Marine Geophysical Research (2022) 43:7
https://doi.org/10.1007/s11001-022-09471-3

ORIGINAL RESEARCH PAPER

Multiple imputation of multibeam angular response data for high resolution full coverage seabed mapping

Benjamin Misiuk[1] · Craig J. Brown[1]

Received: 20 September 2021 / Accepted: 28 February 2022 / Published online: 9 March 2022

Abstract

Acoustic data collected by multibeam echosounders (MBES) are increasingly used for high resolution seabed mapping. The relationships between substrate properties and the acoustic response of the seafloor depends on the acoustic angle of incidence and the operating frequency of the sonar, and these dependencies can be analysed for discrimination of benthic substrates or habitats. An outstanding challenge for angular MBES mapping at a high spatial resolution is discontinuity; acoustic data are seldom represented at a full range of incidence angles across an entire survey area, hindering continuous spatial mapping. Given quantifiable relationships between MBES data at various incidence angles and frequencies, we propose to use multiple imputation to achieve complete estimates of angular MBES data over full survey extents at a high spatial resolution for seabed mapping. The primary goals of this study are (i) to evaluate the effectiveness of multiple imputation for producing accurate estimates of angular backscatter intensity and substrate penetration information, and (ii) to evaluate the usefulness of imputed angular data for benthic habitat and substrate mapping at a high spatial resolution. Using a multi-frequency case study, acoustic soundings were first aggregated to homogenous seabed units at a high spatial resolution via image segmentation. The effectiveness and limitations of imputation were explored in this context by simulating various amounts of missing angular data, and results suggested that a substantial proportion of missing measurements (> 40%) could be imputed with little error using Multiple Imputation by Chained Equations (MICE). The usefulness of imputed angular data for seabed mapping was then evaluated empirically by using MICE to generate multiple stochastic versions of a dataset with missing angular measurements. The complete, imputed datasets were used to model the distribution of substrate properties observed from ground-truth samples using Random Forest and neural networks. Model results were pooled for continuous spatial prediction and estimates of confidence were derived to reflect uncertainty resulting from multiple imputations. In addition to enabling continuous spatial prediction, the high-resolution imputed angular models performed favourably compared to broader segmentations or non-angular data.

Keywords Multibeam echosounder · Acoustic backscatter · Multispectral · Benthic habitat mapping · Angular range analysis · Multiple imputation

Introduction

Seafloor substrate properties are mapped at a high resolution to characterize the benthic environment for a wide range of applications. High-resolution substrate information is necessary to produce detailed maps of surficial geology (e.g., Todd et al. 2014; Stephens and Diesing 2015; Misiuk et al. 2018 Mitchell et al. 2018) and is useful for estimating benthic species habitat suitability (McArthur et al. 2010). This information is often essential for local and regional marine management (Cogan et al. 2009), informing, for example, marine protected area (MPA) design (Howell et al. 2010; Ferrari et al. 2018) and recently, fisheries stock assessment and management (Smith et al. 2017).

Acoustic backscatter is one of the few direct surrogates available for seafloor substrate properties in waters too deep for spectral imaging (> ~ 20–30 m). Seafloor backscatter describes the echo intensity of an acoustic wave that has reflected off the bottom. The intensity reflected is partly a function of the material properties of the substrate, but also depends on internal and external substrate

✉ Benjamin Misiuk
ben.misiuk@dal.ca

[1] Department of Oceanography, Dalhousie University, Halifax, NS, Canada

structure (e.g., surface roughness, sediment bed forms, sediment stratigraphy), and characteristics of the acoustic signal such as the frequency and angle of incidence (Lurton 2010). If the latter factors are properly constrained, backscatter can serve as a useful quantitative indicator of material seabed properties, and has been correlated to specific substrate parameters (e.g., Davis et al. 1996; Goff et al. 2000, 2004; Collier and Brown 2005; Ferrini and Flood 2006; Sutherland et al. 2007; Haris et al. 2012). Recently, angle-varying gain (AVG) corrections have been widely employed to remove the angular dependence of the backscatter intensity from swath sonar systems, producing backscatter "mosaics" that can serve as a non-parametric predictor of seabed substrate and habitat type (Lurton and Lamarche 2015; Schimel et al. 2015). This provides the opportunity to treat backscatter data from multibeam echosounder (MBES) or sidescan sonar (SSS) systems as a raster layer, simplifying its use as a predictor in sediment and habitat modelling, while also enabling additional textural and image processing approaches (e.g., Lucieer and Lamarche 2011; Fakiris et al. 2019; Trzcinska et al. 2020).

Although the AVG backscatter mosaic has proven highly useful for mapping seabed substrate properties and benthic habitats, it requires a reduction of angle-dependent acoustic information that may be useful for discriminating seabed characteristics (Fonseca et al. 2009; Haris et al. 2012). Backscatter intensity co-depends on the angle of incidence and the properties of the substrate (e.g., roughness, grain size; Lamarche and Lurton 2018), and different MBES beam angles may provide complementary, non-redundant information describing substrate characteristics (Hughes Clarke et al. 1996). Scattering of inner acoustic beams, for example, is dominated by specular reflection, which is largely a function of seafloor hardness (Weber and Lurton 2015). At the outer beams, scattering is increasingly sensitive to the interface roughness. Some mapping approaches advocate retaining this rich acoustic information to better discriminate substrate properties. Angular range analysis (ARA) is an approach for calculating the full angular response curve (ARC) across each half of the MBES swath (i.e., from nadir to the outer beam on both the port and starboard sides of the swath), which can be used to estimate substrate properties, for example, by deriving parameters that are used to invert an acoustic backscatter model using calibrated backscatter data (Fonseca and Mayer 2007). The use of data spanning half the swath width results in low horizontal resolution for substrate predictions that are estimated using angular approaches. Fonseca et al. (2009) therefore proposed to aggregate soundings by homogenous patches of seabed identified from the backscatter mosaic rather than by the swath width, increasing the spatial resolution of the angular analysis. Che Hasan et al. (2012, 2014) applied this technique automatically by first segmenting the backscatter mosaic into "image objects", then aggregating soundings according to the segment boundaries.

Alternative approaches have sought spatially continuous solutions for retaining the resolution of the AVG backscatter mosaic while also achieving an estimate of the ARC. Parnum (2007) proposed to derive an "angular cube"—comparable to the hyperspectral cube in terrestrial remote sensing—by interpolating continuous surfaces from discrete incidence angles, thereby producing angular response estimates for each raster cell of a gridded dataset. Huang et al. (2014) suggested this could also be achieved by producing multiple backscatter mosaics using different reference angles for the AVG correction. These two approaches were compared by Alevizos and Greinert (2018), who found they produced similar results, yet noted the latter may be more flexible regarding the dataset overlap and sounding density (e.g., in deeper waters). Simons and Snellen (2009) have presented yet another solution, wherein the backscatter response of each beam is classified independently in an unsupervised Bayesian framework, precluding the need for angular compensation. If the survey overlap is sufficient, this approach can produce near-continuous maps (Alevizos et al. 2015)—otherwise, results can be interpolated (Gaida et al. 2019).

The advent of multi-frequency MBES presents new opportunities for seabed discrimination. Multi-frequency MBES vary the operating frequency on a "ping-by-ping" basis, cycling through a pre-selected range of frequencies (e.g., 100, 200, 400 kHz; Brown et al. 2019). In addition to multiple promising approaches for utilizing the backscatter response of these individual frequencies for substrate discrimination (e.g., Buscombe and Grams 2018; Costa 2019; Gaida et al. 2019), Gaida et al. (2020) demonstrated the potential for measuring differences in substrate penetration from the depth soundings of different frequencies. They recommended analysing the multi-frequency backscatter intensity alongside the difference in depth measurements to obtain a robust estimate of both the surficial sediment composition and water depth. Additionally, they provide evidence that differences in substrate penetration between frequencies vary substantially with both incidence angle and substrate composition.

The interrelationships between backscatter intensity, incidence angle, substrate penetration, and operating frequency for multi-frequency MBES are complex, yet we believe this complexity could be utilized for addressing outstanding challenges regarding the application of high-resolution angle-dependent backscatter analysis. All angular response analyses must contend with the discontinuous nature of angular data. Where soundings are sufficiently dense, this issue is readily mitigated, for example, by simply gridding the soundings at an appropriate cell size (Alevizos et al. 2018). In fact, ensuring a sufficiently dense survey so that all angles of incidence are represented at the desired spatial resolution is the most straightforward and robust solution to the issue of angular data discontinuity. As the depth increases though, or where the survey overlap of pre-existing data is insufficient, such solutions become infeasible, and alternative methods such as object-based aggregation (Che Hasan et al. 2012, 2014), interpolation (Parnum 2007; Alevizos and Greinert 2018), or multiple angular compensations (Huang et al. 2014; Alevizos and Greinert 2018) are required. Furthermore, multi-frequency MBES may also suffer reduced sounding density in the along-track direction, per frequency, as the system must cycle through the frequencies sequentially.

Multicollinearity is also a characteristic of MBES angular response data that is increased in multi-frequency systems. MBES backscatter intensity at any given angle of incidence is expected to be correlated with intensities at other, similar angles, which is one reason why AVG compensation is successful. Though different operating frequencies will produce unique angular response curves (Weber and Lurton 2015), some amount of correlation is still expected between them, depending on the substrate—effectively increasing the dimensionality of multicollinearity. Multicollinearity is, of course, problematic in some modelling contexts, but can also be highly valuable for handling missing data. Data imputation techniques enable the leveraging of collinearity between variables in order to facilitate statistical modelling using incomplete observations (i.e., data points with missing values for some variables).

Data imputation refers to a suite of methods for completing datasets with missing values in order to avoid the listwise deletion of partially complete observations in downstream analyses (Rubin 1987; van Buuren 2018). These techniques are now well-developed and have been applied to missing data problems in public censes (Rubin 1987), psychology (van Ginkel et al. 2010), medicine and epidemiology (e.g., Ambler et al. 2007; Vergouw et al. 2012; Eekhout et al. 2012), and biology (e.g., Troyanskaya et al. 2001; Penone et al. 2014). The general approach of most data imputation techniques is to estimate missing observations of variables using relationships with other variables where observations are not missing. Imputation of univariate missing data is readily accomplished, for example, by modelling the missing values according to the complete observations, or by drawing from observed values. Multivariate imputation—where missing data are imputed for multiple variables simultaneously—is non-trivial, yet can be accomplished using Fully Conditional Specification (FCS; van Buuren et al. 2006), wherein missing data from each variable are modelled iteratively and sequentially until the scope of the imputation is achieved. Furthermore, "multiple imputation" describes methods in which the missing data are imputed multiple times to obtain "proper" estimates of the uncertainty of the missing data (Rubin 1987, 2004; van Buuren 2018). Unlike many modelling applications, imputation stands to benefit from multicollinearity within the dataset, which can be leveraged to estimate missing values where no data occur.

Given the multidimensional multicollinearity of multi-frequency MBES datasets, and the preponderance of missing data at any given angle of incidence at the seafloor, we propose to evaluate the use of multiple imputation to estimate missing angular data at a high spatial resolution. This is motivated primarily by (i) the common need for spatially continuous map predictions that require complete observations of all explanatory variables over the study extent, and (ii) a desire to avoid omitting ground truth data points during statistical modelling at locations where all explanatory angular variables are not represented (i.e., the listwise deletion of partially complete observations; van Buuren 2018). Additionally, we propose to use the difference between measured depths from different acoustic frequencies at particular incidence angles as a proxy for substrate penetration to inform the imputation procedure and increase capacity to map sediment properties (Gaida et al. 2020). To achieve full spatial continuity for angular intensity and depth difference data, aggregation is still required at some level to produce map units that accommodate observations at multiple angles of incidence simultaneously. We investigate the methods proposed by Che Hasan et al. (2012, 2014), wherein discontinuous angular measurements are aggregated according to object-based image segments expected to represent "acoustic themes" (i.e., homogenous substrate units; Fonseca et al. 2009). This method facilitates the co-location of angular response data and can help to initially minimize the amount of missing information per observation. We hypothesize an optimal segmentation scale that is fine enough to delimit substrate boundaries at the scale of interest, but general enough to allow for the simultaneous observation at a range of incidence angles. The goals of this paper are:

(i) to evaluate the effectiveness of multiple imputation for leveraging inter-angular and inter-frequency multicollinearity of MBES data to produce realistic estimates of missing angular acoustic values;
(ii) to evaluate the usefulness of imputed angular response and depth difference data for benthic habitat and substrate mapping at a high spatial resolution.

Methods

Multi-frequency MBES data

The Bedford Basin 2017 MBES dataset was selected to test the utility of imputation methods for multi-frequency angular mapping. This dataset is well-established following its use in the 2017 R2Sonic Multispectral Challenge (e.g., Buscombe and Grams 2018; Gaida et al. 2018; Costa 2019) and is useful as a benchmark for methodological investigations. Brown et al. (2019) provide a detailed description of the data acquisition. Briefly, the survey was conducted on 2 May 2017 to map a heterogeneous patch of seabed that has been studied previously at the Bedford Basin, Halifax, Canada (Fader and Miller 2008). A pole mounted R2Sonic 2026 MBES with a Valeport sound velocity probe and POS MV Wave Master Inertial Navigation System with two Trimble GPS antennas was deployed port-side of a 12 m vessel, and soundings were conducted at 100, 200, and 400 kHz operating frequencies. Sound velocity profiles (SVP) were obtained during the survey using an AML Base X2, and all data were integrated during acquisition using QPS QINSy. The full data processing workflow for this project is summarized in Fig. 1.

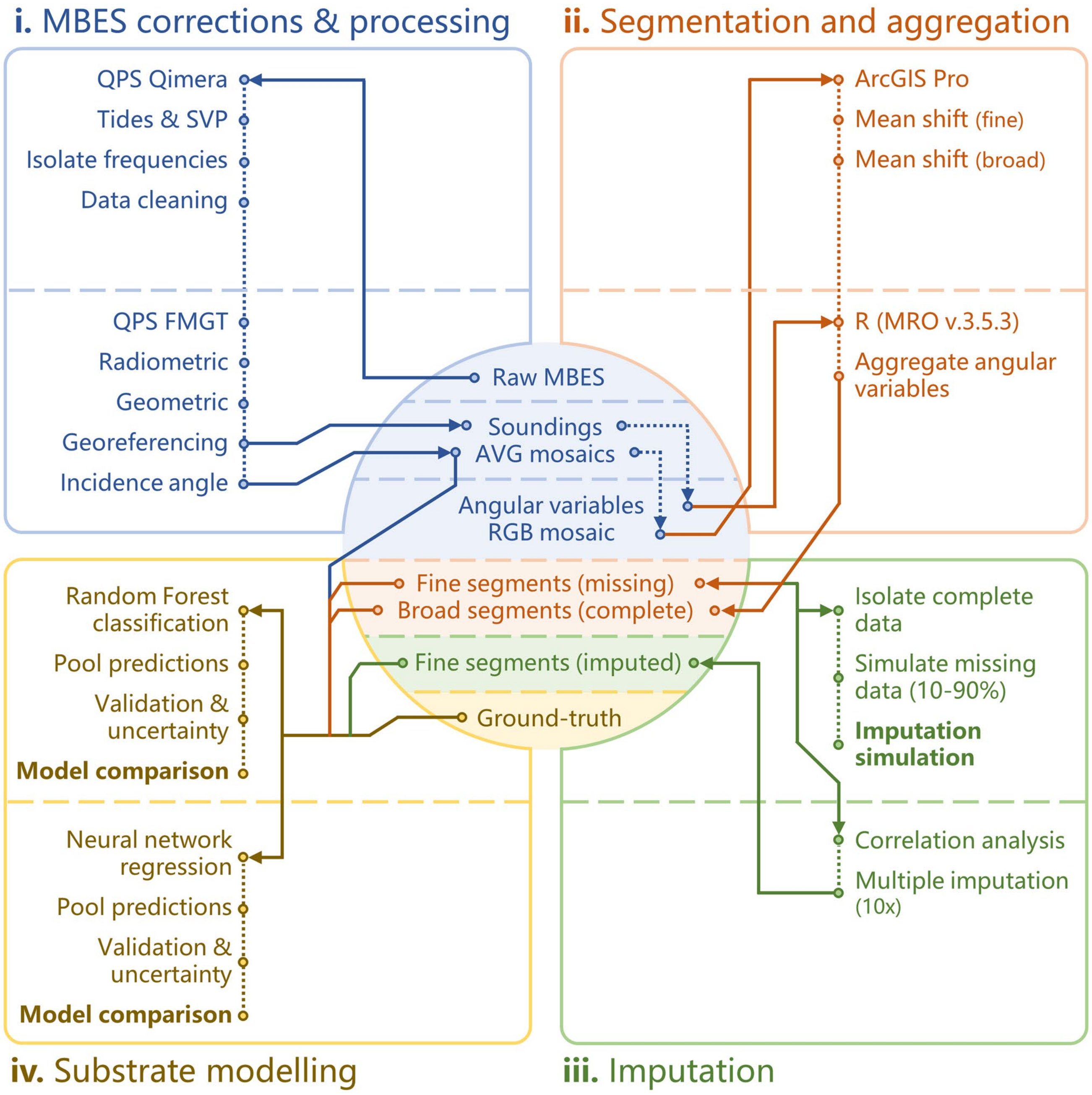


Fig. 1 Summary of methods used to evaluate the effectiveness of multiple imputation for analysis of angular acoustic data. Data are summarized in the center; analyses and processes are on the periphery (i–iv). Solid lines represent inputs and outputs; dashed lines represent intermediate processing

The QPS suite was used to reprocess the multi-frequency MBES data for the purposes of this study. The raw data were loaded into Qimera with corresponding tidal and SVP measurements acquired during the survey. Tidal and SVP corrections were applied to the data, followed by filters to isolate soundings from each MBES frequency in turn (100, 200, 400 kHz). For each set of soundings, spline filters were used to reject erroneous data, followed by minimal manual cleaning of residual artefacts. Cleaned and corrected soundings were exported as.GSF files along with gridded bathymetric surfaces, yielding three single-frequency datasets.

The Fledermaus Geocoder Toolbox (FMGT) was used to apply standard radiometric corrections to the raw MBES backscatter intensity. Cleaned soundings for each of the frequencies were imported into FMGT via the.GSF files output by Qimera, and the gridded bathymetric layers were added as reference grids for slope correction. The R2Sonic 2026 sonar defaults were retained within the software, and absorption coefficients were estimated for each frequency using calculations provided by the National Physical Laboratory (2018), as recommended in the current FMGT software documentation. Default settings were used for all other geometric and radiometric corrections. The corrected soundings without AVG (i.e., level BL_3; Schimel et al. 2018; Malik et al. 2019) were exported as ASCII files, and AVG-compensated mosaics (i.e., BL_4) were also generated using the "flat" algorithm, referencing the mean of the angular interval between 30–60° with a moving window of 300 pings. The echosounder was not calibrated prior to the survey; all backscatter values output from processing were on a relative dB scale.

Backscatter data segmentation and aggregation

Angular acoustic measurements were derived over the extent of the MBES coverage. The ASCII files output from FMGT were parsed by the angle of incidence at the seafloor with a custom R function to achieve a point representation of each sounding with depth and backscatter intensity attributes for each frequency. The soundings were aggregated to 3° bins to acquire the backscatter intensity variables $BS_f(\theta)$ for all frequencies $f = 100, 200, 400$, and angle bins $\theta = [10, 12], [13, 15], \ldots, [55, 57]$. The difference in measured depth, $\Delta d_{ff'}(\theta)$, was calculated between each pair of frequencies f and f' for each angle bin θ by aggregating the depth soundings to a 2-m grid and subtracting co-located soundings of disparate frequency but like angle. These variables have a high angular (3°) and spatial (2 m grid) resolution, but poor continuity across the dataset resulting from insufficient angular coverage.

In order to co-locate the angular frequency-dependent variables that do not occur in the same grid cell, and to increase their spatial coverage, the acoustic data were aggregated to seabed segments using the approach suggested by Che Hasan et al. (2012; modified from Fonseca et al. 2009). Seabed segments were achieved by segmenting a three-band RGB raster containing AVG-compensated backscatter mosaics for each frequency using the mean shift algorithm (Comaniciu and Meer 2002) in ArcGIS Pro. Behaviour of the mean shift in ArcGIS Pro is controlled by manipulating parameters that balance the "spatial detail" (segmentation weighting based on proximity), "spectral detail" (weighting based on attribute values), and minimum or maximum segment size, which are used to merge or split segments at a given number of raster cells. Here, spatial and spectral detail were set to their maximum values and the minimum segment size was used to manipulate the segmentation outcome. The product was a continuous set of segmented units over the study extent (termed "acoustic segments"). The angular frequency-dependent data were then averaged at each acoustic segment (Fig. 2). The average backscatter intensity of frequency f and angle bin θ for a given segment was obtained by first averaging the values of all soundings $i = 1, 2, \ldots, n$ within a given raster cell, then averaging the values of all cells $j = 1, 2, \ldots, m$ within an acoustic segment:

$$BS_f(\theta) = \frac{1}{m_f(\theta)} \sum_{j=1}^{m_f(\theta)} \left(\frac{1}{n_f(\theta)_j} \sum_{i=1}^{n_f(\theta)_j} BS_f(\theta)_{ij} \right) \tag{1}$$

Similarly, the average depth difference between frequencies f and f' at angle θ was calculated for a given segment by first averaging the depth soundings over each raster cell where data exists, then subtracting depth values of differing frequency at co-located cells. The results were averaged over the acoustic segment:

$$\Delta d_{ff'}(\theta) = \frac{1}{m_{ff'}(\theta)} \sum_{j=1}^{m_{ff'}(\theta)} \left[\left(\frac{1}{n_f(\theta)_j} \sum_{i=1}^{n_f(\theta)_j} d_f(\theta)_{ij} \right) - \left(\frac{1}{n_{f'}(\theta)_j} \sum_{i=1}^{n_{f'}(\theta)_j} d_{f'}(\theta)_{ij} \right) \right] \tag{2}$$

Given the three frequencies, 16 angle bins, and two measurements (backscatter and depth difference), this yielded 96 attributes for each acoustic segment of the dataset.

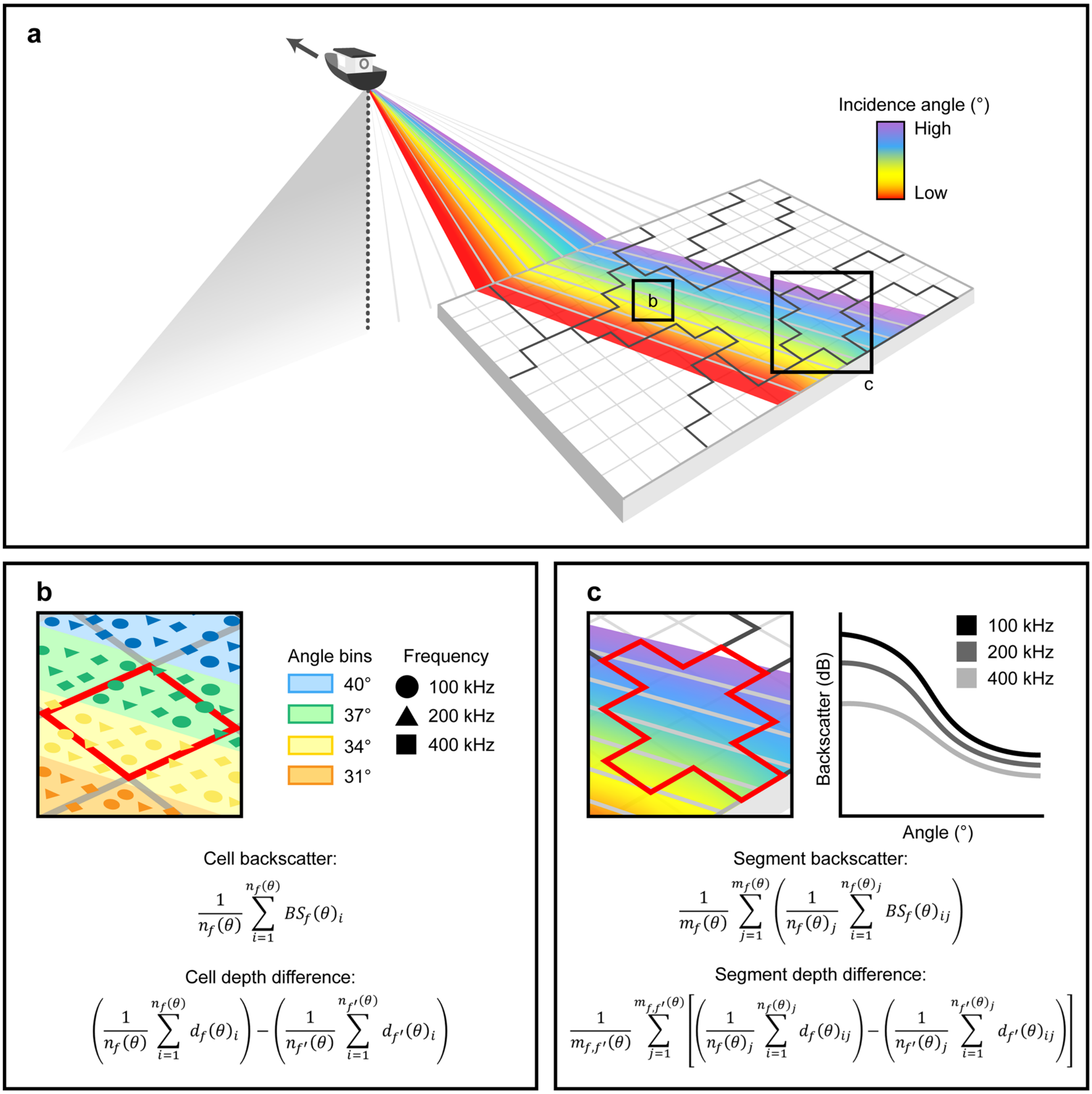

a
Incidence angle (°)
High
Low
b
c
b
Angle bins
40°
37°
34°
31°
Frequency
100 kHz
200 kHz
400 kHz
Cell backscatter:
$\frac{1}{n_f(\theta)}\sum_{i=1}^{n_f(\theta)} BS_f(\theta)_i$
Cell depth difference:
$\left(\frac{1}{n_f(\theta)}\sum_{i=1}^{n_f(\theta)} d_f(\theta)_i\right)-\left(\frac{1}{n_{f'}(\theta)}\sum_{i=1}^{n_{f'}(\theta)} d_{f'}(\theta)_i\right)$
c
100 kHz
200 kHz
400 kHz
Backscatter (dB)
Angle (°)
Segment backscatter:
$\frac{1}{m_f(\theta)}\sum_{j=1}^{m_f(\theta)}\left(\frac{1}{n_f(\theta)_j}\sum_{i=1}^{n_f(\theta)_j} BS_f(\theta)_{ij}\right)$
Segment depth difference:
$\frac{1}{m_{f,f'}(\theta)}\sum_{j=1}^{m_{f,f'}(\theta)}\left[\left(\frac{1}{n_f(\theta)_j}\sum_{i=1}^{n_f(\theta)_j} d_f(\theta)_{ij}\right)-\left(\frac{1}{n_{f'}(\theta)_j}\sum_{i=1}^{n_{f'}(\theta)_j} d_{f'}(\theta)_{ij}\right)\right]$

Fig. 2 **a** Swath multibeam soundings were binned according to the angle of incidence and overlain on a segmented raster grid (segment boundaries in black). **b** The backscatter and depth difference of binned angular soundings were aggregated per 2-m raster cell (example cell in red) for three operating frequencies. **c** Cell values were then aggregated according to acoustic segments (example in red) to achieve estimates over the full range of angles for a homogenous patch of seabed. Missing data may occur where a segment does not contain data at all incidence angles

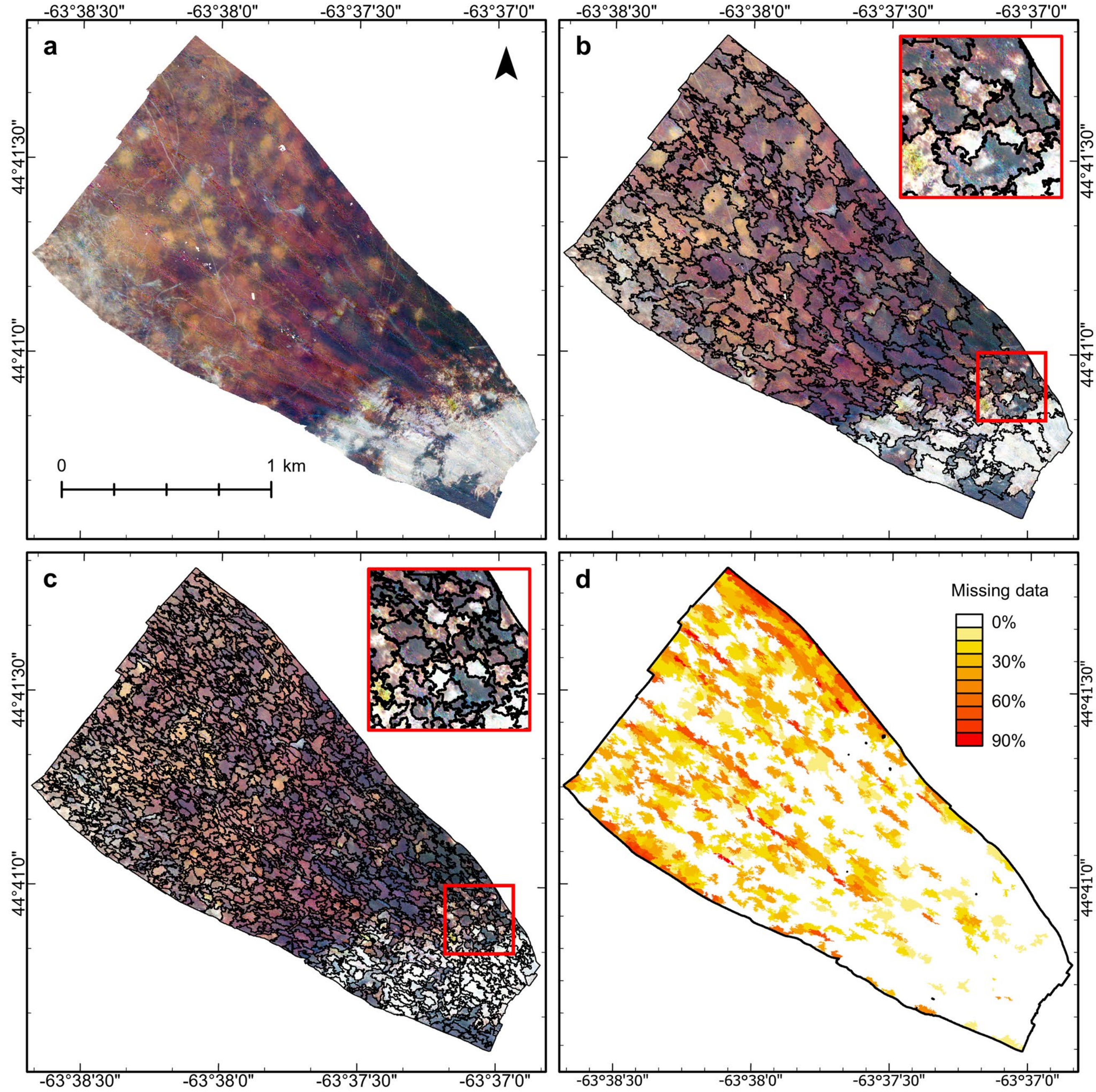


Fig. 3 **a** RGB composite of multi-frequency backscatter ($BL_{4,100}$, $BL_{4,200}$, $BL_{4,400}$) segmented at **b** a broad scale with no missing data, and **c** a fine scale that captures the substrate heterogeneity, which produces **d** acoustic segments with missing data

To explore the trade-off between spatial resolution and data continuity for benthic mapping at the Bedford Basin, two segmentations were generated:

(i) the three-band backscatter mosaic was segmented at a scale broad enough (minimum mean shift segment size 7000 m^2) to yield segments with no missing angular data (i.e., all segments included soundings at all frequencies and incidence angles; Fig. 3b);
(ii) segmentation was performed at the finest scale necessary to capture the apparent local heterogeneity in substrate patches visible in the multi-band backscatter mosaic, allowing missing data to occur (minimum mean shift segment size 1000 m^2; Fig. 3c, d).

Because the former segmentation generated acoustic segments of considerably lower detail than the latter, it is necessary to evaluate the trade-off between dataset continuity and spatial resolution. For comparison, the AVG-compensated backscatter mosaics for each frequency (hereafter $BL_{4,100}$, $BL_{4,200}$, $BL_{4,400}$) and average depth (between all frequencies; $\overline{D}$) were also extracted for each segment.

Imputation

Missing angular data resulting from the fine scale segmentation were completed using multiple imputation. "Multiple imputation" encompasses the entire workflow for achieving statistical inference using an incomplete dataset—from the estimation of missing values to statistical model fitting. The full multiple imputation workflow can be divided into three parts (Rubin 1987; van Buuren 2007):

(i) Multiple versions of the complete dataset are generated by predicting the missing values using statistical relationships with other correlated variables in the dataset. The non-missing data are identical between versions of the full dataset, but the imputed values differ as a function of the imputation models, which incorporates variability in the prediction of missing values (van Buuren 2018).
(ii) Each version of the complete imputed dataset is used to fit the statistical model of interest, resulting in multiple independent analyses and models, which differ according to variability among the imputed data.
(iii) The models are pooled to estimate the parameter(s) of interest, including estimates of uncertainty arising from multiple versions of the imputed dataset. Uncertainty in the modelled parameters is a function of uncertainty regarding the missing values.

The motivation for completing multiple versions of the dataset and analysis is that the single imputation of a value that was missing underestimates uncertainty regarding what the missing value should be. Rubin (1978) noted that, "imputing one value for a missing datum cannot be correct in general, because we don't know what value to impute with certainty (if we did, it wouldn't be missing)". A number of methods exist to introduce proper amounts of variability in the imputed data.

An effective approach to generating multiple multivariate imputations (step i) above) is Fully Conditional Specification (FCS; van Buuren et al. 2006). FCS proceeds iteratively by specifying an imputation model independently for each variable in the dataset with missing values, using the other covariates as predictors. There are many potential imputation models, but generally, these should produce imputed values with an appropriate amount of variability given that which is observed for each variable in the dataset. Multiple Imputation by Chained Equations (MICE) is a Markov Chain Monte Carlo method for FCS—specifically, a Gibbs sampler (van Buuren and Groothuis-Oudshoorn 2011; van Buuren 2018). MICE works by first completing missing data for each of q predictors in the dataset (here, the 96 acoustic attributes at each acoustic segment) using random draws from the observed values as "placeholders" (Azur et al. 2011). Starting with the first of the q variables, β_1 of β_q, the "placeholder" values are removed, and a model is fit between the remaining observed values for β_1 and the other predictors in the dataset, $\beta_2, \ldots, \beta_q$. The missing values are then predicted for β_1 using the model, which also incorporates uncertainty in the prediction using one of several means (e.g., random draws from the data, bootstrapping, Bayesian regression). The variable β_1 has now been updated and the algorithm proceeds with β_2, using the other synthetically complete variables in the dataset, including those resulting from prediction in the previous steps. The entire process of cycling through all q variables is repeated for a set number of iterations for the algorithm to converge (often ~ 10), and the final version of the dataset is retained as one of multiple imputations. Additional details on the specifics of these steps are provided by Azur et al. (2011), van Buuren and Groothuis-Oudshoorn (2011), and van Buuren (2018).

Here, several imputation models were trialed for the prediction step, yet the implementation of Random Forest multiple imputation in the R package 'mice' (van Buuren and Groothuis-Oudshoorn 2011) generally produced imputations that converged quicker, with lower error, than other methods such as predictive mean matching (PMM), Bayesian linear regression, and linear regression using bootstrapping. Random Forest was therefore selected for all imputations presented hereafter. The ability of Random Forest to automatically model variable interactions has been highlighted as one of its strengths for imputation

(Doove et al. 2014). This is expected to be useful in the present case of highly dimensional multi-frequency MBES data. This form of imputation proceeds not by simply predicting missing values using Random Forest, but by using the trees that comprise a Random Forest model to identify “donor” values in a manner similar to PMM (van Buuren 2018). Treating the variable being imputed as the response and the other covariates as predictors, multiple decision trees are grown on bootstrap samples of the dataset according to established Random Forest methods (Breiman 2001). Imputed values are generated by determining the terminal leaf of each decision tree to which the missing values belong according to the tree splits, then randomly selecting from among the observed values at those leaves, which comprise the potential “donors” (Doove et al. 2014). This has the effect of injecting multiple elements of stochasticity into the imputations, while retaining the exclusive selection of “real” values that are drawn from the dataset. Conceptually, this is a non-parametric and non-linear method for generating candidate donor values rather than a predictive model, and relatively few trees are required for each imputation (e.g., 10). Doove et al. (2014) provide a detailed description of the algorithm.

Imputation simulation

To inform on the circumstances under which the methods presented here may be tenable, simulation was used to explore how accurately the angular frequency-dependent backscatter ($BS_f(\theta)$) and depth difference ($\Delta d_{ff'}(\theta)$) measurements can be imputed given variable amounts of missing data. First, all segments resulting from the mean shift algorithm with complete observations at all incidence angles using each frequency ($n = 564$) were isolated to produce a complete dataset (i.e., with no missing data). Missing data were synthetically generated for this dataset by randomly dropping observations within the segments to achieve versions of the dataset with between 10 and 90% of data missing. To realistically simulate the conditions under which missing data may occur using a swath sonar system, the missing data generator algorithm proceeded as follows:

(i) select an acoustic segment at random from the complete dataset;

(ii) for that segment, randomly select one of the available incidence angle bins ($\theta = [10, 12], [13, 15], \ldots, [55, 57]$);

(iii) drop all observations of variables that were measured at the angle bin selected in (ii).

Given the two variables measured at three operating frequencies, the above produces six missing values per iteration. At the end of each iteration, the cumulative proportion of missing data is calculated and the algorithm proceeds until the desired amount of missing data is achieved. Because multiple variables are measured at a given angle-dependent sounding, all variables at a given angle are likely to be either present or missing for each observation, and this method of missing data generation is therefore expected to be more realistic than a random draw.

For each simulated incomplete dataset (10–90% missing data) the missing data were imputed and compared to the actual measured values that were dropped. Ten Random Forest imputations were conducted for each simulation using the default Random Forest parameters in the ‘mice’ package for ten iterations. The mean absolute error (MAE) and variance explained (VE) between multiple imputed values and the real values were calculated to determine the absolute and relative average accuracies of the imputations given increasing amounts of missing data.

Full dataset imputation

Following simulation, the missing angular measurements of all fine scale segments ($n = 1087$) were imputed to produce ten plausible hypotheses of the full dataset. Imputation of the full dataset was performed using the same imputation design as the simulations in "Imputation simulation" section (ten imputations using Random Forest for ten iterations). Using the ‘mice’ package, it is possible to specify a unique set of predictors for each variable undergoing imputation. Random Forest is expected to be robust to uninformative predictors (i.e., they are not selected for tree splitting), yet predictors of a given variable were additionally omitted if the absolute Pearson correlation between the variable and predictor was below 0.1, or if the predictor was present in less than 10% of observations of the variable being imputed. All angular variables at all frequencies were included in the imputation procedure, along with the average water depth measured by all frequencies ($\overline{D}$). The imputation algorithm was run for ten iterations, after which convergence was checked by observing the mean and standard deviation of imputed values plotted against the iteration number, following van Buuren and Groothuis-Oudshoorn (2011). The result was ten versions of the imputed dataset with complete observations of variables using each operating frequency at each incidence angle bin for both backscatter intensity ($BS_f(\theta)$), and depth difference ($\Delta d_{ff'}(\theta)$).

Substrate modelling

The utility of the imputed data for habitat mapping applications was evaluated using empirical modelling. Georeferenced ground truth photographs and video observations

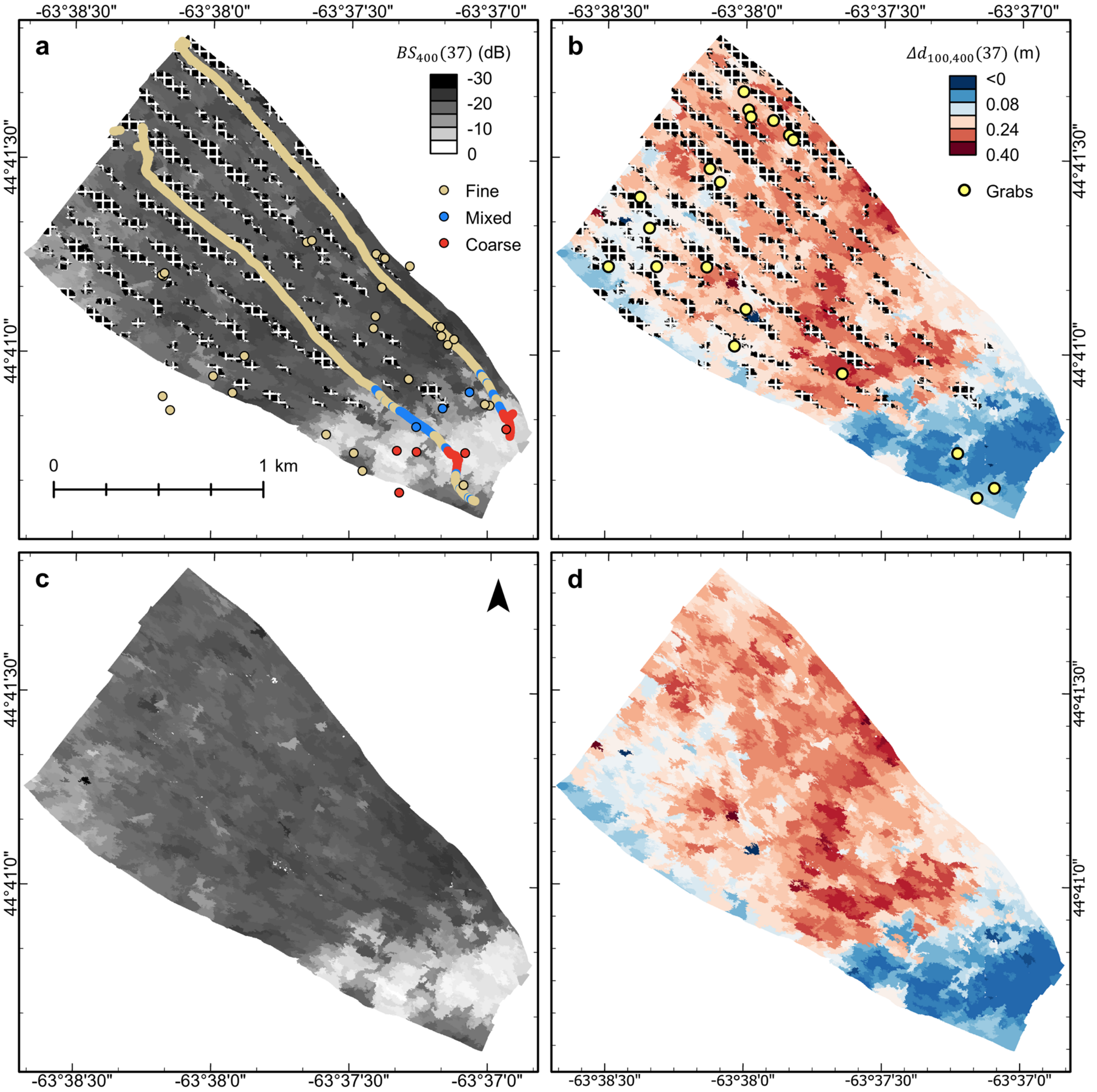


Fig. 4 **a** 400 kHz backscatter intensity for each segment measured at the 37° incidence angle bin ($BS_{400}(37)$) with observed substrate classes from seafloor images and video, and **b** depth difference between 100 and 400 kHz soundings for each segment measured at the 37° incidence angle bin ($\Delta d_{100,400}(37)$) with grab samples. Segments with missing data for the 37° angle bin are shown as hatched areas and were imputed to produce full coverage observations for **c** $BS_{400}(37)$ and **d** $\Delta d_{100,400}(37)$

Table 1 Configurations of acoustic data, segmentation scale, and missing data compared via empirical modelling

Segmentation scale	Acoustic data treatment	Predictors	Missing data
Fine	AVG	$BL_{4,100}+ BL_{\underline{4},200} + BL_{4,400} + D$	No
Fine	Angular	$BS_f(\theta)+ \Delta d_{f,f'}(\theta) + \overline{D}$	Yes
Broad	Angular	$BS_f(\theta)+ \Delta d_{f,f'}(\theta) + \overline{D}$	No
Fine	Angular	$BS_f(\theta)+ \Delta d_{f,f'}(\theta) + \overline{D}$	No (imputed)

collected in 2017 and 2018, described by Brown et al. (2019), were examined in the context of the angular frequency-dependent data (Fig. 4). Where a segment had multiple ground truth observations, the most common class was assigned (producing $n = 171$ ground truth segments). Random Forest models predicting the bottom class using the angular frequency-dependent predictors were trained using the 'randomForest' package (Liaw and Wiener 2002) in the Microsoft R Open version of R (R Core Team 2019). Models were built using 500 trees, default hyperparameters, and no variable reduction. For comparison, several models were generated using various configurations of the acoustic predictors and segmented backscatter layers: (i) fine scale segmentation with AVG compensated frequency-dependent backscatter values and depth; (ii) fine scale segmentation with angular frequency-dependent predictors and depth (with missing data); (iii) broad scale segmentation with angular frequency-dependent predictors and depth; and (iv) fine scale segmentation with imputed angular frequency-dependent predictors and depth (Table 1).

Model predictive performance and uncertainty were evaluated to explore the effects of multiple imputation on classification results. Predictive performance was quantified using the classification accuracy and kappa values calculated from the out-of-bag samples. The votes from the Random Forest represent a discrete event X for each acoustic segment, with three possible outcomes, $\{x_1, x_2, x_3\}$ that represent the ground truth classes. The predictive uncertainty from the ten different models can therefore be represented succinctly for each acoustic segment using their entropy,

$$H(X) = -\sum_{i=1}^{3} P(x_i) \log P(x_i), \tag{3}$$

where $P(x_i)$ is the proportion of votes (i.e., the probability) for each class. Lower values of H represent increasingly unanimous agreement among models and high values indicates low agreement. Entropy values were mapped along with the predicted classes for each acoustic segment to convey predictive uncertainty spatially. The relationship between the proportion of missing data, which were imputed, and the entropy at each segment was also investigated. To control for heterogeneity among the substrate classes, entropy was modelled as a function of the proportion of missing data and also the predicted class using a generalized linear model (GLM). Entropy is a positive and continuous variable, but here contained a preponderance of zero values where model agreement was unanimous, which occurred exclusively for the "fine" sediment class. A hurdle approach was used to partition the model into two parts: (i) the probability that entropy is not zero, and ii) the entropy value, conditional on it not being zero (Cragg 1971; Potts and Elith 2006). The first model was a binomial GLM with a log link function to predict the probability of non-zero entropy as a function of the proportion of missing data and the predicted substrate class. The second was a GLM with a gamma error distribution and identity link to predict non-zero entropy values using missing data proportion and substrate class. Interaction was tested between all predictors.

The capacity for modelling continuous substrate properties was also evaluated. Van Veen grabs obtained from the site in 2018 ($n = 19$; Brown et al. 2019) were analyzed for sediment grain size, providing particle size distributions. The arithmetic mean grain size ($\overline{x}_a$) and sorting (σ_a) in µm were calculated for each sample using GRADISTAT (Blott and Pye 2001) and were assigned to their overlapping acoustic segments (Fig. 4). Initial trials suggested that neural networks outperformed Random Forest at modelling these continuous substrate properties. Neural networks predicting $\overline{x}_a$ and σ_a using each configuration of the acoustic data (Table 1) were trained using Keras TensorFlow in Python 3.7. Both models comprised two dense layers of 128 and 64 units using the rectified linear unit (ReLU) activation function, followed by a single-unit dense output layer with linear activation. Dropout was implemented between all dense layers at a rate of 0.3. The models were optimized using the Adam algorithm (Kingma and Ba 2017) with a mean squared error (MSE) loss function and were trained for 600 epochs using the full batch size. Leave-one-out cross validation (LOO CV) was used to obtain estimates of model performance given the small sample size. Pearson's correlation (r), variance explained (VE), and mean absolute percent error (MAPE) were calculated between the predicted and omitted test values, and were compared between all configurations of the acoustic data (Table 1).

The effects of missing data on the uncertainty of grain size parameter predictions were explored using GLMs. Prediction variability, as measured by the standard deviation, was modelled against missing data proportion for each parameter ($\overline{x}_a$ and σ_a). To control for apparent increases in uncertainty at higher predicted values, the

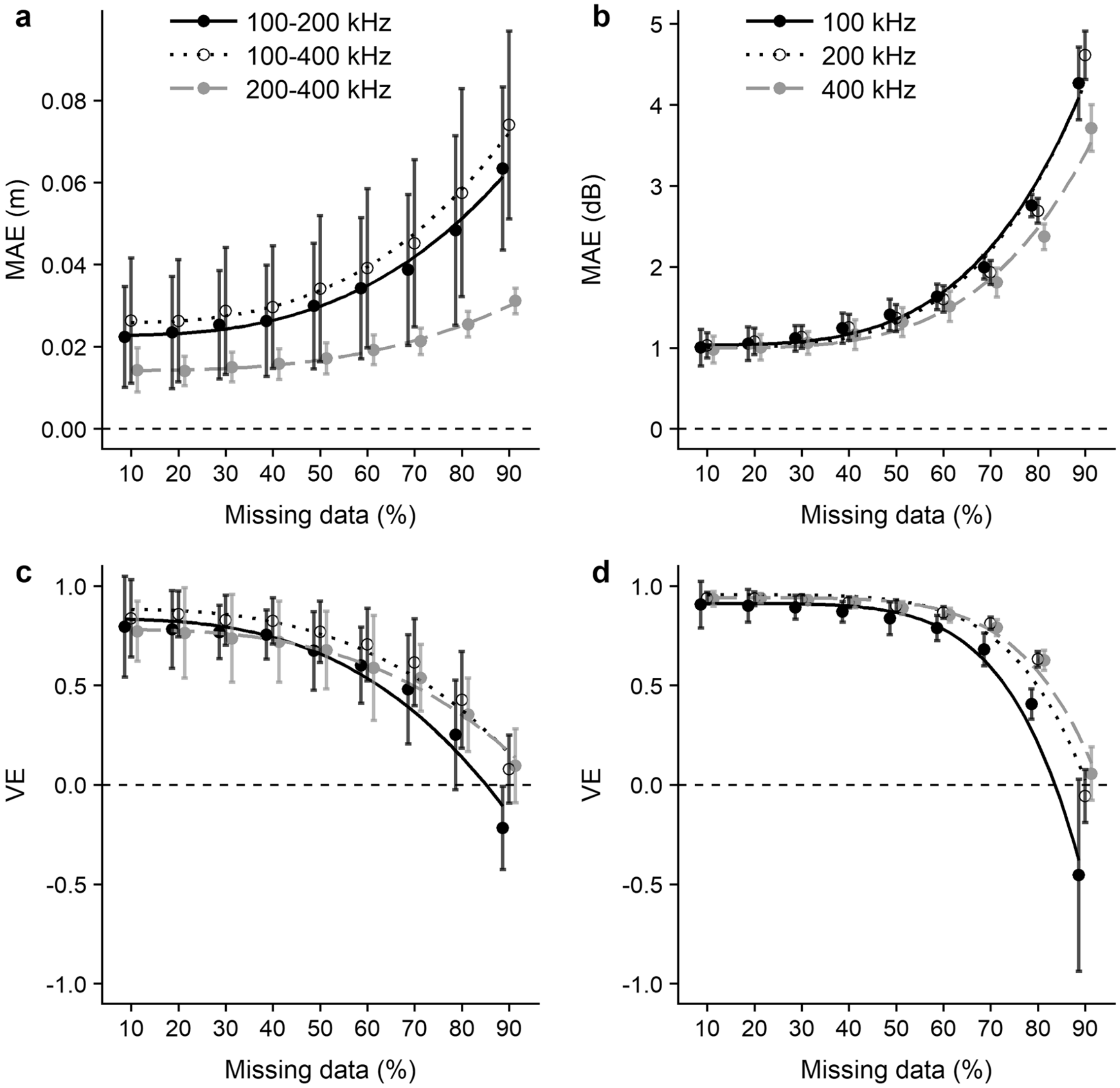
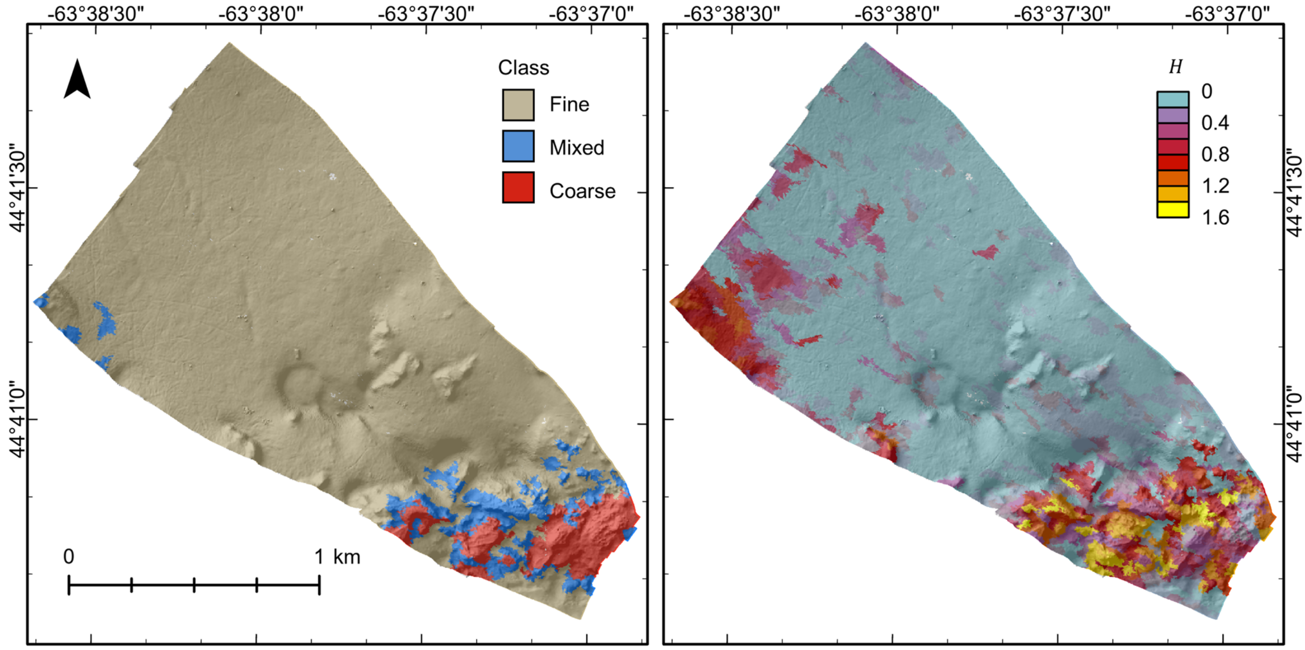


Fig. 5 Mean absolute error **a** and **b** and variance explained **c** and **d** ± 1 SD between simulated missing data and imputed data for angular- and frequency-dependent depth differences (left) and backscatter intensity (right). Lines are fitted exponential curves

predicted parameters at each segment were also included in the model. The standard deviation of predictions (the response) was continuous and greater than zero, yet non-linear relationships were observed with the predictors. Initial models also suggested heteroskedasticity of residuals. GLMs were fit with a gamma error distribution and a log link function, and interaction was tested between all predictors.

Table 2 Out-of-bag Random Forest classification performance using different methods of acoustic data treatment and scale

Acoustic data	Segmentation scale	*n*	Accuracy (± 1 SD)	Kappa (*k*; ± 1 SD)
AVG	Fine	171	0.9374	0.7398
Angular	Fine (missing)	117	0.9487	0.8414
Angular	Broad	71	0.9014	0.6401
Angular	Fine (imputed)	171	0.9538 ± 0.0033	0.8224 ± 0.0110



	Reference			
Prediction	Fine	Mixed	Coarse	UA
Fine	142	3	3	96%
Mixed	3	9	0	75%
Coarse	2	0	12	86%
PA	97%	75%	80%	**95% ($k = 0.82$)**

Model	Coefficient	Value	P
Binomial (log link)	β_0	1.757	< 0.001
	β_{miss}	7.482	< 0.001
Gamma (identity link)	β_0	0.162	< 0.001
	β_{miss}	0.028	0.417
	β_{mixed}	0.928	< 0.001
	β_{coarse}	0.706	< 0.001

Fig. 6 Pooled Random Forest predicted substrate classes and OOB confusion matrix (left), and prediction entropy with hurdle model coefficients (right). Note that the gamma coefficient for fine sediment is the reference level for the predicted sediment factor; coefficients β_{mixed} and β_{coarse} are offsets relative to β_{fine}. Hill shade is from the bathymetric raster

Results

Imputation

Exploratory analysis suggested strong multicollinearity among angular frequency-dependent variables. In short, backscatter measurements were strongly correlated across incidence angles and frequencies. Backscatter measurements were correlated to the depth differences between frequencies at non-oblique angles (e.g., <46°), but were generally uncorrelated with these measurements at increasingly oblique angles. Depth differences were also intercorrelated, with lower correlations at increasingly oblique angles. Detailed analysis of the correlation matrix is provided in Online Resource S1.

Imputation simulation

Given the observed correlations between angular frequency-dependent variables, missing data were simulated to determine what proportions of missing data may be reasonable to impute for a multi-frequency MBES dataset. Results suggested that the error of depth difference ($\Delta d_{f,f'}(\theta)$) imputation tends to accelerate past > 50% missing data (Fig. 5). The low MAE of $\Delta d_{200,400}(\theta)$ imputations, even at high levels of missing data, contrasts with the drop in VE, suggesting the $\Delta d_{200,400}(\theta)$ variables may have comparatively low amounts of variance. The error of angular backscatter ($BS_f(\theta)$) imputation accelerates at levels > 60% missing data, which appears as an elbow in plots of error against the proportion of missing data (Fig. 5). These are likely conservative estimates given the reduced simulation sample size using complete cases ($n = 564$) compared to the full dataset ($n = 1087$).

Table 3 Neural network regression LOO CV performance using different methods of acoustic data treatment and scale

			Mean grain size ($\bar{x}_a$)		
Acoustic data	Segmentation scale	*n*	Pearson *r* (±1 SD)	VE (±1 SD)	MAPE (±1 SD)
AVG	Fine	19	0.7178	0.5153	0.2169
Angular	Fine (missing)	7	0.9305	0.8546	0.1994
Angular	Broad	15	0.8205	0.6731	0.2185
Angular	Fine (imputed)	19	0.8596 ± 0.0169	0.7378 ± 0.0302	0.1688 ± 0.0112
			Sorting (σ_a)		
AVG	Fine	19	0.7757	0.5918	0.3273
Angular	Fine (missing)	7	0.8453	0.6929	0.387
Angular	Broad	15	0.8327	0.6926	0.2853
Angular	Fine (imputed)	19	0.8798 ± 0.0093	0.7737 ± 0.0162	0.2594 ± 0.0149

Full dataset imputation

The proportion of missing data resulting from the fine scale segmentation ("Backscatter data segmentation and aggregation" section) was 14.05%. Results from "Imputation simulation" section suggest this is a reasonable proportion of missing data to impute for both sets of angular frequency-dependent variables ($\Delta d_{f,f'}(\theta)$; $BS_f(\theta)$). Multiple Imputation by Chained Equations was performed for ten iterations to produce ten versions of the complete dataset.

Substrate modelling

Random Forest classification of the bottom type observed in ground truth imagery produced different results depending on the configuration of the acoustic predictor data. The fine scale segmentations using angular data (missing and imputed) produced the most accurate classification results (Table 2)—the differences between these two models are that (i) imputation makes available more ground truth samples for model training ($n = 171$ vs. $n = 117$ acoustic segments), and (ii) model predictions using imputation span the full extent of the study area, while those without imputation leave missing data areas unclassified. The segmentation that was sufficiently broad to yield no missing angular data was too coarse to capture the full heterogeneity of bottom types and produced the least accurate model. The compensated acoustic data with a fine segmentation (and no missing data) was less successful than the fine scale angular methods at predicting rarer classes, as indicated by the kappa score.

The hurdle model provided insight into the effects that the missing data, which were imputed, had on the predictive uncertainty (i.e., entropy). The fitted binomial model was:

$$\ln\left(\frac{p(H > 0)}{1 - p(H > 0)}\right) = \beta_0 + \beta_{miss}miss + \varepsilon, \tag{4}$$

where $p(H > 0)$ is the probability of non-zero entropy, and *miss* is a predictor representing the proportion of missing data. The model suggested that the proportion of missing data was a significant predictor of non-zero entropy ($P < 0.001$; Fig. 6), but we note, again, that the zero values occurred exclusively in the "fine" class (precluding the inclusion of predicted class in the model). Back-transforming predicted values from the logistic model predicts a ~ 15% increase in the probability of non-zero entropy between 0 and 100% missing data.

The fitted gamma model for non-zero entropy was:

$$\ln(H) = \beta_0 + \beta_{miss}miss + \beta_{mixed}mixed + \beta_{coarse}coarse + \varepsilon, \tag{5}$$

where *mixed* and *coarse* are levels of the predicted categorical substrate class (with the "fine" class acting as reference). Results from this model suggested that non-zero entropy was highly dependent on the predicted substrate class ($P < 0.001$), but that the proportion of missing data had no significant effect ($P = 0.417$). Interaction terms between the proportion of missing data and predicted substrate class were non-significant and were dropped. Additional model analysis is provided in Online Resource S2.

Prediction of continuous substrate properties was generally most successful using the angular frequency-dependent predictors. These outperformed the compensated backscatter predictors in all cases, even when using broader acoustic segments (Table 3). The broader segmentation resulted in the aggregation of four grab samples at two of the acoustic segments, reducing the sample size to $n = 15$. Predictions using the angular variables at a fine segment scale with missing data were highly accurate, particularly for the mean grain size predictions, yet the occurrence of missing predictors reduced the sample size by more than half ($n = 7$), bringing into question the representativeness of these results. The angular imputed datasets performed well on average with the lowest error in all cases (according to the MAPE), and the highest correlation and VE scores amongst the sorting models.

GLMs provided insight into the effects of missing data on the uncertainty for the two grain size parameters. For

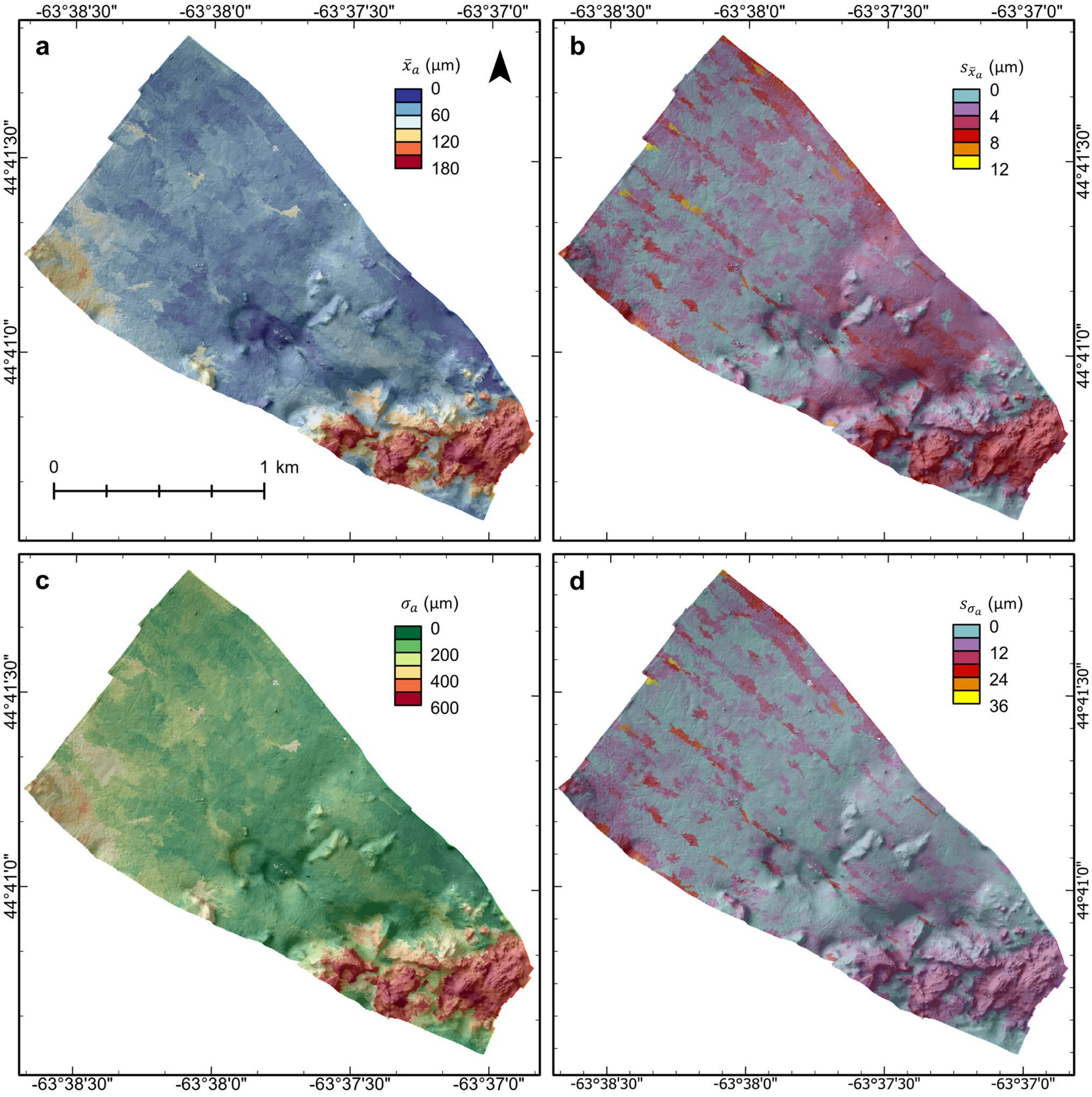


Fig. 7 Pooled neural network mean grain size predictions **a** with the standard deviation of predictions **b**, and sorting predictions **c** and standard deviation **d** from multiple imputation data at the Bedford Basin. Hill shade is from the bathymetric raster

Table 4 GLM model coefficients for mean grain size and sorting standard deviation of predictions ($s_{\bar{x}_a}$, s_{σ_a}). Coefficient values are provided according to the link, and also back-transformed to the scale of the response for interpretation

Model	Coefficient	Value	Back-transform (e^{β})	*P*
$s_{\bar{x}_a}$	β_0	0.3981	1.4890	< 0.001
Gamma	$\beta_{\bar{x}_a}$	0.0070	1.0070	< 0.001
(log link)	β_{miss}	1.2541	3.5046	< 0.001
s_{σ_a}	β_0	1.0747	2.9290	< 0.001
Gamma	β_{σ_a}	0.0022	1.0022	< 0.001
(log link)	β_{miss}	1.8227	6.1886	< 0.001

both predictions, increased uncertainty was apparent in the along-track direction of the survey (NW–SE), corresponding to acoustic segments with high amounts of missing angular data (Fig. 3), which can be conveyed spatially along with the pooled model predictions (Fig. 7). The gamma GLMs fitted to investigate the effects of missing data on the mean grain size and sorting prediction variability were:

$$\ln\left(s_{\bar{x}_a}\right) = \beta_0 + \beta_{\bar{x}_a}\bar{x}_a + \beta_{miss}miss + \varepsilon \tag{6}$$

and

$$\ln\left(s_{\sigma_a}\right) = \beta_0 + \beta_{\sigma_a}\sigma_a + \beta_{miss}miss + \varepsilon. \tag{7}$$

The coefficient for missing data, β_{miss}, suggested an increase in the standard deviation of predictions by a factor of 3.50 per unit for mean grain size, which is a factor of ~ 1.13 per 10% missing data. Similarly, β_{miss} for sorting suggested an increase in the standard deviation of predictions by a factor of 6.19 per unit, or ~ 1.20 per 10% missing data. Areas of finer predicted mean grain size appeared to have lower uncertainty than coarse areas (e.g., $\bar{x}_a > 120$ μm). The fitted coefficient for the predicted mean grain size ($\beta_{\bar{x}_a}$) suggested that the prediction variability increased slightly but significantly with the predicted value of $\bar{x}_a$—by a factor of less than 1.01 per μm, or ~1.07 per 10 μm (Table 4). The coefficient for sorting (β_{σ_a}) predicted a very slight increase in variability at higher sorting values, by a factor of ~1.02 per 10 μm. All interaction terms were non-significant and were dropped from the models. Additional details on the GLMs are provided in Online Resource S2.

Discussion

Novel acoustic technologies obtain increasingly detailed information on the seabed, and complementary analytical approaches facilitate the use of these data for seabed mapping purposes. The recent introduction of multi-frequency MBES provides opportunities to overcome dependencies imposed by the use of a single acoustic frequency, and additionally, to investigate new metrics related to the relationships between frequencies (e.g., differences in substrate penetration; Gaida et al. 2020). Empirical approaches on the use of multi-frequency angular response information for seabed mapping are still sparse, and general research on acoustic angular response and its application for seabed mapping is ongoing (e.g., Alevizos and Greinert 2018; Wendelboe 2018; Fakiris et al. 2019; Fezzani et al. 2021; Fonseca et al. 2021). Ways in which discrete ground truth samples can be characterized using angular data covering the full swath width, and thus the full study area, are of particular interest. The methods presented here propose to leverage the collinearity between angular measurements and multiple operating frequencies to impute well-informed estimates of spatially continuous angular data across the extent of the study area. This enables modelling and prediction of substrate parameters from discrete ground truth samples at a high spatial and acoustic resolution. The uncertainty associated with predictions using imputed data are conveyed spatially to aid in the interpretation of results. We note that this procedure also allows for estimation of the full angular response curve at all acoustic segments across the dataset (Online Resource S3).

Imputation simulations in "Imputation simulation" section suggested that the proportion of missing data resulting from fine scale acoustic segmentation in this study (14.05%) did not approach the upper limits of what may be reasonable to impute. Even with 40% simulated missing data, imputed values explained, on average, > 90 and 75% of observed data variance for backscatter and depth difference variables, respectively (Fig. 5). These results are encouraging, particular for angular backscatter imputation, and are enabled by the consistent collinearity of collocated measurements (Online Resource S1). We suggest that the success of depth difference variable imputation will depend strongly on the substrate properties, as will the usefulness of such variables for habitat and substrate mapping (Gaida et al. 2020). These variables may have little value for predicting coarse or hard surficial sediments that preclude measurable differences in substrate penetration.

The performance of models used to predict seabed substrates varied here depending on the scale and format of acoustic predictors. Regardless of whether the acoustic data were angular or AVG-compensated, Random Forest classification using data segmented at a fine spatial scale performed favourably compared to a broader segmentation. Recall that the fine segmentation scale was selected to capture the apparent substrate heterogeneity observed from the backscatter data, while the broad scale was selected to produce acoustic segments with no missing data. All neural network grain size parameter models using angular data outperformed those using AVG data at a fine spatial scale.

We note the apparent correspondence between the thematic resolution of response and predictors in these cases (i.e., the measurement detail). The sediment types identified from benthic images at a broad thematic resolution (“fine”, “mixed”, “coarse” classes), were predicted well enough using acoustic data of low angular resolution (i.e., AVG-compensated; Table 2). Grain size parameters from physical samples (e.g., $\overline{x}_a$ and σ_a), however, are measured at a high resolution, and were modelled more effectively here using angular predictors (Table 3). The acoustic response of the seabed may be sensitive to these grain size parameters at varying incidence angles—for example, as acoustic scattering is increasingly influenced by interface roughness and less by hardness at oblique angles (Weber and Lurton 2015). This information is lost when performing AVG compensation. In all cases explored here, the fine scale segmentation with angular predictors outperformed the AVG models, and imputation enabled full coverage maps and increased sample size, without the loss of predictive accuracy. It is important to note that these methods are also applicable to single frequency MBES datasets.

Although imputation had no negative impact on model performance, it affected the uncertainty of model predictions. Multiple imputation (i.e., the creation and modelling of multiple entire imputed datasets) necessarily implies that we cannot be certain of the imputed values. Retaining and quantifying this uncertainty is an important part of the imputation procedure (Rubin 1978; van Buuren 2018); predictions based on data that are partially missing *should* be uncertain. Recent emphasis on conveying spatial model uncertainty can also be found in the seabed mapping literature (e.g., Mitchell et al. 2018; Shields et al. 2020; Strong 2020; Diesing et al. 2021). The results of multiple imputation can be incorporated into broader uncertainty analyses and mapped spatially to facilitate interpretation of model results and confidence. The classification results here, for example, incorporate the uncertainty that can be gleaned from an individual Random Forest model with the uncertainty arising from imputation by tallying votes from the ten individual models (Fig. 6). The map of entropy (H; Fig. 6) portrays predictions of the “mixed” class as uncertain compared to the other classes. Statistical analyses support this observation, suggesting that the seabed class was a significant predictor of model uncertainty, while the proportion of missing data had comparatively little impact on uncertainty (Fig. 6; Online Resource S2). The confidence of neural network grain size parameter predictions, on the other hand, appears to be strongly affected by the proportion of missing data according to the uncertainty map (Fig. 7), where increased prediction variability is apparent in the along-track direction. This observation was supported statistically, wherein the proportion of missing (and therefore imputed) data was significantly and positively predictive of the model variability (Online Resource S2).

Machine learning approaches are attractive for analysing angular and multi-frequency MBES datasets given their capacity for handling multidimensionality. Neural networks, for example, tune model weights to regularize or ignore uninformative predictors using backpropagation, which is computed based on a loss function that describes the error between the model and response data. This is highly advantageous for cases with many variables, some of which may be collinear and/or uninformative for the mapping purpose. At the Bedford Basin, previous research has identified subsurface dredge spoil covered by mud (visible in the northwest of Fig. 3a; Fader and Miller 2008; Brown et al. 2019). These subsurface features are clearly detected using the 100 kHz signal, yet show a lower homogenous return with the 400 kHz frequency (Brown et al. 2019). Here, although the subsurface dredge spoil was detected by the 100 kHz signal, and also the 100–400 kHz depth difference measurements (Fig. 4d), this information was not mapped to the prediction of surficial grain size or sediment class—it was effectively ignored by the models, which were trained exclusively using surficial samples (Figs. 6, 7). By generating predictors at a range of frequencies and incidence angles, the amount of potentially useful information is increased, and the model, rather than the analyst, determines the relevance of predictors based on the response variable.

Conclusions

Multiple imputation is a promising approach for achieving continuous estimates of angular response data at a high spatial resolution. Simulations in this study suggested that the proportion of missing angular data resulting from a fine object-based segmentation could be imputed with little error, which is attributable to the high collinearity of angular acoustic data. The implementation of Random Forest as an imputation model allows for automatic variable selection and interaction while retaining uncertainty of the imputed values. This is ideal for highly dimensional angular acoustic data.

The full imputed datasets of multi-frequency angular measurements were effective for producing continuous maps of seabed substrate properties at a high resolution. Machine learning approaches are well suited to modelling imputed multi-frequency angular MBES data, providing an alternative to variable selection or dimensionality reduction. We found no indication that the imputed data decreased the performance of seabed classification models of bottom type or regression of grain size parameters. The latter, though, were predicted with increasing uncertainty as the proportion of missing data increased for a given acoustic segment. Increased uncertainty at areas of sparse sounding density is the cost exacted for continuous high-resolution maps using these methods, and this must be conveyed along with the mapped predictions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11001-022-09471-3.

Acknowledgements Thank you to R2Sonic for providing data that were used in this study, and to Quality Positioning Services (QPS) for access to software licenses. Thanks to Ben DiTrolio for feedback on the manuscript. Fig. 2 was illustrated by Molly Wells (http://www.mollywellsart.ca/) as part of an undergraduate course on Experiential Learning in Marine Biology.

Authors' contributions B. Misiuk conceived of the study design. Field data collection and processing was coordinated by C.J. Brown. Data analysis was performed by B. Misiuk. The manuscript was written by B. Misiuk, and revisions were completed by B. Misiuk and C.J. Brown. Both authors read and approved the final manuscript.

Funding B. Misiuk was supported by an International Postdoctoral Fellowship awarded through the Ocean Frontier Institute/Canada First Research Excellence Fund. This research is part of the Ocean Frontier Institute Benthic Ecosystem Mapping and Engagement (BEcoME) Project.

Data availability Acoustic data were made available by R2Sonic. Ground truth data can be made available by the authors upon request.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare no competing interests.

Consent to participate Not applicable.

Consent for publication This manuscript has been approved for submission by both authors.

Ethical approval Not applicable.

References

Alevizos E, Greinert J (2018) The hyper-angular cube concept for improving the spatial and acoustic resolution of MBES backscatter angular response analysis. Geosciences 8:446. https://doi.org/10.3390/geosciences8120446

Alevizos E, Snellen M, Simons DG, Siemes K, Greinert J (2015) Acoustic discrimination of relatively homogeneous fine sediments using Bayesian classification on MBES data. Mar Geol 370:31–42. https://doi.org/10.1016/j.margeo.2015.10.007

Alevizos E, Snellen M, Simons DG, Siemes K, Greinert J (2018) Multi-angle backscatter classification and sub-bottom profiling for improved seafloor characterization. Mar Geophys Res 39:289–306. https://doi.org/10.1007/s11001-017-9325-4

Ambler G, Omar RZ, Royston P (2007) A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. Stat Methods Med Res 16:277–298. https://doi.org/10.1177/0962280206074466

Azur MJ, Stuart EA, Frangakis C, Leaf PJ (2011) Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 20:40–49. https://doi.org/10.1002/mpr.329

Blott SJ, Pye K (2001) GRADISTAT: a grain size distribution and statistics package for the analysis of unconsolidated sediments. Earth Surf Process Landf 26:1237–1248. https://doi.org/10.1002/esp.261

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Brown CJ, Beaudoin J, Brissette M, Gazzola V (2019) Multispectral multibeam echo sounder backscatter as a tool for improved seafloor characterization. Geosciences 9:126. https://doi.org/10.3390/geosciences9030126

Buscombe D, Grams PE (2018) Probabilistic substrate classification with multispectral acoustic backscatter: a comparison of discriminative and generative models. Geosciences 8:395. https://doi.org/10.3390/geosciences8110395

Che Hasan R, Ierodiaconou D, Laurenson L (2012) Combining angular response classification and backscatter imagery segmentation for benthic biological habitat mapping. Estuar Coast Shelf Sci 97:1–9. https://doi.org/10.1016/j.ecss.2011.10.004

Che Hasan R, Ierodiaconou D, Laurenson L, Schimel A (2014) Integrating multibeam backscatter angular response, mosaic and bathymetry data for benthic habitat mapping. PLoS ONE 9:e97339. https://doi.org/10.1371/journal.pone.0097339

Cogan CB, Todd BJ, Lawton P, Noji TT (2009) The role of marine habitat mapping in ecosystem-based management. ICES J Mar Sci 66:2033–2042. https://doi.org/10.1093/icesjms/fsp214

Collier JS, Brown CJ (2005) Correlation of sidescan backscatter with grain size distribution of surficial seabed sediments. Mar Geol 214:431–449. https://doi.org/10.1016/j.margeo.2004.11.011

Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24:603–619. https://doi.org/10.1109/34.1000236

Costa B (2019) Multispectral acoustic backscatter: how useful is it for marine habitat mapping and management? J Coast Res 35:1062. https://doi.org/10.2112/JCOASTRES-D-18-00103.1

Cragg JG (1971) Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39:829. https://doi.org/10.2307/1909582

Davis KS, Slowey NC, Stender IH, Fiedler H, Bryant WR, Fechner G (1996) Acoustic backscatter and sediment textural properties of inner shelf sands, northeastern Gulf of Mexico. Geo-Mar Lett 16:273–278. https://doi.org/10.1007/BF01204520

Diesing M, Thorsnes T, Bjarnadóttir LR (2021) Organic carbon densities and accumulation rates in surface sediments of the North Sea and Skagerrak. Biogeosciences 18:2139–2160. https://doi.org/10.5194/bg-18-2139-2021

Doove LL, Van Buuren S, Dusseldorp E (2014) Recursive partitioning for missing data imputation in the presence of interaction effects. Comput Stat Data Anal 72:92–104. https://doi.org/10.1016/j.csda.2013.10.025

Eekhout I, de Boer RM, Twisk JWR, de Vet HCW, Heymans MW (2012) Missing data: a systematic review of how they are reported and handled. Epidemiology 23:729–732. https://doi.org/10.1097/EDE.0b013e3182576cdb

Fader GBJ, Miller RO (2008) Surficial geology, Halifax Harbour, Nova Scotia. Geological Survey of Canada

Fakiris E, Blondel P, Papatheodorou G, Christodoulou D, Dimas X, Georgiou N, Kordella S, Dimitriadis C, Rzhanov Y, Geraga M, Ferentinos G (2019) Multi-frequency, multi-sonar mapping of shallow habitats—efficacy and management implications in the national marine park of Zakynthos. Greece Remote Sens 11:461. https://doi.org/10.3390/rs11040461

Ferrari R, Malcolm H, Neilson J, Lucieer V, Jordan A, Ingleton T, Figueira W, Johnstone N, Hill N (2018) Integrating distribution models and habitat classification maps into marine protected area planning. Estuar Coast Shelf Sci 212:40–50. https://doi.org/10.1016/j.ecss.2018.06.015

Ferrini VL, Flood RD (2006) The effects of fine-scale surface roughness and grain size on 300 kHz multibeam backscatter intensity in sandy marine sedimentary environments. Mar Geol 228:153–172. https://doi.org/10.1016/j.margeo.2005.11.010

Fezzani R, Berger L, le Bouffant N, Fonseca L, Lurton X (2021) Multispectral and multiangle measurements of acoustic seabed backscatter acquired with a tilted calibrated echosounder. J Acoust Soc Am 149:4503–4515. https://doi.org/10.1121/10.0005428

Fonseca L, Mayer L (2007) Remote estimation of surficial seafloor properties through the application angular range analysis to multibeam sonar data. Mar Geophys Res 28:119–126. https://doi.org/10.1007/s11001-007-9019-4

Fonseca L, Brown CJ, Calder B, Mayer L, Rzhanov Y (2009) Angular range analysis of acoustic themes from Stanton Banks Ireland: a link between visual interpretation and multibeam echosounder angular signatures. Appl Acoust 70:1298–1304. https://doi.org/10.1016/j.apacoust.2008.09.008

Fonseca L, Lurton X, Fezzani R, Augustin J-M, Berger L (2021) A statistical approach for analyzing and modeling multibeam echosounder backscatter, including the influence of high-amplitude scatterers. J Acoust Soc Am 149:215–228. https://doi.org/10.1121/10.0003045

Gaida TC, Tengku Ali TA, Snellen M, Amiri-Simkooei A, van Dijk TAGP, Simons DG (2018) A multispectral Bayesian classification method for increased acoustic discrimination of seabed sediments using multi-frequency multibeam backscatter data. Geosciences 8:455. https://doi.org/10.3390/geosciences8120455

Gaida TC, Snellen M, van Dijk TAGP, Simons DG (2019) Geostatistical modelling of multibeam backscatter for full-coverage seabed sediment maps. Hydrobiologia 845:55–79. https://doi.org/10.1007/s10750-018-3751-4

Gaida TC, Mohammadloo TH, Snellen M, Simons DG (2020) Mapping the seabed and shallow subsurface with multi-frequency multibeam echosounders. Remote Sens 12:52. https://doi.org/10.3390/rs12010052

Goff JA, Olson HC, Duncan CS (2000) Correlation of side-scan backscatter intensity with grain-size distribution of shelf sediments, New Jersey margin. Geo-Mar Lett 20:43–49. https://doi.org/10.1007/s003670000032

Goff JA, Kraft BJ, Mayer LA, Schock SG, Sommerfield CK, Olson HC, Gulick SPS, Nordfjord S (2004) Seabed characterization on the New Jersey middle and outer shelf: correlatability and spatial variability of seafloor sediment properties. Mar Geol 209:147–172. https://doi.org/10.1016/j.margeo.2004.05.030

Haris K, Chakraborty B, Ingole B, Menezes A, Srivastava R (2012) Seabed habitat mapping employing single and multi-beam backscatter data: a case study from the western continental shelf of India. Cont Shelf Res 48:40–49. https://doi.org/10.1016/j.csr.2012.08.010

Howell KL, Davies JS, Narayanaswamy BE (2010) Identifying deep-sea megafaunal epibenthic assemblages for use in habitat mapping and marine protected area network design. J Mar Biol Assoc U K 90:33–68. https://doi.org/10.1017/S0025315409991299

Huang Z, Siwabessy J, Nichol SL, Brooke BP (2014) Predictive mapping of seabed substrata using high-resolution multibeam sonar data: a case study from a shelf with complex geomorphology. Mar Geol 357:37–52. https://doi.org/10.1016/j.margeo.2014.07.012

Hughes Clarke JE, Mayer LA, Wells DE (1996) Shallow-water imaging multibeam sonars: a new tool for investigating seafloor processes in the coastal zone and on the continental shelf. Mar Geophys Res 18:607–629. https://doi.org/10.1007/BF00313877

Kingma DP, Ba J (2017) Adam: a method for stochastic optimization. arXiv:14126980 [cs]

Lamarche G, Lurton X (2018) Recommendations for improved and coherent acquisition and processing of backscatter data from seafloor-mapping sonars. Mar Geophys Res 39:5–22. https://doi.org/10.1007/s11001-017-9315-6

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2:18–22

Lucieer V, Lamarche G (2011) Unsupervised fuzzy classification and object-based image analysis of multibeam data to map deep water substrates, Cook Strait, New Zealand. Cont Shelf Res 31:1236–1247. https://doi.org/10.1016/j.csr.2011.04.016

Lurton X (2010) An introduction to underwater acoustics: principles and applications, 3rd edn. Springer, Berlin

Lurton X, Lamarche G (2015) Chapter 1: introduction to backscatter measurements by seafloor-mapping sonars. In: Lurton X, Lamarche G (eds) Backscatter measurements by seafloor-mapping sonars: guidelines and recommendations. GeoHab Backscatter Working Group

Malik M, Schimel ACG, Masetti G, Roche M, Le Deunf J, Dolan MFJ, Beaudoin J, Augustin J-M, Hamilton T, Parnum I (2019) Results from the first phase of the seafloor backscatter processing software inter-comparison project. Geosciences 9:516. https://doi.org/10.3390/geosciences9120516

McArthur MA, Brooke BP, Przeslawski R, Ryan DA, Lucieer VL, Nichol S, McCallum AW, Mellin C, Cresswell ID, Radke LC (2010) On the use of abiotic surrogates to describe marine benthic biodiversity. Estuar Coast Shelf Sci 88:21–32. https://doi.org/10.1016/j.ecss.2010.03.003

Misiuk B, Lecours V, Bell T (2018) A multiscale approach to mapping seabed sediments. PLoS ONE 13:e0193647. https://doi.org/10.1371/journal.pone.0193647

Mitchell PJ, Downie A-L, Diesing M (2018) How good is my map? A tool for semi-automated thematic mapping and spatially explicit confidence assessment. Environ Model Softw 108:111–122. https://doi.org/10.1016/j.envsoft.2018.07.014

National Physical Laboratory (2018) http://resource.npl.co.uk/acoustics/techguides/seaabsorption/

Parnum IM (2007) Benthic habitat mapping using multibeam sonar systems. PhD Thesis, Curtin University of Technology

Penone C, Davidson AD, Shoemaker KT, Di Marco M, Rondinini C, Brooks TM, Young BE, Graham CH, Costa GC (2014) Imputation of missing data in life-history trait datasets: which approach performs the best? Methods Ecol Evol 5:961–970. https://doi.org/10.1111/2041-210X.12232

Potts JM, Elith J (2006) Comparing species abundance models. Ecol Model 199:153–163. https://doi.org/10.1016/j.ecolmodel.2006.05.025

R Core Team (2019) R: a language and environment for statistical computing. Version 3.5.3. R Foundation for Statistical Computing, Vienna, Austria

Rubin DB (1978) Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse. In: Proceedings of the survey research methods section of the American Statistical Association. American Statistical Association, pp 20–34

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley-Interscience, Hoboken

Rubin DB (2004) The design of a general and flexible system for handling nonresponse in sample surveys. Am Stat 58:298–302. https://doi.org/10.1198/000313004X6355

Schimel ACG, Beaudoin J, Gaillot A, Keith G, Le Bas T, Parnum I, Schmidt V (2015) Chapter 6: Processing backscatter data: from datagrams to angular responses and mosaics. In: Lurton X, Lamarche G (eds) Backscatter measurements by seafloor-mapping sonars: guidelines and recommendations. GeoHab Backscatter Working Group

Schimel ACG, Beaudoin J, Parnum IM, Le Bas T, Schmidt V, Keith G, Ierodiaconou D (2018) Multibeam sonar backscatter data processing. Mar Geophys Res 39:121–137. https://doi.org/10.1007/s11001-018-9341-z

Shields J, Pizarro O, Williams SB (2020) Towards adaptive benthic habitat mapping. arXiv:200611453 [cs]

Simons DG, Snellen M (2009) A Bayesian approach to seafloor classification using multi-beam echo-sounder backscatter data. Appl Acoust 70:1258–1268. https://doi.org/10.1016/j.apacoust.2008.07.013

Smith SJ, Sameoto JA, Brown CJ (2017) Setting biological reference points for sea scallops (*Placopecten magellanicus*) allowing for the spatial distribution of productivity and fishing effort. Can J Fish Aquat Sci 74:650–667. https://doi.org/10.1139/cjfas-2015-0595

Stephens D, Diesing M (2015) Towards quantitative spatial models of seabed sediment composition. PLoS ONE 10:e0142502. https://doi.org/10.1371/journal.pone.0142502

Strong JA (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

Sutherland TF, Galloway J, Loschiavo R, Levings CD, Hare R (2007) Calibration techniques and sampling resolution requirements for groundtruthing multibeam acoustic backscatter (EM3000) and QTC VIEW™ classification technology. Estuar Coast Shelf Sci 75:447–458. https://doi.org/10.1016/j.ecss.2007.05.045

Todd BJ, Shaw J, Li MZ, Kostylev VE, Wu Y (2014) Distribution of subtidal sedimentary bedforms in a macrotidal setting: the Bay of Fundy, Atlantic Canada. Cont Shelf Res 83:64–85. https://doi.org/10.1016/j.csr.2013.11.017

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Trzcinska K, Janowski L, Nowak J, Rucinska-Zjadacz M, Kruss A, von Deimling JS, Pocwiardowski P, Tegowski J (2020) Spectral features of dual-frequency multibeam echosounder data for benthic habitat mapping. Mar Geol 427:106239. https://doi.org/10.1016/j.margeo.2020.106239

van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 16:219–242. https://doi.org/10.1177/0962280206074463

van Buuren S (2018) Flexible imputation of missing data, 2nd edn. Taylor & Francis Group, Boca Raton

van Buuren S, Groothuis-Oudshoorn K (2011) mice: multivariate imputation by chained equations in *R*. J Stat Soft 45. https://doi.org/10.18637/jss.v045.i03

van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB (2006) Fully conditional specification in multivariate imputation. J Stat Comput Simul 76:1049–1064. https://doi.org/10.1080/10629360600810434

van Ginkel JR, Sijtsma K, van der Ark LA, Vermunt JK (2010) Incidence of missing item scores in personality measurement, and simple item-score imputation. Methodology 6:17–30. https://doi.org/10.1027/1614-2241/a000003

Vergouw D, Heymans MW, van der Windt DAWM, Foster NE, Dunn KM, van der Horst HE, de Vet HCW (2012) Missing data and imputation: a practical illustration in a prognostic study on low back pain. J Manip Physiol Ther 35:464–471. https://doi.org/10.1016/j.jmpt.2012.07.002

Weber TC, Lurton X (2015) Chapter 2: background and fundamentals. In: Lurton X, Lamarche G (eds) Backscatter measurements by seafloor-mapping sonars: guidelines and recommendations. GeoHab Backscatter Working Group

Wendelboe G (2018) Backscattering from a sandy seabed measured by a calibrated multibeam echosounder in the 190–400 kHz frequency range. Mar Geophys Res 39:105–120. https://doi.org/10.1007/s11001-018-9350-y

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.