# Mapping seabed sediments: Comparison of manual, geostatistical, object-based image analysis and machine learning approaches

Markus Diesing[1*], Sophie L. Green[2], David Stephens[1], R. Murray Lark[3], Heather A. Stewart[2], Dayton Dove[2]

[1]Centre for Environment, Fisheries and Aquaculture Science, Pakefield Road, Lowestoft, Suffolk, NR33 0HT, United Kingdom

[2]British Geological Survey, Murchison House, West Mains Road, Edinburgh, EH9 3LA, United Kingdom

[3]British Geological Survey, Environmental Science Centre, Nicker Hill, Keyworth, Nottinghamshire, NG12 5GG, United Kingdom

[*] Corresponding author: markus.diesing@cefas.co.uk; tel. +44 1502 524266; fax +44 1502 513865

***Abstract***

Marine spatial planning and conservation need underpinning with sufficiently detailed and accurate seabed substrate and habitat maps. Although multibeam echosounders enable us to map the seabed with high resolution and spatial accuracy, there is still a lack of fit-for-purpose seabed maps. This is due to the high costs involved in carrying out systematic seabed mapping programmes and the fact that the development of validated, repeatable, quantitative and objective methods of swath acoustic data interpretation is still in its infancy. We compared a wide spectrum of approaches including manual interpretation, geostatistics, object-based image analysis and machine-learning to gain further insights into the accuracy and comparability of acoustic data interpretation approaches based on multibeam echosounder data (bathymetry, backscatter and derivatives) and seabed samples with the aim to derive seabed substrate maps. Sample data were split into a training and validation data set to allow us to carry out an accuracy assessment. Overall thematic classification accuracy ranged from 67% to 76% and Cohen's kappa varied between 0.34 and 0.52. However, these differences were not statistically significant at the 5% level. Misclassifications were mainly associated with uncommon classes, which were rarely sampled. Map outputs were between 68% and 87% identical. To improve classification accuracy in seabed mapping, we suggest that more studies on the effects of factors affecting the classification performance as well as comparative studies testing the performance of different approaches need to be carried out with a view to developing guidelines for selecting an appropriate method for a given dataset. In the meantime, classification accuracy might be improved by combining different techniques to hybrid approaches and multi-method ensembles.

**Keywords**: Marine; Benthic; Habitat; Sediment; Mapping; North Sea

## 1 Introduction

Worldwide, the oceans and marginal seas are under increasing pressure from human activities (Halpern et al., 2008) and there is an ever greater need for good seabed habitat maps, both to underpin environmental and socio-economic impact assessments and to assist in the development of effective management measures that will contribute to our responsible stewardship of the marine environment and the sustainable use of its resources. The development of seabed mapping is now driven more by specific policy needs than our innate desire to explore our world. Several global, European and national initiatives aiming at maintaining biodiversity and conserving habitats and species (Convention on Biological Diversity, OSPAR Convention, EU Habitats, Birds and Marine Strategy Framework Directives, UK Biodiversity Action Plan, Marine and Coastal Access Act etc.) require better seabed habitat maps than exist at present to support assessments of the status of the seabed. In Europe, this need is currently addressed in part, through the European Marine Observation and Data Network (EMODNet), which, among other outputs, has so far compiled and harmonised available seabed sediment information to deliver a map of seabed substrates at a scale of 1:1 million. However, the resultant map has a smallest cartographic unit of approximately 4 $km^2$ and hence might be too generalised for detailed analysis on a more local level, although higher resolution seabed maps will be produced in a subsequent phase.

The advent of swath acoustic techniques has revolutionised seabed mapping science, as we are now able to map the seabed at high spatial resolution and accuracy. The development of swath acoustic systems dates back as early as the 1940s (Kenny et al., 2003)and was initially driven by sidescan sonar. Multibeam echosounders (MBES), with their ability to simultaneously record bathymetry and backscatter strength, have become the system of choice for detailed, high-resolution seabed mapping (Brown et al., 2011a). Despite the fact that we now have the technical ability to map the seabed with high detail and accuracy, we are a long way from achieving accurate and fit-for-purpose seabed habitat maps. This can be attributed to two key reasons:

Firstly, only a limited number of countries have so far initiated and executed large-scale seabed mapping programmes (e.g. 'Irish National Seabed Survey', 'Integrated Mapping for the Sustainable Development of Ireland's Marine Resource' and 'Marine Area Database for Norwegian waters') due to the high costs involved. However, a larger number of countries are collecting swath acoustic data for hydrographic charting purposes in a systematic way and possess a wealth of legacy data sets, mainly physical seabed sampling (grabs and cores) and seabed observations (videos and stills). Making best use of these available data sets is becoming increasingly important due to limited financial resources. In the United Kingdom, the Marine Environmental Mapping Programme (MAREMAP) aims to achieve common, national objectives in seabed and shallow geological mapping addressing themes such as habitat mapping, Quaternary science, coastal and shelf sediment dynamics and the assessment of human impacts and geohazards in the marine environment. MAREMAP makes use of data that are primarily collected for other purposes (e.g. MBES data of the Civil Hydrography Programme) or already existing (e.g. legacy grain-size data of the British Geological Survey (BGS)).

Secondly, the development of validated, repeatable, quantitative and objective methods of swath acoustic data interpretation is lagging behind the ability to collect high-quality swath acoustic data. Anderson et al. (2008) identified a lack of statistical, objective procedures as one of the most 'burning issues' of acoustic seabed classification. Expert interpretation of acoustic data 'by eye' is still

relatively common. More recently, automated methods have been explored, driven largely by the advantages of using objective classification algorithms, thus minimising subjectivity (Brown et al., 2011a). A variety of approaches has been trialled including artificial neural networks (Marsh and Brown, 2009; Ojeda et al., 2004), Bayesian decision rules (Simons and Snellen, 2009), decision trees (Che Hasan et al., 2012a; Dartnell and Gardner, 2004; Ierodiaconou et al., 2011; Rattray et al., 2009; Rooper and Zimmermann, 2007), support vector machines (Che Hasan et al., 2012b), Random Forest (Che Hasan et al., 2012b; Lucieer et al., 2013), Maximum Likelihood Classifier (Buhl-Mortensen et al., 2009; Che Hasan et al., 2012b; Ierodiaconou et al., 2011), clustering (Blondel and Gomez Sichi, 2009; Brown and Collier, 2008; Brown et al., 2012) and Principal Component Analysis within commercial software QTC Multiview (Brown et al., 2011b; McGonigle et al., 2009; Preston, 2009). However, only three studies have been published that attempt to compare different automated seabed mapping approaches (Che Hasan et al., 2012b; Ierodiaconou et al., 2011; Stephens and Diesing, 2014).

It is against this background that we held a "MAREMAP Acoustic Data Interpretation Workshop" in Edinburgh in October 2012. The workshop was centred on a common data set exercise where MBES and physical ground-truthing data were made available prior to the workshop. Participants were asked to apply their preferred methodology to the data sets and derive a map of seabed substrates. Results were discussed and compared during the workshop. This paper summarises the main outcomes of the common data set exercise. The objectives of the exercise, and hence this paper, were i) to compare the different methodologies with respect to their thematic accuracy and spatial representation of predicted seabed substrates and ii) to assess the relative merits and limitations of the different approaches.

## 2 Data

The area selected for the common data set exercise lies in the western North Sea off the Scottish coast of the United Kingdom (Figure 1). The study area is characterised by a complex glacial history with numerous meltwater channels incised into the seabed. The topography exhibits considerable geomorphological variability with drumlins and mega scale glacial lineations also identified (Stewart and Bradwell, 2013). The acoustic data were collected between 2006 and 2008 on behalf of the Maritime and Coastguard Agency (MCA) as part of a Safety of Navigation bathymetric survey; HI1152 Inner - Wee Bankie to Gourdon. The area interpreted covers 2325 $km^2$ and has full MBES bathymetry and backscatter coverage.

Participants were provided with MBES bathymetry, backscatter and ground-truth sample data. The bathymetry data were obtained from the MCA with a resolution of 7 m. Data sets derived from bathymetry (rugosity, aspect and slope) were also made available.

Backscatter was processed using PRISM (Processing of Remotely-sensed Imagery for Seafloor Mapping) software at the National Oceanography Centre, Southampton (LeBas and Hühnerbach, 1998). These data were initially processed at a higher resolution of 1 m. This resolution was chosen after trialling a range of values and can be seen as a balance between retaining as much detail as possible and avoiding a significant amount of gaps in the dataset.

The ground-truthing data set comprised 185 seabed sediment samples taken from the BGS sample archive. The majority of these were taken from grab samples. A subset of samples were selected so that 45 were retained for validation and 140 provided to workshop participants ('training data', Figure

1 and Table 1). These samples were selected using a randomized stratified approach (based on substrate class) so the validation set contained approximately the same class proportions as the training set.

Each sample record included positional information, content of gravel, sand and mud in weight-% and a classification according to Long (2006) in line with the EUNIS (European Nature Information System) habitat classification. This scheme of Long (2006) is a modification of the classification by Folk (1954; 1980) and comprises the four substrate classes 'coarse sediment', 'sand and muddy sand', 'mud and sandy mud' and 'mixed sediment' (Figure 2). A fifth substrate class of the EUNIS classification is 'rock and other hard substrata'; however no samples were available for this class.

The BGS sample archive contains samples collected in the 1970s and 1980s prior to the standard use of Global Positioning System (GPS) and Differential GPS (DGPS). Samples positions using the Decca Main Chain system are accurate to some hundreds of metres (Fannin, 1989).

## 3 Methods

### 3.1 Manual interpretation

Production of seabed sediment distribution maps using the manual interpretation method involves review of MBES bathymetry (including derived datasets) together with the MBES backscatter. These data were analysed in conjunction with samples in a geographic workspace i.e. ArcGIS. An understanding of the geological history and the implications for sediment characteristics and distributions is a crucial part of this methodology. The mapping geologist exercises 'expert' judgment by interpreting these data in the context of the geological (seabed and sub-seabed) and hydrodynamic environment. This judgement also introduces a level of adaptability and enables the interpreter to map areas where ground-truthing is limited, and to identify questionable, or over-represented sample points.

This manual process can be described in three generalized steps. The first is a broadscale review of the acoustic and ground-truthing data to identify characteristic and anomalous features of the map area in light of the geological setting and hydrodynamic environment. This review is conducted at both broad and focussed scales. The interpreter assesses which features will be detectable with respect to the final map's intended scale. Secondly, line work is constructed within ArcGIS to delineate distinct zones of backscatter intensity observed in the MBES backscatter data, i.e. along apparent sediment boundaries. Whilst these boundaries are represented as lines they generally represent gradational boundaries on the seabed. The bathymetry data can also be used to guide this line work, as sediment boundaries are often associated with discrete bedforms and/or broader changes of bathymetry associated with variation in the underlying geology. The derived datasets are viewed concurrently to further constrain the line work. These data are of particular use in areas of coarser sediment and areas of potential rock outcrop with a thin sediment veneer. The final step involves assigning the sediment classes, based on the previous steps with inclusion of the ground-truthing PSA data in accordance with the modified Folk classification scheme (Long, 2006). The line work created in ArcGIS may then be converted to polygons with the associated attribution.

Factors such as the positional accuracy of the samples can be assessed as part of the mapping process in order to prioritise the reliability of the different data layers. Due to imprecise positioning related to the use of the DECCA system, the ground-truthing data may be in conflict with that suggested by the

acoustic data. Also, many of the ground truthing samples were collected using a grab sampler. Where mixed sediments exist, this methodology tends to underestimate the finer sediment fractions which may not be retained. It is therefore necessary for the mapping geologist to apply expert judgement in determining relative merits of the different data in order to utilise the most representative material.

### 3.2 Object-based image analysis

The provided primary acoustic data layers (bathymetry and backscatter) were initially gridded to the same extent and resolution (7 m pixel size). A 2D Fourier filter (Wilken et al., 2012) was applied to the backscatter data to remove apparent 'stripe noise' parallel to the vessel's track. Several derivatives were calculated from both primary data layers (Table 2).

Software eCognition v8.8.0 was used to carry out object-based image analysis (OBIA). OBIA is widely used in terrestrial remote sensing applications (Blaschke, 2010), but was also successfully applied for mapping benthic habitats (Lucieer and Lamarche, 2011; Lucieer, 2008). OBIA is a two-step approach consisting of segmentation and classification. The aim of the segmentation is to divide the image into meaningful objects of variable sizes, based on their spectral and spatial characteristics. The resulting objects can be characterised by various features such as layer values (mean, standard deviation, skewness etc.), geometry (extent, shape etc.), texture and many others. Classification is then based on user-specified combinations of these image object features.

The interpretation was carried out separately for rock and sediment. No rock samples were available and the interpretation had to rely on acoustic signatures that were deemed characteristic of rocky substrate. Visual exploration of the acoustic data layers revealed that seabed roughness and the bathymetric position index (BPI, Lundblad et al., 2006) were useful proxies to differentiate rock from sediment.

#### *3.2.1 Mapping rock*

Segmentation was carried out using the multi-resolution segmentation algorithm in eCognition. This algorithm is an optimisation procedure, which locally minimises the average heterogeneity of image objects for a given resolution of image objects. Starting from an individual pixel (or existing image object), it consecutively merges pixels (or image objects) until a certain threshold, defined by the scale parameter is reached. The scale parameter is an abstract term that determines the maximum allowable heterogeneity for the resulting image objects. The choice of an appropriate scale parameter was aided by the Estimation of Scale Parameter (ESP) tool (Dragut et al., 2010).

The object heterogeneity, to which the scale parameter refers, is defined by the 'composition of homogeneity' criterion. This criterion defines the relative importance 'of colour' (pixel value in this case, e.g. backscatter digital number) versus shape of objects. If high weight is given to colour then the object boundaries will be predominantly determined by variations in colour of the image (e.g. backscatter strength). Further on, the shape criterion has contributions from smoothness and compactness, both of which can be weighted. A high value for smoothness will lead to smoother boundaries of the objects. High values of compactness will increase the overall compactness of image objects. We applied default values of 0.9 for colour, 0.1 for shape, 0.5 for smoothness and 0.5 for compactness. The segmentation was carried out on seabed roughness with a scale parameter of 4 as indicated by the ESP tool.

Rock was initially mapped where the standard deviation of the BPI with a radius of 7 map units (49 m) exceeded 0.2. This simple classification captured the bulk of the rock distribution; however it missed transitional areas at the margins of these rock 'cores', which did stay below the set threshold

for the standard deviation of the BPI7, but were rough enough to be considered as rock. To map such objects as rock they had to fulfil two conditions: The standard deviation of BPI7 had to exceed 0.12 and candidate objects had to share at least half of their border with rock 'core' objects. The resultant classified objects were subsequently exported to a shape file and inspected in a GIS environment. Limited manual edits were carried out to derive the final rock distribution.

#### *3.2.2 Mapping sediment*

An initial segmentation was carried out on backscatter strength with a scale parameter of 7 and default values of 0.9 for colour, 0.1 for shape, 0.5 for smoothness and 0.5 for compactness using the multi-resolution segmentation algorithm. This was followed by a spectral difference segmentation, which joins objects based on set thresholds for the maximum allowable difference in spectral values between neighbouring objects. A maximum allowable backscatter difference of 0.5 was set in this case following trials with a range of values. In this way, the number of objects was reduced from approximately 110,000 to less than 15,000.

Mean object values of bathymetry, aspect, slope, roughness, curvature, Moran's I, BPI3, BPI5, BPI10, BPI25, backscatter, backscatter roughness and backscatter Moran's I were extracted for those objects that coincided with training sample locations. The extracted data sets were explored with the aim to establish threshold values that would allow the discrimination of sediment types based on acoustic data. A threshold of 244 for backscatter allowed distinguishing between coarse sediment and sand. Attempts to find thresholds for mixed sediment and mud were unsuccessful.

### 3.3 Machine Learning (Random Forest)

The provided primary acoustic data layers were pre-processed in the same way as described in Section 3.2. Several derivatives were calculated from both bathymetry and backscatter (Table 2).

Random Forest (Breiman, 2001) is a machine learning approach that belongs to the family of decision tree learning. It has been applied before to the mapping of marine substrates and authors have reported promising results (Che Hasan et al., 2012b; Li et al., 2011; Lucieer et al., 2013). The goal of decision tree learning is to create a model that predicts the value of a target variable based on several input variables. Decision trees are of two main types: Classification trees predict class membership while regression trees predict a quantity. The former was applied in this case.

The decision tree creates classification rules by recursively partitioning the data into increasingly homogenous groups. To measure the homogeneity of each node of the tree, an impurity index is used (Gini impurity index for classification). This is in turn used to choose the best partition at each node (subset of data) of the tree. Random Forests are comprised of an ensemble of randomly constructed decision trees. Two components of randomness are introduced. Firstly each tree is constructed using a random bootstrapped sample of the training data. Secondly, rather than testing all features for the best split, a random subset of variables is tested at each split in each tree. The prediction is made for unobserved data by taking a majority vote of the individual trees. The idea behind introducing the randomness into the construction of the trees is that the final outcome will be less subject to any biases in the training dataset and that the capacity for generalisation will be increased.

The following is a brief description of how the algorithm works (Liaw and Wiener, 2002).

1. Take $n_t$ bootstrap samples (sampling with replacement), equal to the size of original data, from the data set. (where *t* is the number of trees in the forest, set to a high number)

2. A classification tree is grown for each of the bootstrap samples $n_t$. For each node of the tree, $m_{try}$ variables are chosen and tested to find the best split of the data.
3. Predict class at unsampled locations by aggregating the predictions of the $n_t$ trees by majority vote.

### 3.4 Geostatistics

The method that is used here (Lark et al., 2012) is based on compositional cokriging (Lark and Bishop, 2007; Pawlowsky-Glahn and Olea, 2004) in which the data on particle size distribution (proportions by mass of particles in the mud, sand and gravel size classes) are treated, after transformation by the additive log-ratio (alr) transform, as a realization of a multivariate, spatially dependent random function. This function is characterized by statistical parameters, those of the linear model of coregionalisation (Journel and Huijbregts, 1978), which are estimated from data. They can then be used to compute the empirical best linear unbiased prediction (E–BLUP) of the transformed contents of the different particle size classes at unsampled locations. It is possible to evaluate the probability, conditional on the data, of observing each of a set of sediment texture classes at any of the unsampled sites from the E-BLUP prediction distribution.

It is possible in principle to extend the compositional cokriging predictor to incorporate additional covariates, such as backscatter strength, bathymetry etc. However, in this case exploratory analysis found no significant linear relationship between the compositional data and the bathymetry, which would be required for this to give any benefit. How best to integrate acoustic data into the geostatistical framework is a matter for further work. For purposes of the present paper we regard the E–BLUP by compositional cokriging from the sediment compositional data alone as a state-of-the art for spatial prediction against which to measure the benefits from incorporation of acoustic data through alternative statistical or computational models for predicting the sediment texture class.

The methodology, applied to the prediction of sediment texture classes, is described in detail by Lark et al. (2012). We obtained alr transformations of gravel and mud content in the sediment samples with sand used as the denominator. Note that the alr transformation is not defined if the proportion for any particle size class is zero. For this reason we assumed that zero values reflect measurement error, and we followed Martín-Fernandéz and Thió-Henestrosa (2006) by substituting any zero values (percentage scale) with 0.005, half the minimum recorded value, and then renormalizing the values for any observation to sum to 100.0 as described by Lark et al. (2012).

Following Lark et al. (2012) we obtained the E-BLUP joint distribution for the two alr-transformed variables at each target site. We then sampled from each distribution and back-transformed each sample to the gravel-sand-mud trigon (Figure 2). In this way we obtained a Monte Carlo integration of the E-BLUP joint distribution over the respective parts of the sample space that correspond to each texture class, and from this an estimated probability for each texture class. From these probabilities one may identify the class of maximum probability. The probability of that class is a measure of the degree of certainty of the prediction. One might also examine the probabilities for all classes when making decisions about a particular location which depend on the (unknown) sediment texture class.

This approach was used to compute probabilities for all four sediment texture classes from the EUNIS habitat classification at each of the validation sites.

### 3.5 Assessment of the classification approaches

We employed the validation data set comprising 45 samples (24% of the full ground-truth data set) to derive error matrices for three approaches (OBIA, Random Forest and geostatistics). As the manual

interpretation was carried out prior to the workshop based on all ground-truth samples, these results could not be included in the assessment and comparison of accuracy. Despite this, we felt that it was nevertheless desirable to include the results from manual interpretation in this study, as this is probably still the most commonly applied approach to seabed mapping.

Several measures of classification accuracy can be derived from an error matrix, including the overall thematic map accuracy ('percentage correct'), purity, representation and Cohen's (1960) kappa coefficient of agreement (e.g. Foody, 2002; Lark, 1995; Stehman, 1997). The overall accuracy gives the percentage of cases correctly allocated and is calculated by dividing the total number of correct allocations by the total number of ground-truth samples (Congalton, 1991).The use of Cohen's kappa has been advocated by some as it makes some compensation for chance agreement. The representation is the probability that a pixel belonging to the class is mapped as that class. The purity gives the probability that a pixel belongs to a class given that it is mapped as that class. Purity and representation are frequently called 'user's accuracy' and 'producer's accuracy'; however these terms might be considered misnomers (Lark, 1995) and are therefore avoided in the following.

We evaluated the statistical significance of differences in overall accuracy and kappa coefficient between two approaches employing McNemar's $\chi$-squared test with continuity correction based on overall accuracy and a resampling technique described by McKenzie et al. (1996). Both tests are suited to compare accuracy statistics that are not independent, which is the case when they are derived from the same set of ground-truth samples (Foody, 2004).

The resultant maps were compared with the Map Comparison Kit (Visser and de Nijs, 2006). A pixel by pixel comparison of map pairs was carried out and the matching number of pixels per substrate class was calculated. Based on these values the overall percentage agreement of two maps was derived.

## 4 Results

### 4.1 Data exploration

Exploration of the training data shows limited discriminatory power of bathymetry to separate between substrate types (Figure 3). Backscatter allows a distinction between coarse sediment (CS) on the one hand and mud and sandy mud (Mu) and sand and muddy sand (Sa) on the other. The latter two substrate classes are however virtually indistinguishable based on backscatter.

### 4.2 Resulting maps

Figure 4 depicts the four seabed substrate maps resulting from the different approaches. For ease of comparison of the map results with the primary acoustic data layers, bathymetry and backscatter strength are displayed as well.

### 4.3 Thematic accuracy assessment

The results of the thematic accuracy assessment are summarised in Table 3. Overall accuracies were 67% for OBIA and 69% for geostatistics, whilst Random Forest achieved the highest accuracy of 76%. Respective kappa coefficients ranged from 0.34 (geostatistics) to 0.38 (OBIA) and 0.52 (Random Forest). As Random Forest scored highest for both statistics, we compared OBIA and

geostatistics against Random Forest, in order to test whether the results for the two accuracy statistics were significantly different between methods (Table 4). Both tests, McNemar's $\chi$-squared test with continuity correction based on overall accuracy and the test of McKenzie et al. (1996) based on the kappa coefficient, showed that there were no significant differences at the 5% level of significance in the accuracy statistics of the classifications derived by the different methods (Table 4).

### 4.4 Accuracy per substrate type for the different methods

Purity and representation for the different substrates and methods are depicted in Figure 5. All methods achieve reasonable purity for coarse sediments (57% - 78%) and especially sands (67% - 77%). The highest purity of coarse sediments is attained by geostatistics, while Random Forest achieves the highest purity of sand and muddy sand. All methods fail to predict mud and mixed sediment. Values of representation range between 47% and 80% for coarse sediments and between 69% and 92% for sand and muddy sand. Maximum representation is attained by OBIA in the case of coarse sediments and geostatistics for sand. Again, representation is zero for the other two substrate classes.

### 4.5 Spatial comparison of maps

Comparison of the frequency of occurrence of each seabed substrate from each of the applied approaches showed overall consistent patterns, but with marked differences in the absolute percentages (Figure 6). Sand was the most widespread substrate, predicted to occur in 54% to 71% of area. Coarse sediments were the second-most frequent substrate predicted to cover between 24% and 39% of area. The remaining substrates were by far less frequent and not predicted by all four methods. Rock was only included in maps produced by OBIA and manual interpretation with 0.82% and 1.80% area covered respectively. Mixed sediment and mud were not mapped by OBIA. Mud was predicted to occur in 3.96% (manual) and 4.92% (geostatistics) of area. Contrary to this, Random Forest predictions were as low as 0.45%. Manual interpretation mapped 1.82% of seabed area as mixed sediment, while predictions by Random Forest and geostatistics were much lower, yielding 0.05% and 0.09% respectively.

Pairwise pixel-by-pixel comparisons of the spatial distribution of the five substrates between each of the four different maps showed agreements ranging between 68% and 87% (Table 5). The most similar map pair was OBIA and Random Forest. The two least similar maps were those derived by manual interpretation and geostatistics. Those approaches which employed acoustic data layers showed a high degree of similarity, ranging between 77% and 87%. Differences were mainly related to the less frequently occurring classes mud, mixed sediment (manual-OBIA, manual-Random Forest) and rock (manual-Random Forest, OBIA-Random Forest). Besides that, differences were due to the position of boundaries between sand and coarse sediment (Figure 7). The results of the geostatistical approach differed more markedly from the other three approaches. However, there was a broad agreement in the prediction of mud between the geostatistical and manual approaches.

# 5 Discussion

## 5.1 Limitations of the data sets

We decided to choose acoustic and ground-truthing data sets that were initially collected for other purposes for the presented data interpretation and map comparison exercise. This choice was deliberate as these are the likely types of data available for seabed sediment and habitat mapping in times of constrained financial resources. Across the United Kingdom continental shelf there exist seabed samples from more than 25,000 locations collected during a systematic mapping programme carried out by the BGS during the 1970s and 1980s (Fannin, 1989). Similarly, the MCA systematically collects MBES data as part of the Civil Hydrography Programme. Together with MBES data from other groups, these cover currently more the 200,000 $km^2$ or approximately 26% of the UK Exclusive Economic Zone (Wynn et al., 2012). Making best use of available data sets is important, yet the available data have not yet been fully utilised for seabed mapping purposes.

Whilst the call to 'collect data once and use it many times' is understandable and laudable, it has to be acknowledged that data cannot always be collected in a way that makes them ideal for a variety of purposes. For example, MBES data collected for hydrographic charting purposes will be optimised for high-quality bathymetric data during acquisition, which might compromise the quality of the backscatter data. Therefore, it is important to explore the limitations of the utilised data sets.

Arguably, the most important source of error in the data sets we utilised is the positional error of the ground-truth samples. These errors are difficult to estimate, as the accuracy of the DECCA system varies not only with distance from the land-based transmitting station but is also dependent on the season and the time of the day (e.g. Fig. 2 in Kubicki and Diesing, 2006). However, according to Last (1992: Fig. 6) the accuracy within the study site is better than 400 m at all times and seasons with a confidence of 95%. Errors of such order of magnitude might be significant especially in heterogeneous areas.

The seabed is subject to hydrodynamic processes which will cause temporal variability of sediment composition at a particular location. An unknown amount of error can be attributed to the fact that the observations are of varying ages and crucially, the fact that the acoustic data were not collected at the same time as the sediment observations. While the approaches have been optimised to predict what is observed in the data, does this reflect what is currently the reality on the ground? However, water depths within the study area range between 0 m and 97 m making the deeper parts of the site unsusceptible to wave agitation except during extreme and rare events. Additionally, tidally induced bottom shear stresses are relatively weak within the study area (e.g. Pingree and Griffiths, 1979). Consequently, remobilisation of seabed sediments is likely to occur rarely in those areas that are not affected by regular wave impact. There is also evidence that interaction between waves and large roughness elements (wave ripples), present on coarse sediment domains, generates near-bed turbulence that is greatly enhanced relative to that in fine sediment domains. This turbulence inhibits settling of fine material in an area dominated by coarse sediment (Green et al., 2004; Murray and Thieler, 2004). Consequently, sediment distribution patterns with strong grain-size contrast might remain stable for decades (Diesing et al., 2006; Goff et al., 2005) despite frequent remobilisation of sediment. Such a mechanism has been specifically attributed to sorted bedforms (Murray and Thieler, 2004), but high stability has also been described for other sediment distribution patterns with high grain-size contrast (Anthony and Leth, 2002; Schwarzer et al., 2003; Tauber and Emeis, 2005). It might therefore be reasonable to assume that the observed patterns of high grain-size contrast in our study area are generally stable over decadal timescales.

Furthermore, the input acoustic data layers are aggregated to a particular spatial resolution. These are dictated by technical limitations of the sampling equipment (MBES sampling density) and more significantly the computing power for processing the data. The bathymetry and backscatter grids were aggregated to a resolution of 7 m. Considerable variability of sediment type within an area of 49 $m^2$ of seabed can be expected, but the sediment observations are not sampled at the same scale. A sediment grab samples perhaps ≈0.3 $m^2$ of seabed; this is the support size of the grain-size data and it is different from the support size of the acoustic data. Averaging quantities over a large area has the effect of reducing their variance and making the distributions more normal (Isaaks and Srivastava, 1989). The implications are that it is unrealistic to expect to account for all the variability in our observations when the predictor variables have a coarser resolution (larger support size) to the sampled data.

With regard to the MBES data sets, errors are introduced during data collection caused by weather conditions and sea state. Movements of the echo sounder that cannot be fully compensated will lead to artefacts in bathymetry and backscatter data. Further, 'stripe noise' parallel to the course of the vessel might result from backscatter data processing, although significant improvements have been made through the application of Angle Varying Gain correction (Fonseca et al., 2009) and 2D Fourier filtering (Wilken et al., 2012).

All these potential error sources will have an effect on mapping results, although the effects will differ between methods. For example, machine learning approaches might be especially susceptible to imperfections in ground-truth data like imprecise geolocation. As the 'learning' is based on the relationships between ground-truth information and feature values encountered at the ground-truth locations, errors in positioning will lead to imprecise relationships. Maybe more importantly, classes that were not sampled at all were not predicted by the Random Forest algorithm as is the case with rock. Conversely, apparent mismatches between ground-truth information and acoustic data can be accounted for in manual interpretation. There is, however, the danger that perceptions might be wrong, leading to results that match expectations, but might miss novel or unexpected relationships.

Potential errors will also have an impact on accuracy assessment. An underlying implicit assumption of any accuracy assessment is that the ground-truth data used are an accurate representation of reality. As discussed above, there are several sources of error related to the utilised ground-truth data and these limitations need to be kept in mind when comparing the performance of the different approaches.

### 5.2 Quantitative comparison of approaches

We have assessed the thematic accuracy of three substrate maps produced by OBIA, machine learning (Random Forest) and geostatistical approaches. A validation could not be carried out for the manual classification approach, as all samples were used in the construction of the map. However, this could be done in principle, for other areas. The derived accuracy statistics (Table 3) fall within the range of values reported in similar studies (Brown et al., 2012; Che Hasan et al., 2012a; Che Hasan et al., 2012b; Ierodiaconou et al., 2011; Lucieer et al., 2013; Lucieer, 2008; Ojeda et al., 2004; Rattray et al., 2009; Rooper and Zimmermann, 2007; Stephens and Diesing, 2014). Random Forest produces the highest values for overall accuracy and kappa. However, there are no statistically significant differences between the performances of the three tested approaches at the 5% level (Table 4). Results of purity and representation did vary between substrates and methods. Therefore, we cannot single out a particular method as 'best performing' in this study. Conversely, Stephens and Diesing (2014), in a study comparing six machine-learning algorithms, reported that tree-based methods (Random Forest

and Decision Trees) and Naive Bayes achieved the highest accuracies when predicting sediment classes based on acoustic data and grab samples. Ierodiaconou et al. (2011) found that the decision tree method QUEST (Quick Unbiased Efficient Statistical Tree) was producing significantly better results than the decision tree method CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation) and a Maximum Likelihood Classifier. Che Hasan et al. (2012b) evaluated four supervised classification methods and found that Support Vector Machines, Random Forest and QUEST produced significantly better results than the Maximum Likelihood Classifier. Both studies employed a Z statistic to test the significance of the difference in accuracy. This approach might, however, be not appropriate for this study, as Foody (2004) points out that the assumption of the independence of samples needs to be satisfied in order to apply the Z statistic. Studies comparing the performance of different classifiers typically employ the same set of samples, and this might also be true in the case of the studies presented by Ierodiaconou et al. (2011) and Che Hasan et al. (2012b).

To our knowledge, there exist no other studies that have attempted to compare the performance of classification approaches for seabed mapping based on MBES and ground-truth data, although Brown et al. (2005) evaluated and compared habitat maps produced with the acoustic ground-discrimination system (AGDS), RoxAnn, and sidescan sonar. Based on the cited studies and the one presented here, it is currently difficult to draw conclusions on the usefulness of specific approaches for mapping seabed substrates and habitats. Clearly, there is a need for more comparative studies to advance the application of robust, objective methods of seabed classification and evaluate the relative merits of the different techniques.

All three tested methods achieved reasonably accurate predictions of coarse sediment and particularly sand and muddy sand. However, none of the methods accurately predicted mixed sediments and mud and sandy mud. This might be partly due to the low number of samples available for those two substrate types (Table 2). At least equally important is the fact, that none of the acoustic features had discriminatory power to separate these substrates (Figure 3). Mud might often be related to changes in bathymetry and/or backscatter as such fine-grained sediment typically accumulates in deeper basins and exhibits low backscatter, when not consolidated. However, in the study site mud was virtually indistinguishable from sand in terms of bathymetry and backscatter (Figure 3). We can only speculate what the reason for this might be, but it seems likely that it is attributable to the fact that acoustic data were not collected at the same time as the sediment samples.

With regard to mixed sediment, there exist fundamental physical reasons why it might not be separable from coarse sediment based on backscatter strength, as it can be very sensitive to the presence of roughness elements on the order of half the acoustic wavelength (Goff et al., 2004). In this study, the employed frequency was 300 kHz, which equals a wavelength of 5 mm assuming a sound velocity of 1500 m/s. Because of this, it can be expected that gravel content has a crucial impact on backscatter strength. Gravel content exhibits a strong but non-linear relationship with backscatter strength and the presence of even a minor portion of gravel (a few percent) leads to a significant increase in backscatter (Goff et al., 2004; Goff et al., 2000). Mixed and coarse sediments both have gravel contents larger than 5 weight-percent (Figure 2). The backscatter response will be dominated by the gravel content, regardless of the sand: mud ratio, which distinguishes mixed from coarse sediment at a value of 9:1. In this scenario, it might be useful to incorporate geostatistical predictions of the sand: mud ratio and knowledge of the local geology.

The detection and accurate mapping of uncommon habitats with limited spatial extent might become crucial when these are of conservation importance. The common accuracy statistics like those

employed in this study give equal weight to all ground-truth samples. Because of this, the overall accuracy statistic will be heavily influenced by the common classes and might therefore be misleading. As an alternative, the Balanced Error Rate (BER) has been proposed (Luts et al., 2010). The BER is the average of the error rate (inverse of accuracy) for each separate class. Because it gives equal weight to the different class errors, it might be more appropriate when class frequencies are very uneven. One such habitat of high conservation importance but limited spatial extent on a continental shelf-wide scale is rocky reef. Despite the fact that rock is present within our study site, this exercise was however not specifically focused on mapping rock features due to the absence of adequate ground-truth data. It was consequently not possible to measure the accuracy of the different methods in mapping rock. The ability to successfully discriminate rocky reef from sediment by applying OBIA to sidescan sonar data was demonstrated by Lucieer (2008), but future studies might further investigate this important issue employing other mapping techniques and datasets.

The similarity of maps ranged from 68% to 87% (Table 5). A large amount of the disagreement was accounted for by those classes that were difficult to predict accurately (mixed sediments and mud) or were not predicted by some approaches (e.g. rock). Mismatches did also occur along the boundaries between sand and coarse sediment, whilst the 'core' areas of these classes did usually coincide. Although such comparisons are difficult to interpret as there is no 'true' map to compare the other maps against, they are still useful as they indicate where there is agreement in predictions and thus increased confidence in the results. Although a higher overall rate of agreement would be desirable, these values are significantly larger than those cited by Brown et al. (2005), in a study comparing habitat maps produced with AGDS. These authors compared four maps, which were 36% to 43% in agreement. Brown et al. (2005) concluded that AGDS is probably not an appropriate tool for map production and swathe acoustic systems (e.g. MBES) should be used instead if the end use of the habitat map demands a high degree of accuracy in relation to habitat boundaries and discrete seabed features. Although the results are not directly comparable, the significantly larger agreement rates reported in this study indicate the superiority of MBES-derived maps over ADGS.

### 5.3 Qualitative comparison of approaches

Apart from quantifiable indices as presented above it might be useful to consider the principal merits and limitations of the applied approaches. This is carried out by evaluating the different methods against a certain set of attributes. In particular, we are investigating here whether a specific approach is validated, repeatable, quantitative and objective.

The predictions made by geostatistics, OBIA and Random Forest have all been validated applying a validation set (Table 3). A validation could not be carried out for the manual classification approach, as all samples were used to construct the map. However, this could be done in principle, for other areas. In any case, it will be crucial that a sufficient amount of samples is available for training and validation. This does apply to the total amount of samples as well as each individual class to be mapped.

Results of geostatistics, OBIA and Random Forest could be considered as repeatable. Given the same datasets and model parameters of the respective methods (the variogram model for the geostatistics, the rules for the Random Forest and the rule-set for the OBIA) identical results could be achieved. Outputs from traditional manual interpretations are likely to be less repeatable. However, the repeatability is improved through the use of a standardised approach and workspace between different interpreters and application of QA/QC procedures. Repeatability of results will become an even more important issue as the science of seabed mapping moves towards change detection and monitoring.

Quantitative outputs, i.e. sediment composition instead of substrate class were only provided by geostatistcs, here. However, Random Forest and other machine learning methods, are capable of predicting quantities instead of classes(Breiman, 2001). Outputs from manual interpretation and OBIA will invariably be qualitative. Quantitative outputs have the significant advantage that they easily allow reclassification of the data. For example, predicted mud, sand and gravel percentages could be classified following Long (2006) or Folk (1954) with minor effort.

No method is fully objective as a certain amount of human intervention is always necessary. This intervention is minimal in the case of Random Forest, which essentially requires the user select only two parameters, namely the number of trees (t) to be grown and the number of predictors tested at each node ($m_{try}$). In practice, the resulting model is usually fairly insensitive to parameters, providing t is large enough to stabilise predictions. Geostatistical modelling proceeds according to specific and objective criteria, such as those used to cross-validate and select the linear model of coregionalisation. However, it is neither possible nor desirable to automate the procedure, which requires expert judgement about the explicit assumptions that it makes. OBIA, as it was implemented here, does require informed choices by the user and in that sense it is not objective. It does however apply the developed rules in a systematic way via the rule-set. As for OBIA, manual interpretation also involves making informed choices. Whilst this may be considered subjective, it has the benefit of incorporating a geological understanding of the seabed system into the analysis. In this way, the methodology might be more suitable to account for variability in data quality and quirks associated with the use of legacy data, such as positional accuracy and undersampling of certain classes.

### 5.4 Implications and suggestions for future research

What is the best approach for seabed mapping? Unsurprisingly, there is no simple answer to this question. It has to be acknowledged that all approaches have pros and cons, and how well they fare will depend on several factors including, but not limited to, the positioning accuracy of the sample data, the quantity of available samples, the quality of the acoustic data sets, the discriminatory power of the acoustic data etc. In order to improve prediction accuracy of seabed mapping, there exist at least three different strategies: i) Developing guidelines for selecting an appropriate method for a given dataset; ii) hybrid approaches and iii) multi-method ensembles.

The first strategy would require studies on the effects of numerous factors affecting the performance of various approaches. This might include studying the effects of sample size, density and spatial distribution of samples, positioning and classification error, grid size (resolution) of acoustic data etc. One crucial issue is the compatibility of ground-truth with MBES data. Key questions, among others, are: Does the sample allocation adequately represent seabed heterogeneity? What is the adequate pixel size to link ground-truth information with acoustic data? The former question could be addressed by the application of optimum allocation analysis (OAA) to sampling design (Clements et al., 2010). This would however assume that sampling is carried out after MBES data have been collected as these datasets will be required to allocate samples. In our scenario, i.e. working with existing datasets, OAA might be used to better understand the gaps between an optimal sample design and the existing sample distribution. The latter question could be investigated by varying the cell size of the acoustic datasets and measuring the effect on prediction performance. Also, additional comparative studies are needed similar to the one presented here and those from Ierodiaconou et al. (2011), Che Hasan et al.(2012b) and Stephens and Diesing (2014). While such studies are worthwhile and important, it cannot be expected to amass the required body of evidence in the short term. It might therefore be appropriate to look at the other two strategies as well.

Hybrid approaches have been trialled in the context of predicting sediment fractions. Li et al. (2011) compared 14 methods in their ability to predict mud content and found that a hybrid approach combining Random Forest and geostatistics performed best. Such an approach could be extended to predict substrate composition from the acoustic data using Random Forest and subsequently interpolate the residuals from the Random Forest model (i.e. the unexplained variability) using geostatistics. One problem with this approach is that it is not possible to estimate the parameters of the Random Forest model and the random effects (residual process) jointly, as one may in a linear mixed model. Estimates of the random effects model based on residuals from the fitted Random Forest are likely to be biased. Also, Plets et al. (2012) provide an example of a substrate map that has been produced utilising a number of acoustic data interpretation techniques.

Lastly, a multi-method ensemble approach could be taken. An ensemble approach is often viewed as an effective way of improving classification performance (Du et al., 2012; Lu and Weng, 2007). In the ensemble approach, outputs from several methods are combined with the aim to produce an accurate classification. A prerequisite for a successful ensemble classification is the use of accurate approaches that differ in their architecture and hence yield individual classifications that differ to a degree (Foody et al., 2007). Combining multiple classifications is frequently achieved by simple voting, where all voters have equal weight (Alpaydin, 2010).

## 6 Acknowledgements

We would like to thank Tim LeBas (National Oceanography Centre Southampton) for processing MBES backscatter data and all participants of the workshop for their input to discussions. We are grateful to the two anonymous referees who provided valuable comments on the original manuscript. MD and DS were supported by Cefas Research and Development funding (research project DP312). This work has been supported by the ongoing regional mapping programme MAREMAP (www.maremap.ac.uk), a joint initiative led by the British Geological Survey, the National Oceanography Centre Southampton and the Scottish Association for Marine Science. SG, RML, HS and DD publish with permission of the Director, British Geological Survey (Natural Environment Research Council).

## 7 References

Alpaydin, E., 2010. Introduction to Machine Learning, 2nd Edition ed. MIT Press, Cambridge, MA.

Anderson, J.T., Van Holliday, D., Kloser, R., Reid, D.G., Simard, Y., 2008. Acoustic seabed classification: current practice and future directions. ICES Journal of Marine Science 65, 1004-1011.

Anthony, D., Leth, J.O., 2002. Large-scale bedforms, sediment distribution and sand mobility in the eastern North Sea off the Danish west coast. Marine Geology 182, 247-263.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65, 2-16.

Blondel, P., Gomez Sichi, O., 2009. Textural analyses of multibeam sonar imagery from Stanton Banks, Northern Ireland continental shelf. Applied Acoustics 70, 1288-1297.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5-32.

Brown, C.J., Collier, J.S., 2008. Mapping benthic habitat in regions of gradational substrata: An automated approach utilising geophysical, geological, and biological relationships. Estuarine Coastal and Shelf Science 78, 203-214.

Brown, C.J., Mitchell, A., Limpenny, D.S., Robertson, M.R., Service, M., Golding, N., 2005. Mapping seabed habitats in the Firth of Lorn off the west coast of Scotland: evaluation and comparison of habitat maps produced using the acoustic ground-discrimination system, RoxAnn, and sidescan sonar. ICES Journal of Marine Science 62, 790-802.

Brown, C.J., Sameoto, J.A., Smith, S.J., 2012. Multiple methods, maps, and management applications: Purpose made seafloor maps in support of ocean management. Journal of Sea Research 72, 1-13.

Brown, C.J., Smith, S.J., Lawton, P., Anderson, J.T., 2011a. Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. Estuarine, Coastal and Shelf Science 92, 502-520.

Brown, C.J., Todd, B.J., Kostylev, V.E., Pickrill, R.A., 2011b. Image-based classification of multibeam sonar backscatter data for objective surficial sediment mapping of Georges Bank, Canada. Continental Shelf Research 31, S110-S119.

Buhl-Mortensen, P., Dolan, M., Buhl-Mortensen, L., 2009. Prediction of benthic biotopes on a Norwegian offshore bank using a combination of multivariate analysis and GIS classification. ICES Journal of Marine Science 66, 2026-2032.

Che Hasan, R., Ierodiaconou, D., Laurenson, L., 2012a. Combining angular response classification and backscatter imagery segmentation for benthic biological habitat mapping. Estuarine Coastal and Shelf Science 97, 1-9.

Che Hasan, R., Ierodiaconou, D., Monk, J., 2012b. Evaluation of Four Supervised Learning Methods for Benthic Habitat Mapping Using Backscatter from Multi-Beam Sonar. Remote Sensing 4, 3427-3443.

Clements, A.J., Strong, J.A., Flanagan, C., Service, M., 2010. Objective stratification and sampling-effort allocation of ground-truthing in benthic-mapping surveys. ICES Journal of Marine Science 67, 628-637.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37-46.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37, 35-46.

Dartnell, P., Gardner, J.V., 2004. Predicting seafloor facies from multibeam bathymetry and backscatter data. Photogrammetric Engineering and Remote Sensing 70, 1081-1091.

Diesing, M., Kubicki, A., Winter, C., Schwarzer, K., 2006. Decadal scale stability of sorted bedforms, German Bight, southeastern North Sea. Continental Shelf Research 26, 902-916.

Dragut, L., Tiede, D., Levick, S.R., 2010. ESP: a tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. International Journal of Geographical Information Science 24, 859-871.

Du, P., Xia, J., Zhang, W., Tan, K., Liu, Y., Liu, S., 2012. Multiple Classifier System for Remote Sensing Image Classification: A Review. Sensors 12, 4764-4792.

Fannin, N.G.T., 1989. Offshore Investigations 1966-87, British Geological Survey Technical Report WB/89/2. British Geological Survey, Keyworth, Nottingham.

Folk, R.L., 1954. The distinction between grain size and mineral composition in sedimentary-rock nomenclature. Journal of Geology 62, 344-359.

Folk, R.L., 1980. Petrology of sedimentary rocks. Hemphill Publishing Company, Austin, Texas.

Fonseca, L., Brown, C., Calder, B., Mayer, L., Rzhanov, Y., 2009. Angular range analysis of acoustic themes from Stanton Banks Ireland: A link between visual interpretation and multibeam echosounder angular signatures. Applied Acoustics 70, 1298-1304.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sensing of Environment 80, 185-201.

Foody, G.M., 2004. Thematic Map Comparison: Evaluating the Statistical Significance of Differences in Classification Accuracy. Photogrammetric Engineering & Remote Sensing 70, 627-633.

Foody, G.M., Boyd, D.S., Sanchez-Hernandez, C., 2007. Mapping a specific class with an ensemble of classifiers. International Journal of Remote Sensing 28, 1733-1746.

Goff, J.A., Kraft, B.J., Mayer, L.A., Schock, S.G., Sommerfield, C.K., Olson, H.C., Gulick, S.P.S., Nordfjord, S., 2004. Seabed characterization on the New Jersey middle and outer shelf: correlatability and spatial variability of seafloor sediment properties. Marine Geology 209, 147-172.

Goff, J.A., Mayer, L.A., Traykovski, P., Buynevich, I., Wilkens, R., Raymond, R., Glang, G., Evans, R.L., Olson, H., Jenkins, C., 2005. Detailed investigation of sorted bedforms, or "rippled scour depressions", within the Martha's Vineyard Coastal Observatory, Massachusetts. Continental Shelf Research 25, 461-484.

Goff, J.A., Olson, H.C., Duncan, C.S., 2000. Correlation of side-scan backscatter intensity with grain-size distribution of shelf sediments, New Jersey margin. Geo-Marine Letters 20, 43-49.

Green, M.O., Vincent, C.E., Trembanis, A.C., 2004. Suspension of coarse and fine sand on a wave-dominated shoreface, with implications for the development of rippled scour depressions. Continental Shelf Research 24, 317-335.

Halpern, B.S., Walbridge, S., Selkoe, K.A., Kappel, C.V., Micheli, F., D'Agrosa, C., Bruno, J.F., Casey, K.S., Ebert, C., Fox, H.E., Fujita, R., Heinemann, D., Lenihan, H.S., Madin, E.M.P., Perry, M.T., Selig, E.R., Spalding, M., Steneck, R., Watson, R., 2008. A Global Map of Human Impact on Marine Ecosystems. Science 319, 948-952.

Ierodiaconou, D., Monk, J., Rattray, A., Laurenson, L., Versace, V.L., 2011. Comparison of automated classification techniques for predicting benthic biological communities using hydroacoustics and video observations. Continental Shelf Research 31, S28-S38.

Isaaks, E., Srivastava, R., 1989. An introduction to applied geostatisitics. Oxford University Press, New York, Oxford.

Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. Academic Press, London.

Kenny, A.J., Cato, I., Desprez, M., Fader, G., Schuttenhelm, R.T.E., Side, J., 2003. An overview of seabed mapping technologies in the context of marine habitat classification. ICES Journal of Marine Science 60, 411-418.

Kubicki, A., Diesing, M., 2006. Can old analogue sidescan sonar data still be useful? An example of a sonograph mosaic geo-coded by the DECCA navigation system. Continental Shelf Research 26, 1858-1867.

Lark, R.M., 1995. Components of accuracy of maps with special reference to discriminant anlysis on remote sensor data. International Journal of Remote Sensing 16, 1461-1480.

Lark, R.M., Bishop, T.F.A., 2007. Cokriging particle size fractions of the soil. European Journal of Soil Science 58, 763–774.

Lark, R.M., Dove, D., Green, S.L., Richardson, A.E., Stewart, H., Stevenson, A., 2012. Spatial prediction of seabed sediment texture classes by cokriging from a legacy database of point observations. Sedimentary Geology 281, 35-49.

Last, D., 1992. The Accuracy and Coverage of Loran-C and of the Decca Navigator System - and the Fallacy of Fixed Errors. The Journal of Navigation 45, 36-51.

LeBas, T.P., Hühnerbach, V., 1998. PRISM: Processing of Remotely-sensed Imagery for Seafloor Mapping. A Collection of Software for the Processing, Analysis and Enhancement of Sidescan Sonar Imagery, SOC Technical Report, Southampton.

Li, J., Heap, A.D., Potter, A., Huang, Z., Daniell, J.J., 2011. Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation of mud content across the southwest Australian margin. Continental Shelf Research 31, 1365-1376.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2/3, 18-22.

Long, D., 2006. BGS detailed explanation of seabed sediment modified Folk classification.

Lu, D., Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing 28, 823-870.

Lucieer, V., Hill, N.A., Barrett, N.S., Nichol, S., 2013. Do marine substrates 'look' and 'sound' the same? Supervised classification of multibeam acoustic data using autonomous underwater vehicle images. Estuarine, Coastal and Shelf Science 117, 94-106.

Lucieer, V., Lamarche, G., 2011. Unsupervised fuzzy classification and object-based image analysis of multibeam data to map deep water substrates, Cook Strait, New Zealand. Continental Shelf Research 31, 1236-1247.

Lucieer, V.L., 2008. Object-oriented classification of sidescan sonar data for mapping benthic marine habitats. International Journal of Remote Sensing 29, 905 - 921.

Lundblad, E.R., Wright, D.J., Miller, J., Larkin, E.M., Rinehart, R., Naar, D.F., Donahue, B.T., Anderson, S.M., Battista, T., 2006. A Benthic Terrain Classification Scheme for American Samoa. Marine Geodesy 29, 89 - 111.

Luts, J., Ojeda, F., Plas, R., Van De Moor, B., De Huffel, S., Van Suykens, J.A.K., 2010. A tutorial on support vector machine-based methods for classification problems in chemometrics. Analytica Chimica Acta 665, 129-145.

Marsh, I., Brown, C., 2009. Neural network classification of multibeam backscatter and bathymetry data from Stanton Bank (Area IV). Applied Acoustics 70, 1269-1276.

Martín-Fernandéz, J.A., Thió-Henestrosa, S., 2006. Rounded zeros: some practical aspects for compositional data, in: Buccianti, A., Mateu-Figueras, G., Pawlowsky- Glahn, V. (Eds.), Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society of London, pp. 191–201.

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

McKenzie, D.P., Mackinnon, A.J., Peladeau, N., Onghena, P., Bruce, P.C., Clarke, D.M., Harrigan, S., McGorry, P.D., 1996. Comparing correlated kappas by resampling: Is one level of agreement significantly different from another? Journal of Psychiatric Research 30, 483-492.

Murray, A.B., Thieler, E.R., 2004. A new hypothesis and exploratory model for the formation of large-scale inner-shelf sediment sorting and "rippled scour depressions". Continental Shelf Research 24, 295-315.

Ojeda, G.Y., Gayes, P.T., Van Dolah, R.F., Schwab, W.C., 2004. Spatially quantitative seafloor habitat mapping: example from the northern South Carolina inner continental shelf. Estuarine, Coastal and Shelf Science 59, 399-416.

Pawlowsky-Glahn, V., Olea, R.A., 2004. Geostatistical Analysis of Compositional Data. Oxford University Press, New York.

Pingree, R.D., Griffiths, D.K., 1979. Sand transport paths around the British Isles resulting from $M_2$ and $M_4$ tidal interactions. Journal of the Marine Biological Association of the United Kingdom 59, 497-513.

Plets, R., Clements, A., Quinn, R., Strong, J., Breen, J., Edwards, H., 2012. Marine substratum map of the Causeway Coast, Northern Ireland. Journal of Maps 8, 1-13.

Preston, J., 2009. Automated acoustic seabed classification of multibeam images of Stanton Banks. Applied Acoustics 70, 1277-1287.

Rattray, A., Ierodiaconou, D., Laurenson, L., Burq, S., Reston, M., 2009. Hydro-acoustic remote sensing of benthic biological communities on the shallow South East Australian continental shelf. Estuarine Coastal and Shelf Science 84, 237-245.

Rooper, C.N., Zimmermann, M., 2007. A bottom-up methodology for integrating underwater video and acoustic mapping for seafloor substrate classification. Continental Shelf Research 27, 947-957.

Schwarzer, K., Diesing, M., Larson, M., Niedermeyer, R.-O., Schumacher, W., Furmanczyk, K., 2003. Coastline evolution at different time scales - examples from the Pomeranian Bight, southern Baltic Sea. Marine Geology 194, 79-101.

Simons, D.G., Snellen, M., 2009. A Bayesian approach to seafloor classification using multi-beam echo-sounder backscatter data. Applied Acoustics 70, 1258-1268.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment 62, 77-89.

Stephens, D., Diesing, M., 2014. A Comparison of Supervised Classification Methods for the Prediction of Substrate Type Using Multibeam Acoustic and Legacy Grain-Size Data. PLoS ONE 9, e93950.

Stewart, H.A., Bradwell, T., 2013. Seafloor glacial features reveal ice streaming and re-advance of the last British Ice Sheet, offshore eastern UK: new evidence from multibeam echosounder data, American Geophysical Union Fall Meeting, San Francisco, USA.

Tauber, F., Emeis, K.-C., 2005. Sediment mobility in the Pomeranian Bight (Baltic Sea): a case study based on sidescan-sonar images and hydrodynamic modelling. Geo-Marine Letters 25, 221-229.

Visser, H., de Nijs, T., 2006. The Map Comparison Kit. Environmental Modelling & Software 21, 346-358.

Wilken, D., Feldens, P., Wunderlich, T., Heinrich, C., 2012. Application of 2D Fourier filtering for elimination of stripe noise in side-scan sonar mosaics. Geo-Marine Letters 32, 337-347.

Wynn, R.B., Bett, B.J., Evans, A.J., Griffiths, G., Huvenne, V.A.I., Jones, A.R., Palmer, M.R., Dove, D., Howe, J.A., Boyd, T.J., partners, M., 2012. Investigating the feasibility of utilizing AUV and Glider technology for mapping and monitoring of the UK MPA network. National Oceanography Centre, Southampton, p. 244.

**Table 1. Split of ground-truth stations into training and validation sets**

| Substrate class | Training set | Validation set |
|---|---|---|
| **Coarse sediment (CS)** | 43 | 15 |
| **Mixed sediments (Mx)** | 4 | 1 |
| **Mud and sandy mud (Mu)** | 15 | 3 |
| **Sand and muddy sand (Sa)** | 78 | 26 |

**Table 2. Derivatives calculated from primary acoustic data layers.**

| Feature | Description | Neighborhood (pixels) |
|---|---|---|
| **Slope** | Slope gradient | 3 |
| **Aspect** | Expressed as eastness and northness | 3 |
| **Local Moran's I*** | Spatial auto-correlation in a neighbourhood (Moran 1950) | 5 |
| **Curvature** | Curvature in the direction of slope and perpendicular to slope | 3 |
| **BPI** | Bathymetric Position Index (Lundblad et al., 2006) | 3,5,10,25 |
| **Roughness*** | Range of values in neighbourhood (Wilson et al., 2007) | 3 |

* Indicates calculated for backscatter as well as bathymetry

**Table 3. Error matrices for object-based image analysis, Random Forest and geostatistics.**

| | Ground-truth | | | | | |
|---|---|---|---|---|---|---|
| **Mapped classes** | **CS** | **Mx** | **Mu** | **Sa** | **Total** | **Purity** |
| ***a) Object-based image analysis*** | | | | | | |
| **CS** | 12 | 0 | 1 | 8 | 21 | 57.14% |
| **Mx** | 0 | 0 | 0 | 0 | 0 | 0.00% |
| **Mu** | 0 | 0 | 0 | 0 | 0 | 0.00% |
| **Sa** | 3 | 1 | 2 | 18 | 24 | 75.00% |
| **Total** | 15 | 1 | 3 | 26 | 45 | |
| **Representation** | 80.00% | 0.00% | 0.00% | 69.23% | | |
| **Overall accuracy** | | **66.67%** | | **Kappa coefficient** | | **0.378** |
| ***b) Random Forest*** | | | | | | |
| **CS** | 11 | 0 | 1 | 3 | 15 | 73.33% |
| **Mx** | 0 | 0 | 0 | 0 | 0 | 0.00% |
| **Mu** | 0 | 0 | 0 | 0 | 0 | 0.00% |
| **Sa** | 4 | 1 | 2 | 23 | 30 | 76.67% |
| **Total** | 15 | 1 | 3 | 26 | 45 | |
| **Representation** | 73.33% | 0.00% | 0.00% | 88.46% | | |
| **Overall accuracy** | | **75.56%** | | **Kappa coefficient** | | **0.515** |
| ***c) Geostatistics*** | | | | | | |
| **CS** | 7 | 0 | 0 | 2 | 9 | 77.78% |
| **Mx** | 0 | 0 | 0 | 0 | 0 | 0.00% |
| **Mu** | 0 | 0 | 0 | 0 | 0 | 0.00% |
| **Sa** | 8 | 1 | 3 | 24 | 36 | 66.67% |
| **Total** | 15 | 1 | 3 | 26 | 45 | |
| **Representation** | 46.67% | 0.00% | 0.00% | 92.31% | | |
| **Overall accuracy** | | **68.89%** | | **Kappa coefficient** | | **0.340** |

**Table 4. Evaluation of statistical significance of differences in classification accuracy. All p-values are >0.05 and hence there are no significant differences at the 5% level of significance in the accuracy statistics.**

| | **McKenzie et al. (1996)** | | **McNemar χ-squared test with continuity correction** | |
|---|---|---|---|---|
| | Difference in kappa | p-value | $X^2$ | p-value |
| **Random Forest v. Geostatistics** | 0.175 | 0.1134 | 0.5715 | 0.4497 |
| **Random Forest v. OBIA** | 0.137 | 0.2215 | 1.125 | 0.2888 |

**Table 5. Results of the map comparisons expressed as percent agreement.**

| | OBIA | Random Forest | Geostatistics |
|---|---|---|---|
| **Manual** | 82.3% | 76.7% | 68.1% |
| **OBIA** | | 86.9% | 71.4% |
| **Random Forest** | | | 73.1% |

***Highlights***

- We investigate four mapping approaches in their ability to map seabed substrates.
- We assess classification accuracy and similarity of map outputs.
- Differences in accuracy are not statistically significant at the 5% level.
- Map agreement ranged from 68% to 87%.
- We propose strategies to increase classification accuracy.

Figure 1. Area selected for the common data set exercise in the North Sea (inset). Left panel shows bathymetry in relation to Chart Datum (CD) and the locations of training and validation sample data. Right panel shows backscatter strength as digital number (DN) and classified samples (CS –Coarse sediment; Sa – Sand and muddy sand; Mu – Mud and sandy mud; Mx – Mixed sediments).

Figure 2. Ternary diagram showing four sedimentary substrate classes (Long, 2006).

Figure 3. Box plots of bathymetry and backscatter plotted against substrate types (refer to Table 1 Table 1for explanation of abbreviations). Horizontal lines indicate mean values, boxes indicate quartiles, whiskers show standard deviation, and open circles are outliers. An outlier is any point that falls below QL – 1.5*IQR or above QU + 1.5*IQR, where IRQ is the difference between the quartiles, QL is the value of the lower quartile and QU is the value of the upper quartile.

Figure 4. Resulting seabed substrate maps. Refer to Table 1**Table 1** for explanation of abbreviations. Also shown are bathymetry and backscatter strength. Note that the geostatistics method did not make use of acoustic data.

Figure 5. Purity and representation for the different substrates and methods. OBIA – Object-based image analysis; RF – Random Forest; Gstat – Geostatistics. Refer to Table 1**Table 1** for explanation of substrate abbreviations.

Figure 6. Area percentages for the five substrate types and four approaches. Manual – Manual interpretation; OBIA – object-based image analysis; RF – Random Forest; Gstat – Geostatistics.

Figure 7. Results of the map comparisons: A – Manual-OBIA; B – Manual-Random Forest; C – Manual-Geostatistics; D – OBIA-Random Forest; E – OBIA-Geostatistics; F – Random Forest-Geostatistics. Green indicates agreement and red indicates disagreement.

Figure 1

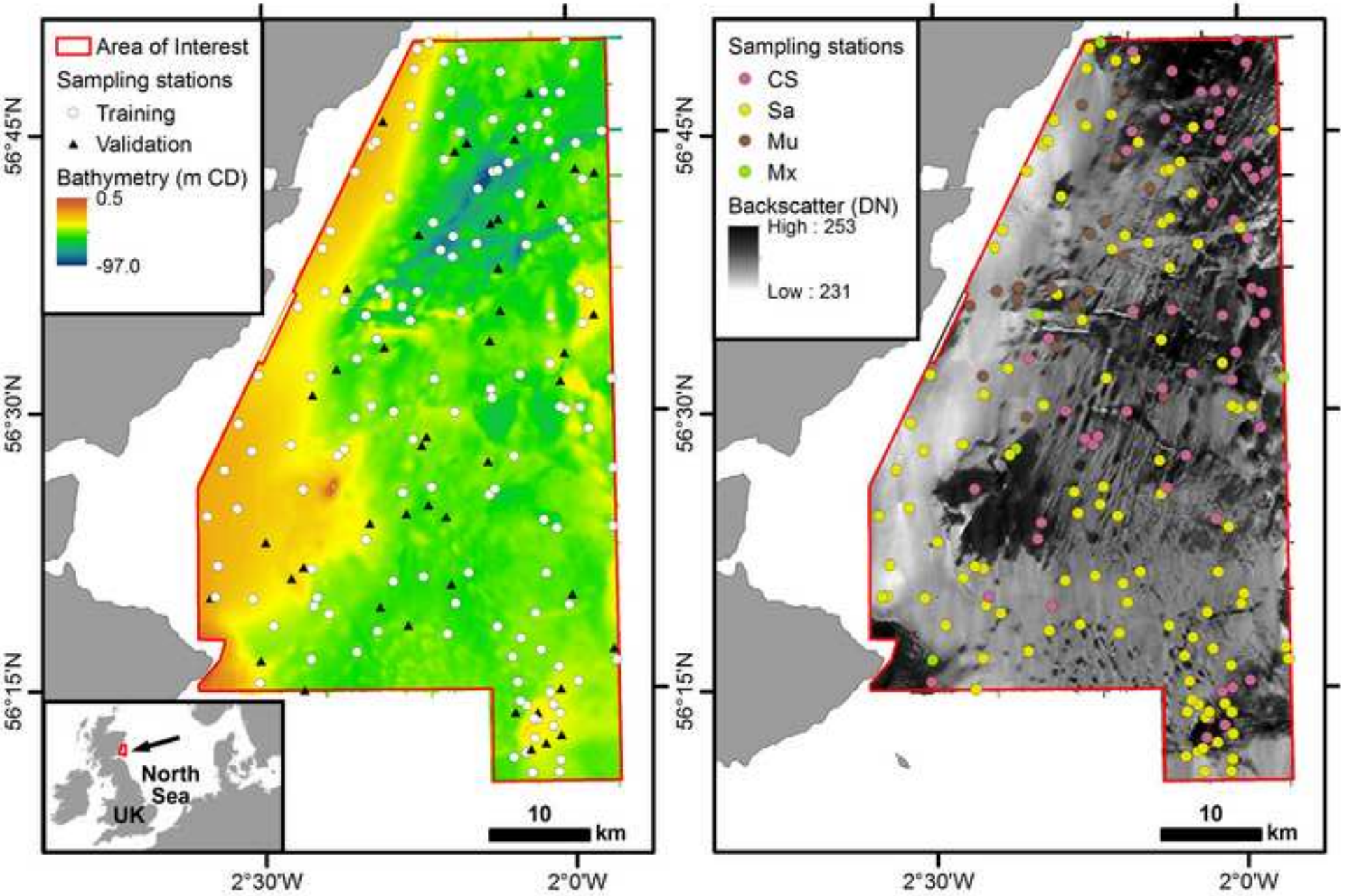
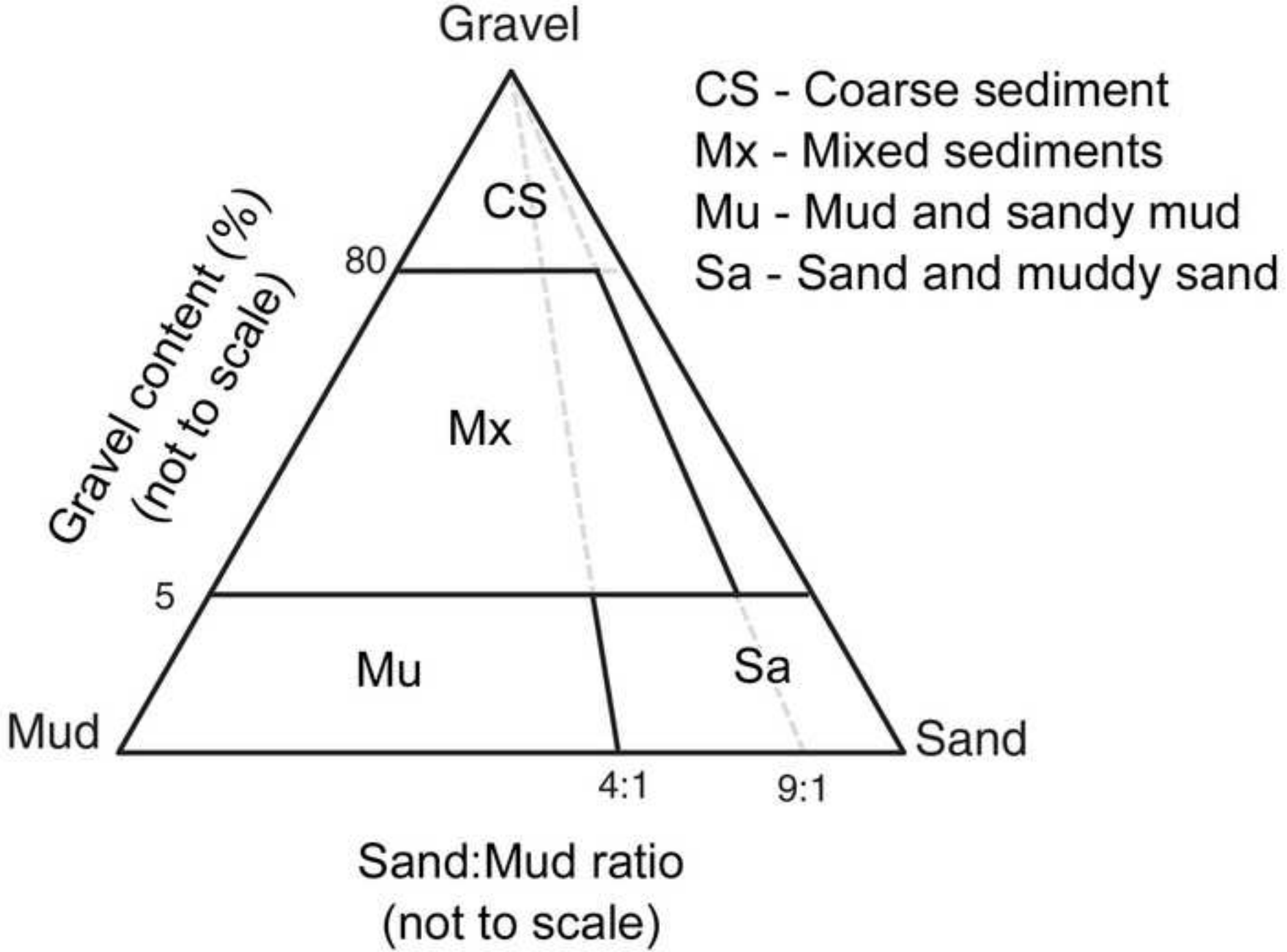
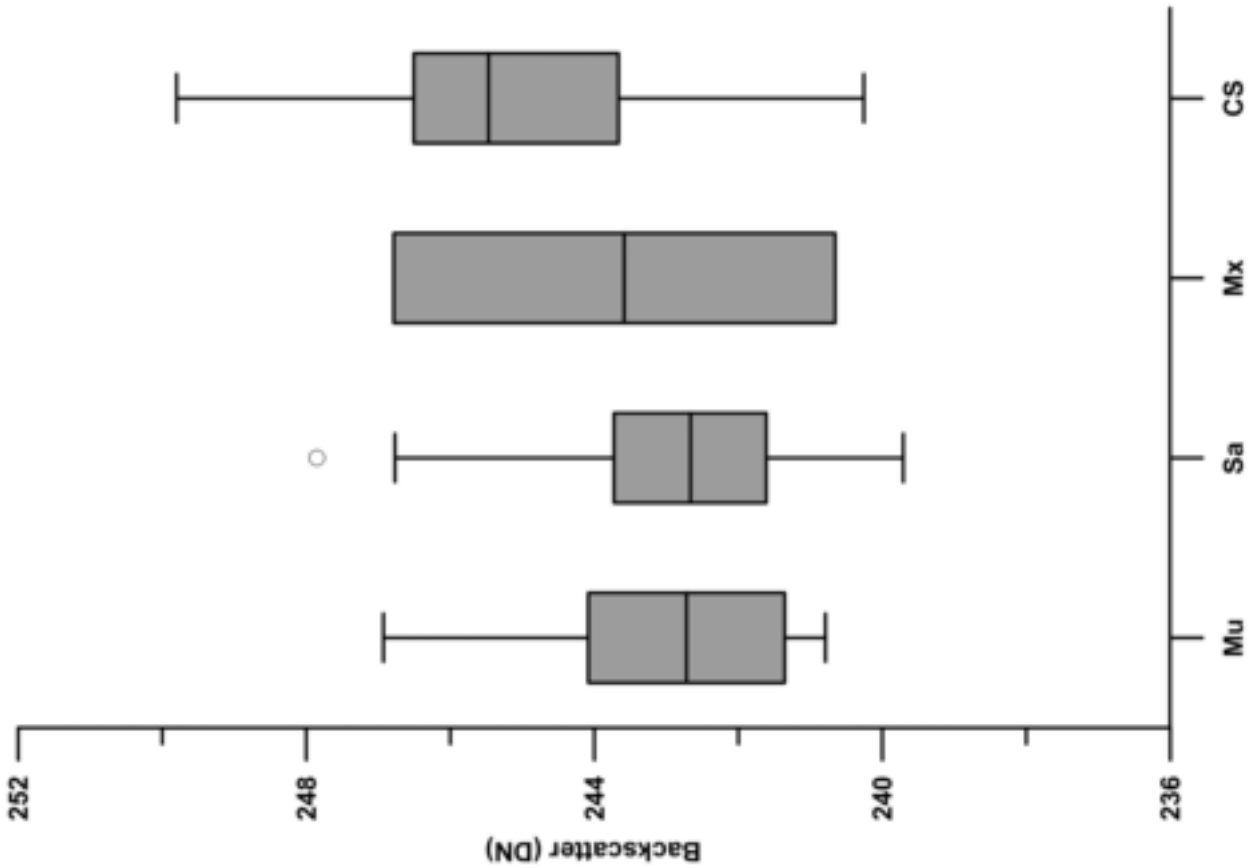


Figure 2





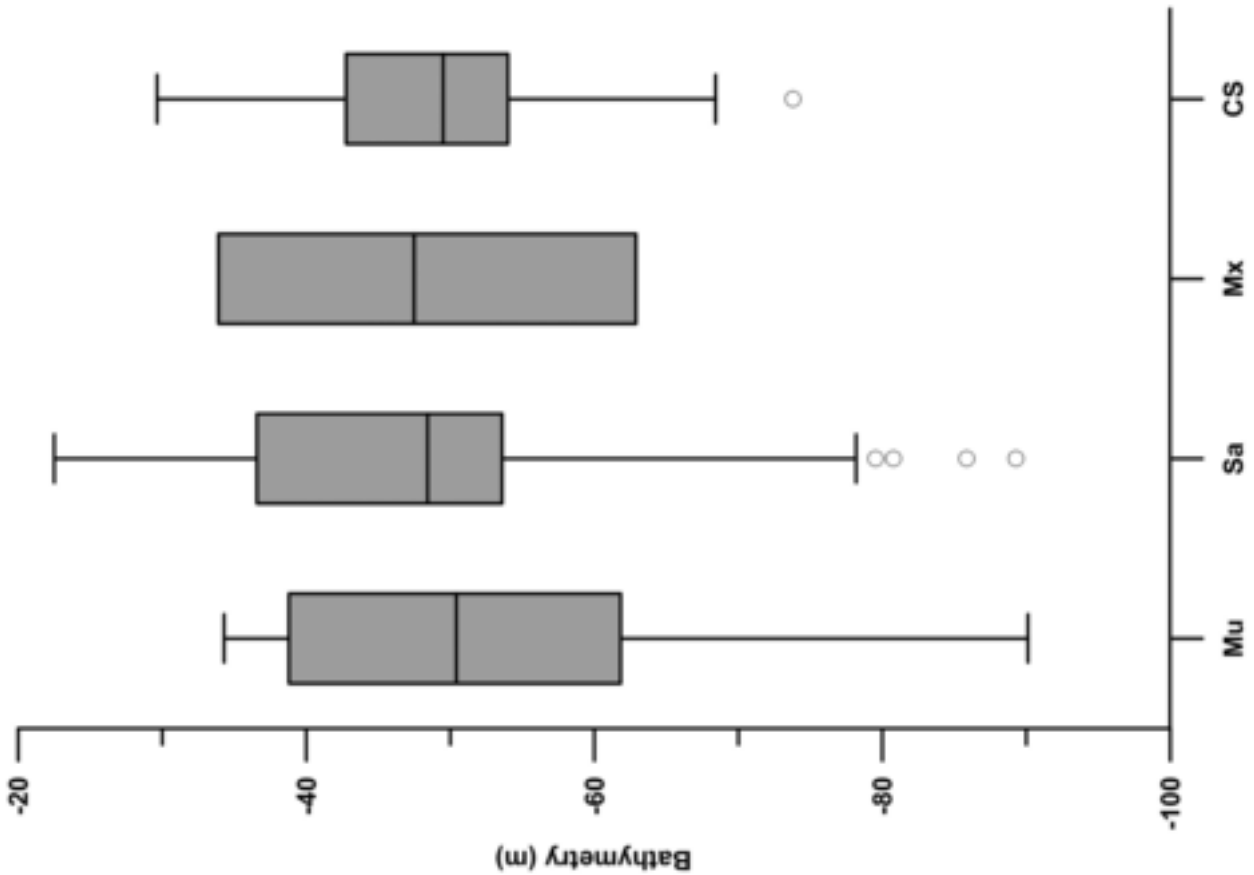
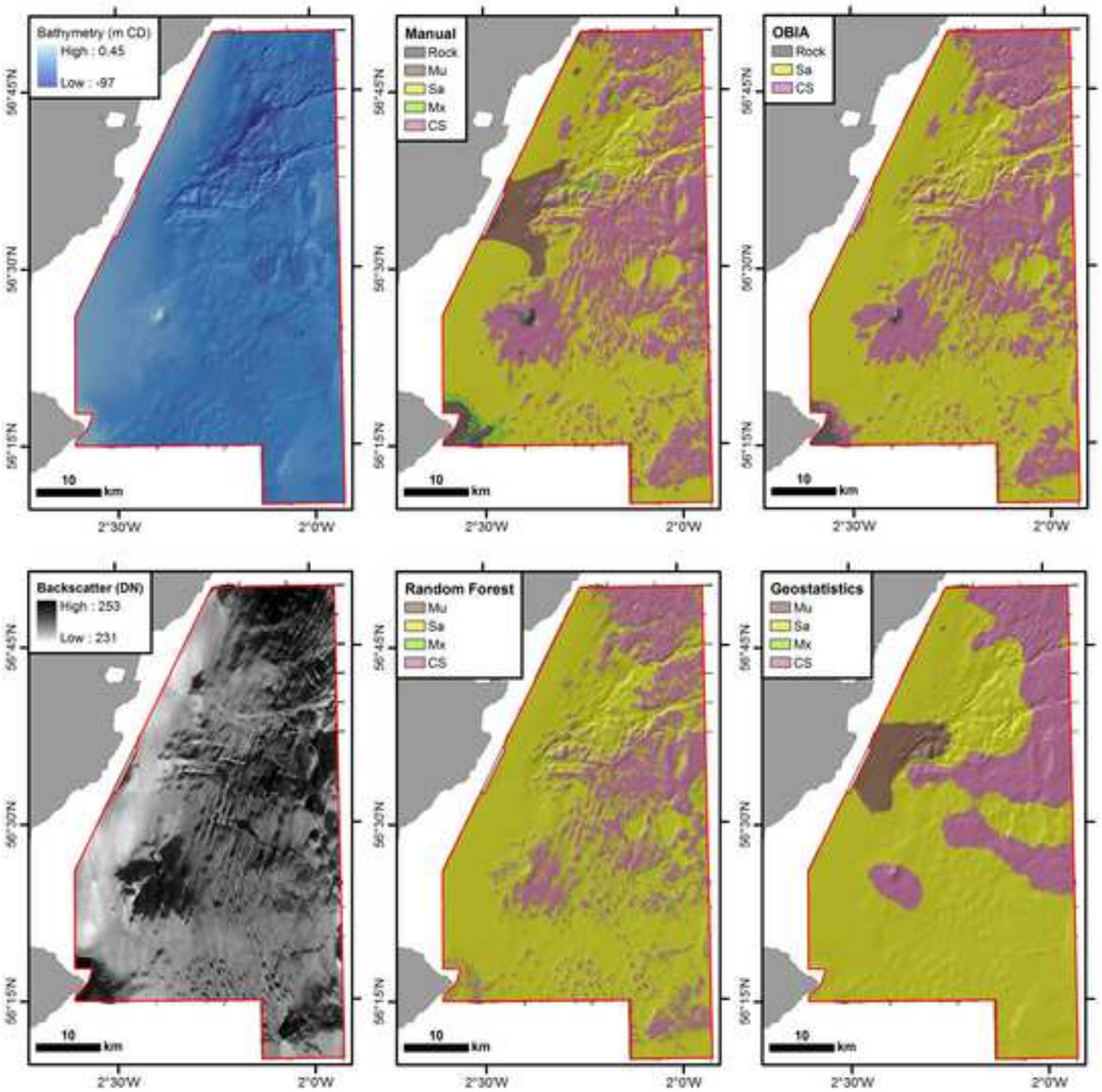




Figure 3

Figure 4





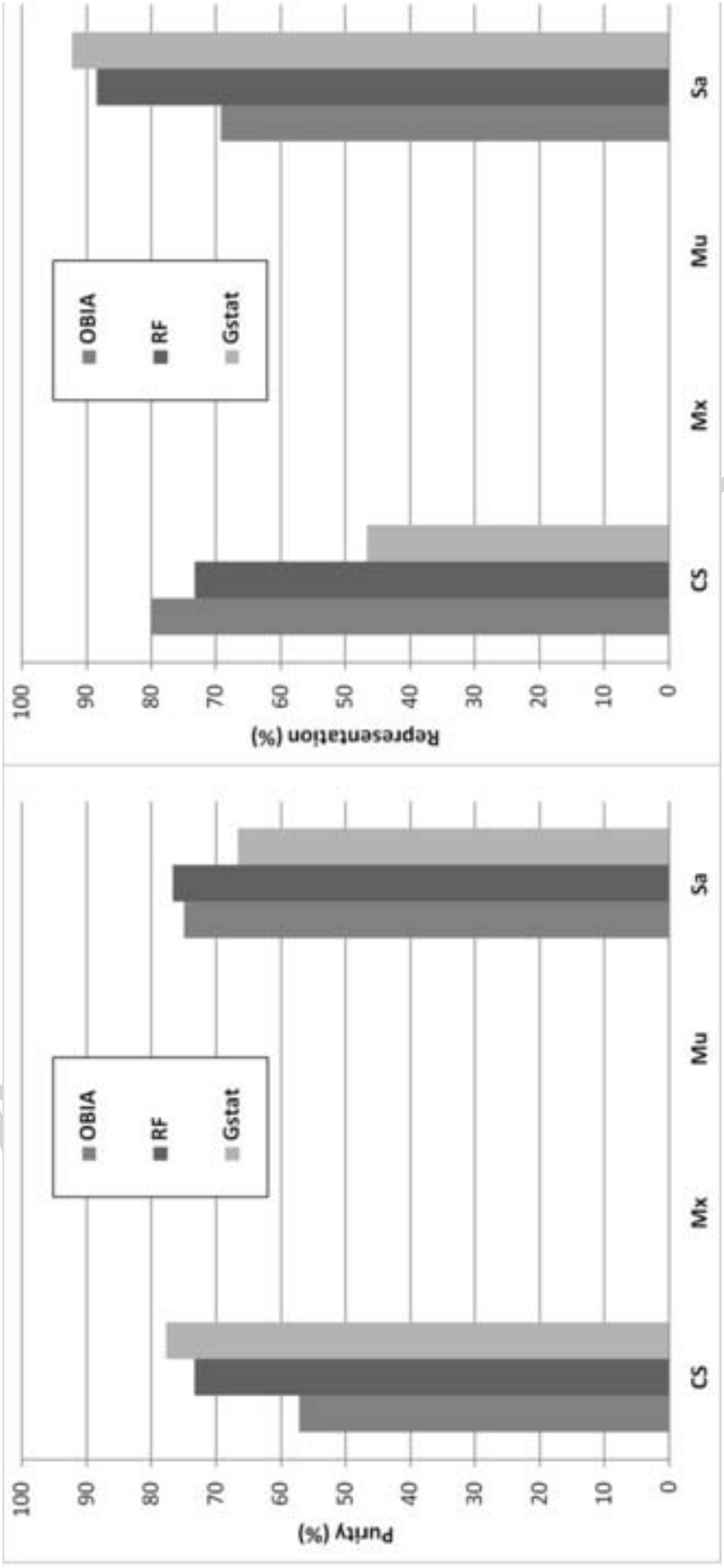

Representation (%)
100
90
80
70
60
50
40
30
20
10
0
CS
Mx
Mu
Sa
OBIA
RF
Gstat
Purity (%)
100
90
80
70
60
50
40
30
20
10
0
CS
Mx
Mu
Sa
OBIA
RF
Gstat

Figure 5

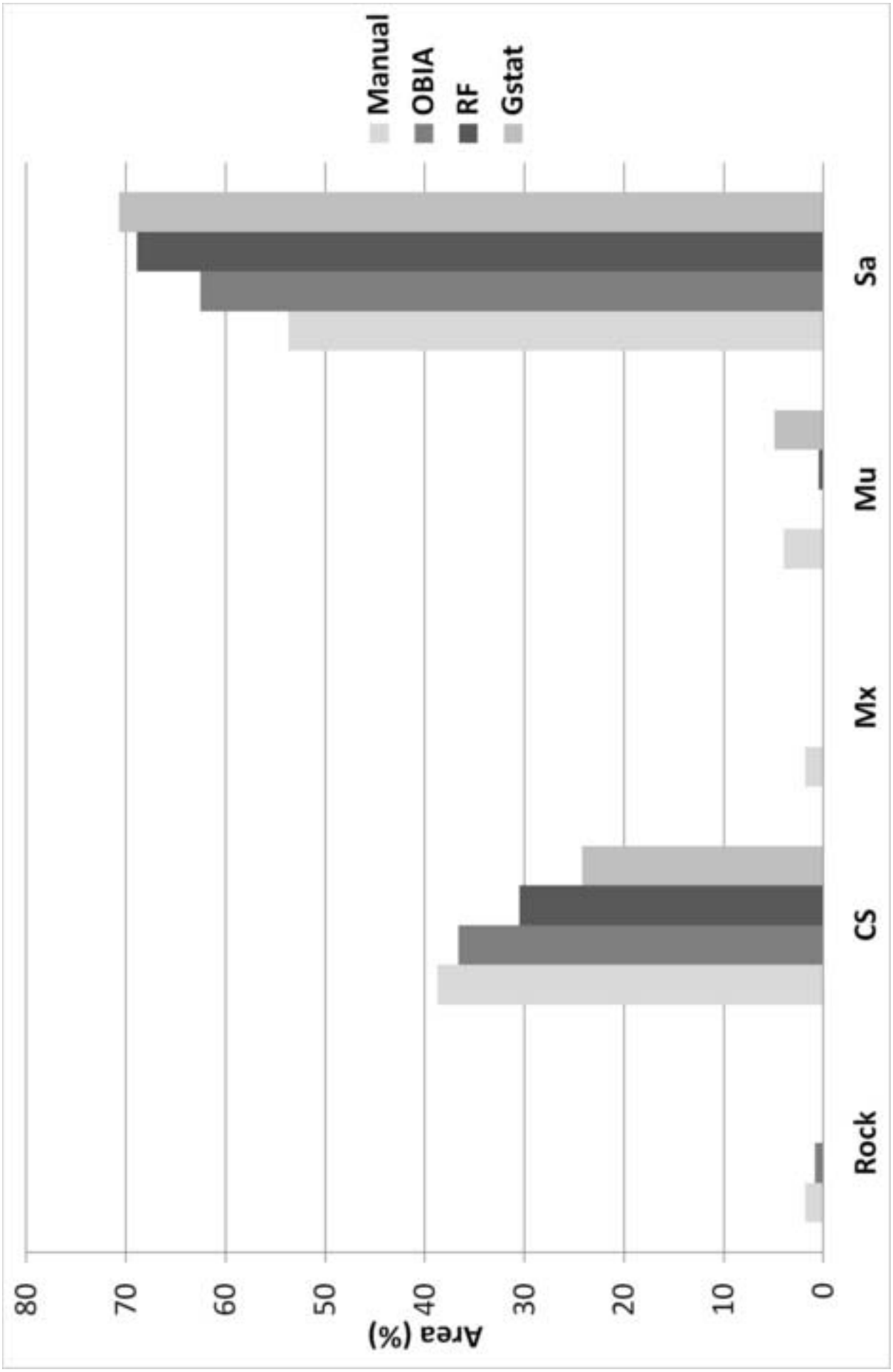

Manual
OBIA
RF
Gstat
Area (%)
0
10
20
30
40
50
60
70
80
Rock
CS
Mx
Mu
Sa

Figure 6

Figure 7