Centre for Environment Fisheries & Aquaculture Science

Cefas

EMODnet Geology – WP3 Case Study

Exploring the suitability of historic datasets to produce robust quantitative sediment maps

Author: Peter Mitchell

Date in format: August 2021

This publication is available at www.gov.uk/government/publications

www.cefas.co.uk

Executive summary

During recent phases of EMODnet Geology the generation of quantitative sediment maps have been a novel and valuable data product. As opposed to traditional maps that present the distribution marine sediments as categorical or thematic classes, the quantitative sediment maps present the relative proportions of each of the sediment components: mud (grain size $d < 63$ µm), sand (63 µm $\leq d < 2$ mm) and gravel ($d \geq 2$ mm) as continuous predictions ranging from 0%-100%. Previous analysis focussed on the northwest European continental shelf, as this was an area with good spatial coverage of both environmental predictor layers and thousands of high quality groundtruth samples. Having demonstrated the application of this analysis approach and subsequently observed the use of those maps by other scientists and institutions, the intention is to repeat these analyses for other European sea regions. However, the amount and quality of input data, i.e. groundtruth samples and full coverage predictor variables, vary between regions which has the potential to limit the use of the quantitative modelling approaches in some areas. Before applying these analyses more extensively there is a need to understand how differences in groundtruth data quality impact the quality of maps produced.

This study explores some of these issues through three case studies. Each case study made use of the previously published northwest European continental shelf data, manipulating the groundtruth data and then running the same analyses to observe how changes to the input data influence the quality of model outputs. The first case study investigated how many training samples are required to produce accurate products. The second case study explored whether to include historical or less robustly analysed groundtruth samples as training data in the modelling process, and if this changed depending on the spatial scale. The third case study investigated how precisely sediment fraction information should be measured in order for the samples to be included in the models.

As expected the most accurate maps were produced from the largest training dataset of robustly measured sediment samples, however, the three case studies provide other useful insights into the number of samples and minimum data standards that can be applied in future analyses in other regions of Europe. It appears preferable to apply strict minimum data standards for which groundtruth training samples to include in the modelling process rather than seek to include as many samples as possible. Case study two showed that use of the historic and visual assessment samples for modelling reduced the quality of the models' predictions. By comparison, in case study one decreasing sample size resulted in relatively minor decreases in model predictive performance, as models derived from 2500 samples produced predictions that explained 83%-87% of the variance and the thematic maps were >90% as accurate as the models derived from >30,000 samples. It was also the case that including sediment samples with less precise sediment fraction data decreased the predictive performance of the models, further supporting the need for rigorous selection of

training data. Where training samples were manipulated in a controlled manner, by rounding sediment fraction data to be less precise, the map products were robust to minor changes (rounding sediment fraction percentages to integer values) but beyond this, decreases in predictive performance were more visible. Finally, should only thematic class groundtruth samples be available, the third case study presented a method for generating quantitative maps from these data. However, given how this approach negatively impacted the models' predictive performance, it would generally be preferable to discard these samples instead. The recommendations that come from these three case studies will guide the methods applied for other sea regions of Europe to produce new robust quantitative sediment map products.

Summary of recommendations for quantitative sediment map analysis:

- It is important to apply strict minimum standards for the quality and vintage of groundtruth sediment samples which are to be included in the modelling process.
- Increasing the number of model training samples will improve the quality of mapped products, but only if the sediment fraction information were precisely measured.
- Filtering model training data to remove sediment fraction samples that were estimated using visual assessment will likely improve the quality of mapped products.
- The precision of groundtruth sample sediment fractions should at least be measured to the nearest percentage point to produce the most accurate mapped products.
- Sediment samples that are only recorded as thematic classes should be discarded from quantitative sediment mapping, rather than an estimate of the fractions of mud, sand and gravel being applied.
- If the information is available, sample metadata should include a record of the method used to measure sediment fractions, thereby allowing the most appropriate samples to be selected for model testing and training.
- To validate the accuracy of mapped products, a subset of the most recent and reliable samples should be used to as testing data.

Cefas Document Control

Submitted to:	EMODnet Geology
Date submitted:	06/08/2021
Project Manager:	Kerry l'Anson
Report compiled by:	Peter Mitchell
Quality control by:	Simeon Archer-Rand
Approved by and date:	04/08/2021
Version:	V1
Recommended citation for this report:	Mitchell, P.J. (2021). Exploring the suitability of historic datasets to produce robust quantitative sediment maps. Cefas Case Study Report for EMODnet Geology – WP3. 25 pp.

Version control history

Version	Author	Date	Comment
V1	Mitchell, P.	04/08/2021	

Contents

1. Introduction

Seabed sediment maps are foundational datasets for government and the commercial sector to make decisions about how to utilise and manage the marine environment. These maps, which depict the distribution of different substrates at the surface of the seabed, are used in a wide range of applications such as: providing a proxy for the distribution of seabed habitats (Brown et al., 2011; Populus et al., 2017); to indicate the provision of ecosystem services (Diesing et al., 2020; Luisetti et al., 2019); to inform the placement of marine infrastructure (Baker and Harris, 2012); and to support the development of designated protection zones (Ware and Downie, 2019). Through years of work by the EMODnet Geology community, the standardisation of substrate classification schemes and harmonisation of maps across national boundaries have facilitated the creation of full coverage substrate maps around European waters (Kaskela et al., 2019). These maps are continually being updated to finer resolutions and adopting new mapping methods where the data allows.

One such method development has been the creation of quantitative sediment maps (Stephens and Diesing, 2015). Traditional sediment mapping approaches have focussed on predicting the distribution of fixed thematic classes. When modelling sediment substrates these have generally separated the classes based on the sediment grain sizes. The classification schemes adopted within EMODnet Geology have been developed from the Folk triangle (Folk, 1954; Long, 2006), which defines different classes based on the relative proportions of mud (grain size $d < 63$ µm), sand (63 µm $\leq d < 2$ mm) and gravel ($d \geq 2$ mm) within a sample. Classification schemes such as Folk 5, Folk 7 and Folk 16 use different numbers of classes to subdivide this triangle, with the number representing the total number of classes (including one ‘Rock and boulders’ class not captured within the triangle) (see Figure 3 in Kaskela et al., 2019). The focus of quantitative sediment maps has instead been to predict these fractions (mud, sand and gravel) as three separate layers. Each output layer has values ranging from 0-1 (or 0%-100%) with the sum of the three layers for any given location equalling one. This is achieved by applying multiple regression models, such as a machine learning algorithm, to predict a likely response variable from a range of predictor variables. The first study of this kind (Stephens and Diesing, 2015) predicted sediment fractions with reasonable success (independent data showed the two models explained 66% and 71% of the variance) over an extensive area of the northwest European continental shelf (that included areas of seabed within Belgium, Denmark, France, Germany, the Netherlands, Norway, Republic of Ireland and the United Kingdom).

Since this study, creating these kinds of map products has been an active area of development within EMODnet Geology Work Package 3. Through access to new predictor layers and a refinement of the modelling approach, these maps were updated by Mitchell et al. (2019) to both produce predictions at a finer spatial scale and expand the extent of coverage. While still focussing on the northwest European continental shelf these maps were expanded to include areas within the national maritime boundaries of Belgium, Denmark, France, Germany, Netherlands, Norway, Republic of Ireland, Sweden and the

United Kingdom and Channel Islands. In a separate case study, Diesing (2015) also explored the suitability of these quantitative approaches at a fine scales (10 m resolution) using an area of the UK where high resolution multibeam echosounder data were available.

Given the increased granularity of the predicted sediment fraction layers these map products have been utilised as input layers for a number of other studies. These include: modelling infaunal distributions (Cooper et al., 2019) and diversity (Thompson et al., 2021); predicting sediment biogeochemistry (Diesing et al., 2020; Luisetti et al., 2019); and modelling hydrodynamics influence on the seabed (Williams et al., 2019). These maps have other advantages over traditional thematic maps as well. For instance, the prediction sediment fractions layers can be simply converted to any classification scheme and maps of prediction uncertainty can be easily extracted (Mitchell et al., 2019).

Following the success of these previous EMODnet Geology case studies, applying these techniques more extensively across all the partner countries sea-basins would be desirable, thereby creating sediment grainsize fraction maps for wider use. The methodology used to produce the predictions of sediment fractions utilise machine learning models. Machine learning models apply a correlative approach whereby the model is trained using a set of predictor variables, that indicate something about the environment (e.g. depth, topography, current speed etc), and reference data, for which the outcome is known (e.g. groundtruth samples). Therefore, the accuracy of the reference data is important in order to appropriately train the models and thereby produce accurate predictions. Hence the first studies used the northwest European continental shelf as a study area given it has most groundtruth samples available that contained sediment fraction information.

Sediment grab samples have been frequently collected by government, commercial and scientific research organisations for a multitude of reasons (e.g. geological, environmental or navigational purposes). While sediment grain size is frequently recorded, the method by which these are measured varies depending on the organisation and application. The Northeast Atlantic Marine Biological Analytical Quality Control (NMBAQC) scheme have developed a standardised set of best practices for processing sediment particle size data (Mason, 2016). These are based on a combination of sieving and laser diffraction methods. However, sieving or laser diffraction methods are costly, time consuming and, depending on the original purpose, potentially unnecessary. So, for many organisations within Europe, including those who are partners of the NMBAQC, visual assessment by experts remains a common approach. This was evident in the 13 datasets that were combined for Mitchell et al (2019). Certain grain size percentages were rounded to the nearest 10% or 25% (such as 50% sand/50% mud or 25% gravel/75% sand) in unusually high frequency within the nationally compiled datasets. While the metadata did not record how the samples had been measured its likely these round numbers had been estimated by visual assessment rather than analysed using quantifiable laboratory methods. Another common issue of sample data quality relates to the age of samples. Sediment type can change over time, particularly following major storm events, so those countries with large historical datasets may find that not all samples are currently relevant. In addition, historical samples collected prior to the adoption of Global Positioning System may be imprecisely positioned. This locational error

has been observed to reduce machine learning model performance (Graham et al., 2008; Mitchell et al., 2017).

While issues like old and imprecisely positioned samples or visually assessed samples were discarded from the previous analyses, their impact on model performance and map accuracy was not explicitly measured. Even if the samples are less precise, these samples still tell us something about the environment where that sample was taken (e.g. a visual assessment of 50% sand/50% mud may in fact be predominantly sand or mud, but it's definitely not 'Gravel'). So, it's possible that inclusion of these samples in the models may actually improve predictions, particularly if the number of precisely measure sediment samples are limited. For other sea regions where recent and robustly measured samples may be limited or not available at all, it's important to understand how or in what situations the less robust samples should not be discarded.

The dataset previously published in Mitchell et al. (2019) provides an ideal dataset to test some of these approaches. While 45,000 samples were used in that previous analyses, this was only after approximately a third of the samples had been discarded due to age or apparent imprecision. Given the readily available combination of robust predictor variables and a large number of both laboratory-analysed samples and less robust samples this study looked to test some of the typical modelling decisions made to understand the trade-offs of using groundtruth data suspected to contain errors. By simulating different data situations this report presents a series of small case studies that examine different trade-offs for generating quantitative sediment maps from less precise groundtruth samples. The results from these case studies are used to produce a number of recommendations about how to produce better maps from less reliable data in other sea-regions. The intent is for these recommendations to be implemented under subsequent EMODnet-Geology phases.

This report presents three case studies which explore how changes to the groundtruth samples used to train the machine learning models influence the final predictions and map accuracy. These case studies explore:

- How many laboratory-assessed sediment samples are required to produce reliable maps;
- Whether the inclusion of less robust samples can increase map accuracy, and whether this varies at different spatial scales; and
- How precise do samples need to be in order to produce reliable maps?

2. Data and Methods

These case studies use the data previously described in Mitchell et al. (2019) which cover an area of the northwest European continental shelf which includes areas within the national maritime boundaries of Belgium, Denmark, France, Germany, Netherlands, Norway, Republic of Ireland, Sweden and the United Kingdom and Channel Islands (Figure 1).

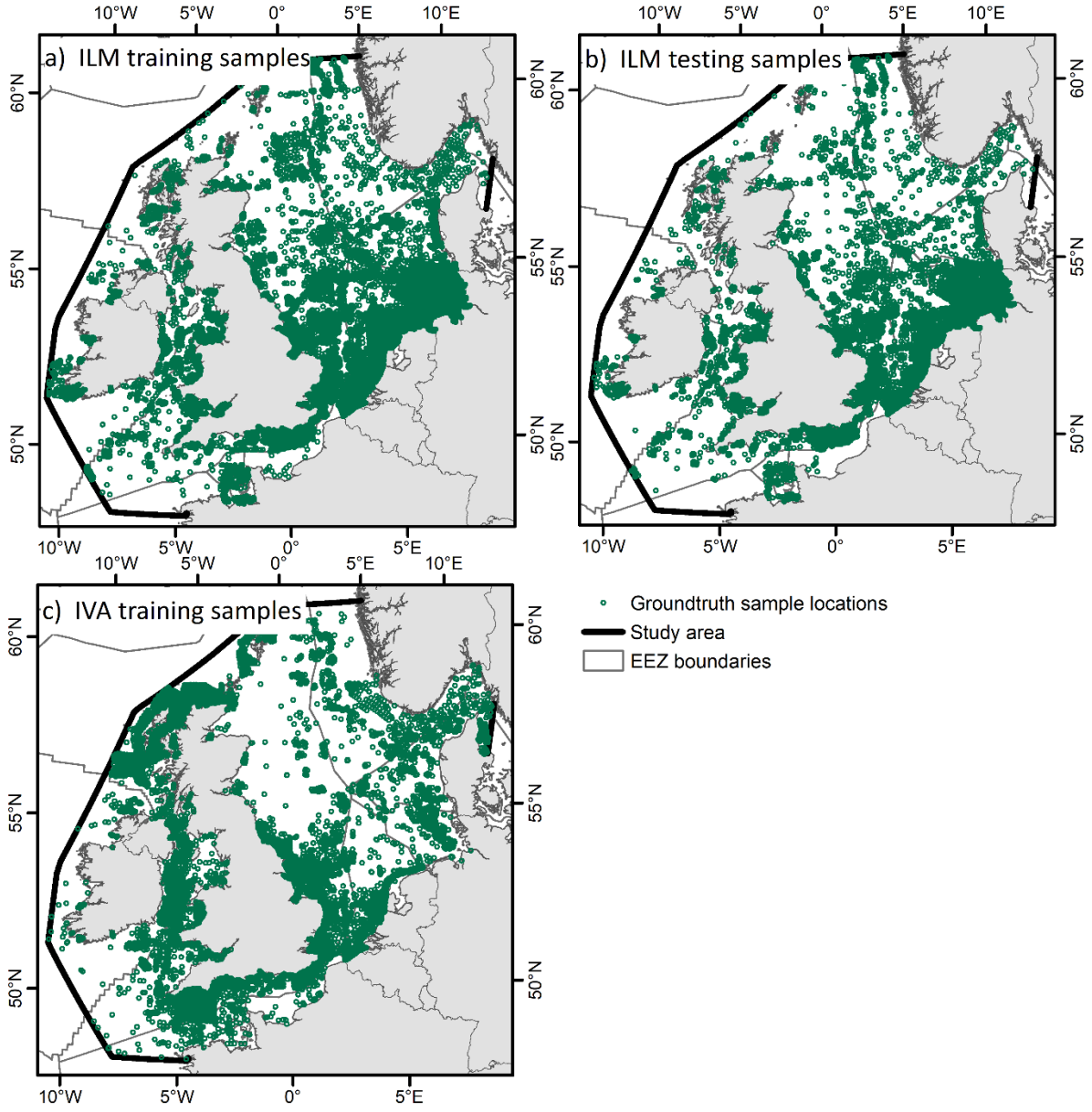


Figure 1. Study area of the northwest European continental shelf and the location of groundtruth samples used in the three case studies. (a) Training and (b) testing samples previously used by Mitchell et al. (2009) were thought to be laboratory measured (Inferred Laboratory Measured). (c) the samples previously discarded by Mitchell et al. (2009) were termed Inferred Visual Assessment and used the second case study only.

Predictive models were trained using the same eight predictor variables as Mitchell et al. (2019). These were: bathymetry, bathymetric position index (calculated at two spatial scales), distance from coast, current speed, orbital velocity of waves at the seabed and suspended inorganic particulate matter (calculated for both the summer and winter). All predictor variables were available in World Geodetic System 1984 with a grid size of 1/8 arc minute * 1/8 arc minute (7.5 arc seconds or approximately 130 m by 230 m).

The sediment sample groundtruth observations were collated from 13 national marine or geological institute databases (Supplement S1; Mitchell et al., 2019). Approximately 68,000 samples were available where particle size composition data had been recorded in the form of percentages of mud, sand and gravel. Samples were initially split into two groups: those used in Mitchell et al. (2019) that were recent and likely to be laboratory analysed; and those that were discarded from the Mitchell el al. (2019) analysis (samples before 1990 or sampled suspected to be based on visual assessment). These are hereon termed Inferred Laboratory Measured (ILM) and Inferred Visual Assessment (IVA) respectively. As the metadata for these samples did not record how particle size composition data had been assessed, the IVA samples were deduced by searching for frequently occurring figures rounded to the nearest five that were common in the data. A subset of the ILM samples were then selected as a testing dataset to be withheld from model training in all the case studies (Table 1). This testing dataset used the same samples that were originally withheld in Mitchell et al. (2019). Where multiple groundtruth samples were available within a single unit of analysis (i.e. a predictor variable pixel) the average percentages of mud, sand and gravel were calculated to produce one set of fractions that was considered representative of that pixel (Table 1).

Table 1. Number of groundtruth samples available.

	No. raw samples	No. available at 7.5 arc second resolution
IVA (training)	12,681	12,114
ILM (training)	36,832	30,679
ILM (testing)	18,754	15,281
Total	68,267	58,275

Compositional data of both the training and testing data were converted to two additive log-ratios (*alr*) (Aitchison, 1986) that could then be independently modelled separately, with any zeros in the mud sand or gravel fractions converted to the lowest observed fraction (0.01) within the data (as per, Lark et al., 2012). As per Mitchell et al. (2019) the gravel fraction was used as the denominator using the formulae:

$$alr_m = \log\left(\frac{mud}{gravel}\right) = \log(mud) - \log(gravel) \qquad (1)$$

$$alr_s = \log\left(\frac{sand}{gravel}\right) = \log(sand) - \log(gravel) \quad (2)$$

The random forest prediction algorithm (Breiman, 2001) was selected for all analyses using the randomForest package (Liaw and Wiener, 2018) in *R* (R Development Team, 2019). All forests were tuned with 500 trees and all other model parameters kept as default. Separate models were fitted for the two response variables (alr_m and alr_s) and the models' quality was assessed using the 'variance explained' and mean of the squared prediction error (MSE), measured against the test dataset observations that had been withheld for all models.

The two additive log-ratios were back-transformed to predict the three sediment fractions (mud, sand and gravel) using the formulae:

$$mud = \frac{exp(alr_m)}{exp(alr_m)+exp(alr_s)+1} \quad (3)$$

$$sand = \frac{exp(alr_s)}{exp(alr_m)+exp(alr_s)+1} \quad (4)$$

$$gravel = 1 - (mud + sand) \quad (5)$$

Model predictions were also converted to thematic classes for comparison between different treatments. This included two of the commonly used classification schemes by the EMODnet Geology community (Kaskela et al., 2019), Folk 5 and Folk 16, as well as the EUNIS Level 3 classification for broadscale sedimentary habitats, which is based on the simplified Folk triangle (Long, 2006). Classification accuracy was measured based on the observed versus predicted sediment type, with the different treatments compared based on overall accuracy. Additional insights into model success may also be gained by comparing different accuracy measures or the spatial predictions. However, for computational speed and simplicity, these were not considered here.

Each case study is described below: providing a brief rationale for how these analyses relate to less precise samples, the methodology of how the data were treated and the results.

3. Case Study 1 – Number of samples

3.1. Rationale

Random forest models are a type of machine learning model which use real world observations to train an algorithm and in turn can then be applied to predict the most probable outcome in new areas based on a set of known predictor variables. Therefore, machine learning models are unreliable when predicting into novel environments which they

do not have training observations to fit the algorithm. Machine learning models are therefore sensitive to the number of samples used (Hernandez et al., 2006; Mitchell et al., 2017; Wisz et al., 2008), particularly when the relationships between the predictor variables and observed outcomes are complex. This sensitivity to the number of samples is not necessarily linear, as model performance may improve rapidly at first with the addition of samples, but then improvements in model performance diminish as the number of samples increase beyond a certain point (Hernandez et al., 2006; Wisz et al., 2008).

Within the northwest European continental shelf there were tens of thousands of samples available, which provided ample samples for both model testing and model training. However, this may not be the case for other European sea regions, where less laboratory measured samples are available. This case study explores the relationship between the number of laboratory analysed samples and predictive performance of the models to provide an indication of the number of samples required for other similar quantitative sediment mapping analyses.

3.2. Data treatment

Using only the ILM training samples, bootstrap sampling with replacement was performed to provide training datasets ranging in size from 100 samples to 30,679 samples (the entire ILM training dataset). The alr_m and alr_s models were then trained with no additional tuning and their predictive performance was tested against the same ILM testing dataset.

3.3. Results

As training sample size decreased the accuracy of model predictions also decreased, however, this change was not linear. More training samples resulted in alr_m and alr_s models' predictions having a greater variance explained (and smaller MSE), when assessed using the testing data. The relationship generally followed a logarithmic scale (Figure 2). Once the sediment fraction maps were translated into thematic maps a similar same relationship was observed (Figure 3). The greatest changes were observed when training sample sizes were less than 1000 samples. By 2500 training samples the explained variances of the models were 83% and 87% of the explained variance of the two additive log-ratio models derived from the full training dataset (30,679 samples), and thematic map accuracy was >90% of the maximum achieved overall accuracy. The minimum sample size assessed was 100 training samples, by which point the overall accuracy of the thematic maps was approximately equivalent or below the no information rate for the three classification schemes.

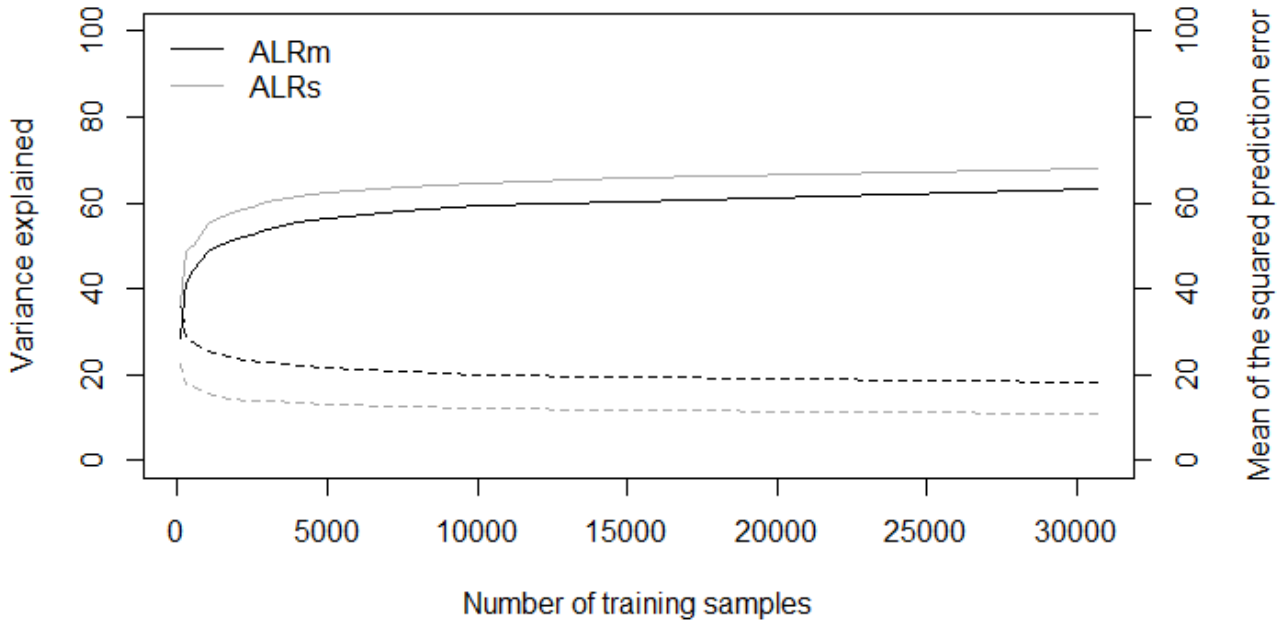


Figure 2. Effect of training sample size on the variance explained (solid lines and scale bar on the left side) and mean of the squared prediction error (MSE) (dashed lines and scale bar on the right side) of the two additive log-ratio models.

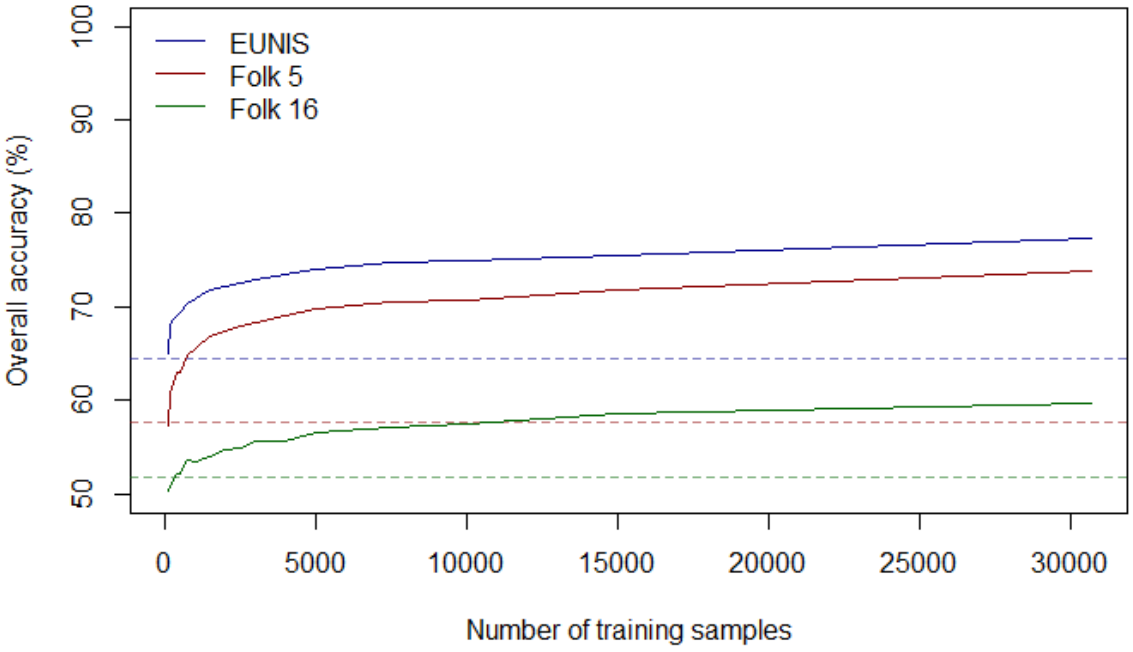


Figure 3. Effect of training sample size on the overall accuracy once the quantitative sediment predictions were converted into thematic maps. Final sediment predictions were classified based on the EUNIS Level 3, Folk 5 and Folk 16 classification schemes (represented by different colours. The dashed lines represent the no information rate of each of the thematic map for the classification scheme of the same colour, calculated from the testing samples.

4. Case Study 2 – Using less robust samples across multiple spatial scales

4.1. Rationale

Presented with the situation of a European marine region that only has a limited number of recent laboratory measured samples but an abundance of samples classified using expert visual assessment and/or samples several decades old, its necessary to consider what can be done with those less robust samples. Is it preferable to reject all samples that do not meet the necessary data quality standards, even if this means the training dataset may be very small (explored in Case Study 1)? Or would it be better to retain the less robust samples in model training? This case study explores the trade-off between developing quantitative sediment maps from less robust samples or a combination of less robust and recent laboratory measured samples.

When considering how best to overcome the limitations of less robust samples this case study also includes spatial scale as a factor that could be altered when less robust samples are to be included in the models. Given the spatial mismatch between the grab samples used (<1 m^2 for all grab types) versus the minimum mapping unit (the predictor variable pixels are ~30,000 m^2), it's unlikely that any one sample is entirely representative of the local conditions. In the previous study, Mitchell et al. (2019) calculated the average mud, sand and gravel fractions for pixels where multiple groundtruth observations were present. This same logic could be applied when including the less robust samples to train the model, as it is likely that the average mud, sand and gravel fractions from multiple samples are more representative of the true fractions than when only one sample is available. In the most detailed resolution (7.5 arc seconds) the number of groundtruth samples that were taken within the same pixel were generally limited. However, that number will increase as the pixels become larger, meaning more training samples will be generated as an average of multiple samples compared to a single sample.

4.2. Data treatment

Analyses used five combinations of the IVA training and ILM training quantitative sediment samples (Table 2). These were: the IVA training samples only; an equal number of IVA training samples and ILM training samples; twice as many ILM training samples as IVA training samples; three times as many ILM training samples as IVA training samples; and the ILM training samples only. In the first four treatments the number of IVA training samples were kept constant and the ILM training samples were varied. For a single spatial scale, the predictions were then generated and tested against the same reference dataset made up entirely of ILM samples.

Table 2. Total number of samples used to train the alr_m and alr_s models for each data treatment. For each of the four spatial resolutions used, models were generated using five different combinations of IVA and ILM data. The ratio of samples used for each treatment was based on the raw samples, prior to averaging within each grid cell.

Ratio of raw samples IVA and ILM datasets	Spatial resolution			
	7.5 arc seconds	15 arc seconds	30 arc seconds	60 arc seconds
	Number of samples			
1:0 (IVA : ILM)	12,114	11,898	11,466	10,416
1:1 (IVA : ILM)	23,161	22,544	21,197	18,269
1:2 (IVA : ILM)	33,334	32,173	29,495	24,157
1:3 (IVA : ILM)	42,539	40,752	36,657	28,842
0:3 (IVA : ILM)	30,679	28,932	25,946	19,687

These quantitative sediment maps were then generated for the five combinations of IVA training and ILM training samples in four spatial scales (7.5, 15, 30 and 60 arc seconds). To prepare the data, the eight predictor variables were resampled onto a common grid for each of the four spatial resolutions using a cubic resampling approach in ArcGIS (ESRI, 2016). Once these grids had been generated the average mud, sand and gravel fractions were calculated for each pixel where multiple samples were present. For those treatments that used a mix of IVA and ILM samples, each sample had an equal weighting. As the density of samples varied across the study area, increasing the pixel size reduced the total number of training cells, but a doubling of size did not result in half the final number of samples.

4.3. Results

Model predictive performance was highest for models derived from the ILM training samples only. Once the IVA samples were added to this training dataset, despite increasing the sample size the predictive performance decreased. The decrease in explained variance of the testing dataset was ~3% in the alr_m and alr_s models and a decrease of ~1% of overall accuracy of the thematic maps, regardless of the spatial scale used. The worst models were those generated from only the IVA samples, which had explained variances that were 29%-34% less than the ILM only models and 20%-25% less of the variance than the 1:1 IVA:ILM training data. Once the sediment fraction predictions were converted to thematic maps these also resulted in decreased overall accuracy when the less robust IVA samples were included in the models. Of note were the overall accuracies of the Folk 16 thematic maps derived from IVA only data. For these maps the overall accuracy was less than the no information rate, i.e. had the whole map been predicted as the most common class (Sand) it would have been more accurate than the IVA only maps.

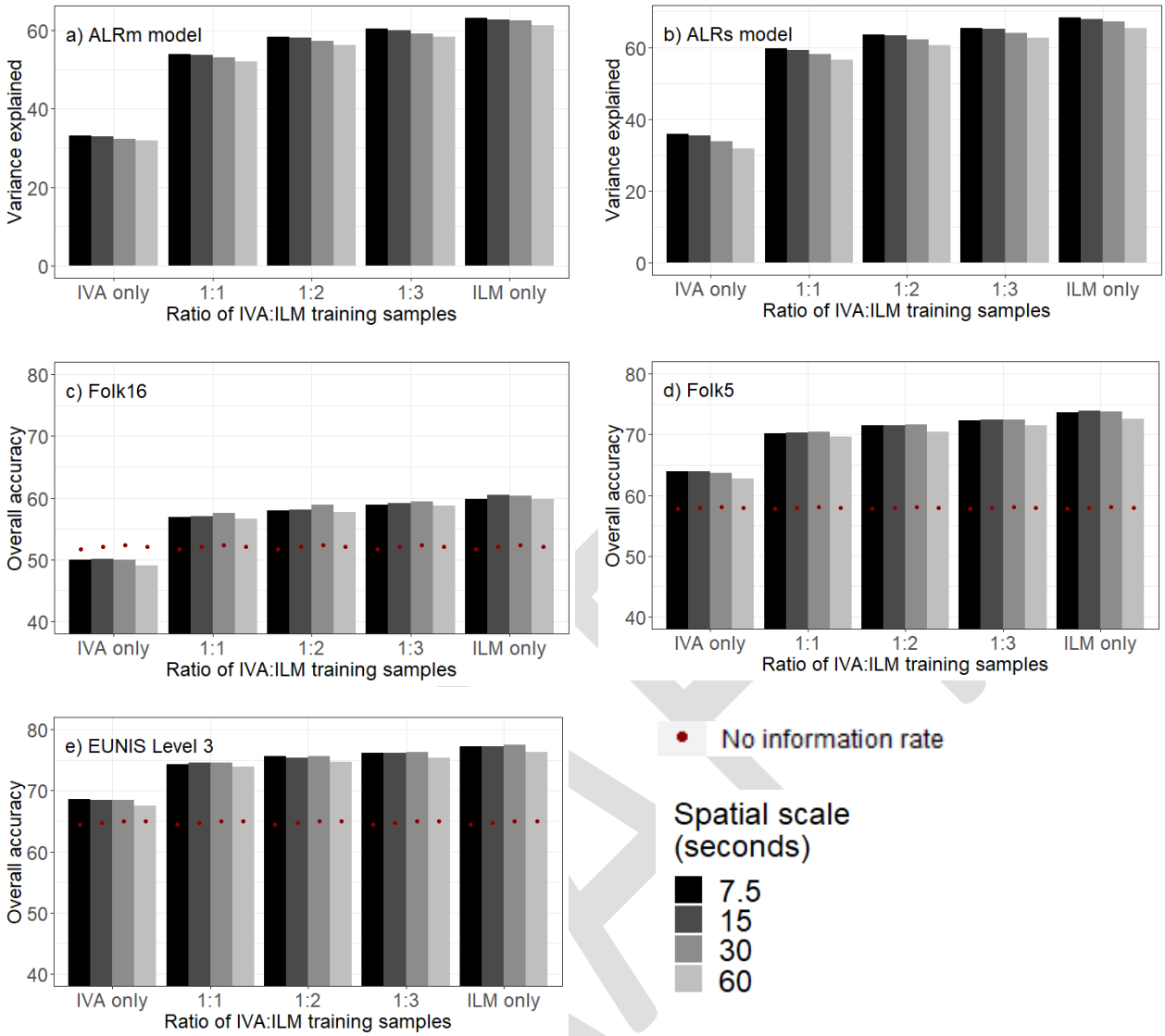


Figure 4. Model performance and map accuracy of quantitative sediment maps derived from different combinations of IVA and ILM training data, generated across four spatial scales. The variance explained of the (a) alr_m and (b) alr_s models used to predict sediment fractions were calculated from withheld testing data. Having generated thematic maps using three different classification schemes (c) Folk 16, (d) Folk 5 and (e) EUNIS level 3, the map accuracy was also calculated against the withheld testing data. Columns are grouped by the ratio of IVA to ILM samples used to train the models and colours represent the different spatial scales. For each model the no information rate is overlaid on the bar.

5. Case Study 3 – Less precise samples

5.1. Rationale

The third case study explores the question of: how precisely do the sediment fractions need to be measured? When compiled at a national level the databases contain samples from

numerous studies that were assessed using a mixture of techniques such as laser diffraction, sieving and expert visual assessment. Different methods, and even different laboratories, can produce slightly different results from the same samples (Konert and Vandenberghe, 1997; Passchier, 2007; Rodríguez and Uriarte, 2009; Silburn et al., 2018). In particular, the expert visual assessment would be the most subjective, and reliably estimating the proportion of mud, sand and gravel would be highly dependent on the analyst. Mitchell et al. (2019) discarded samples where recorded fractions were rounded percentages (e.g. 10%, 20% or 25%), as the assumption was these had been estimated and not accurately measured. Whether these percentages were assigned directly in the field or a written description of the sediment was recorded and then fractions estimated at a later date, these probably represent the most imprecise data. However, given that the analytical methods can produce small differences in the results it would be useful to understand how sensitive the models are to varying levels of imprecision in the training data.

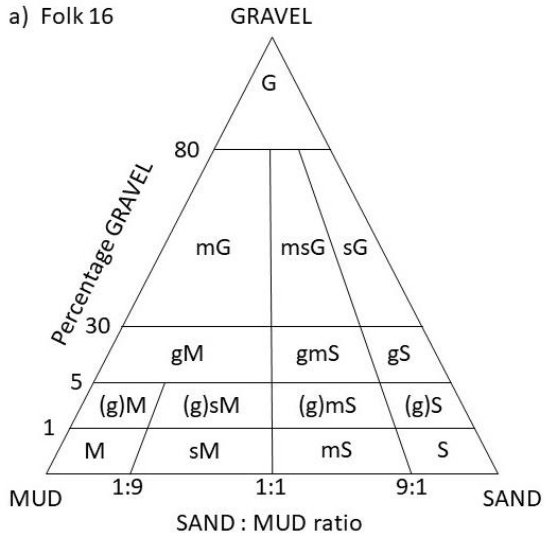
While there are many reasons why measuring highly precise sediment fractions are desirable, it is unknown whether this precision is required for producing quantitative sediment maps given the other sources of error within the modelling process. The models published in Mitchell et al. (2019) explain 63% and 68% of the variance, so there remains a reasonable amount of variation within the data not explained by the models. Strong (2020) provides a detailed assessment of many of the potential sources of error common to habitat mapping approaches. Many of these are likely to occur within the current quantitative sediment analysis and data (e.g. sampling resolution, choice of model algorithm, selection of predictor variables, uncertainty within predictor variables and replication of samples). Therefore, it's worth considering how important sample imprecision is, given the many other possible sources of error present within the entire modelling process.

As many national sediment sample archives from other regions of Europe have a large proportion of samples derived from expert visual assessment it would be valuable to understand how precise data samples should be before they start to reduce model accuracy. This case study explores this question by simulating different levels of precision within the training dataset and then running the same modelling approaches on these less precise data.

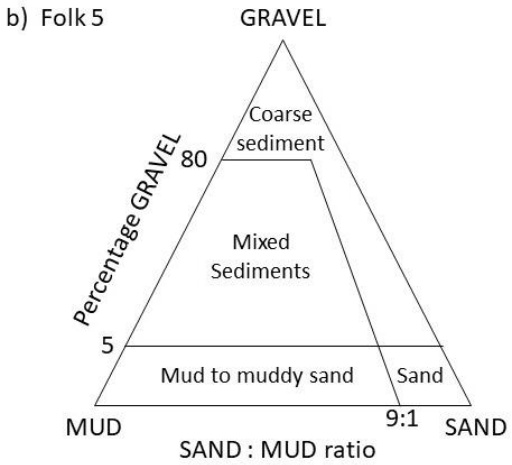
5.2. Data treatment

The ILM training dataset was used to assess different levels of precision by rounding the mud, sand and gravel data to different levels of precision. Seven different levels of precision were assessed by rounding the mud, sand and gravel fractions to the nearest: 0.01%, 0.1%, 1%, 5%, 10% 25% and 50%. This was applied using the full ILM training dataset at 7.5 seconds resolution. From these less precise mud, sand and gravel fractions the two additive log-ratios were calculated, and the same modelling approach described previously was followed.

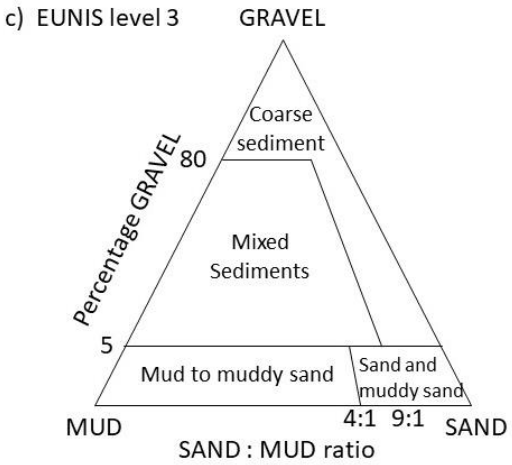
In addition, three further datasets were created to simulate samples where the mud, sand and gravel fractions had been estimated from written descriptions. To do this the centroid values of mud, sand and gravel were calculated from the Folk triangle for the Folk 16, Folk 5 and EUNIS level 3 classification schemes (Figure 5). All samples were from the ILM training dataset were then assigned these centroid values based on the thematic class that that sample would have been assigned. For example, a sample from the ILM training dataset which had 51% mud, 46% sand and 2% gravel would be classified as a ‘Slightly gravelly sandy mud’ in the Folk 16 classification scheme and ‘Mud to muddy sand’ in the Folk 5 and EUNIS level 3 classification schemes. Therefore, it would be reassigned the values 67.9% mud, 29.1% sand and 3% gravel for Folk 16, 53.625% mud, 43.875% sand and 2.5% gravel for Folk 5 and 58.5% mud, 39% sand and 2.5% gravel for EUNIS level 3. Once the mud, sand and gravel fractions had been recalculated based on the thematic classes for each dataset the two additive log-ratios were calculated and the same modelling approach described previously was followed.



Code	Description	Fraction		
		Mud (%)	Sand (%)	Gravel (%)
G	Gravel	5	5	90
mG	Muddy gravel	33.75	11.25	55
msG	Muddy sandy gravel	13.375	31.625	55
sG	Sandy gravel	2.25	42.75	55
gM	Gravelly mud	61.875	20.625	17.5
gmS	Gravelly muddy sand	25.25	57.25	17.5
gS	Gravelly sand	4.125	78.375	17.5
(g)M	Slightly gravelly mud	92.15	4.85	3
(g)sM	Slightly gravelly sandy mud	67.9	29.1	3
(g)mS	Slightly gravelly muddy sand	29.1	67.9	3
(g)S	Slightly gravelly sand	4.85	92.15	3
M	Mud	94.525	4.975	0.5
sM	Sandy mud	69.65	29.85	0.5
mS	Muddy sand	29.85	69.65	0.5
S	Sand	4.975	94.525	0.5



Class	Fraction		
	Mud (%)	Sand (%)	Gravel (%)
Coarse sediment	2.375	45.125	52.5
Mixed sediments	31.625	25.875	42.5
Mud to muddy sand	53.625	43.875	2.5
Sand	4.875	92.625	2.5



Class	Fraction		
	Mud (%)	Sand (%)	Gravel (%)
Coarse sediment	2.375	45.125	52.5
Mixed sediments	31.625	25.875	42.5
Mud to muddy sand	58.5	39	2.5
Sand and muddy sand	9.75	87.75	2.5

Figure 5. All sediment samples from the ILM training dataset were reclassified based on the (a) Folk 16, (b) Folk 5 and (c) EUNIS level 3 classification schemes, shown here on the Folk triangles. Percentages and fractions are not drawn to scale. Each of these sediment samples were assigned the percentages of mud, sand and gravel shown in the adjacent table. Values were calculated based on the centroid of each class within the Folk triangles.

5.3. Results

Rounding the sediment fraction data to the nearest percentage point in the training groundtruth data had a minimal effect on the models' explained variance and thematic map

overall accuracy. In comparison to the models derived from the unaltered data, the explained variance decreased by 4% and 3% for the alr_m and alr_s models respectively, and the overall accuracy increased by <1% for the Folk 16 map but decreased by 3% for the Folk 5 and 1% for the EUNIS level 3 maps (Figure 6). As the sediment fraction data were made less precise through rounding, the decrease in model predictive performance and accuracy became more noticeable. In the most extreme example, where sediment fraction data were rounded to the nearest 50%, the two models' explained variance decreased by 28% and 23% compared to the unaltered data. Once the sediment fraction predictions were concerted to thematic maps the difference in model performance was less dramatic (5% for Folk 16, 8% for Folk 5 and 6% for EUNIS level 3).

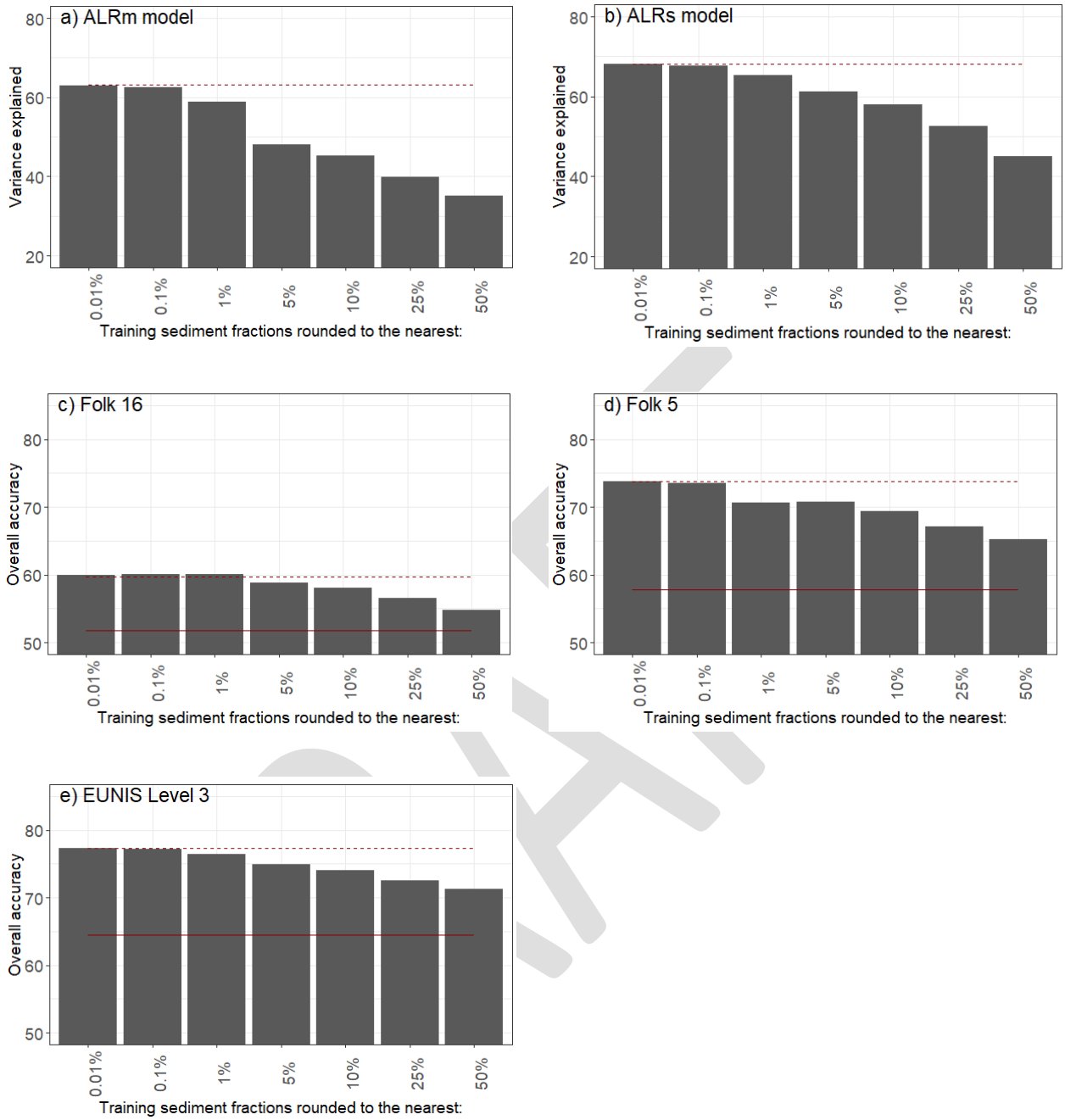


Figure 6. Predictive performance and thematic map accuracy of model outputs that were generated from training data containing varying levels of sediment fraction precision. Different levels of precision were simulated by rounding the sediment fractions (percentages of mud, sand and gravel) in the training data. Predictive performance was assessed using the testing data based on the (a) alr_m and (b) alr_s models' variance explained, and the overall accuracy of the thematic maps using three different classification schemes (c) Folk 16, (d) Folk 5 and (e) EUNIS level 3. The dashed red line represents the variance explained and overall accuracy of the original unaltered data and the solid line represents the no information rate of the thematic maps.

The second component of the case study, which simulated using thematic class groundtruth samples to estimate sediment fractions and use these for analyses, found that these data were of limited use for generating reliable maps. This included the classification scheme that retained the most precise information, Folk 16, and assigned all training samples to one of 15 sediment classes. The quantitative sediment maps that were derived from these reclassified data had much lower predictive performance than the maps derived from the original unaltered sediment data (Figure 7). For the Folk 16 training samples the decrease in the alr_m and alr_s models' variance explained relative to the unaltered data were 10% and 7% respectively. Once translated into thematic classes this had the greatest reduction in overall accuracy at the Folk 16 level (down 6%), while the Folk 5 and EUNIS level 3 overall accuracies decreased by 3% each. Where training samples were converted to less precise classification schemes (i.e. Folk 5 and EUNIS level 3 which only have four sediment classes each), the quality of map products decreased further, and in some instances were even below the no information rate (Figure 7c and 7d).

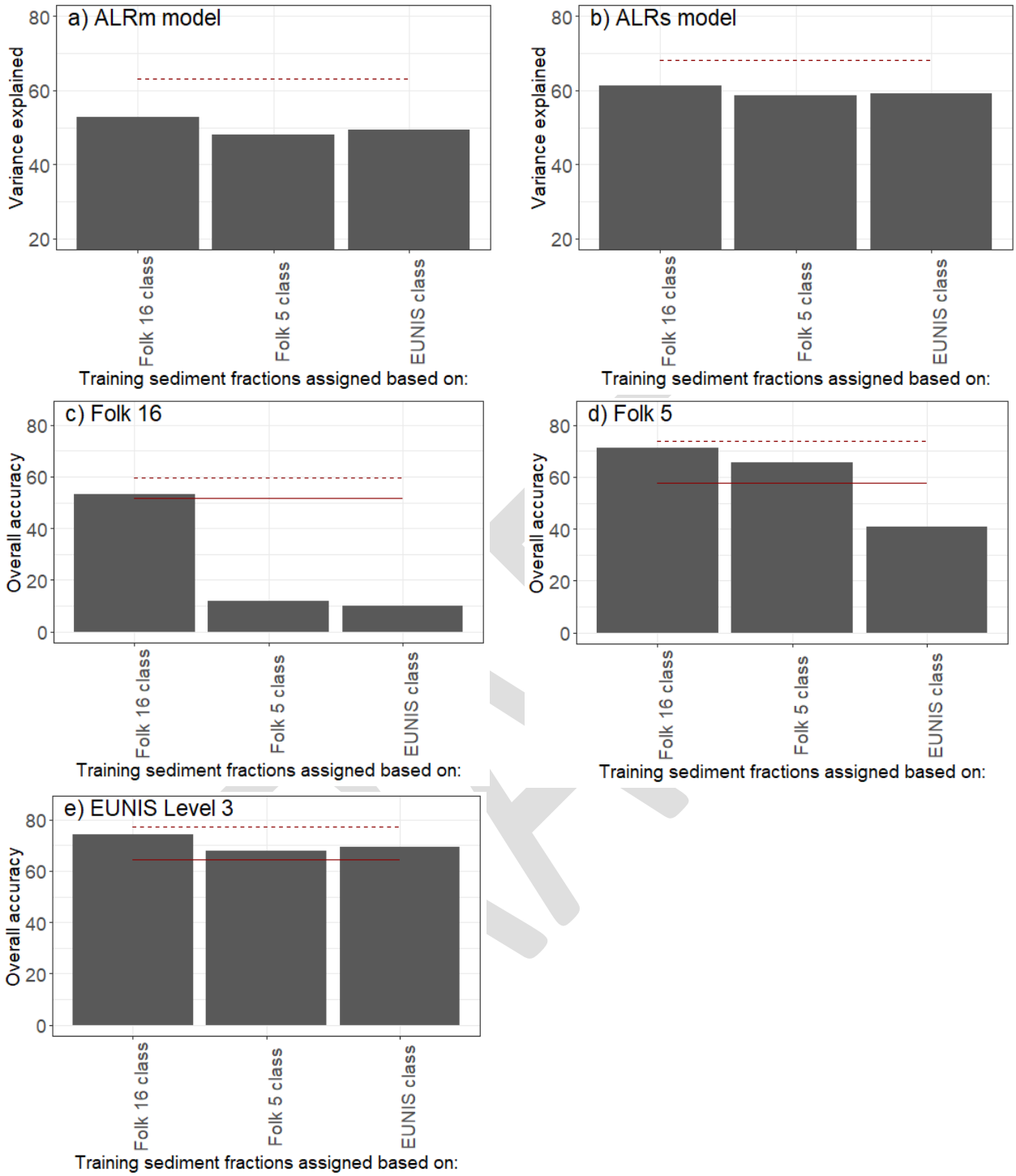


Figure 7. Predictive performance and thematic map accuracy of model outputs that were generated from training data which had been reclassified to different classification schemes. Training sediment fractions were assigned based on the centroid value within the Folk triangles (Figure 5), and this sediment fraction data were then used in the quantitative sediment mapping process. Predictive performance was assessed using the testing data based on the (a) *alr*$_m$ and (b) *alr*$_s$ models' variance explained, and the overall accuracy of the thematic maps using three different classification schemes (c) Folk 16, (d) Folk 5 and (e) EUNIS level 3. The dashed red line represents the variance explained and overall accuracy of

the original unaltered data and the solid line represents the no information rate of the thematic maps.

6. Discussion

The three case studies each provide valuable insights into how changes to the groundtruth samples, used to train the machine learning models, influence the quality of the quantitative sediment maps produced. To briefly summarise: the first case study observed that model predictive performance decreased as training sample size decreased; the second case study observed that at all spatial scales and ratios of IVA to ILM samples assessed, the IVA samples reduced the quality of the models; and the third case study observed that the quality of models withstood minor levels of imprecision in the sediment fraction data. Individually the findings of each of these case studies are not necessarily surprising, but considering these three case studies together provide some interesting lessons for future work. Given the intention of future EMODnet Geology phases is to develop further quantitative sediment maps for other regions within Europe, the results are discussed in this context, and a number of recommendations based on the case studies are identified.

The first case study observed that increased sample size of quantitative samples used for model training produced better predictive performance and more accurate maps. This result was expected as this relationship is well documented (Foody, 2009; Millard and Richardson, 2015; Wisz et al., 2008). What is noteworthy is quantifying the strength of this relationship so we can apply these recommendations to similar analyses in other sea-basins of Europe in the near future. Even as the sample size increased to ~30,000 samples the variance explained of the models and overall accuracy of the thematic maps had not completely plateaued, however, the greatest increases in predictive performance occurred as the training sample size increased to 1000 samples (Figure 2 and Figure 3). Given the number of countries covered, it is unlikely that other regions will have as many samples as the northwest European continental shelf, but it is encouraging that as few as 1000 samples, but preferably 2500, can produce reasonable results. It is also worth considering this in the context of the second case study, as models developed from as few as 300 ILM samples (i.e. ~1% of the training samples) outperformed the models and maps developed from ~12,000 IVA samples (based on variance explained and thematic map overall accuracy). This suggests that each reliable sediment sample – i.e. that is recent, accurately positioned and precisely measured – has far greater value to the modelling process than the historic samples which may have one or a combination of these issues.

Assessing the quality of sediment samples prior to inclusion in the analysis is an important stage in the modelling process. As any model will generally contain some error (here the top performing additive log-ratio models explained 63-68% of the variance), it can be that some inaccuracy in the training data are tolerated if it provides observations in new areas. However, in the second case study the inclusion of IVA samples into the training dataset decreased the models' explained variance by ~3% and thematic maps' overall accuracy by ~1% (comparing 1:3 IVA to ILM samples to ILM samples only (Figure 4)). This was the case

for all spatial scales assessed. This suggests that the IVA data contain too much error to be useful to train the models. The third case study provided an example of less precise data that could be included in the model. The influence on variance explained and thematic map accuracy was minimal where sediment fractions were rounded to the nearest percentage point (Figure 6). However, beyond this the decrease in the quality of model outputs became more noticeable.

The second case study only presented one method of including the less robust groundtruth sediment samples in the modelling process. Models which included IVA samples gave these samples equal weighting to the ILM samples regardless of their spatial location or how densely sampled that environmental space was. This was the simplest approach of including the IVA samples, however, other methodologies might identify ways of improving the map products compared to those ILM only samples. For example, only retaining IVA samples for locations or positions within environmental space that were not sampled by ILM samples. One such area within this study site is off the west coast of Scotland which had a limited number of ILM samples but many IVA samples (Figure 1). Exploring other methods to improve the quantitative sediment maps using IVA samples was beyond the scope of this report, but would be an area for further research.

In national databases, where sediment samples only contain the thematic class of the samples and not the measured sediment fractions, the results of the third case study suggest these samples should be discarded. By estimating the mud, sand and gravel fractions based on the centroid of the Folk triangle, it was possible to use these as training data to generate quantitative sediment maps from these data. However, even for the most detailed descriptions (Folk 16 classes), the model predictions were worse in terms of both variance explained and thematic map accuracy (Figure 7).

These results suggest the need to focus on applying clear minimum standards for the data to be utilised in the model rather than incorporating all samples regardless of vintage or quality. From the northwest European databases, only limited information was available to assess sample quality (such as year collected and instrument type). Therefore, samples were grouped into IVA or ILM samples by inferring the measurement technique rather than this being recorded within the sample metadata. Applying best judgement in determining which samples to include in the modelling process may be the only option. These considerations of data quality should also consider the data withheld in the testing or validation dataset. As reference samples may also contain errors (Strahler et al., 2006), these could result in a misleading estimation of map accuracy. Therefore, strict data standards should also be applied to the testing samples in order to reliably assess map quality.

These case studies only investigated some of the issues associated with the groundtruth samples used to train the models and did not consider the other source of data: the predictor variables. Mitchell et al. (2019) used a combination of bespoke predictive layers generated for the analysis and previously existing data (e.g. EMODnet Bathymetry Consortium, 2016). Countless other variables could be considered for inclusion in the analysis for other areas,

such as terrain variables (Brown et al., 2011; Lecours et al., 2017) or oceanographic variables (e.g. Diesing et al., 2020). Further, simple location specific variables such as distance from rivers (Mitchell et al., 2021), or latitude and longitude could also be considered in any future analysis, depending on the processes suspected to drive sediment distributions within that region.

With the intention to develop further quantitative sediment maps for other regions within the EMODnet Geology project, this report has addressed some of the considerations for identifying suitable groundtruth data. Nevertheless, it may be that for some areas the quantity and quality of the sediment sample data do not meet these standards. Should this be the case, the three presented case studies should provide some indication of how that will impact the quality of mapped outputs.

7. References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data, Journal of the Royal Statistical Society. Series B. Methodological. Chapman and Hall, London. https://doi.org/10.2307/2345821

Baker, E.K., Harris, P.T., 2012. Habitat Mapping and Marine Management. Seafloor Geomorphol. as Benthic Habitat 23–38. https://doi.org/10.1016/B978-0-12-385140-6.00002-5

Breiman, L., 2001. Random forests. Mach. Learn. 5–32. https://doi.org/10.1023/A:1010933404324

Brown, C.J., Smith, S.J., Lawton, P., Anderson, J.T., 2011. Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. Estuar. Coast. Shelf Sci. 92, 502–520. https://doi.org/10.1016/j.ecss.2011.02.007

Cooper, K.M., Bolam, S.G., Downie, A.-L., Barry, J., 2019. Biological-based habitat classification approaches promote cost-efficient monitoring: An example using seabed assemblages. J. Appl. Ecol. 00, 1–14. https://doi.org/https://doi.org/10.1111/1365-2664.13381

Diesing, M., 2015. Case study: Quantitative spatial prediction of seabed sediment composition.

Diesing, M., Thorsnes, T., Bjarnadóttir, L.R., 2020. Organic carbon in surface sediments of the North Sea and Skagerrak. Biogeosciences Discuss. 1–30. https://doi.org/10.5194/bg-2020-352

EMODnet Bathymetry Consortium, 2016. EMODnet Digital Bathymetry (DTM 2016). https://doi.org/10.12770/c7b53704-999d-4721-b1a3-04ec60c87238

ESRI, 2016. ArcGIS Desktop.

Folk, R.L., 1954. The distinction between grain size and mineral composition in sedimentary-

rock nomenclature. J. Geol. 62, 344–359.

Foody, G.M., 2009. Sample size determination for image classification accuracy assessment and comparison. Int. J. Remote Sens. 30, 5273–5291. https://doi.org/10.1080/01431160903130937

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Peterson, A.T., Loiselle, B. a., Anderson, R.P., Dudik, M., Ferrier, S., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Phillips, S.J., Richardson, K.S., Pereira, R.S., Schapire, R.E., Soberón, J., Williams, S.E., Wisz, M.S., Zimmermann, N.E., 2008. The influence of spatial errors in species occurrence data used in distribution models. J. Appl. Ecol. 45, 239–247. https://doi.org/10.1111/j.1365-2664.2007.01408.x

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography (Cop.). 29, 773–785.

Kaskela, A.M., Kotilainen, A.T., Alanen, U., Cooper, R., Green, S.L., Guinan, J.C., Heteren, S. Van, Kihlman, S., Lancker, V. Van, Stevenson, A., EMODnet Geology Partners, 2019. Picking up the pieces—Harmonising and collating seabed substrate data for European maritime areas. Geosciences 9, 1–18. https://doi.org/10.3390/geosciences9020084

Konert, M., Vandenberghe, J., 1997. Comparison of laser grain size analysis with pipette and sieve analysis: a solution for the underestimation of the clay fraction. Sedimentology 44, 523–535. https://doi.org/10.1046/j.1365-3091.1997.d01-38.x

Lark, R.M., Dove, D., Green, S.L., Richardson, A.E., Stewart, H. a., Stevenson, A., 2012. Spatial prediction of seabed sediment texture classes by cokriging from a legacy database of point observations. Sediment. Geol. 281, 35–49. https://doi.org/10.1016/j.sedgeo.2012.07.009

Lecours, V., Devillers, R., Simms, A.E., Lucieer, V.L., Brown, C.J., 2017. Towards a framework for terrain attribute selection in environmental studies. Environ. Model. Softw. 89, 19–30. https://doi.org/10.1016/j.envsoft.2016.11.027

Liaw, A., Wiener, M., 2018. Breiman and Cutler's Random Forests for Classification and Regression.

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

Luisetti, T., Turner, R.K., Andrews, J.E., Jickells, T.D., Kröger, S., Diesing, M., Paltriguera, L., Johnson, M.T., Parker, E.R., Bakker, D.C.E., Weston, K., 2019. Quantifying and valuing carbon flows and stores in coastal and shelf ecosystems in the UK. Ecosyst. Serv. 35, 67–76. https://doi.org/10.1016/J.ECOSER.2018.10.013

Mason, C., 2016. NMBAQC's best practice guidance. Particle Size Analysis (PSA) for supporting biological analysis.

Millard, K., Richardson, M., 2015. On the importance of training data sample selection in Random Forest image classification: A case study in peatland ecosystem mapping.

Remote Sens. 7, 8489–8515. https://doi.org/10.3390/rs70708489

Mitchell, P.J., Aldridge, J.N., Diesing, M., 2019. Legacy data: How decades of seabed sampling can produce robust predictions and versatile products. Geosciences 9, 182. https://doi.org/10.3390/geosciences9040182

Mitchell, P.J., Monk, J., Laurenson, L.J.B., 2017. Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. Methods Ecol. Evol. 8, 12–21. https://doi.org/10.1111/2041-210X.12645

Mitchell, P.J., Spence, M.A., Aldridge, J., Kotilainen, A.T., Diesing, M., 2021. Sedimentation rates in the Baltic Sea: A machine learning approach. Cont. Shelf Res. 214, 104325. https://doi.org/10.1016/j.csr.2020.104325

Passchier, S., 2007. 14 Particle Size Analysis (granulometry) of Sediment Samples, in: Coggan, R.A., Populus, J., White, J., Sheehan, K., Fitzpatrick, F., Piel, S. (Eds.), Review of Standards and Protocols for Seabed Habitat Mapping. MESH, pp. 116–126.

Populus, J., Vasquez, M., Albrecht, J., Manca, E., Agnesi, S., Al Hamdani, Z., Anderson, J., Annunziatellis, A., Bekkby, T., Bruschi, A., Doncheva, V., Drakopoulou, V., Duncan, G., Inghilesi, R., Kyriakidou, C., Lalli, F., Lillis, H., Mo, G., Muresan, M., Salomidi, M., Sakellariou, D., Simboura, M., Teaca, A., Tezcan, D., Todorova, V., Tunesi, L., 2017. EUSeaMap, a European broad-scale seabed habitat map. https://doi.org/http://doi.org/10.13155/49975 3

R Development Team, 2019. R: A language and environment for statistical computing.

Rodríguez, J.G., Uriarte, A., 2009. Laser diffraction and dry-sieving grain size analyses undertaken on fine- and medium-grained sandy marine sediments: A note. J. Coast. Res. 25, 257–264. https://doi.org/10.2112/08-1012.1

Silburn, B., Parker, Ruth, Mason, C., Parker, Reg, 2018. Rapid fines assessment: A quantitative volumetric method to assess fines content in marine soft sediments. Limnol. Oceanogr. Methods 16, 376–389. https://doi.org/10.1002/lom3.10252

Stephens, D., Diesing, M., 2015. Towards quantitative spatial models of seabed sediment composition. PLoS One 1–23. https://doi.org/10.1371/journal.pone.0142502

Strahler, A.H., Boschetti, L., Foody, G.M., Friedl, M. a., Hansen, M.C., Herold, M., Mayaux, P., Morisette, J.T., Stehman, S. V., Woodcock, C.E., 2006. Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps, European Communities, Luxembourg. Luxembourg.

Strong, J.A., 2020. An error analysis of marine habitat mapping methods and prioritised work packages required to reduce errors and improve consistency. Estuar. Coast. Shelf Sci. 240, 106684. https://doi.org/10.1016/j.ecss.2020.106684

Thompson, M.S.A., Couce, E., Webb, T.J., Grace, M., Cooper, K.M., Schratzberger, M., 2021. What's hot and what's not: Making sense of biodiversity 'hotspots.' Glob. Chang. Biol. 27, 521–535. https://doi.org/10.1111/gcb.15443

Ware, S., Downie, A.-L., 2019. Challenges of habitat mapping to inform marine protected

area (MPA) designation and monitoring: An operational perspective. Mar. Policy 111, 103717. https://doi.org/10.1016/j.marpol.2019.103717

Williams, M.E., Amoudry, L.O., Brown, J.M., Thompson, C.E.L., 2019. Fine particle retention and deposition in regions of cyclonic tidal current rotation. Mar. Geol. 410, 122–134. https://doi.org/10.1016/j.margeo.2019.01.006

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., Elith, J., Dudik, M., Ferrier, S., Huettmann, F., Leathwick, J.R., Lehmann, A., Lohmann, L.G., Loiselle, B. a., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Phillips, S.J., Richardson, K.S., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S.E., Zimmermann, N.E., 2008. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14, 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Centre for Environment Fisheries & Aquaculture Science

Cefas

World Class Science for the Marine and Freshwater Environment

We are the government's marine and freshwater science experts. We help keep our seas, oceans and rivers healthy and productive and our seafood safe and sustainable by providing data and advice to the UK Government and our overseas partners. We are passionate about what we do because our work helps tackle the serious global problems of climate change, marine litter, over-fishing and pollution in support of the UK's commitments to a better future (for example the UN Sustainable Development Goals and Defra's 25 year Environment Plan).

We work in partnership with our colleagues in Defra and across UK government, and with international governments, business, maritime and fishing industry, non-governmental organisations, research institutes, universities, civil society and schools to collate and share knowledge. Together we can understand and value our seas to secure a sustainable blue future for us all, and help create a greater place for living.

OGL

Pakefield Road, Lowestoft, Suffolk, NR33 0HT

The Nothe, Barrack Road, Weymouth DT4 8UB

www.cefas.co.uk | +44 (0) 1502 562244

disability confident EMPLOYER

ISO 9001

OHSAS 18001

SGS