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

2.1 Habitat Characterisation

Collecting acoustic data over large areas

TODO

- Talk about SSS, SBES (as older technologies), focus on MBES in particular and improvements it brings
- "multibeam echosounders (MBES) are an increasingly common source of acoustic data for benthic habitat mapping" (Calvert et al., 2015)

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 come in the form of acoustic backscatter data, basically meaning that waves are sent out, and the strength/frequency at which they are returned is measured. The technology for collecting backscatter has improved considerably recently,

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

allowing for higher resolution to be collected this way.

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 et al., 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 et al., 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) (et al., 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. (Snelgrove, 1994)

- how is current flow data obtained...?
- flow data is actually claimed to show relationships more accurately than simply looking at acoustic data together with ground truthing data (Kostylev, 2012)
- with some using extra data sets as well, such as the flow of water in the columns above the benthic area in question. (et al., 2011)
- light refraction data (?)

2.2 Habitat Classification

Studies have used both supervised and unsupervised methods in clustering the initial data, 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 et al., 2011)

2.2.0.1 Supervised methods maximum likelihood estimation (Micallef et al., 2012) 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 et al., 2014)

2.2.0.2 Unsupervised methods unsupervised methods (k-means clustering (Henriques et al., 2014)) to classify data. 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 et al., 2014)

2.3 Map Creation

The final step is map creation, which a large portion of 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.

3 Evolution of Map Creation Methods

More recent studies have 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. This particular study uses Gaussian Processes to extend 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. (Bender et al., 2012)

- the classifications being made naturally involve uncertainty, considering that we are still learning the relationship between different characteristics of benthos with the varying communities of fauna and flora that reside there. Whilst guessing the most likely class for a particular domain has its practical applications, is it not arguably more *natural* to represent the uncertainty in any one of these guesses in our actual result?[2]
- benthic habitats aren't a domain that we humans are very well versed in, so to maximise our ability to form predictions, we want to utilise larger datasets where available/possible - but Gaussian processes involve a matrix inversion process that requires an O(n^3) operation, which definitely doesn't scale well with larger datasets, which have traditionally used non-parametric methods
- our aim is to use large scale Gaussian processes to overcome the problem with the matrix inversion bottleneck by transforming our data matrix into a sparse matrix, which requires considerably less time to invert

[2] http://www.gaussianprocess.org/gpml/chapters/RW3.pdf

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.

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.

CJ. Brown et al. 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.

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 Kostylev. Benthic habitat mapping from seabed acoustic surveys: do implicit assumptions hold? 2012.

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.

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.

NOAA. Exploration, oceanwater.

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.

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.