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[1]. 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.[2]. 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] Oceanservice.noaa.gov. (2016). How much of the ocean have we explored?. [online] Available at: http://oceanservice.noaa.gov/facts/exploration.html

[2] Ozcoasts.gov.au. (2016). Benthic habitat mapping: Mapping Overview. [online] Available at: http://www.ozcoasts.gov.au/geom_geol/toolkit/mapoverview.jsp

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

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. [3]

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 et al., 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. 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)

[3] Unesco.org. (2016). Water column correction techniques. [online] Available at: http://www.unesco.org/csi/pub/source/rs10.htm

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 et al., 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 et al., 2001), at the time (and in fact, even now) integrated more sources of data together than most other studies undertaken - multibeam bathymetric data, geoscientific 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.

Certain organisations, government bodies/etc. will also provide guidelines outlining the classification proess. For example, the European Nature Information System website and the Australian Government's 'Interim Marine and Coastal Regionalisation for Australia'[4] both provide classification schemes for people creating habitat maps or other similar efforts.

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, constantly seeking a deeper understanding through a proportionally increasing amount of sampling creeps 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 and their application in the various stages of benthic habitat mapping along with the benefits they provide.

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, or the actual classification of 'new' (or testing) data in determining their predicted habitat classes.

[4] Unesco.org. (2016). Water column correction techniques. [online] Available at: http://www.unesco.org/csi/pub/source/rs10.htm

3.1.0.1 Multinomial Logistic Regression Multiple Logistical Regression is one of the more basic machine learning algorithms that can 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.

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. In particular, Belanger et al. (2012) used multinomial logistic regression across temperature, salinity, and productivity to correctly predict class membership by a margin of 23-84% more than by pure chance. This is equivalent to an improvement of 1-2x compared to a random guess, which taken at face value would suggest that logistic regression is an undesirable choice of algorithm for this problem domain.

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.

As random forests are a method that is low in complexity but 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, pointing to the need to do a more systematic comparison of different methods in benthic habitat mapping. A further advantage to using random forests as pointed out in (Hasan et al., 2014) is that it can provide insight into which features were more important than others, which can aid future studies to be more successful and efficient by focusing more efforts towards collecting the most influential data. The success met with using random forests make it a good benchmark to compare against for future work that aim to develop methods to create more accurate benthic habitat maps than has been done before.

3.1.0.3 Multi-class Support Vector Machines Although support vector machines fell outside the top three overall supervised learning algorithms in terms of performance (accuracy), their non-parametricity could potentially 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, requiring modification to the original algorithm to support more - an active area of research that has not found any single ’best’ way to perform this algorithm extension yet. To do so, there are two, basic main approaches available, one being a **one vs. all** approach, and the other being a combination of all **one vs. one** approach. We can hence quite clearly see that given C possible classes, the first would require C separate classifiers, whereas the latter would require $\frac{C(C-1)}{2}$ classifiers (Murphy, 2012). Multi-class SVMs were used in Ahsan et al. (2011) for illustrative purposes, using the one vs. one approach as per their use of LibSVM (Chang et al., 2011), outperforming classification trees on certain datasets, and in limited cases (but not overall), Gaussian Mixture Models as well. Again, the mixed results would suggest that under certain conditions (such as size of the dataset and various properties of the data itself, some of which are not known before testing), use of a multi-class SVM could provide a useful benchmark to some extent.

3.1.1 Probabilistic Methods

The classifications being made regarding benthic habitats naturally involve uncertainty, as we are still learning the relationship between diffrent 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. While deterministic methods will create a model that attempts to explicitly account for all variables, probabilistic models deal with joint distributions over all the variables. As we need to better understand ”the complexities of coastal system functioning rather than simplfying and scaling down the system into smaller components” (Diaz et al., 2004), this feature can be esepcially valuable seeing as there is simply not enough ’expert knowledge’ to adequately, explicitly model the relationship across a range of variables.

3.1.1.1 Illustrative Example A simple example of this can be seen when comparing the deterministic approach of a logistic regression classifier, with the probabilistic Naive Bayes classifier. Starting from no data, up until a certain threshold, a Naive Bayes (NB) classifier will actually provide a more accurate classification as it approaches its comparatively higher asymptoptic quicker, after which point, one there is sufficient data, the logistic regressor will provide the better results (Ng and Jordan, 2002). In this simplified example, an analogy can be drawn where the data used up to the threshold when the NB classifier performs better represents a lack of knowledge about the data causing the logistic regressor to underperform, where as the continued addition of data represents more understanding (more data points) of the domain, allowing logistic regression to then outperform the NB classifier.

3.1.1.2 Gaussian Mixture Models Gaussian mixture models (GMMs) are parametric models that ”model the distribution of data as a set of clusters, where each cluster is a multivariate Gaussian” Ahsan et al. (2011). In this particlar paper, GMMs are compared with classification trees, which it is found to perform better than in most cases, but were also predicted classes from unseen data with higher certainty than discriminative methods. This is because of its generative nature that accounts for the distribution of bathymetric features, allowing it to model the joint distribution of the classes as well as features. Moreover, each Gaussian function within the model has its own mean and covariance matrix, which also contributes to its powerful modeling ability. However, the use of GMM may have been hindered by the dimensionality of the data - while only five properties were measured, each was calculated for a varying number of scales for the input vector, meaning the ’features’ were at least some multiple of five. As this exceeds the recommended six dimensions for use with GMM, application to a very large dataset may be beyond reasonable computational ability.[5]. To avoid this, the feature vector may have to be truncated to contain the bathymetric properties for only one particular scale at a time.

3.1.1.3 Using Gaussian Processes 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 of the results obtained. Using Gaussian Processes which do not inherently support classification, Bender et al. (2012) extended the probabilistic least squares classifier to retain the information regarding 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.

3.1.1.4 Gaussian Processes and large datasets However, Gaussian processes involve a matrix inversion process that requires an O(n^3) operation which does not scale well with large datasets. To overcome this whilst reaping the benefits of Gaussian processes, Bender et al. (2012) extracted subsets of the original dataset on which to perform analysis - a small, randomly chosen portion from three Gaussians, of the initial millions of observations. While this has still provided a high accuracy for all methods tested, 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) - something that can be explored in future work. To illustrate how the obstacle can be overcome, the latter paper describes a method whereby, rather than inverting the covariance matrix in its raw form, a threshold is calculated at which point, rather than observing the normal 'tapering' off of covariance values, they are simply set to zero beyond that point. This will result in a significant portion of the covariance matrix being populated with 0s, at which point inversion of the sparse matrix can be performed for which there are known efficient methods. However, there have been more ways of sparse approximation GPs that other studies have explored.

[5] Nickgillian.com. (2016). GMM Classifier NickGillianWiki. [online] Available at: http://www.nickgillian.com/wiki/pmwiki.php/GRT/GMMClassifier

3.1.1.5 Sparse Approximation Gaussian Processes It is of importance that a number of methods of dealing with sparse approximation of GPs are taken into account if the aim is to deal with large GPs in the inversion step. (Quinonero-Candela and Rasmussen, 2005) explores exactly this, immediately discounting the "subset of data" (SoD) method as being non-competitive due to it not being able to represent the original data to a reasonably accurate enough extent, though we have seen that this was the approach taken in (Bender et al., 2012). As all the different methods (bar SoD) have a complexity of $O(nm^2)$ where n is the size of the data, and $n < m$, the authors notably point out that no gross approximations should be made as more competent methods are computationally equivalent, and as such point towards their notes on future work to outperform the existing state of the art. As such, we would wish to explore combining the **Partially Independent Training Conditional approximation** with "the most powerful selection method for the inducing inputs."

4 Conclusion

In the reviewed literature, we can see that advancements in both the technology used to retrieve bathymetric and benthic truthing data, as well as the machine learning/analytic side of habitat mapping, has resulted in higher quality maps being created. The applicabilty of and successful application of deterministic algorithms such as random forests has shown that this area lends well to machine learning techniques. We have seen that one of the gaps present in improving map quality is our understanding of the complexity of benthic habitats. As the extensive surveying of more habitats can get prohibitively expensive, one of the best tested compromises is to employ probabilistic machine learning techniques to create habitat maps that are as accurate as possible. Use of Gaussian processes to do this, which are considered state of the art, face the obstacle of being able to work with large datasets without a filtering step, which potentially removes informative data. To improve on such methods, we need to be able to tweak the data being processed so that we can use GPs on such datasets, for which research exists but simply has not yet been applied to benthic habitat mapping.

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.

Chang, Chih-Chung, and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27, 2011.

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.

Vladimir E. Kostylev, Brian J. Todd, Gordon B. J. Fader, R. C. Courtney, Gordon D. M. Cameron, 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.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 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.

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

Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

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.

Machine Learning in Benthic Habitat Mapping - Outline of Research Approach Research Methods - INFO5993 Assignment 2

Justin Ting, 430203826

April 2016

1 Introduction

This study will be about benthic habitat mapping, which is the ”spatial representation of physically distinct areas of seafloor that are associated with particular groups of plants and animals” (Harris and Baker, 2012). Its significance lies in the fact that there is an increasing need by regulatory bodies such as governments, etc. to better manage bodies of water which are being directly affected by human activity to preserve their state and prevent further damage, etc. Many studies have already collected a range of different data source and tried varying techniques at creating habitat maps using such data. To date, there have not yet been attempts to combine the predictive power of Gaussian proceses (Bender et al., 2012) fully with the increasing amount of bathymetric data that modern technology can collect. The purpose of this study will hence be to capitalise all the data we have access to to create better benthic maps to allow decision making bodies to make more informed decisions.

2 Aims and Research Questions

Our aim is to create a habitat mapping scheme that can provide more accurate habitat maps than current state of the art methods by attempting to solve some of the obstacles that are present in them. For our proposed research, we are specifically looking at exploring the following questions:

- How can we improve the quality of benthic habitat mapping?
- How can we put the existing data we have to better use (i.e. not needing to go on expeditions to obtain new data)?
- How can we use Gaussian Processes (henceforth GP) and the available existing data to improve accuracy of the mapping process?

The former two questions essentially cumulate to form the third, which will be the focus question for our research.

3 Proposed Methodology

3.1 Data Collection

Bathymetric and truthing data from at least Scott Reef and O'Hara Bluffs will be used to verify the performance of the algorithms being tested. Data for both these locations have been used in previous studies(Bender et al. (2012), Ahsan et al. (2011)).

Datasets from different marine contexts are needed as the distribution and variance of properties in different areas can vary considerably, causing some algorithms to work better in certain (physical) conditions. This can be seen in (Ahsan et al., 2011), where there were differences in the accuracy metric depending on the location of the benthic habitat in question - O'Hara, Chevron, and Scott Reef.

3.2 Pre-processing

To address the question, we will begin by exploring possible methods to apply GPs to large datasets. The obstacle faced with the use of GPs is in applying the matrix inversion to the raw covariance matrix which has a worst case complexity of $O(n^3)$ where n is the input size - this is a major bottleneck as it does not scale, preventing use with larger datasets.

One method of overcoming this that has been explored in literature is to 'transform' the full matrix into a sparse one at the inversion step - (Melkumyan and Ramos, 2009) and (Furrer et al., 2006) detail a method that involves a 'cut-off' point within the covariance matrix such that rather than the covariance values tapering off and approaching (but never reaching) zero, after a certain point it is instead actually set to zero. Once this step is done, there are known ways to invert such matrices in much less than $O(n^3)$ time, with a range of libraries in differnet languages implementing such methods. Other ways of doing this will be explored as well as necessary. To supercede the current state of the art in benthic habitat mapping however, it would be prudent to implement the future work suggested (Quinonero-Candela and Rasmussen, 2005) to obtain the best approximation of the large GPs, which involves combining the Partially Independent Training Conditional approximation with "the most powerful selection method for the inducing inputs."

3.3 Measuring Performance of GPs on large dataset

Given that (Bender et al., 2012) is one of the more recent studies employing the use of GPs in benthic habitat mapping, the first metric to compare would be whether including a much larger subset of the original data, if not all of it, improves upon the performance of the work which we are aiming to build on.

Intrinsic accuracy of the method with the stated pre-processing steps will be tested via cross-validation and checked that it exceeds (or not) the probabilistic target least squares classifier (PTLSC) in accuracy.

Performance will also be measured relative to a number of other methods that were found to be highly performant in the *literature review*. These include, in particular, **random forests**, and **boosted decision trees** as well (though the latter method has yet to see much adoption when creating benthic habitat maps).

3.4 Limitations

The two distinct marine locations for which real data is obtainable is likely not representative enough to get a full picture of how well the method proposed will perform in different contexts. During the duration of the study, some effort will be put into obtaining more datasets from other authors of similar studies. Failing that, as a backup/last resort, we will consider generating synthetic data based on information available in past research papers that detail the properties to diversify the data which we use to assess our method.

References

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

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

Reinhard Furrer, Marc G. Genton, and Douglas Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15, 2006.

Peter T. Harris and Elaine K. Baker. *Seafloor Geomorphology as Benthic Habitat*. Elsevier Inc., 2012.

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.

Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6: 1939–1959, 2005.