Quantitative experimental comparison of single-beam, sidescan, and multibeam benthic habitat maps

Alexandre C. G. Schimel, Terry R. Healy†, David Johnson, and Dirk Immenga

Schimel, A. C. G., Healy, T. R., Johnson, D., and Immenga, D. 2010. Quantitative experimental comparison of single-beam, sidescan, and multibeam benthic habitat maps. – ICES Journal of Marine Science, 67: 1766–1779.

Map comparison is a relatively uncommon practice in acoustic seabed classification to date, contrary to the field of land remote sensing, where it has been developed extensively over recent decades. The aim here is to illustrate the benefits of map comparison in the underwater realm with a case study of three maps independently describing the seabed habitats of the Te Matuku Marine Reserve (Hauraki Gulf, New Zealand). The maps are obtained from a QTC View classification of a single-beam echosounder (SBES) dataset, manual segmentation of a sidescan sonar (SSS) mosaic, and automatic classification of a backscatter dataset from a multibeam echosounder (MBES). The maps are compared using pixel-to-pixel similarity measures derived from the literature in land remote sensing. All measures agree in presenting the MBES and SSS maps as the most similar, and the SBES and SSS maps as the least similar. The results are discussed with reference to the potential of MBES backscatter as an alternative to SSS mosaic for imagery segmentation and to the potential of joint SBES–SSS survey for improved habitat mapping. Other applications of map-similarity measures in acoustic classification of the seabed are suggested.

Keywords: accuracy, average of mutual information (AMI), contingency matrix, Cramér's V, Goodman–Kruskal's lambda, kappa statistic, Theil's uncertainty coefficient.

Received 4 November 2009; accepted 23 May 2010; advance access publication 5 August 2010.

A. C. G. Schimel, T. R. Healy, and D. Immenga: Coastal Marine Group, Department of Earth and Ocean Sciences, The University of Waikato, PB3105 Hamilton, New Zealand. D. Johnson: MetOcean Solutions Ltd, 3/17 Nobs Line, New Plymouth, New Zealand. Correspondence to A. C. G. Schimel: tel: +64 7 8384024 ext. 7223; fax: +64 7 8560115; e-mail: acgs1@waikato.ac.nz.

Introduction

In the past 10 years, the human-induced worldwide decline of marine environments has raised awareness of the urgent need to improve the management of marine living resources and triggered an increase in research efforts to understand, classify, and protect ocean habitats (Jackson *et al.*, 2001; Pauly *et al.*, 2002; Pikitch *et al.*, 2004). The mapping of benthic habitats is typically achieved on the basis of direct biological or geological observations combined with data from remote-sensing acoustic systems (Diaz *et al.*, 2004), a practice known as acoustic seabed classification (ASC; Anderson *et al.*, 2008).

Direct observations are obtained from *in situ* techniques such as photography, video, sampling, coring, or scuba diving (Brown and Coggan, 2007). The remote-sensing acoustic systems typically used are single-beam echosounder (SBES), sidescan sonar (SSS), and multibeam echosounder (MBES; Kenny *et al.*, 2003; Michaels, 2007). *In situ* technologies allow the efficient localized description of the seabed but have limited coverage, whereas remote-sensing technologies allow excellent coverage but their output is ambiguous in terms of habitat description. A combination of both approaches allows counter-balancing for the respective flaws of each type and allows cost-effective surveying (Diaz *et al.*, 2004). However, the wide range of approaches to combine *in situ* data and acoustic data into a map testifies to the lack of agreement on a single, optimal habitat-mapping technique.

Many and varied acoustic features can be used for classification. Examples of SBES features include the energy of the first and second bottom echoes (Heald and Pace, 1996; Siwabessy *et al.*, 2000), or parameters describing the spectrum, envelope, or amplitude of the first echo (Anderson *et al.*, 2002; Ellingsen *et al.*, 2002; Preston *et al.*, 2004a). Examples of features derived from MBES or SSS backscatter imagery include statistical moments within a neighbourhood of samples (Preston *et al.*, 2004a; Brown and Collier, 2008), spectral features from Fourier or wavelet transform analysis (Pace and Gao, 1988; Atallah *et al.*, 2002), or indices from grey-level co-occurrence matrices (Huvenne *et al.*, 2002; Blondel and Gómez Sichi, 2009). Examples of features derived from MBES bathymetry include seabed roughness (Ierodiaconou *et al.*, 2007), topographic position index (Iampietro *et al.*, 2005), or local Fourier histogram texture features (Cutter *et al.*, 2003). Examples of features derived from MBES-backscatter angular response include empirical parameters describing the response shape (Hughes Clarke, 1994; Beyer *et al.*, 2007), or solutions to an inverted geoacoustic model fitted to the response curve (Fonseca *et al.*, 2009).

Also, there is a wide range of classification algorithms available. The traditional interpretative approach, in which experts are responsible for manually segmenting an acoustic image, is still often used because of its reliability (Kostylev *et al.*, 2001; Roberts *et al.*, 2005; Ehrhold *et al.*, 2006; Collier and Humber, 2007;

†Deceased 20 July 2010.

Prada *et al.*, 2008), but advances in computer processing capabilities now allow the use of various automated approaches (Simard and Stepnowski, 2007). Examples of automated algorithms used in recent literature include *k*-means clustering (Legendre *et al.*, 2002; Blondel and Gómez Sichi, 2009), decision tree (Ierodiaconou *et al.*, 2007), discriminant analysis (Hutin *et al.*, 2005), Bayes' decision rule (Simons and Snellen, 2009), and neural networks (Marsh and Brown, 2009).

Finally, the design of a given classification methodology is subjective. Different results can be obtained if acoustic data are classified with the help of *in situ* data (supervised approach) or without (unsupervised approach; Simard and Stepnowski, 2007). Other important considerations include the number of categories to work with, whether to run the classification on individual features or coherent localized groups of features (object-orientated analysis; Lucieer, 2008), or whether to run a "hard" or fuzzy classification (Lucieer and Lucieer, 2009).

The increasing number of acoustic systems, data-processing techniques, classification schemes, and methodologies to link acoustic and *in situ* data, some of which are described above, implies a growing need for comparison. Ultimately, comparative studies could lead to the identification of the most appropriate systems (or combinations of systems) and methodologies for given survey objectives and conditions. With this purpose, a number of studies offer a comparison of the theoretical performances of different acoustic-mapping systems (Hamilton *et al.*, 1999; Kenny *et al.*, 2003; Le Bas and Huvenne, 2009). However, such a system-orientated approach ignores the variable results that can be obtained from different processing or classification methodologies.

The conventional approach for comparing different processing or classification methodologies is to produce a case-study map for each, estimate their respective accuracy in reference to a ground-truth dataset, and compare the two estimates. The techniques for estimating the accuracy of a thematic map have their origin in land remote sensing (Congalton, 1991; Foody, 2002, 2008), and their use is gaining momentum in ASC (Foster-Smith and Sotheran, 2003; Brown *et al.*, 2005; Brown and Collier, 2008; Lucieer, 2008; Walker *et al.*, 2008). Obtaining an estimate of map accuracy is now relatively straightforward, but comparing two estimates is difficult because it requires calculation of their respective variances, and this is highly dependent on the size and design of the ground-truth dataset (Stehman and Czaplewski, 1998; Foody, 2009). This is an important issue in ASC, where seabed ground-truthing presents specific challenges including access difficulty, poor visibility, acoustic/ground-truth data-scale difference, position precision, and habitat subjective description (Brown and Coggan, 2007).

A second approach to comparing different processing or classification methodologies is the direct comparison of one map with another, without referring to an *in situ* dataset as ground-truth. Such map-to-map comparison benefits from decades of development in diverse fields involving land mapping (Boots and Csillag, 2006). Techniques for the comparison of land maps include measures derived from pixel-to-pixel comparison (Foody, 2006), features identification and analysis (Dungan, 2006), pattern-based techniques (White, 2006), or fuzzy-logic-based measures that take into account possible vagueness in pixel location or legend category (Hagen-Zanker, 2006). In contrast to land remote sensing, map-to-map comparison is still relatively uncommon in ASC to date, with the notable exception of the works by Foster-Smith *et al.* (2004) and Brown *et al.* (2005).

The main advantage of direct map-to-map comparison is that it allows one to circumvent the complications posed in the first approach by its requirement for a properly designed ground-truth survey (Stehman, 2006). However, the reciprocal drawback is that in the absence of evaluation of map accuracy, the observation of map similarity or dissimilarity is ambiguous. For example, the observation that two given maps A and B differ importantly could be the result of A being accurate and B not, or B being accurate and A not, or both A and B being inaccurate, or both A and B being equally accurate, but happening to depict different ground characteristics. As a result, map-to-map comparison is generally limited to specific study objectives where the accuracy ambiguity is lifted or made irrelevant. Examples of objectives for map-to-map comparison include the basic characterization of the degree of similarity between different mapping algorithms, the detection of changes over time, or the validation of a map produced under the assumption that it is compared with a map that actually represents the ground-truth (Foody, 2007).

Here, we aim to illustrate the potential benefit of map-to-map comparison in ASC for comparing seabed maps produced by different acoustic systems or classification methodologies. As a case study, three maps were created to represent the result of independent, typical benthic habitat-mapping efforts at the same site. They were obtained from SBES, SSS, and MBES datasets, which were acquired at a different time with different resolution and coverage, classified in unsupervised mode using the usual algorithms for each acoustic system, and ground-truthed using different *in situ* surveys. The small size of the ground-truth surveys precluded reliable estimation of map accuracy, but not a direct map-to-map comparison. A number of measures derived from the literature in land remote sensing and selected for their suitability to this study context was applied to estimate map similarity. The similarity results were then examined, the benefits and limits of the selected approach discussed, and other potential applications of map-similarity measures in ASC suggested.

Material and methods

The study site was the Te Matuku Marine Reserve, located south of Waiheke Island in the Hauraki Gulf in New Zealand (~36°51′S 175°08′E; Figure 1a). The 690 ha reserve was established in 2005 to cover the Te Matuku Bay estuary and its subtidal extension in the sheltered Tamaki Strait. The study focuses on the subtidal part of the reserve, which accounts for ~550 ha, including flats off the bay headlands and the entrance of the Waiheke Channel to a depth of 25 m. Early surveys recognized the area as typical of inner Hauraki Gulf sheltered shores: the dominant seabed substratum is fine, silty mud, with extensive bioturbation in places, occasional patches of horse mussel (*Atrina zelandica*) shell debris, and rocky outcrops around headlands and Passage Rock Island (The Royal Forest and Bird Protection Society, 1998).

SBES classification

In 2002, New Zealand's National Institute of Water and Atmospheric Research (NIWA) conducted a habitat survey of the proposed area for the Te Matuku Marine Reserve, as part of a wider programme of habitat identification in the Hauraki Gulf. The habitat mapping was performed with a Simrad EA501P SBES, the signal of which was processed and classified with Quester Tangent software QTC View Series 4 and QTC IMPACT (Morrison *et al.*, 2003).

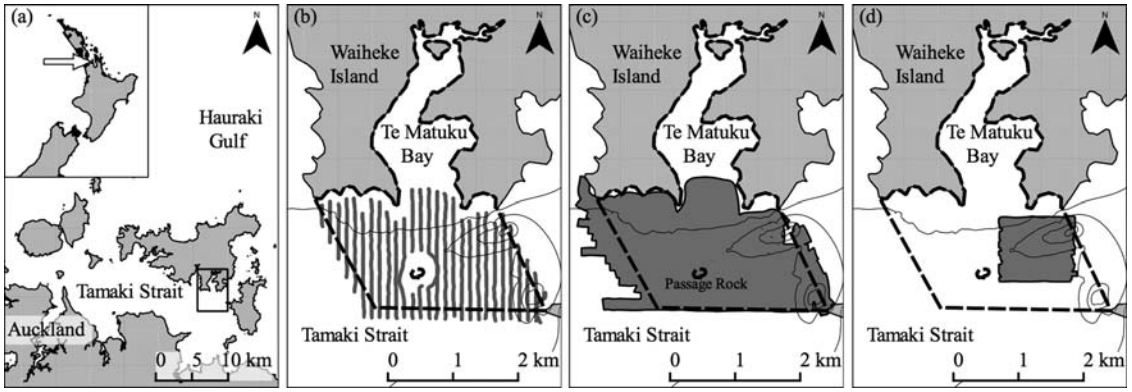

Figure 1. (a) Location of the study site in the Tamaki Strait, Hauraki Gulf, New Zealand (36°51′S 175°08′E). (b) Coverage of the SBES survey (north–south continuous lines), after Morrison *et al.* (2003). (c) Coverage of the SSS survey (dark area). (d) Coverage of the MBES survey (dark area). All panels except the left one also display the extent of the Te Matuku Marine Reserve (dashed contour) and the 5, 10, 15, and 20 m isobaths.

The SBES used in the survey had an operating frequency of 200 kHz, a ping rate of 5 Hz, and a fixed beam width of 7° (Morrison *et al.*, 2001). The acoustic dataset covered the entire subtidal part of the marine reserve (Figure 1b), with a total of 30 lines acquired in a north or south direction at a speed of ~3 m s^{-1}, separated by 120 m on average (Morrison *et al.*, 2003). The QTC software analysed the SBES signal in stacks of consecutive pings to minimize signal variability (Preston *et al.*, 2004a). This process resulted in the generation of one ping-stack every 6 m on average along the lines (Morrison *et al.*, 2003). Therefore, the original dataset for classification had an average spatial resolution of 120 m × 6 m.

The QTC software first extracted 166 features from the ping stacks, then applied a principal component analysis to identify the three principal components, which are termed *Q*-values. The *Q*-values were then clustered using a semi-automatic algorithm, in which the user was responsible for the decision of whether there should be further splitting of the clusters with the help of statistical diagnostics. When the final set of clusters was decided, the *Q*-values were compared with the centroid of each, resulting in a category being assigned to each ping-stack along with a confidence value between 0 and 100% (Preston *et al.*, 2004a, b). This process of classification resulted in an optimal number of four categories (Morrison *et al.*, 2003). An interpolation algorithm was then applied to the ping-stack classification to obtain a thematic map covering the entire site (Morrison *et al.*, 2003). However, the resulting map displayed a general unrealistic "blocky" aspect (Morrison *et al.*, 2003; Schimel *et al.*, 2010). This effect is found frequently when using traditional interpolation algorithms for categorical data on point-based datasets with both an imbalance between along- and across-track resolution and a high point-to-point variability (Foster-Smith and Sotheran, 2003; Reid, 2007).

In this case, the ping-stack classification was interpolated again using an alternative algorithm designed for categorical data and based on an inverse distance calculation, with the aim of obtaining a map with a more realistic aspect. The inverse distance calculation was expected to create a spatial averaging effect to filter out the rapid variations in the original data, and the specific design for categorical data ensured that no artificial categories were created in the process (Reid, 2007).

With any point x, the algorithm would associate the category for which the sum of the inverse distances between x and the points belonging to the set to be interpolated, classified in this category and located within a given threshold distance from x, is maximized over all categories. The resulting category is

$$c(x) = \arg\max_{k\in[1,n]} \sum_{y\in Y_{k,D,x}} \frac{1}{\mathrm{d}(x,y)}, \quad (1)$$

where Y is the entire dataset of points to be interpolated, n the total number of categories in which Y is partitioned, $Y_{k,D,x}$ the subset of Y consisting of the elements classified in category k and located within the threshold distance D from x, and d a distance function. In practice, the QTC ping-stack classification dataset Y was limited to the elements y that scored more than 80% confidence during the classification process, the interpolation was run on a grid of points x set up at a resolution of 1 m, the Euclidian distance was used for d, the threshold distance D was set to 100 m, and the final results were limited to the convex hull of the QTC ping-stack classification dataset Y to remove unnecessary extrapolation.

The interpolated map was ground-truthed with a video and sediment-sampling survey of 12 stations arranged in a stratified design: three stations were selected within patches of "pure" category for each of the four categories (Morrison *et al.*, 2003). At each station, underwater video footage was acquired and a sediment sample obtained with a Smith–McIntyre grab sampler. Primary substratum type, secondary cover, and conspicuous epifauna were described from video footage and sediment sample observation, and grain-size distribution was derived from the analysis of samples using a GALAI (CIS-100) laser particle sizer (Morrison *et al.*, 2003). To complete this original ground-truthing effort, the sediment grain-size analysis was carried on further in this study with the computation of the volume percentage of clay, silt, sand, and gravel-size particles (>2 mm), as well as the mean grain size and sorting of the <2 mm fraction. All 12 stations were used for category identification; none were conserved for map-accuracy estimation.

SSS classification

In 2002, the University of Waikato's Department of Earth and Ocean Sciences conducted an SSS survey of the proposed site for

the Te Matuku Marine Reserve using a Klein 595 SSS for data acquisition and Triton Imaging Inc. ISIS software suite for data processing (Figure 1c). SSS imagery was obtained from mosaicking the 100-kHz data at a resolution of 0.2 m using the assumption of a flat seabed. As the poor quality of the data precluded efficient data conditioning for modern image-analysis techniques to be applied, the mosaic was segmented manually. Segmentation was performed with the digitizing tools of GIS software on the basis of a visual assessment of areas of homogeneous tone and texture. Five categories were identified. The map was then rasterized at a resolution of 1 m.

In 2005, New Zealand's Department of Conservation performed a sediment-sampling survey of the marine reserve. The survey consisted of 146 stations arranged in a simple random design over the entire reserve, including its intertidal part. Sediment samples were collected at each station using a small rectangular dredge described in Grace and Whitten (1974), then analysed for grain-size distribution using a Malvern laser particle sizer (K. Sivaguru, pers. comm.). For each sample, the volume percentage of clay, silt, sand, and gravel (>2 mm), and the mean grain size and sorting of the <2 mm fraction, were calculated. Only 69 of the 146 stations were located within the area covered by the SSS imagery and were used for ground-truthing the SSS map. All 69 stations were used for category identification; none were conserved for map-accuracy estimation.

MBES classification

In 2007, an MBES survey was conducted over a rectangular area of ~100 ha in the Waiheke Channel part of the Te Matuku Marine Reserve (Figure 1d). The specific purpose of the survey was to acquire an MBES dataset for development of a processing methodology and for the preliminary comparison of its results with the SBES and SSS classifications (Schimel *et al.*, 2010). Therefore, the survey was not performed on the entire subtidal part of the marine reserve, as were the previous surveys, but only on an area large enough to cover occurrences of each category from the previous classifications, as well as the full depth range of the area.

The survey was conducted with a Kongsberg EM3000 MBES (300 kHz), planned so that outer beams from two consecutive runlines were slightly overlapping to ensure 100% coverage. The backscatter data were processed to remove the along-track banding effect and gridded at a resolution of 1 m (Schimel *et al.*, 2010). A 10 m × 10 m two-dimensional median filter was then applied to the imagery to remove the high-frequency noise typically present in MBES-backscatter data recorded near the nadir. Observation of the filtered image histogram revealed three main concentrations of pixels at, respectively, high, medium, and low backscatter levels. The filtered image was classified using a *k*-means clustering algorithm, with the number of categories *k* accordingly set to three.

The map produced by this semi-automatic classification was ground-truthed using footage from a drop-video-camera survey carried out in 2008 and comprising 24 stations arranged in a systematic design over the area covered by the MBES. The video camera was fitted on a frame lowered to the seabed, and the vessel was allowed to drift during the length of footage recording on each site. Such drifting minimized the error in the frame position, assumed to be identical to the vessel position, measured with DGPS, and ensured that the habitat observed was representative of its surroundings. Map categories were described on the basis of visual assessment of the video footage. In addition, four of the stations were sampled and observed by a scuba diver. The samples were analysed for sediment grain-size distribution with a Malvern laser particle sizer. All 24 stations were used for category identification; none were conserved for map-accuracy estimation.

Map-comparison measures

As outlined above, a wide range of approaches developed for the comparison of land maps can be used directly in ASC, depending on study context and objectives. The objective of the current study is to estimate the overall similarity of three overlapping maps with identical resolution of 1 m but different legends and different coverage, and for which no samples are available for map-accuracy assessment. In this context, a map-to-map comparison approach can be implemented using similarity measures obtained from the count of pixels shared by the maps, which is usually presented in the form of a contingency matrix (Table 1).

Diverse measures expressing different aspects of map similarity can be computed from the contingency matrix. Here, several measures were selected and applied with the objective of providing an overview of the range of existing measures and of the diverse aspects of map similarity that can be estimated. Following a review by Rees (2008), the measures of categorical agreement A (overall accuracy), Cohen's κ, and Foody's κ^* and the measures of categorical association Theil's U, Cramér's V, and Goodman–Kruskal's λ were selected.

Historically, the first map-similarity measures used in land remote sensing were metrics originally designed for estimating the accuracy of a map produced against a reference ground-truth dataset. Therefore, they require the two maps to be described with the same legend. In reference to the terminology in Table 1, this implies that m and n must be equal, that A_i and B_i must be the same for each row/column i, that the elements on the diagonal represent the count of pixels where the classifications agree, and that the off-diagonal elements represent classification disagreements.

In this specific case, the overall accuracy A is the straightforward proportion of pixels where the two classifications agree. Accordingly, it takes values between 0, indicating no agreement, and 1, indicating complete agreement:

$$A = \frac{1}{N}\sum_{i=1}^{n} c_{ii}. \quad (2)$$

Cohen's κ is a popular measure of agreement that uses the off-diagonal elements to estimate chance agreement and to

Table 1. Contingency matrix for two maps A and B comprising, respectively, m and n categories.

	Map B categories					
Map A categories	B_1	...	B_j	...	B_n	**Total rows**
A_1	c_{11}	...	c_{1j}	...	c_{1n}	c_{1+}
...	...	...	...	...	...	...
A_i	c_{i1}	...	c_{ij}	...	c_{in}	c_{i+}
...	...	...	...	...	...	...
A_m	c_{m1}	...	c_{mj}	...	c_{mn}	c_{m+}
Total columns	c_{+1}	...	c_{+j}	...	c_{+n}	N

The number c_{ij} designates the number of pixels that fall conjointly in category A_i in map A and B_j in map B. The numbers c_{+j} and c_{i+}, respectively, designate the sum of the elements in column j and the sum of elements in row i. N is the total number of pixels shared by the two maps.

compensate A accordingly (Cohen, 1960; Congalton, 1991; Monserud and Leemans, 1992; Couto, 2003):

$$\kappa = \frac{A - 1/N^2 \sum_{i=1}^{n} c_{i+}c_{+i}}{1 - 1/N^2 \sum_{i=1}^{n} c_{i+}c_{+i}}. \quad (3)$$

The estimation of chance agreement in κ has often been criticized, and various alternatives have been suggested (Brennan and Prediger, 1981; Ma and Redmond, 1995). In particular, the estimation of chance agreement assuming that the marginal distributions are not specified *a priori* is considered more suitable in the context of geographical mapping (Foody, 1992; Stehman, 1999). Modifying κ accordingly, this measure becomes

$$\kappa^* = \frac{A - 1/n}{1 - 1/n}. \quad (4)$$

As κ and $\kappa^\star$ are re-scaled versions of A that take into account chance agreement, they systematically take lower values than A. They take a value of 0 if map agreement is equivalent to that expected by chance, a negative value if map agreement is less than would be expected by chance, and a maximum value of 1 for complete agreement.

The requirement that the two maps to be compared must have the same legend to allow using A, κ, or $\kappa^\star$ is an obstacle in many studies where the legends differ in the number of categories and/or category labels. Using A, κ, or $\kappa^\star$ in this context implies aggregating and re-labelling some categories until a common legend is obtained, which is often done subjectively (Foster-Smith *et al.*, 2001; Giri *et al.*, 2005; McCallum *et al.*, 2006). A better approach is to use alternative measures that can be computed regardless of a possible legend mismatch, i.e. from a "not necessarily square" contingency matrix (Boots and Csillag, 2006; Foody, 2006).

Finn (1993), drawing from information theory, suggested a map-similarity measure with this characteristic. If map uncertainty is considered to be the information content of a map, then an estimation of map similarity can be obtained through computing the average mutual information (AMI), which measures the reduction in one map's uncertainty when the other map is known (Theil, 1972; Finn, 1993; Couto, 2003; Foody, 2006; Rees, 2008):

$$\mathrm{AMI} = H(A) + H(B) - H(A, B), \quad (5)$$

where $H(A)$ and $H(B)$ describe the respective entropy (uncertainty) of the two maps, and $H(A, B)$ describes their joint entropy. With a constant term of 1 and in Hartley units, they are, respectively,

$$H(A) = -\sum_{i=1}^{m} \frac{c_{i+}}{N} \log\left(\frac{c_{i+}}{N}\right), \quad (6)$$

$$H(B) = -\sum_{j=1}^{n} \frac{c_{+j}}{N} \log\left(\frac{c_{+j}}{N}\right), \text{ and} \quad (7)$$

$$H(A, B) = -\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{c_{ij}}{N} \log\left(\frac{c_{ij}}{N}\right). \quad (8)$$

Theil's uncertainty coefficient U is a normalized and symmetric estimate of mutual information based on AMI that originated in categorical statistics, where the above concepts apply equally (Theil, 1972). It is written (Press *et al.*, 1992) as

$$U = \frac{2 \times \mathrm{AMI}}{H(A) + H(B)}. \quad (9)$$

More recently, Rees (2008) suggested two other pixel-to-pixel comparison measures drawn from the field of categorical statistics, which can also be computed from the contingency matrix without the requirement of identical legends: Cramér's V and Goodman–Kruskal's λ.

Cramér's V is a normalized version of Pearson's χ^2 statistic (Cramér, 1946; Rees, 2008):

$$V = \sqrt{\frac{\chi^2}{N(\min(m, n) - 1)}}, \quad (10)$$

and Pearson's χ^2 is

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(c_{ij} - c_{i+}c_{+j}/N)^2}{c_{i+}c_{+j}/N}. \quad (11)$$

Goodman–Kruskal's λ is a measure of the proportional reduction in error in one map obtained from knowledge of the other map (Goodman and Kruskal, 1954; Rees, 2008). In its symmetrical version, it is

$$\lambda = \frac{\sum_{i=1}^{m} \max_j(c_{ij}) + \sum_{j=1}^{n} \max_i(c_{ij}) - \max_j(c_{+j}) - \max_i(c_{i+})}{2N - \max_j(c_{+j}) - \max_i(c_{i+})}. \quad (12)$$

U, V, and λ are normalized; they take values between 0, indicating no association, and 1, indicating complete association.

Map-comparison methodology

In this study, the three maps to be compared had different legends because they were ground-truthed independently. The measures of association U, V, and λ were therefore adapted while the measures of agreement, A, κ, and $\kappa^\star$, were not, unless the map legends were modified. A methodology was developed to automate the decision process for legend modification and allow the use of the three measures of agreement in this study.

Consider two maps A and B having the same number of categories m but different or unknown category labels. One could assess the similarity between A and B by computing a measure of agreement for all possible category bijections between A and B and keeping only one of the resulting values, intuitively the largest one. This process is equivalent to forming all the $m!$ possibilities of category permutations in one map.

If A and B have different numbers of categories m and n such that $n > m$, one could still apply the permutation process described above after having formed all the possibilities of aggregating categories from B so that only m categories remained. This category-aggregation process is equivalent to identifying all the possibilities to partition a set of n elements into m non-empty subsets, as given by the Stirling numbers of the second type (Abramowitz and Stegun, 1964):

$$S(n, m) = \frac{1}{m!} \sum_{k=0}^{m} (-1)^{m-k} \frac{m!}{k!(m-k)!} k^n. \quad (13)$$

Accordingly, the total number of values that can be taken by a measure of agreement between two maps A and B having a different number of categories m and n after the aggregation/permutation process is $m!S(n, m)$. The main advantage of this process is that it allows the popular measures of categorical agreement to be used for maps with different legends in an automated manner. A second advantage is that it provides an optimal solution for the comparison of the legends, which is the aggregation/permutation possibility that maximizes the map-similarity measure. This information allows verification that the computed measure is actually an estimate of map agreement rather than the product of a chance association of completely different categories.

In the present study, the SBES, SSS, and MBES maps were compared using the measures of categorical agreement and the measures of categorical association described above. As the three maps were created at a common resolution of 1 m, each pair of maps led to a straightforward contingency matrix. The measures of association U, V, and λ were computed directly from the contingency matrices, and the measures of agreement A, κ, and κ^* were computed after the application of the automatic aggregation/permutation process. At the end of the process, only the maximum value of each measure and the corresponding solution in legend agreement were reported.

This comparison methodology had its limitations. First, the difference in map size (Figure 1) could have an influence on the results. As the MBES map was smaller than the other two, the MBES–SBES and the MBES–SSS comparisons were limited to the size of the MBES map, whereas the SBES–SSS comparison was limited to the area shared by the two maps, i.e. almost the entire study site. This difference may artificially lessen the level of agreement or association of the latter comparison. Second, the level of agreement generally increases as categories are aggregated (Giri *et al.*, 2005; Foody, 2007), implying that comparisons between maps described with fewer categories may artificially show better agreement or association than other comparisons.

To assess the influence of map size and the number of categories in this study, the three maps were compared a second time after being limited to the pixels shared by the three maps, i.e. approximately the MBES area, and after the maps were all reduced to a same number of categories by subjective aggregation. This process was termed map reduction.

Results and discussion

Map results and analysis

Figure 2a shows the SBES ping-stack dataset classified by the QTC software into four categories, labelled A, B, C, and D, and Figure 2b the result of interpolation of the dataset using the categorical inverse distance algorithm. Both figures also show the locations of the ground-truth stations. Table 2 lists the results of the ground-truthing survey.

The video footage and visual assessment of the sediment samples confirmed the dominance of mud as a primary substratum on the entire study site. In contrast, grain-size analysis revealed that sediment samples contained mainly sand-size particles. Despite this discrepancy, both video footage and grain-size analysis agreed that categories A and B proved similar in demonstrating the softer sediment at the site, that category C had a slightly coarser sediment, and that category D was defined mainly by its notable cover of shells and shell fragments.

The origin of this discrepancy was not determined, but the upper layer of the seabed at the site might be stratified so that the samples, which were mostly of subsurface sediment, would naturally yield a different result from the video footage, which only allowed an assessment of the composition of the surface sediment (M. Morrison, pers. comm.). Another hypothesis is that the organic content in the samples, which is high at the site, was not entirely degraded during the analysis, and might have bound silt-size grains into coarser particles.

Earlier studies using the QTC software reported cases of correlation between QTC classification and water depth (Anderson *et al.*, 2002; Hewitt *et al.*, 2004). A similar correlation was found on this site by Schimel *et al.* (2010), who observed that the distribution of category A corresponded to shallow water and those of categories C and D to deeper water. Categories C and D were identified as distinctive habitats from the ground-truthing survey, but categories A and B were identified as similar. This suggests that depth, or another environment factor correlated with it but not measured in the ground-truth survey, might have contributed to separating A and B during the classification process.

Figure 2c shows the SSS imagery and Figure 2d the thematic map resulting from manual classification, and the locations of the 2005 sediment samples used for ground-truthing. From the SSS imagery, the operator identified five categories labelled E, F, G, H, and I, for which tone and texture appeared clearly different from each other. Figure 3 depicts the results of the grain-size analysis for each category. There was a notable variation in the ground surface occupied by each SSS category. A smooth-textured low-reflectivity background covered most of the mosaic (category I), but it was replaced in places by a rougher texture type with greater reflectivity, mainly in a large patch in the centre east of the mosaic and in intermittent, smaller patches in the centre and the south (category E). The extension of rocky headlands and islands on the seabed showed great reflectivity and could be separated into two different texture types (categories F and H), both of which, but particularly category F, were rare. A last texture type presenting a pattern alternating high- and low-reflectivity marks was identified mainly in the eastern part of the site (category G).

As the SSS ground-truthing sampling scheme was devised randomly, the high variability in SSS category surface resulted in a great variability in the number of samples available for each category. In all, 43 were located within the largest category (I), whereas no samples were located within the smallest category (F). Respectively nine, nine, and eight samples were located within categories E, G, and H. The acoustic classification and the grain-size analysis matched poorly, with a substantial variation in grain-size results in categories G, H, and I, and similar grain-size results between all categories (Figure 3). Categories E and I, which showed radically different tone and texture and were therefore particularly distinguishable from each other on the acoustic imagery, proved to be particularly similar in sediment content, i.e. a medium to fine silt poorly to very poorly sorted. They both had a negligible fraction of gravel-size (>2 mm) and clay-size particles. The main difference was that I had a higher sand content than E. Category G was quite similar but with a less sorted, less silty, and sandier content, and its gravel-size fraction was more in evidence. Finally, category H was also similar, but increasing the trend from G into less sorted and larger grain sizes. It is the only category for which the mean volume content was greater for sand than for silt. Accordingly, the categories were further

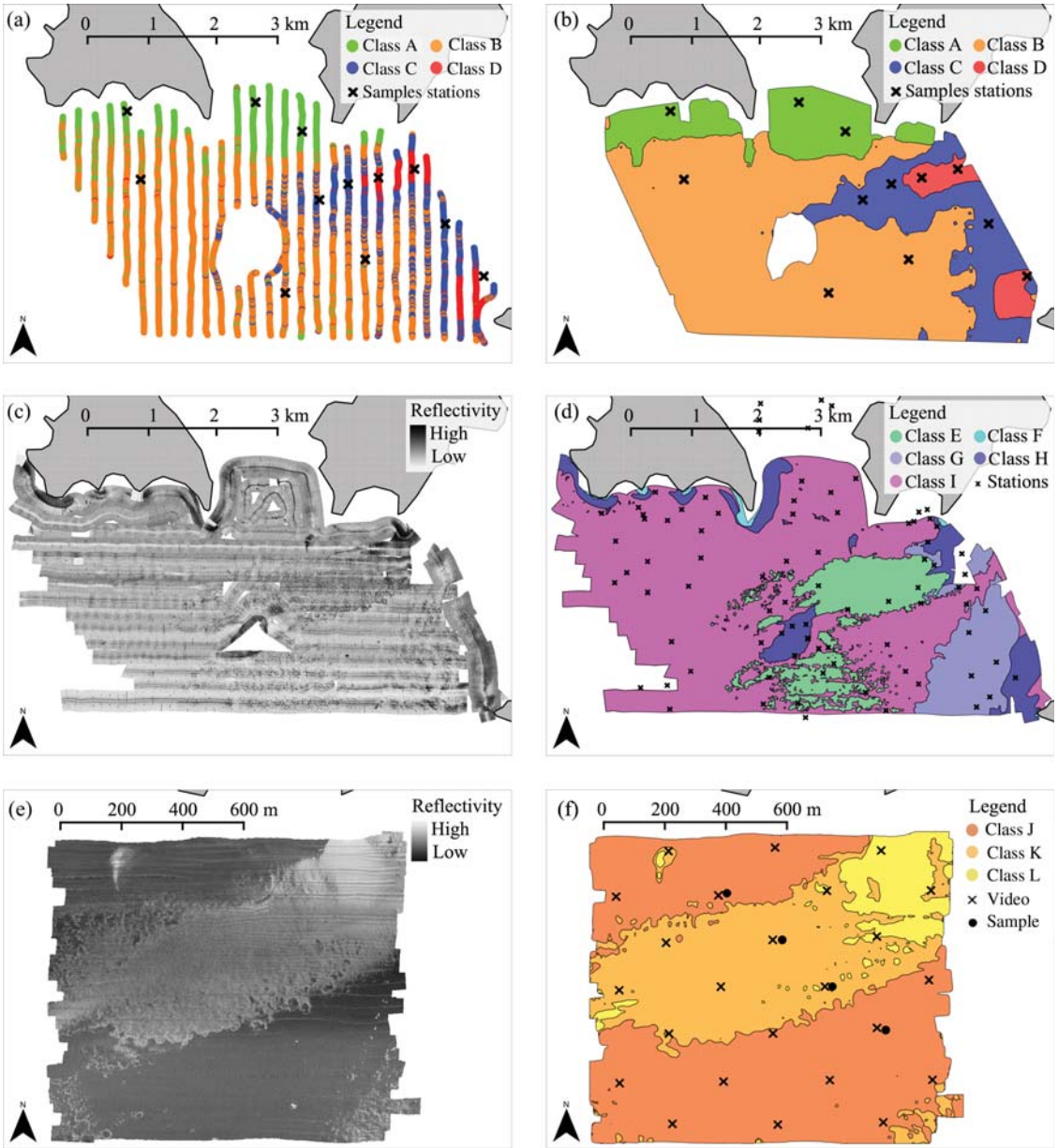

Figure 2. (a) SBES ping-stack classification by QTC View/Impact (after Morrison *et al.*, 2003). (b) Map resulting from the application of the interpolation algorithm to the SBES classification. Both panels also display the location of the ground-truth stations for the SBES map. (c) SSS mosaic. (d) Map resulting from the manual classification of the SSS mosaic and the location of the sampling stations from the 2005 survey. (e) MBES imagery. (f) Map resulting from the automatic classification of the MBES imagery and the location of the ground-truth stations from the 2008 survey. The location of the data displayed in the final two panels is indicated in Figure 1.

labelled as E (mainly fine silt, poorly sorted), F (no stations), G (sandy silt, very poorly sorted), H (silty sand, very poorly sorted), and I (mainly medium silt, with sand occurrence).

Similar difficulties in relating grain-size results and sidescan classification have been observed in other studies on soft-sediment areas with even less homogeneous surficial sediment distribution than in the present case (Zajac *et al.* 2000; Brown *et al.*, 2002). An important variation in tone and texture in the sidescan imagery that cannot be linked clearly to sediment grain size suggests a contribution of other environmental factors, possibly related to seabed roughness. This hypothesis implies that the *in situ* technique selected for ground-truthing the SSS map may not be suitable for all categories.

Figure 2e depicts the MBES reflectivity map after partial correction of the along-track banding effect (Schimel *et al.*, 2010), and Figure 2f the thematic map resulting from the semi-automatic classification of this reflectivity map, and the location of the ground-truth stations. The clustering algorithm was set to split the dataset into three categories labelled J, K, and L. The algorithm attributed category J to the low-reflective, smooth-textured background of the reflectivity map, category K to the medium-reflective, rough-textured features, which were mainly in a band crossing the

Table 2. Results of the ground-truth survey of the SBES classification.

Category	Video observation		Sample observation			
A	Mud and sloped burrows. Cushion stars		Very soft mud. Shell fragments underneath the surface			
B	Mud and sloped burrows		Soft to very soft mud. Few shell fragments on surface			
C	Sandy mud. Dead shells		Soft grey clay. Shell fragments			
D	Poor visibility. Heavy sand and shell in places. Hard mud in others		Soft mud. Many shell fragments on surface and beneath			
	Statistics of the <2 mm content		**% content in volume**			
	Mean (phi)	**Sorting (phi)**	**Clay**	**Silt**	**Sand**	**>2 mm**
A	3.13 (very fine sand)	1.45 (poorly sorted)	0.1	22.9	75.4	1.6
B	3.30 (very fine sand)	1.61 (poorly sorted)	0.3	32.8	64.7	2.2
C	1.41 (medium sand)	1.81 (poorly sorted)	0.2	12.1	83.0	4.7
D	0.24 (coarse sand)	1.07 (poorly sorted)	0.1	2.7	72.1	25.1

The four QTC categories are described on the basis of the observation of the video footage and of the content of the grab samples (Morrison *et al.*, 2003). The grain-size analysis results are averaged for the three ground-truthing sites falling in each category. The analysis results include the mean grain size and sorting of the <2 mm content (both in phi scale), and the percentage content in volume of clay, silt, sand, and gravel-size (>2 mm) particles.

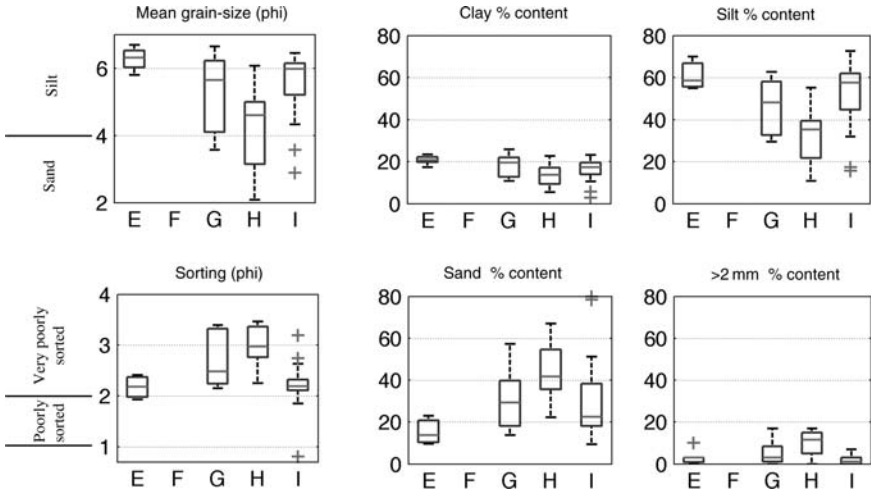

Figure 3. Boxplots describing the content of the 2005 samples within each SSS acoustic class. Measures displayed are the mean grain size and sorting of the <2 mm content (both in phi scale), and the percentage content in volume of clay, silt, sand, and gravel-size (>2 mm) particles.

area from its central west to northeast, and category L to the high-reflective features dominating the northeast corner of the map.

The 2008 video survey confirmed the quasi-homogeneous sediment distribution of the zone, as observed in the previous ground-truth surveys. All videos showed areas completely covered in soft mud, with a variable cover of burrows and shells or shell fragments. This general observation was confirmed by analysis of the four sediment samples, which yielded a similar content primarily dominated by clay-size particles bound into medium-silt-size particles by organic matter. The only notable variation between samples was the size of the >2 mm fraction, which was entirely made up of shell fragments, always. Compilation of video observations for each MBES map category suggested that the cover of either shells or shell fragments was the principal difference between categories. Shell fragments were almost absent in category J, but quite frequent though dispersed in category K. Shell cover was, in contrast, very important in category L. Accordingly, the categories were further labelled as J (medium silt), K (medium silt and sparse shell fragments), and L (medium silt, shells, and shell fragments).

This video-survey analysis supported the previous analysis of the SSS map. In the context of a seabed with a highly homogeneous, very soft sediment type, it is likely that some variations in the SSS or MBS imageries were controlled by environmental factors other than grain size. The density and distribution of burrows and shell fragments, which were reported in earlier studies and confirmed in the 2008 video survey, were possible contributors through their influence on sediment-surface hardness and roughness (Stanton, 2000; Pouliquen and Lyons, 2002). However, traditional ground-truthing techniques such as grab samples or qualitative observation of video footage do not allow their density to be measured precisely, and so confirming their influence.

Here, every sample from each ground-truthing survey was used for category identification. No additional samples were available for measuring map accuracy. The uncertainty on the suitability of the selected ground-truthing techniques for some categories implies that even if more samples had been available, accuracy estimation may have been flawed. In the current state of the ground-truth surveys, it is therefore impossible to quantify the quality of the three maps. Moreover, each dataset could have been classified using different approaches to achieve better map quality, e.g. using supervised approaches or producing a different number of

categories, but quantifying this quality through the computation of map accuracy would have remained impossible.

Map-comparison results and analysis

Figure 4a shows an overlap of the SBES and SSS maps, and Table 3 is the associated contingency matrix. The comparison of the SBES and SSS maps using the measures of agreement required a single step of aggregation of two categories of the SSS map. Figure 4b shows an overlap of the SBES and MBES maps, and Table 4 is the associated contingency matrix. Comparison of the SBES and MBES maps required a single step of aggregation of two categories of the SBES map. Figure 4c is an overlap of the SSS and MBES maps, and Table 5 is the associated contingency matrix. Comparison of the SSS and MBES maps also required a single aggregation step of two categories of the SSS map, because SSS category F did not overlap with the MBES map and had, therefore, to be removed from the computations. Table 6 lists the scores obtained for each measure of agreement and association from the contingency matrices.

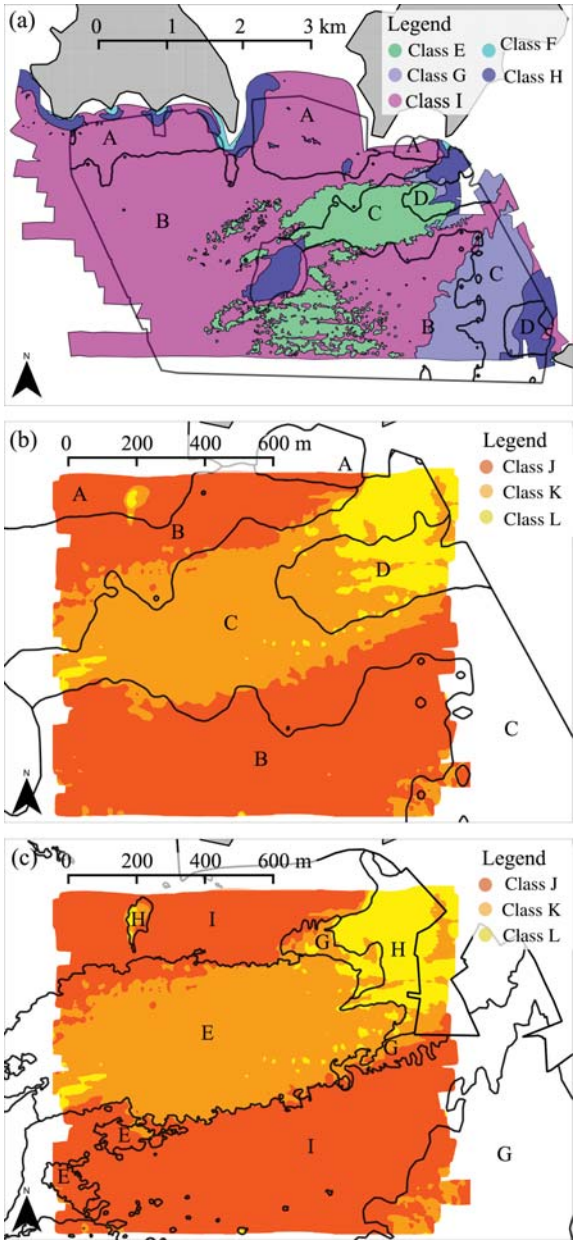

Figure 4. SBES map overlaid on (a) the SSS map and (b) the MBES map, and (c) the SSS map overlaid on the MBES map. In (a) and (b), the SBES segments of importance are labelled with their category, and in (c), the SSS segments of importance are so labelled. In (a), the SSS map categories are given in the legend, and in (b) and (c), the MBES map categories are so given.

Table 3. Contingency matrix of the SBES and SSS maps.

SBES map category	SSS map category					
	E	F	G	H	I	Total
A	0	2 401	0	37 081	690 920	730 402
B	423 370	136	189 776	11 823	2 230 261	2 855 366
C	255 631	353	291 193	85 643	109 186	742 006
D	53 079	0	44 445	86 069	2 355	185 948
Total	732 080	2 890	525 414	220 616	3 032 722	4 513 722

Table 4. Contingency matrix of the SBES and MBES maps.

MBES map category	SBES map category				
	A	B	C	D	Total
J	51 216	482 693	100 527	0	634 436
K	2 920	52 185	263 330	59 000	377 435
L	1 248	1 724	49 126	41 389	93 487
Total	55 384	536 602	412 983	100 389	1 105 358

Table 5. Contingency matrix of the MBES and SSS maps.

MBES map category	SSS map category					
	E	F	G	H	I	Total
J	71 420	0	44 249	2 397	514 254	632 320
K	315 018	0	32 826	10 937	8 763	367 544
L	11 171	0	8 359	56 981	1 288	77 799
Total	397 609	0	85 434	70 315	524 305	1 077 663

Note that the F category column is empty because this SSS category does not overlap the MBES map.

Table 6. Measures of association and measures of agreement obtained from the contingency matrices (Tables 3–5).

Compared maps and contingency matrices	Measures of association			Measures of agreement		
	V	*λ*	*U*	Max *A*	Max *κ*	Max *κ**
SBES/SSS (Table 3)	0.417	0.141	0.247	0.672[a]	0.307[b]	0.563[a]
SBES/MBES (Table 4)	0.545	0.462	0.325	0.759[c]	0.567[c]	0.638[c]
MBES/SSS (Table 5)	0.768	0.661	0.497	0.863[d]	0.746[d]	0.795[d]

For the measures of agreement, the automatic permutation/aggregation procedure was applied and only the maximum values were reported.
[a]A ∼ F, B ∼ E + I, C ∼ G, D ∼ H.
[b]A ∼ F, B ∼ I, C ∼ E + G, D ∼ H.
[c]J ∼ A + B, K ∼ C, L ∼ D.
[d]J ∼ G + I, K ∼ E, L ∼ H.

Each measure in this study provided an assessment of global map similarity in a different manner, so yielded a different range of scores

(Table 6). Some measures independently estimated different aspects of map similarity. *U*, for example, measured the amount of information shared by two maps, often showing the lowest scores, whereas *A*, which measured the overall accuracy of one map in reference to the other, had the highest scores. Other measures were related. For example, κ and κ^* systematically scored lower than *A* because they are only re-scaled versions of *A* to take into account chance agreement. In addition, κ scored systematically lower than κ^* because its estimate of chance agreement was less conservative. Despite these differences in score range, all measures were consistent in indicating the SSS and MBES maps as the most similar, and the SBES and SSS maps as the least similar (Table 6).

The next step was that of testing the influence of map size and the number of categories on the measures. As the MBES map had the fewest categories in the study, the other two maps were reduced to match that number. Using the ground-truth survey results to identify similar categories, categories A and B were aggregated in the SBES map, and categories G and H in the SSS map. After limiting all three maps to their common area, the resulting reduced MBES, SBES, and SSS maps were described by three categories each: J, K, and L for the MBES map, A + B, C, and D for the SBES map, and E, G + H, and I for the SSS map. Tables 7–9 list the contingency matrices for comparing these reduced maps, and Table 10 lists the scores obtained by the measures of categorical association and agreement on these matrices. As all the reduced maps had the same number of categories, computation of the measures of agreement did not require further category aggregation, but still required all possibilities of category permutation.

As the MBES/SBES and MBES/SSS map comparisons were already limited to the small MBES area and included automatic category aggregation to match the lowest number of categories, the map reduction was expected to have an influence only on the SBES/SSS map comparison. This was not observed (Table 10). Only λ and κ indicated that the SBES/SSS map similarity increased notably following the map reduction. The other measures only indicated a very small increase or even a decrease. The reduction actually had a clearer effect on the MBES/SSS map comparison, because all measures indicated that the map similarity decreased as a result. For the MBES/SBES map comparison, the reduction showed no influence on the measures of agreement, but mixed influence on the measures of association, where λ and *U* both increased and *V* decreased. Despite these modifications in the scores, the initial observation that the MBES and SSS maps were the most similar and that the SBES and SSS maps were the least similar remained valid after the reduction.

The very good agreement in location and extent between the SSS categories E and I and the MBES categories K and J (Figure 4c, Table 5) probably contributed to the high similarity scores attained in comparing these two maps. The decrease in similarity observed after map reduction can probably be linked to the forced aggregation of SSS categories G and H, whereas they were previously better associated with separate MBES categories, respectively, J and L (see Table 5 and legend agreement solution in Table 6).

The general confusion between SBES categories B and C and SSS categories E and G probably contributed to the low similarity scores found in comparing these two maps. The scattered SSS E segments in the south of the study site were associated with SBES category B, whereas the main SSS E segment in the centre was associated with SBES category C, which in turn was found too in the southeast in a zone dominated by SSS category G (Figure 4a). This confusion is also apparent in the detail of the optimal solutions resulting from the aggregation/permutation procedure (Table 6): SSS category E appeared better associated with SBES category B for computing the overall accuracy *A* and κ^*, but better associated with SBES category C for computing κ.

Table 7. Contingency matrix of the reduced SBES and SSS maps.

Reduced SBES map category	Reduced SSS map category			
	E	G + H	I	Total
A + B	103 546	43 727	444 701	591 974
C	240 984	82 492	79 134	402 610
D	53 079	28 622	0	81 701
Total	397 609	154 841	523 835	1 076 285

Table 8. Contingency matrix of the reduced SBES and MBES maps.

Reduced MBES map category	Reduced SBES map category			
	A + B	C	D	Total
J	533 909	98 294	0	632 203
K	55 098	257 312	54 484	366 894
L	2 967	47 004	27 217	77 188
Total	591 974	402 610	81 701	1 076 285

Table 9. Contingency matrix of the reduced MBES and SSS maps.

Reduced MBES map category	Reduced SSS map category			
	E	G + H	I	Total
J	71 420	46 646	514 137	632 203
K	315 018	43 247	8 629	366 894
L	11 171	64 948	1 069	77 188
Total	397 609	154 841	523 835	1 076 285

Table 10. Measures of association and measures of agreement obtained from the contingency matrices (Tables 7–9).

Compared maps and contingency matrices	Measures of association			Measures of agreement		
	V	λ	*U*	Max *A*	Max κ	Max κ^*
SBES/SSS (Table 7)	0.423 (+1%)	0.378 (+168%)	0.212 (−14%)	0.664[a] (−1%)	0.423[a] (+38%)	0.496[a] (−12%)
SBES/MBES (Table 8)	0.524 (−4%)	0.495 (+7%)	0.344 (+6%)	0.760[b] (∼0%)	0.560[b] (−1%)	0.641[b] (∼0%)
MBES/SSS (Table 9)	0.675 (−12%)	0.634 (−4%)	0.478 (−4%)	0.831[c] (−4%)	0.707[c] (−5%)	0.746[c] (−6%)

The percentage increase or decrease in the measures compared with their original value in Table 6 is indicated in parenthesis. For the measures of agreement, the automatic permutation procedure was applied, and only the maximum values are reported.
[a]A + B ∼ I, C ∼ E, D ∼ G + H.
[b]J ∼ A + B, K ∼ C, L ∼ D.
[c]J ∼ I, K ∼ E, L ∼ G + H.

In this study, the three maps were obtained independently and showed various differences or similarities in technology (frequency, bandwidth, beam patterns, sonar depth, operating angular sector, etc.), signal processing (calibration, acquisition gains, post-survey processing, etc.), survey design (spatial coverage and resolution), and classification methodology (features to classify, classification algorithm, and analysis design). Therefore, the observed MBES/SSS similarity and SBES/SSS dissimilarity cannot be linked to a single varying parameter, but is rather the result of the combined effects of several parameters with unknown contributions.

The main potential origins for the MBES/SSS map similarity and SBES/SSS map dissimilarity are the map resolution and coverage. The MBES and SSS maps relied on high-resolution, full-coverage acoustic imageries, whereas the SBES map relied on a dataset with such a lower resolution that it required interpolation to be compared with the other maps. The interpolation means that most of the content of the SBES map is predicted rather than measured and that one should remain critical of its results (Foster-Smith and Sotheran, 2003). A second potential explanation is the systems' respective operating angular sectors (Michaels, 2007). The SSS operated from very low to mid-range grazing angles, in the 1–40° range under the assumption of a flat seabed. The MBES operated from low to very-high grazing angles, in the 25–90° range. The SBES operated at very-high grazing angles only, in the 86–90° range. As the contributions of both surface roughness and volume heterogeneity backscattering processes vary considerably with grazing angle (Lurton, 2002), particularly the former in the 70–90° range, perhaps some spatial changes in seabed characteristics are detectable in a signal recorded at certain angles, but invisible at other angles. Therefore, the separation of the SBES and SSS operating angle sectors could lead to different aspects of the seabed being measured, and the partial overlap of MBES and SSS angle sectors may increase the chance that these two systems measure the same seabed variations. A third possible explanation is the choice of the features used for classification. Both MBES and SSS maps were obtained from the classification of the amplitude of their respective signals, which translated into image tone and texture. In comparison, the SBES map was obtained from the classification of three unknown *Q*-values, which can be any of the 166 features the QTC software extracted from the SBES signal cumulative amplitude, amplitude quantiles and histogram, power spectrum, and wavelet packet transform (Preston *et al.*, 2004a). This difference in number and nature of features implies that the resulting SBES map could be based on different seabed characteristics from those of the MBES and SSS maps (Simard and Stepnowski, 2007).

Similar hypotheses can be formulated to explain the greater similarity of SBES with MBES than with SSS. For instance, it is less likely that SBES and MBES measured different seabed characteristics because their operating angular sectors overlap. In addition, the SBES and MBES maps were obtained from a similar automatic clustering-classification algorithm, whereas the SSS map was obtained from subjective interpretation. The first approach the first approach is insensitive to the spatial distribution of the features, but the second implies some degree of spatial analysis as a result of the capabilities of the human brain for object and texture recognition (Russ, 2007).

Conclusions

Three benthic habitat maps covering the same site were created from different acoustic datasets, but the size and the design of the ground-truth surveys rendered estimation of their accuracy impossible. However, a direct map-to-map comparison was possible and performed. Several techniques for map-to-map comparison exist, but in this case, a set of measures for a map pixel-to-pixel comparison originating from the fields of statistics and land remote sensing was used. This approach did not allow any conclusions to be drawn on the accuracy of individual maps, but it did permit estimates to be made of how much the different systems/processing methodologies led to similar results which were, in summary, that the MBES and SSS maps were essentially similar, whereas the SSS and SBES maps were not similar.

The basis for classification of SSS and MBES was their imagery, which appeared to be similar (Schimel *et al.*, 2010). The similarity measured between their respective segmentations confirms this and supports the argument that MBES imagery, even at a lesser resolution, is a viable alternative to SSS imagery to segmentation.

The hypothesis that a SBES map could be representative of different seabed characteristics from those appearing on SSS maps has been suggested in previous comparative studies, which advised that the two systems should be run in tandem so that the output map can benefit from such a multisystem approach (Foster-Smith *et al.*, 2004; Brown *et al.*, 2005; Anderson *et al.*, 2008). The low similarity measured here between the SBES and SSS maps supports this argument. However, it remains unclear whether most of the dissimilarity observed is created by potential SBES map artefacts resulting from its lower resolution or by genuinely different mapped seabed characteristics.

Estimating the respective accuracy of the SBES and SSS maps could have helped clarify this ambiguity. All this shows that, despite the benefits, a map-to-map comparison approach cannot replace the value of a well-designed ground-truth survey accompanying all acoustic-mapping effort and hence allowing estimation of map accuracy and its variance (Foody, 2002, 2009; Anderson *et al.*, 2008). As far as possible, the map-accuracy comparison and map-to-map comparison approaches should be performed together in analyses of overlapping maps.

It is important to note that this study was limited to a specific quasi-homogeneous soft-sediment coastal environment, a specific resolution, and specific segmentation methodologies, so its conclusions must be viewed in this context. Only the repetition of such multisystem experimental comparative studies in different environments would help extend the range of the conclusions.

A wide range of comparative studies in seabed mapping would benefit from the measures presented here, or from other map-comparison tools used in land remote sensing. In contrast to this study, particular focus could be on reducing the variability in the origin of the maps to target the similarity study. For example, comparing maps obtained from

(i) a unique system's output classified with various segmentation methodologies would specifically address the similarity between methodologies;

(ii) different datasets, but classified using a unique segmentation methodology, would specifically estimate the complementarity of different datasets;

(iii) a unique system and methodology, but acquired at different times, would facilitate monitoring the changes at a given site over time;

(iv) a unique system, segmentation methodology, and survey, but classified with different legends in supervised mode, would specifically address the issue of similarity between different classification schemes.

Acknowledgements

We thank Remy Zyngfogel (MetOcean Solutions Ltd), Mark Morrison, Jim Drury (NIWA), Clinton Duffy, Kala Sivaguru (Department of Conservation), Jacinta Parenzee, and Bryna Flaim (University of Waikato) for providing external survey results, help in data acquisition and grain-size analysis or in improving the manuscript. The paper also benefitted from the constructive comments of two anonymous reviewers. The research was conducted in association with MetOcean Solutions Ltd (New Plymouth, New Zealand) and funded by the Foundation for Research, Science and Technology (Technology in Industry Fellowship, contract number METO0602).

References

Abramowitz, M., and Stegun, I. A. 1964. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover, New York. 1046 pp.

Anderson, J. T., Gregory, R. S., and Collins, W. T. 2002. Acoustic classification of marine habitats in coastal Newfoundland. ICES Journal of Marine Science, 59: 156–167.

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

Atallah, L., Probert Smith, P. J., and Bates, C. R. 2002. Wavelet analysis of bathymetric sidescan sonar data for the classification of seafloor sediments in Hopvågen Bay – Norway. Marine Geophysical Researches, 23: 431–442.

Beyer, A., Chakraborty, B., and Schenke, H. W. 2007. Seafloor classification of the mound and channel provinces of the Porcupine Seabight: an application of the multibeam angular backscatter data. International Journal of Earth Sciences, 96: 11–20.

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

Boots, B., and Csillag, F. 2006. Categorical maps, comparisons, and confidence. Journal of Geographical Systems, 8: 109–118.

Brennan, R. L., and Prediger, D. J. 1981. Coefficient kappa: some uses, misuses, and alternatives. Educational and Psychological Measurement, 41: 687–699.

Brown, C. J., and Coggan, R. 2007. Verification methods of acoustic classes. ICES Cooperative Research Report, 286: 116–131.

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

Brown, C. J., Cooper, K. M., Meadows, W. J., Limpenny, D. S., and Rees, H. L. 2002. Small-scale mapping of sea-bed assemblages in the eastern English Channel using sidescan sonar and remote sampling techniques. Estuarine, Coastal and Shelf Science, 54: 263–278.

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

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

Collier, J. S., and Humber, S. R. 2007. Time-lapse side-scan sonar imaging of bleached coral reefs: a case study from the Seychelles. Remote Sensing of Environment, 108: 339–356.

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

Couto, P. 2003. Assessing the accuracy of spatial simulation models. Ecological Modelling, 167: 181–198.

Cramér, H. 1946. Mathematical Methods of Statistics. Princeton University Press, Princeton, NJ. 575 pp.

Cutter, G. R., Rzhanov, Y., and Mayer, L. A. 2003. Automated segmentation of seafloor bathymetry from multibeam echosounder data using local Fourier histogram texture features. Journal of Experimental Marine Biology and Ecology, 285/286: 355–370.

Diaz, R. J., Solan, M., and Valente, R. M. 2004. A review of approaches for classifying benthic habitats and evaluating habitat quality. Journal of Environmental Management, 73: 165–181.

Dungan, J. L. 2006. Focusing on feature-based differences in map comparison. Journal of Geographical Systems, 8: 131–143.

Ehrhold, A., Hamon, D., and Guillaumont, B. 2006. The REBENT monitoring network, a spatially integrated, acoustic approach to surveying nearshore macrobenthic habitats: application to the Bay of Concarneau (South Brittany, France). ICES Journal of Marine Science, 63: 1604–1615.

Ellingsen, K. E., Gray, J. S., and Bjørnbom, E. 2002. Acoustic classification of seabed habitats using the QTC VIEW^{TM} system. ICES Journal of Marine Science, 59: 825–835.

Finn, J. T. 1993. Use of the average mutual information index in evaluating classification error and consistency. International Journal of Geographical Information Science, 7: 349–366.

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

Foody, G. M. 1992. On the compensation for chance agreement in image classification accuracy assessment. Photogrammetric Engineering and Remote Sensing, 58: 1459–1460.

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

Foody, G. M. 2006. What is the difference between two maps? A remote senser's view. Journal of Geographical Systems, 8: 119–130.

Foody, G. M. 2007. Map comparison in GIS. Progress in Physical Geography, 31: 439–445.

Foody, G. M. 2008. Harshness in image classification accuracy assessment. International Journal of Remote Sensing, 29: 3137–3158.

Foody, G. M. 2009. Sample size determination for image classification accuracy assessment and comparison. International Journal of Remote Sensing, 30: 5273–5291.

Foster-Smith, R. L., Brown, C. J., Meadows, W. J., White, W. H., and Limpenny, D. S. 2001. Ensuring continuity in the development of broad-scale mapping methodology—direct comparison of RoxAnn and QTC–VIEW technologies. CEFAS Lowestoft Laboratory Contract Report, AE0908. 113 pp.

Foster-Smith, R. L., Brown, C. J., Meadows, W. J., White, W. H., and Limpenny, D. S. 2004. Mapping seabed biotopes at two spatial scales in the eastern English Channel. 2. Comparison of two acoustic ground discrimination systems. Journal of the Marine Biological Association of the UK, 84: 489–500.

Foster-Smith, R. L., and Sotheran, I. S. 2003. Mapping marine benthic biotopes using acoustic ground discrimination systems. International Journal of Remote Sensing, 24: 2761–2784.

Giri, C., Zhu, Z., and Reed, B. 2005. A comparative analysis of the Global Land Cover 2000 and MODIS land cover data sets. Remote Sensing of Environment, 94: 123–132.

Goodman, L. A., and Kruskal, W. H. 1954. Measures of association for cross classifications. Journal of the American Statistical Association, 49: 732–764.

Grace, R. V., and Whitten, R. F. 1974. Benthic communities west of Slipper Island, north-eastern New Zealand. Tane, 20: 4–20.

Hagen-Zanker, A. 2006. Map comparison methods that simultaneously address overlap and structure. Journal of Geographical Systems, 8: 165–185.

Hamilton, L. J., Mulhearn, P. J., and Poeckert, R. 1999. Comparison of RoxAnn and QTC–View acoustic bottom classification system performance for the Cairns area, Great Barrier Reef, Australia. Continental Shelf Research, 19: 1577–1597.

Heald, G. J., and Pace, N. G. 1996. An analysis of 1st and 2nd backscatter for seabed classification. Proceedings of the 3rd European Conference on Underwater Acoustics, Heraklion, Crete, Greece, 2: 649–654.

Hewitt, J. E., Thrush, S. F., Legendre, P., Funnell, G. A., Ellis, J., and Morrison, M. 2004. Mapping of marine soft-sediment communities: Integrated sampling for ecological interpretation. Ecological Applications, 14: 1203–1216.

Hughes Clarke, J. 1994. Toward remote seafloor classification using the angular response of acoustic backscattering: a case study from multiple overlapping GLORIA data. IEEE Journal of Oceanic Engineering, 19: 112–127.

Hutin, E., Simard, Y., and Archambault, P. 2005. Acoustic detection of a scallop bed from a single-beam echosounder in the St Lawrence. ICES Journal of Marine Science, 62: 966–983.

Huvenne, V. A. I., Blondel, P., and Henriet, J. P. 2002. Textural analyses of sidescan sonar imagery from two mound provinces in the Porcupine Seabight. Marine Geology, 189: 323–341.

Iampietro, P. J., Kvitek, R. G., and Morris, E. 2005. Recent advances in automated genus-specific marine habitat mapping enabled by high-resolution multibeam bathymetry. Marine Technology Society Journal, 39: 83–93.

Ierodiaconou, D., Burq, S., Reston, M., and Laurenson, L. 2007. Marine benthic habitat mapping using multibeam data, georeferenced video and image classification techniques in Victoria, Australia. Journal of Spatial Science, 52: 93–104.

Jackson, J. B. C., Kirby, M. X., Berger, W. H., Bjorndal, K. A., Botsford, L. W., Bourque, B. J., Bradbury, R. H., *et al.* 2001. Historical overfishing and the recent collapse of coastal ecosystems. Science, 293: 629–638.

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

Kostylev, V. E., Todd, B. J., Fader, G. B. J., Courtney, R. C., Cameron, G. D. M., and Pickrill, R. A. 2001. Benthic habitat mapping on the Scotian Shelf based on multibeam bathymetry, surficial geology and sea floor photographs. Marine Ecology Progress Series, 219: 121–137.

Le Bas, T. P., and Huvenne, V. A. I. 2009. Acquisition and processing of backscatter data for habitat mapping—comparison of multibeam and sidescan systems. Applied Acoustics, 70: 1248–1257.

Legendre, P., Ellingsen, K. E., Bjørnbom, E., and Casgrain, P. 2002. Acoustic seabed classification: improved statistical method. Canadian Journal of Fisheries and Aquatic Sciences, 59: 1085–1089.

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

Lucieer, V., and Lucieer, A. 2009. Fuzzy clustering for seafloor classification. Marine Geology, 264: 230–241.

Lurton, X. 2002. An Introduction to Underwater Acoustics: Principles and Applications. Springer, London. 347 pp.

Ma, Z., and Redmond, R. L. 1995. Tau-coefficients for accuracy assessment of classification of remote-sensing data. Photogrammetric Engineering and Remote Sensing, 61: 435–439.

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

McCallum, I., Obersteiner, M., Nilsson, S., and Shvidenko, A. 2006. A spatial comparison of four satellite derived 1km global land cover datasets. International Journal of Applied Earth Observation and Geoinformation, 8: 246–255.

Michaels, W. L. 2007. Review of acoustic seabed classification systems. ICES Cooperative Research Report, 286: 94–115.

Monserud, R. A., and Leemans, R. 1992. Comparing global vegetation maps with the kappa statistic. Ecological Modelling, 62: 275–293.

Morrison, M., Drury, J., Shankar, U., Middleton, C., and Smith, M. 2003. A broad scale, soft sediment habitat assessment of the Hauraki Gulf. National Institute for Water and Atmospheric Research Client Report, AKL2003-64. 56 pp.

Morrison, M. A., Thrush, S. F., and Budd, R. 2001. Detection of acoustic class boundaries in soft sediment systems using the seafloor acoustic discrimination system QTC View. Journal of Sea Research, 46: 233–243.

Pace, N. G., and Gao, H. 1988. Swathe seabed classification. IEEE Journal of Oceanic Engineering, 13: 83–90.

Pauly, D., Christensen, V., Guénette, S., Pitcher, T. J., Sumaila, U. R., Walters, C. J., Watson, R., *et al.* 2002. Towards sustainability in world fisheries. Nature, 418: 689–695.

Pikitch, E. K., Santora, C., Babcock, E. A., Bakun, A., Bonfil, R., Conover, D. O., Dayton, P., *et al.* 2004. Ecosystem-based fishery management. Science, 305: 346–347.

Pouliquen, E., and Lyons, A. P. 2002. Backscattering from bioturbated sediments at very high frequency. IEEE Journal of Oceanic Engineering, 27: 388–402.

Prada, M. C., Appeldoorn, R. S., and Rivera, J. A. 2008. Improving coral reef habitat mapping of the Puerto Rico insular shelf using side scan sonar. Marine Geodesy, 31: 49–73.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. 1992. Numerical Recipes in C: the Art of Scientific Computing. Cambridge University Press, New York. 994 pp.

Preston, J. M., Christney, A. C., Beran, L. S., and Collins, W. T. 2004a. Statistical seabed segmentation—from images and echoes to objective clustering. Proceedings of the Seventh European Conference on Underwater Acoustics, ECUA 2004, Delft, The Netherlands. 6 pp.

Preston, J. M., Christney, A. C., Collins, W. T., McConnaughey, R. A., and Syrjala, S. E. 2004b. Considerations in large-scale acoustic seabed characterization for mapping benthic habitats. ICES Document CM 2004/T: 13. 8 pp.

Rees, W. G. 2008. Comparing the spatial content of thematic maps. International Journal of Remote Sensing, 29: 3833–3844.

Reid, D. 2007. Accounting for spatial and temporal scales and interpolation in acoustic seabed classification surveys. ICES Cooperative Research Report, 286: 73–93.

Roberts, J. M., Brown, C. J., Long, D., and Bates, C. R. 2005. Acoustic mapping using a multibeam echosounder reveals cold-water coral reefs and surrounding habitats. Coral Reefs, 24: 654–669.

Russ, J. C. 2007. The Image Processing Handbook. CRC Press, Boca Raton, FL. 817 pp.

Schimel, A. C. G., Healy, T. R., McComb, P., and Immenga, D. 2010. Comparison of a self-processed EM3000 multibeam echosounder dataset with a QTC View habitat mapping and a sidescan sonar imagery, Tamaki Strait, New Zealand. Journal of Coastal Research, 26: 714–725.

Simard, Y., and Stepnowski, A. 2007. Classification methods and criteria. ICES Cooperative Research Report, 286: 61–72.

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

Siwabessy, P. J. W., Penrose, J. D., Fox, D. R., and Kloser, R. J. 2000. Bottom classification in the continental shelf: a case study for the north-west and south-east shelf of Australia. Proceedings of the Australian Acoustical Society Conference, Joondalup, Australia. 6 pp.

Stanton, T. K. 2000. On acoustic scattering by a shell-covered seafloor. Journal of the Acoustical Society of America, 108: 551–555.

Stehman, S. V. 1999. Comparing thematic maps based on map value. International Journal of Remote Sensing, 20: 2347–2366.

Stehman, S. V. 2006. Design, analysis, and inference for studies comparing thematic accuracy of classified remotely sensed data: a special case of map comparison. Journal of Geographical Systems, 8: 209–226.

Stehman, S. V., and Czaplewski, R. L. 1998. Design and analysis for thematic map accuracy assessment: fundamental principles. Remote Sensing of Environment, 64: 331–344.

The Royal Forest and Bird Protection Society. 1998. Te Matuku marine reserve application Waiheke Island. The Royal Forest and Bird Protection Society of New Zealand Inc., Wellington. 31 pp.

Theil, H. 1972. Statistical Decomposition Analysis with Applications in the Social and Administrative Sciences. North-Holland, London. 337 pp.

Walker, B. K., Riegl, B., and Dodge, R. E. 2008. Mapping coral reef habitats in southeast Florida using a combined technique approach. Journal of Coastal Research, 24: 1138–1150.

White, R. 2006. Pattern based map comparisons. Journal of Geographical Systems, 8: 145–164.

Zajac, R. N., Lewis, R. S., Poppe, L. J., Twichell, D. C., Vozarik, J., and Digiacomo-Cohen, M. L. 2000. Relationships among sea-floor structure and benthic communities in Long Island Sound at regional and benthoscape scales. Journal of Coastal Research, 16: 627–640.

doi:10.1093/icesjms/fsq102