JNCC Report
No. 569

Semi-automated mapping of rock in the English Channel and Celtic Sea

Diesing, M., Green, S.L., Stephens, D., Cooper, R. & Mellett, C.L.

July 2015

© JNCC, Peterborough 2015

ISSN 0963 8901

For further information please contact:

Joint Nature Conservation Committee
Monkstone House
City Road
Peterborough PE1 1JY
www.jncc.defra.gov.uk

This report should be cited as:

Diesing, M. Green, S.L., Stephens, D., Cooper, R. & Mellett, C.L. 2015. Semi-automated mapping of rock in the English Channel and Celtic Sea. *JNCC Report*, No. 569

British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

Cefas

Summary

This report describes the results from a semi-automated approach to the mapping of bedrock outcropping at the seabed. The method consists of two elements, namely 1) the automated spatial prediction of the presence and absence of rock at the seabed using a random forest ensemble model, and 2) manual editing of the model outputs based on ancillary geological data and expert knowledge. The method is applied to Charting Progress 2 regions 3 (Eastern Channel) and 4 (Western Channel and Celtic Sea), but is expected to be applicable to other regional seas as well.

Automated predictions were made based on observations on the presence and absence of rock (response variable) and various predictor variables including bathymetry, several derivatives of bathymetry (slope, rugosity, bathymetric position index etc.), modelled hydrodynamics (depth averaged tidal current speeds and peak wave orbital velocities) and geological information such as the relative resistance to erosion based on bedrock age and lithology, indicators of sediment mobility and presence of hard substrate at or near the seabed. The accuracy of the model output was assessed based on an independent set of test data and accuracy statistics indicated that results were satisfactory (overall accuracy: 83%). Visual inspection did reveal that mis-classifications occurred in places and the model outputs were adjusted accordingly. The confidence in the developed rock layer was assessed based on the type (quality) of bathymetric data, model agreement of the random forest ensemble and agreement between predictions and observations in a spatially explicit way. Confidence scores were amended where manual edits were made in a systematic manner. The final output gives a significantly improved representation of the presence of bedrock at the seabed in the English Channel and Celtic Sea.

Contents

1 Introduction

1.1 Background

In order to prepare a set of data layers to be used by stakeholders of the Marine Conservation Zones (MCZ) regional projects, the Department for Environment, Food and Rural Affairs (Defra) let contract MB0103 in 2009-10 to produce a UK-wide data layer showing areas of rock and hard substrate at or near the seabed surface (Gafeira *et al* 2010). The British Geological Survey (BGS) carried out this work as a subcontractor of ABPmer. The outputs were:

1. Rock and hard substrate polygon layer
2. Rock and cobbles point layer
3. Confidence layer
4. Layer showing areas in which multi-beam bathymetry data has been collected.

In 2011, BGS updated the polygon dataset and named it DigHardSubstrate250, which is provided alongside version 3 of the BGS' offshore seabed sediments map (DigSBS250).

The Joint Nature Conservation Committee (JNCC) has a responsibility for reporting on the status of the UK's reefs, which is a habitat defined under Annex I of the Habitats Directive[1]. Reefs are made up of three sub-types: bedrock, stony and biogenic reef. DigHardSubstrate250 is therefore a useful product for JNCC in that it indicates the potential location and extent of bedrock reef at or near the seabed. In addition, DigHardSubstrate250 forms the rock part of the European Marine Observation and Data Network (EMODnet) Geology[2] seabed substrate layer, which feeds into EUSeaMap – the broad-scale predictive map output of the EMODnet Seabed Habitats[3] project. EUSeaMap is used for regional and national scale assessments, such as assessing the representativeness of marine protected area networks for broad-scale habitats.

There is currently no process in place for periodic updates of DigHardSubstrate250; however, new multibeam echosounder and sample data are being collected every year, meaning that existing data products, and therefore assessments and conservation advice to Government, can quickly become out of date. The Gafeira *et al* (2010) method relied heavily upon expert judgement, which has many benefits but can be time-consuming and not easily repeatable. Therefore, in order to enable future consistent and repeatable updates that benefit from both automated approaches and expert judgement, a new method is required – ideally one that is faster, more objective with an audit trail of expert-based decisions and a consistent, easily understood description of confidence.

A further substantial improvement would be a separation between bedrock outcropping at the seabed surface and bedrock that is covered by a thin veneer of sediment, as the presence of sediment on top of rock might have significant ecological consequences in terms of habitat provision.

[1] Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora. Official Journal of the European Communities No L 206/7.
[2] http://www.emodnet.eu/geology
[3] http://www.emodnet.eu/seabed-habitats

1.2 Aims and objectives

1.2.1 Project aims

The aim of this project is to produce a standard interpretation for rock distribution for Charting Progress 2 (CP2) regions 3 and 4 that can be used by the different agencies of the national and devolved governments. This project is designed as a pilot study with the aim to demonstrate a method that maximises the benefits of automated mapping approaches and in-depth geological knowledge. The developed method should be able to be applied subsequently to all other parts of the UK shelf.

1.2.2 Objectives

1. Develop a semi-automated method that combines the strengths of objective, repeatable spatial predictions with in-depth knowledge of the geology and marine environment of the study site.
2. Develop a vector-based geospatial data product showing the potential extent of rock at, or near (i.e. covered with thin sediment), the sea floor for subtidal areas of CP2 regions 3 and 4 at a spatial scale equivalent to 1:250,000.
3. Keep a record of manual edits made to allow for efficient updates in future.
4. Carry out a three-step confidence assessment for each polygon and include scores in the output data product.

1.2.3 Sub-types of rock

Below is a more detailed description of the meaning of the two sub-types mentioned in point 2 above:

Rock at the surface: Rock present at outcrop. This suggests a habitat dominated by exposed bedrock. Whilst it is unlikely that large areas of exposed rock will exist with zero sediment cover present, this classification should capture areas of negligible or highly mobile, patchy sediments where the veneer is minimal.

Rock with thin sediment: These are essentially sub-crops of bedrock, i.e. areas where bedrock rises to the seabed surface, but remains largely covered by a thin veneer of sediment. This will be derived by subtracting areas predicted as ‘rock at the surface’ from previously mapped rock areas (DigHardSubstrate250).

2 Materials and methods

2.1 Study site

The study site comprises Charting Progress 2 regions 3 (Eastern Channel) and 4 (Western Channel and Celtic Sea), adjusted for the 2014 revised Exclusive Economic Zone boundary (Figure 1). However, suitable bathymetric data do not exist west of approximately 7° 30' W (Western Approaches). Automated spatial predictions are therefore restricted to the seabed area east of 7° 30' W and existing polygons of rock outcrops are used to fill the gap in the Western Approaches.

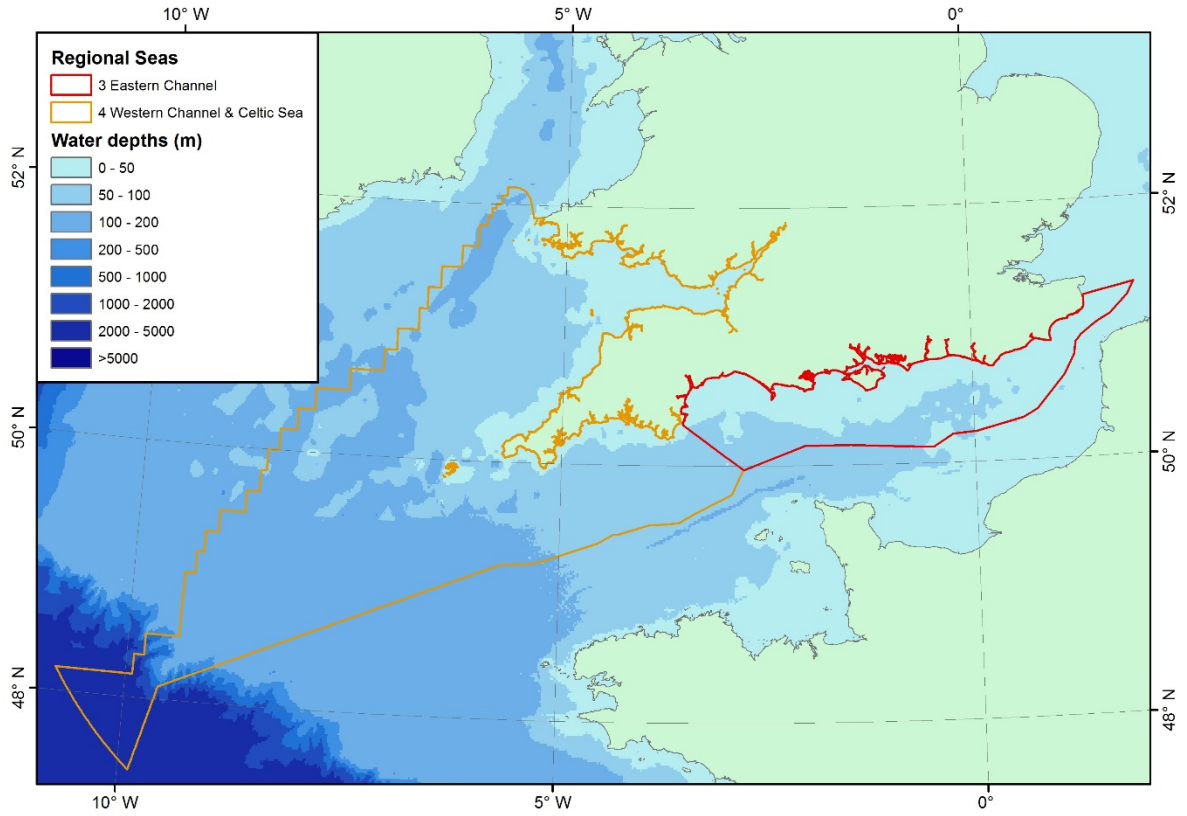

Figure 1. Charting Progress 2 regional sea boundaries, adjusted for the 2014 revised Exclusive Economic Zone boundary. This study is concerned with regions 3 and 4.

2.2 Data

2.2.1 Substrate observations

The input dataset contained 10,590 substrate observations within the study area. These were obtained from the Defra marine vector dataset (JNCC 2011). These data have been successfully used in previous studies involving the mapping of rocky substrates *(Stephens et al* 2014). Of these data, 8,434 (79.6 %) were unambiguous absence of rock. 1,664 (15.7 %) were unambiguous presence of rock, i.e. rock and no other substrate type was recorded. In 212 (2 %) cases rock was indicated as the predominant substrate type but other substrate types were also recorded, while 280 (2.6 %) observations included 'rock' in the substrate type but it was not the dominant type. Only unambiguous presence/absence records were used resulting in 492 (4.6 %) observations being removed. This decision was made following

trials, which indicated that excluding ambiguous samples would give the most accurate predictions.

2.2.2 Predictor features

In order to predict rock presence at unobserved locations, the substrate observations had to be related to auxiliary variables (referred to as features) that have continuous coverage across the study area. These predictor features are comprised of a bathymetry digital elevation model (DEM), topographic characteristics derived from the bathymetry (such as slope and roughness), outputs from hydrodynamic modelling and polygon layers indicating properties of the seabed. Detailed descriptions of all features are given in Table 1.

Table 1. Predictor features.

Feature	Description	Unit	Name	Reference
Bathymetry	Bathymetry (water depth) projected to UTM 30 North at a resolution of 30 m. Available high-resolution multibeam bathymetry, from MPA sites with a presence of rock, was also included.	m	bathy	(Astrium Oceanwise 2011)
Roughness	Derived from bathymetry; the difference between minimum and maximum of cell and its 8 neighbours.	m	rgh	(Wilson *et al* 2007)
Slope	Derived from bathymetry, the maximum slope gradient	degree	slope	(Wilson *et al* 2007)
Aspect	Derived from bathymetry, direction of steepest slope, expressed as Eastness (sine of aspect) and Northness (cosine of aspect)		eastness, northness	(Wilson *et al* 2007)
Curvature	Derived from bathymetry, rate of change of slope. Profile curvature is measured parallel to maximum slope; plan curvature is measured perpendicular to slope.		curv_pl, curv_pr, curv	(Wilson *et al* 2007)
Bathymetric Position Index (BPI)	Derived from bathymetry, vertical position of cell relative to neighbourhood (identifies topographic peaks and troughs). Radii of 10, 20, 30, 40 and 50 pixels were used.	m	BPI10, BPI20, BPI30, BPI40, BPI50	(Lundblad *et al* 2006)
BGS Hard Substrate	DigHardSubstrate250 data product. Delineates areas of rock at outcrop, or overlain by thin (<0.5 m) sediment based on bathymetric data, the BGS legacy sample database and expert interpretation.		BGS_HS	(Gafeira *et al* 2010)
Indicators of Mobile Sediments	Seabed morphologies characteristic of mobile sediments were delineated using hillshade, slope and rugosity data.		BGS_IMS	(Westhead *et al* 2014)
Quaternary Thickness	Data layer detailing thickness of Quaternary cover on the UK Continental Shelf.	m	BGS_QT	(Westhead *et al* 2014)
Relative Resistance	Representation of the relative resistivity of bedrock on the UK Continental Shelf based on age		BGS_BRR	(Clayton & Shamoon 1998)

	and lithology. Derived utilising BGS DigRock250[4] following the method described by Clayton & Shamoon, (1998).			
Distance to Coast	Euclidean distance to nearest coastline	m	DC	
Current Velocity (mean)	Mean tidal current velocity averaged across water column calculated using a Telemac model with an unstructured grid of variable resolution.	m/s	UVmean	
Peak Wave Orbital Velocity	Peak wave orbital velocity at seabed. Surface wave parameters (wave height and period) were output from a POLCOM model for the years 2000 to 2008. The original resolution of the data in 12 km. Bottom orbital velocities were calculated from these and the 6 arcsec Defra DEM. Maximum, mean and standard deviation of peak orbital velocity were calculated.	m/s	MaxPeakUrms, StdDevPeakUrms	(Holt & James 2001; Aldridge *et al* 2015)

2.3 Methods

2.3.1 Pre-processing of observations

The first step was to extract the values of each predictor feature at the location of each substrate observation. The quality and reliability of the bathymetry data is not constant across the study area. The dataset is a collation of all available data, mostly collected since the 1980s and differing techniques were used to collect and process the data. This means that, although the grid resolution is constant at 30m across the study area, the underlying data is of varying quality and this will also affect the topographic variables derived from the bathymetry. Table 2 shows the number of observations in each category of bathymetry quality. The observations falling in the lowest class of bathymetry (Chart) were removed from the dataset as well as where the data quality was unknown (NA).

Table 2. Observations by bathymetry quality class. Ordered by reliability from left to right.

Type	*NA*	*Chart*	*Interpolated*	*Singlebeam echosounder*	*Multibeam echosounder*
Number of observations	12	1540	640	4772	3626
Percent of observations	0.1	14.5	6	45	34.2

Not all predictor features had the same spatial extent or resolution. This meant that for some predictor features there were gaps, mainly around the coast or far offshore, resulting in some observation locations having no data values (NA) for some or all features. Any observations that contained NA values for at least one predictor feature were discarded.

A total of 1,127 cases had NA values in the predictors and were discarded. The breakdown of the number of NA values by feature indicate the highest number was for the wave orbital velocity layer. This resulted in a data set of 7,469 observations with which to train and test the random forest prediction model. Of these, 1,449 (19.4 %) observations were ‘presence’ (P) of bedrock and 6,020 (80.6 %) were ‘absence’ (A).

[4] http://www.bgs.ac.uk/downloads/start.cfm?id=2892

In order to test the model predictions, the data were split randomly into training and test sets. Two-thirds of the data were used to train the model and 33% used to test its predictive performance. The ratio of presence to absence records were approximately equal in both the training and test sets (Table 3).

Table 3. Training and test sets.

	Training	Test
P	951 (19.2 %)	498 (19.6 %)
A	3978 (80.8 %)	2042 (80.4 %)

2.3.2 Model training

A random forest (RF) classification model was trained (Breiman 2001). RF has become one of the most widely used and successful statistical learning models for classification and regression, showing good performance in a large number of domains (Che Hasan *et al* 2014; Diesing *et al* 2014; Cutler *et al* 2007; Prasad *et al* 2006; Pal 2005; Chapman *et al* 2010; Chan & Paelinckx 2008; Oliveira *et al* 2012; Che Hasan *et al* 2012; Stephens & Diesing 2014; Huang *et al* 2014; Huang *et al* 2012; Lucieer *et al* 2013). RF is an ensemble technique which aggregates the results of a large number of classification trees. In this case, the 'forest' includes 2,500 classification trees. The predicted class is chosen based on the majority vote from the individual trees. The fraction of the votes that are given for a class can be interpreted as reliability in the estimate. For example, if 95 % of the votes are for a case being presence of rock we can be more confident versus a case where only 51 % of the votes are for a case being rock. RF is a non-parametric technique, i.e. no assumptions regarding the shape of distributions of the response or predictor variables are made (Cutler *et al* 2007). It can handle complex, non-linear relationships between predictor and response variables. As well as using the test set to validate the model, RF implicitly generates a cross-validated (CV) measure of model accuracy. RF also provides a relative estimate of predictor feature importance. This is a measure of the variability explained by each feature, averaged across every tree in the RF.

Prior to training the model, a feature selection step was implemented to test the statistical significance of the predictor features for the presence/absence prediction of rock. The Boruta algorithm (Kursa & Rudnicki 2010) is a feature selection wrapper (Guyon & Elisseeff 2003) based on the RF model. The algorithm uses the feature importance score generated by RF to test each of the predictor features against the effect of random noise. Only features that have scores significantly higher than random are retained for use in the final model.

To evaluate the predictive performance of the model, three statistics were calculated using the observed versus predicted class labels: 1) Classification accuracy, the percentage of the observations correctly classified; 2) Cohen's kappa statistic (Cohen 1960), which incorporates 'expected' agreement; 3) Balanced error rate (BER), which is the mean of the error rate for each class (Luts *et al* 2010). It is important to calculate a range of statistics to evaluate model performance as each can illuminate different aspects of the model prediction ability. For example, classification accuracy will not tell you whether the model is biased towards a specific class, which can occur especially if the class frequencies are uneven, as is the case with the data in this study.

2.3.3 Knowledge-based enhancements

The output of the RF predictions was then reviewed manually by a mapping geologist, in order to assess its validity in terms of the established geology of the area. The first stage of

the process involved conversion of the modelled output into a readily editable ESRI shape file, in accordance with the project requirement. The steps detailed below were followed:

- Conversion of the RF output to 20m raster in order to perform generalisation.
- Each cell was replaced with a majority of eight neighbouring cells. This essentially reduces smaller areas and increases a large (majority) area.
- Boundary cleaning; smoothing of the boundaries between zones by buffering and debuffering. This results in smaller areas being engulfed into a larger ones, according to boundary length. Large areas have a higher priority to expand into smaller ones.
- Conversion of raster back to polygons.
- Elimination of polygons smaller than 0.015625km^2 (based on a minimum mappable unit feature with a diameter of 125m for 1:250K mapping).
- Aggregate polygons with less than 125m distance between features, then removal of holes.

Following the generalisation process the modelled output was reviewed against published mapping and sample data, and polygons were deleted, edited and added in accordance with the geological understanding of the region.

A number of small, irregular polygons were removed on the basis that they represented artefacts from the bathymetric data.

In the eastern section of the study area numerous elongated features were identified as rock at outcrop from the RF mapping. These features have previously been the subject of extensive study and are well documented as sediment banks *(*James *et al* 2010*;* Hamblin *et al* 1992). Features including Bassurelle, Shingle, Bullock and Varne Banks were edited as part of the manual process and removed from the rock category. It is likely that the high relief associated with these features resulted in mis-classification. Further to the west, in the central part of the study area numerous sediment waves were also included in the rock classification. These were removed on the basis of their morphology and an understanding of the hydrodynamics of the area (Figure 2).

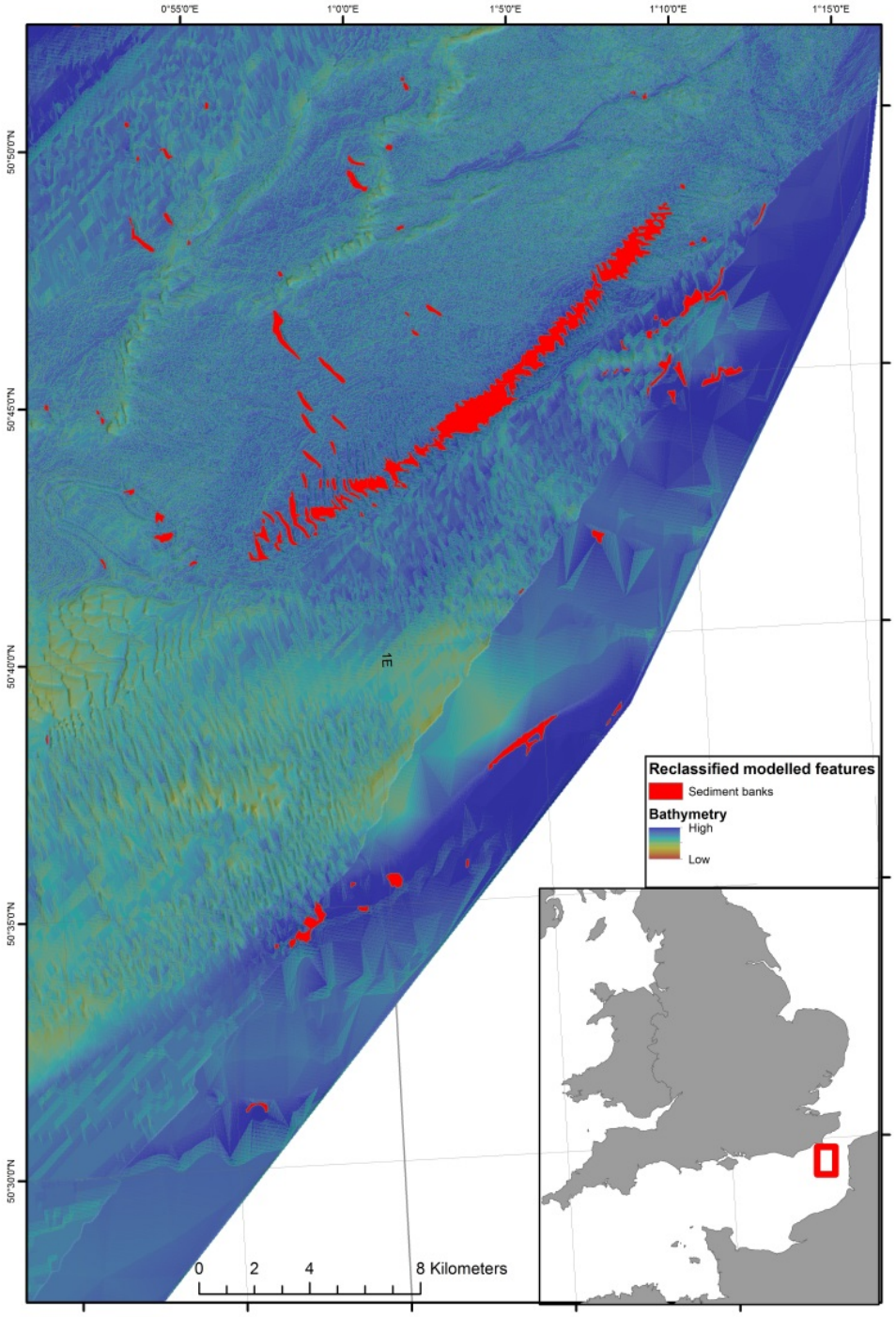

Figure 2. Image showing reclassified Sediment bank feature in the Eastern Channel. Bathymetry from Astrium, Oceanwise (2011).

Additions from a number of studies were also incorporated as part of the expert interpretation phase. Cefas mapping based on monitoring programmes from Lyme Bay and Torbay (Jenkins & Eggleton 2014) and Wight Barfleur Reef (Barrio Froján *et al* 2014), JNCC Reef/Not Reef map compilation version 7 (Ellwood 2013), and MESH South West Approaches Canyons survey data (Davies *et al* 2008) were considered in the mapped output, following review.

The second category required, 'Rock with thin sediment' was not included in the RF approach. These polygons are derived from the BGS DigHardSubstrate250 mapped output.

The polygons included capture areas where rock is anticipated within 0.5m of the seabed. Details of the method applied are recorded in Gafeira *et al* (2010). The output polygons from the 'Rock at outcrop' semi-automated routine were subtracted from the DigRock250 shapefile to form a combined data set indicating rock at outcrop and subcrop.

In addition to the rock outputs, a 'changes' shapefile was also generated in order to document the modifications to the modelled output.

2.3.4 Confidence assessment

The confidence assessment method follows a three-step approach similar to that used to assess confidence in EUNIS habitat maps (Ellwood 2014) but tailored for this project. The assessment was performed on a per-polygon basis due to the possible heterogeneity of inputs into the model across the output area. The method requires the assessor to follow the flow diagram shown in Figure 3 and score the polygon appropriately at each stage.

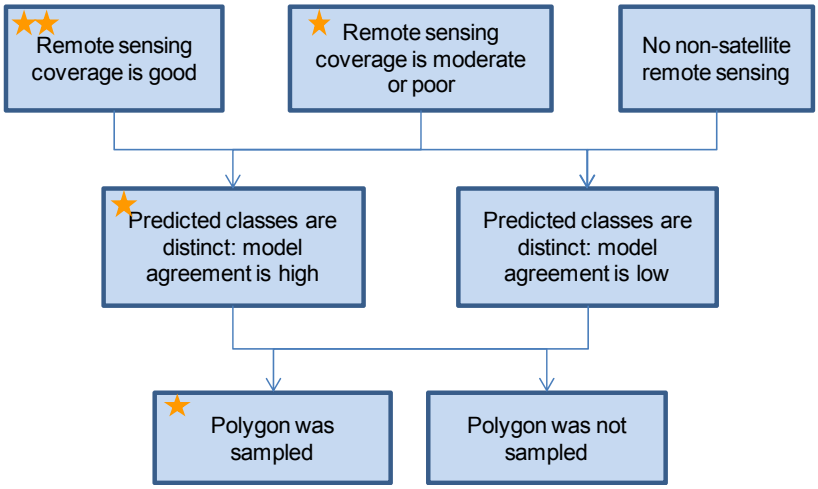

Figure 3. Three-step confidence decision tree; the assessor starts at the top and follows the arrows. Stars/points are awarded according to the answers given and the final score is the sum of the stars/points.

From this method, a maximum qualitative score of 4 can be achieved by a polygon (Table 4). The final score should not be taken as a quantitative probability of the habitat's likelihood in extent or presence, the measurement is a qualitative score based on the data inputs and level of agreement between the predictive models.

Table 4. All possible combinations of scores under the three-step scheme. Polygons with equal scores are therefore assumed to have roughly similar levels of confidence, regardless of the route through the decision tree.

Score	Remote sensing coverage	Distinctness of class boundaries	Amount of sampling
4	★★	★	★
3	★★		★
	★★	★	
	★	★	★
2	★★		
	★		★
	★	★	
		★	★
1	★		
		★	
			★
0			

Application for polygons identified as rock at outcrop

Remote Sensing Coverage was assessed based on the type of acoustic data that were available: A score of two was given where multibeam echosounder data were present, a score of one for singlebeam echosounder data and a score of zero for all other data types.

Class Distinctness was scored in two stages:

1. Initially the agreement of the random forest ensemble outputs were used: a score of one was attained where agreement was high (>75%), indicating high probability of presence of rock, or low (<25%), indicative of high probability of absence of rock. Intermediate values ranging from 25% to 75% were given a score of zero.
2. Following the knowledge-based enhancements, where expert judgement led to modification or addition of a polygon the initial score was overwritten with a score of one. On this basis, additions from previous studies such as Wight Barfleur Reef and Lyme Bay were assigned a value of one. This therefore indicates higher confidence associated with validation of the presence of an area of rock outcrop by more detailed study or assessment by a geologist.

In the case of the Amount of Sampling criterion, a score of one was given if a polygon was sampled and the majority of samples agreed with the prediction. A score of zero was attained if a polygon was not sampled or the majority of samples within the polygon disagree with the prediction.

Application for polygons identified as rock with thin sediment

The BGS DigHardSustrate250 dataset includes an assessment of confidence based on data density. However, for production of the shapefile in this project a standard value of zero was applied for Remote Sensing Coverage as limited bathymetry data was available for the production of this shapefile. A value of one was applied for the Class Distinctness criterion, to reflect the influence of human judgement was also applied. As the same legacy database was used for the samples, the same approach was used for the Amount of Sampling criterion as for the updated ‘Rock at outcrop’ polygons.

3 Results

3.1 Random forest predictions

The feature selection process indicated that only one feature (Quaternary Thickness) did not contribute significantly to the presence/absence predictions; this feature was removed from the model.

The accuracy produced by the cross-validation (CV) showed encouraging results (Table 5). The classification accuracy statistic of >80 % seems to be a good result, however as explained this shouldn't be taken in isolation because of the imbalanced class frequency (80 % of the observations are Absence). The kappa statistic is 0.5 which indicates 'moderate' agreement (Viera & Garrett 2005). The BER for both CV and test sets is 0.23. The statistics for the test set are almost identical to the CV results (Table 6). This shows that the training of the model has not been 'over-fitted' to the training data and it is generalising real patterns in the data.

The error rate for Presence is 0.33, indicating that 2/3 of instances where presence of bedrock is predicted are correct. The error rate for Absence is lower (0.13). This could be a result of bias in the model resulting from the uneven class frequencies. The fact that there are more Absence observations means that when the model is 'unsure' in a sense it defaults to predicting Absence (as this is more likely to be correct in a completely random situation).

Table 5. Cross-validation confusion matrix and performance statistics.

	(Predicted) Absence	*(Predicted) Presence*	*Class Error*
(Observed) A	3463	515	0.13
(Observed) P	310	641	0.33
Accuracy %	83.26		
Kappa	0.503		
BER	0.23		

Table 6. Test set confusion matrix and performance statistics.

	(Predicted) Absence	(Predicted) Presence	Class Error
(Observed) A	1777	265	0.13
(Observed) P	164	334	0.33
Accuracy %	83.11		
Kappa	0.502		
BER	0.23		

Feature importance scores (Figure 4) indicated bathymetric roughness (rgh) as being the most important variable, its score is considerably higher than slope, which is the second most important. Interestingly variability in wave orbital velocity (StdDevPeakUrms) is the 3rd highest scoring feature. The least important feature with substantially lower score than any other (although still indicated as statistically significant by the feature selection process) was the indicator of mobile sediments (BGS_IMS) layer.

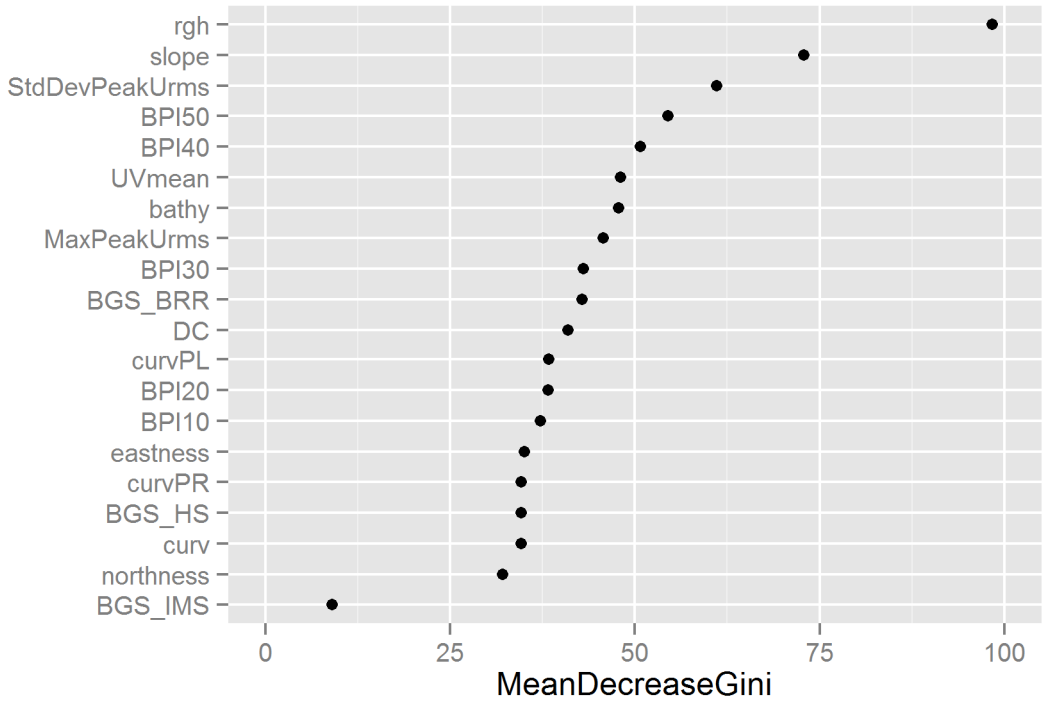

Figure 4. Relative feature importance indicated by RF. MeanDecreaseGini is the cross-validated measure of variability explained by the feature. See Breiman 2003 for details.

3.2 Knowledge-based enhancements

The results of the expert interpretation phase represent a validation of the RF modelled approach. The addition of polygons of rock at outcrop derived from the semi-automated process represent significant added value to the existing 'rock with thin sediment' data layer.

The pilot study area is characterised by extensive areas of rock within 0.5m of the sea bed. The updates to the 'rock at the seabed' data indicate that a relatively small proportion of the area mapped as hard substrate by Gafeira *et al* (2010) is likely to be comprised of outcropping rock.

Extensive areas of outcropping rock are recorded in the Bristol Channel, Haig Fras and Western Channel. As roughness was identified as the most influential parameter, it is possible that flatter platforms of outcropping rock may not be fully quantified. However, they will be captured in the rock with thin sediment data included in the output layer (Figure 5).

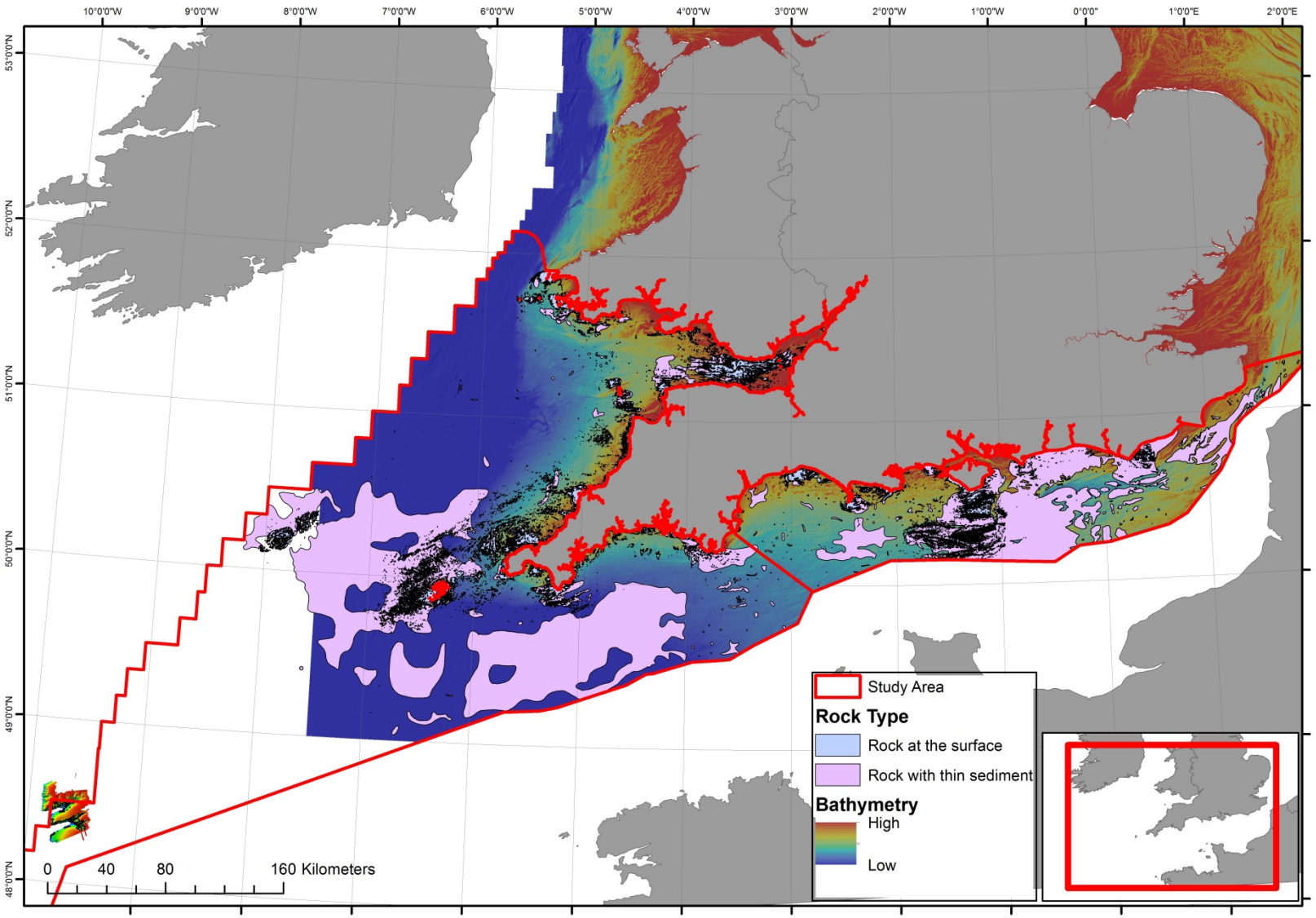

Figure 5. Distribution of rock at the seabed surface and rock covered with thin sediment (<0.5m) within the study area.

3.3 Confidence assessment

The confidence assessment recorded results between zero and four (Figure 6). A limited number of polygons achieved the higher confidence values of three and four. The majority of these polygons were located in the areas where pre-existing interpretation was incorporated (Wight Barfleur Reef, Lyme Bay and Torbay areas and the SW Approaches) reflecting the influence of multibeam bathymetry data and expert interpretation on the scores recorded.

The majority of the mapped polygons recorded confidence values of one or two. The biggest limiting factor on limiting achievement of higher confidence values was the sampling criteria applied. The majority of the large polygons associated with rock with thin sediment, recorded a confidence value of one or two. This is indicative of their inclusion as possible areas of rock based on the previous assessment (Gafeira *et al* 2009), although they have not been fully reviewed as part of this study.

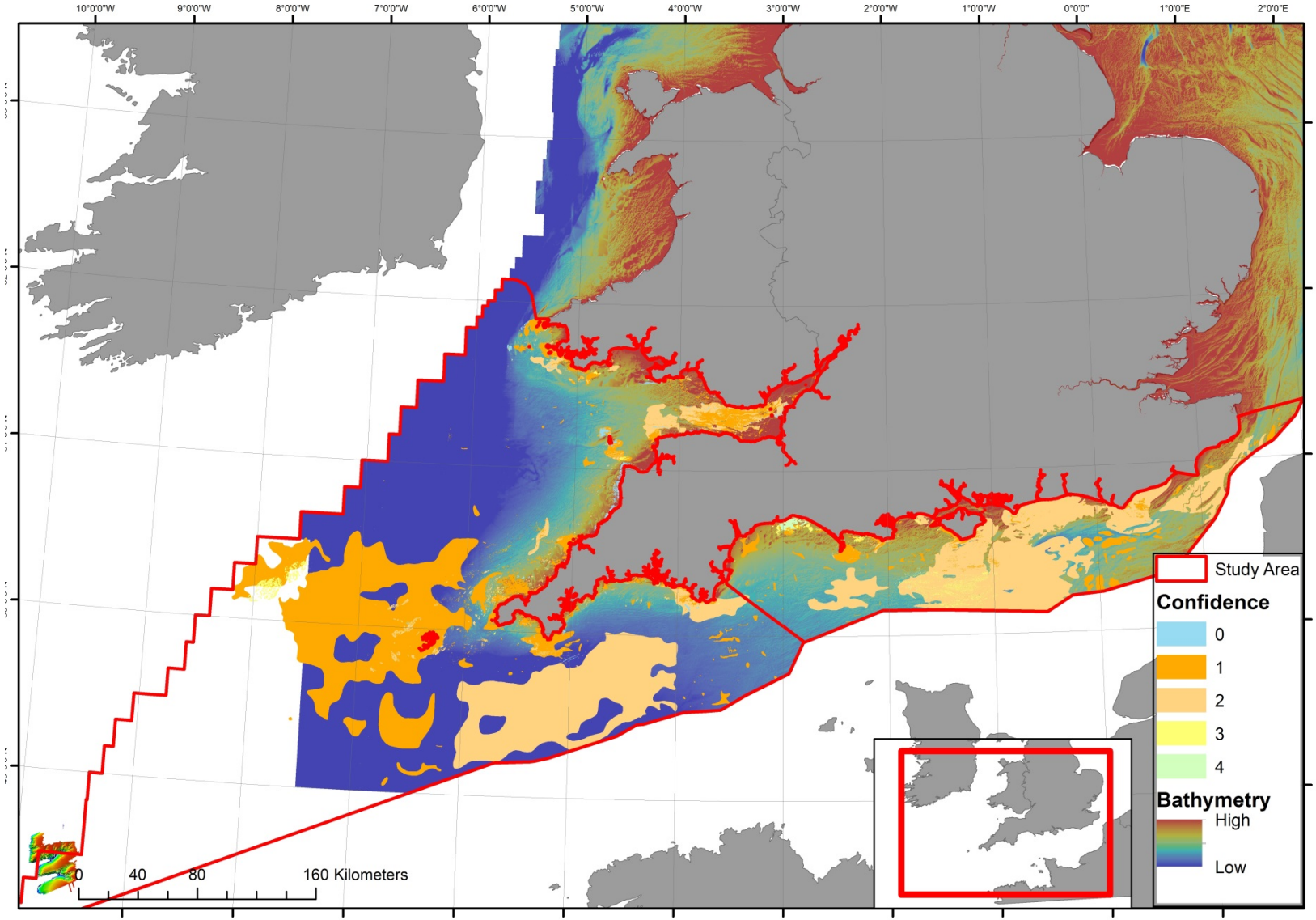

Figure 6. Confidence assessment of the updated map output. Values between zero and four, where four indicates maximum confidence.

4 Discussion

We have derived a new data layer of rock in the English Channel and the Celtic Sea. A significant improvement over previous data layers is the distinction between bedrock outcropping at the seabed and rock covered by a thin veneer of sediment, which was previously not made (Gafeira *et al* 2010). This distinction is however of ecological significance, as even a thin veneer of sediment on top of rock might turn a habitat dominated by epibiota attached to hard substrates into an infauna-dominated habitat. As one example, Pisces reef in the Irish Sea (Callaway *et al* 2009) is a bedrock outcrop smothered with mud, which is thick enough in places to provide suitable habitat for the burrowing Norway lobster (*Nephrops norvegicus*).

The derived data layer has a nominal scale of 1:250,000 and as such gives a sufficiently detailed indication of the distribution of rock at or near the seabed in the English Channel and Celtic Sea at a regional scale. Whilst the data layer was derived by using the best available data sources and methods, it should be noted that the derived results are unlikely to be sufficient for detailed monitoring of change in reef extent, due to the inherent and unavoidable inaccuracies in data and methods.

We have demonstrated how automated approaches to seabed mapping and in-depth geological knowledge can be combined to derive an improved representation of bedrock at and near the seabed in the English Channel and the Celtic Sea. In essence, this means that the applied method could be described as semi-automated, as it contains aspects of automated mapping as well as expert intervention. It would certainly be desirable to develop a fully automated method with the aim to reduce subjectivity and increase reproducibility. However, this would require i) a complete understanding of the underlying processes that lead to exposure of bedrock at the seabed, and ii) exhaustive datasets that fully describe the predictor-response relationships.

With incomplete knowledge and data, the best option to derive meaningful predictions is a combined approach as demonstrated in this report. It is noteworthy that we have made an effort to include as much knowledge as possible at the automated prediction stage by including predictor variables that are known or expected to influence the presence of rock at the seabed. Likewise, it should be noted that tools like variable importance plots are useful in understanding which variables are suitable predictors. Unsurprisingly, terrain parameters (roughness, slope, and BPI) and hydrodynamics were important predictors. However, the most important of the hydrodynamic predictors was the standard deviation of the annual mean peak orbital velocities. This rather unexpected result might indicate that it could be worth investigating relationships between hydrodynamic forcing and resulting substrate type in more detail as well as suggesting that developing higher resolution hydrodynamic models would be useful for improving predictions.

The insights gained from the variable importance plot and the manual reclassification of seemingly misclassified objects could be fed back to the automated classification stage and it could be expected that such an iterative process will improve automated prediction results and reduce the amount of expert intervention required. Such an interaction could be repeated until no further improvements in classification accuracy are achieved. Additionally, new or improved data become available over time (e.g. improvements to the Defra DEM reflecting new hydrographic survey data). It would therefore be desirable to regularly update the predictions in order to reflect improvements in data, methods and knowledge. The general method that was set up as part of the project lends itself to such a task as processes of automated prediction and knowledge-based enhancements have been formalised.

The presented method is also applicable to other parts of the UK continental shelf, for which similar response and predictor datasets are available (samples, bathymetry and geological layers) or could be created (hydrodynamics). This improved method will hopefully lead to more regular updates of the DigHardSubstrate250 (or equivalent) data product and therefore ensure conservation decisions are based on the best available data.

The shapefiles produced by this approach represent a significant update to our previous understanding of the distribution of rock at, or near, the seabed in the English Channel and Celtic Sea area. This can be used to inform future updates to seabed substrate maps such as those included in EMODnet Geology and the BGS map series. The seabed substrate maps are currently being updated, as part of a rolling process within the EMODNet Geology group. It is envisaged that these updates from this project will be incorporated in summer 2015.

5 References

Aldridge, J.N. *et al* 2015. Assessment of the physical disturbance of the northern European Continental shelf seabed by waves and currents. *Continental Shelf Research*. Available at: http://www.sciencedirect.com/science/article/pii/S0278434315000576 [Accessed March 18, 2015].

Astrium Oceanwise. 2011. *Creation of a high resolution digital elevation model (DEM) of the British Isles continental shelf*

Barrio Froján, C., Diesing, M. & Rance, J. 2014. *Characterisation of the Wight-Barfleur Reef cSAC*, Peterborough, UK.

Breiman, L. 2003. Manual On Setting Up, Using, And Understanding Random Forests V3.1.

Breiman, L. 2001. Random Forests. *Machine Learning*, **45**(1), pp.5–32.

Callaway, A. *et al* 2009. The impact of scour processes on a smothered reef system in the Irish Sea. *Estuarine, Coastal and Shelf Science*, **84**(3), pp.409–418. Available at: http://www.sciencedirect.com/science/article/B6WDV-4WV15TX-8/2/e6aeafcd53b106117400aa22aa042a25.

Chan, J.C.W. & Paelinckx, D. 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, **112**(6), pp.2999–3011.

Chapman, D.S. *et al* 2010. Random Forest characterization of upland vegetation and management burning from aerial imagery. *Journal of Biogeography,* **37**(1), pp.37–46.

Che Hasan, R. *et al* 2014. Integrating Multibeam Backscatter Angular Response, Mosaic and Bathymetry Data for Benthic Habitat Mapping. *PLOS ONE*, **9**(5), p.e97339. Available at: http://dx.doi.org/10.1371%2Fjournal.pone.0097339.

Che Hasan, R., Ierodiaconou, D. & Monk, J. 2012. Evaluation of Four Supervised Learning Methods for Benthic Habitat Mapping Using Backscatter from Multi-Beam Sonar. *Remote Sensing*, **4**(11), pp.3427–3443. Available at: http://www.mdpi.com/2072-4292/4/11/3427.

Clayton, K. & Shamoon, N. 1998. A new approach to the relief of Great Britain II. A classification of rocks based on relative resistance to denudation. *Geomorphology*, 25, pp.155–171.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**(1), pp.37–46.

Cutler, D.R. *et al* 2007. Random forests for classification in ecology. *Ecology*, **88**(11), pp.2783–92.

Davies, J., Guinan, J., Howell, K., Stewart, H. & Verling, E. 2008. MESH South West Approaches Canyons Survey (MESH Cruise 01-07-01) Final Report, 156pp. Available at: http://www.emodnet-seabedhabitats.eu/PDF/SWCanyons_FinalReport_v1.4_final.pdf

Diesing, M. *et al* 2014. Mapping seabed sediments: Comparison of manual, geostatistical, object-based image analysis and machine learning approaches. *Continental Shelf Research*,

84, pp.107–119. Available at:
http://www.sciencedirect.com/science/article/pii/S0278434314001629.

Ellwood, H. 2013. *Creating a Composite Map of Annex 1 Reef, Version 2.0.*, Available at:
http://jncc.defra.gov.uk/page-3054.

Ellwood, H. 2014. *Creating a EUNIS level 3 seabed habitat map integrating data originating from maps from field surveys and the EUSeaMap model*, Available at:
http://jncc.defra.gov.uk/page-6655.

Gafeira, J. *et al* 2010. *Developing the necessary data layers for Marine Conservation Zone selection - Distribution of rock/hard substrate on the UK Continental Shelf*, Edinburgh: British Geological Survey.

Guyon, I. & Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**, pp.1157–1182.

Hamblin, R.J.O. *et al* 1992. *United Kingdom offshore regional report: the geology of the English Channel*, London: HMSO for the British Geological Survey.

Holt, J.T. & James, I.D. 2001. An s coordinate density evolving model of the northwest European continental shelf: 1. Model description and density structure. *Journal of Geophysical Research: Oceans*, **106**(C7), pp.14015–14034. Available at:
http://dx.doi.org/10.1029/2000JC000304.

Huang, Z. *et al* 2014. Predictive mapping of seabed substrata using high-resolution multibeam sonar data: A case study from a shelf with complex geomorphology. *Marine Geology*, **357**, pp.37–52. Available at:
http://www.sciencedirect.com/science/article/pii/S0025322714002205.

Huang, Z. *et al* 2012. Predictive modelling of seabed sediment parameters using multibeam acoustic data: a case study on the Carnarvon Shelf, Western Australia. *International Journal of Geographical Information Science*, 26(2), pp.283–307. Available at: <Go to ISI>://WOS:000300611200006.

James, J.W.C. *et al* 2010. *The South Coast Regional Environmental Characterisation*,

Jenkins, C. & Eggleton, J.D. 2014. *Update of Annex 1 habitat mapping in the Lyme Bay and Torbay cSAC*, Lowestoft.

JNCC. 2011. DEFRA Marine Reference Data.

Kursa, M. & Rudnicki, W. 2010. Feature selection with the Boruta Package. *Journal of Statistical Software*, **36**(11), pp.1–11. Available at: http://www.jstatsoft.org/v36/i11/paper/.

Lucieer, V. *et al* 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**, pp.94–106. Available at:
http://www.sciencedirect.com/science/article/pii/S0272771412004246.

Lundblad, E.R. *et al* 2006. A Benthic Terrain Classification Scheme for American Samoa. *Marine Geodesy*, **29**(2), pp.89–111. Available at:
http://www.informaworld.com/10.1080/01490410600738021.

Luts, J. *et al* 2010. A tutorial on support vector machine-based methods for classification problems in chemometrics. *Analytica Chimica Acta*, **665**(2), pp.129–145.

Oliveira, S. *et al* 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *Forest Ecology and Management*, **275**, pp.117–129.

Pal, M. 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, **26**(1), pp.217–222.

Prasad, A.M., Iverson, L.R. & Liaw, A. 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, **9**(2), pp.181–199.

Stephens, D. *et al* 2014. *Potential interactions of seaweed farms with natural nutrient sinks in kelp beds. Marine Estate Research Report.*,

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**(4), p.e93950. Available at: http://dx.doi.org/10.1371%2Fjournal.pone.0093950.

Viera, A.J. & Garrett, J.M. 2005. Understanding interobserver agreement: The Kappa statistic. *Family medicine*, **37**(5), pp.360–3.

Westhead, R.K. *et al* 2014. *Geological Constraints on Development across the UK Continental Shelf: a study for The Crown Estate. British Geological Survey Internal Report.*

Wilson, M.F.J. *et al* 2007. Multiscale Terrain Analysis of Multibeam Bathymetry Data for Habitat Mapping on the Continental Slope. *Marine Geodesy*, **30**(1-2), pp.3–35. Available at: http://dx.doi.org/10.1080/01490410701295962.