Vol. 552: 61–70, 2016
doi: 10.3354/meps11775

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published June 23

Comparison of image annotation data generated by multiple investigators for benthic ecology

Jennifer M. Durden[1,2,*], Brian J. Bett[1], Timm Schoening[3], Kirsty J. Morris[1], Tim W. Nattkemper[4], Henry A. Ruhl[1]

[1]National Oceanography Centre, European Way, Southampton SO14 3ZH, UK
[2]Ocean and Earth Science, University of Southampton, National Oceanography Centre Southampton, European Way, Southampton SO14 3ZH, UK
[3]GEOMAR Helmholtz Centre for Ocean Research Kiel, 24148 Kiel, Germany
[4]Biodata Mining Group, Faculty of Technology, Bielefeld University, 33501 Bielefeld, Germany

ABSTRACT: Multiple investigators often generate data from seabed images within a single image set to reduce the time burden, particularly with the large photographic surveys now available to ecological studies. These data (annotations) are known to vary as a result of differences in investigator opinion on specimen classification and of human factors such as fatigue and cognition. These variations are rarely recorded or quantified, nor are their impacts on derived ecological metrics (density, diversity, composition). We compared the annotations of 3 investigators of 73 megafaunal morphotypes in ~28 000 images, including 650 common images. Successful annotation was defined as both detecting and correctly classifying a specimen. Estimated specimen detection success was 77 %, and classification success was 95 %, giving an annotation success rate of 73 %. Specimen detection success varied substantially by morphotype (12–100 %). Variation in the detection of common taxa resulted in significant differences in apparent faunal density and community composition among investigators. Such bias has the potential to produce spurious ecological interpretations if not appropriately controlled or accounted for. We recommend that photographic studies document the use of multiple annotators and quantify potential inter-investigator bias. Randomisation of the sampling unit (photograph or video clip) is clearly critical to the effective removal of human annotation bias in multiple annotator studies (and indeed single annotator works).

KEY WORDS: Expert knowledge · Scoring · Visual imaging · Multiple investigators · Data quality · Quality assurance/quality control

INTRODUCTION

Visual imaging is increasingly used to assess the marine environment, particularly in ecological studies of the deep sea. In recent years, improvements to subsea and photographic technologies have made the collection of high-quality underwater photographs and video more efficient (e.g. Morris et al. 2014, Durden et al. in press). This has typically resulted in a substantial increase in the number of photographs captured in a single sampling event. For example, at the well-photographed Porcupine Abyssal Plain (PAP) sustained observatory, the number of seabed images captured rose by 2 orders of magnitude, from ~3000 photographs acquired by a conventional towed camera in 2011 (Durden et al. 2015) to ~300 000 photographs captured using an autonomous underwater vehicle in 2012 (Morris et al. 2014, Milligan et al. 2016).

Data extraction from photographs or videos (referred to as 'annotation') is still largely a manual process, with automated annotation processes (e.g.

*Corresponding author: jennifer.durden@noc.soton.ac.uk

Publisher: Inter-Research · www.int-res.com

Schoening et al. 2012) still requiring the input of human annotations or human-mediation of their output. The determination of biological metrics is generally made through manual classification, counting and/or sizing of specimens of interest. This is time-consuming, particularly with large sets of photographs, when annotation encompasses a large group of visually diverse organisms. One approach to acquire data faster is for multiple investigators to annotate images within a single image set.

The use of multiple investigators to assess visual data in ecological studies is believed to be common, but often not an acknowledged aspect of the method. Recently developed frameworks, such as CATAMI (Althaus et al. 2015), and software tools, such as the Video Annotation and Reference System (Schlining & Stout 2006) and BIIGLE (Schoening et al. 2009), have been designed to facilitate the manual annotation of image sets by groups of investigators. In contrast to other methods involving multiple visual assessors, such as crowd-sourced image annotation, investigators often work alone, and consensus among investigators is rarely employed to reduce potential error or bias in the data. However, some studies have found that the differences among experts in biological visual classifications could drastically alter the assessed diversity of a community (e.g. Gobalet 2001). Inconsistencies in taxonomic classification among experts have been documented in studies of both physical samples and imagery (e.g. Culverhouse et al. 2014, Howell et al. 2014). Image annotators also suffer from systematic biases as a result of human factors in visual tasks (Culverhouse et al. 2014). Such factors include organism size: First & Drake (2012) found that the success in detecting plankton was related to size. Several time-related factors affect human performance in visual tasks, including increased errors from an increased speed of labelling and an increased continuous period performing the task (Culverhouse et al. 2014). Howell et al. (2014) suggested that humans learn as they annotate, so annotation reliability increases with experience. Fatigue and boredom potentially decrease human performance in visual tasks, up to 70% after 30 min of work (Colquhoun 1959). Other psychological factors affecting human performance in visual tasks include short-term memory limits, recency effects and positivity bias (Evans 1987). Thus, the methods of annotation should be thoroughly documented in all aspects and planned specifically to be robust to the biases introduced by human investigators.

Here we expand on previous studies of investigator agreement in image-based annotation that have evaluated small numbers of morphotypes in small image datasets (Schoening et al. 2012, Howell et al. 2014). We examine the data quality of human annotations of a large image dataset, encompassing a large number and variety of epibenthic megafauna, as assessed by multiple investigators. First, we directly compare annotation data from 3 investigators of a common set of photographs to quantify the accuracy of their detection and classification of the megafauna. We then assess what impact any differences among investigators may have on resulting ecological metrics, such as density, diversity and composition. We consider 2 factors that may influence investigator detection and classification success: the size of specimens, and time spent annotating. Finally, we recommend methods to reduce human annotator bias.

MATERIALS AND METHODS

Study design and data collection

The megabenthos of the PAP (48° 50′ N, 16° 30′ W; 4850 m water depth), in the vicinity of the sustained observatory (Hartman et al. 2012), was studied. Seabed images were captured within a 1 km^2 area of level seafloor, distant from any significant seabed topography, such that no systematic variations in the density, diversity or taxonomic composition of the megafauna were expected.

Approximately 30 000 vertical photographs were collected in a grid survey, at altitudes of 1.9 to 4.1 m above the seabed, using a 5 megapixel Point Grey Research Inc. Grasshopper 2 camera with a 2/3′ sensor mounted in a downward orientation on the autonomous underwater vehicle Autosub6000 during a single deployment from the RRS ‘Discovery’ research cruise 377 in July 2012 (Ruhl et al. 2012). Images were processed to correct illumination, and for pitch, roll and yaw of the vehicle, then mosaicked into strips of 10 consecutive images (referred to as ‘tiles’; total 2849), using the methodology detailed in Morris et al. (2014).

Tiles were annotated for 73 benthic megafaunal morphotypes (>1 cm in dimension, sensu Grassle et al. 1975; listed in Supplement 1 at www.int-res.com/articles/suppl/m552p061_supp.pdf), which were identified to the lowest taxonomic level possible, counted and measured using a custom-built macro in the image analysis software ImagePro Plus (Media Cybernetics). These morphotypes represented those found in images and trawls in previous studies of the area (Billett et al. 2010, Durden et al. 2015) and comprised taxa from 9 phyla, having a wide range of body sizes (Supplement 1) and morphologies. General categories (e.g. 'Unspecified') were used when an investigator could not assign the specimen to a more detailed taxonomic level. Investigators were all previously experienced in the detection and classification of benthic invertebrates in seabed photographs, rather than specialists in particular taxonomic groups. To ensure agreement on morphological features of the fauna, a catalogue of the megafauna of the area was consulted, and the investigators annotated an initial group of 100 common tiles (not included in the subsequent analysis) and discussed their classifications (e.g. Howell et al. 2014). The morphotype data were assessed at 2 levels: the finest taxonomic resolution achieved, and combined into 16 higher taxonomic categories, previously used by Billett et al. (2010).

The tiles were randomised and divided among 3 investigators for annotation (see Table 1; Supplement 2 at www.int-res.com/articles/suppl/m552p061_supp.pdf). Of these, a sub-group of 65 were randomly selected to be annotated by all 3 investigators (referred to as 'common tiles'). The remaining tiles (referred to as the 'large tile set') were divided for annotation by a single investigator only. The tiles assigned to each investigator (including tiles from the large tile set and the common tiles) were randomised prior to annotation. Annotation was halted periodically, and a subset of classifications to that point was reviewed to ensure continued agreement on morphological characterization of taxa. Following such reviews, existing annotations were revised to reflect changes in the agreed classification and to add newly recognised morphotypes.

Comparison of common tiles

'Annotation success' was defined as a combination of both detection and classification of a given specimen and was assessed in the common tiles (see example in Supplement 2). 'Detection success' (DS) was computed as the number of specimens detected by an investigator as a fraction of the total number of specimens detected by any investigator (n). We estimated the true number of specimens present, including the probability of joint non-detection (nd) by all 3 investigators ($nd = (1 - DS)^3$), as $N = (1 + nd) \times n$. The corrected detection success was then estimated as n/N. 'Classification success' was calculated as the number of specimens that were identically classified by all investigators as a fraction of the number of specimens detected by all 3 investigators. 'Annotation success' was computed as the number of specimens that were both detected and identically classified by all investigators as a fraction of the total number of specimens detected by at least one investigator.

Ecological metrics

To evaluate the impact of multiple annotators on ecological metrics, annotation data from individual tiles were aggregated to produce groups of replicate samples (Supplements 2 & 3) as follows:

(1) Common tiles. In the 'common tiles' annotated by all 3 investigators, tiles were randomly assigned to 1 of 4 replicates. Note that a single randomisation was applied across the investigators, such that data comparisons among investigators represented repeated measures.

(2) Large tile set. In the 'large tile set' (all tiles excluding the 'common tiles'), tiles were randomly assigned to replicates of ~100 tiles, to yield a typical sample size of 900 specimens per replicate. A further group of 10 replicates was randomly selected from the large tile set, without regard to the identity of the investigator, to serve as an example of multi-investigator data.

(3) Small tile set. For each investigator, a set of 65 tiles was randomly selected from the portion of the large tile set annotated and assigned to 4 replicates to match the treatment of the common tiles. A further set was selected without regard for investigator identity to represent multi-investigator data.

Specimen counts were converted to densities using the calculated area of seabed represented by each tile. Instances where only a portion of a specimen was visible in an image were counted as 0.5 in terms of abundance. Density data in each tile were $\log(x + 1)$ transformed prior to parametric statistical analyses and were assessed per replicate set and reported as geometric mean density and 95% confidence interval. Densities in common tile replicates were compared among investigators using repeated measures ANOVA to account for all investigators annotating the same tiles and with conventional ANOVA for the large tile set. Univariate diversity indices (Shannon H'_2 and Simpson Index of Diversity $1\text{-}D$; e.g. Magurran 2013) were calculated using the vegan package in R (Oksanen et al. 2012). The expected number of morphotypes was calculated by rarefaction using EstimateS (Colwell 2013). Total specimen counts per morphotype were rounded up to the nearest integer prior to diversity calculations.

Differences in the apparent community composition among investigators were assessed using multivariate statistics (Bray-Curtis dissimilarity measure and 2-dimensional non-metric multidimensional scaling ordination), with comparisons tested using ANOSIM and SIMPER routines implemented with PRIMER6 (Clarke & Warwick 2008). Faunal data were subject to a range of transformations (none, log(x+1) and presence–absence) prior to the calculation of dissimilarity measures to assess different aspects of inter-investigator variations (e.g. detection and classification success). To assess the potential impact of rare taxa, community analyses were also completed using only morphotypes that were recorded in all replicates by all investigators (6 morphotypes in the common tiles and small tile set, 16 morphotypes in the large tile set; Supplement 1).

Quantifying bias and precision in ecological metrics

Precision within an investigator's annotations was quantified by calculating the coefficient of variation of the univariate measures for each set of tiles. The bias of an investigator's annotations was estimated using the overall mean among investigators for each parameter. Bias in community composition estimates was calculated using ANOSIM analyses, and precision was assessed using the autosimilarity method described by Schneck & Melo (2010). In brief, Bray-Curtis dissimilarity was computed between 2 groups of 'x' tiles randomly selected without replacement from the large tile set (where x = 1, 2, … half the number of tiles in the set), facilitating an assessment of the impact of sample size (number of tiles) on the apparent value and precision of faunal similarity estimates.

Human factors in annotation

Spearman's rank correlations (r_S) between morphotype characteristics (size and number of individuals) and annotation success were investigated with the annotations from the common tiles, using morphotypes with more than one successful annotation. The median size dimension of each morphotype in pixels was converted to an area, either as a circle with the diameter of this dimension, or as dimension^2 × 0.25 for elongate morphotypes (see Supplement 1).

Time-related biases in annotation were assessed by comparing faunal densities to the time spent annotating a tile. The time at completion of annotation was extracted from the timestamp stored with each tile. The total time spent annotating tiles was estimated using the number of tiles and median time spent per tile. Note that time spent compiling the database of morphotypes for annotation was not considered, nor was time spent training the investigators on the trial 100 tiles. The morphotypes *Amperima/Ellipinion/Kolga*, *Iosactis vagabunda* and Ophiuroidea were selected for detailed analysis as a result of their high densities and differing detection rates in the common tiles.

RESULTS

Direct comparison of annotations in common tiles

In total, the 3 investigators made 1648 annotations in the common tiles, approximately equally split among investigators (Table 1). A total of 692 distinct specimens were detected by at least one investigator: 399 were detected by all investigators (58%), 146 (21%) by 2 investigators, and 147 (21%) by 1 investigator only. The apparent detection success was relatively consistent among investigators (74–82%). The mean detection success was 78%, yielding an apparent joint non-detection probability of 0.01, equating to some 7 potentially undetected specimens. The mean corrected detection success was 77% (73–81% across the investigators). The corrected probability of detection of a specimen by all investigators was calculated as 47% (322 specimens), appreciably lower than the fraction of specimens actually detected by all 3 investigators. These data suggested that there was significant variation in the detectability of individual morphotypes but that detection success was relatively consistent among investigators across the fauna as a whole.

Morphotype discrimination among investigators was largely consistent: investigators found similar numbers of morphotypes (Table 1), although none of the investigators found all morphotypes, and 5 morphotypes were recorded by only 1 investigator (Supplement 1). Of the 399 specimens detected by all 3 investigators, full agreement of the classification occurred in 378 cases (95%). Of the 21 cases where full agreement was not achieved, 2 investigators agreed in 18 cases (5%), and all 3 investigators disagreed in 3 cases (<1%). Combining detection success (77%) and classification success (95%), overall annotation success was estimated to be 73%. At the higher taxonomic group level, classification success was 98%, and annotation success was 75%.

Table 1. Ecological metrics calculated by replicate tile set and investigator. Area, density, and Shannon and Simpson's indices are given as geometric means (and 95 % confidence interval). Richness was rarefied (n = 500 for common tiles and small tile set, n = 5000 for large tile set). Annots: number of annotations; M: number of morphotypes; Multi: Multi-investigator

	Annots	M	Area (ha)	No of ind.	Density (ind. ha^{-1})	Shannon index (H'_2)	Simpson's index (1-D)	Rarefied richness $E(S)_n$
Common tiles								
Investigator 1	541	40	0.0203 (0.0196, 0.0211)	535.5	6562 (5937, 7211)	2.75 (2.42, 3.09)	0.71 (0.63, 0.77)	39.3 (35.8, 42.9)
Investigator 2	565	37	0.0203 (0.0196, 0.0211)	560.5	6889 (6462, 7328)	2.78 (2.65, 2.92)	0.71 (0.68, 0.75)	35.6 (31.8, 39.9)
Investigator 3	542	41	0.0203 (0.0196, 0.0211)	535.5	6574 (5948, 7224)	2.76 (2.44, 3.08)	0.71 (0.66, 0.75)	39.9 (33.4, 46.5)
Small tile set								
Investigator 1		42	0.0203 (0.0196, 0.0211)	548	6662 (5534, 7873)	2.68 (2.24, 3.12)	0.68 (0.59, 0.77)	40.5 (32.9, 48.0)
Investigator 2		42	0.0204 (0.0197, 0.0212)	545	6648 (6465, 7145)	2.76 (2.58, 2.95)	0.70 (0.68, 0.71)	40.7 (35.3, 46.1)
Investigator 3		37	0.0205 (0.0197, 0.0213)	575.5	6698 (6548, 7460)	2.62 (2.37, 2.87)	0.68 (0.62, 0.74)	35.1 (27.7, 42.6)
Multi		41	0.0204 (0.0197, 0.0210)	557	6749 (6155, 7364)	2.91 (2.70, 3.11)	0.73 (0.70, 0.75)	40.7 (36.8, 44.6)
Large tile set								
Investigator 1		65	0.1235 (0.1232, 0.1238)	6812	6889 (6523, 7263)	2.92 (2.83, 3.00)	0.70 (0.68, 0.72)	63.0 (60.4, 65.6)
Investigator 2		65	0.1253 (0.1247, 0.1259)	8106.5	6466 (6260, 6676)	2.92 (2.84, 3.00)	0.71 (0.70, 0.72)	61.2 (55.2, 67.1)
Investigator 3		68	0.1230 (0.1212, 0.1249)	8179.5	6472 (6163, 6787)	2.96 (2.90, 3.03)	0.71 (0.70, 0.72)	63.3 (58.3, 68.3)
Multi		71	0.1259 (0.1254, 0.1264)	8352.5	6633 (6556, 6712)	2.97 (2.91, 3.03)	0.71 (0.70, 0.72)	66.2 (60.8, 71.6)

Detection success varied greatly among morphotypes (12–100 %; Table 2). Of the 147 specimens detected by a single investigator, the most common morphotypes were Ophiuroidea (28 %), *Iosactis vagabunda* (20 %), Indeterminate - 'Tube-dwelling invertebrate' (13 %), and *Amperima/Ellipinion/Kolga* (8 %). Assessed at the higher taxonomic level, detection success ranged from 14 % (Octocorallia) to 66 % (Actiniaria).

Table 2. Annotation success by morphotype (defined as organisms detected and identified similarly by all 3 investigators as a fraction of instances of the organism annotated by at least 1 investigator) in the common tiles. Only morphotypes with more than 1 annotation by all 3 investigators are listed. Indet.: Indeterminate

	Annotation success (%)	
<50	50–80	>80
Amperima/Ellipinion/Kolga	*Bathycrinus* sp.	*Oneirophanta* sp.
Actiniaria sp. 9	*Iosactis vagabunda*	*Molpadiedemas villosus*
Porifera type 3	*Amphianthus* sp.	
Ophiuroidea	Porifera type 4	
Indet. - 'Hydroid'	*Psychropotes longicauda*	
Indet. - 'Scaphopod'	Aphroditid	
Indet. - 'Tube-dwelling invertebrate'	Stalked tunicate	
	Cerianthid sp. 1	
	Cerianthid sp. 3	
	Echiura	

Impact to ecological metrics

Faunal density

There was no significant difference in the estimated faunal density among investigators in the common tiles (repeated measures ANOVA, $p > 0.05$; Table 1), or in the small tile set (ANOVA, $p > 0.05$). However, an investigator bias was detected in the large tile set (ANOVA $F_{3,34} = 8.9$, $p < 0.001$). The coefficients of variation for density decreased with an increase in sample size, while bias in the density estimates increased with sample size (maximal bias per investigator in the small tile set was 0.4, and 4.2 in the large tile set; Table 3).

Morphotype diversity

A total of 44 morphotypes were recorded in the common tiles, 53 in the small tile set, and 73 in the large tile set (Supplement 1). The number of morphotypes recorded was generally consistent among investigators (Table 1). Shannon and Simpson's diversity indices were not significantly different among investigators in the small or large tile sets (ANOVA, $p > 0.05$), nor was rarefied richness (Table 1; Supplement 4 at www.int-res.com/articles/suppl/m552p061_supp.pdf). As the sample size increased, the coefficients of variation and bias in the diversity indices and estimated richness decreased (Table 3).

Table 3. Precision (coefficient of variation = CV, %) and bias (%) in the density, diversity metrics and richness determined, as introduced by investigators. Multi: Multi-investigator

	Density		Shannon index		Simpson index		Rarefied richness	
	CV	Bias	CV	Bias	CV	Bias	CV	Bias
Common tiles								
Investigator 1	8.1	−1.5	10.3	−0.5	8.3	0.0	4.6	2.7
Investigator 2	5.3	3.1	4.1	0.6	4.4	0.0	5.8	−7.0
Investigator 3	8.1	−1.5	9.8	−0.1	5.9	0.0	8.4	4.3
Small tile set								
Investigator 1	14.5	−0.1	13.9	−0.2	11.5	−1.0	9.5	4.5
Investigator 2	6.3	−0.3	5.6	2.8	2.0	1.9	6.7	5.0
Investigator 3	5.5	0.4	8.0	−2.5	7.6	−1.0	10.9	−9.5
Multi	15.0	−1.1	6.1	8.3	1.9	6.3	4.9	5.0
Large tile set								
Investigator 1	3.7	4.2	4.4	−0.5	3.4	−0.9	2.1	0.8
Investigator 2	1.9	−2.2	4.8	−0.5	3.2	0.5	4.9	−2.1
Investigator 3	3.1	−1.5	3.9	0.9	2.6	0.5	4.0	1.3
Multi	2.0	0.4	3.5	1.6	2.4	0.5	4.2	5.9

Faunal composition

Faunal composition was not significantly different among investigators in the common tiles (log(x + 1)-transformed data; Fig. 1a; ANOSIM, $p > 0.05$), but significant differences in apparent faunal composition were detected among investigators in both the small and large tile sets (Fig. 1b,c). In the small tile set, a significant difference among investigators existed when density was allowed the maximum contribution (untransformed data; ANOSIM R = 0.16, $p < 0.05$), and in the log(x + 1)-transformed data (ANOSIM R = 0.18, $p < 0.05$) but not in the presence–absence data (ANOSIM, $p > 0.05$), suggesting that differences among investigators were primarily related to differences in the estimated densities of some taxa (i.e. variations in detection success). The significant difference among investigators in the large tile set was apparent whether based on presence-absence (ANOSIM R = 0.20, $p < 0.001$), log(x + 1)-transformed-data (ANOSIM R = 0.53, $p < 0.001$) or untransformed data (ANOSIM R = 0.51, $p < 0.001$). This result suggests that investigator bias in faunal composition is magnified with sample size, for example, for log(x + 1)-transformed data the ANOSIM R value (effect size) increased from 0.18 to 0.53 as replicate sample size increased from 65 to 100 tiles (Fig. 2). Differences in similarity among experts became significant at ~500 tiles, while precision in estimated community composition (as mean within-investigator community similarity) for the multi-investigator data was not fully asymptotic at 1400 tiles (14 000 photos, Fig. 2).

Density-driven variations in apparent faunal composition were detected among investigators. The morphotype contributing most (>10 %) to the dissimilarity among investigators in the common tiles and the small tile set was *Iosactis vagabunda*, with Ophiuroidea and *Amperima/Ellipinion/Kolga* contributing at least 10 % to dissimilarity. These 3 morphotypes were the highest ranked in terms of density and contributed at least 5 % to dissimilarities among investigators in the large tile set. The bias among investigators persisted even when the analysis was restricted to common morphotypes, in both the log(x + 1)-transformed small tile set (6 morphotypes; ANOSIM R = 0.18, $p < 0.05$) and the large tile set (Fig. 1; 16 morphotypes; ANOSIM R = 0.44, $p < 0.001$). These results suggest that variation in detection success among investigators is the primary cause of apparent variations in faunal composition. A comparison of the densities of these key taxa estimated by different investigators corroborates this: significant differences were found in the estimated density of Ophiuroidea in the small and large tile sets (ANOVA $F_{3,12} = 5.59$, $p < 0.05$ and $F_{3,34} = 18.79$, $p < 0.0001$, respectively). Investigators also differed significantly in the estimated density of *Amperima/Ellipinion/Kolga* in both the small (ANOVA $F_{3,12} = 11.27$, $p < 0.001$) and large tile sets (ANOVA $F_{3,34} = 14.88$, $p < 0.0001$). In the case of *Iosactis vagabunda*, investigators differed significantly in the estimated density in the large tile set (ANOVA $F_{3,34} = 3.57$, $p < 0.05$) but not in the small tile set.

Human factors

Annotation success in the common tiles was not significantly related to the pixel area of a morphotype ($p > 0.05$), nor to the total number of specimens of a particular morphotype ($p > 0.05$). Estimated total faunal density was not significantly correlated with time spent per tile in either the common tiles, the small tile set, or the large tile set (all $p > 0.05$; Supplement 5 at www.int-res.com/articles/suppl/m552p061_supp.pdf). However, the estimated density of Ophiuroidea in the large tile set was significantly corre-

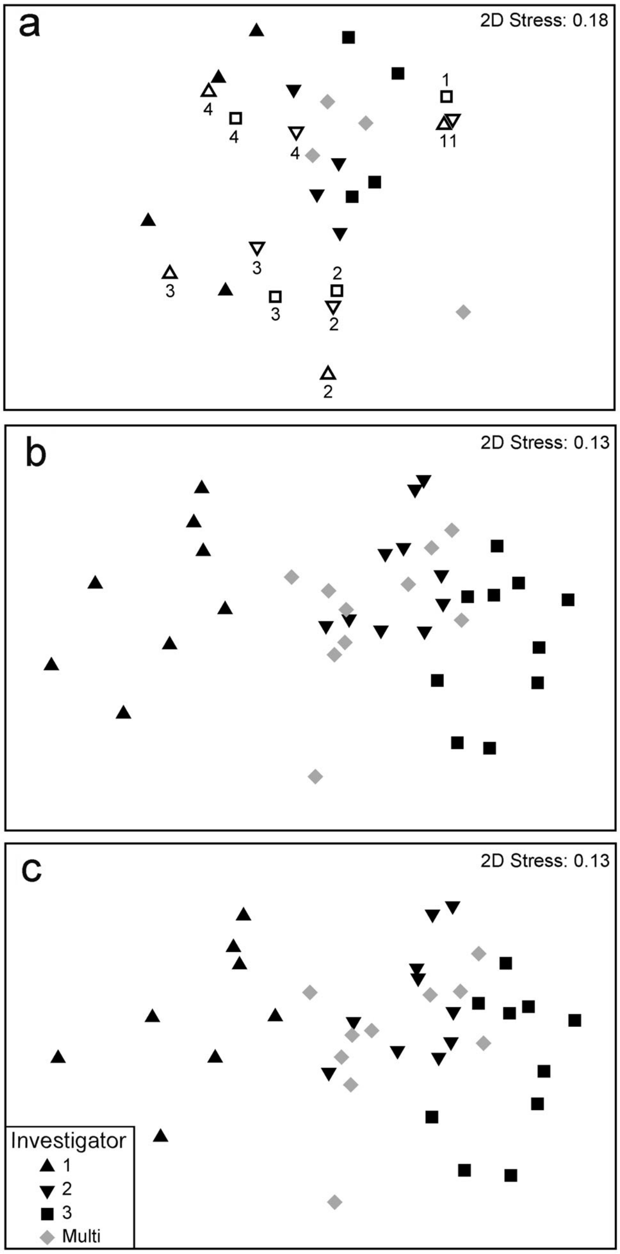

Fig. 1. Two-dimensional non-metric multidimensional scaling ordination plots of $\log(x + 1)$-transformed megafaunal community data as represented by investigators in (a) the common tiles (open symbols, with replicates labeled 1 to 4) and the small tile set (by replicate including the multi-investigator data and all morphotypes); and the large tile set by replicate, (b) including the multi-investigator data and all morphotypes and (c) including only the 16 morphotypes found in every replicate

lated with time spent per tile ($r_S[28] = 0.48$, $p < 0.05$; Fig. 3). Time spent per tile differed significantly among investigators in both the common tiles and the large tile set (ANOVA $F_{2,11} = 12.9$, $p < 0.01$ and $F_{2,27} = 38.4$, $p < 0.001$; Supplement 5).

DISCUSSION

The direct comparison of annotations among investigators revealed that it was specimen detection, rather than classification, that was the primary source of bias in the annotation data. Although investigators had high and similar specimen detection rates overall, the variable detection success of individual morphotypes introduced significant scope for bias. In the present study, classification success was high. Nevertheless, it was clear that without appropriate randomisation of images among investigators, the use of multiple annotators could introduce (statistically backed) illusory ecological conclusions.

While some variation in annotation among investigators is expected, it is the magnitude of the potential bias that requires consideration. The classification success rate in the common tiles (95%) was higher than that found in a study of fewer number of morphotypes by Culverhouse et al. (2003; 43%). In another study, a wide range of inter-observer agreement was found for 13 morphotypes assessed by 5 investigators (0–97%; Schoening et al. 2012). Culverhouse et al. (2003) suggested that some investigators are more consistent at categorisation (classification) while inconsistent at counting (detection), or vice versa, and that bias may be related to increased familiarity with certain morphotypes by particular investigators. This familiarity has been suggested to result in investigators assigning more detailed classification to some groups of species than others (Gobalet 2001) and may result in some faunal groups being more successfully annotated by some investigators than others, as we encountered in the present study.

Investigator bias on ecological metrics was not apparent in the common tiles; that is, no significant differences among investigators were detected in terms of density, diversity or community composition. However, significant differences in density and community composition were detected in the small and large tile sets, exceeding background ecological variation. These differences appeared to be driven by low detection success in the common morphotypes, rather than differences in the detection of rare morphotypes. In addition, it is worth noting that the use of multiple investigators resulted in a small positive bias of species richness; for example, in both the small and large tile sets, rarefied morphotype richness was inflated by 5% over the single investigator average.

The megafaunal assemblage of the PAP is one of high dominance by a few morphotypes. The dominant morphotype (*Iosactis vagabunda*) contributed 55% of the individuals annotated, while Ophiuroidea

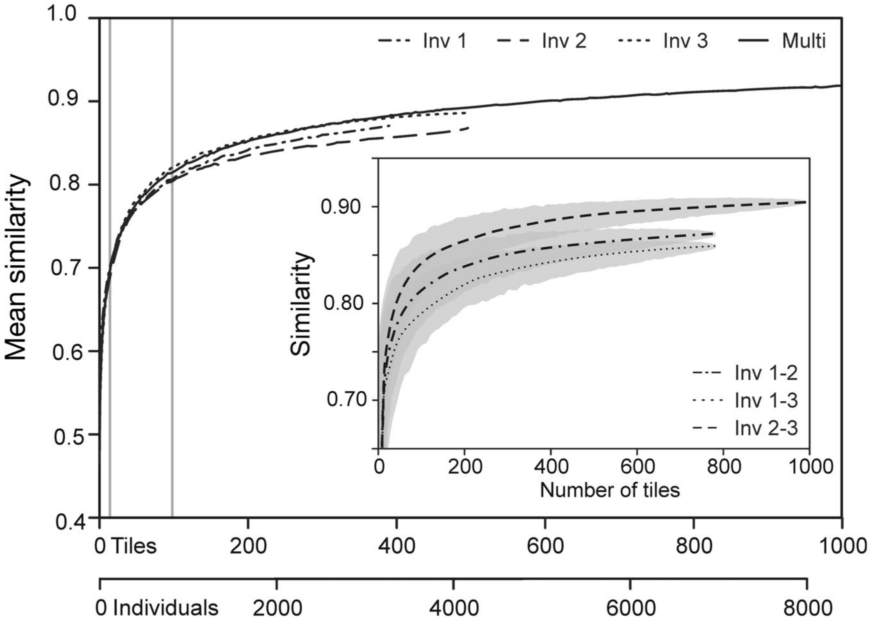

Fig. 2. Precision in the community structure determined by each investigator (Inv 1–3), and in the multi-investigator data (Multi), computed as autosimilarity of log(x + 1)-transformed data. Grey vertical lines indicate the numbers of tiles in replicates of each of the small and large tile sets (16 and 100). Inset: Inter-investigator similarity, with 95% confidence intervals

and *Amperima*/*Ellipinion*/*Kolga* contributed a further 12 and 9%, respectively, despite these 3 morphotypes all having low detection successes. Small variations in the detection success of these morphotypes substantially influenced apparent density and community composition.

Human factors exerted some influence on the resulting annotation data. The time spent annotating was positively correlated with the number of Ophiuroidea annotated. Time is an important consideration in all studies based on annotations by humans, regardless of the number of annotators. It is a known issue with human-based data gathering. Correlations between time spent annotating and apparent density (e.g. for Ophiuroidea) are consistent with Megaw's (1979) suggestion that time available was a source of error in visual inspection tasks.

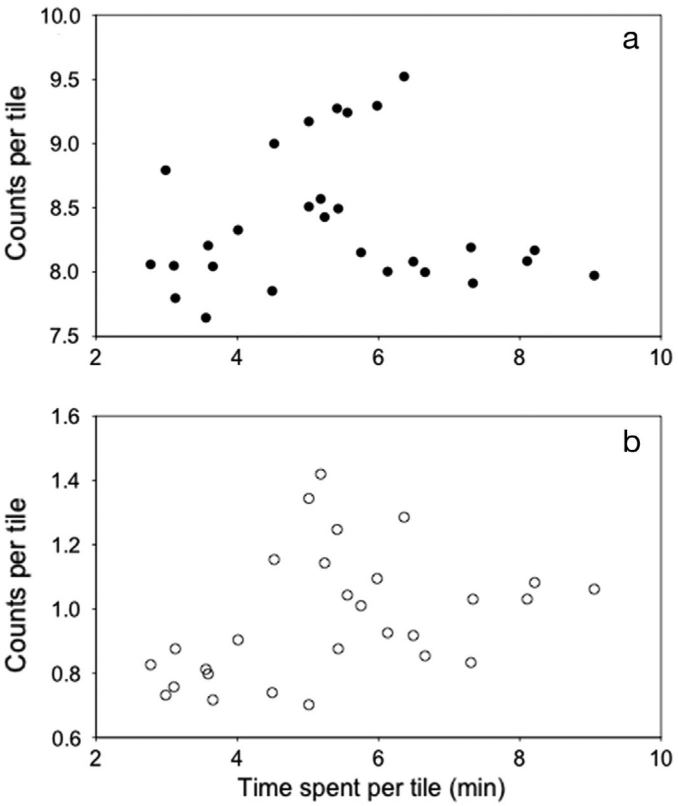

Fig. 3. Human factors influencing annotation data. Counts per tile of (a) all megafauna and (b) Ophiuroidea with time spent annotating in replicates of the large tile set. Note difference in y-axis scale

Given the apparent importance of variable specimen detection success in driving inter-investigator bias, computer-aided specimen detection may be of particular value. Schoening et al. (2012) developed the iSIS (intelligent Screening of underwater Images System) software that detected more megafauna in a set of training images than were in the set of gold standard annotations generated by 5 investigators, suggesting that the system could be used to generate a set of detected objects for humans to review and accept or reject. Such human-mediated machine annotation could allow human investigators to focus on the classification of the detected objects, a task achieved with high success (95%) in the present study.

Having identified the primary source of bias as variable specimen detection, both among morphotypes and to a lesser extent among investigators, the nature and occurrence of bias in particular ecological parameters can be readily understood. Apparent density is impacted only by specimen detection rates, with bias becoming evident at large sample sizes. Apparent diversity is less impacted by specimen detection rates, because diversity indices deal with relative rather than absolute abundances. Low inter-investigator bias in species relative abundance estimates and very low non-consensus rates in classification success do, nevertheless, combine to slightly inflate diversity estimates in multi-investigator data.

In contrast, apparent faunal composition is directly impacted by both specimen detection (where density is or relative abundance are included in the similarity estimates) and classification success.

Sample size-related change in the relative levels of within-investigator variation and among-investigator bias in annotation data has important implications for the conduct of image-based ecological studies. We note that many published studies that have employed seabed photography deal with sample areas equivalent to that of the common tiles or small tile set of the present study (e.g. Soltwedel et al. 2009, De Leo et al. 2010), where the impact of bias may be low. However, the complete large tile set studied here is representative of the very extensive sets of photographs now becoming available (e.g. Morris et al. 2014, Wynn et al. 2014), particularly for ecological studies of the benthos, and thus there may be significant issues with bias and precision in annotation data.

CONCLUSIONS

It is clear that different annotators give different results on either per image or per annotation bases and that these differences may impact the ecological metrics (i.e. community structure) derived from this annotation data. The use of multiple annotators is a reasonable way to reduce annotation time in large image or video datasets, but the 'effect size' of investigator bias is likely to increase with sample size, such that it may become a particular concern in the large datasets. The statistical power to detect change increases with sample size (and number of replicates); consequently the risk of spurious ecological interpretations of investigator bias is similarly increased. In addition, all human investigators may introduce bias to annotation data as a result of human factors, such as time, as found here.

We recommend the following actions to reduce investigator-related bias in image annotation data (whether video clips or still images) and to allow fair comparisons to other annotation data generated by single or multiple annotators:

(1) Randomise the order of image or video clips annotation, both in single and multiple-investigator studies. This reduces both annotator bias in studies with multiple investigators and time-related bias in studies with either single- or multiple-investigator annotation. Annotator or time-related bias becomes spatial or temporal bias in studies where contiguous blocks of images are annotated.

(2) Quantify potential inter-investigator bias by directly comparing annotators in a randomly selected subset of the imagery, as has been done in the present study and also suggested by Howell et al. (2014).

(3) Report the use of multiple investigators and the inter-investigator agreement achieved, including the detection success and classification consensus.

(4) When comparing to existing data, consider the apparent ecological effect size and potential bias effect size.

Acknowledgements. We thank the captain, crew and scientists of the RRS 'Discovery' research cruise 377 and the Autosub6000 engineers, in particular, S. McPhail, M. Furlong and M. Pebody. This work was funded by the UK Natural Environment Research Council (NERC) and contributes to the Autonomous Ecological Surveying of the Abyss project (NERC NE/H021787/1), the NERC Marine Environmental Mapping Programme (MAREMAP), and the Porcupine Abyssal Plain – Sustained Observatory (NERC National Capability programme). T.S. was supported by the Bundesministerium für Wirtschaft und Energie (BMWI) as part of the IS2U project (FKZ 03SX344A).

LITERATURE CITED

Althaus F, Hill N, Ferrari R, Edwards L and others (2015) A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: the CATAMI classification scheme. PLoS ONE 10:e0141039

Billett DSM, Bett BJ, Reid WDK, Boorman B, Priede IG (2010) Long-term change in the abyssal NE Atlantic: the 'Amperima Event' revisited. Deep-Sea Res II 57:1406–1417

Clarke KR, Warwick RM (2008) Changes in marine communities: an approach to statistical analysis and interpretation. PRIMER-E, Plymouth

Colquhoun WP (1959) The effect of a short rest-pause on inspection efficiency. Ergonomics 2:367–372

Colwell RK (2013) EstimateS: statistical estimation of species richness and shared species from samples. Version 9. User's guide and application published at: http://purl.oclc.org/estimates

Culverhouse PF, Williams R, Reguera B, Herry V, González-Gil S (2003) Do experts make mistakes? A comparison of human and machine indentification of dinoflagellates. Mar Ecol Prog Ser 247:17–25

Culverhouse PF, Macleod N, Williams R, Benfield MC, Lopes RM, Picheral M (2014) An empirical assessment of the consistency of taxonomic identifications. Mar Biol Res 10:73–84

De Leo FC, Smith CR, Rowden AA, Bowden DA, Clark MR (2010) Submarine canyons: hotspots of benthic biomass and productivity in the deep sea. Proc R Soc B 277: 2783–2792

Durden JM, Bett BJ, Jones DOB, Huvenne VAI, Ruhl HA (2015) Abyssal hills — hidden source of increased habitat heterogeneity, benthic megafaunal biomass and diversity in the deep sea. Prog Oceanogr Part A 137:209–218

Durden JM, Schoening T, Althaus F, Friedman A and others (in press) Perspectives in visual imaging for marine biology and ecology: from acquisition to understanding. In:

Hughes RN, Hughes DJ, Smith IP, Dale AC (eds) Oceanography and marine biology: an annual review, Book 54. CRC Press, Boca Raton, FL

Evans JStBT (1989) Bias in human reasoning: causes and consequences. Laurence Erlbuam, Hove

➤ First MR, Drake LA (2012) Performance of the human 'counting machine': evaluation of manual microscopy for enumerating plankton. J Plankton Res 34:1028–1041

➤ Gobalet KW (2001) A critique of faunal analysis; inconsistency among experts in blind tests. J Archaeol Sci 28: 377–386

➤ Grassle JF, Sanders HL, Hessler RR, Rowe GT, Mclellan T (1975) Pattern and zonation — study of bathyal megafauna using research submersible *Alvin*. Deep-Sea Res 22:457–481

➤ Hartman SE, Lampitt RS, Larkin KE, Pagnani M and others (2012) The Porcupine Abyssal Plain fixed-point sustained observatory (PAP-SO): variations and trends from the Northeast Atlantic fixed-point time-series. ICES J Mar Sci 69:776–783

➤ Howell KL, Bullimore RD, Foster NL (2014) Quality assurance in the identification of deep-sea taxa from video and image analysis: response to Henry and Roberts. ICES J Mar Sci 71:899–906

Magurran AE (2013) Measuring biological diversity. Blackwell Publishing, Oxford

➤ Megaw ED (1979) Factors affecting visual inspection accuracy. Appl Ergon 10:27–32

Milligan RJ, Morris KJ, Bett BJ, Durden JM and others (2016) High resolution study of the spatial distributions of abyssal fishes by autonomous underwater vehicle. Sci Rep 6:26095

Milligan RJ, Morris KJ, Bett BJ, Durden JM and others (in press) First high resolution study of the spatial distributions of abyssal fish. Nat Sci Rep

Morris KJ, Bett BJ, Durden JM, Huvenne VAI and others (2014) A new method for ecological surveying of the abyss using autonomous underwater vehicle photography. Limnol Oceanogr Methods 12:795–809

Oksanen J, Blanchet FG, Kindt R, Legendre P and others (2012) vegan: community ecology package. Version 2.0−5. https://cran.r-project.org/package=vegan

Ruhl H (principal scientist) and others (2012) *RRS James Cook* Cruise 62, 24 Jul–29 Aug 2011. Porcupine Abyssal Plain– sustained observatory research, National Oceanography Centre, Southampton. Southampton Cruise Report 12:119

Schlining BM, Stout NJ (2006) MBARI's video annotation and reference system. Proc OCEANS 2006. IEEE, Boston, MA

➤ Schneck F, Melo AS (2010) Reliable sample sizes for estimating similarity among macroinvertebrate assemblages in tropical streams. Ann Limnol Int J Limnol 46:93–100

Schoening T, Ehnert N, Ontrup J, Nattkemper TW (2009) BIIGLE Tools — a web 2.0 approach for visual bioimage database mining. In: Banissi E, Stuart L, Wyeld TG, Jern M and others (eds) Proc 2009 13th International Conference Information Visualisation. IEEE, Barcelona, p 51–56

➤ Schoening T, Bergmann M, Ontrup J, Taylor J and others (2012) Semi-automated image analysis for the assessment of megafaunal densities at the Arctic deep-sea observatory HAUSGARTEN. PLoS ONE 7:e38179

➤ Soltwedel T, Jaeckisch N, Ritter N, Hasemann C, Bergmann M, Klages M (2009) Bathymetric patterns of megafaunal assemblages from the arctic deep-sea observatory HAUSGARTEN. Deep-Sea Res I 56:1856–1872

➤ Wynn RB, Huvenne VA, Le Bas TP, Murton BJ and others (2014) Autonomous underwater vehicles (AUVs): their past, present and future contributions to the advancement of marine geoscience. Mar Geol 352:451–468

*Editorial responsibility: Charles Peterson,
Morehead City, North Carolina, USA*

*Submitted: December 10, 2015; Accepted: May 13, 2016
Proofs received from author(s): June 10, 2016*