kurtosis, values and positions of modes, quantiles, Fourier transform components, and any other number of statistical parameters. Principal Components Analysis is then used, either on the curves themselves, or more usually on features. The first few principal components are placed into vectors, and these statistical objects are regarded as vectors of co-ordinates in a multi-dimensional space. Statistical clustering is then made on the vectors of features. The clustering groups the data into several sets or classes having the same general properties.

This processing sequence can introduce distortions, losses, and biases to the data, particularly when curve fitting of multimodal data is involved. A far simpler and model free technique is to directly cluster the curves themselves. In this direct approach the curves are essentially treated as geometric entities, and calculation of proxies (feature extraction) is unnecessary. Use of the whole curve in grouping or clustering potentially provides a better description in some statistical sense (e.g. least squares) than simplified representations or segmentations.

Although it is a very simple concept, classification of single valued curves by direct clustering does not seem to be generally known in the literature. This may be because some classification and clustering methods are influenced by self-correlation effects in adjacent data values, and are not suitable techniques for direct clustering of curves. Additionally, some clustering metrics may not be suitable for curves, and some types of curves are not amenable to clustering. Largely, however, it seems not to have occurred to researchers that direct clustering of curves was possible, particularly when high dimensionality is involved. The concept of feature extraction and reduction of dimensionality has been all pervasive, but it may not always be necessary.

The present paper aims to introduce the concept of direct clustering of curves to a wider audience. The methodology for direct clustering of curves developed in Hamilton [1][2] is outlined. The usefulness of the technique is demonstrated by application to several different data types, different spatial scales, and to time series data.

II. METHODOLOGY FOR DIRECT CLUSTERING OF CURVES

The CLARA Statistical Clustering Algorithm

The statistical clustering algorithm employed is the CLARA (Clustering LARge Applications) algorithm of Kaufman and Rousseeuw [3]. CLARA is a general clustering algorithm, and was not designed to cluster curves. It was shown to be suitable

978-1-4244-5222-4/10/$26.00 ©2010 CROWN

for this purpose by [1][2], subject to some limitations. This approach differs from the usual applications of clustering, which consider data objects as vectors of co-ordinates in a multi-dimensional space, not as curves or sets of connected points. For the present application the number of dimensions is equal to the number of points in each curve. CLARA is designed to cluster a minimum of 100 objects. FORTRAN code for CLARA was obtained from the internet at http://lib.stat.cmu.edu/general/clusfind. For R code see http://stat.ethz.ch/R- manual/R- devel/library/cluster/html/ clara.html. CLARA is also available in several statistics packages.

A Brief Description Of Statistical Clustering

Clustering is a tool of multidimensional analysis which partitions a population of objects into groups having some particular measure of similarity. Clustering uses multidimensional distance metrics, e.g. CLARA implements the Euclidian ($\sqrt{(\sum_i(c_i-d_i)^2}$) or Manhattan ($\sum_i|(c_i-d_i)|$) metrics to decide cluster membership, where in the present paper c_i and d_i are curves, e.g. wind-wave spectra with i frequency bins. Typically samples are first randomly assigned to clusters, and are then re-assigned to other clusters only if this minimises within-cluster variability and maximises between-cluster variability. Clustering stops when reorganisation of samples fails to produce a statistical improvement in the assessment method. CLARA provides one curve from each cluster, termed a medoid [3], as representative of the cluster properties.

CLARA has an option to standardize data in each dimension, or to leave it non-standardised. Optimal configuration for use of CLARA to cluster large data sets consisting of curves was established by [1] as Manhattan distance metric with non-standardisation of parameters. Non-standardisation maintains the relativity of scale between curve values, which is desirable since for the present application they all have the same units. Outliers may be isolated using standardisation of parameters. Outliers should be removed before final clustering to prevent the formation of single member clusters, or very small clusters, which reduce the definition of the other clusters.

Subsampling Techniques Of The CLARA Algorithm

Many clustering algorithms require too much processing power, computer memory, or processing time to be tenable for analysis of large data sets. Software program CLARA overcomes these problems by coupling statistical sampling and clustering techniques. The algorithm first clusters several sets of randomly chosen subsamples (typically five sets), then uses the particular subsampling returning best results in a least squares sense to cluster the entire data set. This provides a fast algorithm, at the possible expense of accuracy. Sensitivity tests by the present author have shown the sampling scheme is robust, and that inaccuracy arising from the CLARA sampling methods is not an issue past some critical (and comparatively low) choice of number of objects in the subsamplings.

Using CLARA to process over 20,000 objects can take from a few minutes to 4 to 6 hours depending on the sampling scheme used. Testing for 20,000 objects indicates that much the same results are returned whether subsamples of the data set contain small (hundreds) or large (thousands) of objects, so that reliable initial results can be returned in very short processing times. This provides a quick-look facility for examination of large data sets [1].

Production Of Clusters With Small Populations

A key point is that testing also indicates CLARA can return very small clusters from large data sets consisting of curves [1][2], with cluster membership effectively determined by the geometrical properties or shapes of the curves or objects, not by numerical dominance of particular shapes. Standardisation enhances this effect, allowing outlier identification. In one case for cumulative grain size curves, forming 10 clusters with the standardised option and Manhattan metric isolated four convex curves to one cluster, from a total number of 1281 objects, other curves being concave [1]. Provided a sufficient number of clusters are formed, this behaviour of CLARA ensures that real characteristics of large data sets are not lost.

The Number Of Clusters

CLARA provides a 'Silhouette Coefficient' for the user to estimate the optimal number of clusters. Progressively more clusters are requested from two upwards, until the highest value of the coefficient has been passed. This highest value in principle occurs for the optimal number of clusters. However [1] demonstrates that CLARA can sensibly form a much greater number of clusters than indicated by the Silhouette Coefficient, even when distinctly different clusters do not exist. Instead of relying on statistical indicators of cluster numbers, the user is recommended to cluster from two groups upwards until it is considered that no more useful information is obtained in the clustering space, or e.g. in the geospatial domain. Use of CLARA in quick-look mode aids this task.

The effectiveness of the clustering and assessment of the number of clusters chosen can be checked by examining overplots of the curves forming each cluster for uniformity of properties (shape, location, and central tendency), and by examining overplots of cluster medoids to check they differ appreciably from each other for the particular application. It is usually better to request many clusters, rather than a few, for large data sets. If too many clusters have been requested, indicated e.g. if closely related curves begin to be placed into different clusters, then similar clusters can be amalgamated. Not requesting enough clusters can force curves with different properties into the same cluster, and can suppress outlier

detection. Data exploration is an essential part of any examination of large data sets.

CLARA Clustering Of Quasi-continuums

When data sets contain distinct groupings of curves very different in geometrical shape or location from each other, the clustering technique returns results which can be regarded as absolute, in the sense that different observers will agree that only one set of clusters should be formed, and that any particular curve can obviously belong to only one particular cluster. However, this is not always the case. Properties may evolve, rather than jump from one state to another, with curves then effectively forming a quasi-continuum in the cluster space. References [1][2] demonstrate that CLARA can produce meaningful clusters in this situation. An entropy clustering algorithm did not always satisfy this condition.

Types Of Curves Unsuitable For CLARA Clustering

The Manhattan distance metric is a very simple measure of similarity or dissimilarity, which for the present application essentially measures the area of non-overlap between two curves. This metric does not work well to classify distributions containing curves which have minimal to no overlap with other curves [1]. An extreme example is for curves consisting of a single spike (i.e. having data in only one bin), where no two curves have the spike in the same bin, and where each spike has numerically the same value. This can conceivably occur for grain size data and other physical phenomena. The distance between any two objects in this example is statistically the same for the Manhattan (or Euclidian) metric, and reliable separation or classification is not possible. For datasets of this type another metric should be used (entropy is suitable), or the data should be transformed, e.g. by forming cumulative curves [1].

Independence Of Classifications On Spatial Location And Spatial Scales

Many oceanographical data sets sample a spatial field. The present methodology does not input the geographical locations of curves to the clustering. Curves are grouped solely by their geometrical properties, *independent of spatial considerations*. It is expected, although not guaranteed, that profiles assigned to a particular cluster will generally group together in a particular spatial area or spatial configuration. Charting the classifications in geographical space then provides a spatial classification free of the restraints and biases imposed by grid construction. A range of spatial scales can be processed and revealed in a single clustering operation. If spatial decorrelation scales or grids are required for some purpose, they can be estimated from the geographical configurations of clusters, providing a more natural analysis sequence than attempting to estimate them beforehand, the procedure followed when gridding is used. Decorrelation scales are then not restricted to fixed values or orthogonal directions. Nor is a particular area or spatial bin restricted to a single classification, since spatial averaging is not made. For oceanographic applications, variability in areas with shifting fronts, interleaving, or mixing of water types is not removed, but remains available to the analysis.

Constrained clustering methods known as regionalised analyses incorporate geographical positions and gridding in geostatistical interpolation methods. These methods may assist spatial prediction and map making, but whether or not an object is similar to another object should be a function of the properties of the objects, not of their relative geographical position. Linear features with a short width scale and a large length scale may not be well characterised by grids. Properties of water bodies on opposite sides of fronts (or geological properties across faults) may bear little or no relation to one another. If the fronts move, then gridding can not adequately characterise the data.

Other Factors

Different statistical clustering algorithms and metrics may produce different partitionings of a data set, and a particular algorithm may work well for some data curves, but not for others.

III. EXAMPLES

Seabed Echoes – Acoustic Seabed Classification

Acoustic seabed classification is the process of inferring seabed type by its acoustic backscatter response to different types of active sonars, e.g. echosounders. The commonly used method to obtain acoustic seabed classification from echosounder returns (seabed echoes) is to use feature extraction, Principal Components Analysis, and statistical clustering or simulated annealing [4]. One commercial system forms 166 features [4]. However, direct clustering of the echoes is a more effective procedure.

Echosounder data were obtained from Balls Head Bay, a small embayment in Sydney Harbour, Australia with a wide range of seabed types. Echoes are compensated for attenuation and beam spreading effects, and standardized to a common seabed depth [4]. Amplitude or energy normalization is then applied. Along track averaging of echoes is used to reduce noise and variability (3149 averaged echoes were formed).

Medoids for an 8-clustering have different shapes and different locations of their mode, and span the clustering space efficiently (Fig. 1). Overplots of echoes in each cluster have good central tendency (Fig. 2), allowing confidence to be

placed in the statistical classification. A coherent seabed classification was produced (Fig. 3) in agreement with groundtruth.

It is easily deduced from the clustering results that the principal shape parameters appearing to govern the classification are echo width, time from the echo start to the echo peak (rise time), and echo peak height. These parameters have well established relations to the seabed backscattering process. For example, backscatter curves with longer decay tails are associated with rougher seabeds. Echoes with short rise times are typically sandy sediments. This type of knowledge allows more informed feature extraction. A full discussion is given in [5].

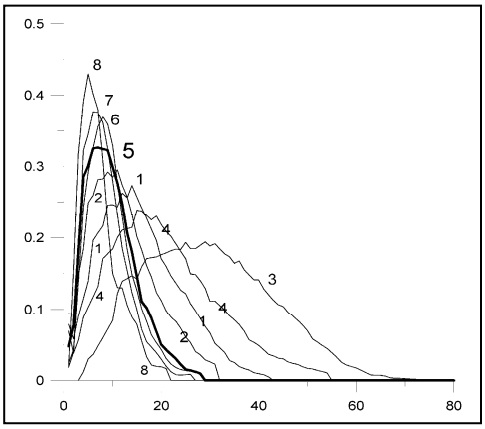

Figure 1. Medoids (representative curves) for an 8-clustering of seabed echoes in Balls Head Bay, Australia.

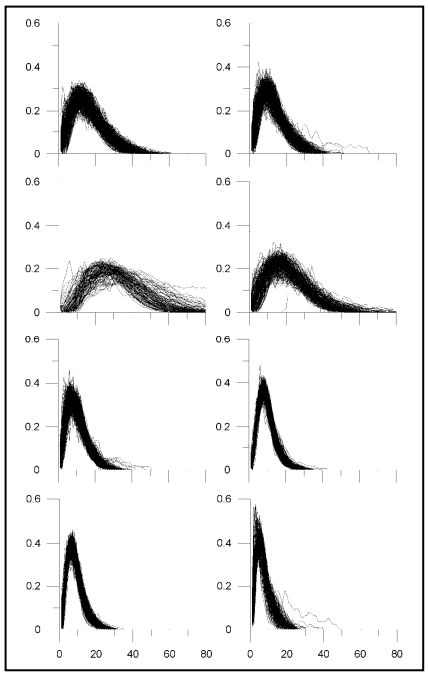

Figure 2. Overplots of seabed echoes for an 8-clustering of seabed echoes for Balls Head Bay.

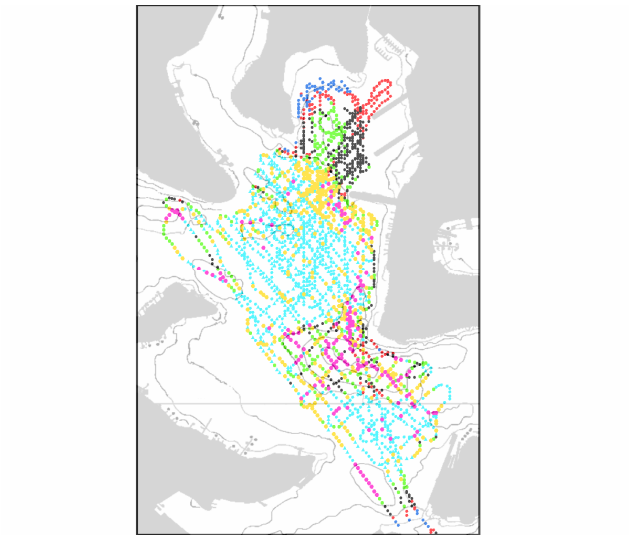

Figure 3. Eight acoustic classes for Balls Head Bay. The mapped area is about 1500 m x 750 m.

Multibeam Sonar Seabed backscatter Curves – Acoustic Seabed Classification

Multibeam sonars estimate bathymetry from time of flight of hundreds of acoustic beams which ensonify the seabed in a thin swathe across track. Some multibeams also record the backscatter received by the beams. For the present application it is assumed the backscatter data have been fully corrected for attenuation, system parameters and gains, effect of local seabed slope and other factors. High variability in the preprocessed backscatter curves is removed by along track averaging prior to clustering. Backscatter data were obtained with a RESON 8125 multibeam sonar in a complex area of inter island sand, reef, seagrass, and rhodolith beds in Esperance Bay, Western Australia.

Processing of multibeam backscatter data to infer seabed type is usually made by feature extraction and dimensional reduction methods, followed by clustering. Image processing methods may also be used, and typically also use dimensional reduction methods [4]. Modelling of the backscatter process is showing promising results, but is not yet an exact science. Supervised classification has been used, but problems occurred for slopes, heterogeneous areas, and habitat or grain size boundaries.

Direct clustering of the 37,484 average multibeam sonar backscatter curves provided results in about 10 minutes. Multibeam processing is often limited to five or six clusters. Use of 10, 20, 30, and 50 clusters all provided useful results. Medoids for the 10-cluster case are shown in Fig. 4. The medoids are more closely spaced at high backscatter values where the curves exhibit greatest variability. The spatial coherency of patterns in the 10-clustering of Fig. 5 is

remarkable. This made it easy to assign seabed descriptors to the acoustic classes. Outliers were revealed, and the clustering was able to map boundaries between some seabed types, a stumbling block for some other methods.

Full details are given in [6].

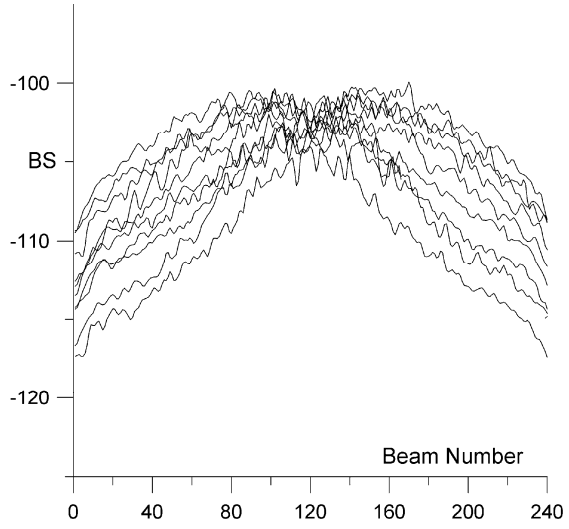

Figure 4. Medoids (representative backscatter curves). For a 10-clustering of multibeam scans for Esperance Bay.

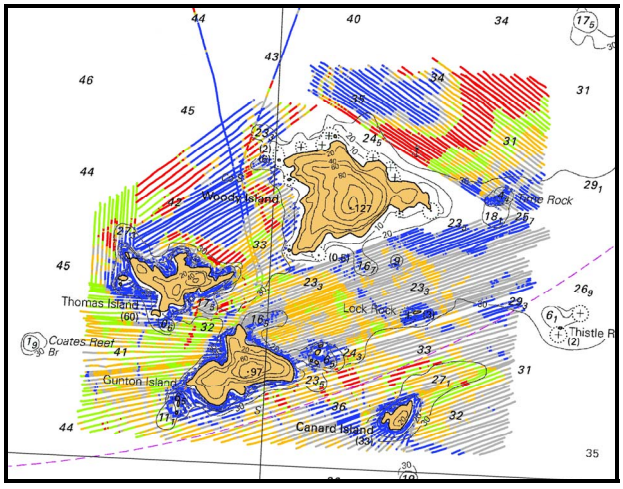

Figure 5. Acoustic seabed classification for Esperance Bay from direct clustering of multibeam sonar backscatter scans. The area is roughly 6.3 km wide and 4.8 km in north-south extent.

Grain Size Distributions – Local Scale

Following difficulties experienced for a manual clustering of 119 high resolution cumulative grain size curves from Sydney Harbour (Fig. 6), the curves were clustered into 20 groups (Fig. 7). The samples range from high clay, to muds, muddy sands, sandy muds, clean sands, and shell gravels. Interpolation was used to produce 76 phi sizes from -5.52 to +9.48. The CLARA clustering was similar to, but better than, the manual clustering. The Silhouette Coefficient estimated only three clusters were present in the data, but the 20 clusters are all well formed. A full discussion is provided in [1].

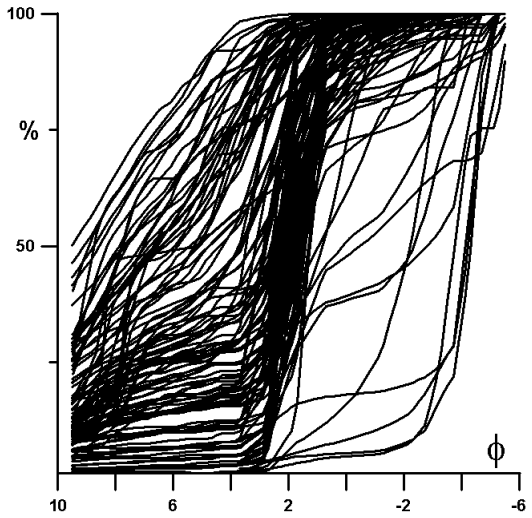

Figure 6. Cumulative grainsize curves for 119 seabed samples from Sydney Harbour.

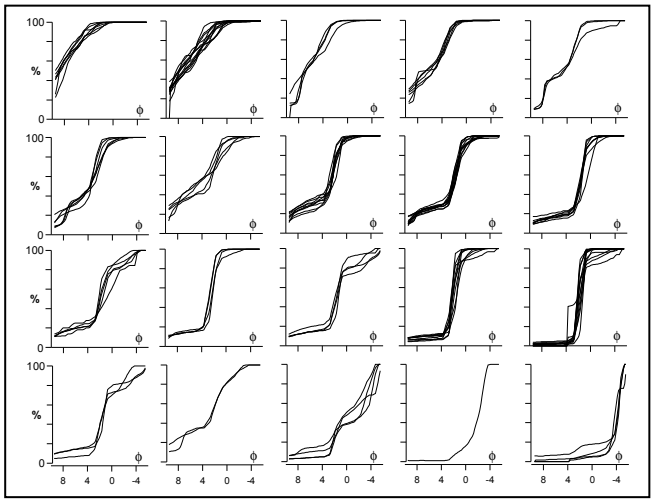

Figure 7. A 20-clustering of the 119 cumulative curves of Fig. 6.

Grain Size Distributions – Regional Scale

A well known data set of 1281 seabed samples from the Darss Sill area of the Baltic Sea was clustered with the CLARA algorithm. The samples are predominantly silts and sands, and most of the cumulative grain size curves have similar shape (Fig. 8). Curves have eight size fractions only. Ten clusters with standardisation of variables isolated a four-member cluster containing curves with convex curvature, other curves being concave [1]. An entropy metric could not match this behaviour. A 20-clustering produced very tight groups (Fig. 9). Spatial mappings showed consistent patterns for two clusters upwards (Fig. 10), matching known seabed characteristics. See [1] for more details.

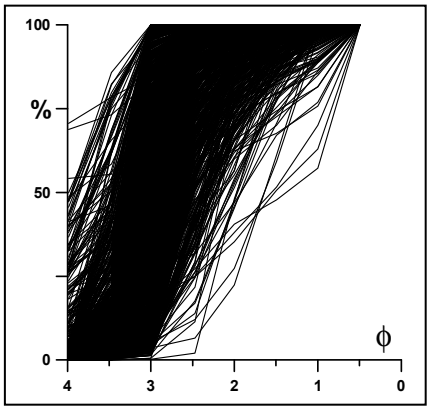

Figure 8. Cumulative grainsize curves for 1281 seabed samples from Darss Sill.

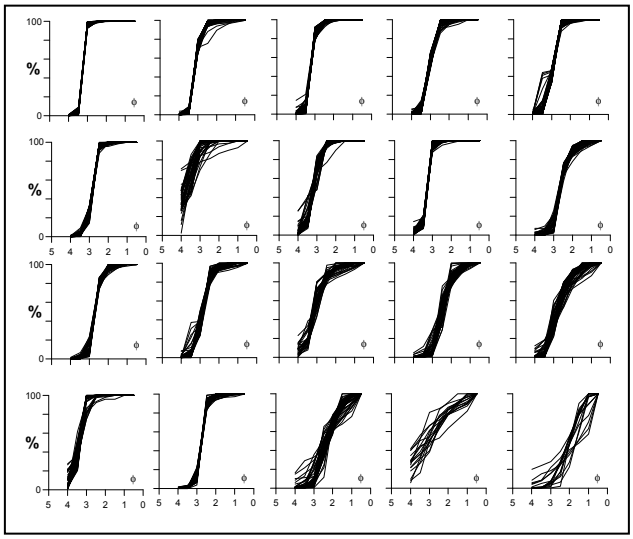

Figure 9. A 20-clustering for the 1281 cumulative curves of Fig. 8.

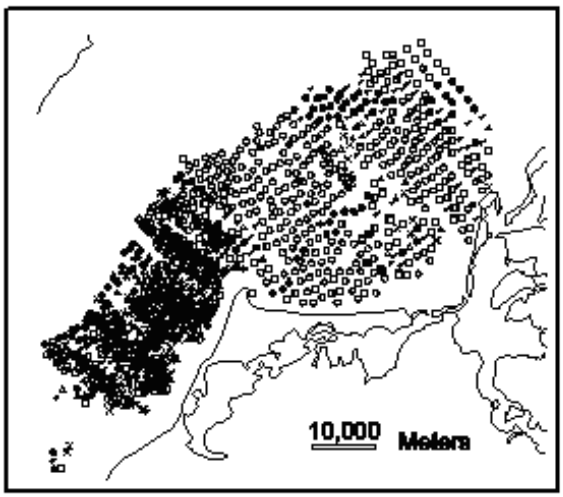

Figure 10. Distributions of eight clusters for Darss Sill obtained by a direct clustering of cumulative curves.

Grain Size Distributions – Continental Shelf Scale

Is the direct clustering technique of any real benefit when applied to very large grain size data sets? It is possible that too much variation occurs in curve shapes to provide anything other than a coarse analysis. The ECSTDB2005 data set for the continental shelf of the eastern USA [7] has 21,399 useable samples at phi sizes of -4 to +11 (Fig. 11). Because of the dominance of pronounced modes in this data, clustering was made on the frequency data, not on cumulative curves, following findings of [2]. Manhattan metric and non-standardisation of variables were used, for five subsamplings of 7000 samples. Ten 20, and 30 clusters were formed.

Overplots for a 30-clustering have good central tendency, sharply depicting the prevalence of pronounced modes in the data (Fig. 12). Geographical distributions of clusters show coherent patterns over a wide range of spatial scales.

One sequence of medoids (22-13-14-19-25-15) in the 30-clustering with the same dominant mode (phi = 3) indicated a possible sequence of properties from a clean fine sand to a sediment with increasing fines. These clusters occurred in geographical proximity to each other in two scenarios, one predominantly along the coast and another along the shelf break. However, over wide areas elsewhere some quite different clusters share the same general geographical distribution. Whether this means spatial heterogeneity, or that samples need a finer clustering, or a finer spatial examination, has not been examined.

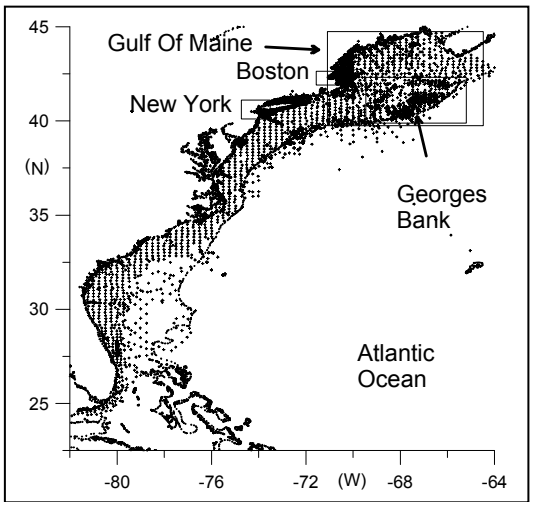

Figure 11. Distribution of 21,399 seabed samples on the US east coast.

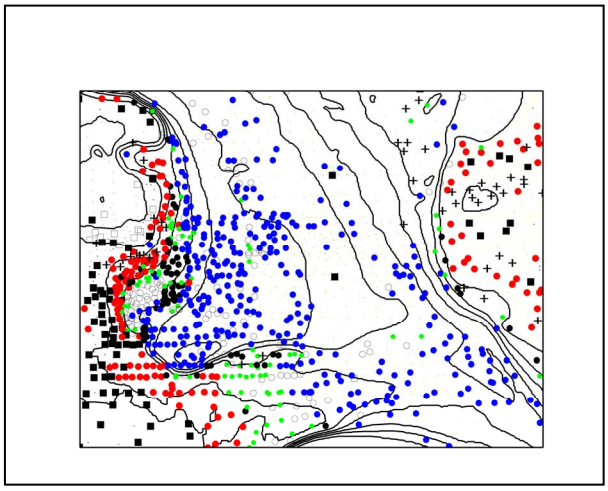

Figure 13. Cluster mapping for Stellwagen Bank, 25 km east of Boston. Clusters are aligned with depth contours on both eastern and western flanks of the bank. The bank is about 30 km in width.

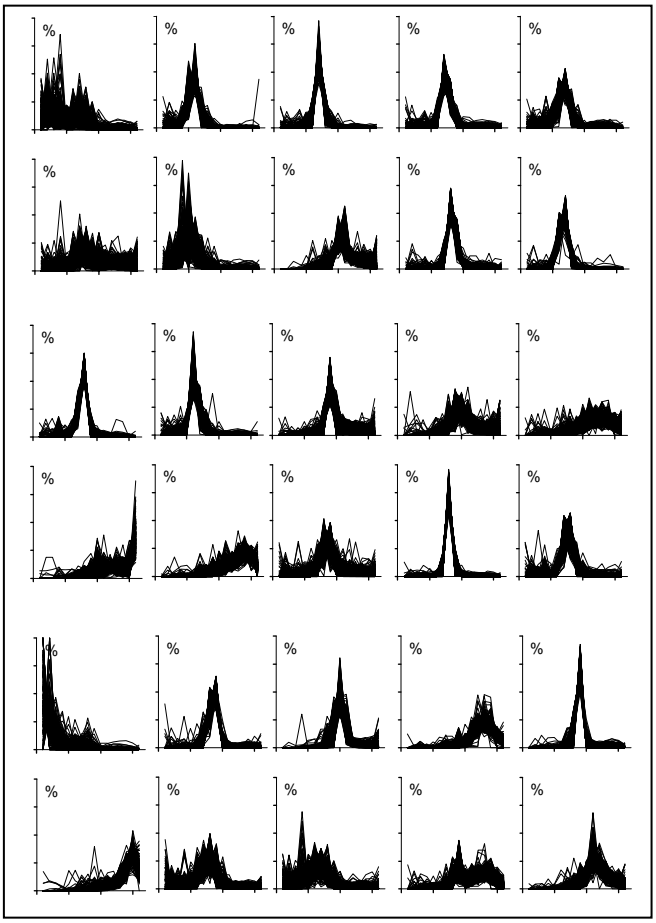

Figure 12. A 30-clustering of grain size curves for the 21,399 samples of Fig. 11. Clusters are numbered from 1 to 30 from the top left and across.

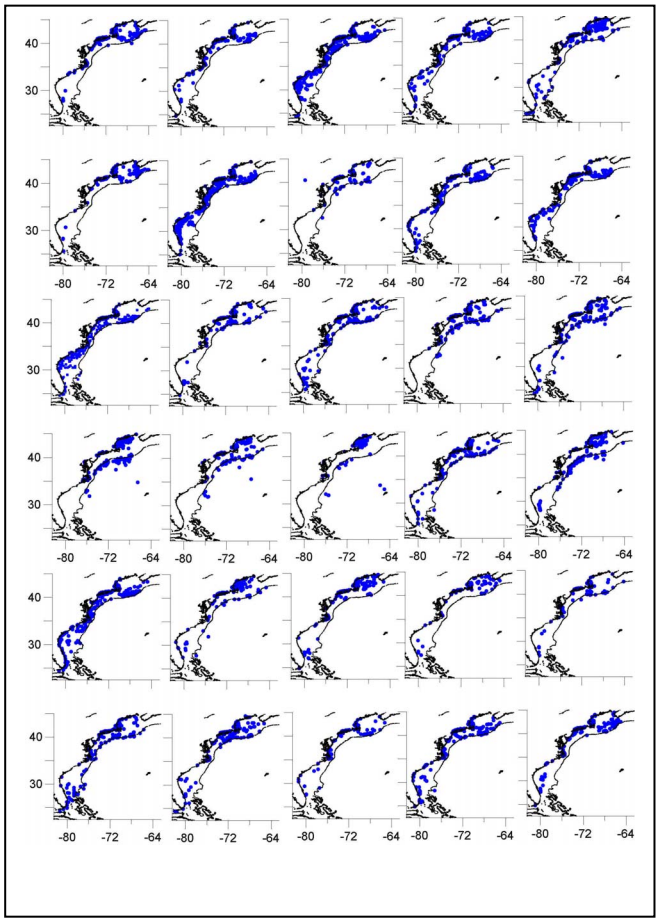

Figure 14. Distribution of the 30 clusters of Fig. 12 for the US east coast.

Over Stellwagen Bank (Fig. 13), the clustering provides far higher resolution of sediment patterns than grain size triangle classifications, which classify the sediments as sand. Particularly on the western side of Stellwagen Bank, the cluster distributions are closely aligned with the bathymetry, and indicate a downslope progression in sediment type (the bank deepens to both east and west) (Figure 13).

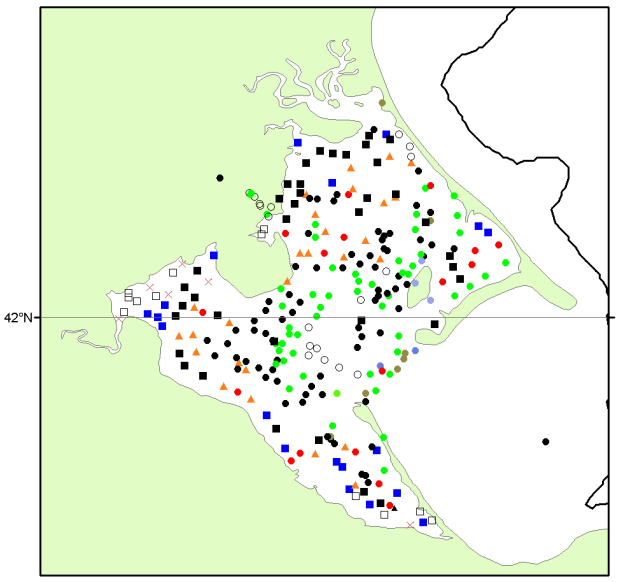

Figure 15. Cluster mapping for Plymouth Bay. The chart area is 10 km x 10 km.

The 30-clustering returns a pattern of sediments in Plymouth Harbour which exhibit spatial coherence at small scales, e.g. along the northwestern shores (Fig. 15).

For this large data set, a 30-clustering of the 21,399 grain size distributions has apparently produced a mapping useful at a wide range of scales. This may not always be the case, e.g. highly sampled smaller areas may require more detailed examination than coarsely sampled large areas, or may have different grain size distributions. However, it is of benefit when a large area has irregular spatial sampling. This data set illustrates the essential immunity of curve based clustering to spatial scales. This is not the case for mappings beginning with gridding methods.

Oceanographic Profiles – Ocean Basin Scale

Antarctic Intermediate Water (AAIW) is a water mass believed to be formed mainly in the Pacific Ocean off the southwest of South America. It occupies an extended vertical interval in the water column. Formation and circulation of AAIW are key components of the ventilation of the world oceans and the global climate. Through its large volumes and ubiquitous travel paths in the oceans AAIW forms a major agent in returning fresh water to the tropics from subpolar regions and in balancing the heat input to the oceans.

The distinguishing characteristic or core property of the AAIW is its salinity minimum in the vertical, acquired in the rain affected formation area of subpolar latitudes. As AAIW cools, sinks and moves through the oceans as a thermohaline circulation (rather than a wind driven one), the minimum salinity increases by mixing with higher salinity waters, both horizontally and vertically. A method known as core layer analysis assumes that horizontal mixing is dominant, and charts the salinity minimum on a single surface. The flow path of AAIW is then inferred by following increases in the salinity minimum on the chart.

In using only a single inflection point from each profile to construct a horizontal property surface, core layer analysis forms an effective data reduction procedure for AAIW examinations. However, much valuable information is discarded, including the three dimensional nature and interactions of water masses. Having the same core salinity does not mean that column properties are the same.

Clustering of salinity-density curves can be used to classify the column properties of AAIW, and to trace its path. As for the core layer method the results can be projected onto a surface, but the projection now indicates representative vertical structure, not a single point value. Cluster medoids can be used as building blocks to trace the evolution of AAIW over its entire vertical density range, rather than using points on a single surface.

A full explanation of the semi-automated water mass analysis techniques enabled by the clustering is beyond the scope of the present paper. Here we are simply interested in showing that it can be done. The data set contains 16,764 Conductivity-Temperature-Depth (CTD) profiles from Argo profiling floats and ships in the NODC database. Because the clustering is shape based, the potentially lower accuracy Argo profiles can be used with the laboratory calibrated ship data with a fair degree of impunity. This circumvents a great deal of exploratory error analysis. The spatial results derived from the clustering show if the two data sets are compatible enough to be used together. Error analysis can be made through image processing techniques involving estimates of spatial coherency and degree of speckle.

A family of medoids for one AAIW path in the South Pacific Ocean is shown in Fig. 16, and the inferred path is shown in Fig. 17. Fig. 18 charts various clusters from other paths. Coherent patterns are obtained which are aligned with various major circulations, e.g. equatorial gyre, main gyre, and the Antarctic Circumpolar Current. Note the extended basin wide distributions of some clusters, and the variegated shapes of others.

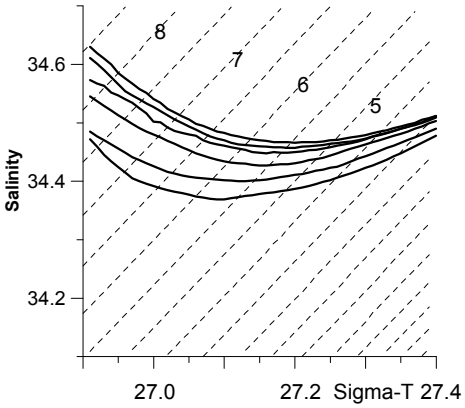

Figure 16. A family of similarly shaped Salinity-density medoids, indicative of a flow path of AAIW. The AAIW salinity minimum increases in the direction of flow.

Paths may be constructed from geographically adjacent clusters with similar medoids, for which the salinity minimum increases in value along the path. The clustering provides a new and simple technique for water mass analysis, with far more capability than core layer analysis. A full analysis of the data has not been finalized.

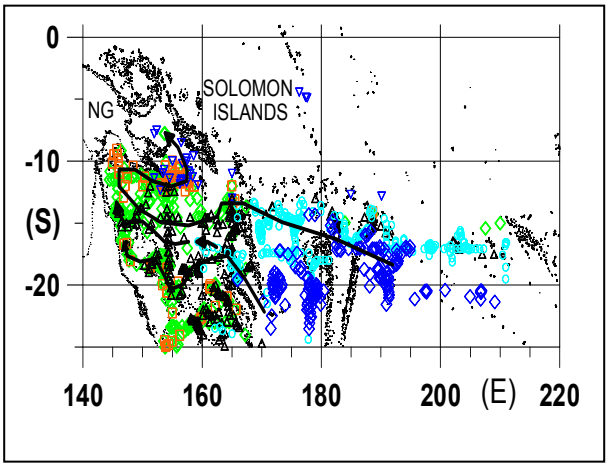

Figure 17. The flow path for the Salinity-density medoids of Fig. 16.

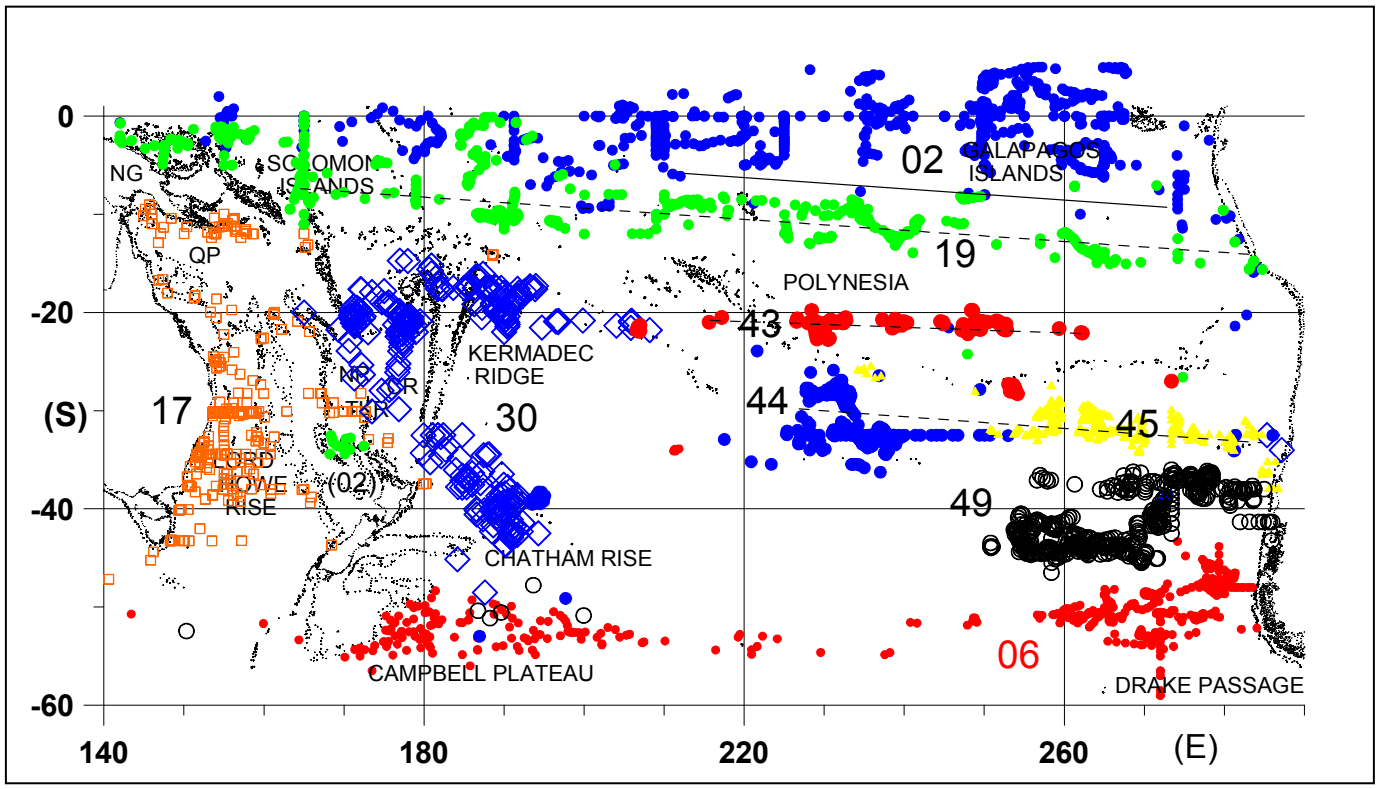

Figure 18. Geographical distribution of selected clusters from a 50-clustering of 16,764 AAIW Salinity-density profiles from ship and Argo profiling floats.

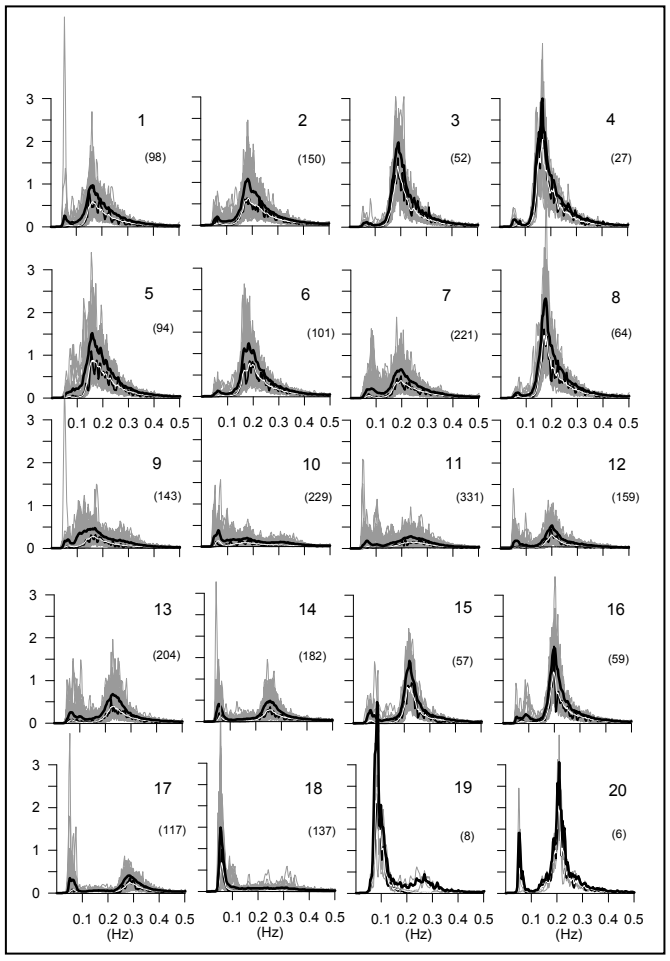

Figure 19. A 20-clustering of 2439 wind wave spectra for Port Hedland for year 1992 for significant wave height range 0.5 to 1.5 m. Thick black lies are tercile curves. Thin white curves are medians.

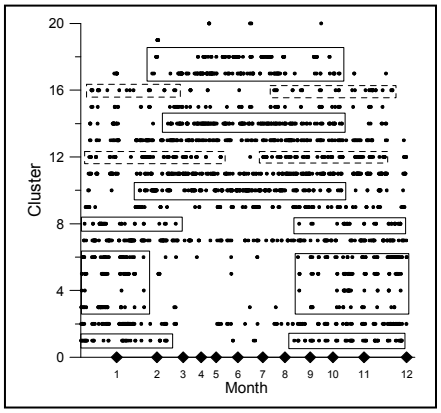

Figure 20. Seasonal changes in spectral type for Port Hedland in year 1992. Rectangles highlight the seasonal occurrence of spectral types.

Wind Wave Spectra – A Time Series Example

Wind wave spectra change or evolve at a wide range of temporal scales from hours (in response to local winds and seabreezes) to seasonal and beyond. The World Meteorological Organization method to classify wind wave spectra is to reduce up to three dominant spectral peaks to five parameters each (wave height, zero crossing period, a spectral shape parameter, peak spectral frequency, and dominant wave direction) [8]. The parameters are obtained by modelling with an imposed spectral form and user decisions, e.g. on separation of modes, and on what is sea and what is swell in the spectrum. The parameters are clustered into five or six groups, e.g. sea, swell, sea + swell, two swells, multimodal spectra. The groups are described as climatological spectra, and are based on numerical dominance of spectral types, e.g. sea+swell. Average spectra for the groups are formed from median values.

Clustering of entire spectra may be used instead to produce a classification based on spectral shape, without feature extraction or dimensional reduction. Physically this should be more meaningful than numerical dominance and a classification based on five parameters. In this application terciles were formed to describe the central tendency of clusters, rather than medoids or medians [3].

A classification of shallow water spectra for Port Hedland, Australia is shown in Fig. 19. The 2439 spectra are for significant height range 0.5 to 1.5 m, and were collected over the year 1992. A plot of spectral type with time (Fig. 20) shows distinct relations of spectral types to the summer monsoon season (November to January), and the South East Trade winds season (April to October). Some spectra occurring all year round are associated with seabreezes.

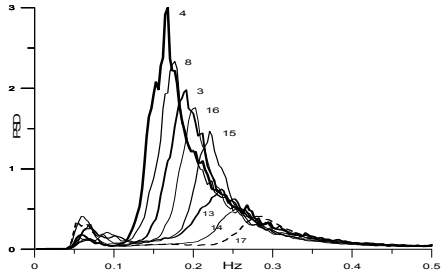

Figure 21. Arrangement of eight of the 20 clusters of Fig. 19 by envelope of spectra.

An overplot of selected spectra grouped by envelope of spectral shapes (Fig. 21) shows a typical wind wave growth pattern. This envelope occurred for summer. These few results show the power of the direct clustering technique to classify

wave spectra. The analyst is not restricted to a few heavily average spectra, but may assess the full data set on its merits. Full details of this application are given in [2].

DISCUSSION

Direct clustering of data sets consisting of single valued curves is a simple but powerful and versatile data exploration technique. Use of metrics other than Manhattan and Euclidian may further extend the usefulness of the method. The technique is certainly not the answer to everything, but as a front end it removes initial processing difficulties encountered in many different areas of the geosciences and other fields of research. Coupling this concept with the statistical sampling and clustering CLARA algorithm of [3] provides a fast way to evaluate and analyse large data sets.

ACKNOWLEDGEMENT

The multibeam sonar data were collected by the Co-operative Research Centre for Estuary, Coast and Waterway Management. The wind wave data were made available by the Port Hedland Port Authority.

REFERENCES

[1] L.J. Hamilton, "Clustering Of Cumulative Grain Size Distribution Curves For Shallow-Marine Samples With Software Program CLARA". *Australian Journal of Earth Sciences* 54, 503-519, 2007.

[2] L.J. Hamilton, "Characterising spectral sea wave conditions with statistical clustering of actual spectra". *Applied Ocean Research*. doi:10.1016/j.apor.2009.12.003, December 2009.

[3] L. Kaufman, and P.J. Rousseeuw. Finding Groups In Data: An Introduction To Cluster Analysis. John Wiley, New York, 1990.

[4] J.M. Preston, A.C. Christney, L.S. Beran, and W.T. Collins. Statistical seabed segmentation – from images and echoes to objective clustering. Proc. 7th European Conf. On Underwater Acoustics, pp813-816. 2004

[5] L.J. Hamilton. Acoustic Seabed Classification For Echosounders Through Direct Clustering Of Seabed Echoes. Unpublished document.

[6] L.J. Hamilton and I. Parnum. Seabed Classification From Unsupervised Statistical Clustering Of Entire Multibeam Sonar Backscatter Curves. Unpublished document.

[7] USGS. ECSTDB2005 - USGS East Coast Sediment Texture Database (2005). [http://pubs.usgs.gov/of/2005/2005-1001/data/surficial_sediments/ecstdb2005.txt]

[8] L.J. Lopatoukhin, V.A. Rozhkov, V.E. Ryabinin, V.R. Swail, A.V. Boukhanovsky, and A.B. Degtyarev. Estimation of extreme wind wave heights. World Meteorological Organisation WMO/TD-No. 1041. JCOMM Technical Report No. 9. 73pp. 2000.