


Article

# Assessing the Repeatability of Automated Seafloor Classification Algorithms, with Application in Marine Protected Area Monitoring

America Zelada Leon <sup>1</sup>, Veerle A.I. Huvenne <sup>2,\*</sup>, Noëlie M.A. Benoist <sup>2</sup>, Matthew Ferguson <sup>3</sup>, Brian J. Bett <sup>2</sup>  and Russell B. Wynn <sup>4</sup>

<sup>1</sup> University of Southampton, University Road, Southampton SO17 1BJ, UK; azl1n17@soton.ac.uk

<sup>2</sup> National Oceanography Centre, European Way, Southampton SO14 3ZH, UK; noelie.benoist@noc.ac.uk (N.M.A.B.); bjb@noc.ac.uk (B.J.B.)

<sup>3</sup> Joint Nature Conservation Committee, Monkstone House, City Road, Peterborough PE1 1JY, UK; Matthew.Ferguson@jncc.gov.uk

<sup>4</sup> Wild New Forest CIC, 252 Woodlands Road, Woodlands, SO40 7GH, UK; russ@wildnewforest.co.uk

\* Correspondence: vaih@noc.ac.uk

Received: 1 April 2020; Accepted: 11 May 2020; Published: 15 May 2020



**Abstract:** The number and areal extent of marine protected areas worldwide is rapidly increasing as a result of numerous national targets that aim to see up to 30% of their waters protected by 2030. Automated seabed classification algorithms are arising as faster and objective methods to generate benthic habitat maps to monitor these areas. However, no study has yet systematically compared their repeatability. Here we aim to address that problem by comparing the repeatability of maps derived from acoustic datasets collected on consecutive days using three automated seafloor classification algorithms: (1) Random Forest (RF), (2) K–Nearest Neighbour (KNN) and (3) K means (KMEANS). The most robust and repeatable approach is then used to evaluate the change in seafloor habitats between 2012 and 2015 within the Greater Haig Fras Marine Conservation Zone, Celtic Sea, UK. Our results demonstrate that only RF and KNN provide statistically repeatable maps, with 60.3% and 47.2% agreement between consecutive days. Additionally, this study suggests that in low-relief areas, bathymetric derivatives are non-essential input parameters, while backscatter textural features, in particular Grey Level Co-occurrence Matrices, are substantially more effective in the detection of different habitats. Habitat persistence in the test area between 2012 and 2015 was 48.8%, with swapping of habitats driving the changes in 38.2% of the area. Overall, this study highlights the importance of investigating the repeatability of automated seafloor classification methods before they can be fully used in the monitoring of benthic habitats.

**Keywords:** automated seafloor classification; machine learning algorithms; benthic habitat maps; autonomous underwater vehicles; Grey Level Co-occurrence Matrices; sidescan sonar

## 1. Introduction

One of the consequences of the expanding human population is the increased exploitation of marine resources through industrial activities such as fishing and seabed mining. In many cases, these activities degrade marine habitats and reduce biodiversity [1]. A growing awareness of these problems is driving nations to develop strategies and resolutions—e.g., the World Heritage Convention, the International Coral Reef Initiative, the United Nations Convention on the Law of the Sea, The Convention on Biodiversity—to help manage resources and reduce human impacts [2]. These agreements often result in the establishment of Marine Protected Areas (MPAs) that aim to restrict human activity in order to protect natural resources. Today, protected areas add up to 7.9% of global waters-equivalent to more

than 28.5 million km<sup>2</sup>—and many more areas are expected to join that list in response to targets set by many countries, ranging from 10 to 30% of their national waters to be protected by 2020—a pledge that now has been updated to 2030, in many cases [2,3].

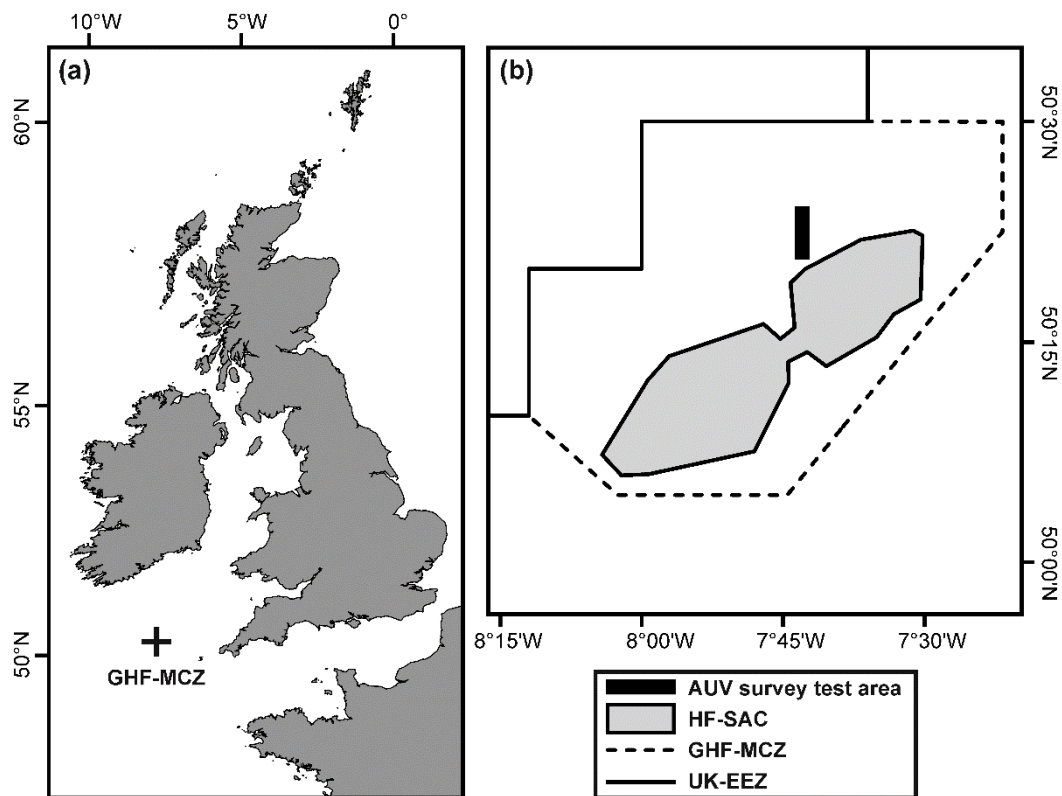
National decisions on the design and management of MPAs, and in particular their long-term monitoring, are commonly supported by benthic habitat maps [4]. Benthic habitat maps aim to characterise areas of the seabed based on the distribution of its habitats. Traditionally, these maps have been created through expert interpretation of environmental and biological data, a process that is not only time-consuming, but also subjective and lacking in consistency and repeatability [5]. Therefore, in order to effectively support decision makers in developing conservation and management plans for the growing number of protected areas, it is crucial to develop faster, repeatable, and objective methods to generate maps that accurately represent the seabed environment and the associated ecological patterns [6].

In response to this need, automated habitat mapping methods have been increasingly explored as an objective alternative [5]. Different approaches to automate habitat mapping have been studied—image-based analysis of acoustic backscatter, pixel-based (PB) and object-based (OB) image analysis [7–9], supervised and unsupervised image classification [10–12], etc. However, there is still little agreement on which is the most effective method. Instead, the key seems to be understanding the parameters used within these methods, for example, an adequate choice of spatial scale for the analysis, and a selection of truly descriptive features prior to the classification step, have been shown to impact greatly on the accuracy of the models [12,13].

Yet, these novel automated methods have scarcely been applied to the assessment of temporal changes in deep-water habitats [14–17], with little evidence of any attempt to conduct a systematic study to assess the suitability and repeatability of different methods in temporal change detection. To validate the use of automated methods in the monitoring of MPAs, it is necessary not only to assess their accuracy, but also to establish that they are repeatable, such that users can be confident that any differences in the resultant benthic habitat maps are the result of true changes to seafloor habitats, and not simply artefacts of the method.

Repeatability addresses the agreement between successive results of a measurement obtained under the same conditions within a short period of time. In automated habitat mapping methods, this concept becomes even more important when addressing change detection and monitoring [18,19]. This was acknowledged by Rattray et al. [14], who identified the unknown repeatability of the classification process as a difficult to control source of error in the classification step when researching benthic habitat change. It has been acknowledged that current acoustic technologies provide limited repeatability depending on the instrument, equipment settings, processing algorithms, operational ranges, environmental conditions, and survey methods [17,20]. Nevertheless, there are few if any studies that have attempted to compare the repeatability of different classification algorithms.

Recent technological advances in autonomous underwater vehicles (AUVs) and other robotic systems also offer the prospect of more efficient seafloor mapping opportunities, and will be of increasing importance in the monitoring of the ever-growing total extent of MPAs [21,22] and other marine monitoring challenges [23]. It is in this context that the UK National Oceanography Centre, in partnership with Defra and the NERC National Capability programs MAREMAP and CLASS, coordinated a project examining the use of AUVs for repeat mapping and monitoring of MPAs in UK waters [21]. As part of this project, acoustic surveys have been carried out periodically within the Greater Haig Fras Marine Conservation Zone (Figure 1) in the years 2012 and 2015, including repeat surveys carried out on two consecutive days in 2015. The surveys offer the ideal dataset to further the investigation of automated methods of benthic habitat mapping, and their repeatability. Hence, the aim of this study is to investigate the repeatability of three automated seafloor classification algorithms and to use the most robust one for the detection of changes in habitat distribution, thus exploring their suitability for the monitoring of MPAs and other marine spatial management measures.



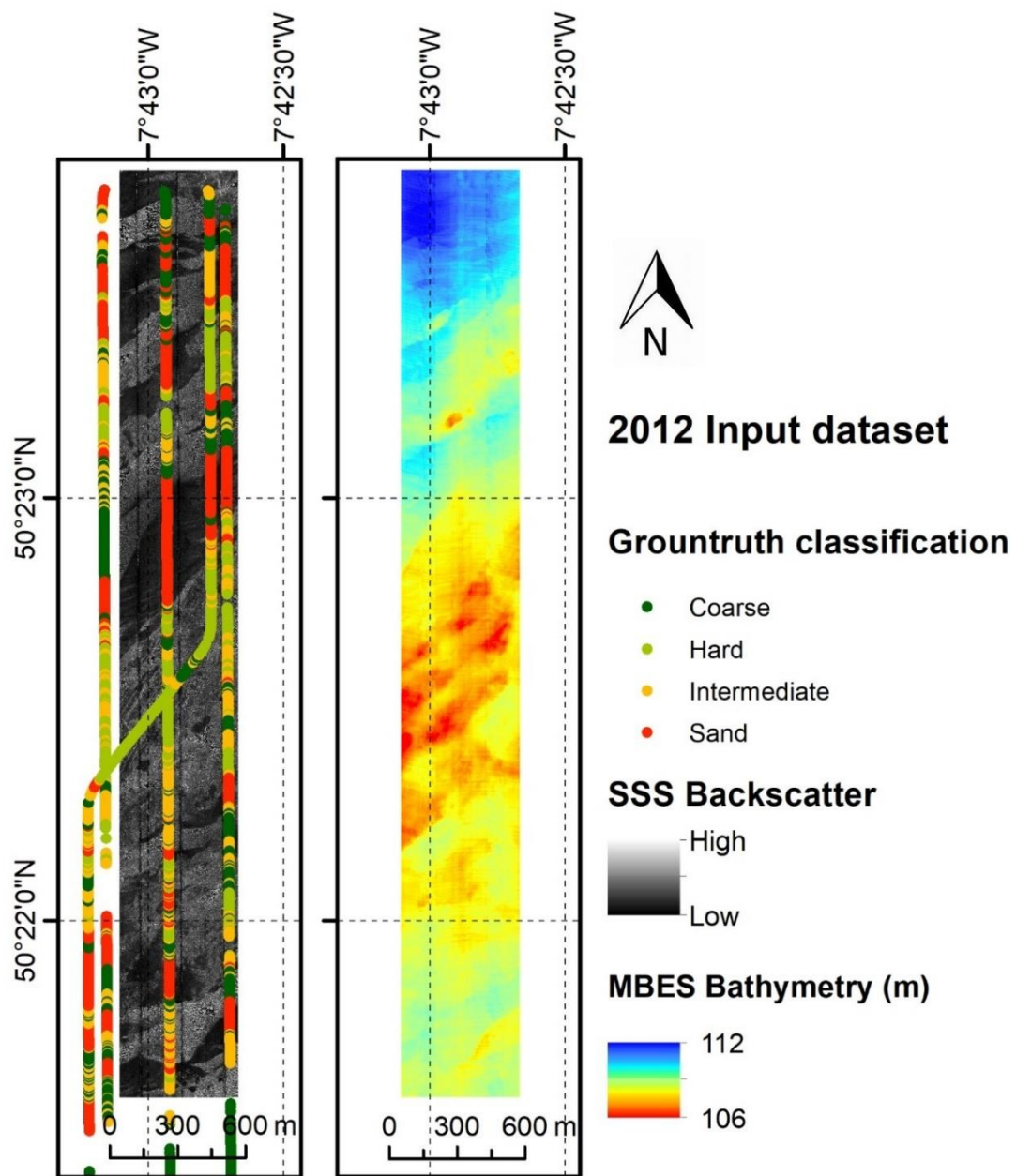
**Figure 1.** (a) General location of study area (+) in the Celtic Sea. (b) Boundaries of the Greater Haig Fras Marine Conservation Zone (GHF-MCZ), partly defined by the United Kingdom exclusive economic zone (UK-EEZ), the Haig Fras Special Area of Conservation (HF-SAC), and the autonomous underwater vehicle (AUV) survey test area.

## 2. Material and Methods

### 2.1. Study Site

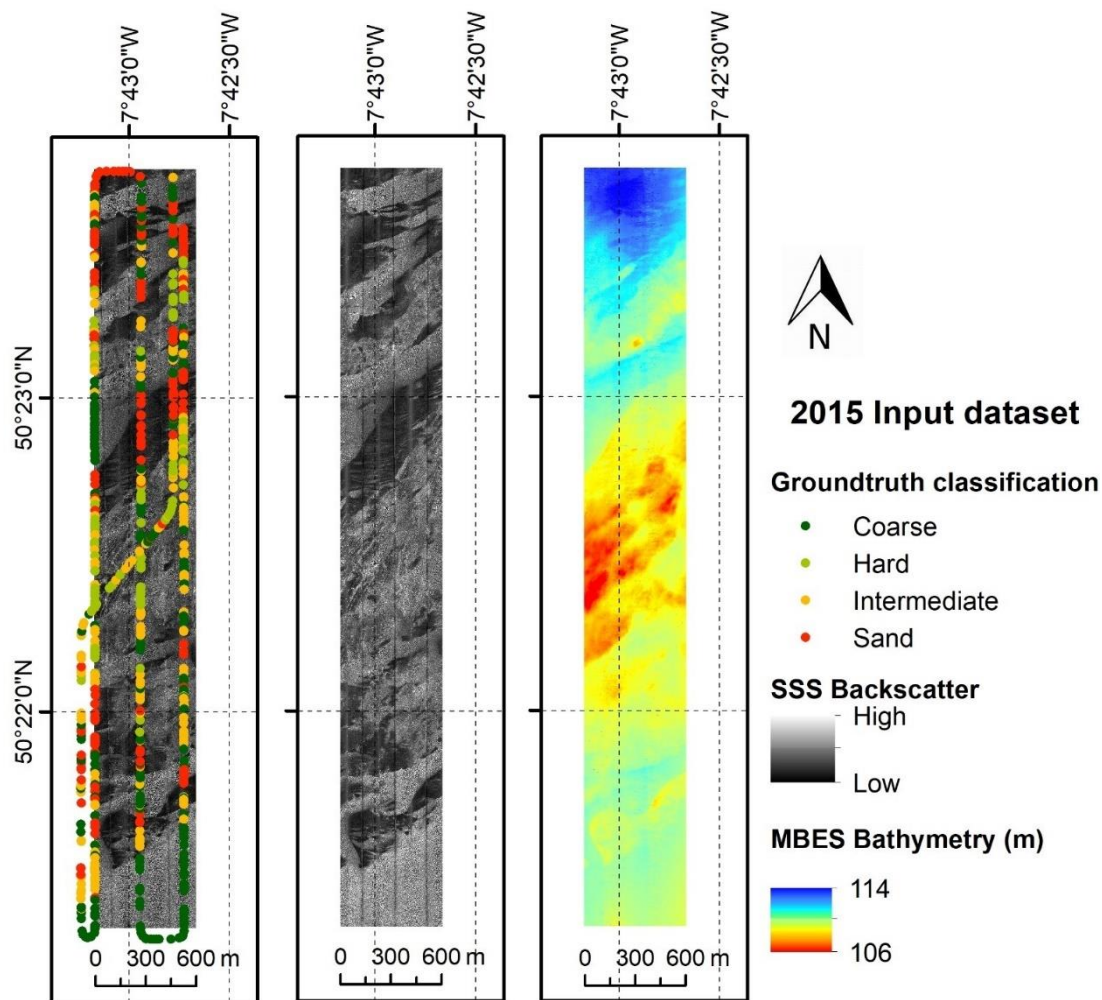
The Greater Haig Fras area was designated a Marine Conservation Zone (GHF-MCZ) in January 2016 (Figure 1). It is located in the Celtic Sea, around 120 km west of Land's End in Cornwall, UK, and covers approximately 2048 km<sup>2</sup>. This MCZ includes the Haig Fras Special Area of Conservation (HF-SAC), a granite outcrop approximately 45 km long and 15 km wide, regarded as the only substantial area of rocky reef in the Celtic Sea (Figure 1). With its western boundary aligned with the UK exclusive economic zone (UK-EEZ), the Greater Haig Fras MCZ varies in depth from 40 m over the rock outcrop, to 118 m depth over the surrounding seabed [24]. These surrounding sediments, varying from coarse to sand and mud, support a variety of marine life including burrowing polychaetes and bivalve molluscs, echinoids, asteroids and epifaunal crustaceans [25].

In July 2012, the NOC Autosub6000 AUV was deployed to collect high-resolution Multibeam Echosounder (MBES) bathymetry, Sidescan Sonar (SSS) backscatter, and seafloor photography data from the Greater Haig Fras area [21,26]. The follow-up surveys conducted over the same area in August 2015 allow evaluation of the amount of benthic habitat change over the three-year time window [24]. The surveyed area, on which this study is based, is approximately 0.7 × 5 km, with a seafloor depth ranging from 106 to 114 m. The area's morphology is primarily dominated by flat sandy and coarse sediments, with some exposed rocky substrata (Figures 2 and 3).



**Figure 2.** Geophysical datasets and groundtruth data points from 2012 survey. Left, AUV sidescan sonar map with classified groundtruth points. Right, AUV multibeam echosounder bathymetry data of the same area. Mercator projection with standard parallel 50. WGS84 datum.





**Figure 3.** Geophysical datasets and groundtruth data points from 2015 survey. Left, AUV sidescan sonar map from Day 1, with classified groundtruth points. Middle, sidescan sonar map from Day 2. Right, shipboard bathymetry data of the same area. Mercator projection with standard parallel 50. WGS84 datum.

## 2.2. Geophysical Data

The first survey was conducted between the 25th and 26th of July 2012, during RRS *Discovery* cruise 377 (D377). All data were collected in a single AUV Autosub6000 deployment (mission 58; Figure 2). Navigation of the Autosub6000 AUV relies on GPS when the vehicle is at the surface, and on a combination of a PHINS Inertial Navigation System and bottom-tracking Doppler Velocity Log (DVL) when close to the seabed. Given the shallow depth of the study area, the AUV could transfer immediately from the GPS to inertial navigation, and drift in the water column during descent was minimal, requiring only small adjustments to the navigation data (see also Section 2.4). MBES data were collected at 50 m altitude, while SSS data were acquired at 15 m altitude. This, together with the higher frequency of the SSS, provided SSS backscatter data that was of far superior resolution and image quality compared to the MBES backscatter. Being collected by AUV over a relatively flat seabed, both datasets had similar positional accuracy. Hence, for the analysis in this study, which is heavily based on backscatter texture analysis, only SSS backscatter was used, together with the MBES bathymetry. Further information on acquisition and post-processing parameters can be found in the corresponding cruise report [26] and is summarized in Tables 1 and 2.

**Table 1.** Sidescan sonar backscatter acquisition parameters.

Acquisition Dates	Platform	System	Operating Frequency (kHz)	Survey Altitude (m)	Processing Software	Output
25–26 July 2012	Autosub6000	EdgeTech 2200-FS	410	15	Chesapeake Sonarwiz6	8-bit greyscale geotif 0.15 m res.
10–12 Aug. 2015	Autosub6000	Edgetech 2200-M	410	15	Chesapeake Sonarwiz6	8-bit greyscale geotif 0.15 m res.

**Table 2.** Multibeam echosounder bathymetry acquisition parameters.

Acquisition Dates	Platform	System	Operating Frequency (kHz)	Survey Altitude (m)	Processing Software	Output
25–26 July 2012	Autosub6000	Kongsberg EM 2000	200	50	CARIS HIPS and SIPS	2 m grid
10–12 Aug. 2015	Autosub6000	Kongsberg EM 2040	200	50	IFREMÉR CARAIBES	3 m grid
10–12 Aug. 2015	RRS <i>James Cook</i>	Kongsberg EM710	70–100	-	CARIS HIPS and SIPS	2 m grid

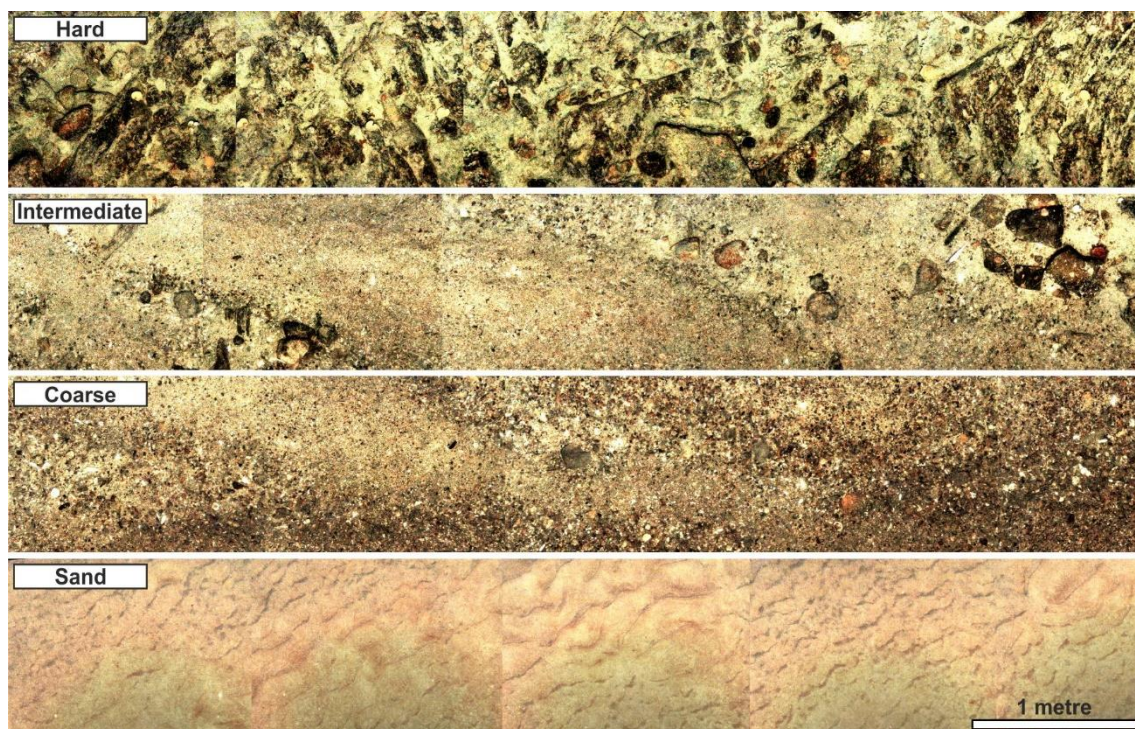
In 2015, the second survey was undertaken between the 10th and 12th of August during RRS *James Cook* cruise 124 (JC124). Swath bathymetry data were collected during AUV Autosub6000 mission 86. Two SSS surveys were also undertaken over the study area, an initial survey during mission 86 and a second survey immediately after, during mission 87, referred to respectively as the ‘Day 1’ and ‘Day 2’ missions (Figure 3). Additionally, shipboard bathymetry data were collected over part of the Greater Haig Fras MCZ, covering the area of study. Further information on acquisition and post-processing parameters can be found in the corresponding cruise report [24] and is summarised in Tables 1 and 2.

### 2.3. Groundtruth Data

Photographic datasets were also acquired on the surveys from D377 and JC124, each using the same survey plan of four parallel lines and one obliquely crossing line. The data were acquired with the AUV Autosub6000 at a target altitude of 3 m, using a FLIR Grasshopper2 camera (mission 58 in 2012 and mission 87 in 2015) (Table 3). General field methodology and subsequent image processing and assessment were as described by Morris et al. [27]. The images were processed, mosaicked in groups of five, and visually classified as per Benoist et al. [28]. As with the geophysical data, a small manual adjustment had to be applied to the groundtruth navigation data, to compensate for the minimal drift of the AUV during its descent. Four habitat types were used to classify the seabed, as described in Figure 4—Hard habitats with hard primary substratum, Intermediate habitats with hard secondary substratum, Coarse habitats, and Sand habitats (see details in [28]). These habitat classes are known to support statistically distinct faunal assemblages.

**Table 3.** Groundtruth data, as the number of image tiles (mosaics of five consecutive images) used to groundtruth the sidescan sonar classifications, separated into training (T) and validation (V) sets by year and survey day [28]. Images to groundtruth the 2012 classification were taken in 2012, all images to groundtruth the 2015 classifications were acquired during one AUV mission in 2015.

Habitat Class	2012		2015 Day 1		2015 Day 2	
	T	V	T	V	T	V
Sand	216	215	80	82	87	87
Coarse	146	145	90	89	95	95
Hard	216	216	86	86	102	102
Intermediate	199	199	56	56	57	58
<b>Total</b>	<b>777</b>	<b>775</b>	<b>312</b>	<b>313</b>	<b>341</b>	<b>342</b>



**Figure 4.** Seabed habitat classes identified in the groundtruth data.

#### 2.4. Data Preparation

The steps taken in the data analysis are summarised in the flowchart in Figure 5. Before assessing the repeatability of the different methods, the bathymetry and backscatter data from each year's survey had to be analysed to check for completeness, correctness, and relevance. This quality control procedure was conducted in ArcGIS v10.4 (Esri). Preparation steps included removal of outliers, infilling of data gaps, setting of projection and resolution (0.5 m) and final clipping of the area of interest. Although the navigational drift of the AUV during its descent was minimal, a slight offset was noted between the acoustic datasets. This was manually corrected, using the 2015 Day 1 SSS mosaic as reference, and georeferencing the other mosaics to this baseline using exactly two tie-points for each mosaic (i.e., allowing only translation and rotation but maintaining internal coherence). As tie-points clearly identifiable rock features were used at the two extremes of the study area.

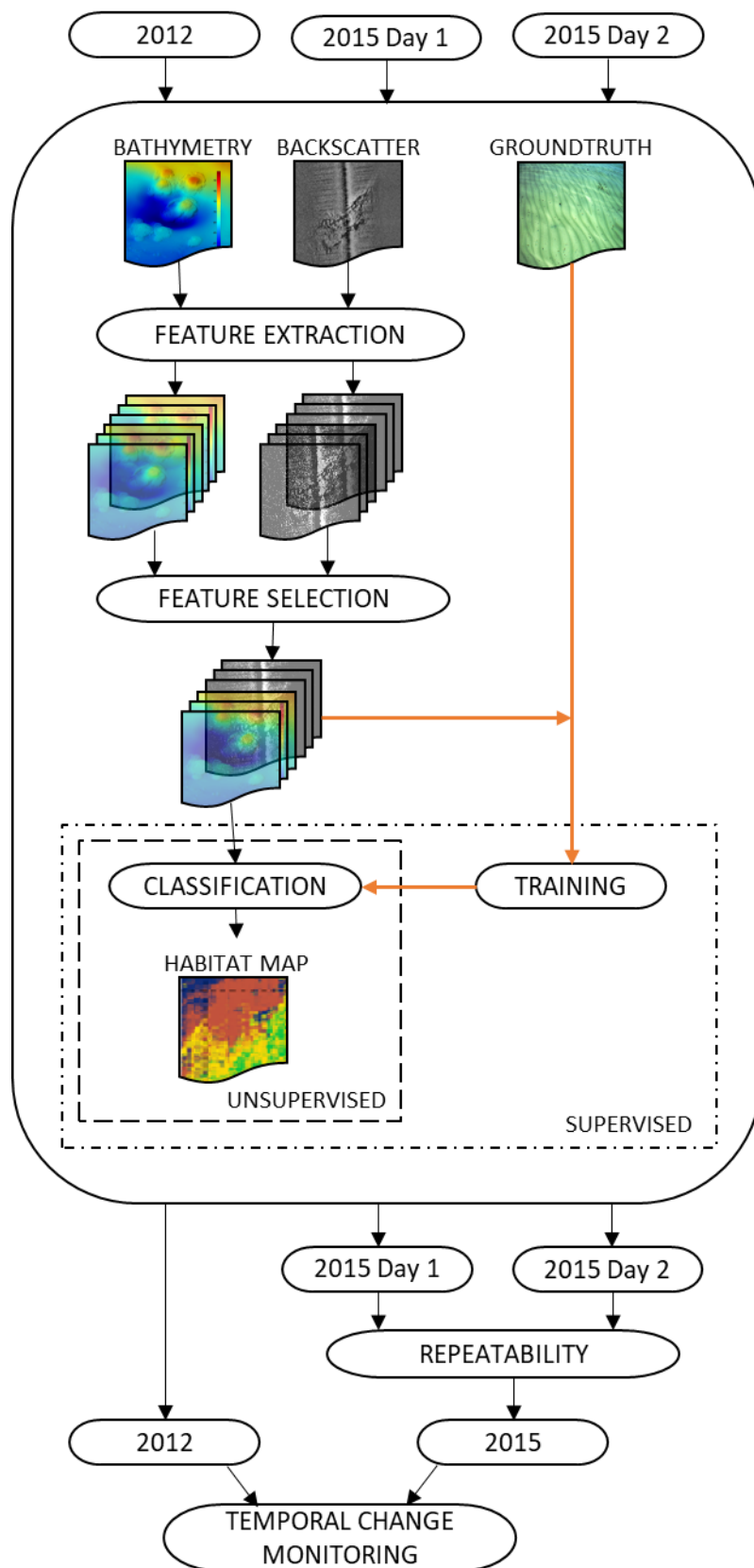


Figure 5. Flowchart summarising the automated seafloor classification and repeatability assessment.



## 2.5. Feature Extraction

### 2.5.1. Bathymetry Derivatives: Multiscale Terrain Analysis

Secondary features were derived from the primary bathymetric datasets to contextualise each pixel within its neighbourhood. Ship-borne MBES bathymetry was used in this study for the 2015 mapping to deal with an observed bias from erroneous vehicle depth recorded in the AUV bathymetry. Different features were calculated at multiple spatial scales, using ArcGIS v10.4 (Esri Inc.) and the Benthic Terrain Modeler Toolbox [29,30]. An initial list of terrain features, comprising depth, aspect (converted into northness and eastness), slope, Bathymetric Position Index (BPI), curvature (planar and profile), and Vector Ruggedness Measure (VRM), was assessed considering their successful application in previous studies [13]. However, the study area is primarily dominated by flat sandy sediments with seafloor depth gently ranging from 106 m to 114 m, and neither BPI, VRM, nor the curvature features showed great variation in the survey area; therefore, they were not used in the modelling.

Northness, eastness and slope, however, were selected together with depth to describe the terrain variability of the study area. They represent terrain variations associated with potential exposure to wave and current energy (northness, eastness), and the likelihood of sediment accumulation (slope), all of which may influence the distribution of biological assemblages.

A multiscale analysis simultaneously incorporated features calculated at different spatial scales aiming to detect variations in the terrain that corresponded with environmental processes and faunal distribution patterns occurring at different scales [13,31,32]. After visually evaluating several cell sizes, this study used four spatial scales to calculate the selected bathymetric derivatives: 2, 5, 10 and 15 m, giving a total of twelve terrain derivatives.

### 2.5.2. Backscatter Derivatives: Multiscale Textural Analysis

This study combined two textural analysis methods: Grey Level Co-occurrence Matrices (GLCM) and Gabor filters (Supplementary Materials: Scripts S1–S3). The Gabor filter bank applied was based on the method proposed by Jain et al. [33], where four orientations: 0°, 45°, 90°, and 135°, and a set of wavelengths starting from  $4\sqrt{2}$ , increasing in powers of two, up to the hypotenuse length of the input image were used to create a filter bank. This combination between four orientations and increasing wavelengths generated a 40-dimensional feature space for each SSS mosaic.

To extract textural information from SSS backscatter using GLCMs, five parameters are needed: the statistical descriptors, number of grey levels, window size, inter-pixel distance, and orientation. The settings of these parameters affect the textural discrimination capabilities [34,35] (Supplementary Materials: Figures S1–S3). The parameterisation of these variables is discussed below.

**GLCM statistical descriptor.** Different statistics can be derived from a GLCM. Haralick et al. [36] proposed fourteen different features, but many are highly correlated [37,38]. Among the least correlated variables, Contrast, Correlation and Entropy combined have proved to yield better results than those obtained using single statistics or using the entire set of statistics [37]. Six texture features derived from GLCM were initially explored in this study: Energy, Contrast, Correlation, Homogeneity, Dissimilarity, and Entropy. Correlation, Contrast and Entropy indeed proved to be the least correlated features. However, after a visual inspection, Entropy yielded similar results when calculated under different scales and orientations and it was considered less valuable given its computational cost (Supplementary Materials: Figure S2). Hence, only two independent features—Contrast and Correlation—were further used in the model. The definitions of these features are given in Table 4.

**Table 4.** Description of GLCM features used in the Textural Analysis, with  $i, j$ : row and column number,  $\delta$ : inter-pixel distance,  $\theta$ : orientation angle,  $\mu, \sigma$ : mean and standard deviation of the row/column values,  $L$ : window size.

Texture Statistic	Description	Calculation Tool
<b>Contrast</b>	Correlation measures the linear dependency of grey levels of neighbouring pixels. $\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (i-j)^2 P(i, j, \delta, \theta)$	MATLAB functions graycomatrix() graycoprops()
<b>Correlation</b>	Contrast measures how regular the pixel value differences within the window are. $\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} \frac{(ij)P(i, j, \delta, \theta) - \mu_i \mu_j}{\sigma_i \sigma_j}$	MATLAB functions graycomatrix() graycoprops()

**Number of Grey Levels.** The selection of the appropriate number of grey levels impacts on the accuracy of the results as it is related to how much information is available for the separation of the textures. A comparison of image histograms with different numbers of grey levels was undertaken, and the optimum number of grey levels required to capture different textures in the images, while remaining computationally efficient, was thirty-two (32) (Supplementary Materials: Figure S1). This is consistent with several other studies that have found that 32 grey levels are a good compromise between texture detection and computation time [39,40].

**Window Size.** The choice of the window size used to extract the texture features is driven by the scale of the textures of interest. A large window can capture broader textures but may poorly distinguish the boundaries between them. A smaller window will preserve the boundaries, but at the expense of potentially failing to capture textures. There is also a computational cost to large windows. A compromise must be found between those parameters. The study aimed to assess multi-scale textural analysis, and trials on a small test area determined that window sizes of 11, 21, 51 pixels were a good compromise between texture capture ability and computation time (Supplementary Material: Figure S3).

**Inter-pixel distance.** Within each window size, different textures can be captured depending on the separation, or inter-pixel distance, between pixels to be compared. Considering the computational cost of calculating several offsets, this study used a discrete set of inter-pixel distances, starting from 5 up to a maximum of half the window size used, with 5-pixel increments in between.

**Orientation.** As there is a potential for anisotropic textures in SSS mosaics, orientations of  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$  and  $135^\circ$  were used instead of their mean. The use of two statistical descriptors, four orientations, three windows sizes, and a range of inter-pixel distances generated a 64-dimensional feature space for each SSS image, as summarised in Table 5.

**Table 5.** Summary of parameters used in GLCM analysis.

Grey Levels	Window Size	Interpixel Distances	Orientations	Static Descriptors	Total Features
32	11	5	$0^\circ, 45^\circ, 90^\circ, 135^\circ$	Contrast, Correlation	8
32	21	5, 10	$0^\circ, 45^\circ, 90^\circ, 135^\circ$	Contrast, Correlation	16
32	51	5, 10, 15, 20, 25	$0^\circ, 45^\circ, 90^\circ, 135^\circ$	Contrast, Correlation	40
Total					64

## 2.6. Feature Reduction

In this study, Principal Component Analysis (PCA) was applied separately to the two previously generated feature subsets, the bathymetric derivatives and the textural features, producing two PCA subsets. As a result of the differences in units and variances between features, all variables were standardised to have zero-mean and unit-variance prior to application of PCA. The eigenvectors of

the PCA indicated the direction and the eigenvalues explained the variance of the data along the new axes. Following Ismail et al. [32], principal components (PCs) with eigenvalues greater than one were retained, as they explain more variance than the original variables, in this way achieving a reduction from the original number of features.

The final feature dataset included: original bathymetry and SSS backscatter, spatial information in the form of raster column and row number, PCs from bathymetric derivatives and PCs from backscatter textural analysis, a total of 27 features to be used by each classification method.

### 2.7. Classification

Three different classification methods were applied: (1) supervised classification using the Random Forest (RF) algorithm [41]; (2) supervised classification using the k-nearest neighbours (KNN) algorithm [42]; and (3) unsupervised classification using k-means (KMEANS) clustering [43]. They were selected considering their frequent use in the study of automated habitat maps, attempting to cover both supervised and unsupervised methods, from simple to more comprehensive methods.

Given the relative simplicity of KMEANS, and its broad use in habitat mapping studies (e.g., [15,32,44,45]), the unsupervised classification was performed using the 'kmeans' function available in MATLAB, using the Euclidean distance. Using a supervised–unsupervised hybrid approach in this study, the cluster number was set to four as suggested by the groundtruth photographic observations.

Among the supervised classification algorithms, KNN is the simplest and quickest to implement (e.g., [12,46–48]). The KNN algorithm used was the Machine Learning App inbuilt in MATLAB. The training data were scaled so that they had zero mean and unit variance. This was necessary such that each feature was considered equally important by the algorithm. MATLAB's Machine Learning App 'Classifier Type' was set to 'Medium', with a default number of neighbours set as 10, as it was considered a balance between model performance and computational cost.

Finally, RF is among the most comprehensive supervised algorithms (e.g., [12,47,49–52]). It was applied using the Machine Learning App inbuilt in MATLAB using the 'Classification Learner' option and selecting a Bagged Trees model type setting (this function invokes RF algorithm [41]), with the number of decision trees set to 300 and the maximum number of predictor variables set to the equivalent of half the total training data points on each split.

### 2.8. Model Assessment

The accuracy of each individual map was assessed by comparing the predicted classification for each model against an independent test set of groundtruth observations, not used in the model training (Table 3). The groundtruth data was split, and fifty percent (50%) of the observations were withheld from the training process to use as validation data in the final model. The subsampling was done using a random selection for each seafloor habitat type (sand; coarse; intermediate; hard) so that the validation set contained the same class frequencies as the training set (Supplementary Materials: Script S4).

As the 2015 survey dataset yielded two maps for consecutive days, one half of the groundtruth data was used as training data for the map of Day 1 and validation data for the map of Day 2 (and vice versa), to ensure independence of the corresponding analyses.

### 2.9. Accuracy Assessment

The model performance for each classification was summarised using a confusion matrix, and several statistical indicators were calculated. The Kappa coefficient [53], also known as Cohen's Kappa, is a measurement of agreement between classification and reference data, while evaluating the possibility of the agreement occurring by chance. It can range from -1 to 1, where 1 means complete agreement, and negative values represent accuracies worse than by chance. For most purposes, values greater than 0.75 represent excellent agreement beyond chance, values below 0.4 may be taken to represent poor agreement beyond chance, any values in between may be taken to represent fair or

good agreement beyond chance [54]. The interpretation of the coefficient has proved difficult, and some studies suggest using disagreement instead [55]. Nevertheless, motivated by its popularity in similar studies, the Kappa coefficient was used in combination with other statistics such as Overall Accuracy (OA), Producer's Accuracy (PA), User's Accuracy (UA), Balanced Error Rate (BER) and No Information Rate (NIR). OA indicates the total proportion of correctly classified data among the total number of points, PA indicates the probability of a real feature on the ground being correctly classified on the map, UA indicates the probability that the class on the map will actually be presented on the ground, BER denotes the average of the proportion of wrong classifications in each class and NIR informs the greatest accuracy achievable by always predicting the majority class label and it is used as a baseline to compare the OA.

The significance of the difference in accuracy between seafloor classifications from different algorithms for the same days was tested using a McNemar test [56] for non-independent classifications as suggested by Foody [57].

#### 2.10. Repeatability Assessment

To assess the repeatability of each algorithm, datasets from consecutive survey days from 2015 were processed using the same methodology, obtaining two predicted habitat maps per algorithm. The 2015 Day 1 predicted map was set as baseline. A pixel-by-pixel comparison of both predictions was undertaken, and confusion matrices were prepared. From these matrices, a set of statistical parameters were calculated and tabulated: Overall Agreement (i.e., Overall Accuracy), Kappa coefficient and BER, and to assess the repeatability per class, PA and UA were also determined.

As consecutive day classification maps were obtained from independent datasets, with independent input acoustic data and independent groundtruth points, the significance of the difference in accuracy between two maps was tested using a Z-test for independent Kappa coefficients as suggested by Foody [57]. The best automated method was defined as the one with the higher agreement in accuracy, and the higher Overall Agreement and Kappa coefficient between consecutive days. This method was used in the habitat change assessment step.

#### 2.11. Habitat Change Assessment

Finally, the best automated scheme identified was applied to the sonar images from 2012 and 2015. Following a modified method, based on that of Rattray et al. [14], this study assessed the habitat changes between years using a pixel-by-pixel comparison of both classification maps, using the predicted habitat map from 2012 as the baseline to compare the changes in 2015.

#### 2.12. Sensitivity Analysis

Sensitivity of the classification accuracies to the input variables was tested by progressively discarding different subsets of features (bathymetry derivatives, all textural features, GLCM features, Gabor features, spatial information, etc.). The statistical significance of the accuracy differences was tested against the original results including all the features, using a McNemar test [56] for non-independent datasets, as described by Foody [57].

### 3. Results

#### 3.1. Feature Selection: PCA

The PCA transformation was conducted over the two previously generated feature subsets: the bathymetric derivatives and the textural features. From the original twelve terrain derivatives, only three terrain PCs with eigenvalues larger than one were retained. Together they explain up to 64% of the total variance from the original features. The coefficients of each PC for both 2012 and 2015 Day 2 datasets are shown in the Supplementary Materials Table S1. They suggest that Terrain\_PC1 is consistently correlated with the derivative eastness, Terrain\_PC2 with slope, and Terrain\_PC3 is



correlated with northness (Table 6). Different spatial scales also exhibited different correlations, with the largest coefficients consistently associated with the 10-m scale and the lowest coefficients associated with the 2-m scale.

**Table 6.** Interpretation of the coefficients of the main Terrain and Texture PCs (see Supplementary Materials Tables S1–S3 for full details).

2012	2015	Interpretation
Terrain_PC1	Terrain_PC1	Eastness
Terrain_PC2	Terrain_PC2	Slope
Terrain_PC3	Terrain_PC3	Northness
Texture_PC1	Texture_PC1	GLCM Contrast at $51 \times 51$ pixel windows
Texture_PC2	Texture_PC2	GLCM Correlation at $51 \times 51$ pixel windows
Texture_PC3	Texture_PC3	Gabor and GLCM features calculated at $0^\circ$ orientation
Texture_PC5	Texture_PC4	Gabor features of short wavelength (up to 91 pixels)
Texture_PC4	-	Gabor features of long wavelengths (362, 724 and 1448 pixels)
-	Texture_PC5	GLCM Correlation at $135^\circ$ orientation

In the same way, from the original 40 Gabor features and the 64 GLCM features, only 19 textural PCs with eigenvalues larger than one were retained. They explain up to 76% of the total variance from the original textural derivatives. The analysis of the textural PC coefficients in a matrix  $104 \times 19$  is less straightforward than for the bathymetric PCs, so only the first 5 PCs are presented for illustration purposes (Supplementary Materials: Tables S2 and S3, see also interpretation in Table 6). These coefficients suggest that Texture\_PC1 is mostly correlated with features associated with the GLCM Contrast statistic on a sliding window of  $51 \times 51$  pixels. This response is consistent in both 2012 and 2015. Similarly, Texture\_PC2 in both years is correlated with the GLCM Correlation statistic on a sliding window of  $51 \times 51$  pixels. The third component, Texture\_PC3, seems to be mostly associated with both Gabor and GLCM features calculated at  $0^\circ$  orientation, for both years. The Texture\_PC4 differs for both years, and in 2012 seems to be mostly correlated with Gabor features of long wavelength (362, 724 and 1448 pixels). Conversely, Texture\_PC5 from 2012 and the Texture\_PC4 from 2015 seem correlated with Gabor features of short wavelengths (up to 91 pixels). Finally, 2015 Texture\_PC5 appears mostly correlated with the GLCM Correlation statistics at  $135^\circ$  orientation (Table 6).

### 3.2. Model Comparison

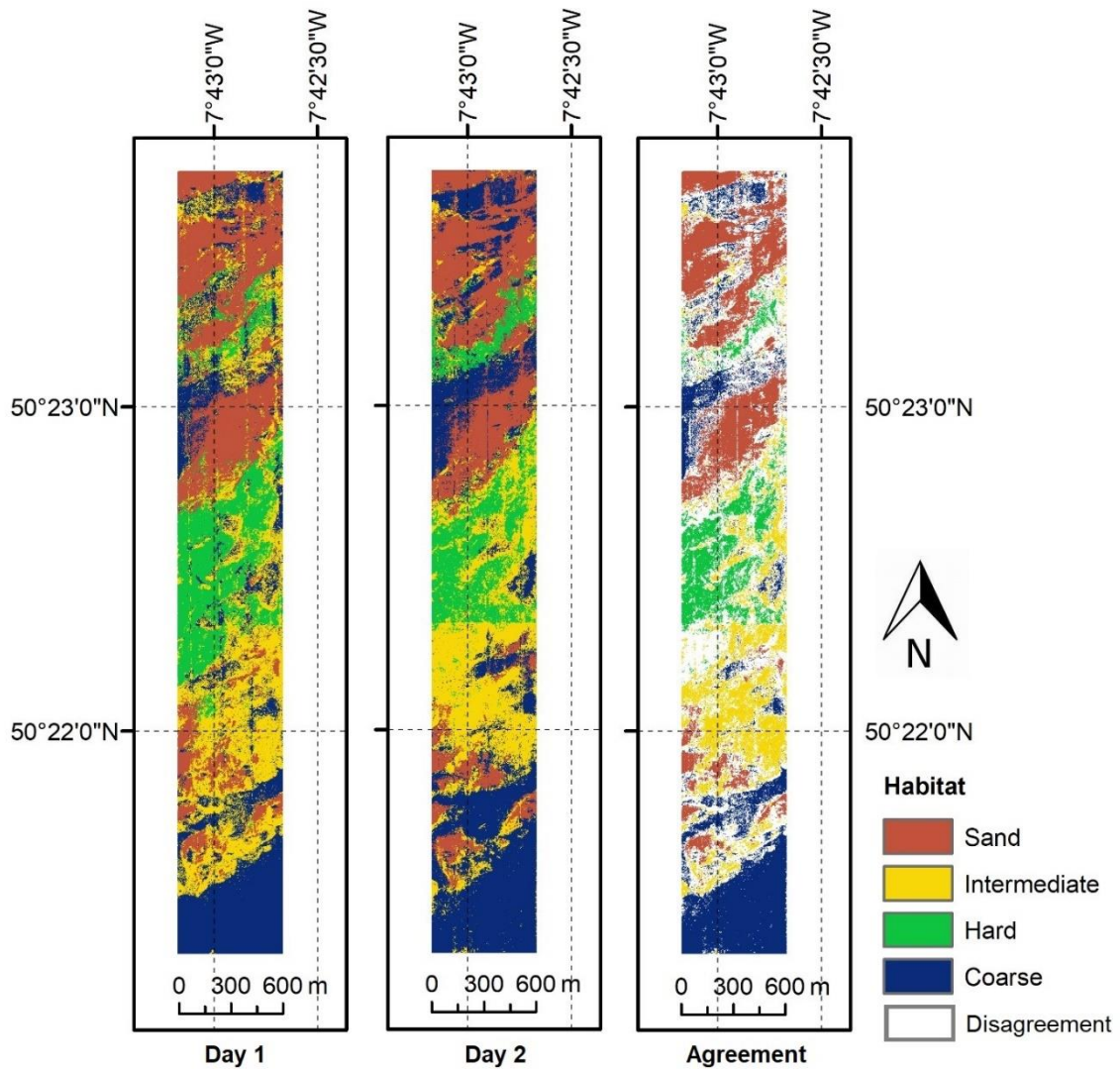
#### 3.2.1. Accuracy of Individual Models

Three different automated habitat mapping solutions were considered in this study, and for each technique, two maps were obtained, 2015 Day 1 and 2015 Day 2 (Figures 6–8; See confusion matrices in Supplementary Materials: Tables S4–S6). The OA of each of these maps varied from 31.7% to 63.7%, with the highest accuracy achieved by RF, followed by KNN and KMEANS. Although accuracies per model varied from Day 1 to Day 2, the ranking of accuracy was consistent in both days and across all the performance indicators including Kappa coefficient and BER (Table 7). A one-tailed McNemar test was used to statistically compare the two better performing classification models [56,57], testing the alternative hypothesis that the RF Kappa coefficient is significantly greater than KNN Kappa coefficient. The test rejected the null hypothesis at the 0.05 significance level each day, indicating that the RF classification algorithm outperformed the other algorithms.

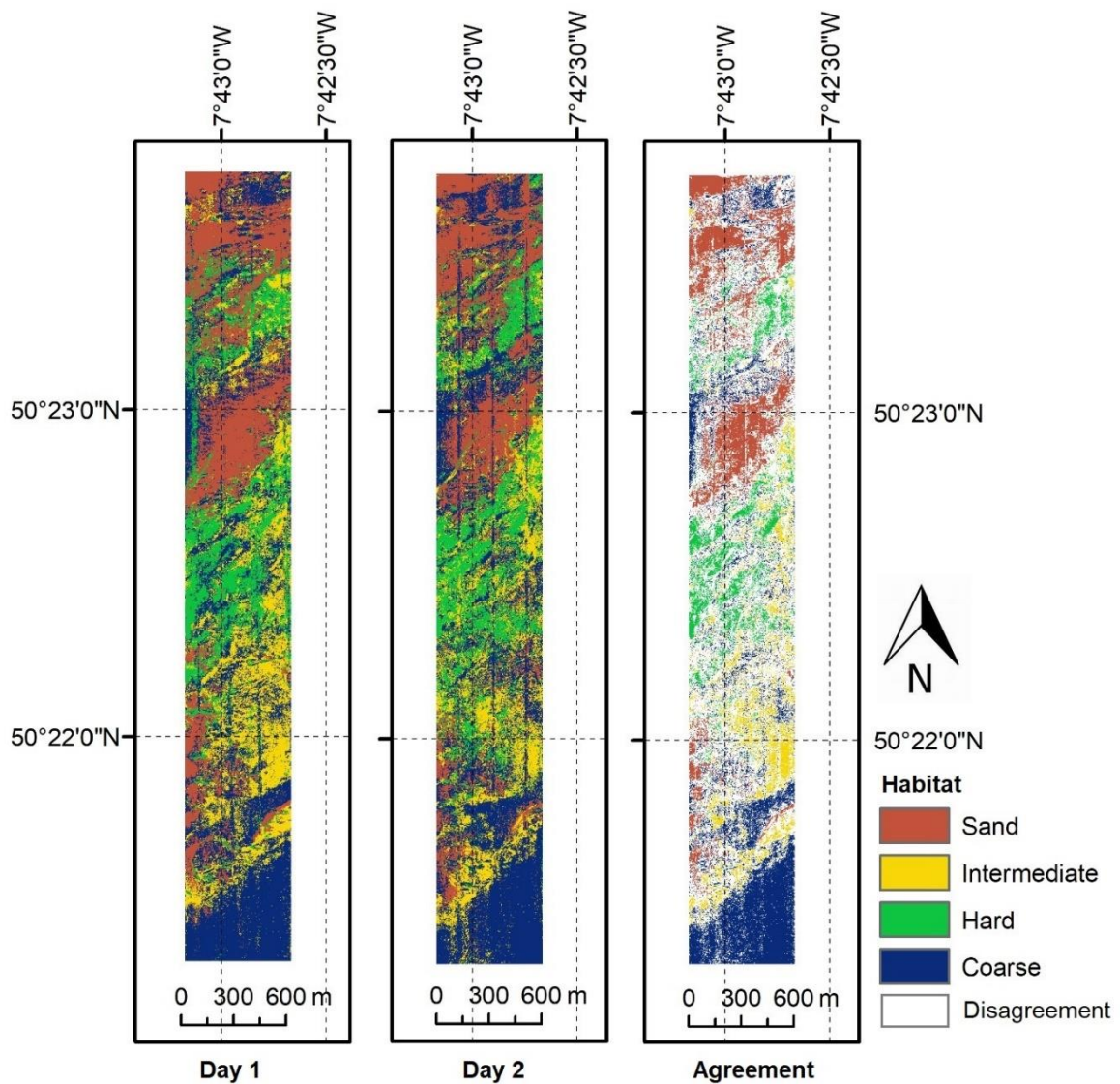
**Table 7.** Performance comparison of Random Forest (RF), k-nearest neighbours (KNN), and k-means (KMEANS) classifications on consecutive days in 2015.

Model	2015 Day 1			2015 Day 2			Comparison	
	Overall Accuracy	Kappa Coeff.	BER	Overall Accuracy	Kappa Coeff.	BER	Agree-ment	Kappa Coeff.
RF	59.40%	0.46	0.4	63.70%	0.51	0.36	60.3%	0.46
KNN	53.70%	0.38	0.47	52.30%	0.35	0.48	47.2%	0.27
KMEANS	45.50%	0.27 *	0.54	31.70%	0.11 *	0.64	53.4%	0.39

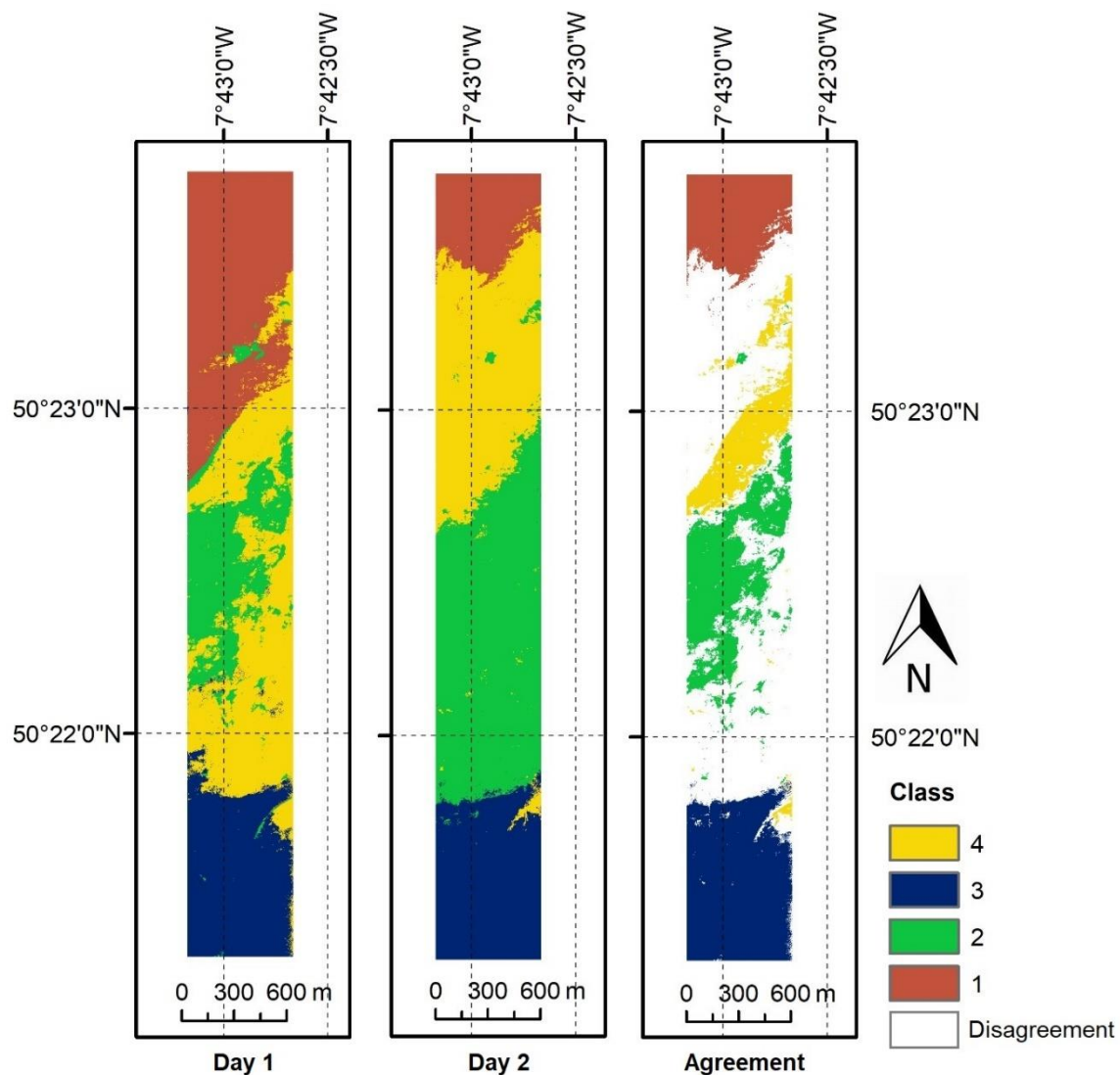
\* Z-test significant difference ( $p < 0.01$ ) Day 1 vs. Day 2 Kappa coefficients.



**Figure 6.** 2015 resulting habitat maps using Random Forest classification algorithm. Left, 2015 Day 1 map. Middle, 2015 Day 2 map. Right, agreement between maps. Mercator projection with standard parallel 50. WGS84 datum.



**Figure 7.** 2015 resulting habitat maps using k-nearest neighbours classification algorithm. Left, 2015 Day 1 map. Middle, 2015 Day 2 map. Right, agreement between maps. Mercator projection with standard parallel 50. WGS84 datum.



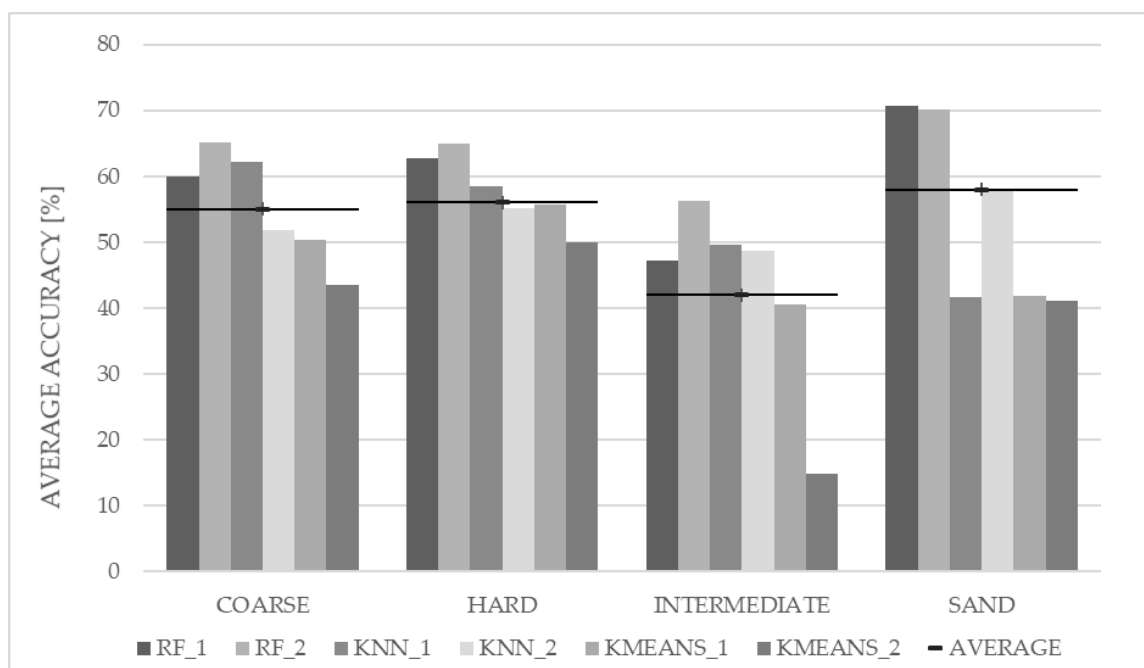
**Figure 8.** 2015 resulting habitat maps using k-means clustering algorithm. Left, 2015 Day 1 map. Middle, 2015 Day 2 map. Right, agreement between maps. Mercator projection with standard parallel 50. WGS84 datum.

### 3.2.2. Per Class Accuracy

Per class accuracy was assessed using the average between UA and PA (Figure 9; See confusion matrices in Supplementary Materials: Tables S4–S6). Averaging across the models the classification of sand obtained the highest accuracy with 58%, followed by hard and coarse with similar accuracies of 56% and 55%, respectively, with the lowest accuracy achieved by the intermediate habitat at 42%. However, these results varied substantially by model, with RF classification accuracies exceeding the average in every case, outperforming the other models.

A closer inspection of the confusion matrix for the RF classified map, shown in Table 8, confirmed that the intermediate habitat achieved the poorest classification accuracy. Some 26.5% (27 out of 102) of the groundtruth points identified as intermediate were confused and classified as coarse habitats. Moreover, the confusion was consistent in the opposite direction, where 14.4% (14 out of 97) and 18.6% (18 out of 97) of the points classified as intermediate were, in fact, coarse or hard habitats.





**Figure 9.** Per class accuracy (mean value between user’s and producer’s accuracy) for each habitat class and the three classification models, for both 2015 Day 1 (\_1) and 2015 Day 2 (\_2) classifications.

**Table 8.** Confusion matrix for Random Forest classification of 2015 Day 2 data. (BER: balanced error rate; NIR: no information rate; values in the body of table indicate the number of validation groundtruth points).

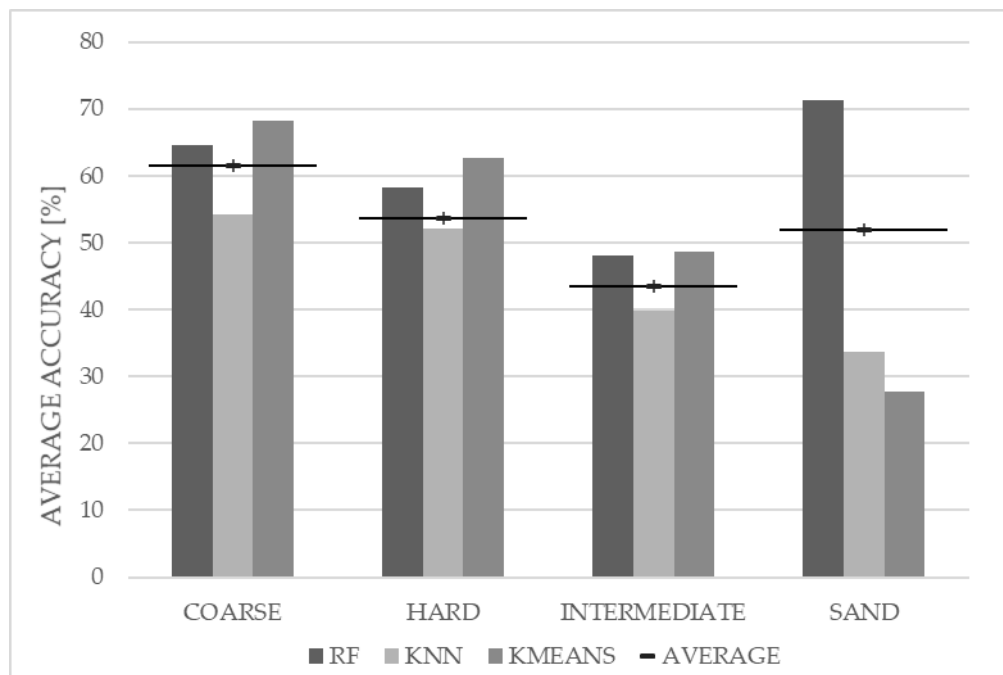
		Groundtruth Class				Total	User’s Accuracy
		Coarse	Hard	Intermediate	Sand		
Predicted Class	Coarse	64	0	27	11	102	62.7%
	Hard	3	36	9	5	53	67.9%
	Intermediate	14	18	56	9	97	57.7%
	Sand	14	4	10	62	90	68.9%
	Total	95	58	102	87	342	
Producer’s Accuracy		67.4%	62.1%	54.9%	71.3%		
Overall Accuracy = 63.7%		NIR = 29.8%		Kappa Coeff. = 0.51		BER = 0.36	

### 3.2.3. Model Repeatability

The maps from consecutive days were compared pixel by pixel to assess the repeatability of the model (See confusion matrices in Supplementary Materials: Tables S4–S6). The highest agreement was achieved by the RF classification algorithm, with 60.30% of the map’s pixels agreeing, and the lowest agreement obtained by the KNN algorithm, with only 47.20% (Table 7). The Z-test illustrated that the maps created using RF and KNN had similar accuracy and were not significantly different, on the contrary, the accuracy of the maps created with KMEANS, despite a reasonable pixel-by-pixel agreement of 53.4%, was significantly different between Day 1 and Day 2 (Table 7). These results suggest that only RF and KNN algorithms delivered repeatable maps.

The agreement per class shows that, on average, the coarse habitat achieved the highest agreement among maps (61.6%), followed by hard and sand habitats with an average agreement around 50%,

and that the lowest agreement occurred in the intermediate class with only 43.5% average agreement. However, these average values are exceeded by the RF algorithm, where the sand class achieved the highest repeatability with 71.3% of agreement (Figure 10). Given that the RF classification algorithm yielded both repeatable and significantly more accurate results than the other models and in particular that the 2015 Day 2 model achieved a better individual accuracy, it was used in the following calculations and comparisons, and it is referred to as the 2015 map.



**Figure 10.** Per class agreement (mean value between user's and producer's accuracy) for each habitat class and the three classification models, comparing the 2015 Day 1 and 2015 Day 2 classifications.

### 3.3. Change Assessment

#### 3.3.1. 2012. Habitat Map

Having identified the RF classification algorithm as both the most accurate and the most repeatable, that method was applied to the 2012 dataset, aiming to assess any temporal change in the survey area benthic habitats. The resultant map for 2012 showed distinctive boundaries between habitats, identifying a central exposed bedrock area and sandy areas in the north of the survey. Coarse sediments seemed more abundant to the very north of the area, with the southern area dominated by intermediate habitat. Validation of the 2012 map yields the confusion matrix shown in Table 9, with an OA of 71.3% and a Kappa coefficient of 0.61, indicating a good agreement beyond chance. Per class accuracies were similar to the 2015 results, with higher accuracies achieved for the hard and sand classes, averaging 81.7% and 80.6% accuracy, respectively, followed by intermediate with 62.4% and coarse with 57.4% average accuracy. Confusion among classes repeated the behaviour of 2015 data with 16% (35 out of 215) and 11% (25 out of 215) of the intermediate groundtruth points classified as coarse and sand, respectively. Similarly, 16% (38 out of 235), 14.5% (34 out of 235), and 10% (23 out of 235) of points classified as intermediate were in fact coarse, hard, and sand, respectively (Table 9).

**Table 9.** Confusion matrix for Random Forest classification of 2012 data. (BER: balanced error rate; NIR: no information rate; values in the body of the table indicate number of validation groundtruth points).

		Groundtruth Class					User's Accuracy
		Coarse	Hard	Intermediate	Sand	Total	
Predicted Class	Coarse	76	4	35	7	122	62.3%
	Hard	7	154	15	3	179	86.0%
	Intermediate	38	34	140	23	235	59.6%
	Sand	24	7	25	182	238	76.5%
	Total	145	199	215	215	774	
Producer's Accuracy		52.4%	77.4%	65.1%	84.7%		
Overall Accuracy = 71.3%		NIR = 27.8%		Kappa Coeff. = 0.61		BER = 0.30	

### 3.3.2. Habitat Transitions

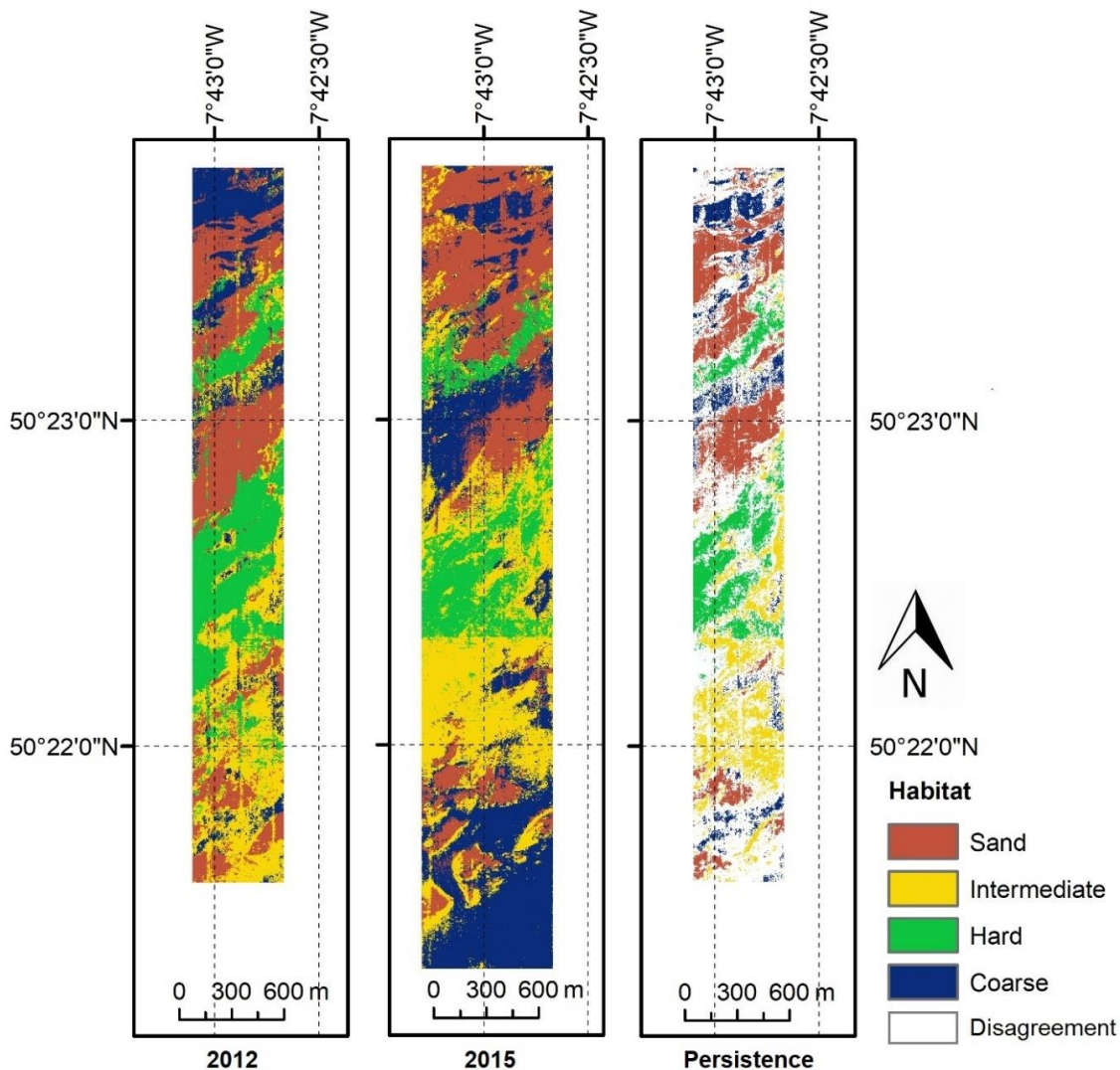
To assess the change in time, the 2015 map was compared against the 2012 map. Figure 11 shows the maps from both years, and a third map depicting habitat persistence between years. The confusion matrix for the pixel-to-pixel comparison between years is given in Table 10, showing an overall habitat persistence of 48.8% in the survey area. The Kappa value is 0.31, indicating a poor agreement between the two maps and suggesting appreciable change between the two years.

**Table 10.** Confusion matrix (percentage of pixels) for comparison between 2012 and 2015 habitat maps. (BER, balanced error rate).

		2012					User's Accuracy
		Coarse	Hard	Intermediate	Sand	Total	
2015	Coarse	7.6	1.0	11.4	5.3	25.3	30.0%
	Hard	0.3	10.8	1.9	0.7	13.7	78.8%
	Intermediate.	1.4	10.9	12.4	7.3	32.0	38.8%
	Sand	6.4	1.7	2.9	18.0	29.0	62.1%
	Total	15.7	24.4	28.6	31.3	100.0	
Producer's accuracy		48.4%	44.3%	43.4%	57.5%		
Persistence = 48.8%		BER = 0.52		Kappa Coeff. = 0.31			

Persistence per class indicates that sand habitat had persisted in 57.5% of the cases, hence exhibiting the least change in time, followed by coarse, hard and intermediate habitat with similar persistence values of around 45%. Habitat change was summarised in terms of gains and losses, allowing to determine both swap and net change proportions (Table 11). These results show that sand and intermediate habitats had the lowest absolute net changes and the highest swapping rates, suggesting that these habitats had simultaneously gained and lost presence in a systematic transition to other classes. Hard habitat exhibited the lowest net change, with a loss of 10.7% in the study area, and the lowest swap rate, suggesting that this habitat was less likely to be systematically swapped by other habitats. Conversely, coarse habitat had the highest net change with an overall gain of 9.6%. Overall, a total change of 51.2% (1 - Overall persistence) was suggested by the pixel-by-pixel comparison, 38.2% resulting from swapping among habitats, and 13% attributable to net change of one habitat to another. The biggest habitat changes appeared to be commonly at the boundaries of large areas of homogeneous

habitat. Nevertheless, some large areas of swapping were also observed, for example, at the northern end of the survey, a large area of sand seemed to have replaced coarse habitat, and conversely, at the southern end of the survey, intermediate habitat appeared to have been replaced by coarse habitat (Figure 11).



**Figure 11.** Habitat change assessment using the Random Forest classification algorithm. Left, 2012 map. Middle, 2015 map. Right, habitat persistence between 2012 and 2015. Mercator projection with standard parallel 50. WGS84 datum.

**Table 11.** Changes in habitat classification between 2012 and 2015, expressed as percentage of the total study area.

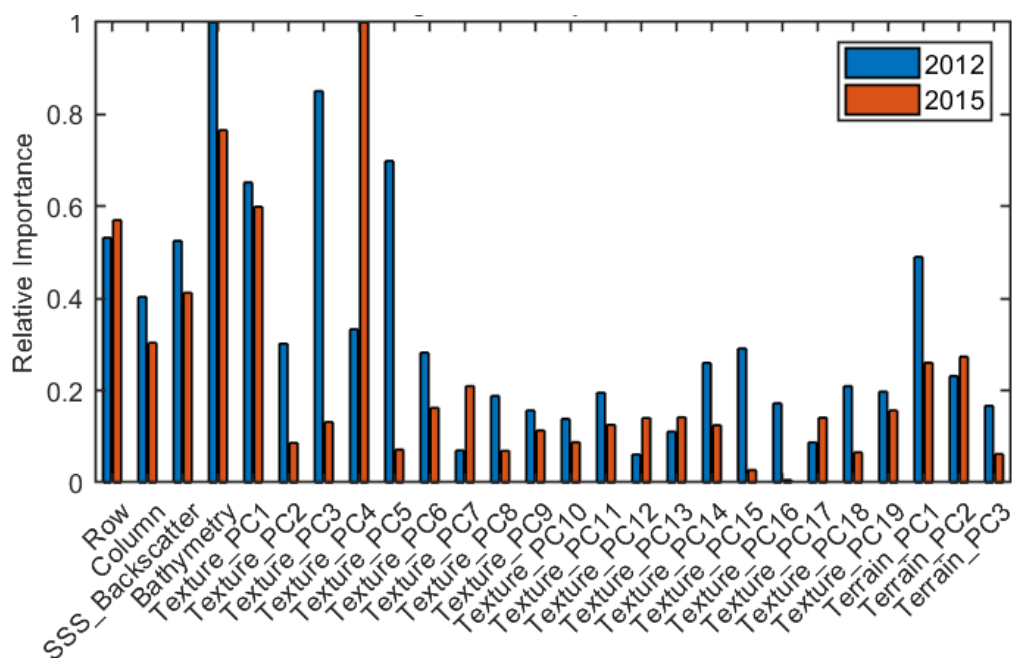
	Gain	Loss	Total Change	Swap	Net Change
<b>Coarse</b>	17.7	8.1	25.8	16.2	9.6
<b>Hard</b>	2.9	13.6	16.5	5.8	−10.7
<b>Intermediate</b>	19.6	16.2	35.8	32.4	3.4
<b>Sand</b>	11.0	13.3	24.3	22.0	−2.3
<b>Total</b>			51.2	38.2	13.0



### 3.4. Model Exploration

#### 3.4.1. Feature Importance

Permutation of the out-of-bag features by the RF algorithm allowed the estimation of feature importance for each model. In both years, spatial information in the form of column and row numbers, plus the original acoustic information (MBES bathymetry, SSS backscatter), proved to be consistently among the most important features. However, the textural PCs and the bathymetric PCs achieved different importance among both years (Figure 12). Eastness (Terrain\_PC1), achieved the highest importance in the 2012 model, while in 2015 both eastness (Terrain\_PC1) and slope (Terrain\_PC2) appear equally important. Examination of the textural PCs showed that in both years GLCM Contrast statistics (Texture\_PC1) and Gabor features at small scales (Texture\_PC4 in 2012 and Texture\_PC5 in 2015) were among the most important features. The 2012 model also considers Gabor and GLCM features calculated at  $0^\circ$  orientation (Texture\_PC3) to be the second most important features of the model.



**Figure 12.** Estimated feature importance in the 2012 and 2015 Random Forest (RF) models. Calculated by out-of-bag permutation of features by RF algorithm.

#### 3.4.2. Sensitivity Analysis

To understand the relevance of different subsets of features in the results, the models from each year were recalculated discarding different subsets of features from the input (Table 12). For both models, 2012 and 2015, the results ignoring the textural information derived from GLCM and the spatial information reported significant differences. Ignoring the Gabor textures, and the terrain derivatives reported slightly less accuracy than using all the features, but the statistical test indicated that these differences were not significant. The sensitivity analysis confirmed that additional information, such as textural statistics or terrain derivatives, reported better accuracy than a model only built with MBES bathymetry and SSS backscatter.

**Table 12.** Sensitivity analysis of model accuracy to the exclusion of different features subsets.

	2012		2015	
	OA	Kappa	OA	Kappa
All	72.0%	0.62	67.8%	0.56
No Terrain PCs	70.8%	0.61	66.1%	0.54
No Gabor PCs	72.7%	0.63	65.8%	0.54
No GLCM PCs	65.9%	0.54 *	60.5%	0.46 *
No spatial context	68.7%	0.58 *	64.6%	0.52 *
No Textural PCs	69.4%	0.59	58.8%	0.44 *
Only Bathymetry and Backscatter	46.3%	0.28 *	41.8%	0.21 *

\* McNemar test statistically significant ( $p < 0.05$ ) from full model ('All').

## 4. Discussion

### 4.1. Comparability of the Results

In this study, we tested the repeatability of three different seafloor classification algorithms: Random Forest (RF) supervised algorithm, k-nearest neighbours (KNN) supervised algorithm, and k-means (KMEANS) unsupervised algorithm. This was done under the assumption that a classification algorithm applied to two acoustic datasets collected on consecutive days should result in identical habitat maps, and that any differences would indicate errors and variability in the methodology. The results suggest that different seabed classification methods varied, not only in the accuracy that they can yield, but more importantly, in the repeatability of their classifications (Table 7). Among the three algorithms tested, RF and KNN proved statistically repeatable, furthermore, RF accuracy (63.7%) and agreement between consecutive days' maps (60.3%) (i.e., repeatability) was the highest among the tested models. This means that RF outperformed both the supervised KNN and the unsupervised KMEANS. This result is consistent with previous studies comparing different automated seafloor classification algorithms [11,12,58]. Moreover, the OA of 71.3% for the 2012 map and 63.7% for the 2015 map, obtained with RF, was also comparable with results reported in similar studies [7,47,59]. Overall, our results highlight that not all seafloor classification algorithms are equally repeatable, and furthermore that the repeatability of these methods should not be assumed.

### 4.2. Inter-Class Confusion

There were differences in the accuracies achieved per class among the three classification models (Figure 9). For example, RF was able to correctly classify sand habitat, consistently across different days' datasets, in over 70% of the cases, well above the 41% median that the other models achieved. Conversely, intermediate habitat not only achieved the lowest classification accuracies, but was confused with coarse and hard habitats (Tables 8 and 9) and appeared to be the least consistently classified habitat (Figure 10). One reason for this confusion might be the variety of the acoustic responses from intermediate habitats. Our definition for the visual assessment of intermediate habitat [28] allows for scenarios ranging from 90% coarse/10% hard to 90% sand/10% hard, passing through all transitional scenarios until hard habitat occupies more than 50% of the imaged area (Table 11). It seems likely that the 'extreme' scenarios would have acoustic signatures more similar to 'pure' coarse and 'pure' sand habitats respectively. Similar confusion in mixed, or mosaic, sedimentary habitats has been observed and discussed in previous studies [18,47,50,59]. The question as to whether marine substrata 'look' (visual classification) and 'sound' (acoustic classification) the same has been addressed previously. For example, Lucieir et al. [47] identified challenges when trying to acoustically distinguish seabed characterised by a sediment veneer on a harder subsurface substratum using groundtruth data from photography. Any seafloor classification exercise carried out for monitoring purposes will have to take this into account and will have to explicitly acknowledge this likely difference between acoustic and visual classifications.

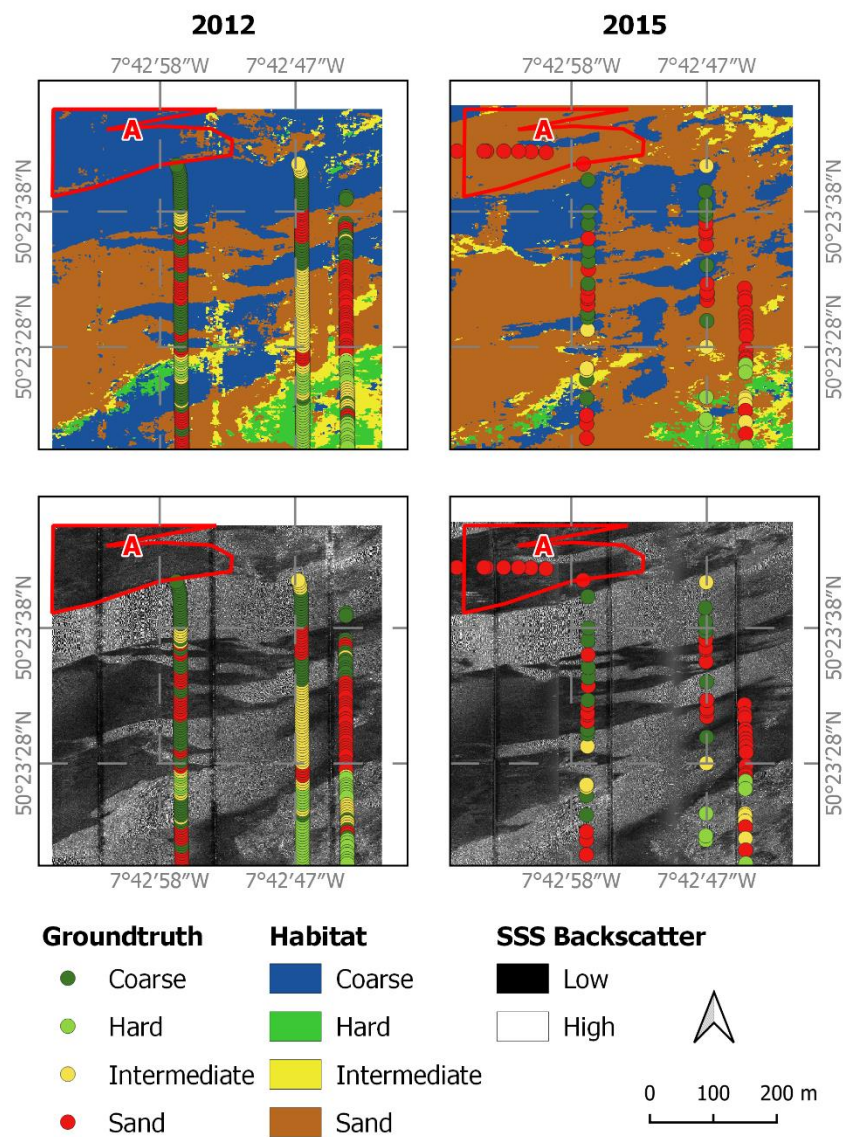
The main aim of this study was to assess the repeatability of several commonly used methods. The optimum selection of classes is a topic of study on its own, and recent studies are proposing novel methods to answer this [60]. In this study, a Top-Down approach or ‘predict first—assemble later’ strategy was used; four habitats were identified from the groundtruth [28]; therefore, four clusters were used in all classification algorithms. From the ecological perspective, this strategy is especially robust when aiming to map species assemblages [61]. However, the choice of habitat classification scheme has been shown to affect the accuracy of the results [62]. Additionally, as these classes were selected without considering the available acoustic data, it may not necessarily have resulted in class signatures that are fully separable in the acoustic features (Supplementary Materials: Figure S5). On the other hand, any other class number determined by their acoustic signature would have lacked meaning compared with the groundtruth’s habitats observed in Benoist et al. [28].

### 4.3. Habitat Transitions

Understanding and constraining inter-class confusion becomes even more important when attempting to assess temporal change in benthic habitats. The results of our 2012 to 2015 comparison indeed indicate a difference between the resultant maps, with a persistence of habitats in only 48.8% of the area. The changes consisted of swapping of habitats in 38.2%, and net change of habitats in 13% of the area (Table 11). Most of the swapping was attributable to intermediate habitat, followed by sand and coarse. This may well be related to the visual classification of intermediate habitat being defined as a combination of sand and/or coarse with hard substratum, making the intermediate class challenging to identify acoustically. However, it may also be related to the mobile nature of both sand and coarse, such that the swapping of habitat could be indicative of sediment movement. Conversely, hard habitat represented the smallest proportion of swapping habitats, but the highest in terms of net change, with 10.7% of the habitat lost, presumably covered by sand or coarse habitats. Similarly, there is photographic (groundtruth) evidence that in 2012 there were more frequent patches of hard substratum in between coarse and intermediate habitats than observed in 2015, when they appeared to have been covered by mobile sediments (Figures 13 and 14). Benoist et al. [28] studied the faunal assemblages in the survey area and found that each habitat supported a statistically distinct assemblage, within which indicator species (taxa) could be identified. Benoist et al. (loc. cit.) noted that hard habitat supported enhanced faunal stocks, while intermediate habitat supported an enhanced faunal diversity. Consequently, the habitat change that we have detected has the potential to modify both faunal standing stock and diversity.

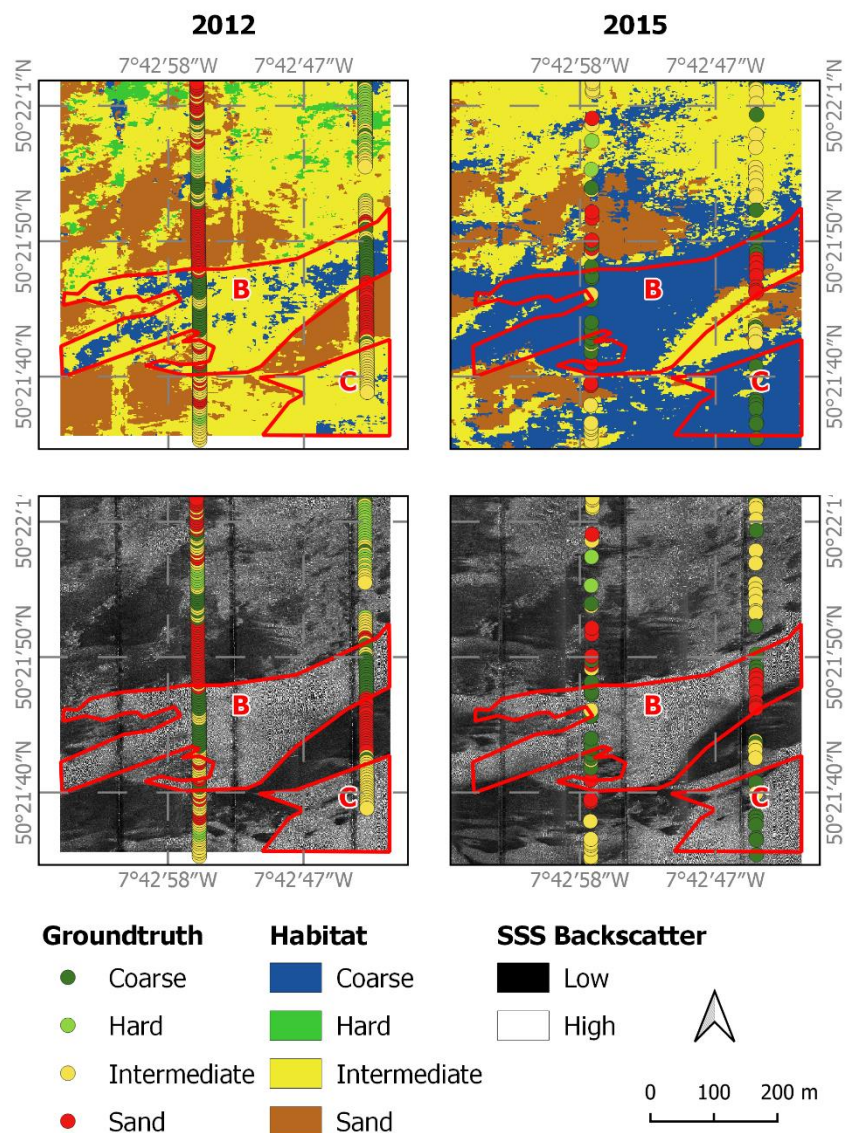
It is important to note that methods of assessing the accuracy of change detection have not been well studied [14,63]. An accepted, but perhaps too general, method involves multiplying the overall classification accuracies for each single classification map [64], which for this study yields a change detection accuracy of 45.5% ( $71.3 \times 63.7\%$ ). However, this accuracy calculation does not consider the repeatability of the method, in the present case estimated at 60.3%. Therefore, if this repeatability were to be properly accounted for, it would be expected that the total change detection accuracy would be lower. Identifying how much of the change between habitats is linked to real habitat transitions and how much is linked to the various potential sources of error in the method is a challenge to be addressed [63,65].

How sensitive these classification algorithms are to the input data is another question that must be resolved before automated habitat mapping can be relied upon as a monitoring tool. Stephens et al. [12] discussed that machine learning approaches might be more susceptible to imperfections in groundtruth data, such as positioning errors. This is on the grounds that these algorithms ‘learn’ from the relationship between groundtruth and feature values at those points, forcing results that are coherent with the information they are given.



**Figure 13.** Example of classification changes between 2012 and 2015 at the northern end of the study area, note variation in delimited region identified as ‘A’. The difference in classification could be partly the result of a lack of groundtruthing points in the 2012 dataset. Mercator projection with standard parallel 50. WGS84 datum.





**Figure 14.** Example of classification changes between 2012 and 2015 at the southern end of the study site, note variation in delimited regions identified as ‘B’ and ‘C’. Groundtruthing confirmed that a large part of the intermediate habitat had been transformed into coarse substratum (area ‘C’), a potential indication of sediment mobility in the area. Mercator projection with standard parallel 50. WGS84 datum.

#### 4.4. Dataset Advantages and Limitations

Despite representing a unique opportunity to test the repeatability of automated habitat mapping approaches, the present dataset has limitations. Possibly the most important source of error in the datasets were the navigational errors in both the SSS and the groundtruth photography. Autosub6000 uses an inertial navigation system, coupled to a 300 kHz Doppler Velocity Log (DVL) able to bottom track within a 200 m range. However, the accuracy of the navigation is of the order of 0.1% of the distance travelled since the last fix, due to drifting [21], which, for an average mission of 15 km length, could accumulate a drift of up to 15 m by the end of the mission. This navigation uncertainty becomes more important for high-resolution studies; in the present case, the pixel resolution was set at  $0.5 \times 0.5$  m in order to retain the richness of the SSS backscatter texture. Although a manual minor shifting of the groundtruth data was done to correct what was considered to be a misalignment of the groundtruth with the acoustic dataset, it is likely that spatially incorrect raster values were

extracted, relative to the groundtruth data, at some locations. This would have been more important in transitions between habitat classes and in areas with high habitat variability, but less relevant in relatively large areas of homogeneous habitat. This observation corresponds with the distribution of areas of agreement and disagreement between maps of consecutive days, with higher disagreement at habitat boundaries (Figure 6).

Furthermore, the input acoustic layers and the groundtruth data were processed at different spatial scales, defined by the characteristics of the equipment used and the survey plan. MBES bathymetry was aggregated at a 2 m resolution, while SSS backscatter was processed at a 0.15 m resolution. During the quality control and the re-processing, it was decided that a maximum 0.5 m resolution was required to preserve the richness of backscatter texture for the forthcoming textural analysis, and both acoustic layers were aggregated to a common 0.5 m resolution. However, photographic images for groundtruthing were mosaicked into elongated tiles (typically order 1.2 × 6.0 m) covering an area of c. 7.3 m<sup>2</sup> of seabed. It is, therefore, expected that the variability in the acoustic layers has a finer resolution than the visual classification method adopted for the groundtruth data.

Some studies have suggested that by collecting a larger amount of groundtruth data, some of the inter-class confusion could be eliminated [59]. Our study relied on an abundant groundtruth dataset with more than 1500 data points in 2012, and almost 700 in 2015. Yet, some areas lacking groundtruth have shown apparent misclassification, for example in the north of the 2012 SSS backscatter mosaic, there is an area of low backscatter intensity visually interpreted as sand but classified as coarse (Figure 13, area A). The acoustic signature in the 2015 mosaic is very similar, but this time there are groundtruth data points across the area, and it is classified as sand. This suggests that simply achieving a large volume of groundtruth data might not be sufficient to ensure greater classification accuracies [66]. Future surveys should also consider the distribution of groundtruth effort in space, and across the required habitat classes, perhaps following geostatistical methods aimed at achieving a stratified and spatially balanced groundtruth dataset [67,68], or applying a form of a posteriori stratified random sampling as the scheme implemented in the ecological analysis of the 2012 groundtruth data [23]. For example, an initial unsupervised classification of the acoustic data could be used to segment the area into potentially important seabed types; groundtruth visual classification effort could then be balanced across these acoustic strata in an attempt to avoid class imbalance [69] in the subsequent supervised classification.

In future studies, one way to account for the positional uncertainties and the scale differences could be to create a buffer around each groundtruth point when extracting the raster information, hence generating a larger training and validation dataset where we expect the signal-to-noise ratio would be higher and the classifier could 'learn' more accurately. Alternatively, one could attempt to measure the sensitivity of the results to purposely misaligned groundtruth data, shifting them in different directions and at different distances and then recalculating the accuracy achieved.

#### 4.5. On the Method: Multiscale Terrain and Textural Analysis

An initial list of terrain features, including BPI, curvature (planar and profile), VRM, eastness, northness and slope was assessed, considering their successful application in previous studies [7,13,51,52,70]. These terrain features are so popular that they have been incorporated in an ArcGIS toolbox named Benthic Terrain Modeler [29,30]. However, the study area is primarily dominated by low-relief sandy/mobile sediments with seafloor depth gently ranging from 106 to 114 m, and only eastness, northness and slope were found to be useful bathymetric derivatives. Although the spatial scale changes relevant to biology are expected to be continuous, these terrain features were calculated over four discrete spatial scales: 2, 5, 10, and 15 m (Table 4). A close inspection of the terrain PC coefficients (Supplementary Materials: Table S1) showed that the features calculated at 2 m were considerably less well correlated with the first three PCs, suggesting that they explained less variation than the features calculated at other scales.

The selection of these particular scales was arbitrary; the 2 m scale aimed to represent the original scale at which the bathymetric data was processed, and the other scales were simple increments of 5 m. We observed that the signal-to-noise ratio increased at broader spatial scales. However, this effect was replaced by 'blurred' results at a 20 m scale and above, hence 15 m was the maximum scale tested (Supplementary Materials: Figure S4). Recent studies have proposed selecting these scales using automated and objective techniques [32,49].

GLCM has mostly been used and thought of as a textural feature with a fixed scale [39,44,58]. However, we employed a way of calculating GLCMs at multiple scales by varying the window size, and, within each window, the inter-pixel distance. Additionally, to capture anisotropic textures in SSS mosaics, we calculated the statistical descriptors in four directions  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$  and  $135^\circ$  rather than using their mean as suggested by Haralick et al. [36]. This technique generated sixty-four different features in a computationally expensive procedure. However, the PCA results suggested that some windows and inter-pixel distance combinations explained more variability than others. For example, Texture\_PC1, in both 2012 and 2015, was mostly correlated with GLCM Contrast calculated at a  $51 \times 51$ -pixel window, and by itself explained 26% and 28% of each year's textural variability. Similarly, Texture\_PC2 was primarily correlated with GLCM Correlation. The use of different orientations also seemed to be important, as Texture\_PC3, in both years, appeared to be particularly correlated with GLCM Correlation at  $0^\circ$  (Table 6 and Supplementary Materials: Tables S2 and S3). The sedimentary bedforms in the area surveyed have a distinct north-south elongation, influencing the outcome of the GLCM results depending on the analysis angle chosen. Between these three first PCs they explain up to 42% and 43% of each year's textural variability.

Other studies in habitat mapping have proposed alternative ways to consider multiple scales in GLCM by using training zones and then determining the optimal combination of grey levels, window size and inter-pixel distance [35,46,71]. In any case, our results support and highlight the importance of considering different scales when undertaking a textural analysis using GLCM.

#### 4.6. Relative Importance of Texture and Terrain Features

In automated seafloor classification, every abiotic feature included will affect the resultant map, and more features do not necessarily result in better classifications [12,52]. We incorporated a feature selection step, by using PCA. Principal Component Analysis is not inherently a feature selection technique, but by keeping only a few PCs that explain the most variance it is possible to retain the most important uncorrelated features. This is an objective approach that has been used in several benthic habitat mapping studies [32,59,72,73]. This step enabled us to reduce one hundred and four (104) textural features to nineteen (19), and twelve (12) bathymetric derivatives to three (3). This constituted an 80% reduction in data size, with a consequent reduction in calculation time. Nevertheless, RF out-of-bag estimation of feature importance did not report all features as equally important, and several textural PCs were among the least relevant features. Alternative feature selection algorithms have been used in similar studies such as the Boruta algorithm [16,74], or the Trimble in-built Feature Space Optimization tool [10].

A sensitivity test was undertaken to understand what subset of features was most relevant in assuring high accuracy of the model. It was motivated by the considerable difference in calculation time for each subset of features (i.e., 60 GLCM features: ~36 h; 40 Gabor features: ~2 h; 12 Terrain derivatives: ~1 h), aiming to study if the time invested was valuable to the prediction results. The analyses showed that maps excluding Gabor features or Terrain derivatives yielded around 2% less accuracy in the predictions, but that this difference was not statistically significant. Conversely, the exclusion of GLCM features or the spatial information was deemed to yield significantly less accurate results (Table 12).

We found that by not including GLCM textural information the accuracy dropped significantly. These results are consistent with the long-established popularity of the method in several seafloor classification studies [35,39,40,46,71,73,75]. Interestingly, Gabor filters were found to be less important, suggesting that not all textural analysis methods are equally suitable for seafloor backscatter analysis.

This finding is in agreement with Karoui et al. [76], who compared GLCM, Gabor filters, and other wavelet methods for seafloor backscatter segmentation, and found that GLCM accounted for more than 90% of the most important features. Similarly, GLCM were found to yield better classification accuracies than Gabor filters in a study of Synthetic Aperture Radar imagery [37]. Nevertheless, it is important to note that within the computational capacities of this study, the calculation of 60 different GLCM features was, by far, the most costly process in terms of time, taking about 36 h to complete, and perhaps more efficient GLCM calculation scripts must be developed.

Despite terrain derivatives appearing not to be significantly important in our results, this should not be directly extrapolated to other studies. Our study area only has low relief, predominantly comprising unconsolidated surficial sediments, having no more than 8 m total range in the bathymetry. For example, results from study areas with more complex seafloor topographies (e.g., reef drops [50], reef peaks [47], presence of large granitic boulders [7]) have included a broader subset of terrain derivatives, several of which have been found to be among the most relevant features. Therefore, while the overall methodology applied by this study can be used in different seafloor topographies—extraction of terrain and SSS backscatter derivatives at different scales, followed by a feature reduction step using PCA—it must be noted that the selection of the terrain derivatives to be used should be adapted based on efficacy and known area characteristics.

In summary, our sensitivity testing work suggested that in an area of low relief, bathymetric derivatives could be excluded from the analysis, but textural features, in particular GLCM features, were the most powerful information in the classification of the benthic habitats.

## 5. Conclusions

This study investigated the repeatability of three automated seafloor classification algorithms and used the most robust one for the detection of changes in an area of the MPA Haig Fras between 2012 and 2015.

Overall, the Random Forest supervised algorithm (RF, OA: 63.7%) outperformed both the supervised k-nearest neighbours (KNN, OA: 52.3%) and the unsupervised k-means (KMEANS, OA: 31.7%) approaches, confirming RF as a robust classification algorithm. Moreover, our results highlighted the importance of integrating multiscale analyses, not only for the bathymetric features but also of textural features. In the low-relief seabed test area, bathymetric derivatives seemed to be non-essential, and conversely textural features were fundamental to the classification of different habitats present. In particular, the use of Grey Level Co-occurrence Matrices (GLCM) was the most appropriate technique to extract textures from the sidescan sonar backscatter data.

Our results suggested that not all algorithms were equally repeatable. Of the algorithms tested, only RF and KNN proved to be statistically repeatable, and the agreement between data collected on consecutive days (i.e., repeatability) was significantly higher with RF (60.3%) than with KNN (47.2%).

Assessing benthic habitat change between 2012 and 2015 using the RF approach, 48.8% habitat persistence was determined, with swapping of habitats driving the change in 38.2% of the area. Most of the swapping occurred in intermediate habitats, followed by sand and coarse. We conclude that this results from the mobility of the seabed sediments. However, assessing the accuracy of these estimates of habitat change remains a challenge, given that current, commonly used indices do not consider the true repeatability of the classification methods.

We highlight the importance of further investigation of the repeatability of seabed classification methods before they can be reliably implemented in the monitoring of benthic habitats. Groundtruth data play a key role in the accuracy of supervised algorithms, and we suggest that procuring a spatially balanced and class-balanced (stratified) distribution of samples with high positional certainty is likely key to improving classification accuracy and repeatability.



**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2072-4292/12/10/1572/s1>, Figure S1: Histogram example of SSS mosaic at different grey levels, Figure S2: Example results of textural analysis with GLCM, using different statistics and different orientations at windows size of 51 pixels and inter-pixel distance of 12 pixels, Figure S3: Example results of textural analysis with GLCM, calculated Correlation at different windows size and inter-pixel distance, Figure S4: Example of terrain derivative slope at different scales, Table S1: Coefficients of the first three Terrain Derivative PCs for each year, Table S2: Coefficients of the first five Gabor Textural PCs for each year, Table S3: Coefficients of the first five GLCM Textural PCs for each year, Figure S5: Box-plot showing habitat discrimination of 2012 and 2015 groundtruth data for SSS backscatter, Bathymetry, three top relevant Textural PCs (Figure 12) and first two Terrain PCs, Table S4: RF algorithm confusion matrices from 2015 Day 1 and Day 2 models, and their pixel-by-pixel comparison, Table S5: KNN algorithm confusion matrices from 2015 Day 1 and Day 2 models, and their pixel-by-pixel comparison, Table S6: KMEANS algorithm confusion matrices from 2015 Day 1 and Day 2 models, and their pixel-by-pixel comparison, Script S1: GLCM Matlab Script, Script S2: GLCM Function, Script S3: Gabor Filter Matlab Script, Script S4: Groundtruth split Matlab script.

**Author Contributions:** Conceptualization, V.A.I.H., B.J.B. and R.B.W.; methodology, A.Z.L. and V.A.I.H.; software, A.Z.L.; validation, A.Z.L., N.M.A.B., and B.J.B.; formal analysis, A.Z.L., N.M.A.B. and M.F.; investigation, A.Z.L.; resources, V.A.I.H. and R.B.W.; data curation, A.Z.L. and N.M.A.B.; writing—original draft preparation, A.Z.L.; writing—review and editing, V.A.I.H., N.M.A.B., M.F. and B.J.B.; visualization, A.Z.L.; supervision, V.A.I.H.; funding acquisition, A.Z.L., V.A.I.H., B.J.B. and R.B.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the NERC Marine Environmental Mapping Programme (MAREMAP), the DEFRA project “Investigating the feasibility of using AUV and Glider technology for mapping and monitoring of the UK MPA network (MB0118)”, and the NERC Climate Linked Atlantic Sector Science (CLASS) project (Grant no: NE/R015953/1); A.Z.L. was funded by the CONICYT PFCHA/MAGISTER BECAS CHILE/2017—73180206 Fellowship program.

**Acknowledgments:** We are grateful to the marine crews and science parties of RRS *Discovery* cruise D377 and RRS *James Cook* cruise JC124, together with the Marine Autonomous and Robotic Systems group at the National Oceanography Centre. Also, our gratitude to T. Le Bas (NOC) for his assistance in the understanding of object-based analysis. We thank the four anonymous reviewers for their constructive comments.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Clarke Murray, C.; Agbayani, S.; Ban, N.C. Cumulative effects of planned industrial development and climate change on marine ecosystems. *Glob. Ecol. Conserv.* **2015**, *4*, 110–116. [[CrossRef](#)]
- Wells, S.; Ray, G.C.; Gjerde, K.M.; White, A.T.; Muthiga, N.; Bezaury Creel, J.E.; Causey, B.D.; McCormick-Ray, J.; Salm, R.; Gubbay, S.; et al. Building the future of MPAs—Lessons from history. *Aquat. Conserv. Mar. Freshw. Ecosyst.* **2016**, *26*, 101–125. [[CrossRef](#)]
- UNEP-WCMC and IUCN Marine Protected Planet. Available online: <https://www.protectedplanet.net> (accessed on 20 January 2020).
- Harris, P.T.; Baker, E.K. Why Map Benthic Habitats? In *Seafloor Geomorphology as Benthic Habitat*; Elsevier: Amsterdam, The Netherlands, 2012; pp. 3–22.
- Brown, C.J.; Smith, S.J.; Lawton, P.; Anderson, J.T. Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. *Estuar. Coast. Shelf Sci.* **2011**, *92*, 502–520. [[CrossRef](#)]
- Lecours, V. On the Use of Maps and Models in Conservation and Resource Management (Warning: Results May Vary). *Front. Mar. Sci.* **2017**, *4*, 1–18. [[CrossRef](#)]
- Ierodiaconou, D.; Alexandre, C.G.S.; Schimel, C.G.; Kennedy, D.; Monk, J.; Gaylard, G.; Young, M.; Diesing, M.; Rattray, A. Combining pixel and object based image analysis of ultra-high resolution multibeam bathymetry and backscatter for habitat mapping in shallow marine waters. *Mar. Geophys. Res.* **2018**, *39*, 271–288. [[CrossRef](#)]
- Innangi, S.; Tonielli, R.; Romagnoli, C.; Budillon, F.; Di Martino, G.; Innangi, M.; Laterza, R.; Le Bas, T.; Lo Iacono, C. Seabed mapping in the Pelagie Islands marine protected area (Sicily Channel, southern Mediterranean) using Remote Sensing Object Based Image Analysis (RSOBIA). *Mar. Geophys. Res.* **2019**, *40*, 333–355. [[CrossRef](#)]

9. Lacharité, M.; Brown, C.J.; Gazzola, V. Multisource multibeam backscatter data: Developing a strategy for the production of benthic habitat maps using semi-automated seafloor classification methods. *Mar. Geophys. Res.* **2018**, *39*, 307–322. [[CrossRef](#)]
10. Lucieer, V.; Lamarche, G. Unsupervised fuzzy classification and object-based image analysis of multibeam data to map deep water substrates, Cook Strait, New Zealand. *Cont. Shelf Res.* **2011**, *31*, 1236–1247. [[CrossRef](#)]
11. Che Hasan, R.; Ierodiaconou, D.; Monk, J. Evaluation of Four Supervised Learning Methods for Benthic Habitat Mapping Using Backscatter from Multi-Beam Sonar. *Remote Sens.* **2012**, *4*, 3427–3443. [[CrossRef](#)]
12. Stephens, D.; Diesing, M. A Comparison of Supervised Classification Methods for the Prediction of Substrate Type Using Multibeam Acoustic and Legacy Grain-Size Data. *PLoS ONE* **2014**, *9*, e93950. [[CrossRef](#)]
13. Wilson, M.F.J.; O'connell, B.; Brown, C.; Guinan, J.C.; Grehan, A.J. Multiscale Terrain Analysis of Multibeam Bathymetry for Habitat Mapping on the Continental Slope. *Mar. Geod.* **2007**, *30*, 3–35. [[CrossRef](#)]
14. Rattray, A.; Ierodiaconou, D.; Monk, J.; Versace, V.L.; Laurenson, L.J.B. Detecting patterns of change in benthic habitats by acoustic remote sensing. *Mar. Ecol. Prog. Ser.* **2013**, *477*, 1–13. [[CrossRef](#)]
15. Snellen, M.; Gaida, T.C.; Koop, L.; Alevizos, E.; Simons, D.G. Performance of Multibeam Echosounder Backscatter-Based Classification for Monitoring Sediment Distributions Using Multitemporal Large-Scale Ocean Data Sets. *IEEE J. Ocean. Eng.* **2018**, 1–14. [[CrossRef](#)]
16. Montereale Gavazzi, G.; Roche, M.; Lurton, X.; Degrendele, K.; Terseleer, N.; Van Lancker, V. Seafloor change detection using multibeam echosounder backscatter: Case study on the Belgian part of the North Sea. *Mar. Geophys. Res.* **2018**, *39*, 229–247. [[CrossRef](#)]
17. Montereale-Gavazzi, G.; Roche, M.; Degrendele, K.; Lurton, X.; Terseleer, N.; Baeye, M.; Francken, F.; Van Lancker, V. Insights into the Short-Term Tidal Variability of Multibeam Backscatter from Field Experiments on Different Seafloor Types. *Geosciences* **2019**, *9*, 34. [[CrossRef](#)]
18. Diesing, M.; Green, S.L.; Stephens, D.; Lark, R.M.; Stewart, H.A.; Dove, D. Mapping seabed sediments: Comparison of manual, geostatistical, object-based image analysis and machine learning approaches. *Cont. Shelf Res.* **2014**, *84*, 107–119. [[CrossRef](#)]
19. Frost, M.; Sanderson, W.G.; Vina-Herbon, C.; Lowe, R.J. *The Potential Use of Mapped Extent and Distribution of Habitats as Indicators of Good Environmental Status (GES)*. Healthy and Biologically Diverse Seas Evidence Group Workshop Report; Joint Nature Conservation Committee: Peterborough, UK, 2013.
20. Anderson, J.T.; Van Holliday, D.; Kloser, R.; Reid, D.G.; Simard, Y. Acoustic seabed classification: Current practice and future directions. *ICES J. Mar. Sci.* **2008**, *65*, 1004–1011. [[CrossRef](#)]
21. 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.; et al. *Investigating the Feasibility of Utilizing AUV and Glider Technology for Mapping and Monitoring of the UK MPA Network*; Final report for Defra project MB0118; National Oceanography Centre: Southampton, UK, 2012.
22. Wynn, R.B.; Huvenne, V.A.I.; Le Bas, T.P.; Murton, B.J.; Connelly, D.P.; Bett, B.J.; Ruhl, H.A.; Morris, K.J.; Peakall, J.; Parsons, D.R.; et al. Autonomous Underwater Vehicles (AUVs): Their past, present and future contributions to the advancement of marine geoscience. *Mar. Geol.* **2014**, *352*, 451–468. [[CrossRef](#)]
23. Jones, D.O.B.; Gates, A.R.; Huvenne, V.A.I.; Phillips, A.B.; Bett, B.J. Autonomous marine environmental monitoring: Application in decommissioned oil fields. *Sci. Total Environ.* **2019**, *668*, 835–853. [[CrossRef](#)]
24. Huvenne, V.A.I.; Wynn, R.B.; Gales, J.A. *RRS James Cook Cruise 124-125-126 09 Aug-12 Sep 2016. CODEMAP2015: Habitat mapping and ROV vibrocorer trials around Whittard Canyon and Haig Fras*; National Oceanography Centre: Southampton, UK, 2016.
25. *The Greater Haig Fras Marine Conservation Zone Designation Order 2016*; Ministerial Order 2016, No. 9; Wildlife Environmental Protection Marine Management: London, UK, 2016.
26. Ruhl, H.A. RRS Discovery Cruise 377 & 378, 05–27 Jul 2012. In *Autonomous Ecological Surveying Of the abyss: Understanding Mesoscale Spatical Heterogeneity at the Porcupine Abyssal Plain*; National Oceanography Centre: Southampton, UK, 2013.
27. Morris, K.J.; Bett, B.J.; Durden, J.M.; Huvenne, V.A.I.; Milligan, R.; Jones, D.O.B.; McPhail, S.; Robert, K.; Bailey, D.M.; Ruhl, H.A. A new method for ecological surveying of the abyss using autonomous underwater vehicle photography. *Limnol. Oceanogr. Methods* **2014**, *12*, 795–809. [[CrossRef](#)]
28. Benoist, N.M.A.; Morris, K.J.; Bett, B.J.; Durden, J.M.; Huvenne, V.A.I.; Le Bas, T.P.; Wynn, R.B.; Ware, S.J.; Ruhl, H.A. Monitoring mosaic biotopes in a marine conservation zone by autonomous underwater vehicle. *Conserv. Biol.* **2019**, *33*, 1174–1186. [[CrossRef](#)] [[PubMed](#)]

29. Lundblad, E.R.; Wright, D.J.; Miller, J.; Larkin, E.M.; Rinehart, R.; Naar, D.F.; Donahue, B.T.; Anderson, S.M.; Battista, T. A Benthic Terrain Classification Scheme for American Samoa. *Mar. Geol.* **2006**, *29*, 89–111. [[CrossRef](#)]
30. Walbridge, S.; Slocum, N.; Pobuda, M.; Wright, D.J. Unified Geomorphological Analysis Workflows with Benthic Terrain Modeler. *Geosciences* **2018**, *8*, 94. [[CrossRef](#)]
31. Misiuk, B.; Lecours, V.; Bell, T. A multiscale approach to mapping seabed sediments. *PLoS ONE* **2018**, *13*, e0193647. [[CrossRef](#)] [[PubMed](#)]
32. Ismail, K.; Huvenne, V.A.I.; Masson, D.G. Objective automated classification technique for marine landscape mapping in submarine canyons. *Mar. Geol.* **2015**, *362*, 17–32. [[CrossRef](#)]
33. Jain, A.K.; Farrokhnia, F. Unsupervised Texture Segmentation Using Gabor Filters. In Proceedings of the 1990 IEEE International Conference On Systems, Man, and Cybernetics Conference Proceedings, Los Angeles, CA, USA, 4–7 November 1990; pp. 14–19.
34. Barber, D.G.; LeDrew, E.F. SAR Sea Ice Discrimination Using Texture Statistics: A Multivariate Approach. *Photogramm. Eng. Remote Sens.* **1991**, *57*, 385–395.
35. Prampolini, M.; Blondel, P.; Foglini, F.; Madricardo, F. Habitat mapping of the Maltese continental shelf using acoustic textures and bathymetric analyses. *Estuar. Coast. Shelf Sci.* **2016**, 1–16. [[CrossRef](#)]
36. Haralick, R.M.; Shanmugam, K.; Dinstein, I. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *3*, 610–621. [[CrossRef](#)]
37. Clausi, D.A. Comparison and fusion of co-occurrence, Gabor and MRF texture features for classification of SAR sea-ice imagery. *Atmos. Ocean* **2001**, *39*, 183–194. [[CrossRef](#)]
38. Ulaby, F.; Kouyate, F.; Brisco, B.; Williams, T.H. Textural Information in SAR Images. *IEEE Trans. Geosci. Remote Sens.* **1986**, *GE-24*, 235–245. [[CrossRef](#)]
39. Blondel, P. Segmentation of the Mid-Atlantic Ridge south of the Azores, based on acoustic classification of TOBI data. In *Tectonic, Magmatic, Hydrothermal and Biological Segmentation of Mid-Ocean Ridges*; MacLeod, C.J., Tyler, P.A., Walker, C.L., Eds.; Geological Society: London, UK, 1996; Volume 118, pp. 17–28.
40. Huvenne, V.A.I.; Blondel, P.; Henriot, J.P. Textural analyses of sidescan sonar imagery from two mound provinces in the Porcupine Seabight. *Mar. Geol.* **2002**, *189*, 323–341. [[CrossRef](#)]
41. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
42. Bremner, D.; Demaine, E.; Erickson, J.; Iacono, J.; Langerman, S.; Morin, P.; Toussaint, G. Output-sensitive algorithms for computing nearest-neighbour decision boundaries. *Discret. Comput. Geom.* **2005**, *33*, 593–604. [[CrossRef](#)]
43. Lloyd, S.P. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [[CrossRef](#)]
44. Huo, G.; Li, Q.; Zhou, Y. Seafloor Segmentation Using Combined Texture Features of Sidescan Sonar Images. In Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 9–12 October 2016; pp. 3794–3799.
45. Alevizos, E.; Snellen, M.; Simons, D.G.; Siemes, K.; Greinert, J. Acoustic discrimination of relatively homogeneous fine sediments using Bayesian classification on MBES data. *Mar. Geol.* **2015**, *370*, 31–42. [[CrossRef](#)]
46. Montereale Gavazzi, G.; Madricardo, F.; Janowski, L.; Kruss, A.; Blondel, P.; Sigovini, M.; Foglini, F. Evaluation of seabed mapping methods for fine-scale classification of extremely shallow benthic habitats—Application to the Venice Lagoon, Italy. *Estuar. Coast. Shelf Sci.* **2016**, *170*, 45–60. [[CrossRef](#)]
47. Lucieer, V.; Hill, N.A.; Barrett, N.S.; Nichol, S. Do marine substrates “look” and “sound” the same? Supervised classification of multibeam acoustic data using autonomous underwater vehicle images. *Estuar. Coast. Shelf Sci.* **2013**, *117*, 94–106. [[CrossRef](#)]
48. Diesing, M.; Stephens, D. A multi-model ensemble approach to seabed mapping. *J. Sea Res.* **2015**, *100*, 62–69. [[CrossRef](#)]
49. Belgiu, M.; Drăguț, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [[CrossRef](#)]
50. Turner, J.A.; Babcock, R.C.; Hovey, R.; Kendrick, G.A. Can single classifiers be as useful as model ensembles to produce benthic seabed substratum maps? *Estuar. Coast. Shelf Sci.* **2018**, *204*, 149–163. [[CrossRef](#)]
51. Duro, D.C.; Franklin, S.E.; Dubé, M.G. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sens. Environ.* **2012**, *118*, 259–272. [[CrossRef](#)]

52. Che Hasan, R.; Ierodiaconou, D.; Laurenson, L.; Schimel, A. Integrating Multibeam Backscatter Angular Response, Mosaic and Bathymetry Data for Benthic Habitat Mapping. *PLoS ONE* **2014**, *9*, e97339. [[CrossRef](#)] [[PubMed](#)]
53. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [[CrossRef](#)]
54. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [[CrossRef](#)] [[PubMed](#)]
55. Pontius, R.G.; Millones, M. Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *Int. J. Remote Sens.* **2011**, *32*, 4407–4429. [[CrossRef](#)]
56. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [[CrossRef](#)]
57. Foody, G.M. Thematic Map Comparison: Evaluating the Statistical Significance of Differences in Classification Accuracy. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 627–633. [[CrossRef](#)]
58. Herkül, K.; Peterson, A.; Paekivi, S. Applying multibeam sonar and mathematical modeling for mapping seabed substrate and biota of offshore shallows. *Estuar. Coast. Shelf Sci.* **2017**, *192*, 57–71. [[CrossRef](#)]
59. Calvert, J.; Strong, J.A.; Service, M.; McGonigle, C.; Quinn, R. An evaluation of supervised and unsupervised classification techniques for marine benthic habitat mapping using multibeam echosounder data. *ICES J. Mar. Sci.* **2015**, *72*, 1498–1513. [[CrossRef](#)]
60. Kågesten, G.; Fiorentino, D.; Baumgartner, F.; Zillén, L. How do continuous high-resolution models of patchy seabed habitats enhance classification schemes? *Geosciences* **2019**, *9*, 237. [[CrossRef](#)]
61. Ferrier, S.; Guisan, A. Spatial modelling of biodiversity at the community level. *J. Appl. Ecol.* **2006**, *43*, 393–404. [[CrossRef](#)]
62. Strong, J.A.; Clements, A.; Lillis, H.; Galparsoro, I.; Bildstein, T.; Pesch, R. A review of the influence of marine habitat classification schemes on mapping studies: Inherent assumptions, influence on end products, and suggestions for future developments. *ICES J. Mar. Sci.* **2019**, *76*, 10–22. [[CrossRef](#)]
63. Foody, G.M. Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sens. Environ.* **2010**, *114*, 2271–2285. [[CrossRef](#)]
64. Coppin, P.; Jonckheere, I.; Nackaerts, K.; Muys, B.; Lambin, E. Digital change detection methods in ecosystem monitoring: A review. *Int. J. Remote Sens.* **2004**, *25*, 1565–1596. [[CrossRef](#)]
65. Foody, G.; Pal, M.; Rocchini, D.; Garzon-Lopez, C.; Bastin, L. The Sensitivity of Mapping Methods to Reference Data Quality: Training Supervised Image Classifications with Imperfect Reference Data. *ISPRS Int. J. Geo Inf.* **2016**, *5*, 199. [[CrossRef](#)]
66. Foody, G.M. Sample size determination for image classification accuracy assessment and comparison. *Int. J. Remote Sens.* **2009**, *30*, 5273–5291. [[CrossRef](#)]
67. Xiao, X.; Gertner, G.; Wang, G.; Anderson, A.B. Optimal sampling scheme for estimation landscape mapping of vegetation cover. *Landsc. Ecol.* **2004**, *20*, 375–387. [[CrossRef](#)]
68. Foster, S.D.; Hosack, G.R.; Hill, N.A.; Barrett, N.S.; Lucieer, V.L. Choosing between strategies for designing surveys: Autonomous underwater vehicles. *Methods Ecol. Evol.* **2014**, *5*, 287–297. [[CrossRef](#)]
69. Luque, A.; Carrasco, A.; Martín, A.; De las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, *91*, 216–231. [[CrossRef](#)]
70. Li, D.; Tang, C.; Xia, C.; Zhang, H. Acoustic mapping and classification of benthic habitat using unsupervised learning in artificial reef water. *Estuar. Coast. Shelf Sci.* **2017**, *185*, 11–21. [[CrossRef](#)]
71. Huvenne, V.A.I.; Huhnerbach, V.; Blondel, P.; Gomez Sichi, O.; Le Bas, T. Detailed Mapping of Shallow-Water Environments Using Image Texture Analysis on Sidescan Sonar and Multibeam Backscatter Imagery. In Proceedings of the 2nd International Conference & Exhibition on Underwater Acoustic Measurements: Technologies & Results, Heraklion, Greece, 25–29 June 2007; pp. 879–886.
72. Hogg, O.T.; Huvenne, V.A.I.; Griffiths, H.J.; Linse, K. On the ecological relevance of landscape mapping and its application in the spatial planning of very large marine protected areas. *Sci. Total Environ.* **2018**, *626*, 384–398. [[CrossRef](#)] [[PubMed](#)]
73. Preston, J. Automated acoustic seabed classification of multibeam images of Stanton Banks. *Appl. Acoust.* **2009**, *70*, 1277–1287. [[CrossRef](#)]
74. Stephens, D.; Diesing, M. Towards Quantitative Spatial Models of Seabed Sediment Composition. *PLoS ONE* **2015**, *10*, e0142502. [[CrossRef](#)] [[PubMed](#)]

75. Lucieer, V.L. Object-oriented classification of sidescan sonar data for mapping benthic marine habitats. *Int. J. Remote Sens.* **2008**, *29*, 905–921. [[CrossRef](#)]
76. Karoui, I.; Fablet, R.; Boucher, J.M.; Augustin, J.M. Seabed segmentation using optimized statistics of sonar textures. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 1621–1631. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).