Ecological Indicators 140 (2022) 109029

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Original Articles

The annotation approach used for marine imagery impacts the detection of temporal trends in seafloor biota

Nicholas Perkins *, Zelin Zhang, Jacquomo Monk, Neville Barrett

Institute for Marine and Antarctic Studies, University of Tasmania, Private Bag 49, Hobart TAS 7001, Australia

ARTICLE INFO

Keywords:
Autonomous underwater vehicle
Benthic monitoring
Indicator selection
Mesophotic ecosystems
Size distribution
Spatial modeling

ABSTRACT

Image-based surveys of the marine benthos are being increasingly adopted as a monitoring tool for habitats and biota, particularly in mesophotic depths (~30–150 m) which are technically difficult to survey. Many modern tools for these surveys, such as remotely operated vehicles and autonomous underwater vehicles, can capture thousands of images in a single deployment. Turning this into quantitative data typically involves human annotation which is often time-consuming and costly. Percent cover of organisms is one of the most common metrics for monitoring changes in abundance, which may be attained through visual estimation, digitization, or point-count approaches. However, alternative metrics of abundance such as density (direct counts) or presence-absence, as well as metrics that quantify condition (e.g., bleaching) and size-structure can also provide quantitative information for tracking change. Understanding the statistical power of different approaches is critical to designing effective image-based monitoring programs. Given the differing statistical power and time taken using different approaches, program managers need to decide where to allocate resources. Here we use benthic imagery from two long-term monitoring sites in south-eastern Australia to annotate three example morphospecies (morphologically distinct organisms) using three approaches: point-count (percent cover), full count and presence-absence within imagery. Also, we compare the performance of the point count and full count approaches for monitoring bleaching in one of our morphospecies. We use spatio-temporal models to quantify trends in the empirical data and simulations to quantify the power of these approaches to detect different levels of temporal change using different sampling efforts (either 100 or 200 images). Additionally, we examine the additional insights that size-structure information can provide for two morphospecies. We find that the full count approach provides a higher statistical power to detect change than the other approaches for our example morphospecies, including tracking bleaching status. Size-structure information can provide additional insights such as the occurrence of recruitment or mortality events or growth of individuals. We recommend that monitoring programs using benthic imagery should consider the choice of annotation approach as this is likely to impact observed temporal patterns, particularly when the focus is specific indicator species rather than total biodiversity.

1. Introduction

Image-based surveys are being increasingly adopted for benthic monitoring programs, particularly as interest expands into deeper mesophotic (~30–150 m) ecosystems (e.g. Bridge et al., 2011; Karpov et al., 2012; Enrichetti et al., 2019). Technologies such as autonomous underwater vehicles (AUVs) and remotely operated vehicles (ROVs) are particularly attractive for monitoring mesophotic ecosystems due to their ability to operate for extended periods at depth, collecting large amounts of imagery over *meso*-scales (100 s m – km). In many regions these depths are being surveyed for the first time, and the data collected has provided initial insights into the biological communities present. While automated approaches hold great potential, sufficient training data for machine learning algorithms first needs to be acquired (Schoening et al., 2016). Building up biodiversity inventories across newly surveyed regions is time-consuming, and consequently, only a small subset of imagery is typically annotated. As these monitoring programs begin to conduct repeat surveys, the emphasis shifts from baseline descriptions to tracking changes through time. For the purposes of tracking trends, indicator species or groups are often chosen to act as surrogates

* Corresponding author.
E-mail address: Nicholas.Perkins@utas.edu.au (N. Perkins).

https://doi.org/10.1016/j.ecolind.2022.109029
Received 13 March 2022; Received in revised form 17 May 2022; Accepted 31 May 2022
Available online 6 June 2022

for overall ecosystem responses (Niemi and McDonald, 2004; Van Rein et al., 2009). This shift in focus means that different sampling approaches and associated metrics can be explored and assessed, with the potential to acquire more information that may allow earlier detection of changes or additional insights into ecosystem dynamics. Understanding the implications of different potential approaches, both in their ability to track changes and in terms of resource allocation is crucial for management decision-making.

Tracking trends in the abundance of indicator species has long been the cornerstone of many ecological monitoring programs (Noss, 1990; Niemi and McDonald, 2004; Butler et al., 2012). For mesophotic ecosystems where there is a lack of historical baselines, tracking change in species is important to build an understanding of ecosystem dynamics and how species naturally fluctuate in abundance through time. Understanding scales of spatial and temporal variation is crucial for informing the suitability of species or groups of species to be used as indicators for ongoing monitoring (Larsen et al., 2001; Perkins et al., 2017). For example, species with high 'natural' variability are likely to have fluctuations in abundance that may not be related to stressors of interest and thus detecting chronic impacts caused by stressors will take longer. Therefore, tracking trends in the abundance of species in mesophotic ecosystems will remain high priority in the early stages of long term monitoring. However, there are several potential metrics for tracking trends in abundance with imagery, such as quantifying percent cover, direct counts of individuals (densities), and presence-absence (PA) within images. Here we explore the trade-offs between these different approaches to quantifying abundance, as well as how other metrics may be beneficial in detecting change.

For many image-based benthic monitoring programs, percent cover across a surveyed area is the most common method used to quantify and track trends in indicator species or groups (e.g., Ryan, 2004; Drummond and Connell, 2005; Monk et al., 2018; Pawlik et al., 2022). This method was adopted from approaches developed for terrestrial plant surveys, and in the marine environment, initially used in intertidal and shallow SCUBA-based surveys. Percent cover can be calculated by visual estimation, digitization approaches, the overlay of grids, or most commonly by the overlay of points. In the point-count approach, points are overlain across images and the species (or group) under each point are annotated, with the proportion of points falling on a species/group being used to quantify cover. Cover is then used as a surrogate for the abundance or biomass of target organisms across the survey area. Studies comparing percent cover estimates with both biomass (Chiarucci et al., 1999) and frequency counts (Parravicini et al., 2010) show that percent cover can act as a proxy for both abundance and biomass, with similar trends detected for both individual species and community metrics.

A number of studies have looked into sampling design issues around the point count approach (e.g. Ryan, 2004; Perkins et al., 2016; Montilla et al., 2020), making recommendations around the level of sampling required, with sample sizes often being prohibitively high when target species are rare or extremely patchy. Research in mesophotic depths in temperate regions is highlighting that most species are rare in terms of cover (<2%; e.g. Monk et al., 2018; Perkins et al., 2020), indicating relatively high sampling will be required to track changes in individual species reliably using this methodology (Perkins et al., 2016). Furthermore, research has shown that tracking cover alone may mask important underlying ecological processes that may be of interest or provide alternative metrics for tracking trends (e.g., Brito-Millán et al., 2019). For example, the mortality of larger individuals may be balanced by the recruitment of many smaller individuals resulting in stable cover, or growth shrinkage or fusion of colonies may occur in the case of corals, and so changes in abundance and underlying population dynamics may be missed when quantifying cover alone.

There has been an emerging push to understand key demographic process such as rates of mortality and recruitment, physiological, and genetic responses to help understand the impacts of stressors on ecosystem functioning and health (Bell et al., 2017). The need to incorporate a wider range of metrics than just abundance of individual species when monitoring benthic communities has been identified by a number of authors (e.g. Cánovas-Molina et al., 2016; Enrichetti et al., 2019), especially when the aim is to detect chronic impacts in species that are long-lived (Bell et al., 2017). Alternative metrics for tracking change in benthic indicator species with imagery may include measurements of ecosystem functioning and health such as measurement of the size of individuals (lengths or areas) and the recording of the condition of individuals or colonies (e.g., the level of bleaching in corals). Metrics which quantify condition provide additional information and may be useful for tracking perturbations such as warming events that may cause bleaching or die-offs. Size information can help quantify the occurrence of recruitment events, growth, and mortality or changes in these which may indicate chronic effects of stressors on populations.

Monitoring program goals will ultimately dictate the choice of indicators and metrics. For example, if monitoring climate change impacts is key, then abundance of indicator species sensitive to warming may be chosen or metrics which measure ecosystem health such as the proportion of bleaching in indicator species or groups may prove more effective. However, in mesophotic depths where baseline knowledge of system dynamics and life-history traits of species is often lacking, gathering additional insights through exploration of different metrics early in a monitoring program will aid in helping researchers and managers to decide on the most appropriate metrics for tracking trends. An assessment of the time taken to annotate data using different approaches can then allow a cost-benefit analysis across metrics. However, other aspects such as the magnitude of future changes, and spatial and temporal variability also need to be considered when exploring different approaches.

Simulation-based "virtual ecologist" (Zurell et al., 2010) approaches allow the exploration of the impact of different sampling designs, effect sizes and statistical analyses on monitoring outcomes. Such approaches have the added benefit of allowing an exploration of how different sources of variance influence monitoring outcomes and can harness information from baseline surveys to estimate these sources of variance. For example, differing levels of sampling error, the spatial distribution of species, natural temporal variation, or differences in responses between sites can be incorporated into simulations and their effects quantified. In particular, the statistical power to detect different magnitudes of change can be quantified, information that is crucial to ensuring the effectiveness of monitoring programs. Despite this, power analyses for monitoring programs are surprisingly rare and often do not examine the implications of the use of different metrics for tracking change.

Here we use a time series of benthic imagery collected by an AUV in the early stages of a long-term monitoring program to explore the use of different annotation approaches and their ability to track change. We use data collected from different annotation approaches for three example morphospecies (morphologically distinct taxa readily identified in imagery) that are reasonably abundant in cross-shelf regional studies across south-eastern Australia in mesophotic depths (30 – 150 m) and quantify their statistical power of detecting simulated levels of change. We contrast the more typically used point-count annotation approach, taking full counts of individuals, and annotating the PA of morphospecies within images. We also test differing site-level sampling effort, using either 100 or 200 images with each approach. Additionally, we examine how the metrics of condition and size structure provide extra information when tracking change by exploring bleaching in one of our morphospecies and size structure in two morphospecies. We provide information regarding the time taken for these different approaches, allowing for a cost-benefit analysis of different methods. We aim to highlight the large amount of often unharnessed information available from image-based benthic surveys and provide management recommendations for the allocation of resources for ongoing monitoring.

2. Methods

To illustrate how different image annotation approaches (point counts to calculate percentage cover, full counts, PA, condition, and size structure) perform, we use three example morphospecies annotated in AUV imagery collected between 2011 and 2017 at two long-term monitoring sites in the Australian Marine Parks (AMPs) on the east coast of Tasmania, Australia (Fig. 1). Our three example morphospecies are a conspicuous and abundant cup sponge ("Cup red smooth") with a mean diameter of 11.9 ± 7.5 cm, and a small (approximately 5 – 10 cm tall, but not measured in this study) bramble coral ("Bramble coral") at the Flinders site; and a massive sponge ("Massive Purple") of roughly oblong shape with a mean length along the long axis of 17.4 ± 8.6 cm at the Freycinet site. All three morphospecies have been considered as potential indicators for ongoing monitoring in the region because they are widely distributed and thus are likely to experience a gradient of pressure such as climate change impacts, and can be quantified with reasonable precision (Perkins et al., 2020). Furthermore, these three example morphospecies provide a range of contrasting distributions, both within images and across sites. The cup sponge is dominant at the Flinders site and is a relatively large morphospecies (large specimens > 20 cm in diameter), occurs in a high proportion of images, and typically with multiple individuals in each image when present. The bramble coral is a relatively small morphospecies, sometimes hard to detect, but likely to have multiple small colonies in images where present. The massive sponge is a large morphospecies, that is less dominant, and more likely to be represented as solitary individuals in an image, but often captured with point count annotating due to its larger size.

We compare how point count data (i.e., percent cover), collected as part of the baseline characterization of sites within these AMPs, compares with full counts of individuals and PA annotating when aiming to detect changes. For our cup sponge morphospecies, we also assess how tracking observed bleaching in this morphospecies compares using different approaches. Finally, for both the cup sponge and massive sponge morphospecies we explore the additional information that size structure data gives when examining temporal trends. Spatio-temporal models are used to quantify the abundance and spatial distribution of these morphospecies, and in the case of the cup sponge the occurrence of bleaching across sites in the baseline surveys. This information is then used as the basis for simulations under differing scenarios of change with

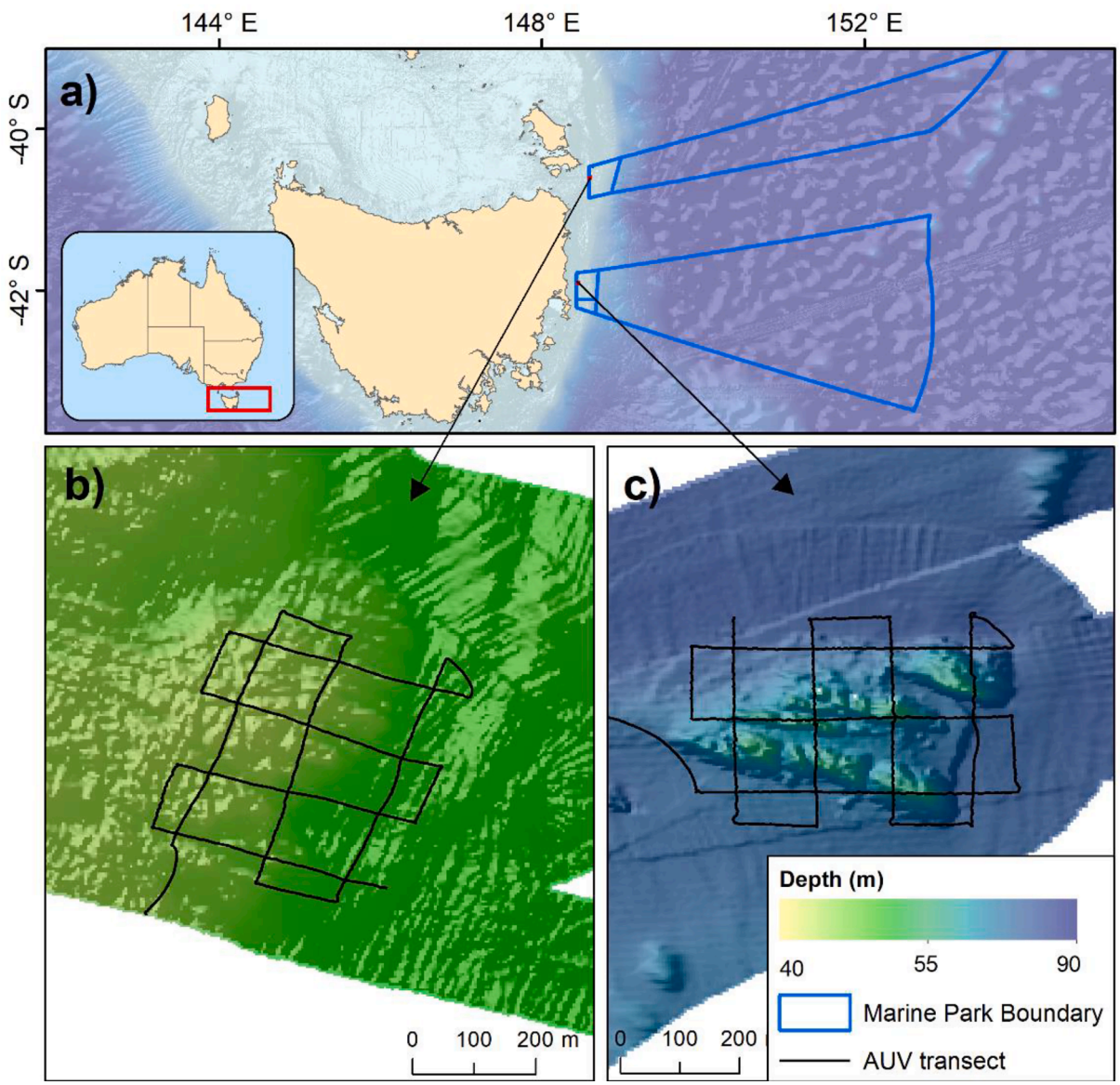

Fig. 1. Map showing survey sites: a) two Australian Marine Parks on the east coast of Tasmania, and detailed view of AUV transects in b) Flinders and c) Freycinet Marine Parks used in the study.

different sampling efforts (number of images) to allow a comparison between different annotation approaches.

2.1. Data collection and image annotation

The AUV imagery-based surveys of seabed biota have been conducted at a range of locations around Tasmania across multiple years as part of a cross-shelf monitoring program. The imagery used in this study is from surveys of two sites located within two AMPs: the Western site within Flinders Marine Park (referred to as the Flinders site) in 2011, 2013 and 2017 and a central site known as "Joe's Reef" in Freycinet Marine Park (referred to as the Freycinet site) in 2011, 2014 and 2016 (Fig. 1). All imagery used in the study is viewable on the online annotation platform Squidle (squidle.org) under the 'Targeted Scoring' dataset. Surveys at each location consist of multiple transect lines laid out in a grid pattern, with the AUV capturing overlapping stereo imagery along the transect lines. Onboard sensors allow the AUV to maintain a relatively constant altitude of ~ 2 m resulting in a stable image footprint of approximately 1.2 X 1.6 m.

All non-overlapping images that contained rocky reef were utilized for the full count and PA approaches at each site to allow trends to be examined and contrasted when using all available imagery compared to a subset. For the point count approach 200 randomly selected reef images were annotated from each site as part of routine monitoring. To compare approaches, the same subset of 200 images annotated with the point count approach was used. Two hundred images were chosen as the maximum level of sampling, as previous work has shown that this level of sampling is likely to detect changes in dominant morphospecies (Perkins et al., 2016; Perkins et al., 2020), and due to our primary aim being to compare annotation approaches rather than sampling effort. All annotations were conducted in Squidle+ (https://squidle.org/), a dedicated online annotation platform for marine benthic imagery. As all our target morphospecies are associated with rocky reef features, images that were wholly sand were also noted and excluded from being included as "annotated" images in the simulations (see below). The point count approach involved the random allocation of 25 points over the image and the annotating of all biota under each point. For the full count approach, all target individuals within each image were tagged. For the red cup sponge at Flinders, whether each individual had > 50% of its surface bleached was also recorded in both annotation approaches. The amount of time taken for each approach was recorded to allow comparison.

Size measurements were conducted for two examples: the cup red smooth morphospecies at Flinders, and the massive purple morphospecies at Freycinet. Size measurements were not conducted for bramble coral due to: (i) measurements could not be taken when looking top-down on erect colonies; and (ii) the complex morphology of this morphospecies would make area measurements problematic time consuming. Size measurements of 200 individuals were made from a random selection of images for each survey year at each site. ImageJ software (Schneider et al., 2012) was used to measure both length and area of each individual so that the quality of the two data sources could be contrasted. Length was measured as the longest diameter across the surface of each morphospecies, whereas area was calculated by tracing the perimeter of each morphospecies with a stylus on a tablet. All measurements were conducted by a single scorer to allow comparisons between time taken to be consistent for analysis.

2.2. Spatio-temporal models

Bayesian spatiotemporal models were used for estimation from the empirical data and for prediction under different sampling scenarios (new images locations), effort (number of images) and future levels of change. The Integrated Nested Laplace Approximation (INLA; Lindgren et al., 2011) approach was used. In the INLA framework, estimation and prediction are performed simultaneously. We used the last year of surveys at each site as the basis for estimation and then simulation of future change. The same Bayesian priors for spatial and temporal correlation outlined in Perkins et al. (2021a) were used.

For the point count data, a binomial model was used where each point that landed on a target individual counted as a success and the total number of points used in each image was the number of binomial trials. For the full count approach, a count-based Poisson model was used. For the occurrence of bleaching, a binomial model was once again used with each cup sponge showing bleaching within an image being a 'success' and the total number of cup sponges being the number of trials (full count approach); or the number of points landing on bleached cup sponge in an image being a 'success' and the number of points landing on the cup sponge morphospecies in an image being the number of trials (point count approach). Note that in the point count approach more than one point could land on the same individual.

2.3. Simulations

For the simulations, a random subset of 100 and 200 images was taken from all possible non-overlapping reef images along each transect. This mimics the approach taken with real imagery, whereby a larger random subset of imagery is selected than required and sand images are skipped over until the target number of reef images are annotated. This level of sampling was chosen as it is typical of the effort previously used when sampling these sites. At the randomly selected image locations, the predicted probability of presence/count of individuals within each image was subsequently adjusted to simulate varying levels of decline, from 10% to 90% in steps of 10%. Declines were simulated as the sites are subjected to marine heatwave events (see Perkins et al., 2022), and it was thought declines were more likely. Declines were simulated by adjusting the binomial probability (for cover estimation) or mean estimated count (density estimation). For cover estimation a random binomial draw was then conducted using the rbinom function in R with 25 trial draws to simulate 25 random points on an image. This simulates the random probability of a point landing on a target individual in an image. For the PA models, a logistic model was used whereby any target individual being present within an image counted as a presence.

For the bleaching of the cup sponge at our Flinders site, we first compared the estimated trajectories across the three surveys using both the full count and point count approaches using 200 images. We then examined how decreasing the number of images (down from 200 by 20 images at a time to 20 images) affected the ability to detect the change in bleaching evident between the 2013 and 2017 surveys.

For each sampling effort and annotation approach 500 simulations were conducted. Simulations were computationally intensive, and it was decided that 500 simulations would give sufficient samples to test the ability of each approach to detect change (e.g. Perkins et al., 2021b). For each simulation spatial models were refit with initial empirical data and the simulated data, and we tested whether an effect could be detected. As our models are Bayesian, we assessed whether the change could be detected by examining the posterior distribution of the 'year' effect which was treated as a categorical variable. Posterior distributions where the 95% credible interval didn't include zero, and were distributed in the direction simulated (i.e., either a decline in cover/abundance or increase in bleaching) were treated as providing evidence of change. Statistical power was calculated as the proportion of simulations where evidence of change was detected out of the total 500 simulations.

2.4. Analysis of size structure data

ImageJ returns values in terms of the number of pixels, so sizes were converted to length (mm) and area (mm^2) based on the altitude of the AUV, half the fixed field-of-view of the camera (21^0), and the number of pixels in the horizontal direction (1380) using the following formulae:

$$\text{length (mm)} = \text{pixel length}* \frac{\left(2\tan\left(\frac{\pi}{180*21}\right)*Altitude\right)*1000}{1380} \quad (1)$$

$$\text{area}\left(\text{mm}^2\right) = \text{pixel } area* \left[\frac{\left(2\tan\left(\frac{\pi}{180*21}\right)*Altitude\right)*1000}{1380}\right]^2 \quad (2)$$

As size data is often right skewed, data was log transformed prior to analysis and for plotting. Despite transformation, size data for the massive purple morphospecies still showed evidence of non-normality. Therefore, size structure data was analyzed with a non-parametric Kruskal-Wallis rank sum test, which does not assume a normal distribution, in order to test for significant differences of both length and area data on survey year. Survey year was treated as a factor. Pairwise comparison of survey years was conducted with a pairwise Wilcoxon rank sum test with a Bonferroni correction for multiple comparisons used to adjust reported p values. The skewness of the size data was examined for both morphospecies through time at each site by comparing the skewness values of length data for each year. The skewness function from the moments R package (Komsta and Novomestky, 2015) was used for all calculations. Positive skewness values indicate a predominance of larger individuals in the data, whereas negative skewness indicates a predominance of smaller individuals.

3. Results

3.1. Trends in empirical data using different approaches

The three morphospecies varied considerably in terms of their overall abundance and how common they were across images annotated at each site (Table 1). Cup red smooth was the most abundant and common morphospecies, occurring in 65–75% of reef images annotated across the three survey years. Cup red smooth was also commonly captured within images with the point count approach, with a point landing on an individual in over half of images where they were present. Bramble coral occurred in 47.0% of images in 2011 but underwent a dramatic decline in 2013 (5.4% of images) before recovering to 38.6% in 2017. The point count approach did not capture the presence within images or trends across time well for bramble coral, with a point landing on it in only about a quarter of images where it was present and the same number of points landing on it in both 2013 and 2017, despite large differences in abundance between these two years (Table 1). Massive purple sponges had relatively low but stable abundance through time but occurred in a relatively large proportion of reef images (20.7–34.1%, Table 1). Like bramble coral, one or more of the 25 points used only landed on massive purple sponges in approximately one in every four images that it occurred.

Model-based estimates of the temporal trend for the three morphospecies suggested different trends in relative abundance between the count-based approaches and the percent cover approach (Fig. 2). For bramble coral the full count approach showed strong evidence of a large decline between 2011 and 2013 and then a recovery in 2017 to similar abundance levels observed in 2011. This same trend was evident when using all images (~1000 per year, Table 1) or a random subset of 200 images. When using the point count approach for bramble coral the large decline between 2011 and 2013 was statistically detected; however, the recovery in 2017 was not, with analysis indicating a similar level of cover between 2013 and 2017. For massive purple the trend for the full count approaches were once again the same when using all images or a subset of 200 images, with an increase in abundance detected between 2011 and 2014. There was no evidence for differences in the cover of massive purple across the time series, although general trends indicated a decline in 2014 from 2011 rather than the increase found in the count-based approaches. For cup red smooth, no evidence was found of significant changes in abundance or percent cover across the survey period, with similar trends found across the annotation approaches.

3.2. The power to detect declines in abundance using different annotation approaches

In detecting decreases in cover/abundance of the three tested morphospecies, the full count approach consistently provided the highest power to detect smaller levels of change with a given level of image sampling (Fig. 3). Also, as expected, annotating 200 images consistently provided higher power than annotating 100 images in all cases, except for full count and PA annotating of the cup red smooth morphospecies where 100 and 200 images provided the same power (Fig. 3). The point count approach was able to detect smaller levels of change than the PA approach for both bramble coral and cup red smooth morphospecies when using 200 images; however, the PA approach was able to detect a 70% decline in bramble coral with 100 images, but only an 80% decline with the point count approach. PA annotating was able to detect relatively small amounts of change (20–30% declines) with high power for the massive purple morphospecies, whereas the point count approach was unable to detect any level of change simulated for the massive purple morphospecies.

The full count approach was able to detect the smallest amount of change of all approaches, with declines of > 20% for cup red smooth and massive purple morphospecies and > 30% for bramble coral being detectable with high power with both 100 and 200 images (Fig. 3). The point count approach performed best for cup red smooth, able to detect ~ 35–40% declines with high power with 200 and 100 images respectively. For bramble coral, declines of ~ 70–80% could be detected with high power with 200 and 100 images respectively. For massive purple only the largest decline (90%) could be detected with the point count approach, and only when using 200 images. Conversely, the PA approach performed best for massive purple, with declines of ~ 20–30%

Table 1
Summary statistics for different annotation approaches used in the study.

		Annotation approach						
		Percent Cover		Full Count All Images			Full Count 200 Images	
Morphospecies (Site)	Year	No. points (total = 5000)	Propn. images presence detected	No. reef images	Count of individuals	Propn. images present	Count of individuals	Propn. Images present
Cup Red Smooth (Flinders)	2011	120	36.5%	975	2138	72.6%	440	72.0%
	2013	140	43.0%	1025	2192	64.0%	417	65.0%
	2017	143	40.0%	1383	3691	75.5%	514	75.0%
Bramble Coral (Flinders)	2011	43	13.0%	975	1630	47.3%	337	47.5%
	2013	18	8.0%	1025	185	5.4%	47	6.0%
	2017	18	8.0%	1383	1904	38.6%	301	40.5%
Massive Purple (Freycinet)	2011	22	7.0%	963	311	20.7%	62	20.0%
	2014	13	4.5%	739	370	34.1%	95	33.0%
	2016	20	7.0%	1086	524	32.8%	91	30.5%

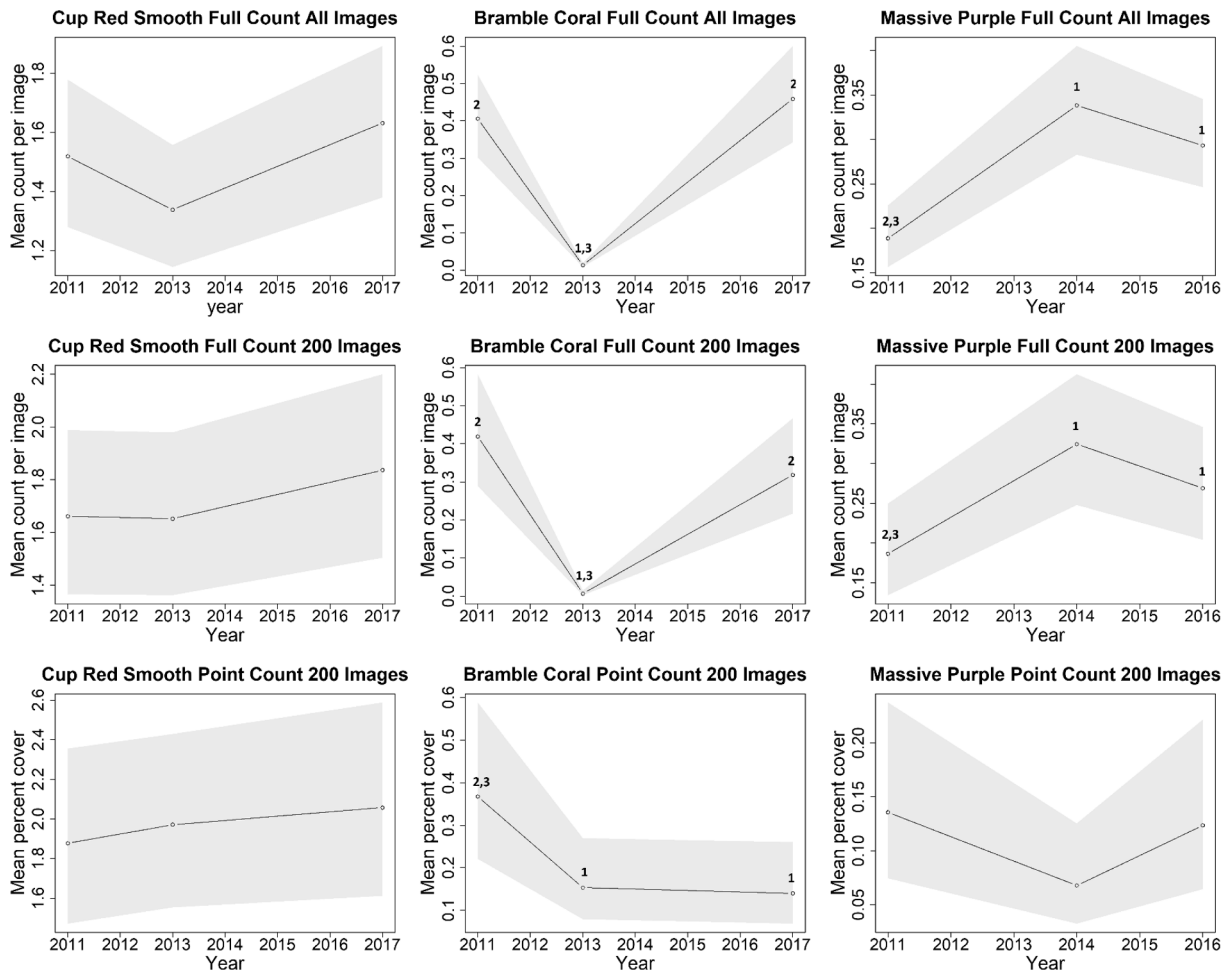

Fig. 2. Model-based estimates of trends for the three morphospecies used in the study annotated using different approaches. The line shows the mean trend and the shading the 95% credible intervals. Numbers above each time point indicate evidence was found for a difference in post hoc pairwise comparisons: 1 = a difference to the first survey time, 2 = a difference to the second survey time and 3 = a difference to the third survey time.

detectable with high power when using 200 and 100 images respectively. The PA approach was able to detect declines of ~ 65–70% with high power for bramble coral with 200 and 100 images respectively. For cup red smooth only declines of > 75% were detectable with high power using the PA approach.

3.3. Time taken using different annotation approaches

Annotation using the full count approach took on average 11 h per sample year for bramble coral and 5 h per year for cup red smooth at Flinders Western Boundary, and 10 h per year for massive purple at Freycinet. This equates to approximately 35 s per image for bramble coral, 16 s per image for cup red smooth and 39 s per image for massive purple. Annotating bramble coral took longer as colonies were often small, and images had to be scanned in more detail by annotators. Annotating massive purple also took relatively longer amount of time as the reef was more complex at the Freycinet site and so each image also had to be scanned more thoroughly. Cup red smooth is very conspicuous, and the Flinders site is relatively flat, making annotation more rapid. On average, full count annotation took 30 s per image across the three morphospecies. In contrast, the point count approach was conducted as part of annotating all biota, which takes an experienced annotator on average 5–7 min per image to annotate 25 points. Therefore, annotation with the point count approach takes approximately 10 h per 100 images. Annotating 100 images with the full count approach takes on average 50 min, which means 12 morphospecies could be annotated per 100 images in 10 h, or perhaps more if multiple morphospecies are targeted in each image. Time taken to annotate bleaching was not recorded but is likely to take longer with the full count approach as all individuals need to be assessed rather than just those the point landed on. However, this was relatively rapid for bleaching in cup red smooth, with the time per image reported here including the assessment of bleaching.

3.4. The power to detect increases in bleaching using different annotation approaches

Both the full count and point count approaches detected an increase in bleaching of red cup smooth between 2013 and 2017 when using 200 images (Fig. 4), with the full count approach suggesting a slighting larger magnitude of increased bleaching. Model-based estimates from the full count approach suggested that the proportion of bleached cup red smooth increased from approximately 8% in 2013 to 21% in 2017. Model-based results from the point count approach suggested an increase from 8% in 2013 to 16% in 2017. Decreasing the number of images sampled resulted in the full count approach being able to detect the change with high power with only 60 images, whereas the full 200 images or more are likely to be required to detect the change with the point count approach (Fig. 5).

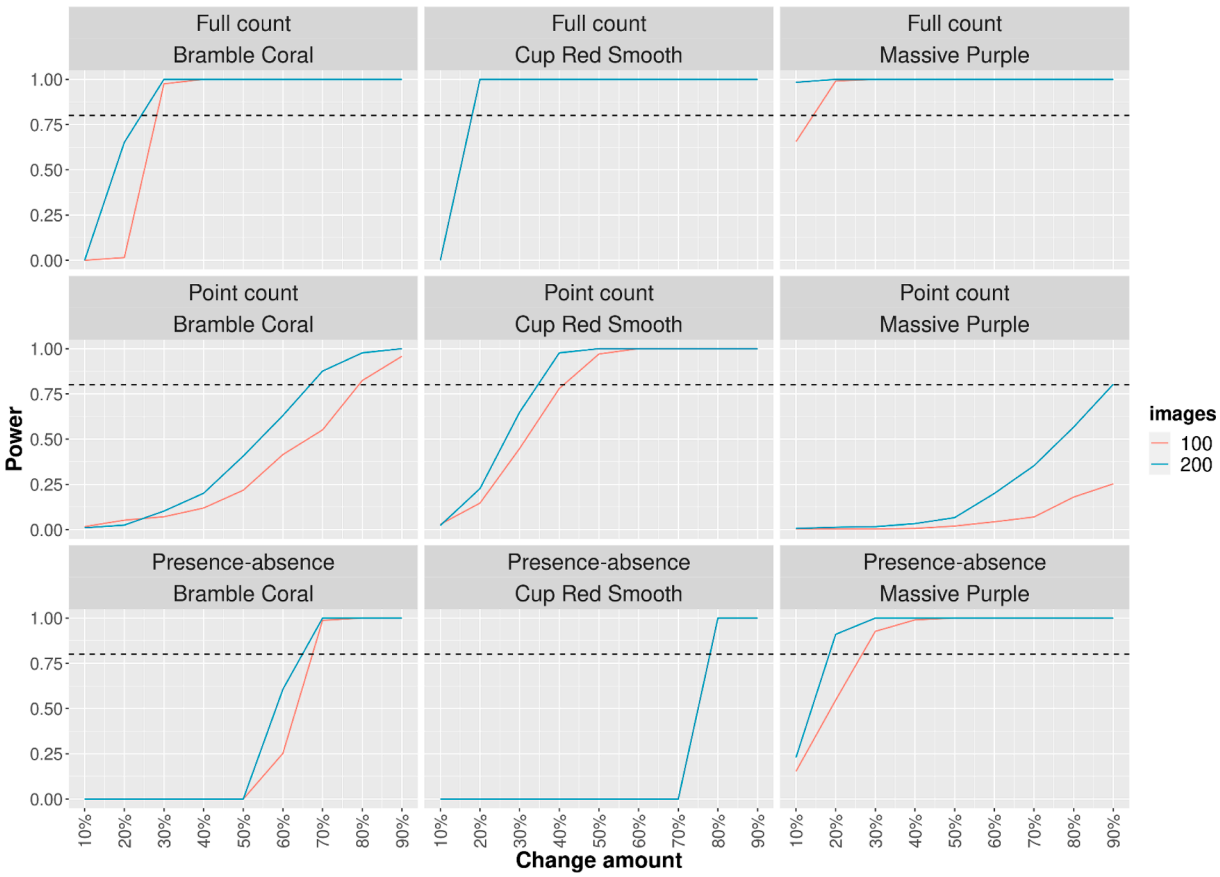

Fig. 3. Power to detect simulated changes (10 – 90% declines in abundance) for the three example morphospecies using three different annotation approaches with 100 and 200 images. Dashed line indicates 80% power. For full count and presence-absence approaches for cup red smooth, trajectories were identical using either 100 or 200 images. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

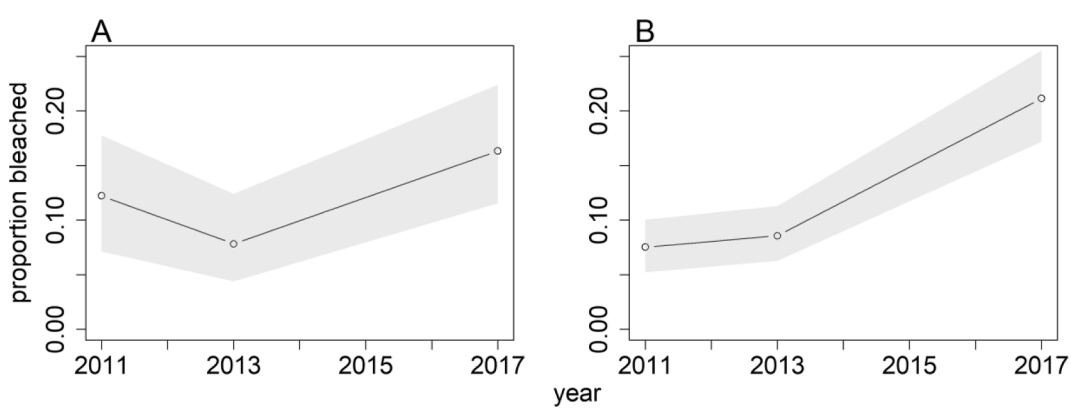

Fig. 4. Trends in the proportion of bleaching in cup red smooth over the survey period when using (A) the point count approach, and (B) the full count approach. Two hundred images were used in both approaches to allow comparison with the same sampling effort.

3.5. Size structure data

Visual inspection of the size structure data for cup red smooth at the Flinders site shows that there was a lack of recruitment over the survey period and growth in the mean size of individuals, with an increased proportion of larger sponges with each successive survey (Fig. 6). A Kruskal-Wallis test revealed significant differences across years for both length ($p = 0.02$) and area ($p = 0.01$). Pairwise Wilcoxon tests showed there were no significant differences in length and area data between 2011 and 2013 (length, $p = 0.62$; area, $p = 0.5$) or 2013 and 2017 (length, $p = 0.29$; area, $p = 0.33$). However, significant differences in both length and area occurred between 2011 and 2017 (length, $p = 0.02$; area, $p = 0.01$). Skewness decreased in the length data (2011 = 1.8, 2013 = 1.6, 2017 = 1.2) through time, indicating that size distributions became more positively skewed with more large individuals through time.

Visual inspection of massive purple at the Freycinet site on the other hand shows that a recruitment event occurred between the 2014 and 2016 surveys, with a larger proportion of smaller individuals in the 2016 survey (Fig. 6). A Kruskal-Wallis test revealed significant differences across years for both length ($p < 0.01$) and area ($p < 0.01$). Pairwise Wilcoxon tests showed there were no significant differences in length

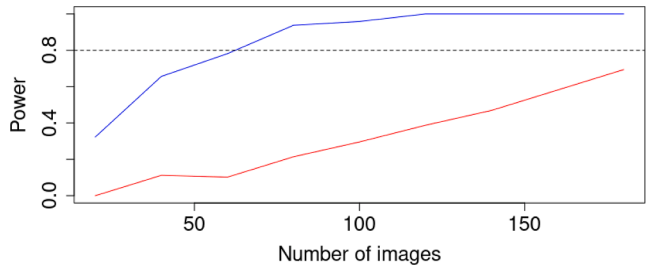

Fig. 5. Power to detect the increase in bleaching of red cup smooth between 2013 and 2017 at Flinders with varying image sampling efforts. The blue line shows power for the full count approach, the red line for the point count approach. The dashed line is at 80% power. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and area data between 2011 and 2014 (length, p = 1.0; area, p = 1.0). However, significant differences in both length and area occurred between 2014 and 2016 (length, p = 0.01; area, p = 0.01), and there were also significant differences in both length and area between 2011 and 2016 (length, p = 0.01; area, p = 0.01). Skewness increased in length data between 2011 and 2014 and then decreased between 2014 and 2016 (2011 = 0.4, 2014 = 1.2, 2016 = 1.1), in line with the observable increase in smaller size classes between 2014 and 2016.

Qualitatively, length and area measurements provide similar trajectories of size structure of the two morphospecies across surveys at each site (Fig. 6). Analysis of area and length data also showed similar trajectories, with both data sources detecting the increasing mean size of red cup smooth at Flinders and the decrease in size of massive purple at Freycinet between 2014 and 2016.

On average, length measurements in ImageJ took approximately 2 s whereas area measurements took approximately 10 s per individual.

4. Discussion

Here we clearly demonstrate the improved ability of a targeted annotation approach to detect temporal trends, both real and simulated, for three temperate mesophotic morphospecies annotated in marine imagery over the more traditionally used point count approach. We show that this approach can be conducted in a relatively short amount of time for our example morphospecies and provides higher statistical power to detect change. In a monitoring context, the point count approach, which has been adopted from terrestrial plant surveys to marine benthic surveys, is likely to be more appropriate in shallower marine ecosystems where dominant space occupiers such as hard corals in the tropics or kelp in temperate regions have relatively high percent covers. However, for temperate mesophotic ecosystems such as are the focus of the monitoring program used in this study, most individual morphospecies have < 2% cover (e.g. Monk et al., 2018; Perkins et al., 2020). Importantly, many potential indicator morphospecies are smaller discreet individuals or colonies that occupy a relatively small proportion of an image despite often being relatively common across images. For some morphospecies, such as our massive purple morphospecies, PA annotating within images may provide a better means of tracking changes than annotating percent cover with the point count approach. However, the full count approach outperformed the PA approach for all our example morphospecies and would take little extra time to achieve during annotation for morphospecies such as massive purple, which occur as discreet easily identifiable individuals. Our results suggest that in systems that consist of smaller and rarer target organisms, alternative annotation approaches should be considered to best utilize resources for ongoing monitoring of indicator morphospecies/groups and to enable tracking of important changes with high statistical power. In some cases, hybrid approaches may be required to track changes in abundance, such as when benthic biota is dominated by encrusting or colonial organisms that aren't well quantified as counts of individuals. Also, additional metrics that track the health or condition of target organisms such as

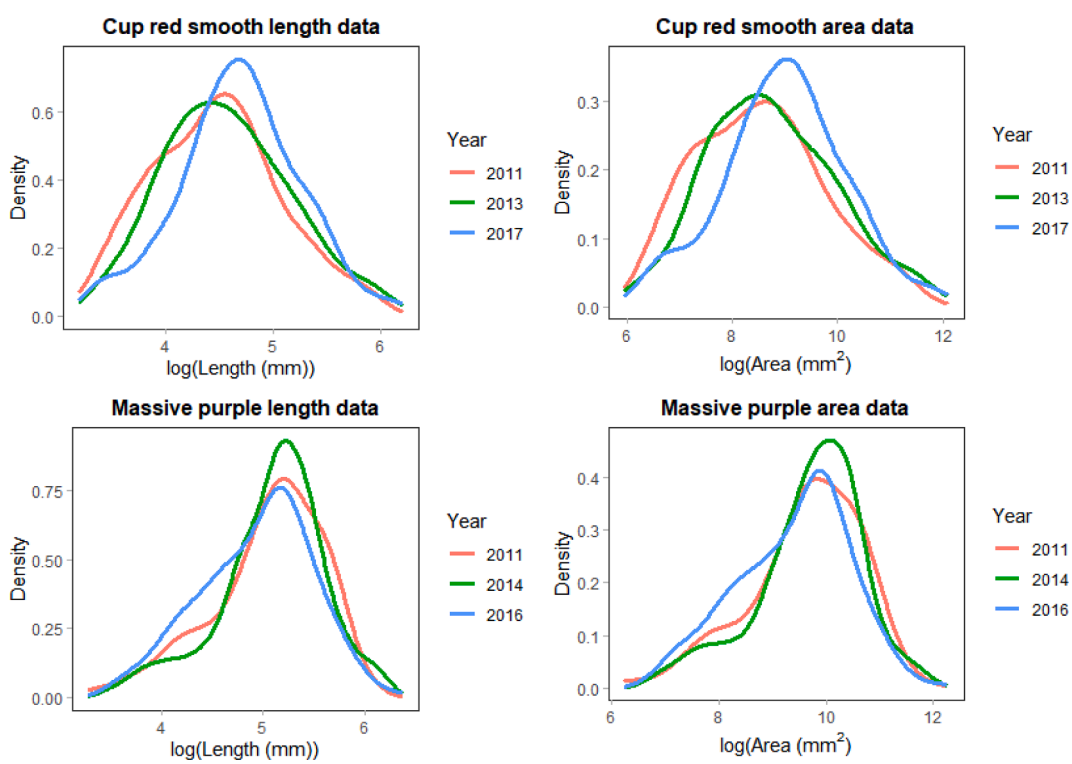

Fig. 6. Size structure data, both length and area, for the cup red smooth and massive purple morphospecies at Flinders and Freycinet sites respectively. Measurements were made for 200 individuals selected from random images for each year. Data has been log-transformed. Note different scales on each axis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

bleaching, and size structure data should be considered as potential indicators and may also be best quantified in combination with a targeted scoring approach. However, these alternative approaches are ideally a refinement after first completing broader biodiversity-based description of biota in new survey areas, such as via comprehensive point-counts of all taxa within imagery, to initially understand the likely relative abundance of candidate indicator morphospecies for ongoing monitoring.

4.1. Tracking changes in metrics of abundance of indicators

Several metrics for tracking changes in abundance are achievable with benthic imagery and we show that a full count approach is likely to provide higher power to detect change over percent cover or PA approaches, particularly for morphospecies that are erect solitary organisms rather than encrusting. Previous work has shown that the point count approach can provide precise estimates of cover for relatively abundant and evenly distributed target organisms such as algae, seagrasses and hard corals (e.g. Drummond and Connell, 2005; Dumas et al., 2009; Trygonis and Sini, 2012), and that annotating a large number of images is generally preferable to annotating a large number of points per image (Drummond and Connell, 2005; Perkins et al., 2016). However, for rarer target organisms, and especially where they show strong spatial aggregation, annotating with the point count approach may require prohibitively high sampling effort to gain sufficient precision in cover estimates to track changes (Dethier et al., 1993; Perkins et al., 2016). In some cases, this can lead to erroneous conclusions about trends due to the high uncertainty around estimates. For example, Dethier et al. (1993) found for both field-based surveys and simulations, target organisms with < 5% cover were often missed, even when a larger number of points were used (50 or 100). In our study, for bramble coral and massive purple morphospecies, which were both low in cover (<0.5% cover over the time series), these morphospecies were not missed, but very different conclusions would be drawn about population trajectories when using the point count approach compared to the full count approach. The decline in abundance of bramble coral between 2011 and 2013 was detected with both approaches; whereas only the full count approach detected the recovery in abundance that occurred in 2017. For the massive purple sponge morphospecies opposing trends were detected, with a low in 2013 when using the point count approach compared to a high with the full count approach. For monitoring programs, detecting important trends when they occur is critical, and power analysis provides a means of testing the ability of different designs and methodological approaches.

Our results show that a full count approach provides higher statistical power to track changes in the abundance of our example morphospecies, can detect smaller levels of change than either the point count approach or PA annotating, and when used in conjunction with quantifying condition (e.g., bleaching) can detect changes with smaller sampling effort. The obvious trade-off is the shift in focus to a smaller subset of target indicators rather than annotating all biota as is typically done with the point-count approach. This suggests that switching to a full count approach may be particularly advantageous when a refined set of indicators are decided upon. However, decisions regarding indicators may be made without full prior knowledge of how different indicators may respond to pressures of interest, and there is thus the potential to select morphospecies or groups that are not the best indicators for tracking changes of management interest. It is therefore prudent to assess change at least periodically across a broader range of morphospecies or groups than just targeted indicators. In our study system, monitoring has now been underway for approximately 12 years, with 3 or more repeated surveys conducted at most sites. Baseline assessments of the diversity present and analysis of temporal changes to date are suggesting a refined set of indicator morphospecies (Perkins et al., 2020). Based on the analysis presented here, approximately 12 indicator morphospecies could be annotated with the full count approach across 100 images in the same amount of time that all biota could be annotated with 25 random points per image. However, this sampling effort would detect changes of > 30% for any of our three example morphospecies with the full count approach; whereas only changes of > 40% could be detected for the most abundant morphospecies (cup red smooth), > 80% for bramble coral, and even a 90% decline could not be detected with high power for massive purple when using the point count approach.

4.2. Tracking changes in other metrics: Bleaching and size structure

Metrics other than abundance, such as measures of ecosystem health (e.g., quantifying changes in condition such as bleaching or the effects of disease) or population dynamics (e.g., measures or recruitment or mortality) provide a more holistic view of ecosystem functioning and may provide more sensitive indicators for specific stressors on ecosystems (Bell et al., 2017). For example, the level of bleaching in corals is often used to track heat stress on coral reefs (e.g., Lachs et al., 2021); whereas size-structure of fish populations can prove to be a more sensitive measure of the effectiveness of marine protected areas than abundance (e.g., White et al., 2021). We found that the annotation approach also affected the ability to track bleaching in the red cup smooth morphospecies, which may be an important indicator of climate change impacts across the region (Perkins et al., 2022). High power to detect changes in bleaching between 2013 and 2017 could be achieved with 60 images using the full count approach, whereas > 200 images were required for high power with the point count approach. Considering that cup red smooth is one of the more abundant morphospecies across sites in the study region (Perkins et al., 2020), it appears that > 200 images are likely to be required with the point count approach to track important changes such as bleaching events or moderate (<50%) changes in abundance for all indicators. The trade-off between collecting data that can provide low precision to track changes in the entire community versus data that provide high certainty of detecting change in a smaller subset of indicators is ultimately a management decision that will be dependent on monitoring program goals. Thorough assessments of the implications of these choices, as presented here, are crucial for informed decision-making.

Size structure data can provide additional insights into ecosystem dynamics and the impacts of stressors that other data sources may be unable to provide (e.g. Bak and Meesters, 1998; Meesters et al., 2001; Lachs et al., 2021). Demographic processes such as growth, mortality and recruitment can be captured when incorporating size information, insights that may be masked or totally hidden when measuring just percent cover or abundance. For example, the full count approach may capture a pulse of new recruits as an increase in abundance, but without knowledge of the mortality of larger individuals or colonies counts alone cannot provide information to discern underlying dynamics. Likewise, percent cover may capture the overall dominance of a morphospecies across a reef, but with percent cover data alone it is impossible to know whether changes are due to growth of existing individuals/colonies, mortality rates or recruitment events. The size-structure data collected here was attainable in a relatively short amount of time and provides valuable extra information on indicators in mesophotic depths where information on life histories is often lacking. Length measurements were achievable in a much shorter time frame and tracked similar changes to area data and so may be preferable where morphospecies are of a reasonably uniform shape such as our example morphospecies. Collecting size structure data for all individuals within an image, such as was completed here, also provides counts of individuals and therefore provides two sources of data simultaneously. Size structure data for the red cup smooth morphospecies show that no recruitment events occurred over a period of increased bleaching following marine heatwave events in the region (Perkins et al., 2022), showing that bleaching may be impacting reproductive output for this morphospecies. For the massive purple morphospecies, a recruitment event was detected

between 2014 and 2016. Ongoing measurement of size information can provide insight into recruitment, growth and mortality dynamics providing information about population dynamics that is missing for most species in mesophotic ecosystems. Given the small amount of additional effort, we therefore recommend that size structure data be collected for selected indicators where possible.

5. Conclusion

Monitoring programs should undergo early and ongoing assessments of their ability to meet program goals (Lindenmayer and Likens, 2010). Such assessments can help refine survey and sampling designs and in directing resources where they are most needed. For benthic monitoring programs capable of collecting large amounts of imagery using technologies such as AUVs and ROVs, the annotation stage typically proves to be the bottleneck in data acquisition. Therefore, optimizing annotation approaches to provide data that can reliably detect trends while minimizing time spent is a high priority. We advocate here, that in systems such as temperate mesophotic shelf reefs which typically consist of highly diverse organisms with low percent cover, a targeted annotation approach should be considered whereby direct counts are conduced within images. Targeted scoring is likely to be particularly advantageous when indicators are discreet individuals/colonies rather than creeping or encrusting organisms. Such an approach can be developed following initial broadly-based biodiversity surveys based on point count methods to help inform likely candidate indicator morphospecies for the system in question. Our results demonstrate that a targeted annotation approach allows tracking relatively small changes (<30%) with high statistical power and annotation can be completed relatively quickly when considering a smaller subset (~10) of indicator organisms. Where resources allow, additional metrics such as condition and size structure can also be collected with relatively small amounts of additional effort and provide valuable insights into ecosystem dynamics.

CRediT authorship contribution statement

Nicholas Perkins: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Zelin Zhang:** Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Jacquomo Monk:** Data curation, Writing – review & editing, Visualization. **Neville Barrett:** Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was undertaken through support from Parks Australia for the monitoring of Australian Marine Parks across the region. In particular, we thank Suzie Kropman from National Oceans Office (now Parks Australia) and Cath Samson for their support of the AUV monitoring program. Data was sourced from Australia's Integrated Marine Observing System (IMOS) – IMOS is enabled by the National Collaborative Research Infrastructure Strategy (NCRIS). AUV deployments were also supported by the Marine Biodiversity Hub, a collaborative partnership supported through funding from the National Environmental Science Programme. We thank Prof Stefan Williams and his IMOS AUV facility team, and collaborating research agencies in eastern Australia (NSW Marine Parks Authority and IMAS, University of Tasmania) for supporting the collection and collation of deep reef imagery in the region and making it available for this study. We also thank Justin Hulls and the numerous technical staff that have assisted in the field programs, and post-processing of the acquired imagery to allow the analysis presented here.

References

Bak, R.P.M., Meesters, E.H., 1998. Coral population structure: the hidden information of colony size-frequency distributions. Mar. Ecol. Prog. Ser. 162, 301–306.
Bell, J.J., Biggerstaff, A., Bates, T., Bennett, H., Marlow, J., McGrath, E., Shaffer, M., 2017. Sponge monitoring: Moving beyond diversity and abundance measures. Ecol. Ind. 78, 470–488.
Bridge, T.C.L., Done, T.J., Friedman, A., Beaman, R.J., Williams, S.B., Pizarro, O., Webster, J.M., 2011. Variability in mesophotic coral reef communities along the great barrier reef, Australia. Mar. Ecol. Prog. Ser. 428, 63–75.
Brito-Millán, M., Vermeij, M.J.A., Alcantar, E.A., Sandin, S.A., 2019. Coral reef assessments based on cover alone mask active dynamics of coral communities. Mar. Ecol. Prog. Ser. 630, 55–68.
Butler, S.J., Freckleton, R.P., Renwick, A.R., Norris, K., 2012. An objective, niche-based approach to indicator species selection. Methods Ecol. Evol. 3, 317–326.
Cánovas-Molina, A., Montefalcone, M., Bavestrello, G., Cau, A., Bianchi, C.N., Morri, C., Canese, S., Bo, M., 2016. A new ecological index for the status of mesophotic megabenthic assemblages in the mediterranean based on ROV photography and video footage. Cont. Shelf Res. 121, 13–20.
Chiarucci, A., Wilson, J.B., Anderson, B.J., Dominicis, V., 1999. Cover versus biomass as an estimate of species abundance: does it make a difference to the conclusions? J. Veg. Sci. 10 (1), 35–42.
Dethier, M.N., Graham, E.S., Cohen, S., Tear, L.M., 1993. Visual versus random-point percent cover estimations: 'objective' is not always better. Mar. Ecol. Prog. Ser. 96, 93–100.
Drummond, S.P., Connell, S.D., 2005. Quantifying percentage cover of subtidal organisms on rocky coasts: a comparison of the costs and benefits of standard methods. Mar. Freshw. Res. 56, 865–876.
Dumas, P., Bertaud, A., Peignon, C., Léopold, M., Pelletier, D., 2009. A "quick and clean" photographic method for the description of coral reef habitats. J. Exp. Mar. Biol. Ecol. 368 (2), 161–168.
Enrichetti, F., Bo, M., Morri, C., Montefalcone, M., Toma, M., Bavestrello, G., Tunesi, L., Canese, S., Giusti, M., Salvati, E., Bertolotto, R.M., Bianchi, C.N., 2019. Assessing the environmental status of temperate mesophotic reefs: A new, integrated methodological approach. Ecol. Ind. 102, 218–229.
Karpov, K.A., Bergen, M., Geibel, J.J., 2012. Monitoring fish in California Channel Islands marine protected areas with a remotely operated vehicle: the first five years. Mar. Ecol. Prog. Ser. 453, 159–172.
Komsta, L., Novomestky, F., 2015. moments: Moments, cumulants, skewness, kurtosis and related tests.
Lachs, L., Sommer, B., Cant, J., Hodge, J.M., Malcolm, H.A., Pandolfi, J.M., Beger, M., 2021. Linking population size structure, heat stress and bleaching responses in a subtropical endemic coral. Coral Reefs 40 (3), 777–790.
Larsen, D.P., Kincaid, T.M., Jacobs, S.E., Urquhart, N.S., 2001. Designs for evaluating local and regional scale trends. Bioscience 51, 1069–1078.
Lindenmayer, D.B., Likens, G.E., 2010. The science and application of ecological monitoring. Biol. Conserv. 143 (6), 1317–1328.
Lindgren, F., Rue, H., Lindstrom, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. Royal Stat. Soc. Ser. B-Stat. Methodol. 73, 423–498.
Meesters, E.H., Hilterman, M., Kardinaal, E., Keetman, M., de Vries, M., Bak, R.P.M., 2001. Colony size-frequency distributions of scleractinian coral populations: spatial and interspecific variation. Mar. Ecol. Prog. Ser. 209, 43–54.
Monk, J., Barrett, N.S., Peel, D., Lawrence, E., Hill, N.A., Lucieer, V., Hayes, K.R., 2018. An evaluation of the error and uncertainty in epibenthos cover estimates from AUV images collected with an efficient, spatially-balanced design. PLoS ONE 13, e0203827.
Montilla, L.M., Miyazawa, E., Ascanio, A., Lopez-Hernandez, M., Marino-Briceno, G., Rebolledo-Sanchez, Z., Rivera, A., Mancilla, D.S., Verde, A., Croquer, A., 2020. The use of pseudo-multivariate standard error to improve the sampling design of coral monitoring programs. PeerJ 8, e8429.
Niemi, G.J., McDonald, M.E., 2004. Application of ecological indicators. Annu. Rev. Ecol. Evol. Syst. 35 (1), 89–111.
Noss, R.F., 1990. Indicators for monitoring biodiversity - a heirachical approach. Conserv. Biol. 4, 355–364.
Parravicini, V., Micheli, F., Montefalcone, M., Villa, E., Morri, C., Bianchi, C.N., 2010. Rapid assessment of epibenthic communities: A comparison between two visual sampling techniques. J. Exp. Mar. Biol. Ecol. 395 (1-2), 21–29.
Pawlik, J.R., Armstrong, R.A., Farrington, S., Reed, J., Rivero-Calle, S., Singh, H., Walker, B.K., White, J., 2022. Comparison of recent survey techniques for estimating benthic cover on Caribbean mesophotic reefs. Mar. Ecol. Prog. Ser. 686, 201–211.
Perkins, N.R., Foster, S.D., Hill, N.A., Barrett, N.S., 2016. Image subsampling and point scoring approaches for large-scale marine benthic monitoring programs Estuarine Coastal and Shelf. Science 176, 36–46.
Perkins, N.R., Foster, S.D., Hill, N.A., Marzloff, M.P., Barrett, N.S., 2017. Temporal and spatial variability in the cover of deep reef species: Implications for monitoring. Ecol. Ind. 77, 337–347.

Perkins, N.R., Hosack, G.R., Foster, S.D., Monk, J., Barrett, N.S., 2020. Monitoring the resilience of a no-take marine reserve to a range extending species using benthic imagery. PLoS ONE 15.
Perkins, N.R., Monk, J., Barrett, N., 2021a. Analysis of a time-series of benthic imagery from the South-east Marine Parks Network Institute of Marine and Antarctic Sciences. University of Tasmania, Hobart, Tasmania.
Perkins, N.R., Monk, J., Soler, G., Gallagher, P., Barrett, N.S., 2022. Bleaching in sponges on temperate mesophotic reefs observed following marine heatwave events. Clim. Change Ecol. 3, 100046.
Perkins, N.R., Prall, M., Chakraborty, A., White, J.W., Baskett, M.L., Morgan, S.G., 2021b. Quantifying the statistical power of monitoring programs for marine protected areas. Ecol. Appl. 31 (1), e2215.
Ryan, D.A.J., 2004. Point sampling strategies for estimating coverage from benthic video transects. Environmetrics 15 (3), 193–207.
Schneider, C.A., Rasband, W.S., Eliceiri, K.W., 2012. NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 9 (7), 671–675.
Schoening, T., Osterloff, J., Nattkemper, T.W., 2016. RecoMIA—Recommendations for Marine Image Annotation: Lessons Learned and Future Directions. Front. Mar. Sci. 3.
Trygonis, V., Sini, M., 2012. photoQuad: A dedicated seabed image processing software, and a comparative error analysis of four photoquadrat methods. J. Exp. Mar. Biol. Ecol. 424–425, 99–108.
Van Rein, H.B., Brown, C.J., Quinn, R., Breen, J., 2009. A review of sublittoral monitoring methods in temperate waters: a focus on scale. Underwater Technol. 28 (3), 99–113.
White, J.W., Yamane, M.T., Nickols, K.J., Caselle, J.E., 2021. Analysis of fish population size distributions confirms cessation of fishing in marine protected areas. Conserv. Lett. 14 (2).
Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B., Grimm, V., 2010. The virtual ecologist approach: simulating data and observers. Oikos 119 (4), 622–635.