21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

Predicting the spatial distribution of seabed hardness based on multiple categorical data using random forest

Jin Li [a], Maggie Tran [a] and Justy Siwabessy [a]

[a] *Geoscience Australia, GPO Box 378, Canberra, ACT 2601, Australia*

Email: Jin.Li@ga.gov.au

Abstract: Seabed hardness is an important character of seabed substrate as it may influence the nature of attachment of an organism to the seabed. Hence, spatially continuous predictions of seabed hardness are important baseline environmental information for sustainable management of Australia's marine jurisdiction. Seabed hardness is usually inferred from multibeam backscatter data with unknown accuracy and can be inferred from underwater video footage or directly measured at limited locations. It can be predicted based on two-class hardness data derived from video footage and environmental predictors, but no study has been undertaken for predicting multiple-classes of hardness data.

In this study, we classified the seabed hardness into four classes based on underwater video images that were extracted from the underwater video footage. We developed an optimal predictive model to predict the spatial distribution of seabed hardness using random forest (RF) based on the point data of the hardness classes and spatially continuous multibeam bathymetry, backscatter and other derived predictors. A novel model selection measure that is the averaged variable importance (AVI) was used based on predictive accuracy that was acquired from averaging the results of 100 times replication of 10-fold cross validation. Finally, the spatial predictions generated using the most accurate model were visually examined and analyzed in comparison with previously published predictions based on two-class hardness data.

This study confirmed that:

1) seabed hardness of four classes can be predicted into a spatially continuous layer with a high degree of accuracy (i.e., with a correct classification rate of 86.27%);
2) model selection for RF is essential for identifying an optimal predictive model in environmental sciences and AVI selects the most accurate predictive model(s) instead of the most parsimonious ones, and is recommended for future studies;
3) caution should be taken when using the correlation coefficient to select predictors for RF in marine environmental sciences;
4) RF is an effective modelling method with high predictive accuracy for multi-level categorical data and can be applied to 'small p and large n' problems in the environmental sciences;
5) the spatial predictions for four-class hardness data were similar with the predictions based on two hardness classes, with high match rates; and
6) RF and AVI are recommended for generating spatially continuous predictions of categorical variables in future studies.

In summary, this is the first attempt to predict the spatial distribution of seabed hardness of four classes. AVI shows its effectiveness in searching for the most accurate predictive models and is recommended for future studies. This study further confirms the superior performance of RF in marine environmental sciences. RF is an effective modelling method with high predictive accuracy not only for presence/absence data but also for multi-level categorical data. RF and AVI are recommended for generating spatially continuous predictions of categorical variables in future studies.

Keywords: *Machine learning, model selection, cross validation, multibeam backscatter and bathymetry, spatial prediction*

1. INTRODUCTION

Seabed hardness is an important character of seabed substrate as it may influence the nature of attachment of an organism to the seabed. Hard substrates provide environments that generally support sessile suspension feeders, while soft (unconsolidated) substrates generally support discrete motile invertebrates (McArthur et al., 2010). Hence, a spatially continuous measurement of seabed hardness would be a significant aid in predicting the spatial distribution of benthic marine communities and thereby to marine ecosystem management. Despite its importance, seabed hardness data is difficult to acquire. It can be 1) directly measured at point locations, 2) inferred from underwater video footage at discrete locations over small areas (Stein et al., 1992), or 3) inferred from multibeam backscatter data (Kloser et al., 2010). However, there are disadvantages associated with these methods (Li et al., 2013). Therefore, predictive modelling provides an alternative approach to generate spatially continuous data of seabed hardness (Li et al., 2013), where seabed hardness was classified into two classes. Moreover, all mixed classes were classified into hard class (Li et al., 2013), so the relevant information of hardness is missing for seabed with the mixed classes of hardness. However, there is no study on predicting seabed hardness based on four classes data.

Random forest (RF), due to its proven predictive accuracy in data mining and many other disciplines (Cutler et al., 2007; Diaz-Uriarte and de Andres, 2006; Marmion et al., 2009) and in the marine environmental sciences (Li et al., 2011c; Li et al., 2010), was applied to predicting seabed hardness based on two classes (Li et al., 2013). Model selection is essential for identifying an optimal predictive model and various methods have been developed (Saeys et al., 2007). However, model selection was argued to be less important for RF, because: 1) RF selects the most important variable at each node split thus is insensitive to un-important variables (Okun and Priisalu, 2007); 2) it is of high predictive performance even when most predictive variables are noisy (Diaz-Uriarte and de Andres, 2006); and 3) if sample sizes are large (500 to 1000), its accuracy depends only on the number of strong features and not on the number of noisy variables (Biau, 2012). It was found that excluding the correlated variables improved the predictive accuracy (Li et al., 2011a; Li et al., 2011b). In contrast, it was observed that including some correlated variables improved the predictive accuracy (Li, 2013b; Li et al., 2012a; Li et al., 2013). These findings demonstrate that model selection is necessary for selecting an optimal predictive model for RF.

A model selection procedure was previously developed by Li *et al.* (2013) for RF based on the variable importance. In the environmental sciences, predictive variables are often correlated, which may affect the observed variable importance for the predictors when using RF. To deal with this, an R package '*extendedForest*' (Smith et al., 2011) was developed to compensate for the shortcomings in the existing RF package by Liaw and Wiener (2002). These studies provide fundamental tools for this study.

In this study, we aim to select an optimal model to predict seabed hardness based on multi-level categorical hardness data and seabed biophysical variables. To achieve this, we also tested the effects of various predictor sets on the predictive accuracy of RF models. Finally, the most accurate model was used to generate a spatially continuous layer of seabed hardness and the predictions were visually examined and compared with the previously published predictions of hardness in two classes (Li et al., 2013).

2. METHODS

2.1. Study region and seabed hardness classification

The study region is located in the eastern Joseph Bonaparte Gulf, northern Australian marine margin, with four areas (A - D) used in this study (Figure 1), which were surveyed in 2009 (Heap et al., 2010) and 2010 (Anderson et al., 2011). Multibeam bathymetry and backscatter data and co-located underwater video data were acquired. The video footage was analyzed based on a 15-second window for each transect to classify seabed substrates and biological presence was used to assist the classification as detailed previously (Li et al., 2013). The substratum composition was visually estimated to 5% precision (Mortensen and Buhl-Mortensen, 2004) in terms of rock, boulders, cobble, rubble, gravel, sand and mud as defined by Wentworth (1922). Anything larger than gravel (i.e. rubble, cobbles, boulders and bedrock) was classified as 'hard' material, while mud, sand and gravel were classified as 'soft' material according to Stein *et al.* (1992). In total, 140 samples of seabed hardness were considered in this study. On the basis of Stein *et al.* (1992), we classified the seabed substrate into four categories: hard, hard-soft, soft-hard and soft. If a substratum consisted of >70% hard material, it was classed as 'hard'; if it consisted of ≤70% and >50 % hard material, it was classed as 'hard-soft'; if it consisted of <50 % and ≥30% hard material, it was classed as 'soft-hard'; and if it consisted of <30 % hard material, it was classed as 'soft'. Substratum consisting of 50% 'hard' and 50% 'soft' materials was not present in this study. Of the 140 samples, 9 samples were recorded as hard, 11 hard-soft, 6 soft-hard and 114 soft. The resultant datasets were used to predict seabed hardness, with hardness classes presented in Fig. 1.

2.2. Predictive variables

Following a preliminary analysis based on data availability and the relationships with seabed hardness as discussed above and in previous studies (Kloser et al., 2010; Siwabessy et al., 2013), 41 predictive variables were initially collected at the locations of video transects. Since there were strong correlations among some predictive variables based on Spearman's rank correlation (ρ), we removed 21 backscatter (bs) variables that were perfectly correlated with other variables or with a ρ=0.99. Thus 20 variables were retained and used in this study (Table 1). The bs25 should have been removed according to the above selection criteria, but was retained because it was used in a previous study (Li et al., 2013). The Pearson's correlation (r) was also derived for the remaining bs variables. Acquisition of these variables were detailed in previous studies (Li et al., 2013). All these variables were available at each grid cell to a 10 m resolution.

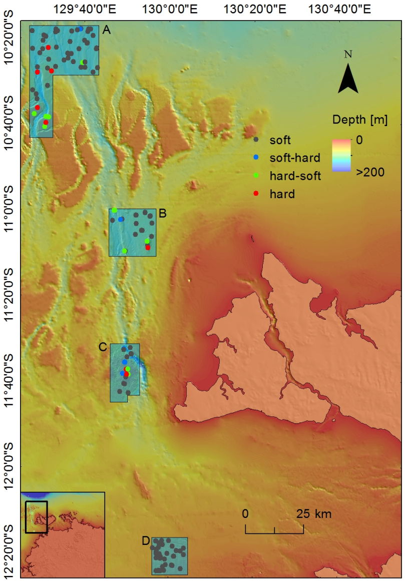

Figure 1. Location of the four study areas (A, B, C, and D) in the study region and seabed hardness classes (hard, hard-soft, soft-hard and soft) overlaid on bathymetry at video transect stations.

2.3. Application of RF and model selection

The R function, *randomForest* by Liaw and Wiener (2002), was used to develop a model to predict the spatial distribution of seabed hardness. The default values of *mtry*, *ntree* and *nodesize* were used for these parameters because they were often proven to be good options (Liaw and Wiener, 2002) as was also evidenced in previous studies (Li et al., 2012b; Li et al., 2013).

Table 1. Predictive variables and their corresponding number.

No.	Predictive variable	No.	Predictive variable
1	easting	11	topographic position index (tpi)
2	northing	12	backscatter 13 o (bs13)*
3	probability of hard substrate	13	bs21
4	bathymetry (bathy)	14	bs25
5	local Moran I of bathy	15	bs27
6	planar curvature (planar.curv)	16	bs32
7	profile curvature (profile.curv)	17	bs35
8	topographic relief (relief)	18	homogeneity of bs (homogeneity)
9	seabed slope (slope)	19	variance of bs (variance)
10	surface area (surface)	20	local Moran I of bs (bs.moran)

* Backscatter normalized to 13^o incidence angle.

The model selection was based on a procedure developed for RF in previous studies (Li, 2013a, b; Li et al., 2013). To identify the most accurate predictive model, a cross-validation function, *rf.cv*, was developed (Li et al., 2013). This function enabled us to remove the least important variables based on variable importance (VI) of predictive variables, and keep the predictors unchanged among iterations. It is a stepwise method using both forward and backward selection to add or eliminate predictors, and uses predictive accuracy to select each predictive variable. A novel model selection measure, averaged variable importance (AVI), was used to select predictors in this study. AVI was based on VI (Li et al., 2013) and an R package '*extendedForest*' (Smith et al., 2011). Due to the randomness associated with the importance of predictive variables generated by RF algorithm, the order of important variable(s) may change with individual iterations; meanwhile, correlated variables may also affect the reliability of VI; so we used the AVI method based on the 100 times replication to generate the average values of

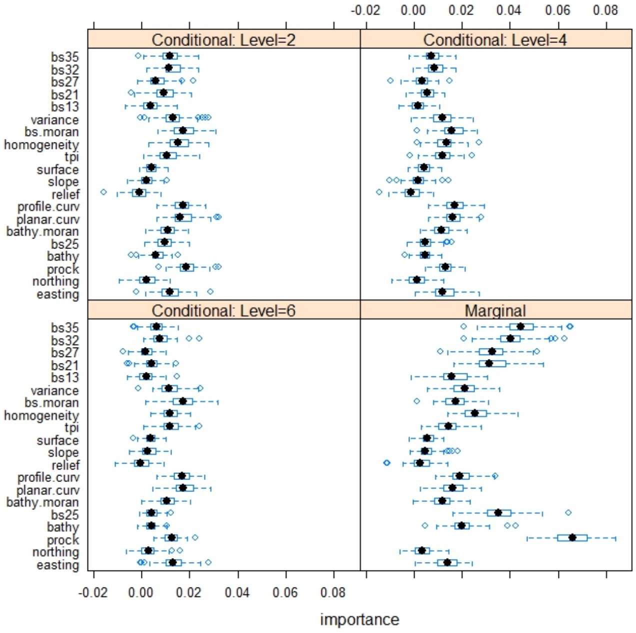

Figure 2. AVI based on 100 iterations of RF using the *extendedForest* package.

VI (Figure 2) that were used to select the predictors.

2.4. Model validation, comparison and spatial predictions

To assess the predictive accuracy of each model, we used 10-fold cross-validation (Hastie et al., 2009). To deal with the random error associated with each 10-fold cross validation as observed in the previous studies (Li, 2013a, b; Li et al., 2013), we repeated the cross validation procedure 100 times. The final results were based on the average of 100 iterations of the 10-fold cross validation. The correct classification rate (*ccr*) and *kappa* were used to measure the predictive accuracy.

The most accurate predictive model was used to predict seabed hardness at each 10m grid cell in the study areas. A portion of area A (A1) that comprises a variety of seabed geomorphic features was selected to illustrate and compare the predictions.

All modelling work was implemented in R 2.15.2 (R Development Core Team, 2012). Relevant maps were produced using ArcGIS (ESRI ® ArcMap ™ 10.0).

Table 2. A brief summary of RF modelling process based on AVI. The corresponding predictor for each number is listed in Table 1.

Model	Modelling process	Predictors
1	All 20 predictive variables	All 20 variables
2	model 1: -relief	1-7, 9-20
3	model 2: -northing	1, 3-7, 9-20
4	model 3: -bs13	1, 3-7, 9-11, 13-20
5	model 4: -bs27	1, 3-7, 9-11, 13-14, 16-
6	model 5: -slope	1, 3, 4-7, 10-11, 13-
7	model 6: -bs25	1, 3-7, 10-11, 13, 16-20
8	model 7: -bs21	1, 3-7, 10-11, 16-20
9	model 8: -bathy.moran	1, 3-4, 6-7, 10-11, 16-20
10	model 9: -surface	1, 3-4, 6-7, 11, 16-20
11	model 10: -bathy	1, 3, 6-7, 11, 16-20
12	model 11: -tpi	1, 3, 6-7, 16-20
13	model 12: -bs35	1, 3, 6-7, 16, 18-20
14	model 13: -variance	1, 3, 6-7, 16, 18, 20
15	model 14: -bs.moran	1, 3, 6-7, 16, 18
16	model 15: -bs32	1, 3, 6-7, 18
17	model 16: -easting	3, 6-7, 18
18	model 17: -profile.curv	3, 6, 18
19	model 18: -homogeneity	3, 6
20	model 19: -planar.curv	3
21	model 11: +bs21	1, 3, 6-7, 11, 13, 16-20
22	model 21: +bathy.moran	1, 3, 5-7, 11, 13, 16-20
23	model 21: -profile.curv	1, 3, 6, 11, 13, 16-20
24	model 21: -bs.moran	1, 3, 6-7, 11, 13, 16-19
25	model 24: -variance	1, 3, 6, 7, 11, 13, 16-18

3. RESULTS

3.1. Model selection using AVI

Twenty five models were developed based on the AVI for 20 predictive variables (Table 2, Figure 3). The first twenty models were developed by removing the least important variable based on AVI. Correct classification rates reached a local maximum for model 11. Two predictors (i.e. bs21 and bathy.moran) were identified as important variables and some predictors were identified as unimportant variables (e.g. profile.curv, bs.moran and variance). After further adding the important variables (i.e. their exclusions resulted in a decrease in the predictive accuracy) to model 11 and removing the unimportant predictors (i.e. their exclusions resulted in an increase in the predictive accuracy) from subsequent models, *ccr* increased and reached the highest mean value of 86.27% for model 24. *Kappa* displayed a similar pattern as *ccr* and reached the highest mean value of 0.4905 for model 24. Overall, model 24 with 10 predictors was the most accurate model compared to other models.

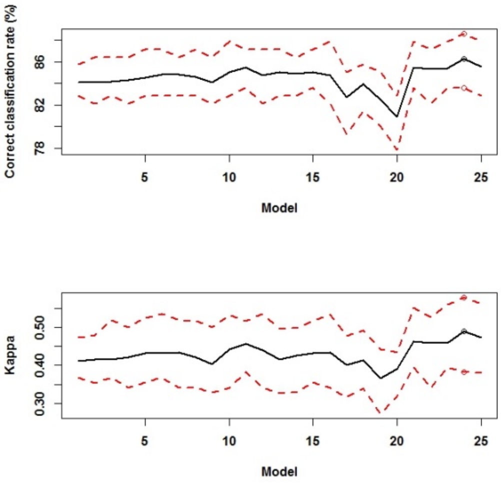

Figure 3. ccr (%) and kappa (mean: black line; minimum and maximum: dash red lines) of 25 RF models based on the averages over 100 iterations of 10-fold cross validation; and the model with the maximum mean ccr and mean kappa (circle).

3.2. Comparison of spatial predictions

The predictions for four hardness classes were similar with the predictions based on two classes (Li et al., 2013). Their match rate was 93.1% when the predictions of hard, hard-soft and soft-hard classes for four-class hardness data were pooled into one category (i.e. hard). The spatial predictions for four-class hardness data in area A1 were similar with the predictions based on two hardness classes (Li et al., 2013) (Figure 4), with a match rate of 88.9% when the predictions of hard, hard-soft and soft-hard for four-class hardness data were combined into a single category (i.e. hard).

4. DISCUSSION

4.1. Predictive accuracy of four-class seabed hardness

The predictive accuracy of the 25 models developed is high (Figure 3). The *kappa* of the most accurate model is 0.49 that is good according to Fielding & Bell (1997). The *ccr* of the most accurate models is even more notable. This demonstrates that the predictive accuracy of the model developed for predicting the seabed hardness is high. The high predictive accuracy suggests that: 1) seabed substrate was properly classified, and with high quality; 2) the predictors used were informative. Furthermore, the predictive accuracy for each model was stabilized and reliable because it is an averaged predictive accuracy based on 100 repetitions of 10-fold cross-validation. Hence we can confirm that:

- RF is an effective modelling method with high predictive accuracy for categorical data in multiple levels;
- a robust predictive model was developed;
- seabed hardness of four classes can be predicted with a high accuracy; and
- RF can be applied to 'small *p* and large *n*' problems in environmental sciences, with a *p* as small as one, as also observed in the previous study (Li et al., 2013).

The high accuracy of RF has been attributed to a number of features associated with RF as previously discussed (Li, 2013b; Li et al., 2011b; Li et al., 2011c).

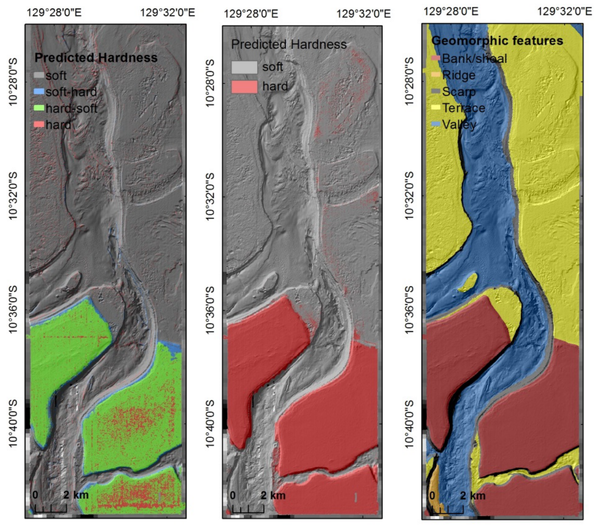

Figure 4. Spatial predictions of seabed hardness for a section of area A (A1) in the Joseph Bonaparte Gulf: a) hardness with four classes (left), b) hardness with two classes (middle), and c) geomorphic. features (right).

4.2. Predictive accuracy with correlated predictive variables and AVI

The predictive accuracy of models developed for four-class hardness data changes with highly correlated predictors. The influence on the accuracy varies with individual predictors. The inclusion of highly correlated predictors (i.e. bs32 and bs35, with a $\rho = 0.96$ and $r = 0.93$; bs21 and bs35, with a $\rho = 0.96$ and $r = 0.94$) was observed to improve the predictive accuracy (i.e., model 24) in this study. This could be explained by the fact that these correlated predictors are informative as they have relatively high variable importance (Figure 2) and their inclusion can increase the number of informative predictors selected for each individual tree in RF, thus improving the predictive accuracy, which is consistent with the findings in previous studies (Li, 2013b; Li et al., 2012b; Li et al., 2013). This suggests that correlated variables may be able to compensate for the small number of proxy predictive variables in environmental sciences. In contrast, the exclusion of some highly correlated predictors can also improve the predictive accuracy. It was observed that bs27 and bs25 were highly correlated with bs21 ($\rho \geq 0.98$ and $r = 0.99$), and the exclusion of bs27 from model 4 and the exclusion of bs25 from model 6 resulted in slight improvement in predictive accuracy. Similar findings were also observed in previous studies (Li et al., 2011b; Li et al., 2012a). These opposite effects imply that not all highly correlated predictors should be used even if they are of high VI or excluded, and highlight the fact that there are no short-cuts in identifying the optimal predictive model. The *extendedForest* (Smith et al., 2011) package can efficiently deal with the correlated variables in terms of the variable importance, but how to select predictors that improve predictive accuracy from correlated predictive variables is still a challenging task. This finding suggests that caution should be taken when using correlation coefficient to select predictors for RF in marine environmental sciences. These applications further demonstrate that model selection is necessary for RF in marine environmental sciences (Li et al., 2011b; Li et al., 2012b).

AVI has a couple of advantages over VI, AIC and BIC. AVI helps to produce a stable order of predictors and thus is preferable to VI (Li et al., 2013). AVI is based on the predictive accuracy and will select a model that is the most accurate or optimal instead of the most parsimonious as discussed above and in a previous study (Li et al., 2013). Traditional model selection methods such as AIC and BIC select the most parsimonious models that are not necessarily with most predictive accuracy. Since improving predictive accuracy is the ultimate goal of predictive modelling (Li et al., 2013), AVI is more appropriate for selecting predictive model(s). The principles underpinning the AVI can be easily applied to other machine learning methods as well as regression models. Therefore, it is recommended for selecting predictive model(s) in future studies.

4.3. Hardness classification methods and prediction maps of seabed hardness

The predictive accuracy of four hardness classes in this study was less than that of two classes in the previous study (Li et al., 2013; Siwabessy et al., 2013). A few factors were expected to negatively affect the predictive accuracy. These include that 1) divided hard class into three classes in this study was expected to reduce the accuracy; and 2) bathymetry was found to be no longer an important predictor in this study because bathymetry could no longer differentiate the classes derived from the previous hard class which is located at similar water depths. The spatial predictions for four-class hardness data were similar with the predictions based on two hardness classes (Li et al., 2013), with high match rates for all four areas and A1. These findings show that major patterns were captured in their predictions.

The predicted maps reflect the influence of various geomorphic features such as banks, terraces, and valleys (Figures 4). The associations of the predicted seabed hardness with geomorphic features are similar to what have been discussed in previous studies (Li et al., 2013; Siwabessy et al., 2013). These associations were supported by ecological studies as certain organisms were expected to be found on hard (Anderson et al., 2011; Przeslawski et al., 2011) and soft substratum (Anderson et al., 2011; Przeslawski et al., 2011) as observed in the corresponding substratum in this study.

5. CONCLUSIONS

This is the first attempt to predict the spatial distribution of seabed hardness to four classes. Seabed hardness of four classes is predictable and can be predicted into a spatially continuous layer with a high accuracy, especially for large areas where multibeam acoustic data exist and predictions of seabed classes are needed for marine planning and management. Model selection is essential for identifying an optimal predictive model for RF in environmental sciences. AVI demonstrates its effectiveness in searching for the most accurate predictive models and are recommended for future studies. This study further confirms the superior performance of RF in marine environmental sciences. RF is an effective modelling method with high predictive accuracy not only for presence/absence data and but also for multi-level categorical data. RF can be applied to ‘small *p* and large *n*’ problems in environmental sciences. RF and AVI are recommended for generating spatially continuous predictions of categorical variables in future studies.

ACKNOWLEDGEMENTS

We would like to thank Dale Roberts and Peter Tan for their valuable comments and suggestions. We thank the Master and crew of the RV Solander and scientific staff at the Australian Institute of Marine Science (AIMS) for their support in conducting the survey, Marcus Stowar (AIMS) for towed-video and still photography acquisition, and the Field and Engineering Support staff at Geoscience Australia. This study is published with the permission of the Chief Executive Officer, Geoscience Australia.

REFERENCES

Anderson, T.J., Nichol, S., Radke, L., Heap, A.D., Battershill, C., Hughes, M., Siwabessy, P.J., Barrie, V., Alvarez de Glasby, B., Tran, M., Daniell, J., Party, S., 2011. Seabed Environments of the Eastern Joseph Bonaparte Gulf, Northern Australia: GA0325/Sol5117 - Post-Survey Report. Geoscience Australia, Record 2011/08, 59pp.

Biau, G., 2012. Analysis of a random forest method. Journal of Machine Learning Research 13 1063-1095.

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

Diaz-Uriarte, R., de Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7(3) 1-13.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24(1) 38-49.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, New York.

Heap, A.D., Przeslawski, R., Radke, L., Trafford, J., Battershill, C., Party, S., 2010. Seabed Environments of the Eastern Joseph Bonaparte Gulf, Northern Australia. Sol4934 - Post-survey Report. Geoscience Australia, Record 2010/09, 78pp.

Kloser, R.J., Penrose, J.D., Butler, A.J., 2010. Multi-beam backscatter measurements used to infer seabed habitats. Continental Shelf Research 30 1772-1782.

Li, J., 2013a. Predicting the spatial distribution of seabed gravel content using random forest, spatial interpolation methods and their hybrid methods, The International Congress on Modelling and Simulation (MODSIM) 2013: Adelaide.

Li, J., 2013b. Predictive Modelling Using Random Forest and Its Hybrid Methods with Geostatistical Techniques in Marine Environmental Geosciences, In: Christen, P., Kennedy, P., Liu, L., Ong, K.-L., Stranieri, A., Zhao, Y. (Eds.), The proceedings of the Eleventh Australasian Data Mining Conference (AusDM 2013), Canberra, Australia, 13-15 November 2013. Conferences in Research and Practice in Information Technology, Vol. 146.

Li, J., Heap, A., Potter, A., Daniell, J.J., 2011a. Predicting Seabed Mud Content across the Australian Margin II: Performance of Machine Learning Methods and Their Combination with Ordinary Kriging and Inverse Distance Squared. Geoscience Australia, Record 2011/07, 69pp.

Li, J., Heap, A.D., Potter, A., Daniell, J., 2011b. Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software 26 1647-1659.

Li, J., Heap, A.D., Potter, A., Huang, Z., Daniell, J., 2011c. Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation of mud content across the southwest Australian margin. Continental Shelf Research 31 1365-1376.

Li, J., Potter, A., Heap, A., 2012a. Irrelevant Inputs and Parameter Choices: Do They Matter to Random Forest for Predicting Marine Environmental Variables?, Australian Statistical Conference 2012: Adelaide.

Li, J., Potter, A., Huang, Z., Daniell, J.J., Heap, A., 2010. Predicting Seabed Mud Content across the Australian Margin: Comparison of Statistical and Mathematical Techniques Using a Simulation Experiment. Geoscience Australia, 2010/11, 146pp.

Li, J., Potter, A., Huang, Z., Heap, A., 2012b. Predicting Seabed Sand Content across the Australian Margin Using Machine Learning and Geostatistical Methods. Geoscience Australia, Record 2012/48, 115pp.

Li, J., Siwabessy, J., Tran, M., Huang, Z., Heap, A., 2013. Predicting Seabed Hardness Using Random Forest in R, In: Zhao, Y., Cen, Y. (Eds.), Data Mining Applications with R. Elsevier, pp. 299-329.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2(3) 18-22.

Marmion, M., Luoto, M., Heikkinen, R.K., Thuiller, W., 2009. The performance of state-of-the-art modelling techniques depends on geographical distribution of species. Ecological Modelling 220(24) 3512-3520.

McArthur, M.A., Brooke, B.P., Przeslawski, R., Ryan, D.A., Lucieer, V.L., Nichol, S., McCallum, A.W., Mellin, C., Cresswell, I.D., Radke, L.C., 2010. On the use of abiotic surrogates to describe marine benthic biodiversity. Estuarine, Coastal and Shelf Science 88 21-32.

Mortensen, P., Buhl-Mortensen, L., 2004. Distribution of deep-water gorgonian corals in relation to benthic habitat features in the Northeast Channel (Atlantic Canada). Marine Biology 144(6) 1223-1238.

Okun, O., Priisalu, H., 2007. Random forest for gene expression based cancer classification: overlooked issues, In: Martí, J., Benedí, J.M., Mendonça, A.M., Serrat, J. (Eds.), Pattern Recognition and Image Analysis: Third Iberian Conference, IbPRIA 2007 Lecture Notes in Computer Science 4478, Springer-Verlag, Berlin: Girona, Spain, pp. 483-490.

Przeslawski, R., Daniell, J., Anderson, T., Vaughn Barrie, J., Heap, A., Hughes, M., Li, J., Potter, A., Radke, L., Siwabessy, J., Tran, M., Whiteway, T., Nichol, S., 2011. Seabed Habitats and Hazards of the Joseph Bonaparte Gulf and Timor Sea, Northern Australia. Geoscience Australia, Record 2008/23, 69pp.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna.

Saeys, Y., Inza, I., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23(19) 2507-2517.

Siwabessy, P.J.W., Daniell, J., Li, J., Huang, Z., Heap, A.D., Nichol, S., Anderson, T.J., Tran, M., 2013. Methodologies for seabed substrate characterisation using multibeam bathymetry, backscatter and video data: A case study from the carbonate banks of the Timor Sea, Northern Australia: Geoscience Australia, Record 2013/11, 82pp.

Smith, S.J., Ellis, N., Pitcher, C.R., 2011. Conditional variable importance in R package extendedForest. R vignette 〈http://gradientforest.r-forge.r-project.org/Conditional-importance.pdf〉.

Stein, D.L., Tissot, B.N., Hixon, M.A., Barss, W.H., 1992. Fish–habitat associations on a deep reef at the edge of the Oregon continental shelf. Fisheries Bulletin 90 540-551.

Wentworth, C.K., 1922. A scale of grade and class terms for clastic sediments. Journal of Geology 30 377-392.