Spatial benthic habitat model for the Oceanic Shoals CMR

Ben Radford, Marji Puotinen

Introduction

Environment Australia is tasked with managing the networks of Commonwealth Marine Reserves (CMRs) and Key Ecological Features (KEFs). Effective management requires first establishing a baseline of what key biota are located in a given CMR or KEF, and then regularly monitoring their status to ensure all is well. A particular challenge is posed by CMRs and KEFs located in the remote and relatively poorly known N and NW regions, where it is logistically impossible to survey every area in detail. One way to 'fill in the gaps' between field observations is to build spatial predictive habitat models. For the marine environment, such modelling involves collecting and integrating spatial datasets to build realistic representations of both the topography and composition of the seafloor and major biotic groups (Brown et al 2011, Holmes et al 2008). Such models are widely useful to scientists and managers, for example to:

- determine the spatial heterogeneity of the benthic environment and key classes of organisms,
- evaluate the physical and biological controls on individual and joint habitat distributions,
- discover relationships among habitats and various species of interest, and
- investigate how habitats and organisms respond to disturbance from human activities.

The Oceanic Shoals CMR is a classic example of a case where spatial predictive benthic models can be useful. Multiple field campaigns have collected high resolution survey data in seven study areas within or near the Oceanic Shoals, but these studies collectively cover only a small fraction of the total area of the CMR. While this fine-scale data is valuable, managers need a map showing where benthic habitat types exist across the entire Oceanic Shoals CMR. We aimed to address this need by producing a benthic habitat map for the CMR using spatial predictive modelling, along with a guide to its limitations and how it should be used appropriately.

Methods

Benthic spatial predictive habitat models aim to map the spatial distribution of types of bottom-dwelling organisms across an area of interest in as much spatial detail as robustly possible. In producing such models for NESP, we aim to ensure they are:

- Ecologically meaningful on relevant spatial and temporal scales,
- Sufficiently accurate for the intended use, and
- Communicated to stakeholders clearly so that their limits and likely errors are clearly understood.

We build such models following the basic process outlined below (Figure 1).

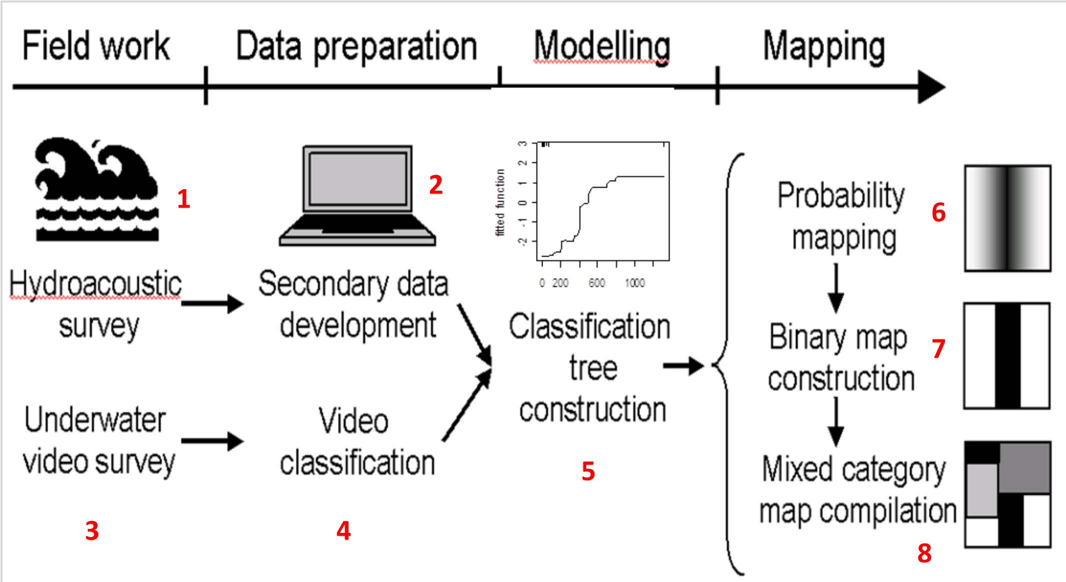

Figure 1. Diagram illustrating the process of producing a spatial benthic prediction model.

Developing predictors

Developing environmental surrogates for the existence and abundance of classes of benthic organisms (step 2) is possible with high resolution bathymetric data (Brown et al 2011). Where such data do not exist in an area of interest, they can be developed from multi-beam sonar data via hydro-acoustic surveys (Holmes et al 2008, Lehmann et al. 2002- step 1 on Figure 1). For the Oceanic Shoals, hydro-acoustic data has been collected at seven locations (Figure 2).

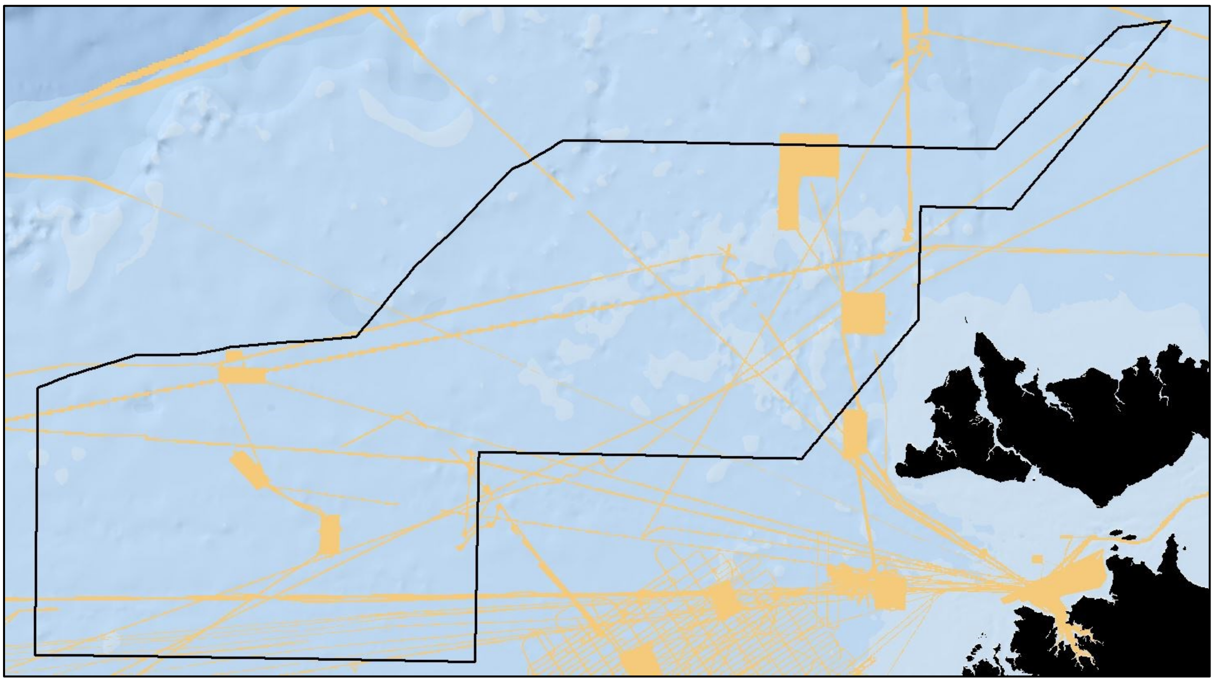

Figure 2. Very high resolution multi-beam sonar coverage of the Oceanic Shoals CMR. The CMR is outlined in black. High resolution multibeam data coverage is shown in orange. Data courtesy of Geoscience Australia.

We used this combined with Geoscience Australia's 250 metre bathymetry coverage of Australian waters to develop the following potential predictors of benthic habitat at 250 m resolution (Figure 1, step 2):

- Depth
- Aspect
- Overall curvature
- Profile curvature
- Plan curvature
- Depth range (5, 10, 25, 50 m windows)
- Standard deviation of depth (5, 10, 25, 50 m windows)
- Mean depth (5, 10, 25, 50 m windows)

Although ocean parameters are also important drivers of benthic community composition and structure (Brown et al 2011), relevant data was not available at spatial and temporal scales sufficient to make it worthwhile to include for the Oceanic Shoals.

Developing training and test data

Building predictive models is not possible without verified field data to document where biota of various types actually exist. The data we use to build and test a given model (Figure 1 – steps 3 & 4) comes from AIMS underwater towed video (http://www.aims.gov.au/docs/research/monitoring/seabed/video-monitoring.html) surveys. It includes both real-time coding of benthic communities from Go-Pro video footage and detailed analysis of downward facing still photos (Figure 3) taken along each of a series of 1.5 km long transects. The former is adept at identifying benthic types like sponge communities that may have a minor profile on a downward facing image while still representing notable biomass. The latter excel at identifying benthic types that do not protrude far into the water column (eg, burrowers, hard corals). For a given field survey, the towed video transects are placed within the study area using a GRTS (Generalized Random Tesselation Stratified) sample design structured to spread transects across a priori classes of habitat complexity while ensuring they are evenly distributed spatially (https://science.nature.nps.gov/im/datamgmt/statistics/r/advanced/grts.cfm).

Figure 3. The AIMS towed video system tow body with mounted video camera and down-ward facing camera for stills (lower), with an example of a still photo taken just above the sea floor (above).

Ecologists at AIMS use the CATAMI classification scheme (http://catami.org/classification) to assign benthic categories to what they observe both in real-time via video and by analysing still photos in the lab.

Building models and mapping habitats

We then explore the statistical relationship between the predictors and field data of the presence / absence of benthic classes across the area of interest using a non-parametric statistical method - classification trees (Figure 1, step 5 – Breiman et al 1984). For the Oceanic Shoals CMR model, we used an innovative version of this called random forest (Breiman 2001, Cutler et al 2007). A random forest model first builds hundreds of classification trees that identify all the unique combinations of variable that could predict the distribution of a given class. Those trees that are not useful in predicting that class cancel each other out. This method outperforms standard classification trees that are defined a priori because it ensures that valid relationships in the data are not missed (Cutler et al 2007).

In our process, we withhold a random sample of one-third of the field data to use for model performance estimates (testing set) and use two-thirds of it to establish how various benthic classes of organisms vary with the potential predictors (training set) to enable building a model. When establishing the testing and training sets, we also test of spatial autocorrelation. Where spatial autocorrelation exists, we retain a representative data point for each cluster of auto-correlated points. After each model is built with the training data, we use the testing data to assess its performance (based on the AUC – 'area under curve' parameter in ROC analysis – Faucet 2006). Models whose AUC values are less than 0.7 are discarded, as is standard.

Typically, we build a separate model for each class of benthos that predicts the likelihood that class exists (Figure 1, step 6) in each pixel across the study area from 0 (no chance it exists) to 1 (100% certainty that it exists). Below is an example of this for sponges (Figure 4).

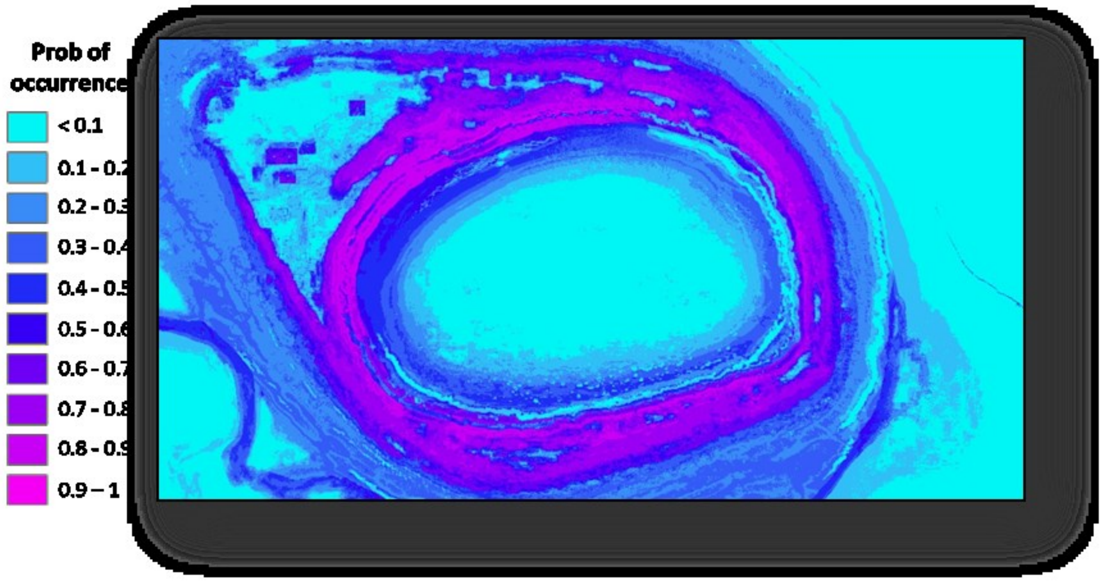

Figure 4. An example of a 'probability map' (Figure 1, step 6) of sponges for part of the Oceanic Shoals CMR at a 2 metre spatial resolution. Sponge is most likely to exist in pixels shaded bright pink, and least likely to exist in pixels shaded turquoise.

We then identify the probability at which errors in misclassifying pixels are balanced between incorrectly assuming the biota doesn't exist (false negative or 'misses') and incorrectly assuming the biota does exist (false positives or false 'hits'). For each biota, we use this probability to simplify the data into two classes: 1 – where the biota might exist and 2 – where the biota is unlikely to exist. This creates a binary map for each class of biota (Figure 1, step 7). We can then combine the binary maps for each class of biota to find out where different classes may co-exist in the same pixel, and where single 'pure' pixels of only one class of biota may exist. This is the 'mixed category map compilation' in step 8 of Figure 1. Below is an example for part of the Oceanic Shoals (Figure 5). Note that no pixels were predicted to contain Hard Coral or Gorgonians or Sponges except in combination with other classes. In contrast, multiple ‘pure pixels’ were predicted for whips, Alcyons, Filterers and Burrowers.

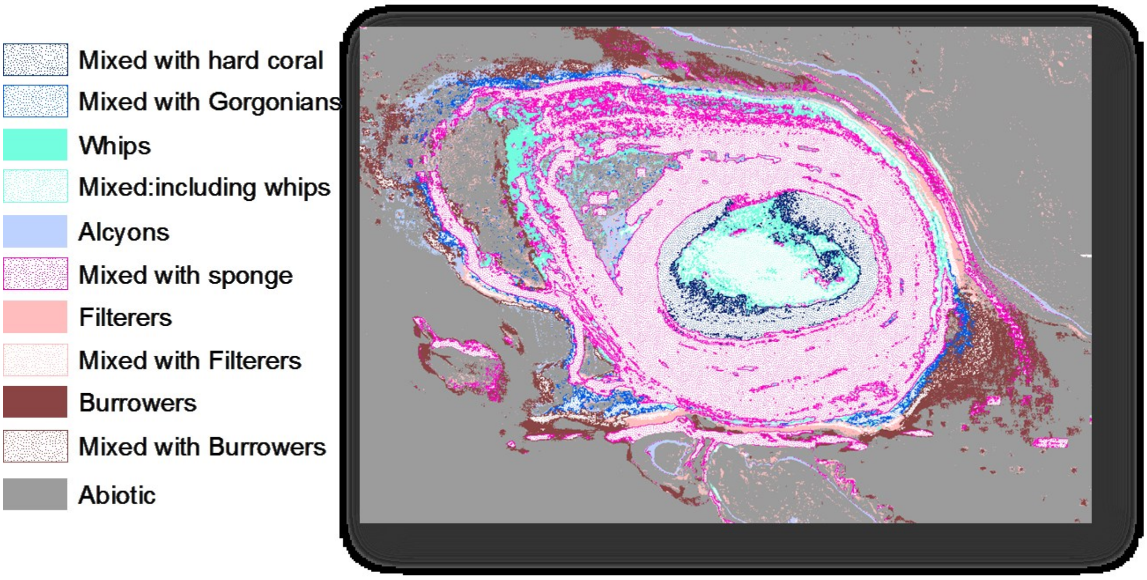

Figure 5. An example of a ‘mixed category map compilation’ (Figure 1, step 7) for part of the Oceanic Shoals CMR at a 2 metre spatial resolution. Note that living organisms may exist in the ‘abiotic’ class, but were not detectable using survey methods. White outlines within the CMR indicate recent zoning designations.

We typically only retain mixed classes that cover at least 5% of the total area of the study area. We will soon trial a method of assigning class names that names a mixed class based on the biota that is most likely to exist in that mix. This can be extracted from the single class probability maps on a pixel by pixel basis.

An alternative approach not shown in Figure 1 is to consider all benthic classes at once and task the model with identifying the most likely class to exist in each pixel. This ‘most likely class’ approach is orders of magnitude faster computationally, and is particularly appropriate when you expect to see clear distinctions between where different benthic classes are likely to occur (that is, you don’t expect them to occur together that often). The two approaches can produce very different results (Figure 6). Where many classes are likely to exist together in a given location, the most likely class model will tend to predict a greater area where no biota is detected (longer grey bar to the left) while the mixed class model will predict a greater area of most biotic classes (longer coloured bars to the right).

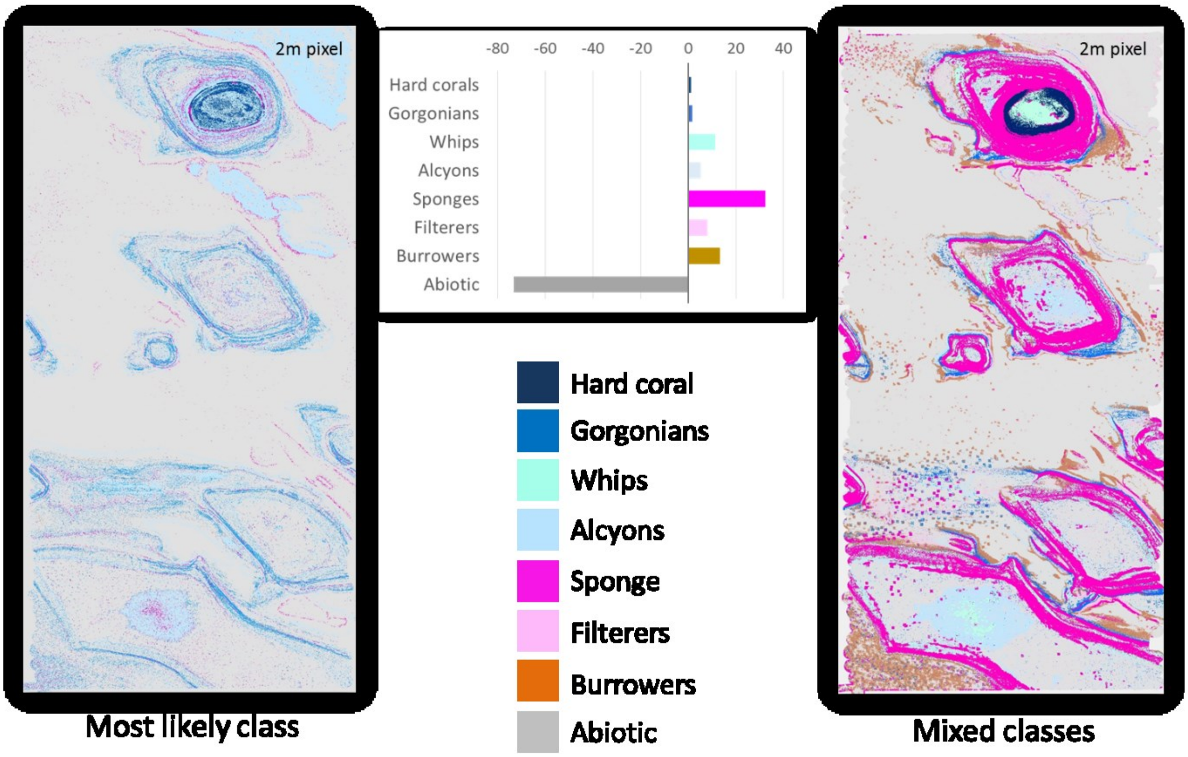

Figure 6. Comparison of a 'most likely class' model (left) versus a 'mixed class' model for part of the Oceanic Shoals CMR at a 2m resolution. The graph in the centre shows the % area difference between the two models for each of 7 classes.

For our initial analysis of the Oceanic Shoals CMR, we aimed to identify the most likely of a set of benthic classes to exist in each 280 m pixel across the study area. This is slightly coarser than the 250 m pixel of the input bathymetry data – some resolution was lost due to the use of a kernel to generate some of the predictors and by projecting the data into flat map coordinates.

Assessing map accuracy

Once we've built a statistical model and used it to predict where a class or classes of biota occur across a study area, it is vital to estimate the accuracy of those predictions (Mumby & Harborne 1999; Holmes at al., 2008; Gray 2001). This is done using the testing data points we randomly withheld when building the model. For each point, we know what actually exists there (the observed value), and we know what the model predicts should exist there (the predicted value). Plotting these by benthic class yields what is called a 'confusion matrix' (Figure 7). In a confusion matrix, the number of data points where the observed class matches the predicted class is shown for each class in the boxes along the black diagonal. All the other boxes in the diagram (that are not on the diagonal) indicate misclassification errors – essentially showing all the ways in which the model failed, broken down by class. For example, for the 'hard coral' row below, values in the boxes other than on the diagonal show the number of test data points where the benthic class was actually hard coral, but the model predicted something else (misses). Most commonly this was either Alcyon or Abiotic (green shaded boxes). For the 'hard coral' column, values in the non-diagonal boxes show the number of test data points where the model predicted hard coral, but the benthic class was actually something else (false hits). Most commonly this was Alcyon (orange shaded box). The relative proportion of false positives and misses given the sample size can be used to estimate overall accuracy of the classification.

		Predicted							
		Alycon	Burrowers	Filter Feeders	Gorgonians	Hard Corals	Abiotic	Sponges	Whips
Observed	Alycon	923	0	1	43	29	163	42	3
	Burrowers	0	18	0	0	0	5	0	0
	Filter Feeders	0	0	33	5	0	19	0	0
	Gorgonians	53	0	3	524	0	152	18	6
	Hard Corals	44	0	0	0	176	12	2	0
	Abiotic	98	6	2	99	6	3712	21	6
	Sponges	59	0	2	15	2	74	214	4
	Whips	12	0	0	18	1	29	1	57
	Total accuracy	78	75	80	74	82	**89**	72	75
	% false hits	23	22	42	31	25	**6**	42	52
	% missed	22	25	20	26	18	**11**	28	25

33 Predicted = Observed

False hits

Misses

Total accuracy balances hits and misses.

Confusion matrix

Figure 7. Example of a confusion matrix for a most likely class model of part of the Oceanic Shoals. The top two-thirds of the diagram show how well the observed (rows) versus predicted (columns) values at each of the testing data points matched for each of eight benthic classes. The black diagonal line indicates the number of testing data points for each class where the predicted class matched what was observed (eg, the model was correct). Each box not on the diagonal line indicates a misclassification error.

Results

We successfully modelled 10 benthic classes across the entire Oceanic Shoals CMR (Figure 7):

1. Alcyons
2. Gorgonians
3. Soft corals
4. Hard corals
5. Halimeda
6. Macroalgae
7. Seagrasses
8. Filterers
9. Burrowers
10. Abiotic

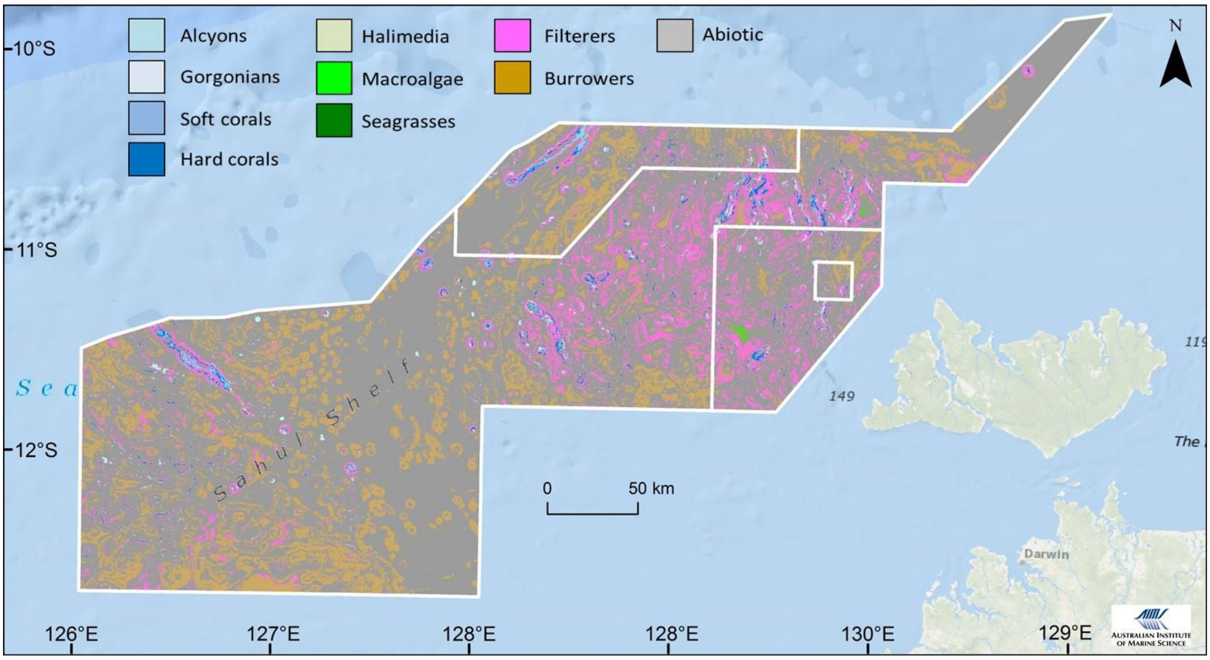

Figure 7. Spatial predictive model of the Oceanic Shoals Commonwealth Marine Reserve for ten classes of biota. Note that living organisms may exist in the 'abiotic' class, but were not detectable using survey methods. White outlines within the CMR indicate recent zoning designations.

Across all classes, the model accuracy was high (82.97% total accuracy, 0.76 of 1 when adjusted for sample sizes to generate a Kappa statistic). Despite this, examining the confusion matrix (Figure 8) shows that total accuracy estimates for four individual classes was poor. These are abiotic, filter feeders, macroalgae and seagrasses. Data points that were actually abiotic were most often mistakenly predicted to be whips. Those that were actually filter feeders were most often mistakenly predicted to be sponges. Those that were actually macroalgae were most often mistakenly predicted to be Halimeda. Those that were actually seagrass were most often mistakenly predicted to be filter feeders. The above should be kept in mind when using these data.

		Predicted											
		Abiotic	Alcyon	Burrowers	Filter Feeders	Gorgonians	Halimeda	Hard Corals	Macro algae	Soft Corals	Seagrass	Sponges	Whips
Observed	Abiotic	1040	1	96	225	0	185	22	35	3	7	433	51
	Alcyon	0	4942	174	0	0	0	0	0	1	0	390	2
	Burrowers	1	119	13042	0	8	1461	48	162	0	0	1708	7
	Filter Feeders	183	0	31	1268	0	37	1	6	0	22	371	101
	Gorgonians	0	0	42	0	383	410	4	36	0	0	314	0
	Halimeda	203	0	747	1	73	42194	66	633	47	0	1468	0
	Hard Corals	8	0	61	0	0	21	1419	0	0	0	184	0
	Macroalgae	9	0	145	0	75	1081	22	2341	0	0	241	0
	Soft Corals	0	0	0	0	0	607	0	0	169	0	24	0
	Seagrass	66	0	3	132	0	63	0	0	0	54	71	15
	Sponges	188	182	1522	357	12	2906	252	302	5	29	27085	19
	Whips	241	22	161	206	0	0	0	0	0	0	212	506
	Total accuracy	54 X	94	81	58 X	70	86	77	67 X	75	48 X	83	72
	% false hits	46	6	19	42	30	14	23	33	25	52	17	28
	% misses	50	10	21	37	68	7	16	40	40	87	18	62

Figure 8. Confusion matrix of the 'most likely class' model of benthic classes across the entire Oceanic Shoals CMR. Note that living organisms may exist in the 'abiotic' class, but were not detectable using survey methods. Red x's denote classes with unacceptable classification accuracy (less than 75%).

Although the estimated accuracy of the 'most likely class' model for the Oceanic Shoals CMR was high, it is important to realise that the training and testing observed data points were not evenly distributed across the study area (Figure 9).

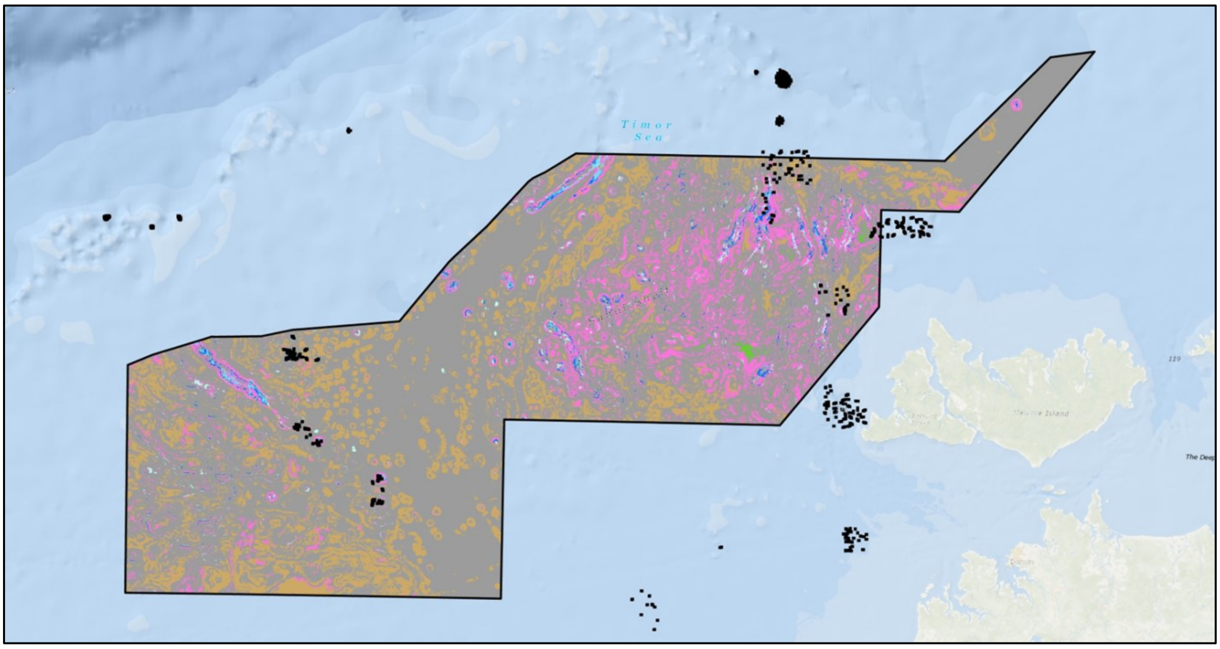

Figure 9. Field data for model building and testing (black dots) beyond and within the Oceanic Shoals CMR.

This means that it is possible that model quality may be lower in areas far from testing and training data points if the relationship between the benthic classes and the predictor variables is not uniform across the CMR. The extent to which this is the case can only be determined by collecting additional field data.

Also important is to consider the spatial scale at which we were able to model the Oceanic Shoals CMR. Due to vast size of the CMR, fine scale bathymetry was too sparse to build a high resolution bathymetric model of the study area. The most detailed dataset covering the entire study area was at a spatial resolution of 250 m. Comparing this for selected areas where fine scale existed (and for which high resolution models were built) illustrates the implications of using the coarser scale bathymetry data (Figure 10).

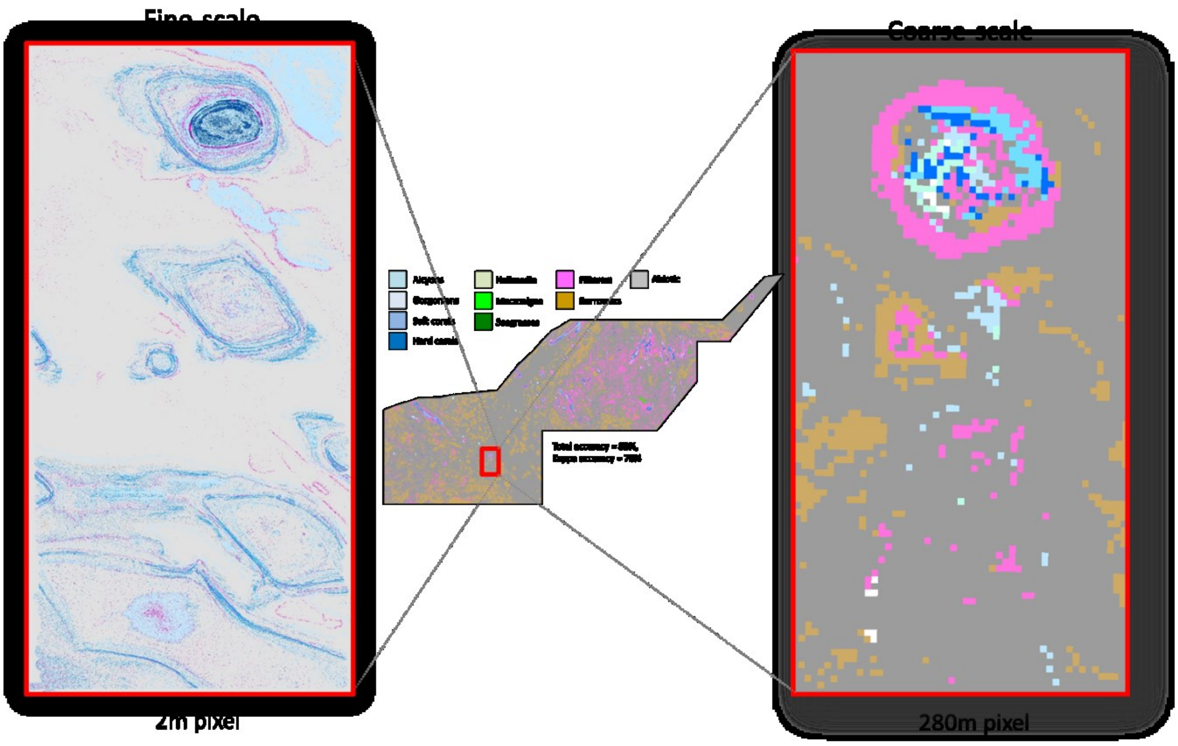

Figure 10. Comparison of fine scale versus coarse scale habitat 'most likely class' model results for a small section of the Oceanic Shoals CMR.

Most notably, the coarse-scale data not only predicts a different relative proportion of the class types, but misses entire features evident in the fine-scale data. The coarse-scale model is still useful, but the implications of using it need to be kept in mind. In particular, if designing a monitoring program based on it, you'll need to make an array of observations within a given 280 by 280 metre pixel classified as 'sponge', for example, to ensure that at least one of those observations contains sponge.

Recommendations

- This coarse-scale habitat map of the entire Oceanic Shoals should be used to target future field surveys in areas of particular interest where validation data is currently missing to collect additional field data. This will enable the development of fine scale habitat models of higher quality.

- Mixed class models should be developed for this region as many of the benthic classes are likely to co-occur. The current 'most likely class' model may underestimate the spatial prevalence of some benthic classes that may exist in mixed assemblages.

- Decisions about poorly modelled habitat types (abiotic, filter feeders, macroalgae and seagrasses) should be made with care, and should consider how the model typically misclassified these types, as shown in the confusion matrix.

- Single class probability models of benthic classes of particular interest (hard coral?) may be of interest to stakeholders for particular applications and can be developed.

- A more detailed analysis of the ecological processes driving the spatial distribution of different habitat types would help to understand the risks posed by various stressors, and aid in the development of appropriate monitoring strategies.

References

Andrefouet, S. et al (2003), 'Multi-site evaluation of IKONOS data for classification of tropical coral reef environments', Remote Sensing of Environment, 88, 128-143;

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brown, C. J., Smith, S. J., Lawton, P., & Anderson, J. T. (2011). Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. *Estuarine, Coastal and Shelf Science*, *92*(3), 502-520.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27, 861-374.

Gray, J. S. (1997). "Marine biodiversity: patterns, threats and conservation needs." Biodiversity and Conservation **6**: 153-175.

Gray, J. S. (2001). "Antarctic marine benthic biodiversity in a world-wide latitudinal context." Polar Biology **24**(9): 1432-2056.

Holmes, K. W., et al. (2008). "Modelling distribution of marine benthos from hydroacoustics and underwater video." Continental Shelf Research **28**(14): 1800-1810.

Lehmann, A., et al. (2002). "Regression models for spatial prediction: their role for biodiversity and conservation." Biodiversity and Conservation **11**: 2085-2092.

Mumby, P. J., et al. (1997). "Coral reef habitat mapping: how much detail can remote sensing provide?" Marine Biology **130**: 193-202.

Mumby, P. J., et al. (1999). "The cost-effectiveness of remote sensing for tropical coastal resources assessment and management." Journal of Environmental Management **3**: 157-166.

Mumby, P. J. and A. R. Harborne (1999). "Development of a systematic classification scheme of marine habitats to facilitate regional management and mapping of Caribbean coral reefs." Biological Conservation **88**(2): 155-163.