Environmental Modelling & Software 97 (2017) 112–129

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Application of random forest, generalised linear model and their hybrid methods with geostatistical techniques to count data: Predicting sponge species richness

Jin Li [a, *], Belinda Alvarez [b, 1], Justy Siwabessy [a], Maggie Tran [a], Zhi Huang [a], Rachel Przeslawski [a], Lynda Radke [a], Floyd Howard [a], Scott Nichol [a]

[a] Geoscience Australia, GPO Box 378, Canberra, ACT 2601, Australia
[b] Museum and Art Gallery of the Northern Territory, PO Box 4646, Darwin, NT 0801, Australia

ARTICLE INFO

Article history:
Received 13 February 2017
Received in revised form
19 May 2017
Accepted 27 July 2017
Available online 6 August 2017

Keywords:
Machine learning
Feature selection
Model selection
Predictive accuracy
Spatial predictive model
Spatial prediction

ABSTRACT

Spatial distribution of sponge species richness (SSR) and its relationship with environment are important for marine ecosystem management, but they are either unavailable or unknown. Hence we applied random forest (RF), generalised linear model (GLM) and their hybrid methods with geostatistical techniques to SSR data by addressing relevant issues with variable selection and model selection. It was found that: 1) of five variable selection methods, one is suitable for selecting optimal RF predictive models; 2) traditional model selection methods are unsuitable for identifying GLM predictive models and joint application of RF and AIC can select accuracy-improved models; 3) highly correlated predictors may improve RF predictive accuracy; 4) hybrid methods for RF can accurately predict count data; and 5) effects of model averaging are method-dependent. This study depicted the non-linear relationships of SSR and predictors, generated spatial distribution of SSR with high accuracy and revealed the association of high SSR with hard seabed features.

1. Introduction

The broad continental shelf offshore northern Australia is characterised by extensive areas of carbonate banks, terraces and isolated pinnacles. These raised features are of important conservation value because they provide potential habitats for sponge communities and have been assigned as Key Ecological Features (KEFs) of regional significance within the Oceanic Shoals Commonwealth Marine Reserve (CMR) (Australia, 2012a, b) (Fig. 1). Information on the distribution of the communities is limited (Huang et al., 2011), however. Previous assessments of the linear relationship of sponge species with environmental variables are provided for the communities on the Van Diemen Rise and the importance of carbonate banks and other raised geomorphic features as biodiversity hotspots was studied (Przeslawski et al., 2014, 2015). An improved understanding of the spatial patterns of sponge species richness (SSR) is important for refining knowledge regarding the ecological significance of the KEFs and for the informed monitoring of ecosystem health and marine environmental management and conservation. The spatially continuous data of SSR across the region is not readily available and the relationships between the richness and environmental variables across the region are largely unknown. Therefore, predictive models for species richness may address the spatial data gaps and could be used to investigate the ecological relationships.

Many statistical and mathematical techniques can be used for generating spatially continuous predictions for numerical variables (Li and Heap, 2014; Li et al., 2011b, 2011c), but they are often data specific and their performance depends on many factors (Li and Heap, 2011). The accuracy of the predictions is crucial for informed monitoring design and evidence-based policy for marine environmental management and conservation of the CMR. Due to its high predictive accuracy in data mining and other disciplines (Cutler et al., 2007; Diaz-Uriarte and de Andres, 2006; Prasad et al., 2006; Shan et al., 2006), random forests (RF) method was introduced to spatial statistics by combining it with commonly used geostatistical methods to predict the spatial distribution of seabed

* Corresponding author.
E-mail address: Jin.Li@ga.gov.au (J. Li).
[1] Current address, Lund University, Department of Geology, Sölvegatan 12, SE-223 62 LUND, Sweden.

http://dx.doi.org/10.1016/j.envsoft.2017.07.016

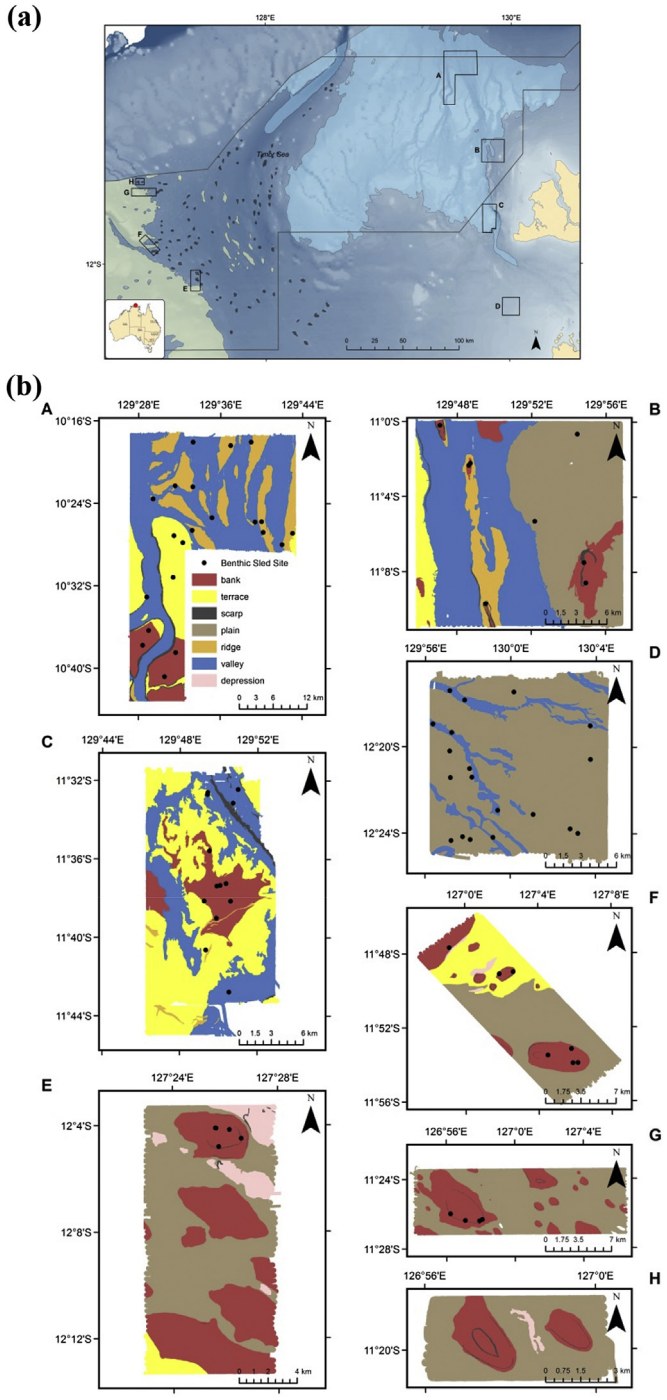


Fig. 1. a) Location of the study areas (A, B, C, D, E, F, G and H) and associated geomorphic features in the Timor Sea region, northern Australian marine margin overlaid with bathymetry; the border of Oceanic Shoals Commonwealth Marine Reserve is indicated by grey line; and Key Ecological Features (KEFs) include: the carbonate banks and terraces of the Van Diemen Rise (North Marine Region) (blue); the carbonate banks and terraces of the Sahul Shelf (Northwest Marine Region) (yellow), and the pinnacles of the Bonaparte Basin (North and Northwest Marine Regions) (black). b) The sampling transects (black dots) within each study area overlaid with associated geomorphic features; and in total 77 sponge species richness samples were available for this study. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sediments (Li, 2011; Li et al., 2010b). This development opened an alternative source of methods for spatial prediction. The hybrid methods, RFOK and RFIDW (i.e. the hybrids of RF with ordinary kriging (OK) or inverse distance weighting (IDW)), have shown high predictive capacity in the marine environmental sciences (Li et al., 2011b, 2011c, 2012b) and terrestrial environmental sciences (Sanabria et al., 2013a, 2013b). However, these methods have not been applied to count data (e.g. SSR data) and may also be data-type specific like other spatial prediction methods.

Although generalised linear models (GLM) and its hybrid methods with a geostatistical technique (i.e., RKGLM) were applied to percentage data (Li et al., 2011b, 2011c, 2010b), they were proven to be less accurate than RF and its hybrid methods with geostatistical techniques. For count data, the commonly used statistical modelling method is GLM, so it was used in this study as a control to be compared with other methods. The hybrid methods of GLM with geostatistical methods have not yet been applied to any other types. This is the first attempt to apply them to count data and test their predictive accuracy for such data type.

Variable selection (or feature selection) is essential for selecting an optimal RF predictive model from a number of candidate models (Li, 2013a, b; Li et al., 2016), although RF is often argued to be insensitive to non-important variables (Okun and Priisalu, 2007) and can deliver good predictive performance even when most predictive variables are noisy (Diaz-Uriarte and de Andres, 2006). The performance of RF is also argued to depend only on the number of strong features and not on the number of noisy variables if sample size is large (500–1000) (Biau, 2012). Variable selection is also essential for its hybrid methods (Li et al., 2011a, b, 2012a, b). Several variable selection methods (i.e. variable importance (VI), averaged VI (AVI), knowledge-informed AVI (KIAVI), Boruta and regularized RF (RRF)) were tested for RF based on a model selection procedure developed previously by Li et al. (2013), where predictive accuracy was used to determine the selection of each predictive variable; and AVI and Boruta (Kursa and Rudnicki, 2010) were recommended for selecting RF and other predictive models (Li et al., 2016).

For GLM, besides the conventional model selection approaches such as stepAIC (Venables and Ripley, 2002), RF has been shown to be able to select useful predictors for GLM (Arthur et al., 2010). This may provide a useful approach for selecting informative predictors for GLM and its hybrid methods with geostatistical methods in this study.

It has been argued that model averaging can often improve predictive accuracy (Marmion et al., 2009). This was tested in the geostatistical context, but findings amongst studies have not been consistent. For example, model averaging did not significantly improve seabed mud predictions (Li et al., 2011b) and seabed gravel predictions (Li, 2013a) although it was found to improve seabed sand predictions (Li et al., 2012b). Therefore, further study is warranted.

In this study, we aim to select the most accurate model to predict the spatial distribution of SSR within the Oceanic Shoals CMR in the Timor Sea offshore, northern Australia, based on samples of SSR using acoustic multibeam data and their derived variables as predictive variables. To achieve this, we: (1) tested the predictive accuracy of models based on GLM, RF and their hybrid methods with OK and IDW; (2) tested the effects of various predictor sets on the predictive accuracy of these methods; (3) examined the influence of a few feature selection methods on the most accurate predictive model identified; and (4) predicted the spatial distribution of sponge richness using the most accurate model and visually examined the predictions.

2. Data processing and methodology

2.1. Study region

The study region is located in the Timor Sea region, northern

Australia (Fig. 1). Eight areas (A - H) in the region were surveyed in 2009 (Heap et al., 2010), 2010 (Anderson et al., 2011) and 2012 (Nichol et al., 2013). These areas were selected to encompass a variety of seabed geomorphic features and water depths. In these surveys, high-resolution multibeam bathymetry and backscatter data and co-located sampling transects across the eight areas were acquired. The areas comprise a spatially complex suite of geomorphic features including shallow flat-topped banks, terraces, ridges, depressions and plains. In survey areas A – D, sampling sites were selected to cover all of the seabed features and water depths; and in the survey areas E –H, sampling sites were selected using a spatially-balanced random stratified method as stated in above post-survey reports and further detailed in relevant studies (Przeslawski et al., 2011; Radke et al., 2015, 2017).

2.2. Sponge samples

Sponges were collected with an epibenthic sled towed for approximately 50–100 m at 1.5–2 knots in each sampling site. The sled was 1.5 × 1 m (width x height) and fitted with a 6 m long, 45 mm stretch diamond net. Specimens were sorted to phylum, weighed, and preserved in ethanol for taxonomic identification. Taxonomic vouchers of sponges were deposited at the Museum and Art Gallery of the Northern Territory (MAGNT, formerly the Northern Territory Museum). A thick section and spicule slide was prepared from each sponge voucher using standard methods (Hooper, 1996; Rützler, 1978), identified to genus following Hooper and Van Soest (2002), and assigned to valid species as listed in the current version of the World Porifera database (Van Soest et al., 2014) using available taxonomic literature. A unique code or operational taxonomic unit (OTU) was assigned to unknown or undescribed taxa (e.g. *Scleritoderma* sp. NT0205). We used presence/absence data due to potential issues arising from sled sampling such as fragmentation and unstandardised effectiveness of collection (Schlacher et al., 2007). There were 85 samples collected, and of which eight samples were excluded due to the uncertainty about transect length. In total, 77 samples were selected and used in this study. SSR is count data based on the presence/absence data, ranging from 1 to 39, with a mean of 10.48 and a standard deviation of 10.53. The point locations of samples are the mid-point of each transect.

2.3. Predictive variables

Following a preliminary analysis based on data availability and the relationships with seabed hardness as discussed above and in previous studies, 80 predictive variables were available for this study. They are:

1) Two location variables: latitude (lat) and longitude (long),
2) Three sediment variables: mud, sand and gravel,
3) Bathymetry (bathy),
4) Twenty-seven backscatter (bs) variables (bs10 to bs36): a diffused reflection of acoustic energy due to scattering process back to the direction from which it's been generated, measured as the ratio of the acoustic energy sent to a seabed to that returned from the seabed, normalised to incidence angles between 10° and 36°,
5) Seventeen derived variables from bs25 based on object and windows (30 m, 50 m and 70 m) approach:
   a. bs_o,
   b. homogeneity (bs_homo_o, bs_homo3, bs_homo5, bs_homo7),
   c. entropy (bs_entro_o, bs_entro3, bs_entro5, bs_entro7),
   d. Local Moran I (bs_lmi_o, bs_lmi3, bs_lmi5, bs_lmi7),
   e. Variance (bs_var_o, bs_var3, bs_var5, bs_var7).
6) Twenty-nine derived variables from bathy using object and windows (30 m, 50 m and 70 m) approach:
   a. bathy_o,
   b. lmi_o, lmi3, lmi5, lmi7,
   c. Topographic position index (tpi_o, tpi3, tpi5, tpi7),
   d. Seabed slope (slope_o, slope3, slope5, slope7),
   e. Planar curvature (plan_cur_o, plan_cur3, plan_cur5, plan_cur7),
   f. Profile curvature (prof_cur_o, prof_cur3, prof_cur5, prof_cur7),
   g. Topographic relief (relief_o, relief3, relief5, relief7),
   h. Seabed rugosity (rugosity_o, rugosity3, rugosity5, rugosity7).
7) Distance to coast (dist.coast)

Acquisition and processing of multibeam bathymetry, backscatter and their derived variables have been detailed in previous studies (Li et al., 2013; Siwabessy et al., 2013) and in Appendix A. All these variables were numerical and available for each grid cell at 250 m resolution in the eight study areas for generating the spatial predictions of SSR. They were also available at the 77 sample locations for developing models to predict SSR (Appendix B) and some summary statistics of these variables were also provided (Appendix B1).

2.4. Preliminary selection of predictive variables

There were strong correlations among some predictive variables based on Spearman's rank correlation (ρ) that was used due to non-linear relationships between some variables. Amongst the highly correlated predictors (i.e., $\rho \geq 0.99$), the variable with the highest ρ with species richness was retained. As a result, in total 49 variables were retained (Table 1). The bs25 was retained as it is the default predictor for Geoscience Australia. The Spearman's rank correlations of SSR and these variables were presented in Appendix C. The Pearson's correlation coefficients (r) were also calculated (Appendix D). In addition, a categorical variable, geomorphic features (geom), was also considered in this study and was the 50th predictor (Table 1).

Table 1
Predictive variables and their corresponding number.

No	Predictive variable	No	Predictive variable
1	long	26	tpi5
2	lat	27	tpi7
3	sand	28	lmi3
4	gravel	29	plan_curv3
5	bs25	30	plan_curv5
6	bs11	31	plan_curv7
7	bs14	32	relief3
8	bs34	33	relief5
9	bs_o	34	relief7
10	bs_homo_o	35	slope3
11	bs_entro_o	36	slope5
12	bs_var_o	37	prof_curv3
13	bs_lmi_o	38	prof_curv5
14	bathy_o	39	prof_curv7
15	tpi_o	40	bs_entro3
16	slope_o	41	bs_entro5
17	plan_cur_o	42	bs_entro7
18	prof_cur_o	43	bs_homo3
19	relief_o	44	bs_homo5
20	rugosity_o	45	bs_homo7
21	dist.coast	46	bs_var3
22	rugosity3	47	bs_var5
23	rugosity5	48	bs_var7
24	rugosity7	49	bs_lmi5
25	tpi3	50	geom

2.5. Predictive methods

In this study, we used GLM and RF and their hybrid methods with geostatistical techniques (Table 2) to develop an optimal predictive model to predict SSR. GLM was used as a control for the count data. As regression kriging was argued to be more accurate than other methods (Hengl, 2007; Li and Heap, 2008), GLMOK, also known as RKGLM (i.e., the combination of GLM and OK) (Li et al., 2010a, 2011b, 2010b), was then used in this study. We also developed a new combination of GLM and IDW (GLMIDW) for this study. Poisson distribution was used in GLM.

The R function, randomForest developed by Liaw and Wiener (2002), was used to develop a model to predict the spatial distribution of SSR. The default values of mtry, ntree and nodesize are often good options (Diaz-Uriarte and de Andres, 2006; Liaw and Wiener, 2002), which were also observed in studies in marine environmental sciences (Li et al., 2012b, 2013), therefore the default values were used for these parameters. The parameters for RFOK and RFIDW were based on the minimum number of samples per region and findings of previous studies (Li, 2013b; Li et al., 2011a). That is, of these hybrid methods, a distance power of 2 and a searching window size of 7 were used for IDW, and a Spherical model and a searching window size of 7 were used for OK. Although a thorough test was not performed, seven was identified as a better choice than 4 or 5 based on a preliminary test of the full model (i.e. the model with all 49 numerical predictors).

We also tested if model averaging could improve the predictive accuracy. In this study, we averaged the predictions of RFOK and RFIDW (RFOKRFIDW), of RF, RFOK and RFIDW (RFRFOKRFIDW), of GLMOK and GLMIDW (GLMOKGLMIDW), of GLM, GLMOK and GLMIDW (GLMGLMOKGLMIDW) to examine the effects of model averaging on predictive accuracy.

2.6. Variable selection and model development

2.6.1. Random forest

The variable selection was based on a procedure developed for RF in previous studies (Li, 2013a, b; Li et al., 2013), where the features were selected based on variable importance and more importantly on the accuracy of the resultant predictive model. The final selection of a predictor was based on its contribution to predictive accuracy, that is, only those predictors that could improve the predictive accuracy were selected. This procedure involved two steps. One step involved selecting model predictors (i.e., feature selection), and the other was to estimate the predictive accuracy of the model (addressed in the next section). To select predictive variables, we adopted the same principle as used in *rfcv* in randomForest package in R (Liaw and Wiener, 2002), that is, identifying and removing the least important variables based on the importance of predictive variables.

Five feature selection methods were used to select predictors in this study: (1) averaged variable importance (AVI), (2) Boruta, (3) knowledge informed AVI (KIAVI) as detailed in previous studies (Li et al., 2013, 2016), (4) recursive feature selection (rfe) (Kuhn, 2014) and (5) variable selection using RF (VSURF) (Genuer et al., 2015). The AVI and Boruta were chosen because they produced the most accurate predictive models in a previous study (Li et al., 2016); the remaining three methods were used as they may lead to accuracy improved RF models for count data. Due to the randomness associated with the importance of predictive variables generated by RF algorithm, the least important variable(s) may change with individual iterations; meanwhile correlated variables may also affect the order of the least important variable(s). Therefore, the R package 'extendedForest' (Smith et al., 2011) was used with 100 repetitions to stabilise the variable importance of RF and to generate the average values of variable importance that were used to select the predictors. The process of the model development was detailed in Table 3. We reduced the full model that used all numerical predictors (Table 1) by progressively removing the least important variable(s) from the previous model based on AVI, which resulted in 23 models.

The next step was to identify the important and unimportant predictors based on the predictive accuracy of the models developed so far (Li et al., 2016). The important predictors were added back to the most accurate model based on AVI (i.e. model 22) to determine if further improvement was possible, which was referred to as KIAVI. We also examined if the inclusion of a categorical variable, geomorphic features (geom), could improve the accuracy because geom was considered to be a potentially important predictor. We then repeated the above procedure, which resulted in a further 10 models (i.e. models 24–33).

From model 24 onwards, we tested to see if the accuracy could be improved by removing geom from model 24. This generated a further 7 models (i.e. models 34–40) and also identified further unimportant predictors.

We then removed the unimportant predictors from the most accurate model identified so far (i.e. model 34), which resulted in three models (i.e. models 41–43). Model 43 was further simplified by removing the least important variables based on AVI, which generated another three models (i.e. models 44–46). For model 46, removal the least important variable would have led to a model identical to model 40, and so no further model reduction was pursued.

We used Boruta to search for the important predictors for RF. The default value (i.e., 100) and the values of 2000 and 5000 were used for the maximal number of importance source runs in Boruta, which resulted in three models. For model 1, long, lat, sand, bs34, bathy_o, slope_o, prof_cur_o, rugosity_o, dist.coast, rugosity3, tpi3, tpi5, tpi7, relief3 and prof_curv5, bs_entro7 were used. Model 2 was similar to model 1, with removal of bs34 and tpi7 and adding bs_var7 to the predictors. For model 3, long, lat, prof_cur_o, rugosity_o, tpi3, tpi5 and relief3 were used.

The rfe selected all 49 variables, which produced a model identical to the full model. The VSURF selected three variables that

Table 2
Full name for the abbreviations of all modelling methods, feature selection methods and measures of model performance in this study.

Abbreviation	Full name
GLM	Generalised linear model
IDW	Inverse distance weighting
OK	Ordinary kriging
RF	Random forest
rfe	Recursive feature selection
RMAE	Relative mean absolute error
RRMSE	Relative root mean square error
SSR	Sponge species richness
VEcv	Variance explained by predictive models
VI	Variable importance
AVI	Averaged VI
GLMGLMOKGLMIDW	The average of GLM, GLMOK and GLMIDW
GLMIDW	The hybrids of GLM with IDW
GLMOK	The hybrids of GLM with OK
GLMOKGLMIDW	The average of GLMOK and GLMIDW
KIAVI	Knowledge-informed AVI
RFIDW	The hybrids of RF with IDW
RFOK	The hybrids of RF with OK
RFOKRFIDW	The average of RFOK and RFIDW
RFRFOKRFIDW	The average of RF, RFOK and RFIDW
RKGLM	The hybrids of GLM with OK
RRF	Regularized RF
VSURF	Variable selection using RF

Table 3
A brief summary of RF modelling process for sponge species richness data using various feature selection methods and predictive variables. 1) models 1–23 based on the AVI using 49 variables; 2) models 24–33 based on KIAVI using the important predictors identified from models 1–23 and also included geomorphic features (geom) as an additional predictor; 3) models 34–40 based on KIAVI using model 24 and the AVI; 4) models 41–43 based on KIAVI by removing the unimportant predictors identified from model 34–40; and 5) model 44–46 based on KIAVI by using model 43 and the AVI. The corresponding predictor for each number is listed in Table 1.

Model	Modelling process	Predictors	No. of predictors
1	Full model: all 49 numerical predictors	All 49 numerical variables	49
2	model 1: - bs_entro_o, bs_var_o, plan_cur_o, plan_cur7, slope5, bs_entro3, bs_homo5, bs_var3, bs_lmi5	1:10, 13:16, 18:30, 32:35, 37:39, 41:43, 45, 47:48	40
3	model 2: - bs_lmi_o, relief_o, rugosity7, plan_curv5	1:10, 14:16, 18, 20:23, 25:29, 32:35, 37:39, 41:43, 45, 47:48	36
4	model 3: - gravel, bs_homo_o, lmi3, plan_curv3, relief7, prof_curv3, prof_curv7, bs_homo3, bs_var5	1:3, 5:9, 14:16, 18, 20:23, 25:27, 32:33, 35, 38, 41:42, 45, 48	27
5	model 4: - bs14, tpi_o, relief5, slope3, bs_homo7	1:3, 5:6, 8:9, 14, 16, 18, 20:23, 25:27, 32, 38, 41:42, 48	22
6	model 5: - rugosity5, tpi7, bs_entro5	1:3, 5:6, 8:9, 14, 16, 18, 20:22, 25:26, 32, 38, 42, 48	19
7	model 6: - bs25	1:3, 6, 8:9, 14, 16, 18, 20:22, 25:26, 32, 38, 42, 48	18
8	model 7: - slope_o	1:3, 6, 8:9, 14, 18, 20:22, 25:26, 32, 38, 42, 48	17
9	model 8: - bs_var7	1:3, 6, 8:9, 14, 18, 20:22, 25:26, 32, 38, 42	16
10	model 9: - prof_curv5	1:3, 6, 8:9, 14, 18, 20:22, 25:26, 32, 42	15
11	model 10: - bs_entro7	1:3, 6, 8:9, 14, 18, 20:22, 25:26, 32	14
12	model 11: - sand	1:2, 6, 8:9, 14, 18, 20:22, 25:26, 32	13
13	model 12: - dist.coast	1:2, 6, 8:9, 14, 18, 20, 22, 25:26, 32	12
14	model 13: - bs_o	1:2, 6, 8, 14, 18, 20, 22, 25:26, 32	11
15	model 14: - rugosity_o	1:2, 6, 8, 14, 18, 22, 25:26, 32	10
16	model 15: - bs34	1:2, 6, 14, 18, 22, 25:26, 32	9
17	model 16: - relief3	1:2, 6, 14, 18, 22, 25:26	8
18	model 17: - bathy_o	1:2, 6, 18, 22, 25:26	7
19	model 18: - bs11	1:2, 18, 22, 25:26	6
20	model 19: - prof_curv_o	1:2, 22, 25:26	5
21	model 20: - rugosity3	1:2, 25:26	4
22	model 21: - tpi5	1:2, 25	3
23	model 22: - tpi3	1, 2	2
24	model 21: + bs_var7, bs_entro7, dist.coast, bs_o, bs34, bathy_o, bs11, geom	1:2, 25, 48, 42, 21, 9, 8, 14, 6, 50	11
25	model 24: - bs_var7	1:2, 25, 42, 21, 9, 8, 14, 6, 50	10
26	model 25: - bs_entro7	1:2, 25, 21, 9, 8, 14, 6, 50	9
27	model 26: - bs_o	1:2, 25, 21, 8, 14, 6, 50	8
28	model 27: - tpi3	1:2, 21, 8, 14, 6, 50	7
29	model 28: - bathy_o	1:2, 21, 8, 6, 50	6
30	model 29: - bs34	1:2, 21, 6, 50	5
31	model 30: - bs11	1:2, 21, 50	4
32	model 31: - geom	1:2, 21	3
33	model 32: - lat	1, 21	2
34	model 21: + bs_var7, bs_entro7, dist.coast, bs_o, bs34, bathy_o, bs11	1:2, 25, 48, 42, 21, 9, 8, 14, 6	10
35	model 34: bs_var7	1:2, 25, 42, 21, 9, 8, 14, 6	9
36	model 35: - bs_entro7	1:2, 25, 21, 9, 8, 14, 6	8
37	model 36: - tpi3	1:2, 21, 9, 8, 14, 6	7
38	model 37: - bs_o	1:2, 21, 8, 14, 6	6
39	model 38: - bs34	1:2, 21, 14, 6	5
40	model 39: - bathy_o	1:2, 21, 6	4
41	model 34: - bs_o	1:2, 25, 48, 42, 21, 8, 14, 6	9
42	model 34: - bathy_o	1:2, 25, 48, 42, 21, 9, 8, 6	9
43	model 34: - bs_o, bathy_o	1:2, 25, 48, 42, 21, 8, 6	8
44	model 43: - bs_var7	1:2, 25, 42, 21, 8, 6	7
45	model 44: - bs_entro7	1:2, 25, 21, 8, 6	6
46	model 45: - tpi3	1:2, 21, 8, 6	5

were identical to those in model 22 based on AVI. Thus no further modelling work was needed for these two methods.

Finally, we examined the residuals of RF for the most accurate model developed so far (i.e., model 43) and found the square root transformation could normalise the residuals. We then applied OK to the transformed the residuals and tested whether the predictive accuracy could be further improved.

2.6.2. Generalised linear model

Three traditional model selection approaches for GLM were used to select predictive models: (1) stepAIC; (2) dropterm and (3) anova (Venables and Ripley, 2002). We used a backward direction with a $k = \log(n)$ for stepAIC, chi-square test with a $k = \log(n)$ for dropterm, and chi-square test for anova. Firstly, we used stepAIC to choose a model (i.e. GLM1) from a full model, containing all 49 numerical predictors. We then simplified GLM1 using dropterm and anova to remove non-significant predictors and developed a further model (i.e. GLM2). We then considered possible two-way interactions of remaining predictors in the model with lowest AIC (i.e. GLM1) and simplified this newly formed model using stepAIC; and we then added a few second orders based on the relationships of species richness with relevant predictors to this model and further simplified it using stepAIC, dropterm and anova, which led to the third model (i.e. GLM3).

Given that the above modelling effort using GLM produced models with poor predictive accuracy and RF was found to be able to select useful predictors for GLM (Arthur et al., 2010), we used the predictors in the most accurate RF model and developed the fourth GLM model (i.e. GLM4). We developed one further model (i.e. GLM5) by simplifying GLM4 based on stepAIC, and the sixth model

(i.e. GLM6) by simplifying GLM5 using and dropterm and anova respectively. We added two-way interactions and one second order term to the most accurate GLM model so far (i.e. GLM6) and used the stepAIC and further developed a model (i.e. GLM7). Moreover, we only added the second order to GLM6 and simplified the model using stepAIC, resulting in model GLM8. Finally, we considered two-way interactions of predictors in GLM8 and developed a further model GLM9 based on stepAIC. The details of the resultant 9 models were presented in Table 4. The model selection approaches for GLM were summarised in Table 5.

2.7. Model validation

To evaluate the performance of the models developed using above selection methods, a 10-fold cross-validation was used (Hastie et al., 2009; Kohavi, 1995). To reduce the influence of randomness associated with the 10-fold cross-validation, it was repeated 100 times (Li, 2013a, b; Li et al., 2013). The choice of this iteration number was based on findings in previous studies (Li, 2013b; Li et al., 2013). Relative mean absolute error (RMAE) and relative root mean square error (RRMSE) (Li and Heap, 2011) were used to assess the performance of the methods tested. Variance explained by predictive models (VEcv) (Li, 2016) was used to assess the predictive accuracy of the models. MAE, RMAE and RMSE were also provided for readers interested.

2.8. Model comparison and spatial predictions

Since the VEcv values were either not normally distributed based on the Shapiro-Wilk normality test, with heterogeneous variance based on Fligner-Killeen test of homogeneity of variances, or both, Mann-Whitney tests were used to compare the difference in accuracy between the most accurate models developed for GLM and RF using various model and variable selection methods and the effects of model averaging.

2.9. Software and data availability

The modelling was implemented in R 3.0.2 (2013), using the 'raster' packages for extracting data from different data layers (Hijmans, 2014), 'gstat' for geostatistical modelling (Pebesma, 2004), 'MASS' for generalised linear models (Venables and Ripley, 2002), and 'randomForest' for random forest modelling (Liaw and Wiener, 2002). Finally, the most accurate predictive models were used to predict SSR at each 250 m grid cell in the study areas. Relevant maps were then produced using ArcGIS (ESRI® ArcMap TM 10.0).

A dataset of SSR and all predictive variables at 77 sample locations was provided as a spreadsheet in Appendix B.

3. Results

3.1. Predictive model using RF, RFIDW, RFOK, RFOKRFIDW and RFRFOKRFIDW

3.1.1. Variable selection using AVI

The RRMSE of RF, RFIDW, RFOK, RFOKRFIDW and RFRFOKRFIDW fluctuated from model 1 to model 23 and reached a minimum mean for model 22, especially for RFIDW with a RRMSE of 78.12% (models 1–23 in Table 3, Fig. 2). Seven important predictors were identified: bs_var7, bs_entro7, dist.coast, bs_o, bs34, bathy_o, and bs11. After reaching the model (i.e., model 23) with only two predictors remaining, no further model reduction was proceeded as these two predictors were location information and there is dramatic drop in RRMSE. Overall, model 22 for RFIDW was more accurate than all other models. It contained three predictors (Table 3).

3.1.2. Variable selection based on KIAVI

In total, 10 models were developed for each of RF, RFIDW, RFOK, RFOKRFIDW and RFRFOKRFIDW based on above important variables and an additional predictor: geomorphic features (geom) (models 24–33 in Fig. 2, Table 3). RRMSE increased from model 24 to model 33, except model 32. Model 24 was of a minimum mean of RRMSE, especially for RFOKRFIDW with a RRMSE of 75.99%. One

Table 4
A brief summary of GLM modelling process for sponge species richness.

Model	Modelling process	Predictors	No. of predictors
1	full model: - 18 predictors including long	2, 3, 5:8, 11:12, 15:16, 18, 20, 22:25, 27:30, 32:36, 41, 43:47	31
2	full model: - 36 predictors including long	2, 3, 12, 15:16, 18, 20, 22, 24, 28, 29, 45, 47	13
3	full model: - 40 predictors including long, + lat:rugosity5, sand:prof_cur_o	2, 3, 12, 16, 20, 22, 23, 28, 47, 2*23, 3*18	11
4	As in model 43 for rf	1:2, 25, 48, 42, 21, 8, 6	8
5	model 4: - bs_entro7	1:2, 25, 48, 21, 8, 6	7
6	model 5: - tpi3, bs11	1:2, 48, 21, 8	5
7	model 6: + lat^2, long:bs_var7, long:dist.coast, long:bs34, lat:bs34, dist.coast:bs34	1:2, 48, 21, 8, 2*2, 1*48, 1*21, 1*8, 2*8, 21*8	11
8	model 6: - dist.coast, + lat^2	1:2, 48, 8, 2*2	5
9	model 8: + longlat, long:bs_var7, long:bs34, lat:bs34	1:2, 48, 8, 2*2, 1*2, 1*48, 1*8	8

Table 5
The selection of GLM predictive models.

Predictor	Selection method	Selection criteria	Resultant model
All 49 variables	stepAIC	BIC	GLM1
Predictors in GLM1	dropterm & anova	*p*-value	GLM2
Predictors in GLM1 & two-way interactions & second orders	stepAIC, dropterm & anova	BIC & *p*-value	GLM3
Predictors in RF43	n/a	n/a	GLM4
Predictors in GLM4	stepAIC	BIC	GLM5
Predictors in GLM5	dropterm & anova	*p*-value	GLM6
Predictors in GLM6 & two-way interactions & second order	stepAIC	BIC	GLM7
Predictors in GLM6 & second orders	stepAIC	BIC	GLM8
Predictors in GLM8 & two-way interactions	stepAIC	BIC	GLM9

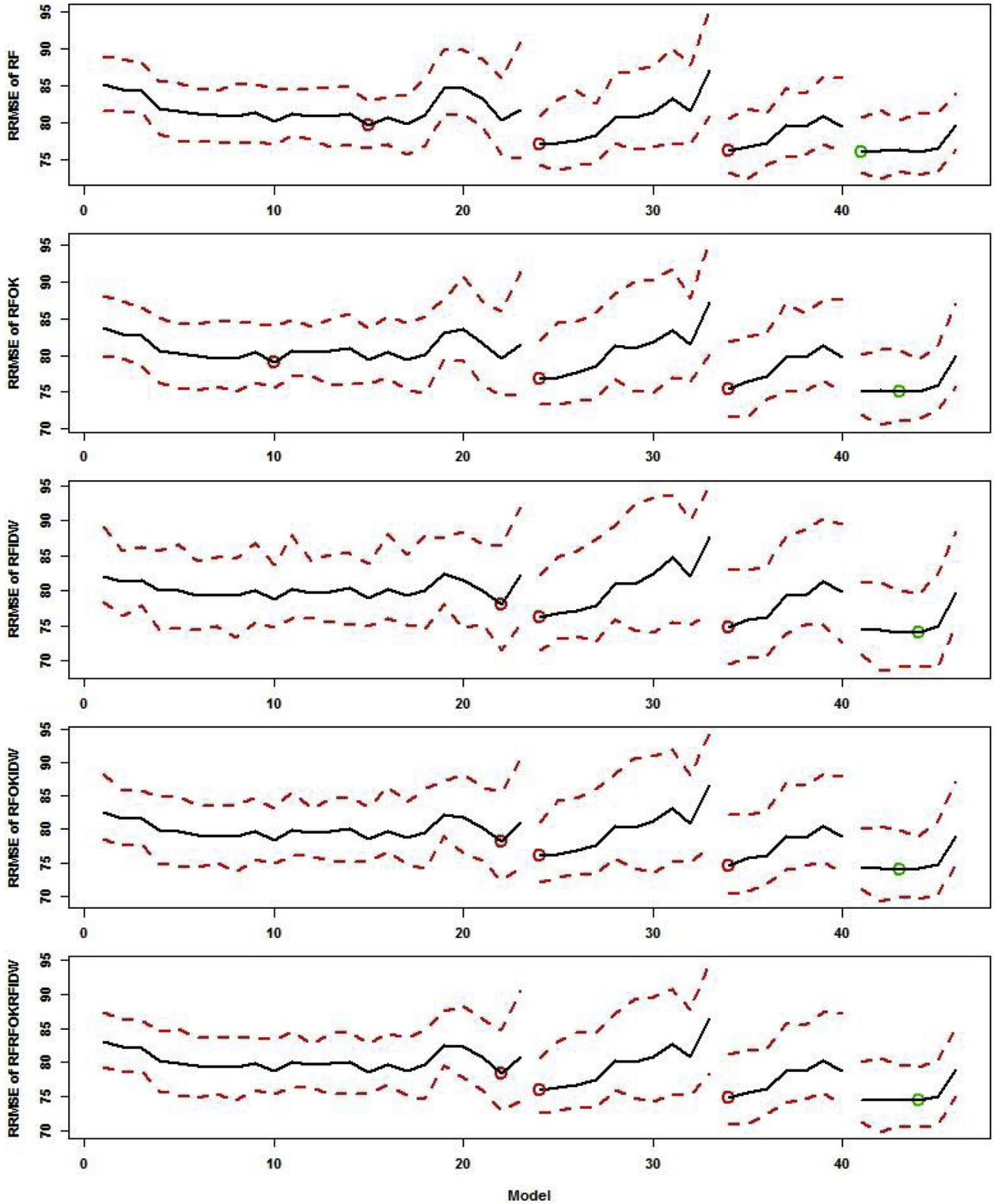


Fig. 2. RRMSE (mean: black line; minimum and maximum: dash red lines) of RF models 1–46 with different predictor sets based on the averages over 100 iterations of 10-fold cross validation for sponge species richness. The minimum mean RRMSE (red circle) for models in: a) models 1–23 (based on the AVI); b) models 24–33 (based on KIAVI); c) models 34–40 (based on KIAVI using model 24 and the AVI) and d) models 41–43 (based on KIAVI by removing the unimportant predictors identified from model 34–40) and then models 44–46 (based on KIAVI by using model 43 and the AVI). The green circle indicates the model with minimum mean RRMSE in models 1–46. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Mean predictive errors of the most accurate models among models 1–46 for each of RF, RFIDW, RFOK, RFOKRFIDW and RFRFOKRFIDW.

Method	MAE	Model	RMAE	Model	RMSE	Model	RRMSE	Model
RF	5.8728	26	56.0406	26	7.9575	41	75.9314	41
RFOK	5.8011	42	55.3486	42	7.8684	43	75.0777	43
RFIDW	5.7301	42	54.6756	42	7.7668	44	74.1069	44
RFOKRFIDW	5.7308	42	54.6814	42	7.7599	43	74.0425	43
RFRFOKRFIDW	5.7787	42	55.137	42	7.7989	44	74.4152	44

Table 7
Mean predictive errors of the most accurate models among models 1–3 for Boruta (with macRuns of 100 (default), 2000 and 5000) based on 49 variables for each of RF, RFIDW, RFOK, RFOKRFIDW and RFRFOKRFIDW.

Method	MAE	Model	RMAE	Model	RMSE	Model	RRMSE	Model
RF	6.2748	2	59.8748	2	8.6110	2	82.1617	2
RFOK	5.9563	2	56.8362	2	8.4197	2	80.3391	2
RFIDW	6.1194	2	58.3879	2	8.3900	2	80.0505	2
RFOKRFIDW	5.9919	2	57.1731	2	8.3382	2	79.5620	2
RFRFOKRFIDW	6.0403	2	57.6323	2	8.3956	2	80.1027	2

unimportant predictor, geom, was identified because its removal reduced the predictive error.

A further seven models (i.e., models 34–40) were developed for each of RF, RFIDW, RFOK, RFOKRFIDW and RFRFOKRFIDW based on the most accurate model identified so far (i.e., model 24) by excluding the unimportant variable (i.e. geom) (Fig. 2, Table 3). There was a general trend of increasing RRMSE from model 34 and to model 40. Model 34 had a minimum RRMSE, especially for RFOKRFIDW with a RRMSE of 74.60%. This modelling process also identified two unimportant predictors: bs_o and bathy_o. For model 40, bs11 was the least important variables and removal of bs11 would result in a model that is identical to model 32, so no further model simplification was pursued.

Three models (i.e., models 41–43) were developed for each of RF, RFIDW, RFOK, RFOKRFIDW and RFRFOKRFIDW based on the most accurate model identified so far (i.e., model 34) by excluding the unimportant variables identified above (Fig. 2, Table 3). RRMSE decreased from model 41 to model 43. Model 43 was of a minimum RRMSE, especially for RFOKRFIDW with a RRMSE of 74.04%.

Three models (i.e., models 44–46) were developed for each of RF, RFIDW, RFOK, RFOKRFIDW and RFRFOKRFIDW based on the variable selection approach of AVI using predictors in the most accurate model identified so far (i.e., model 43) (Fig. 2, Table 3). RRMSE increased from model 44 and to model 46 for RFOKRFIDW and reached a minimum mean for model 44 for RFRFOKRFIDW, but the RRMSE of these models are higher than that of model 43 for RFOKRFIDW. For model 46, bs34 is the least important variables and removal of bs34 would lead to a model identical to model 40, so no further model simplification was pursued.

In total, 46 models were developed for each of RF, RFIDW, RFOK, RFOKRFIDW and RFRFOKRFIDW. Of these models and methods, model 43 of RFOKRFIDW had the lowest RRMSE (Table 6).

3.1.3. Variable selection using boruta, rfe and VSURF

Three models (i.e., models 1–3) were developed for each of RF, RFIDW, RFOK, RFOKRFIDW and RFRFOKRFIDW based on the variable selection approach of Boruta using 49 predictors (Table 7). Seven, 15 and 16 variables were selected respectively for these models; and some important variables identified for model 43 like lat, long and tpi3 were selected but some such as bs11 and/or bs34 were missed out. Of these three models, model 2 was the most accurate for all five methods, and RFOKRFIDW was the most accurate method. Since all variables were selected using rfe, the resultant model was identical to model 1. Only three variables were selected using VSURF, and the resultant model was identical to model 22.

3.1.4. The transformation of the residuals of RF

The transformation of the residuals of RF for model 43 resulted in model 47 that further reduced the predictive error with a RRMSE of 73.69% (Table 8).

3.2. Predictive model using GLM, GLMIDW, GLMOK, GLMOKGLMIDW and GLMGLMOKGLMIDW

For each of GLM, GLMIDW, GLMOK, GLMOKGLMIDW and GLMGLMOKGLMIDW, 9 models were developed (Tables 4 and 9). RRMSE values of GLM1, GLM2 and GLM3, which were based on traditional model selection approaches for GLM, were exceptionally high. RRMSE values of GLM4, GLM5 and GLM6, which were based on both RF model and traditional model selection approaches for GLM, were much lower than that of models 1 to 3 and reached the minimum for model 6, especially for GLMGLMOKGLMIDW with a RRMSE of 85.54%. The last three models were based on the most accurate model so far for GLM (i.e. GLM6) and traditional model selection approaches for GLM (Table 5) and their RRMSE values were much higher than that of model 6. Overall, model 6 for GLMGLMOKGLMIDW was more accurate than all other models. This model contained five predictors (Table 4).

3.3. Comparison of the most accurate predictive methods and the effects of model averaging

The accuracy of the most accurate predictive models for SSR developed for GLM and RF using various model selection and variable selection methods (i.e. model 1 for RFIDW (RF1), model 43 for RFOKRFIDW (RF43), model 47 for RFOKRFIDW (RF47), model 2 for RFOKRFIDW using Boruta (RF.Boruta2) and model 6 for

Table 8
Mean predictive errors of model 47 for RF, RFIDW, RFOK, RFOKRFIDW and RFRFOKRFIDW.

Method	MAE	RMAE	RMSE	RRMSE
RF	5.9596	56.8624	7.9926	76.2642
RFOK	5.6974	54.3599	7.8397	74.8024
RFIDW	5.7410	54.7765	7.7645	74.0868
RFOKRFIDW	5.6855	54.2473	7.7233	73.6909
RFRFOKRFIDW	5.7475	54.8344	7.7776	74.2080

Table 9
The minimum, mean and maximum of RRMSE (%) for GLM models 1 to 9 for GLM, GLMIDW, GLMOK, GLMOKGLMIDW and GLMGLMOKGLMIDW. The most accurate model highlighted in bold.

models	GLM			GLMOK			GLMIDW			GLMOKGLMIDW			GLMGLMOKGLMIDW		
	min	mean	max	min	mean	max	min	mean	max	min	mean	max	min	mean	max
1	94.87	1.69E+09	1.68E+11	93.65	1.69E+09	1.68E+11	94.09	1.69E+09	1.68E+11	93.81	1.69E+09	1.68E+11	94.07	1.69E+09	1.68E+11
2	175.56	2.57E+07	1.79E+09	174.92	2.57E+07	1.79E+09	179.19	2.57E+07	1.79E+09	176.63	2.57E+07	1.79E+09	175.19	2.57E+07	1.79E+09
3	103.27	4.89E+05	2.59E+07	95.91	4.89E+05	2.59E+07	99.27	4.89E+05	2.59E+07	97.14	4.89E+05	2.59E+07	96.97	4.89E+05	2.59E+07
4	91.96	97.59	142.36	81.88	89.02	133.17	82.33	91	135.57	80.24	88.16	133.27	81.84	87.95	134.24
5	**91.09**	96.36	143.24	81.4	88.12	135.42	80.61	90.15	137.84	79.12	87.34	135.59	80.8	86.92	136.11
6	91.14	**93.72**	**100.16**	**80.59**	**87.95**	**99.62**	**78.59**	**89.21**	**97.24**	**77.38**	**86.45**	**96.39**	**79.1**	**85.54**	**94.37**
7	92.04	113.41	195.8	92.51	114.28	197.85	95.59	116.23	200.2	92.29	113.67	198.16	90.96	112.19	196.5
8	90.64	94.39	101.19	83	89.19	100.74	81.74	91.1	99.29	80.35	87.93	97.93	81.15	86.72	96.66
9	93.1	135.72	194.53	93.94	137.51	199.33	92.12	138.35	202.21	91.04	136.45	199.78	90.2	135.12	197.23

GLMGLMOKGLMIDW (GLM6)) was summarised in Table 10 and Fig. 3. The models developed for RF were significantly more accurate than the model for GLM based on the Mann-Whitney tests (with p values < 0.0001). Among the models for RF, the model (i.e. RF43) based on KIAVI was significantly more accurate than the full model (RF1) and the model based on Boruta (i.e. RF.b2) in terms of the Mann-Whitney tests (with p values < 0.0001). Although the transformation of RF residuals has slightly increased the accuracy (i.e. RF47), such improvement was marginal in comparison with the non-transformed model (i.e. RF43) in terms of the Mann-Whitney tests (with p value = 0.1818).

The effects of model averaging on the most accurate predictive models for RF (i.e. model 47) and GLM (i.e. model 6) were summarised in Table 11 and Fig. 4. It is apparent that RFOKRFIDW was more accurate than other methods including RFRFOKRFIDW, but the difference was not significant in comparison with RFIDW that was also not significantly different to RFRFOKRFIDW; additionally, hybrid methods (i.e. RFOK, RFIDW) significantly improved the accuracy in comparison with RF (Table 11). For GLM, the averaged methods were significantly more accurate than other methods and GLMGLMOKGLMIDW was also significantly more accurate than GLMOKGLMIDW; the hybrid methods (i.e. GLMOK, GLMIDW) significantly improved the accuracy in comparison with GLM (Table 11).

Table 10
Comparison of VEcv (%) of the most accurate predictive models for sponge species richness developed for GLM and RF using various model selection methods (i.e. model 1 for RFIDW (RF1), model 43 for RFOKRFIDW (RF43), model 47 for RFOKRFIDW (RF47), model 2 for RFOKRFIDW using Botura (RF.b2) and model 6 for GLMGLMOKGLMIDW (GLM6)) based on the averages over 100 iterations of 10-fold cross validation. The differences between these comparisons based on the Mann-Whitney tests (n = 100 for each model).

Model	VEcv (%)	RF1	RF43	RF47	RF.b2
RF1	32.20				
RF43	44.89	0.0000			
RF47	45.41	0.0000	0.1818		
RF.b2	36.38	0.0000	0.0000	0.0000	
GLM6	26.50	0.0000	0.0000	0.0000	0.0000

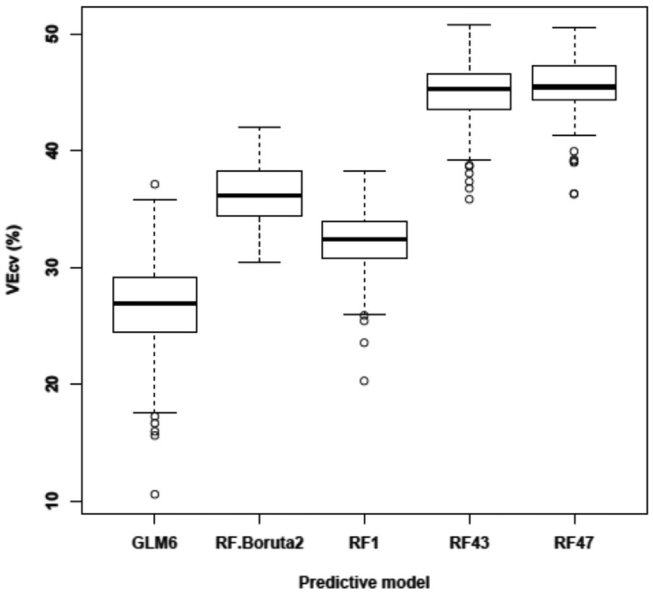


Fig. 3. The VEcv (%) of the most accurate predictive models based on the averages over 100 iterations of 10-fold cross validation for sponge species richness developed for GLM and RF using various model selection methods.

Table 11
Effects of model averaging on the most accurate predictive models for RF (i.e. model 47) and GLM (i.e. model 6) for sponge species richness in terms of VEcv (%) based on the averages over 100 iterations of 10-fold cross validation. The differences between these comparisons based on the Mann-Whitney tests (n = 100 for each model).

Model	RF	RFOK	RFIDW	RFOKRFIDW
RFOK	0.0000			
RFIDW	0.0000	0.0013		
RFOKRFIDW	0.0000	0.0000	0.1641	
RFRFOKRFIDW	0.0000	0.0000	0.4040	0.0000
Model	GLM	GLMOK	GLMIDW	GLMOKGLMIDW
GLMOK	0.0000			
GLMIDW	0.0000	0.0004		
GLMOKGLMIDW	0.0000	0.0000	0.0000	
GLMGLMOKGLMIDW	0.0000	0.0000	0.0000	0.0000

3.4. Goodness of fit, model selection criteria and VEcv

Goodness of fit of models 1 and 43 (or 47) for RF and models 1 to 9 for GLM were depicted in Fig. 5. The goodness of fit for two RF models were similar or the later one is even better (Fig. 5a), but the VEcv of models 43 and 47 were significantly higher than that of model 1 for RF (Table 10).

The goodness of fit of model 1 for GLM was the best and then became progressively worse from models 2–6; the goodness of fit of model 7 for GLM was slightly improved and then reduced for model 8 and slightly improved for model 9. However, the VEcv of models 1 to 9 for GLM displayed an opposite pattern in comparison with the goodness of fit. That is, the VEcv increased from model 1 to model 6, decreased for model 7, increased for model 8 and then

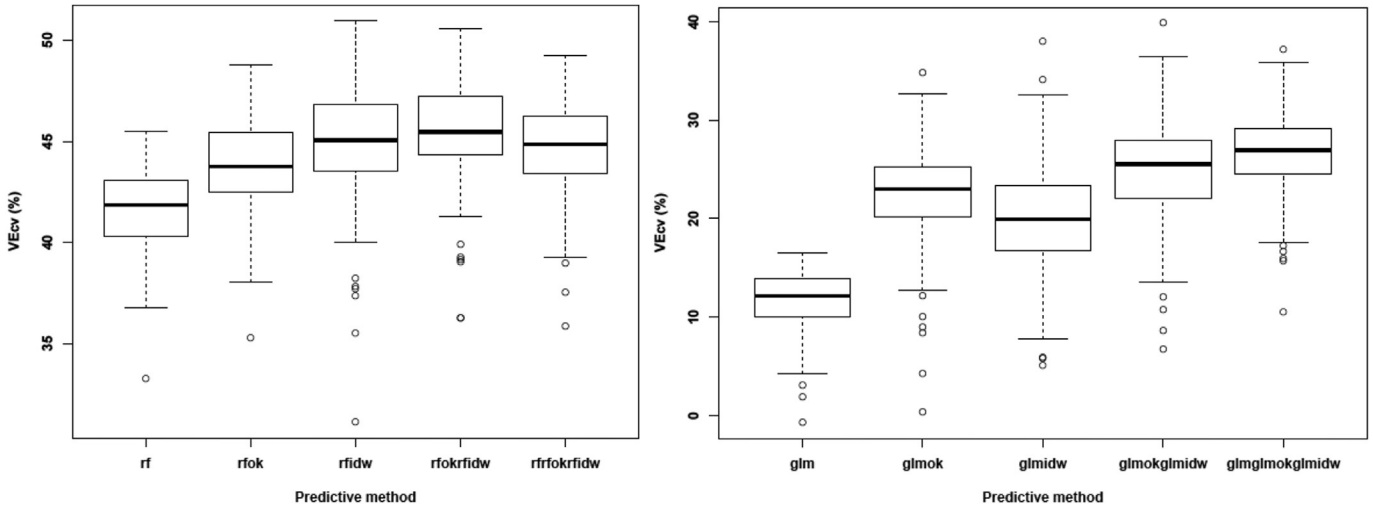
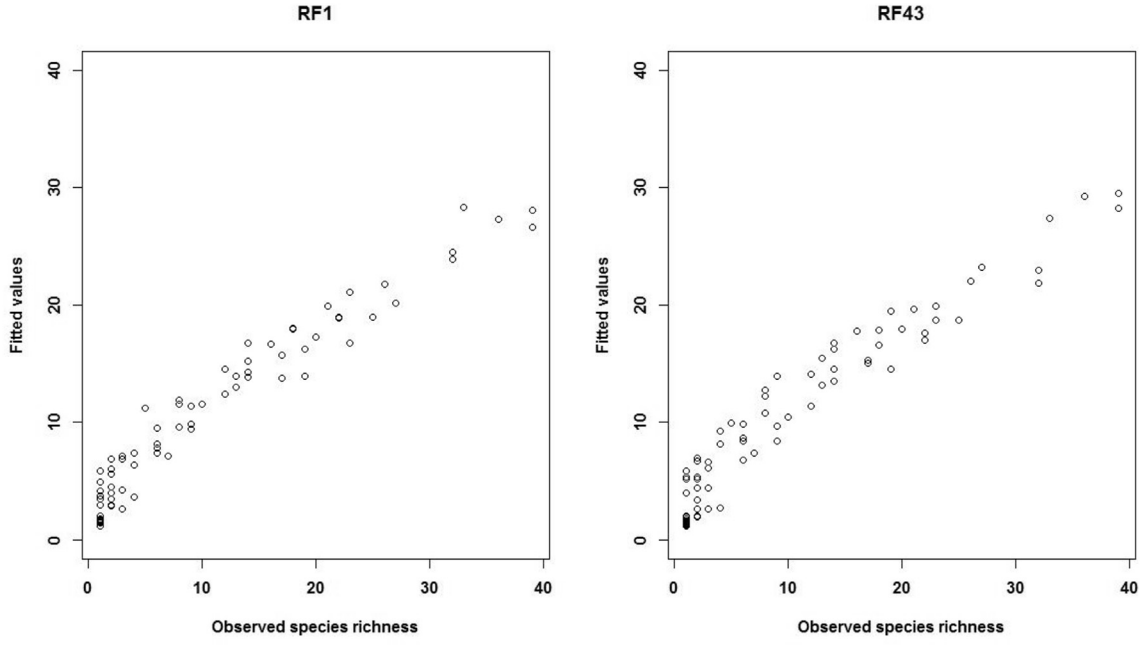
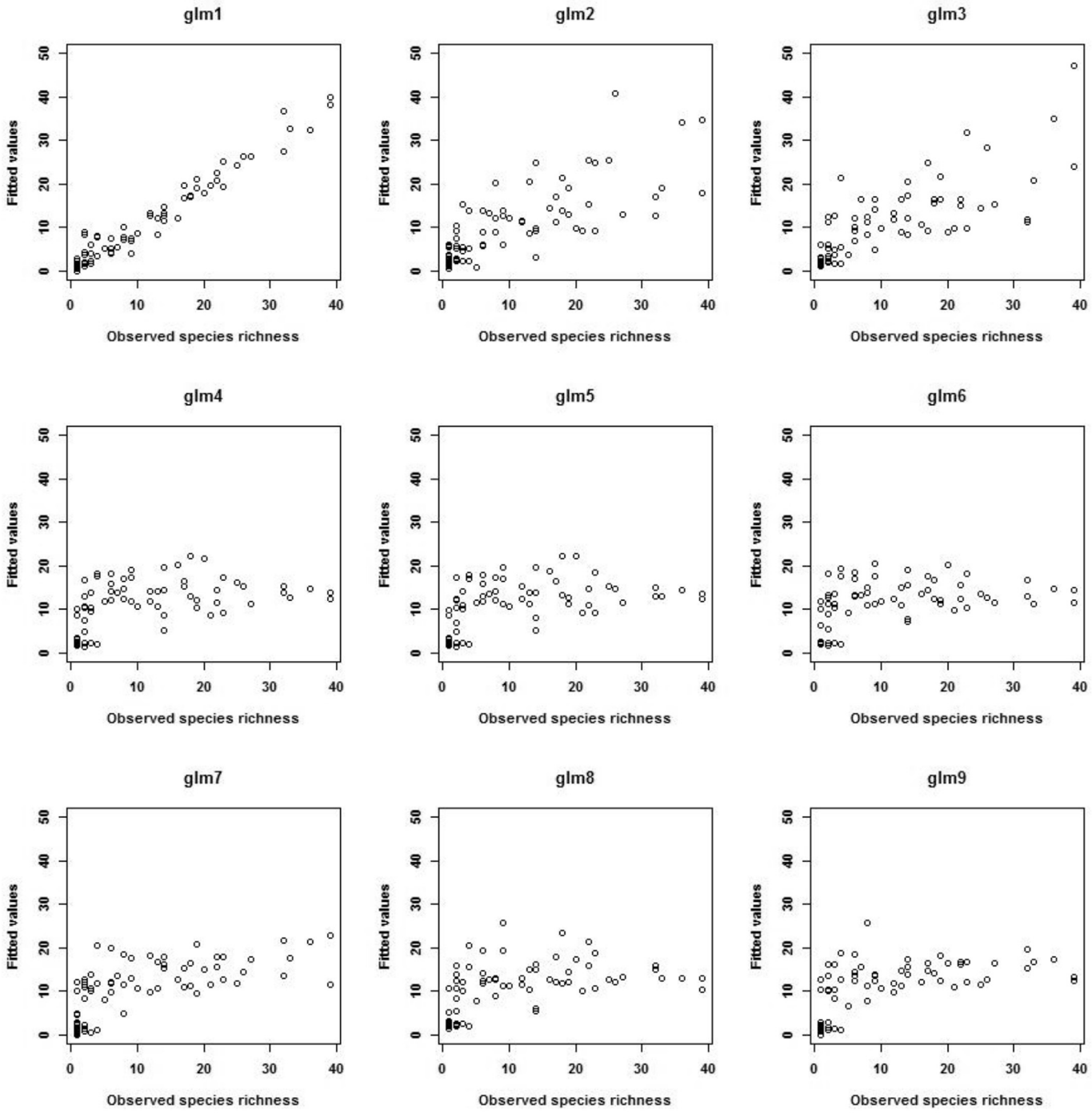


Fig. 4. The effects of model averaging on the most accurate predictive models for RF (i.e. model 47) and GLM (i.e. model 6) for sponge species richness in terms of VEcv (%) based on the averages over 100 iterations of 10-fold cross validation.



a

Fig. 5. The fitted values vs. observed values for sponge species richness: a) models 1 and 43 for RF, and b) models 1 to 9 for GLM.



b

Fig. 5. (*continued*).

decreased for model 9 (Table 12).

Moreover, it was apparent that the traditionally used BIC and deviance explained had little association with VEcv (Table 12). Model 1 for GLM had the lowest BIC, but with lowest VEcv (Table 12). Model 4 had the lowest deviance explained adjusted but its predictive accuracy (i.e. VEcv) was 22.3% that was not the highest. Model 6 was the most accurate with a VEcv of 26.5%, but with moderately high BIC and deviance explained adjusted.

3.5. The predictions of SSR

The predictions were generated using model 47 for RFOKRFIDW that was the most accurate predictive model. The influence of the eight predictors in model 47 for RFOKRFIDW on the predictions on the basis of their importance was as follows: long > lat > dist.coast > bs11 > tpi3 > bs34 > bs_entro7 > bs_var7. The relationships of SSR with these predictors are illustrated in Fig. 6 and it is apparent that these relationships were non-linear.

Table 12
The BIC, deviance explained (%), deviance explained adjusted (%) and VEcv (%) of GLM models 1 to 9.

Model	BIC	Deviance explained (%)	Deviance explained adjusted (%)	VEcv (%)
1	468.63	93.16	88.45	−2.86E+16
2	622.86	62.91	55.26	−6.65E+12
3	587.78	66.35	60.65	−2.40E+09
4	779.10	39.75	32.66	22.30
5	776.15	39.56	33.43	24.11
6	783.27	37.51	33.11	26.50
7	702.79	51.38	43.15	−26.43
8	781.94	37.68	33.29	24.46
9	722.58	47.10	40.88	−83.40

For example, SSR was relatively high along longitude and started to decline after 129.5° with a dramatic decrease around 129.8°; and an opposite pattern was observed in relation to tpi3. SSR increased with bs34 and bs11, reached a plateau when bs34 was higher than −24 db, and decreased when bs11 was higher than −19 db and reached a plateau when bs11 was close to −10 db, which was equivalent to what was observed when bs11 was −21 db.

The predicted SSR was found to be high on banks and terraces and low on plains and valleys. The predictions are illustrated in Fig. 7 for area H in Oceanic Shoals. This area was chosen as an example as it contains highly contrasting geomorphic features. In this area, the influence of the first three important variables (i.e. long, lat and dist.coast) were largely unnoticeable as the area was small and only covered a short range of these variables (i.e. 126.9364° – 127.0022° for long, −11.3457° – −11.3126° for lat, and 264,947–268,605 m for dist.coast) and their influence was barely noticeable over the short range (Fig. 6). The patterns of the predictions mainly reflected the local variations associated with the rest predictors as: (1) the highest SSR was corresponding to the high values of bs11 and bs34 on the bank (Fig. 7b and f), (2) the ring patterns were associated with high SSR and mimicked the patterns of tpi3 and also bs11 and bs34 (Fig. 7b and c), and 3) the blue horizontal broken strips were associated with low SSR and reflected the influence of bs11, bs34, bs_entro7 and bs_var7 (Fig. 7b, d and 7e). In general, low species richness was mostly found in the plains and depressions with low bs values (Fig. 7f).

4. Discussion

4.1. Issues with model valuation and selection criteria for RF

The set of initial input predictors may affect the final model selection. From models 24–33 based on RF, only the exclusion of geom increased the predictive accuracy. That is, with the existence of geom, a few important variables were excluded during the modelling process, and the accuracy of all corresponding models was reduced. This suggests that the set of initial input predictors affects the model selected as previously observed (Li, 2013a).

The status of important and unimportant variables may change with the sets of initial input predictors. Two predictors, bathy_o and bs_o, were identified as unimportant for models 40 and 41, but were identified as important variables for models 1–23. This change suggests that unimportant variables can be identified as important variables if some unimportant variables (i.e., prof_curv_o, rugosity3 and tpi3) exist. This phenomenon was also found in previous studies (Li, 2013a; Li et al., 2016). It was also found that the inclusion of noisy or irrelevant predictors may reduce the possibility of the selection of the important variables at each node split for each individual tree and thus reduce the predictive accuracy (Li et al., 2011a, 2011b). This suggests that pre-selection of predictors for RF is important for predictive modelling, although it was argued that RF can deal with noisy predictors well (Diaz-Uriarte and de Andres, 2006). This further suggests that AVI is not always reliable for selecting predictors or simplifying predictive models. This presents a challenge for selecting an optimal predictive model. Repeating the selection procedure based on important and unimportant variables by using KIAVI may help to resolve this issue although it is time consuming. This is an area worth further investigation in the future.

Although features selected using Boruta can improve the accuracy in comparison with the full model, the accuracy of resultant model is significantly less than that of the most accurate model (i.e., model 47 for RFOKRFIDW). This finding is consistent with previous findings (Li et al., 2016). The accuracy of the model selected using Boruta highly depends on the choice of number of runs, i.e. maxRuns. This may further suggest that for some datasets, more runs are required. However, this may still lead to sub-optimal model. Moreover, some features in the most accurate predictive model were not included in the features selected using Boruta. That is, some important features in terms of predictive accuracy were missed out during its selection. This suggests that Boruta should be used with caution in selecting features for predictive models, which is against previous recommendation (Li et al., 2016).

Features selected from rfe were not optimal and can even be misleading. This may be due to the fact that we do not have causal predictors in this study, which is often the case in the environmental sciences (Li, 2013b). It can be further argued that this may be also due to the small number of predictors used in this study and that rfe may be best used with a large number of predictors. These findings regarding rfe may be limited to this study. This selection method may be useful when the number of predictors or the dimension of feature space gets large. This is largely speculative and further studies are needed.

Features selected using VSURF were too parsimonious in comparison with other selection methods, but the selected features were the important ones. Thus this method could be used to identify a few important features but it can lead to sub-optimal predictive models. Furthermore, we would argue that any feature selection methods based on variable importance only would lead to sub-optimal predictive models. This is because the resultant models from such selection methods are not based on their predictive accuracy. Therefore, caution should be taken when using them to select predictive models.

A complete search for the global optimal model(s) would identify the most accurate predictive model for given samples but it is time consuming and becomes impossible when the number of predictors is large. This is because the computational requirements have a factorial increase with the number (Li et al., 2016), highlighting the importance of variable selection methods as previously discussed.

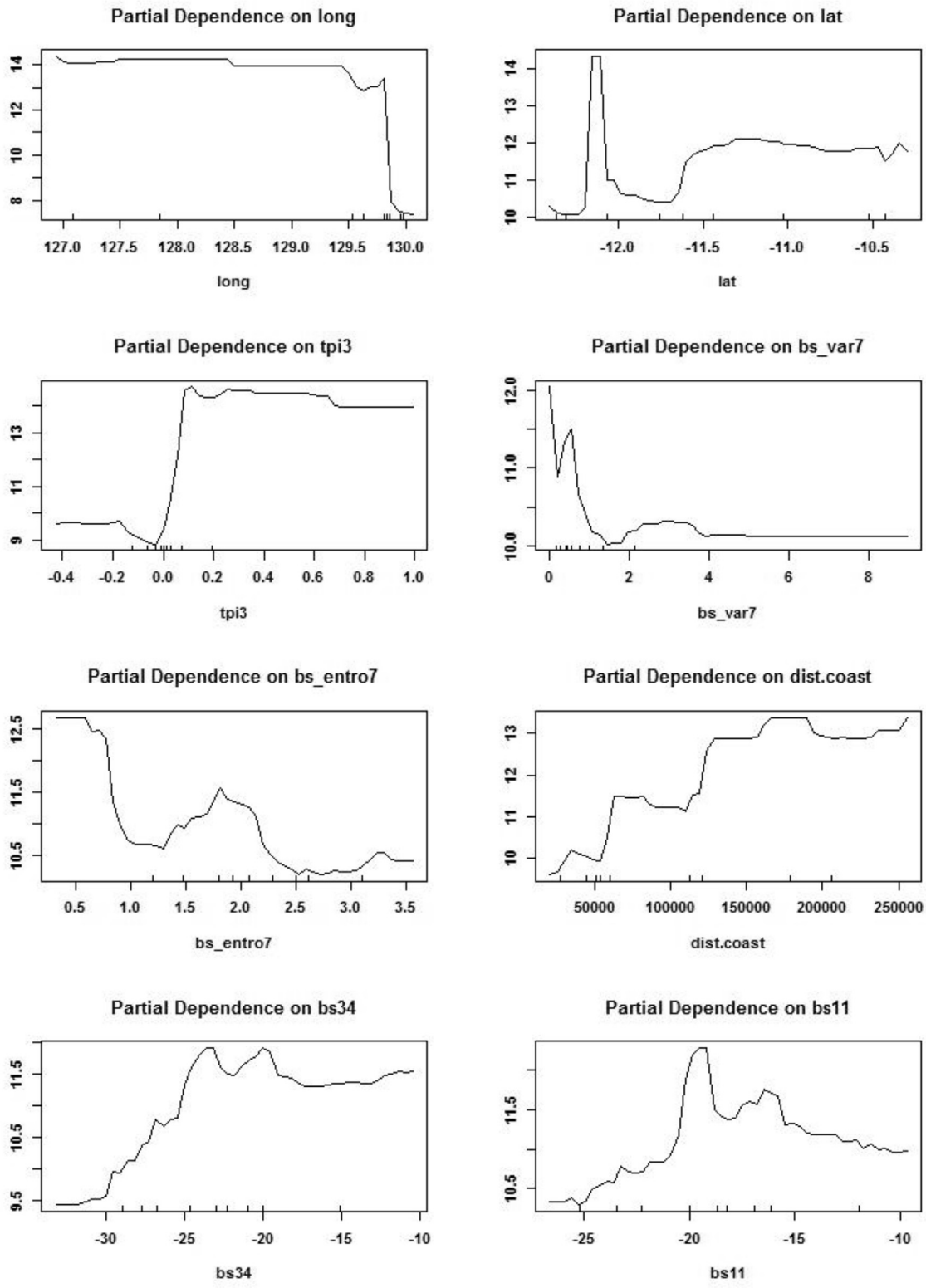


Fig. 6. Partial plot of RF model 47, indicating the relationships of sponge species richness to the eight predictors in the RF model.

In addition, the inclusion of highly correlated predictors (i.e. $\rho = 0.95$ and $r = 0.95$ for bs11 and bs34) could improve predictive accuracy. This is consistent with findings regarding the applications of RF in other studies (Li, 2013b; Li et al., 2012b, 2013, 2016). This implies that correlated variables may be able to compensate for the small number of predictors in environmental sciences. This provides important guidelines for pre-selecting predictors using correlation methods because in environmental sciences we usually only have correlated proxy predictive variables instead of causal predictors or drivers as seen in simulation studies (Biau, 2012; Li, 2013b).

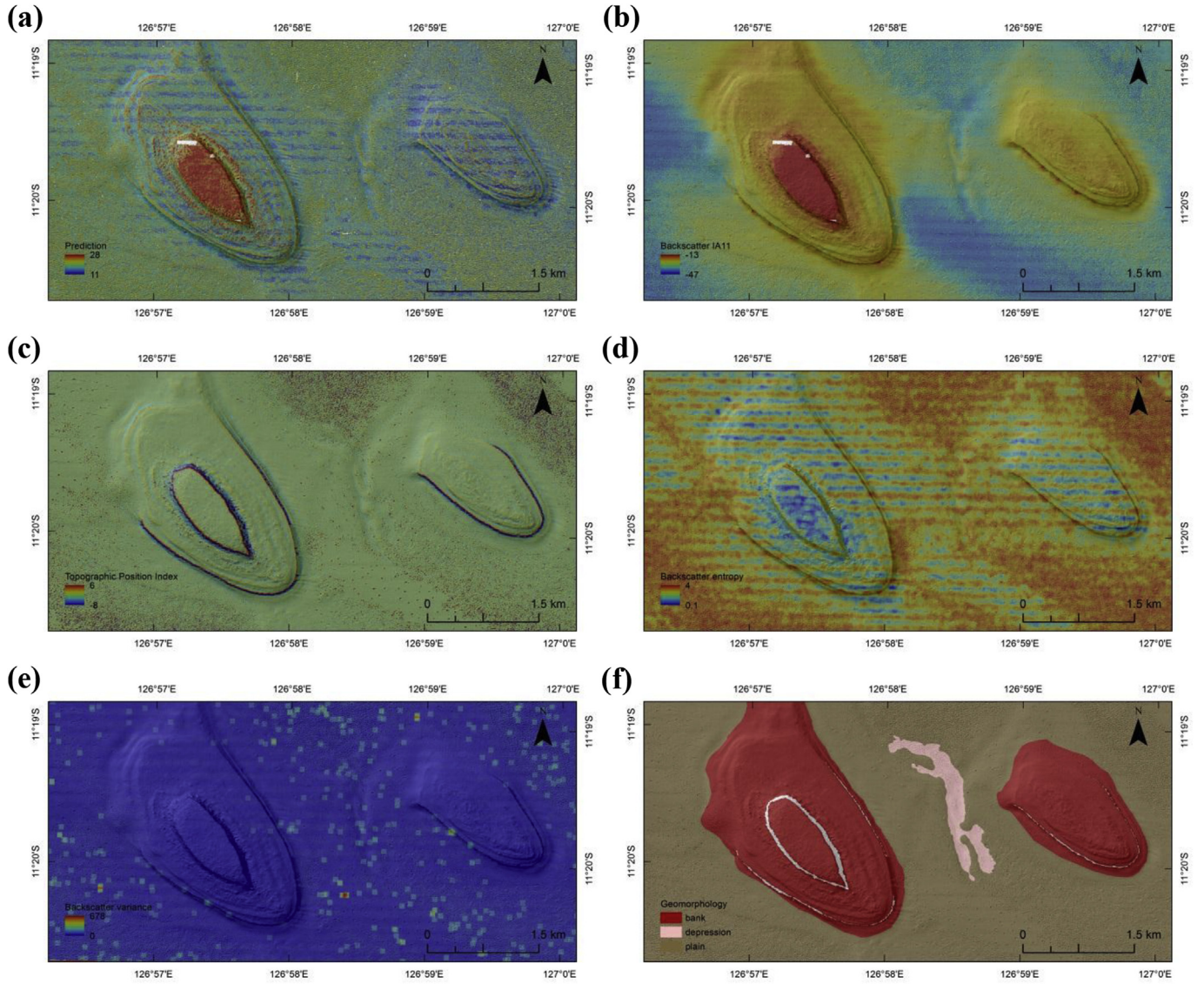


Fig. 7. Spatial predictions of sponge species richness using model 47 for RFOKRFIDW in Oceanic Shoals area 4: a) predictions, b) bs11, c) tpi3, d) bs_entro7, e) bs_var7 and f) geomorphic features overlaid on bathymetry. Since the spatial patterns of bs34 were similar to that of bs11, it was not presented. The blank spaces in the predictions were resulted from the missing values in bs11.

4.2. Issues with model valuation and selection criteria for GLM

The accuracy (i.e. *VEcv*) of GLM predictive model did not align well with BIC, deviance explained (%), and deviance explained adjusted (%). The results suggest that conventional model selection approaches based on AIC, anova and dropterm and their combination for GLM may lead to models with the lowest BIC values, highest deviance explained (%) and highest deviance explained adjusted (%). These models may be the most parsimonious models, but do not necessarily have the highest predictive accuracy. In fact, the accuracies of the resultant models (i.e. GLM1, GLM2 and GLM3) were the most inaccurate and unacceptably low in this study. This finding suggests that using conventional model selection approaches was unable to identify reliable predictive models and they should be used with care for developing predictive models, although they are useful approaches for inferential or exploratory analyses (Leek and Peng, 2015). Overall, this finding highlights that selecting predictive models is highly challenging and that to select predictive models, predictive accuracy should be used instead of AIC or deviance explained (%) because they are misleading and should not be used to select predictive models. This finding confirms that the traditional model selection methods such as AIC and BIC for regression models (e.g. linear model, generalised linear model) attempt to select the most parsimonious models that are not necessarily the most accurate models, especially when proxy variables are used as predictors instead of causal variables (Li et al., 2016).

The most accurate GLM predictive model was built from the predictors of the most accurate RF model, although with a further simplification based on AIC. This finding suggests that the information from RF model is helpful and joint application of variable selection using RF and conventional model selection approaches for GLM can further improve the predictive accuracy of GLM models. This finding is consistent with previous studies (Arthur et al., 2010) where RF was used to select important predictors for GLM and a generalised linear mixed model. This is proven to be a useful model selection approach for developing GLM predictive models.

The assessments of goodness of fit based on the observed values

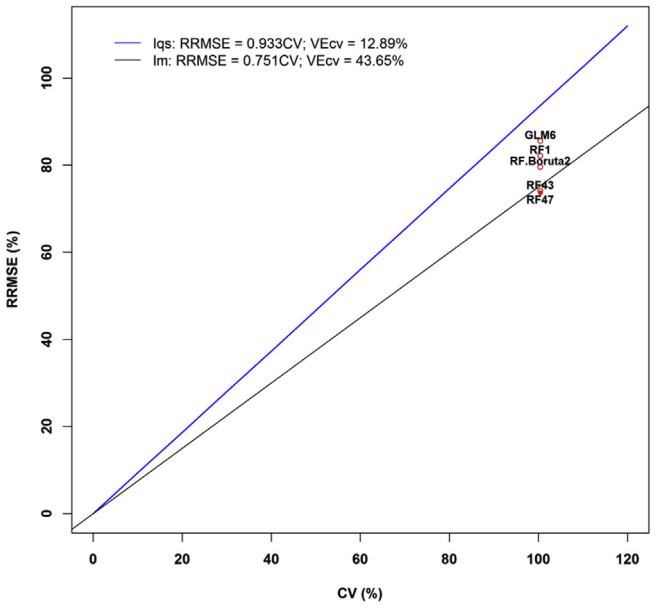


Fig. 8. The relative root mean square error (RRMSE) of the most accurate predictive models identified (i.e., GLM6, RF1, RF.Boruta2, RF43 and RF47: red circles) using various model and variable selection methods and the average accuracy of predictive models published in the environmental sciences (i.e., linear model (lm): black line, and resistant regression (lqs): blue line) (modified from Li, 2016) in relation to coefficient variation (CV (%)). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and fitted values of a model are useful for statistical analysis like inferential or exploratory (Leek and Peng, 2015). However, they are unreliable for identifying predictive models as demonstrated by the relationship of observed species richness and the fitted values (Fig. 5b) and by the corresponding predictive accuracy. How the goodness of fit relates to predictive accuracy is worthy of further investigation. In contrast to GLM, the most accurate RF model with 8 predictors not only improved the predictive accuracy, but also delivered a goodness of fit that was as good as the RF model with all 49 predictors. This may be a further advantage of RF and its hybrid methods.

4.3. Model accuracy

Model averaging significantly improved the accuracy in comparison with RFOK but only marginally improved the accuracy in comparison with RFIDW. Such disparate effects of model averaging on the accuracy improvement have been previously observed as it improved the accuracy (Li et al., 2012b), but showed little effect (Li et al., 2011a, 2011b) or even negative effects (Li, 2013a). The possible reasons for such marginal or negative effects by model averaging have been discussed in depth (Li, 2013a). In comparison to RF, model averaging have significantly improved the accuracy for GLM and its hybrids, which is consistent with findings for other modelling methods (Goswami and O'Connor, 2007; Marmion et al., 2009; Raftery et al., 2005). These findings may imply that the effects of model averaging are method dependent or even data dependent.

It is apparent that the hybrid methods can significantly improve predictive accuracy in comparison with RF (Table 11). This is supported by many previous studies (Li, 2013b; Li et al., 2011c; Sanabria et al., 2013b), although the opposite was observed for RFIDW (Li, 2013a). As for GLM, it was the first time such comparisons have been made between GLM with GLMOK and GLMIDW. The latter is a newly developed hybrid method in this study. Although GLMOK was compared with other methods in previous studies (Li et al., 2011b, 2011c, 2010b), this is the first time it was applied to count data. Evidently the hybrid methods have significantly improved the accuracy in comparison with the GLM model. Furthermore, the hybrid methods of RF and geostatistical methods are considerably more accurate than the hybrid methods of GLM and geostatistical methods. This finding confirms that hybrid methods of RF and geostatistical methods can effectively model count data and are not data-type specific, further demonstrating the potential of the methods for making spatial predictions. Although the hybrid methods were developed and applied to marine sediment data since 2008 with proven high predictive accuracy (Li et al., 2010b, 2011a, b, 2012b), their applications to other data type or terrestrial data are still rare (Sanabria et al., 2013b; Tadić et al., 2015). These methods are recommended for further testing using relevant data types for spatial predictions in the future.

The prediction accuracy (VEcv) of the most accurate model (i.e. model 47 for RFOKRFIDW) is 45.41% and its RRMSE is 73.69%, which is higher than the average accuracy of predictive models published in the environmental sciences (Fig. 8) (Li, 2016; Li and Heap, 2008; Li et al., 2012b). It is worth mentioning that the predictors used in this study are proxies, but they are usually causal variables or drivers in simulated studies (Biau, 2012) and most likely causal variables or drivers for studies in the terrestrial environmental sciences (Austin et al., 2006; Sanabria et al., 2013b). This further demonstrates the capacity of the hybrid methods. The high performance of the hybrid methods could be attributed to features of RF (Li, 2013a; Li et al., 2011b; Li et al., 2011c) and the ability to deal with local variation by geostatistical component in the hybrid methods. This demonstrates an advantage of the hybrid methods. On one hand, they can effectively deal with the global trend either spatially, environmentally or both and with non-linear relationships with predictors, and on the other hand, they can deal with local variations if the residuals contain useful information of local variation.

Finally, the predictive models using proxy predictors may provide useful clues for identifying causal variables. Using proxy predictors in predictive models developed using RF can lead to highly accurate predictive models. These models however are often referred to as 'black box' in the respect that they are unable to directly inform how the dependent variable is related to causal variables or drivers. The model depicts the relationship of the dependent variable with the proxy predictors as shown in this study, which may be able to shed some light on where we should look for the possible causal variables. This is because the predictors that remained in the model may be good surrogates of the causal variables, which may help to narrow the scope where we can start to find the casual variables and provide insight about the dependent variable on how and why it displayed the observed patterns. Professional knowledge can then play a significant role in searching for the causal variables based on the proxy predictors used in the predictive model. This can then lead to the discovery of reliable information for management. For example, we may use the identified relationships of SSR with the proxy predictors in this study to narrow the scope of possible causal variables based on expert knowledge if we can explain ecologically why such a relationship was observed. Therefore, an accurate predictive model can not only produce reliable spatial predictions, but also provide clues for identifying causal variables.

4.4. Predictions of species richness

SSR was generally high in the region west of 129.8° and north of −12.7 and further from coastline (Fig. 7), supporting previous

research in which sponge diversity and community structure were linked to distance offshore (Sorokin et al., 2007; Wilkinson and Cheshire, 1989). The SSR was generally higher when bs variables were higher although non-linearly. The higher bs is, the harder the seabed substrates. This implies that hard seabed substrates provide habitats for sponge species and support high SSR. This finding is supported by previous studies (Beaman et al., 2005; Fromont et al., 2012) as well as by those conducted in the same region (Przeslawski et al., 2014, 2015). The relationships of SSR with the predictors are non-linear in this study, which explains why no (Przeslawski et al., 2015) or only marginally significant relationship (Przeslawski et al., 2014) could be found when a linear relationship was assumed. The nonlinear relationships were expected because many factors contribute to such relationships such as interactions of limiting resources and competition (Austin, 1987; Austin et al., 2006; Huston, 2002). The predictions also depend on the length of environmental gradient captured by samples (Li et al., 2009), which explains why GLM models performed poorly in this study. This can be further explained as that samples used as training for cross-validation may cover a short gradient and thus lead to abnormally high or low predictions for validation samples. The relationship of SSR with latitude was also apparent for many species including sponge in terms of their abundance in a previous study (Smale et al., 2010). In addition, the bathymetry was found not an important predictor of SSR, which is consistent with that Australian sponge communities have shown no relationship to water depth (Schönberg and Fromont, 2011), although not consistent with previous observations by Wilkinson and Cheshire (1989) and Sorokin et al. (2007). These findings largely delineate the region where habitats of sponge species are likely to be found, providing important information for future field validation and monitoring design and highlighting areas where management and conservation of sponge gardens should be focused. Furthermore, our findings should be considered in conjunction with results of species assemblages or functional groups (Cadotte et al., 2011), as well as temporal variability (Piacenza et al., 2015), to provide reliable information for the management and conservation of sponge gardens.

4.5. Limitations

Many modelling strategies and predictive approaches for spatial predictions of species richness have been reviewed (D'Amen et al., 2015). Of these, only the assembly first and predict later strategy and correlative macro-ecological models are applicable to this study. The spatial predictions generated in this study were based on modelling techniques using proxy environmental predictors not necessarily based on solid biological foundations. That is, these predictors were proxies that were not necessarily the causal variables or drivers. Despite this, the modelling approaches could be largely regarded as correlative macro-ecological models and thus may share all limitations associated with such models as detailed in D'Amen et al. (2015). These limitations may include the difficulty of inferring causality from observed patterns, the loss of species' identities, and the problematic assumption of stationarity through space and time. All these limitations should be considered in forming decisions for the management and conservation of the sponge gardens modelled in this study.

In addition, the randomForest function is known for its variable importance measure exhibiting bias towards correlated variables, continuous variables, and variables with many categories (Strobl et al., 2007, 2008). Of these issues, the correlation issue is the major concern because some of the 49 numerical variables used are highly correlated. The 'extendedForest' (Smith et al., 2011), which was developed based on Strobl et al. (2008), was used to address the correlation related issue. The only categorical predictor was considered in the middle of the modelling process and was not selected in the final predictive model. Furthermore, the final selection of a predictor was based on its contribution to predictive accuracy instead of its variable importance. So the effects of any bias in variable importance caused by relevant issues on the selection of final predictive model should be minimal in this study.

5. Conclusions

This is the first application of the hybrid methods of RF with OK and IDW, and GLM with OK and IDW and their averaged methods to count data. Initial input predictors should be pre-selected to minimise their impact on model and variable selection and on the status of important and unimportant variables in future studies. Joint application of KIAVI and cross-validation, where the features were selected based on variable importance and more importantly on the predictive accuracy of the resultant predictive model, is recommended for selecting RF predictive models in the future. Selecting an optimal RF predictive model is challenging and worthy of further investigation. The conventional model selection approaches based on both AIC, anova and dropterm and their combination should be used with care for developing GLM predictive models; AIC or deviance explained (%) should not be used to select GLM predictive models. Joint application of variable selection using RF and conventional model selection approaches for GLM is proven to be useful for selecting GLM predictive models. Criteria for assessing the goodness of fit are unreliable for selecting predictive models and should not be used to select GLM predictive models. The effects of model averaging are method dependent or even data-dependent. The hybrid methods of RF and geostatistical methods can effectively model count data and are not data-type specific; and they can effectively deal with the global trend either spatially, environmentally or both and with non-linear relationships with predictors, and with local variations if the residuals contain useful information of local variation. They are recommended for further testing using relevant data types for spatial predictions in the future. Moreover, an accurate predictive model based on proxy predictors can not only produce reliable spatial predictions, but also provide clues for identifying causal variables. Finally, the relationships of SSR with the predictors are non-linear and the habitats of sponge species are delineated for future monitoring design, management and conservation of sponge gardens in the study region.

Acknowledgements

This work was undertaken for the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Research Program (NERP). Funding for samples collected in 2009 and 2010 was provided through the Australian Government's Offshore Energy Security Program. 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 surveys and the Field and Engineering Support staff at GA. We also thank Johnathan Kool, Wenping Jiang and two anonymous reviewers for their valuable comments and suggestions. This paper is published with the permission of the Chief Executive Officer, Geoscience Australia.

Appendix A-D. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2017.07.016.

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.

Arthur, A.D., Li, J., Henry, S., Cunningham, S.A., 2010. Influence of woody vegetation on pollinator densities in oilseed *Brassica* fields in an Australian temperate landscape. Basic Appl. Ecol. 11, 406–414.

Austin, M.P., 1987. Models for the analysis of species' response to environmental gradients. Vegetatio 69, 35–45.

Austin, M.P., Belbin, L., Meyers, J.A., Doherty, M.D., Luoto, M., 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. Ecol. Model. 199, 197–216.

Beaman, R.J., Daniell, J.J., Harris, P.T., 2005. Geology-benthos relationships on a temperate rocky bank, eastern Bass Strait, Australia. Mar. Freshw. Res. 56, 943–958.

Biau, G., 2012. Analysis of a random forest method. J. Mach. Learn. Res. 13, 1063–1095.

Cadotte, M.W., Carscadden, K., Mirotchnick, N., 2011. Beyond species: functional diversity and the maintenance of ecological processes and services. J. Appl. Ecol. 48 (5), 1079–1087.

Commonwealth of Australia, 2012a. Marine bioregional Plan for the North-west Marine Region. Department of Sustainability, Environment, Water, Population and Communities, p. 260.

Commonwealth of Australia, 2012b. Marine bioregional Plan for the North Marine Region. Department of Sustainability, Environment, Water, Population and Communities, Canberra, p. 191.

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.

D'Amen, M., Rahbek, C., Zimmermann, N.E., Guisan, A., 2015. Spatial predictions at the community level: from current approaches to future frameworks. Biol. Rev. 92, 169–187.

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

Fromont, J., Althaus, F., McEnnulty, F.R., Williams, A., Salotti, M., Gomez, O., Gowlett-Holmes, K., 2012. Living on the edge: the sponge fauna of Australia's southwestern and northwestern deep continental margin. Hydrobiologia 687 (1), 127–142.

Genuer, R., Poggi, J.M., Tuleau-Malot, C., 2015. VSURF: Variable Selection Using Random Forests. R package version 1.0.2 ed.

Goswami, M., O'Connor, K.M., 2007. Real-time flow forecasting in the absence of quantitative precipitation forecasts: a multi-model approach. J. Hydrology 334, 125–140.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second 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.

Hengl, T., 2007. A Practical Guide to Geostatistical Mapping of Environmental Variables. Office for Official Publication of the European Communities, Luxembourg, p. 143.

Hijmans, R.J., 2014. Raster: Geographic Data Analysis and Modeling. http://CRAN.R-project.org/package=raster.

Hooper, J.N.A., 1996. Revision of microcionidae (Porifera: poecilosclerida: Demospongiae), with description of Australian species. Memoirs Qld. Mus. 40, 1–626.

Hooper, J.N.A., Van Soest, R.W.M., 2002. Systema Porifera: a Guide to the Supraspecific Classification of the Phylum Porifera. Kluwer Academic/Plenum Publishers, New York.

Huang, Z., Brooke, B., Li, J., 2011. Performance of predictive models in marine benthic environments based on predictions of sponge distribution on the Australian continental shelf. Ecol. Inf. 6, 205–216.

Huston, M.A., 2002. Introductory essay: critical issues for improving predictions. In: Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A., Samson, F.B. (Eds.), Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Washington, pp. 7–24.

Kohavi, R., 1995. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection, International Joint Conference on Artificial Intelligence (IJCAI). Morgan Kaufmann, pp. 1137–1143.

Kuhn, M., 2014. Caret: Classification and Regression Training. R package version 6.0-30. http://CRAN.R-project.org/package=caret.

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta package. J. Stat. Softw. 36 (11), 1–13.

Leek, J.T., Peng, R.D., 2015. What is the question? Science 347, 1314–1315.

Li, J., 2011. Novel spatial interpolation methods for environmental properties: using point samples of mud content as an example. Surv. Statistician Newsl. Int. Assoc. Surv. Statisticians No. 63, 15–16.

Li, J., 2013a. Predicting the Spatial Distribution of Seabed Gravel Content Using Random Forest, Spatial Interpolation Methods and Their Hybrid Methods, pp. 394–400. 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., 2016. Assessing spatial predictive models in the environmental sciences: accuracy measures, data variation and variance explained. Environ. Model. Softw. 80, 1–8.

Li, J., Heap, A., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists. Geoscience Australia. Record 2008/23, 137pp.

Li, J., Heap, A., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. Ecol. Inf. 6, 228–241.

Li, J., Heap, A., Potter, A., Daniell, J., 2010a. Can Machine Learning Methods Be Applied for Spatial Predictions of Environmental Properties? (Australian Statistical Conference: Perth).

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., 2014. Spatial interpolation methods applied in the environmental sciences: a review. Environ. Model. Softw. 53, 173–189.

Li, J., Heap, A.D., Potter, A., Daniell, J., 2011b. Application of machine learning methods to spatial interpolation of environmental variables. Environ. Model. Softw. 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. Cont. Shelf Res. 31, 1365–1376.

Li, J., Hilbert, D.W., Parker, T., Williams, S., 2009. How do species respond to climate change along an elevation gradient? A case study of the Grey-headed Robin (*Heteromyias albispecularis*). Glob. Change Biol. 15, 255–267.

Li, J., Potter, A., Huang, Z., Daniell, J.J., Heap, A., 2010b. 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., 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.

Li, J., Tran, M., Siwabessy, J., 2016. Selecting optimal random forest predictive models: a case study on predicting the spatial distribution of seabed hardness. PLoS One 11 (2), e0149089.

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

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K., Thuiller, W., 2009. Evaluation of consensus methods in predictive species distribution modelling. Divers. Distributions 15, 59–69.

Nichol, S., Howard, F., Kool, J., Stowar, M., Bouchet, P., Radke, L., Siwabessy, J., Przeslawski, R., Picard, K., Alvarez de Glasby, B., Colquhoun, J., Letessier, T., Heyward, A., 2013. Oceanic Shoals Commonwealth Marine Reserve (Timor Sea) Biodiveristy Survey. GA0339/SOL5650 Post-Survey Report. Geoscience Australia: Canberra.

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.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosciences 30, 683–691.

Piacenza, S.E., Thurman, L.L., Barner, A.K., Benkwitt, C.E., Boersma, K.S., Cerny-Chipman, E.B., Ingeman, K.E., Kindinger, T.L., Lindsley, A.J., Nelson, J., Reimer, J.N., Rowe, J.C., Shen, C., Thompson, K.A., Heppell, S.S., 2015. Evaluating temporal consistency in marine biodiversity hotspots. PLoS One 10 (7), e0133301.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.

Przeslawski, R., Alvarez, B., Battershill, C., Smith, T., 2014. Sponge biodiversity and ecology of the van diemen Rise and eastern Joseph Bonaparte gulf, northern Australia. Hydrobiologia 730 (1), 1–16.

Przeslawski, R., Alvarez, B., Kool, J., Bridge, T., Caley, J., Nichol, S., 2015. Implications of sponge biodiversity patterns for the management of a marine reserve in northern Australia. PLoS One 10 (11), e0141813.

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, 2013. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Radke, L., Nicholas, T., Thompson, P., Li, J., Raes, E., Carey, M., Atkinson, I., Huang, Z.,

Trafford, J., Nichol, S., 2017. Baseline biogeochemical data from Australia's continental margin links seabed sediments to water column characteristics. Mar. Freshw. Res. http://dx.doi.org/10.1071/MF16219.
Radke, L.C., Li, J., Douglas, G., Przeslawski, R., Nichol, S., Siwabessy, J., Huang, Z., Trafford, J., Watson, T., Whiteway, T., 2015. Characterising sediments for a tropical sediment-starved shelf using cluster analysis of physical and geochemical variables. Environ. Chem. 12 (2), 204–226.
Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155–1174.
Rützler, K., 1978. Sponges in coral reefs. In: Stoddart, D.E., Johannes, J.E. (Eds.), Coral Reefs: Research Methods, Monographs on Oceanographic Methodology, vol. 5. UNESCO, Paris, pp. 299–313.
Sanabria, L.A., Cechet, R.P., Li, J., 2013a. Mapping of Australian fire weather potential: observational and modelling studies. In: The 20th International Congress on Modelling and Simulation (MODSIM2013): Adelaide, pp. 242–248.
Sanabria, L.A., Qin, X., Li, J., Cechet, R.P., Lucas, C., 2013b. Spatial interpolation of McArthur's forest fire danger index across Australia: observational study. Environ. Model. Softw. 50, 37–50.
Schlacher, T.A., Schlacher-Hoenlinger, M.A., Williams, A., Althaus, F., Hooper, J.N.A., Kloser, R., 2007. Richness and distribution of sponge megabenthos in continental margin canyons off southeastern Australia. Mar. Ecology-Progress Ser. 340, 73–88.
Schönberg, C.H.L., Fromont, J., 2011. Sponge gardens of ningaloo reef (carnarvon shelf, western Australia) are biodiversity hotspots. Hydrobiologia 687, 143–161.
Shan, Y., Paull, D., McKay, R.I., 2006. Machine learning of poorly predictable ecological data. Ecol. Model. 195, 129–138.
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.
Smale, D.A., Kendrick, G.A., Waddington, K.I., Van Niel, K.P., Meeuwig, J.J., Harvey, E.S., 2010. Benthic assemblage composition on subtidal reefs along a latitudinal gradient in Western Australia. Estuar. Coast. Shelf Sci. 86 (1), 83–92.
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.
Sorokin, S.J., Fromont, J., Currie, D., 2007. Demosponge biodiversity in the benthic protection zone of the Great Australian Bight. Trans. R. Soc. S. Aust. 132, 192–204.
Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forest. BMC Bioinforma. 9, 307.
Strobl, C., Boulesteix, A., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinforma. 8 (25), 25.
Tadić, J.M., Ilić, V., Biraud, S., 2015. Examination of geostatistical and machine-learning techniques as interpolaters in anisotropic atmospheric environments. Atmos. Environ. 111, 28–38.
Van Soest, R.W.M., Boury-Esnault, N., Hooper, J., Rützler, K., de Voogd, N.J., Alvarez, B., Hajdu, E., Pisera, A., Manconi, R., Schoenberg, C., Janussen, D., Tabachnick, K.R., Klautau, M., Picton, B.E., Kelly, M., Vacelet, J., 2014. World Porifera Database. Available online at: http://www.marinespecies.org/porifera. Last consulted on 24 September 2014.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S-plus, fourth ed. Springer-Verlag, New York.
Wilkinson, C.R., Cheshire, A.C., 1989. Patterns in the distribution of sponge populations across the central great barrier reef. Coral Reefs 8, 127–134.