A Literature Review in Machine Learning in Benthic Habitat Mapping
Research Methods - INFO5993 Assignment 2

Justin Ting, 430203826

April 2016

1 Introduction

Earth's oceans cover 70% of its surface, but only less than 10% of the Earth's oceans have been explored to date. (NOAA) There have been increasing efforts over the past few decades to map out these unexplored areas to monitor marine ecosystems to be able to track the state of them over time for management, preseveration, etc. purposes. The process used is called benthic habitat mapping, which is the predicting of what exists at the bottom of a body of water. Most recent studies looking to create benthic habitat maps share some basic key steps - acoustic data is used to estimate properties about the surface of the water, which are then mapped to, using machine learning algorithms, *in situ* data such as still images, videos, or samples of the area in question. It is the relationship which is inferred between the different data sets inferred using machine learning techniques that varies between studies.

2 Overview

The process of benthic habitat mapping involves three key steps which the large majority of all studies in the area go through.[1]. In this section, we will give a brief overview of each of these steps, along with common procedures used in them across studies in this area.

1. **Habitat Characterisation** - extracting properties of the environment such as rugosity (roughness), aspect (direction of slope), depth
2. **Habitat Classification** - grouping the raw information about the environment into categories, such as sand, granite, etc.
3. **Habitat Mapping** - using classifications with the larger scale bathymetry data to extrapolate habitat maps

[1] http://www.ozcoasts.gov.au/geom_geol/toolkit/mapoverview.jsp

2.1 Habitat Characterisation

If we were able to collect high resolution data for the entire ocean's benthos - the job of creating benthic habitats for any given area would be (relatively) trivial. As this is prohibitively expensive, we instead collect large amounts of low resolution data, and small samples of high resolution data (between which we model a relationship). This subsection provides a brief summary of data collected and methods used to do so.

2.1.0.1 Remote-sensing data Due to the cost of sea expeditions, it is economically infeasible to have marine vehicles (autonomous or otherwise) explore the entire ocean floor to confirm the ecological properties of all of Earth's benthos. However, we do need to collect sufficiently detailed data of large areas at a time, partiulcarly those of which we are mapping, and for this, remote-sensing data is used. These usually come in the form of acoustic backscatter data, which involes the firing of sound waves towards the benthos, whereby their frequency and strength upon returning is used to deduce the depth at which a particular material was, as well the density of said material (from which a guess at the actual substance can be made - e.g. sand, mud, etc.).

Multibeam echosounders (MBES) are becoming a more frequently used method of collecting acoustic backscatter data (Calvert, Strong, McGonigle, and Quinn, 2015) despite older methods involving single beam echo sounders (SBES) being cheaper and easier to segment. This stems from the fact that the reduced cost comes at the expense of (potentially) accuracy, as well as lower resolution data. This is due to SBES' beam angle, i.e. the angle formed by the 2D flattening of the 'cone' shape of the emitted beams, ranging from 15-25°, whereas MBES' is 0.5-3°, depending on the particular system (Brown, Smith, and Lawton, 2011). The difference in angle means that data returned via SBES devices are more 'coarse', reprsenting less accuracy and granularity, whereas that of MBES is more detailed and can present more information. However, there is overhead associated with use of MBES, in that the considerably decreased angles means much more 'overalpping' data which adds complexity to the segmentation process.

2.1.0.2 Truthing Data The most common methods to be able to obtain a sufficiently large truthing data set (but still trivially small compared to the area covered by remote-sensing data) are videos or images - though the former still requires post-processing to extract the needed images. The advantage that can be provided here, however, is the redundancy in data points (Rattray, Ierodiaconou, J. Monk, and Kennedy, 2014) - but there is extra cost in time required to convert videos into the needed images (pre-proessing before feeding into algorithms for habitat mapping), which is in itself worth of research within the field. (Lucieera, Hilla, Barretta, and Nichol, 2013)

2.1.0.3 Other data Other data which is less common, but also used to map habitats, is patterns in the water movement (such as tidal currents, wave action) (Brown, Smith, and Lawton, 2011) in the column of water above the area of benthos being mapped - a feature which has proved to provide useful input in arriving at an accurate benthic habitat map (in addition to sediment analysis). (Snelgrove, 1994) Other sources such as UNESCO have also verified the importance and significance of using water column correction techniques to obtain more accurate habitat maps, particularly when correlating

images with segrass standing crop. [2]

2.2 Habitat Classification

Almost all studies use *in situ* 'truthing' data to complement the acoustic data to be able to build a model between the acoustic data and truthing data (creation of these models are explained in following sections). However, we need to know the labels of this data considering that the final goal is to create a habitat map, where any one habitual zone is given its prospetive label - to do this, we also need to label the clusters of truthing data. These categories may be, for example, 'bedrock covered by discontinuous seagrass cover', 'Maerl interspersed with sand and gravel', 'superficially coarse sand to fine gravel covered by dense patches of seagrass', etc. (Micallef, Bas, Huvenne, Blondel, Huhnerbach, and Deidun, 2012). The two overarching ways to perform this classification are in the form of supervised and unsupervised algorithms.

Studies have used both supervised and unsupervised methods in clustering the initial data for the training step, though in many cases. Often, there may be large amounts of visual data, beyond that which any human or even team can reasonably, manually cluster - and as such, unsupervised algorithms are first used to create these clusters, after which an expert may be brought in to verify/simplify (or otherwise) the resulting clusters. (Steinberg, Friedman, Pizarro, Williams, and S.B., 2011)

2.3 Map Creation

The final step is map creation, which many papers related to benthic habitat mapping focus on - and also where the most variation occurs in terms of the method used. The various approaches used can be categorised into two broad categories. The first is a top down approach whereby the classification of the habitat characterisation data is validated (or otherwise) with the truthing data, and the second is a bottom up approach where the characterisation data is similarly clustered into classes, but not to directly represent a particular habitat - instead, the aim is to find a relationship between the acoustic data clusters and the truthing data clusters which we can model. Using this model, we can then extrapolate the acoustic data which doesn't have corresponding truthing data to create the habitat map. (Ahsan, Williams, and Pizarro, 2011) We will explore this aspect more when looking at how the mapping process has evolved over time and the improvements that it has brought about.

2.4 Non-Machine Learning Approaches

While the majority of modern papers in benthic habitat mapping employ machine learning techniques for map creation, this doesn't exclude those that do not from providing useful information and insight. An important study was undertaken in 2001 that employs relatively basic statistical analysis, employing different forms of variance as its main tool of analysis. (Kostylev, Todd, Fader, Courtney, Cmaeron, and Pickrill, 2001), at the time (and in fact, even now) integrated more sources of data together than most other studies undertaken - multibeam bathymetric data, geoscientific

[2] http://www.unesco.org/csi/pub/source/rs10.htm

data, seafloor photographs, habitat complexity, and relative current strength. Rather than drawing broad conclusions about the effectiveness of a collection of tools in creating habiatat maps, deeper analysis is done on subsets of the data to attempt to clarify some of the complexities and intrinsic properties of benthic habitats and ecosystems themselves. Although little is done to address and verify accuracy of the actual results/map in this paper, it provides value through the analysis of variance and covariance performed on and between different benthic/marine properties, establishing relationships typically taken for granted or ignored. For example, it is established that while sediment type contirubted heavily to a higher taxonomic group count, there was little relationship between sediment type and depth. However, this only indicates that there is no linear relationship between the two, and doesn't necessarily preclude a non-linear relationship between them, perhaps with the inclusion of other parameters as well. In particular, Kostylev establishes that gravel subrates are more abundant with varying taxonomic groups than their sand counterparts.

TODO deepen exploration of non-ML habitat mapping here, or something else to fill in the page

Such findings provide useful insights for future studies that will allow a better assessment of data, or to be able to apply initial assumptions in obtaining better results. However, constnatly seeking a deeper understanding through a proportionally increasing amount of sampling errs towards 'exploring' the entire Earth's oceans manually. To obtain economically feasible yet reliable predictions from the limited data that we have, we need to employ machine learning techniques to fully utilise the information that we gather.

3 Machine Learning in Benthic Habitat Mapping

As benthic habitat mapping covers a diverse range of discplines, namely "marine biology, ecology, geology, hydrography, oceanography and geophysics" (Brown et al., 2011), in addition to statistics and machine learning, it is logical that it would take considerable effort and vast resources to give each relevant discpline an equal, and large amount of attention within any single study. Thus, different papers can rely on collective findings of others to launch their own research and look further into particular lines of inquiry, start new ones altogether, or evaluate effectiveness of methods used in the field/etc. Within benthic habitat mapping, one prevalent line of inquiry is how to create more accurate, higher quality maps by employing machine learning techniques. For the remainder of this review, we will be revisiting common machine learning techniques, their application in the various stages of benthic habitat mapping along with the benefits they provide, and where such methods can be used to better effect and improved as well.

3.1 Deterministic Machine Learning Algorithms

In this subsection, we will review some machine algorithms that can be used in benthic habitat mapping processes - whether that be in the initial clustering stages of (ideally) independently gathered datasets such as acoustic backscatter data and collections of high resolution images.

3.1.0.1 Multinomial Logistic Regression Multiple Logistical Regression is one of the more basic machine learning algorithms that can (but should not necessarily) be used to predict habitat classes, and falls under the 'supervised learning' category as we have the 'output' for the feature vector in the intial data. Regression, broadly, involves the estimation of relationships between variables, and logistic regression involves the prediction of likelihood of class membership given a number of variables (that are assumed to have low collinearity). This only applies to domains with two classes, however - to use this technique for classification where we have an unbounded (though usually still relatively low) number of classes, we need to use multinomial logistic regression, which is able to account for more than two distinct, unordered (i.e., sand vs. mud has no relative ordering) classes, where class membership is predicted using maximum likelihood estimation (MLE), similarly to logistic regression.

However, the difference is that whereas logistic regression only requiring a single logit function as its nominal variable is dichotomous, multinomial logistic regression requires comparison between $k - 1$ (where k is the number of possible dependent variables) logit functions. **TODO explain properly...**

Even though Caruana and Niculescu-Mizil (2006) show that logistic regression methods achieve on average worse results than most other approaches available, it recognises that in certain cases the models that perform most poorly on average still display exceptional performance, and as such, this method is still worth exploration and experimentation.

TODO Belanger et al. (2012) uses MLR

3.1.0.2 Random Forests In contrast to logistic regression, random forests were shown in Caruana and Niculescu-Mizil (2006) to be state of the art, only just falling short of boosted decision trees

after callibration. Random forests are an ensemble method, meaning that it uses a collection of estimators, before aggregating their results to obtain some sort of average. The aim of this is to minimise the variance and hence error that any single one of these estimators would otherwise result in.

From the initial dataset, some number B is chosen which represents the *number* of trees to build (as a part of our random forest), after which, B random, unique subsamples of the full dataset are taken. Within each decision tree in our random forest, some constant number m of features is taken at each node of the tree, such that the split at each node only takes into account the m randomly chosen features. Each of the decision trees in our forest will hence have a 'result' (that may be a class or some continuous value). Typically, the final decision of the random forest will be made by a vote count for classification, and an average of each decision tree's result in regression problems.

TODO As random forests are not a very complex method that provides very good results on average, we can see that it is used in quite a few studies (Lucieera et al. (2013), Seiler et al. (2012), Hasan et al. (2014)), where the random forest classifier provided the best results over other methods relating to at least a significant subset of the explored data. However, Lucieera et al. (2013) found that while random forest classifiers were most able to classify substratum and rugosity, K-nearest neighbour classifiers most accurately classified sponge structure classes.

3.1.0.3 Multi-class (fuzzy?) Support Vector Machines

Although support vector machines fell outside the top three overall supervised learning algorithms in terms of performance (accuracy), their non-parametricity may be of benefit given that our knowledge of the complex relationships between elements of benthic habitats are limited. Moreover, despite SVMs being rarely used anywhere in the field, they are "acknowledged to be very competitive descriminative classifiers in machine learning literature" (Ahsan et al., 2011).

However, SVMs in their base form only support classification into two classes

3.1.0.4 k-means Clustering

3.2 Machine Learning in Initial Clustering/Classification

3.2.0.1 Supervised methods TODO

- using established 'classification' schemes from third party bodies such as EUNIS or the Australian Government's 'Interim Marine and Coastal Regionalisation for Australia'.[3]
- maximum likelihood estimation (Micallef, Bas, Huvenne, Blondel, Huhnerbach, and Deidun, 2012) - **TODO cite the following** - bad, requires data to fit with prescribed probabilistic distribution. not data-driven - in many cases, pre-speicified Gaussian distribution is bad fit for data [4]

[3] http://www.environment.gov.au/resource/interim-marine-and-coastal-regionalisation-australia-version-33

[4] http://www.analyticbridge.com/profiles/blogs/the-8-worst-predictive-modeling-techniques

3.2.0.2 Unsupervised methods TODO

- unsupervised methods (k-means clustering (Henriques, Guerra, Mendes, Gaudencio, and Fonseca, 2014)) to classify data. - **TODO find citation for the following** bad - k-means tends to produce circular clusters, doesn't work well with data points that are not a mixture of Gaussian distributions [5]
- henriques14 - custom (?) deterministic method on page 79 - supervised classification, wave model used from (Simoes et al. 2012), multivariate data analysis - similarity profile permutation test, simlarity percentages used to determine a species' contribution to groups, BVstepwise to search for relationships between fauna and environmental variables - used depth, median grain size, % content of different sediment fractions classified according to Udden/Wentworth scale (Henriques, Guerra, Mendes, Gaudencio, and Fonseca, 2014)

3.2.0.3 Using machine learning As a progression from basic statistical analysis to machine learning techniques, we can see that numerous studies use the ArcGIS suite of software tools to perform analysis on data typical in benthic habitat mapping studies such as bathymetry data, while others such as Micallef, Bas, Huvenne, Blondel, Huhnerbach, and Deidun (2012) use it to actually carry out machine learning algorithms such as maximum likelihood classification.

TODO cover a few papers that improve on MBH using ML hasan14 - supervised random forest decision trees using two models - first with bathymetry + backscatter mosaic only, and second with angular response derivatives as well (page 4), with extra layers of decreasing importance gradually added to both, with the accuracy of the models assessed using an error matrix, overall actual accuracy, and Kappa coefficient (Hasan, Ierodiaconou, Laurenson, and Schimel, 2014)

3.2.1 Probabilistic Methods

The classifications being made regarding benthic habitats naturally involve uncertainty, as we are still learning the relationship between differnt characteristics of benthos with the varying communities of fauna and flora that reside there. Whilst guessing the most likely class for a particular domain deterministically has its practical applications, it is arguably more *natural* to represent the uncertainty (Rasmussen and Williams, 2006). As our understanding of marine environments is still quite weak (of the United Nations, 2004), it is debatable whether deterministic results are always appropriate when being used to make high level management decisions relating to marine environments. It may be more valuable to provide a percentage certainty to decision-makers regarding the information they need to know, as the probabilistic representation arguably provides **more** information.

TODO in a simple comparison of logistic regression and naive bayes (neither of which are proposed for use in habitat mapping, but rather, for illustrative purposes), naive bayes performs better for smaller data sets, reaching its (usually higher) asymptoptic error higher, while the logistic regressor will perform better with enough data (Ng and Jordan, 2002)

A recent study used probabilistic methods to develop a mapping between the clustered acoustic data to continuous cluster probabilities, as opposed to discrete cluster labels, thus representing the certainty (or otherwise) of the results obtained. Bender, B., Williams, and Pizarro (2012) used Gaussian Processes to extend the probabilistic least squares classifier to retain the information regarding

[5] http://www.analyticbridge.com/profiles/blogs/the-8-worst-predictive-modeling-techniques

certainty of class membership that exists during the classfication process, rather than discarding it in the traditional method. By evaluating the probabilistic results of PTLSC by comparing its results with the actual cluster probabilities obtained in the classification of the images via an unsupervised variational Dirichlet process model, it was shown that the PTLSC method performed better than a PLSC trained directly on the discrete cluster labels in terms of accuracy, mean squared error, and mean variance as well. This demonstrates that while both PTLSC and PLSC err in their predictions when dealing with the transition different boundaries, by maintaining probabilistic information in the PTLSC, it is able to make slightly better judgements in such cases.

TODO*make some mention about kernel methods generally being perceived as unscalable* However, Gaussian processes involve a matrix inversion process that requires an $O(n^3)$ operation which does not scale well with large datasets, which traditionally use non-parametric methods (**TODO cite this**). To overcome this whilst reaping the benefits of Gaussian processes, Bender, B., Williams, and Pizarro (2012) extracted subsets of the original dataset on which to perform analysis - specifically, a sample size of 500 randomly chosen from three Gaussians, of the initial millions of observations. While this (**TODO - brief explanation of how this is representative**), there is likely information to be gained by being able to use a considerably larger portion of the dataset. To do this, a method would be required to generate sparse covariance matrices through approximations (Bickel and Levina, 2008), or use of functions that guarantee sparseness as a property (Melkumyan and Ramos, 2009).

(**TODO mention use of Gaussian processes in adaptive survey design too if relevant**)

Ahsan et al. (2011) uses Gaussian Mixture Models against Classification Trees "more confident in their predictions of new classes - motivates its use over discriminative classifiers in habitat mapping...particularly beneficial in future dive planning" "GMM performs comparably, or better for all training set sizes on all the datasets" "greater certainty of the GMM on unseen data is explained by fact that GMM is a generative model and hence takes into account the distributino of bathymetric features, enabling it to model the joint distribution of the class and features"

4 Conclusion?

we need to better understand "the complexities of coastal system functioning rather than simplifying and scaling down the system into smaller components" (Diaz, Solan, and Valente, 2004) - **NOTE** doesn't really help our point - need to spin this to make it support our point of including uncertainty given that we *still* don't understand all the complexities of marine habitats

References

Nasir Ahsan, Stefan B. Williams, and Oscar Pizarro. Robust broad-scale benthic habitat mapping when training data is scarce. 2011.

Christina L. Belanger, David Jablonski, Kaustuv Roy, Sarah K. Berke, Andrew Z. Krug, and James W. Valentine. Global environment predictors of benthic marine biogeographic structure. *Proceedings of the National Academy of Sciences*, 109, 2012.

Asher Bender, Stefan B., Williams, and Oscar Pizarro. Classification with probabilistic targets. 2012.

Peter J. Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008.

Craig Brown, Stephen J Smith, and Peter Lawton. 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, 2011.

J. Calvert, J.A. Strong, C. McGonigle, and R. Quinn. An evaluation of supervised and unsupervised classification techniques for marine benthic habitat mapping using multibeam echosounder data. *ICES Journal of Marine Science: Journal du Conseil*, 72:1498–1513, 2015.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International conference on Machine learning*. Association for Computer Machinery, 2006.

Robert J. Diaz, Martin Solan, and Raymond M. Valente. A review of approaches for classifying benthic habitats and evaluating habitat quality. *Journal of Environmental Management*, 73:161–181, 2004.

Rozaimi Che Hasan, Daniel Ierodiaconou, Laurie Laurenson, and Alexandre Schimel. Integrating multibeam backscatter angular response, mosaic and bathymetry data for benthic habitat mapping. *PLoS ONE*, 9, 2014.

Victor Henriques, Miriam Tuaty Guerra, Beatriz Mendes, Maria Jose Gaudencio, and Paulo Fonseca. Benthic habitat mapping in a portuguese marine protected area using eunis: An integrated approach. 2014.

Vladimir E. Kostylev, Brian J. Todd, Gordon B. J. Fader, R. C. Courtney, Gordon D. M. Cmaeron, and Richard A. Pickrill. Benthic habitat mapping on the scotian shelf based on multibeam bathymetry, surficial geology and sea floor photographs. *Marine Ecology Progress Series*, 219: 121–137, 2001.

Vanessa Lucieera, Nicole A. Hilla, Neville S. Barretta, and Scott Nichol. Do marine substrates look and sound the same? supervised classification of multibeam acoustic data using autonomous underwater vehicle images. *Estuarine, Coastal and Shelf Science*, 117:94–106, 2013.

Arman Melkumyan and Fabio Ramos. A sparse covariance function for exact gaussian process inference in large datasets. *International Joint Conference on Artificial Intelligence*, 9, 2009.

Aaron Micallef, Timothy P. Le Bas, Veerle A.I. Huvenne, Philippe Blondel, Veit Huhnerbach, and Alan Deidun. A multi-method approach for benthic habitat mapping of shallow costal areas with high-resolution multibeam data. *Continental Shelf Research*, pages 14–26, 2012.

Andrew Y. Ng and Michael I. Jordan. On discriminative vs.generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841, 2002.

NOAA. Exploration, oceanwater.

Food & Agriculture Organisation of the United Nations. *The State of World Fisheries and Aquaculture*. 2004.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

A. Rattray, D. Ierodiaconou, L. J. B. Laurenson J. Monk, and P. Kennedy. Quantification of spatial and thematic uncertainty in the application of underwater video for benthic habitat mapping. *Marine Geodesy*, 37:315–336, 2014.

Jan Seiler, Ariell Friedman, Daniel Steinberg, Neville Barrett, Alan Williams, and Neil J. Holbrook. Image-based continental shelf habitat mapping using novel automated data extraction techniques. *Continental Shelf Research*, 45:87–97, 2012.

Paul V R Snelgrove. Animal-sediment relationships revisited: Cause versus effect. *Oceanography and marine biology*, 1994.

D. Steinberg, A. Friedman, O. Pizarro, Williams, and S.B. A bayesian nonparametric approach to clustering data from underwater robotic surveys. International Symposium on Robotics Research, 2011.