Continental Shelf Research 31 (2011) 1365–1376

Contents lists available at ScienceDirect

Continental Shelf Research

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

Research papers

Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation of mud content across the southwest Australian margin

Jin Li *, Andrew D. Heap, Anna Potter, Zhi Huang, James J. Daniell

Geoscience Australia, GPO Box 378, Canberra, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 16 December 2010
Received in revised form
26 May 2011
Accepted 27 May 2011
Available online 13 June 2011

Keywords:
Geostatistics
Inverse distance weighting (IDW)
Machine learning method
Ordinary kriging (OK)
Random forest
Regression kriging

ABSTRACT

Spatially continuous data of environmental variables is often required for marine conservation and management. However, information for environmental variables is usually collected by point sampling, particularly for the marine region. Thus, methods generating such spatially continuous data by using point samples to estimate values for unknown locations become essential tools. Such methods are, however, often data- or even variable-specific and it is difficult to select an appropriate method for any given dataset. In this study, 14 methods (37 sub-methods) are compared using samples of mud content with five levels of sample density across the southwest Australian margin. Bathymetry, distance-to-coast, slope and geomorphic province were used as secondary variables. Ten-fold cross validation with relative mean absolute error (RMAE) and visual examination were used to assess the performance of these methods. A total of 1850 prediction datasets are produced and used to assess the performance of the methods and the effects of other factors considered. Considering both the accuracy and the visual examination, we found that a combined method (i.e., random forest and ordinary kriging: RKrf) is the most robust method. This method is novel, with a RMAE up to 17% less than that of the control. No threshold in sample density was detected in relation to prediction accuracy. No consistent patterns are observed between the performance of the methods and data variation. The RMAE of three most accurate methods is about 30% lower than that of the best methods in previous publications, highlighting the robustness of the methods selected in this study. The implications and limitations of this study are discussed and a number of suggestions are provided for further studies.

1. Introduction

Spatially continuous data for a range of variables are required for seabed mapping and characterisation, statistical modelling and surrogacy research for the prediction of biodiversity (Pitcher et al., 2008; Whiteway et al., 2007). However, the spatial continuous data are usually not available and the information of environmental variables is usually collected by point sampling. Spatially continuous data are derived for unknown locations from often sparsely spatial distributed point samples. This is particularly true of seabed data due to the expense and practical limitations of acquiring samples.

Statistical/mathematical methods used to derive spatially continuous data are often data, or even variable, specific and their performance is influenced by many factors (Li and Heap, 2008). Existing research provides no consistent findings on how these factors affect the performance of spatial prediction methods, making it difficult to select an appropriate method for any given dataset (Li and Heap, 2008). Spatial interpolation methods such as inverse distance squared (IDS) are commonly applied because of their relative simplicity and availability. For example, IDS was used for spatial prediction in Geoscience Australia. However, predictions using IDS are usually associated with large errors (Li and Heap, 2008). Therefore, it is often a challenge to select an appropriate spatial interpolation method for a given dataset.

In this study, we aim to identify the most robust methods to predict mud content across the southwest Australian margin. Mud content is used because it has been shown to be a major controlling factor on benthic biodiversity in soft sediments (Pitcher et al., 2008). Samples for this study come from Geoscience Australia's marine samples database (MARS; www.ga.gov.au/oracle/mars). The performance of 14 techniques is experimentally compared. We also examine the effects of sample density and data variation on the performance of these methods. The performance of these methods is also visually examined.

2. Methods

2.1. Study area

The study area is in the southwest region of the Australian margin, which covers 523,350 km^2 (Fig. 1). The region covers

* Corresponding author. Tel.: +61 2 6249 9899; fax: +61 2 6249 9956.
E-mail address: Jin.Li@ga.gov.au (J. Li).

0278-4343/$ - see front matter Crown Copyright © 2011 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.csr.2011.05.015

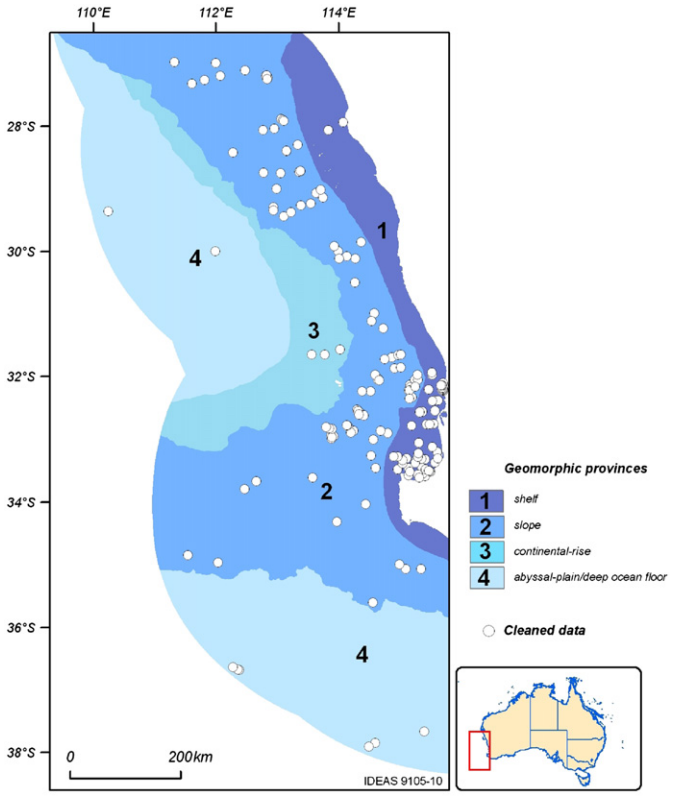

Fig. 1. Spatial distribution of mud sediment samples in the southwest Australian margin.

Table 1
Geomorphic provinces and their area (km^2) in the southwest Australian margin.

Geomorphic province	Shelf	Slope	Rise	Abyssal plain/ Deep ocean floor
Area	52,932	214,938	52,237	203,233

water depths from 0 to 5539 m and all four geomorphic provinces (Heap and Harris, 2008). This region is oriented north–south (Fig. 1). A total of 177 cleaned samples (Li et al., 2010) with sample density of 0.34 samples per 1000 km^2 (or 3000 km^2 per sample) are used. The spatial distribution of samples was also uneven, with most samples being acquired from the shelf and slope (Fig. 1, Table 1).

2.2. Statistical/mathematical methods

A total of 14 methods were compared in this study (Table 2). These methods fall into five categories: (1) non-geostatistical spatial interpolation methods, (2) geostatistical methods, (3) spatial statistical method, (4) machine leaning methods, and (5) combined methods. These methods were selected mainly according to the review of over 40 spatial interpolation methods by Li and Heap (2008) and also on the basis of the applications of machine learning methods in previous studies (Drake et al., 2006; Shan et al., 2006) and our previous modelling experience (Arthur et al., 2010).

2.2.1. Non-geostatistical spatial interpolation methods

Although the inverse distance weighting (IDW) method performs poorly in most cases, it can be used as a control because it is a standard tool used for predicting geospatial data at Geoscience Australia and it is a commonly compared method in spatial interpolation studies (Li and Heap, 2008). Thin plate splines (TPS) method was also included in the study because of its good performance in some studies (Hartkamp et al., 1999; Jarvis and Stuart, 2001; Laslett et al., 1987).

Table 2
Methods compared for predicting mud content.

No.	Method
1	Inverse distance weighting (IDW)
2	Generalised least squares trend estimation (GLS)
3	Kriging with an external drift (KED)
4	Ordinary cokriging (OCK)
5	Ordinary kriging (OK)
6	Universal kriging (UK)
7	Regression tree (RT)
8	Thin plate splines (TPS)
9	General Regression Neural Network (GRNN)
10	Support vector machine (SVM)
11	Linear models and OK (RKlm)
12	Generalised linear models and OK (RKglm)
13	Generalised least squares and OK (RKgls)
14	RandomForest and OK (RKrf)

2.2.2. Geostatistical methods

Kriging with an external drift (KED) and Ordinary cokriging (OCK) (Goovaerts, 1997) were compared because they have been proven to obtain high accuracy when appropriate high quality secondary information is available (Li and Heap, 2008). Ordinary kriging (OK) was considered as it is one of the most commonly compared methods in spatial prediction (Li and Heap, 2008). Universal kriging (UK) was also employed in this study as a trend in the data over space was detected in a preliminary analysis.

2.2.3. Spatial statistical method

Generalised least squares trend estimation (GLS) (Bivand et al., 2008) is used because it allows errors to be correlated (Pinheiro and Bates, 2000; Venables and Ripley, 2002).

2.2.4. Machine leaning methods

Three machine learning approaches were also considered in this study, namely: Regression tree (RT) (Breiman et al., 1984), General Regression Neural Network (GRNN) (Specht, 1990), and Support vector machine (SVM) (Cortes and Vapnik, 1995). The fourth method, random forest (Breiman, 2001; Strobl et al., 2007), was initially considered and outperformed these three methods and also IDW2, so it was used in combination with OK as a combined method below to test if it could further improve the accuracy. The application of SVM and random forest to spatial prediction has not been reported previously (Li and Heap, 2008).

2.2.5. Combined methods

Combined approaches include linear regression models and OK (RKlm), generalised linear models and OK (RKglm), generalised least squares and OK (RKgls) and random forest and OK (RKrf). These four combined approaches are modified versions of regression kriging type C (RK-C) (Asli and Marcotte, 1995; Odeh et al., 1995) that is less sensitive to variation in data and more accurate than other methods (Li and Heap, 2008). For these combined methods, firstly linear regression models (lm), generalised linear models (glm), generalised least squares (gls) and random forest were applied, then OK was applied to the residuals of these models, and finally the predicted values of each model and the corresponding kriged values were added together to produce the final predictions of each combined method. RKrf is

a new combined method that has not been applied in previous studies (Li and Heap, 2008).

2.3. Sample density

Five sample densities were used to test the performance of these methods, namely: 20%, 40%, 60%, 80% and 100% of the total samples collected (Table 3). For densities less than 100% samples were random sampled from the full dataset. The basic summary statistics for mud samples of these datasets are listed in Table 4.

2.4. Secondary information

A number of variables can be used as secondary information to improve the performance of the methods for spatial prediction (Li and Heap 2008). Following a preliminary analysis, geomorphic provinces and bathymetry data that were available at a resolution of 0.01° were used as secondary information. Bathymetry has already been used to improve the performance of spatial interpolators (Verfaillie et al., 2006). The relationship between the bathymetry and sediment grain-size depends on the morphology, topography, and the substrate type (Verfaillie et al., 2006), so the inclusion of such information was expected to improve the predictions. Distance-to-coast and slope might have some influence on the transportation and deposit of mud from onshore sources, so they were also considered as important secondary information in this study.

2.5. Simulation modelling

2.5.1. Data transformation

Geostatistical methods and linear regression models (lm) assume data stationarity of the primary variable, which requires normal distribution and homogeneous variance of samples. This assumption is also necessary for secondary variables when OCK is applied because in OCK secondary variables are modelled as if they were the primary variable. Distributions of mud, bathymetry, distant to coast and slope are left-skewed and non-normal, so appropriate transformations were identified for each variable: arcsine for mud content, and double square root for bathymetry, distance-to-coast and slope. To transform the data for linear models, we often hope to achieve both normality and homogeneous variance. This is usually done in numerous previous studies for methods with requirement of such assumptions. Of course, it cannot be always successful, so we included glm and gls in this study to address this issue. Mud content was also square-root transformed for generalised least squares and transformed to between 0 and 1 for generalised linear models.

Table 3
Sample number and area per sample (km^2/sample) for each sample density.

Sample density	20%	40%	60%	80%	100%
Sample number	35	71	106	142	177
Area per sample	14,953	7371	4937	3685	2957

Table 4
Summary statistics of mud content (%) by sample density.

Sample density (%)	Sample size	Minimum	Mean	Maximum	Standard deviation
20	35	0.01	48.25	97.86	37.84
40	71	0.01	48.09	98.25	37.65
60	106	0.01	47.73	98.25	37.58
80	142	0.01	45.66	98.25	37.1
100	177	0.01	46.23	98.25	37.03

2.5.2. Correlation between mud content and secondary variables

Correlation between the primary and secondary variables is critical to the methods for spatial prediction that use auxiliary information. As the correlation increases, the information brought from the secondary variable on to the primary value increases (Goovaerts, 1997). In this study although the correlation of mud content and the secondary information changes with variables in terms of Pearson's product–moment correlation (r) and Spearman's rank correlation rho (ρ) (Table 5), bathymetry, distance-to-coast and slope all have a high correlation with mud content (Table 5). The relationships between mud content and the secondary variables are non-linear (Fig. 2).

2.5.3. Data projection

The coordinates of mud data were in latitude and longitude based on World Geodetic System 1984 (WGS84) without any projection in this study. The consequence of this choice, knowingly made by the authors, is an implicit assumption of "one degree latitude equals one degree longitude" in the whole study region. Geostatistical methods such as those in the gstat (Pebesma, 2004) package in *R* (R Development Core Team, 2007) expect the input data having been appropriately projected. For data that have no projection information specified, gstat assumes a unit difference in longitude reflects approximately the same distance in latitude and ignores the changes in distance along the latitude, meaning that it use Pythagoras to compute a distance between two points in variogram modelling and kriging (personal communication with Edzer Pebesma in May 2009). Although we were aware of these limitations in using gstat, we decided to model our data in WGS84, because: (1) the study area spans many UTM zones and (2) equal distance projections do not produce satisfactory projections for the purposes of our study.

2.5.4. Variogram modelling

There is no obvious anisotropy detected in the semivariogram maps (Fig. 3). There are a number of variogram models that could be employed and different variogram models may lead to different predictions (Li and Heap, 2008). Thus selecting an appropriate model to capture the features of the data is critical. In this study, variogram model was selected based on the fitted values of range nugget and sill from a range of models including Bessel, Circular, Exponential, Exponential class, Gaussian, Linear, Logarithmic, Pentaspherical, Periodic and Spherical using gstat package (Pebesma, 2004) in *R* (R Development Core Team, 2007). Three models fitted the data more closely than the others, which were then compared (Fig. 3). Of these models, Spherical model was selected for method using and not using secondary variable(s) as it fitted the data better in terms of range, nugget and sill (Table 6), and was then applied to residuals of RK related methods and also for KED for each of the 50 datasets.

Table 5
Pearson's product–moment correlation (r) and Spearman's rank correlation (ρ) of mud content with bathymetry, distance-to-coast and slope. The test for correlation between paired samples was conducted in *R* (R Development Core Team, 2007).

Variable	r	p-Value for r	ρ	p-Value for ρ
Bathymetry	−0.6268	0	−0.6286	0
Distance-to-coast	0.5325	0	0.6171	0
Slope	0.4228	0	0.5522	0

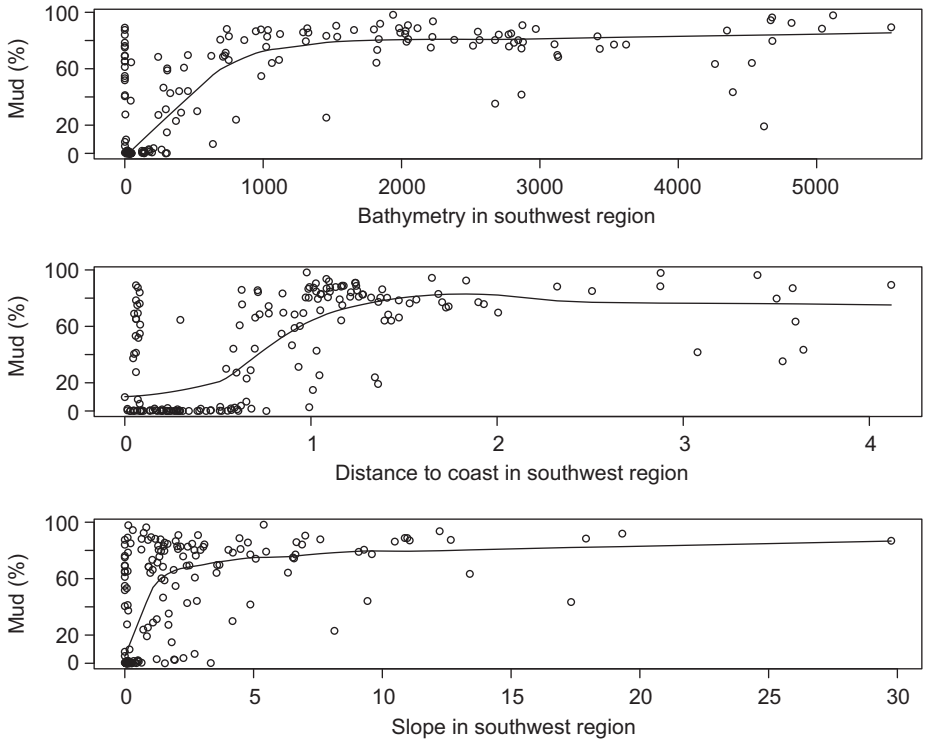

Fig. 2. Relation between mud content and secondary variables (i.e., bathymetry (m), distance-to-coast (deg.) and slope (deg.)). The curve was fitted using lowess in *R* (R Development Core Team, 2007).

2.5.5. Parameters specification

The parameters specified for each of the 14 methods are summarised in Table 7, which leads to 37 sub-methods for comparison. A distance power of 1, 2, 3 and 4 was used in IDW. Of which IDW2 (i.e., IDW with power 2 or IDS), a method used in Geoscience Australia (e.g., Whiteway et al., 2007), was used as the control. Given that bathymetry was the most strongly correlated variable with mud content, it was used as a secondary variable in all methods that consider secondary information. Distance-to-coast and slope were also used in relevant models. The inclusion of latitude and longitude up to third-order polynomial (i.e., the terms in Legendre and Legendre's (1998) equation) was used in UK, KED4, RKlm5, RKglm5, and RKgls5. Appropriate data transformation of secondary information was used for OCK3 and OCK4. In RKrf, KED3, KED4, RKlm5, RKglm5 and RKgls5, we used all possible secondary variables, their second and third power, and latitude and longitude up to third-order polynomial. For generalised least squares, spherical spatial correlation was specified. For generalised linear models, a quasibinomial family with a logit link was used. All these modelling work was implemented using packages including gstat, randomForest and nlme in *R* (R Development Core Team, 2007) with a searching neighbourhood size of 20, except for TPS, GRNN, RT and SVM. TPS was implemented in ArcGIS desktop with two sets of parameters with a searching neighbourhood size of 12. The first set used the spline type of "regularised" with a weight of 0.1, and the second set used the spline type of "tension" with a weight of 5. GRNN, RT and SVM were implemented in DTREG. Predictions were corrected by resetting the faulty estimates to the nearest bound of the data range (i.e., 0% or 100%) if applicable (Goovaerts, 1997).

2.6. Assessment of method performance

To compare the performance of the methods with different sample densities, a 10-fold cross-validation was used according to the findings in Hastie et al. (2001) and Kohavi (1995). The existing cross-validation programme could be used to do this task, but due to random sampling, each method may receive different samples for prediction and validation. To avoid this random error, we randomly split each of the five datasets (for sample density) into 10 sub-datasets. Nine sub-datasets were combined and used for model development and making predictions and the remaining one was used to validate the predictions. This was repeated, varying the validation dataset, until all 10 sub-datasets had been allocated for validation. Consequently, 50 datasets were generated for model development and another 50 datasets for model validation. All methods were applied to the same data in making their predictions. Each method generated 50 prediction datasets. A total of 1850 prediction datasets were produced for assessing the performance of these methods. These data manipulations were implemented in *R* (R Development Core Team, 2007).

Performance of these methods was assessed by identifying the error in the predictions. For each method, the predictions from the 50 datasets were compared to the observed values in the 50 corresponding validation datasets. Relative mean absolute error (RMAE) is not sensitive to the changes in unit/scale (Li and Heap, 2008). Therefore RMAE was used to assess the performance of various methods:

$$RMAE = \frac{1}{n}\sum_{i=1}^{n} |(p_i - o_i)/o_m| 100 \quad (1)$$

where n is the number of observations or samples, o is the observed value, p is the predicted or estimated values, and o_m is the mean of the observed values.

3. Results

Since we aimed to identify the best method for spatial prediction, we first compared the performance of the methods

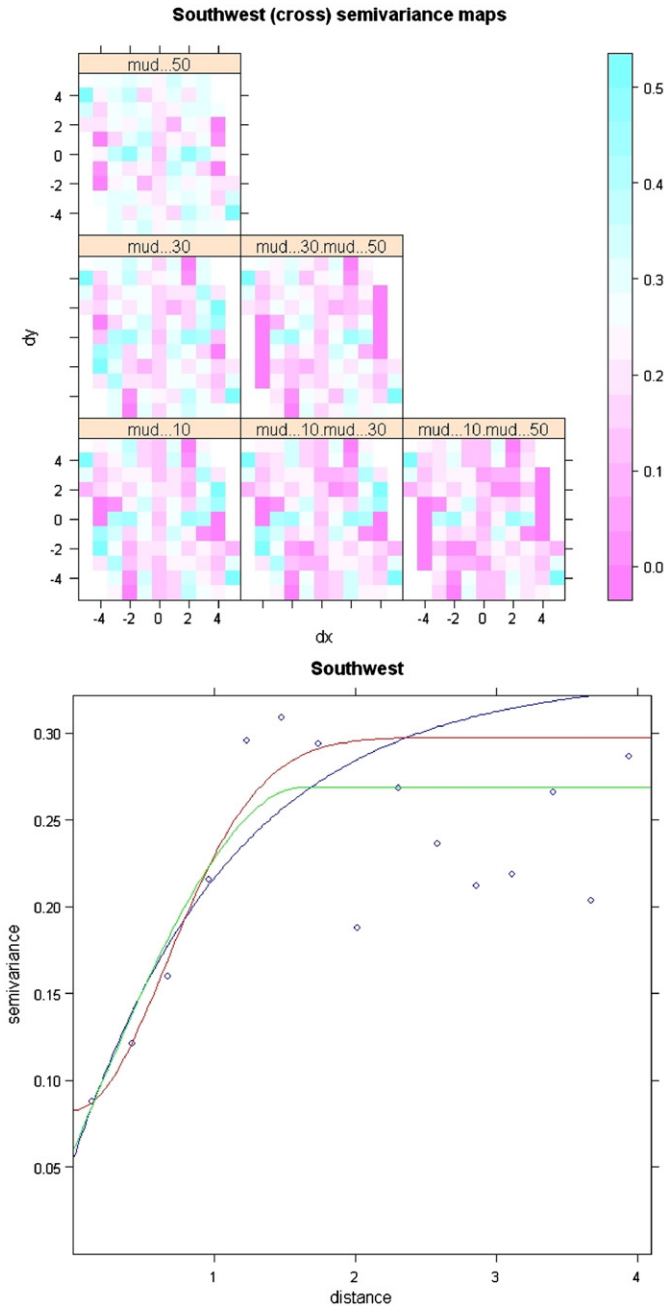

Fig. 3. Variogram maps (top) and variogram models (bottom; exponential: blue, Gaussian: red, and spherical: green) compared using gstat in *R* (R Development Core Team, 2007). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Fitted values by three variogram models for the study region.

Variogram model	Nugget	Partial sill	Range
Exponential	0.0548	0.2781	1.1453
Gaussian	0.0828	0.2149	0.9317
Spherical	0.0599	0.209	1.6286

based on the full dataset. Then we examined the effects of sample density and compared the possible effects of data variation on the most accurate methods and the control method. Finally we visually examined the predictions of these methods.

3.1. Performance of the methods

The performance of all 37 methods/sub-methods was compared in Fig. 4. Three groups of methods were identified: methods outperformed the control (IDW2), methods performed similarly to the control, and methods underperformed the control (Fig. 4, Table 8). The predictive errors of three most accurate methods (i.e., RKrf, IDW4 and IDW3) were significantly less than that of the control. IDW2 was not significantly more accurate than OK, RT, KED1 and TPSt; and it was more accurate than RKgls3 and apparently more accurate than all the remaining methods.

Among the three most accurate methods, the predictive errors of the most accurate method(s) were not significantly less than that of the remaining method(s). The prediction error of RKrf is 17% lower than that of the control.

Non-geostatistical spatial interpolation methods like IDW3 and IDW4 were significantly more accurate than geostatistical methods like OK and KED.

Within geostatistical methods OK performed better than the methods using secondary information such as KED (although not significantly for KED1), OCK and UK.

Except RKrf, all combined methods performed more poorly than the control.

Inclusion of latitude and longitude up to third-order polynomial as secondary information reduced the prediction accuracy of KED and all combined methods with an exception of RKrf.

TPSt performed much better than TPSr, but was less accurate than RKrf, IDW4 and IDW3.

Among the machine learning methods RT performed better than SVM and GRNN. RT was not significantly less accurate than IDW2.

3.2. Effects of sample density

The most accurate method, RKrf, displayed the best performance when sample density was 100%, although it performed better at a sample density of 40% than 60% (Fig. 5). A similar pattern was observed for IDW4, IDW3 and IDW2 (the control). In general, as sample density increased from 20% to 100%, the accuracy of the three selected methods increased by 13–17% in terms of RMAE.

3.3. Data variation

There was no obvious relationship detected between CV and the performance of these four methods, although they all displayed similar patterns (Fig. 6).

3.4. Visual comparison

The spatial distribution of mud samples and the predictions of the three most accurate methods and the control are illustrated in Fig. 7. In comparison with the spatial distribution of mud samples and mud content, the predictions adjacent to the coastal region of the control, IDW4 and IDW3 were higher than those of RKrf. Predictions of IDW2 (the control), IDW4 and IDW3 displayed "bull's-eye" patterns at locations where samples comprised either high or low mud content, which were most apparent for IDW4 and IDW3.

4. Discussion

4.1. Performance of the methods

The most accurate methods, RKrf, can reduce the prediction error by up to 17% in comparison with the control (IDW2).

Table 7
Parameters used for each of 37 methods: bathy, bathymetry; dist.coast, distance-to-coast; lat, latitude; lon, longitude; prov, geomorphic province.

Method	Distance power/trend/secondary variables
IDW2	2
IDW1	1
IDW3	3
IDW4	4
GLS1, KED1, OCK1, RKglm1, RKgls1, RKlm1	bathy
GLS2, KED2, OCK2, RKglm2, RKgls2, RKlm2	bathy, dist.coast, slope
KED3	bathy, dist.coast, slope, bathy^2, bathy^3, dist.coast^2, dist.coast^3, slope^2, slope^3
KED4, RKglm5, RKgls5, RKlm5, RKrf	bathy, dist.coast, slope, bathy^2, bathy^3, dist.coast^2, dist.coast^3, slope^2, slope^3, lat, lat^2, lon, lon^2, lat*lon, lat*lon^2, lon*lat^2, lat^3, lon^3
OCK3	sqrt(sqrt(bathy*(−1)))
OCK4	sqrt(sqrt(bathy*(−1))), sqrt(sqrt(dist.coast)), sqrt(sqrt(slope))
OK	na
UK	lat, lat^2, lon, lon^2, lat*lon, lat*lon^2, lon*lat^2, lat^3, lon^3
RT, GRNN, SVM	lat, lon, bathy, slope, dist.coast
TPSr	Spline.regularised
TPSt	Spline.tension
RKglm3, RKgls3, RKlm3	bathy, dist.coast, slope, prov
RKglm4, RKgls4, RKlm4	bathy, dist.coast, slope, lat, lon, prov

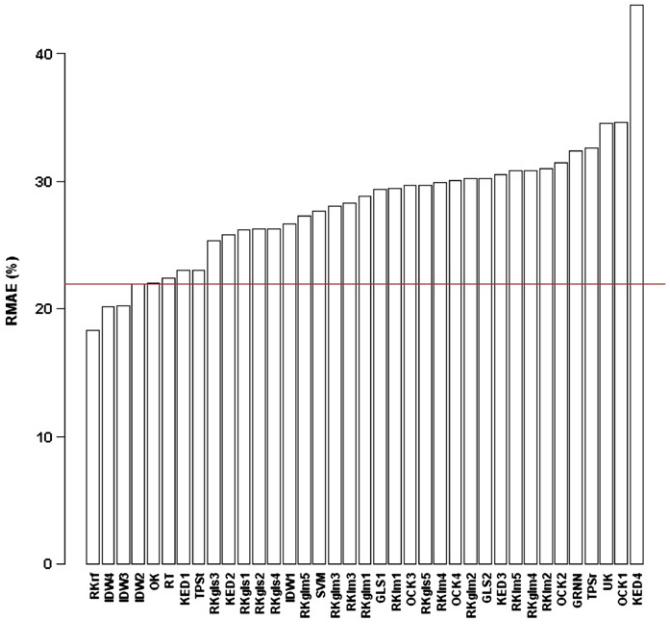

Fig. 4. The relative mean absolute error (RMAE (%)) of the methods for a dataset with 100% sample density. Horizontal line indicates the accuracy of the control (IDW2).

The high accuracy of RKrf could be attributed to the method itself and the high correlation of mud content to the secondary variables, because the accuracy of regression modelling depends on how well the data are sampled and how significant the correlation is between the primary variable and secondary variable (Hengl, 2007). However, the latter explanation may not be true in this study because all methods using secondary information like OCK, KED and RK (except RKrf) performed more poorly than IDW and OK, despite of the findings of previous studies that show stronger correlations result in more accurate predictions by CK and OCK (Goovaerts, 1997), by OCK over OK and RK-C (Martínez-Cob, 1996) and by SKlm, KED and OCCK (Goovaerts, 2000). It was also argued that a threshold exists because for a correlation >0.4 SCK and OCK performed better than other methods (SK, OK, LM) (Asli and Marcotte, 1995). However, this still cannot explain the results observed in this study as the correlation coefficients are above this threshold (Table 5). Therefore, it is the method itself (RKrf) that attributed to its superior performance observed.

Table 8
Comparison of the eight most accurate methods with the control method (IDW2), among three most accurate methods, and of the three methods with OK. *p*-Values were derived from paired *t*-test of the RMAE of a method with higher accuracy versus a method with lower accuracy based on the results of 10-fold cross validation to show whether the predictive errors of the former were significantly less than that of the later.

Methods	*p*-Value	Methods	*p*-Value
IDW3 vs. IDW2	0.0092	**IDW2 vs. RKgls3**	0.0300
IDW4 vs. IDW2	0.0439	**RKrf vs. IDW3**	0.1132
RKrf vs. IDW2	0.0147	**RKrf vs. IDW4**	0.1558
IDW2 vs. OK	0.6091	**IDW4 vs. IDW3**	0.3611
IDW2 vs. RT	0.5018	**IDW3 vs. OK**	0.0126
IDW2 vs. KED1	0.1845	**IDW4 vs. OK**	0.0420
IDW2 vs. TPSt	0.1314	**RKrf vs. OK**	0.0101

The excellent performance of RKrf may be attributed to the following factors. The first is that it uses both bagging (bootstrap aggregation), a successful approach for combining unstable learners, and random variable selection for tree building; and it thus yields an ensemble that can achieve both low bias and low variance (Breiman, 2001; Diaz-Uriarte and de Andres, 2006). The next is because each tree is unpruned, so as to obtain low-bias trees; and it can also deliver good predictive performance even when predictive variables are noise (Diaz-Uriarte and de Andres, 2006). Although decision trees, neural networks, support vector machines were able to deal with poorly predictable data (Shan et al., 2006), they are outperformed by RKrf in this study. RandomForest can model complex interactions among predictive variables (Cutler et al., 2007; Diaz-Uriarte and de Andres, 2006; Okun and Priisalu, 2007) and is relatively robust to outliers and noise (Breiman, 2001), while linear models such as KED, RKlm, RKglm and RKgls are not able to handle such complex interactions and also sensitive to outliers and noise. Finally, the prediction residuals of random forest are interpolated using ordinary kriging and the interpolated values are added to the predictions of random forest, which further reduces the prediction error.

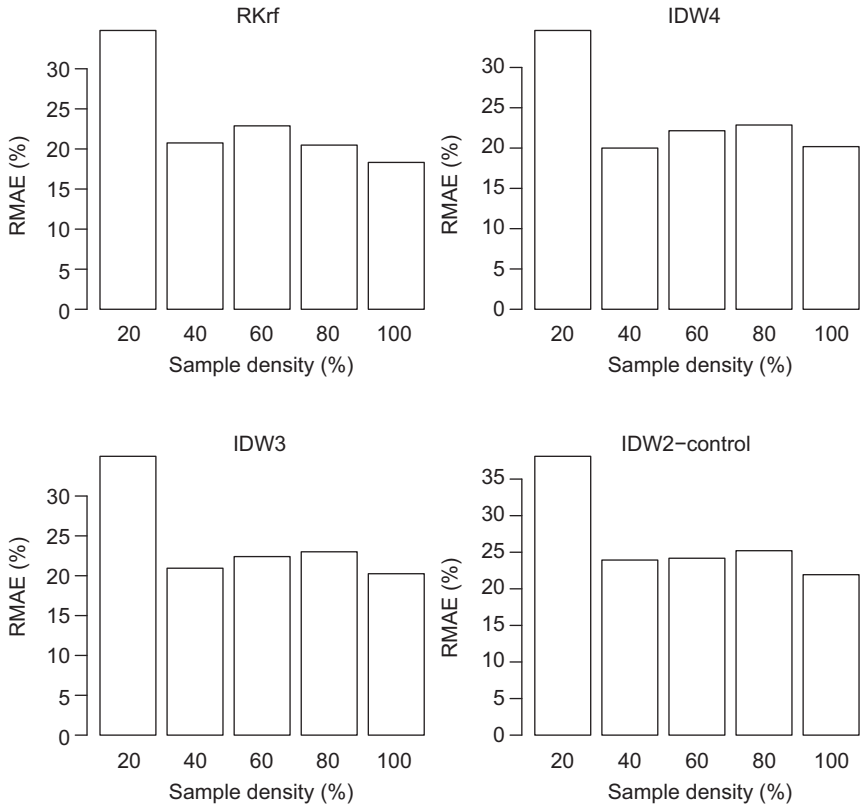

Fig. 5. The relative mean absolute error (RMAE (%)) of top three methods and the control in relation to sample density.

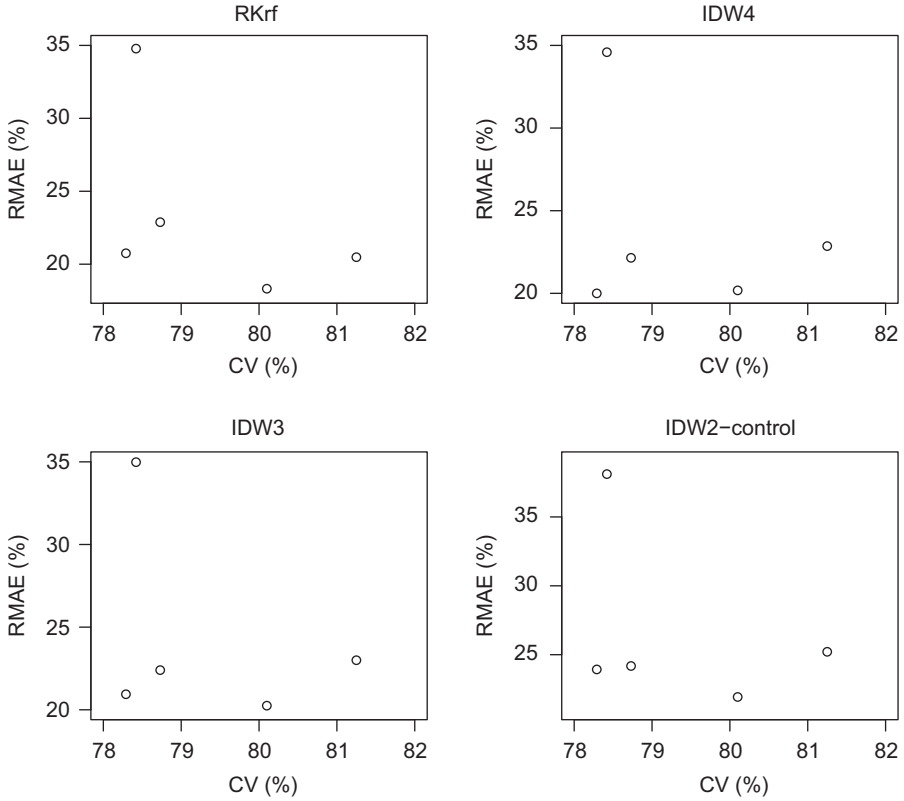

Fig. 6. The relative mean absolute error (RMAE (%)) of the three most accurate methods and the control in relation to CV (%).

OK usually performs better than IDW and is superior at least in theory (Li and Heap, 2008). However, in this study OK performed better than IDW1 but more poorly than IDW2, IDW3 and IDW4. A similar finding was also reported previously that optimal IDW (OIDW) was found to be superior over kriging when data were isotropic and the primary variable was not correlated with the

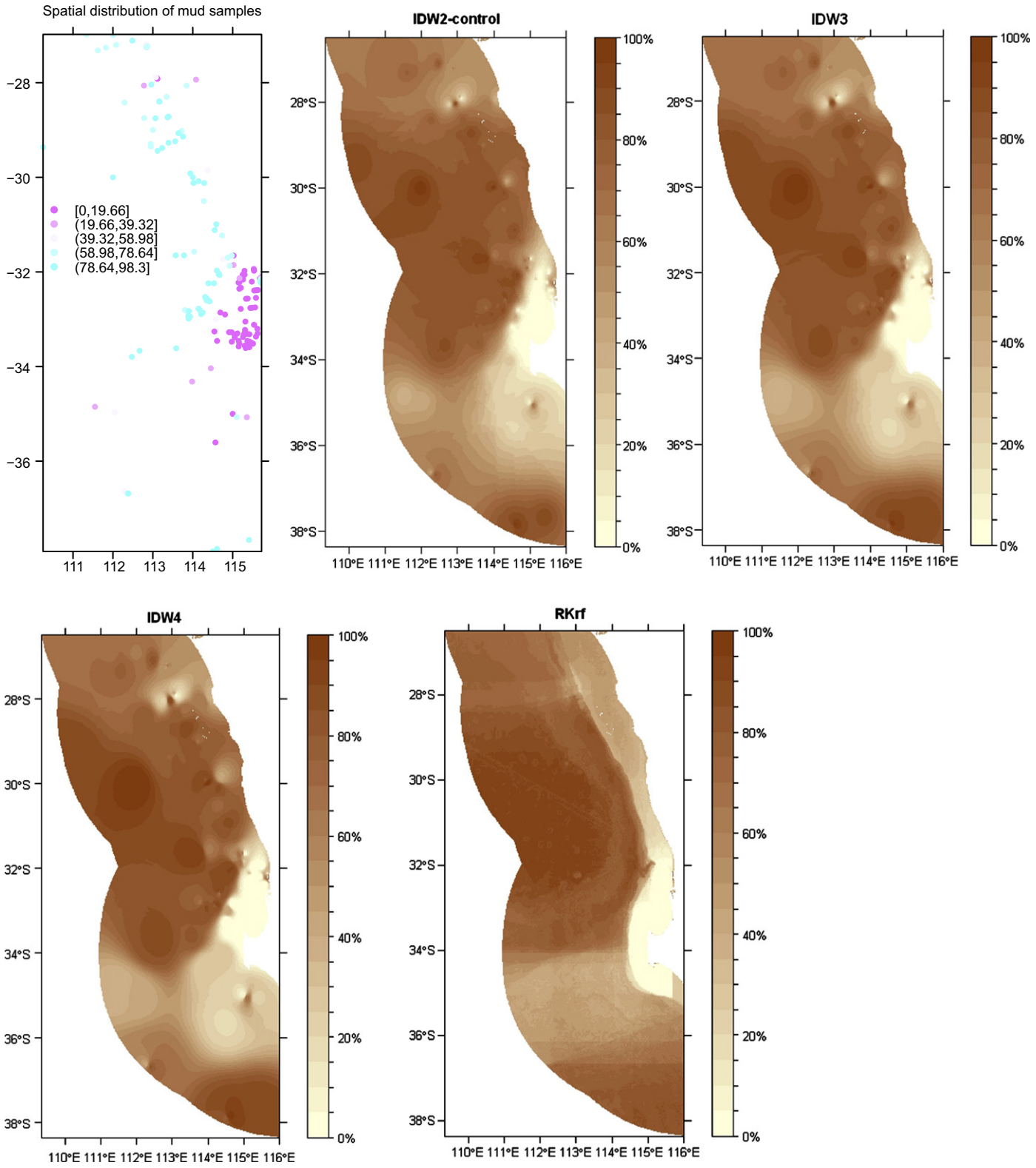

Fig. 7. The predicted spatial distribution of mud in southwest region. IDW2-control; RKrf: the most accurate; IDW4: the second most accurate; and IDW3: the third most accurate.

secondary variable (Collins and Bolstad, 1996). The poor performance of OK in this study could be attributed to: (1) the fact that the data were not projected; (2) that the data stationarity required by OK was not fully satisfied although relevant transformation was employed, and although the data transformation was based on full dataset this might not be the most appropriate for all of the sub-datasets; (3) that the variogram model (i.e., spherical model) selected was based on the full dataset and might also not be the most appropriate for all of the sub-datasets used for prediction; and selection of data-transformation and variogram model for each sub-dataset was not practical for such a comprehensive simulation experiment, which is perhaps a

disadvantage of simulation automation; and (4) the direct back-transformation approach adopted, which may result in a biased estimation of the primary variable because similar phenomena have been reported for lognormal and square-root transformed data (Dambolena et al., 2009; Schuurmans et al., 2007; Yamamoto, 2007). However, for the transformations used in this study, the unbiased procedures of backtransformation are not readily available. Obviously, this is a field worth further research.

Comparing the accuracy of methods using secondary information (e.g., KED, OCK, RK, and UK) and methods without using secondary information (e.g., IDW and OK) (Fig. 5) shows that the effects of including secondary information are method dependent. This suggests that inclusion of secondary information does not always improve the prediction accuracy. However, for RKrf, secondary information is essential.

A singular model in variogram fit (Pebesma, 2004) was observed for KED3, KED4 and RKgls, which may contribute to their poor performance. Overfit may also contribute to the poor performance of some regression kriging methods such as RKglm4 and RKlm5 because the models were not simplified.

The negative effects of including latitude and longitude up to third-order polynomial as secondary information on the prediction accuracy of KED and all combine methods (except RKrf) probably result from the un-projected coordinates of mud content as explained in Section 2.5.3. For data in longitude and latitude, gstat assumes a unit difference in longitude axis reflects approximately the same distance in latitude axis. Ignoring changes in distance along the latitude will certainly reduce the accuracy of distance estimated and thus may reduce the reliability of the predictions.

TPSt and TPSr were found to be less accurate than RKrf, IDW4, IDW3, IDW2, OK, RT and KED1. This is consistent with previous studies that OK and IDW appeared more accurate than TPS (Brus et al., 1996; Schloeder et al., 2001) and for precipitation(Hartkamp et al., 1999). However, TPS performed slightly better than IDW and OCK for temperature (Hartkamp et al., 1999), than IDW and OK (Jarvis and Stuart, 2001; Laslett et al., 1987). These findings indicate that the performance of these methods varies between studies and depends on other variables such as the primary variables.

Although the machine learning methods like RT and SVM performed more poorly than IDW and OK, another machine learning method, random forest, outperformed all other methods when it was combined with OK. A combined method of RT and OK has been shown to be more accurate than IDW and OK (Martínez-Cob, 1996) and SVM was found to be superior to random forest in other disciplines (Statnikov et al., 2008). Therefore, RT and SVM may have great potential if they are combined with other methods like OK for the spatial prediction of marine environmental variables.

4.2. Sample density

In general, as sample density increases the accuracy of the methods for spatial prediction increases. This finding is consistent with other studies (Englund et al., 1992; Isaaks and Srivastava, 1989; Stahl et al., 2006). The general trend and observed interactive effects at lower densities agree with findings of Li et al. (2007) and Wang et al. (2005).

These findings suggest that any increase in sample size (or sample density) could improve the prediction accuracy of the methods for spatial prediction. No threshold in sample size was found, which suggests that the number of samples collected so far is still below the threshold if it exists. Therefore more samples need to be collected.

4.3. Data variation

There is no consistent relationship between the accuracy of the methods for spatial prediction and data variation, which is inconsistent with the patterns observed by Li and Heap (2008), Collins and Bolstad (1996), Martínez-Cob (1996) and Schloeder et al. (2001). This phenomenon might be attributed to the very short gradient of data variation in this study, which was $\leq 3\%$.

The accuracy of the three most accurate methods in this study is much higher than those reported in previous studies (Li and Heap, 2011) (Fig. 8). The data CV in this study is around 80%, and the RMAE of these three best methods decreased by approximately 30% compared with that of the best methods reported in previous publications (Li and Heap, 2011). This finding demonstrates that the methods selected in this study are far more reliable than the methods identified in previously published studies.

4.4. Visual examination

The spatial patterns of predictions of RKrf, the most accurate method, mainly reflect the effect of bathymetry and its related variables like geomorphic features as the patterns were similar to those identified by Heap and Harris (2008) (Fig. 9). This can be explained by the fact that bathymetry is the most important variable for random forest (Fig. 10). Influences of geomorphic features (shelf, slope, continental rise, abyssal plain, canyons and Naturaliste Plateau) (Heap and Harris, 2008) were obvious, despite the fact that the effects of latitude were also apparent in area between -34° and -37°. However, IDW2, IDW3 and IDW4 failed to capture these patterns and resulted in the clear artefacts, "bull's-eye" patterns, in the prediction maps. Therefore, the predictions of RKrf are more reliable than those of IDW2, IDW3 and IDW4 based on visual examination.

4.5. Limitations and uncertainty

4.5.1. Data projection

Data in this study was in latitude and longitude (i.e., WGS84), and not projected, which could reduce the accuracy of geostatistical methods and all other methods using latitude and longitude as secondary information. Since RKrf, the selected method, used secondary information, its accuracy might also be affected. This limitation needs to be taken into account in assessing the

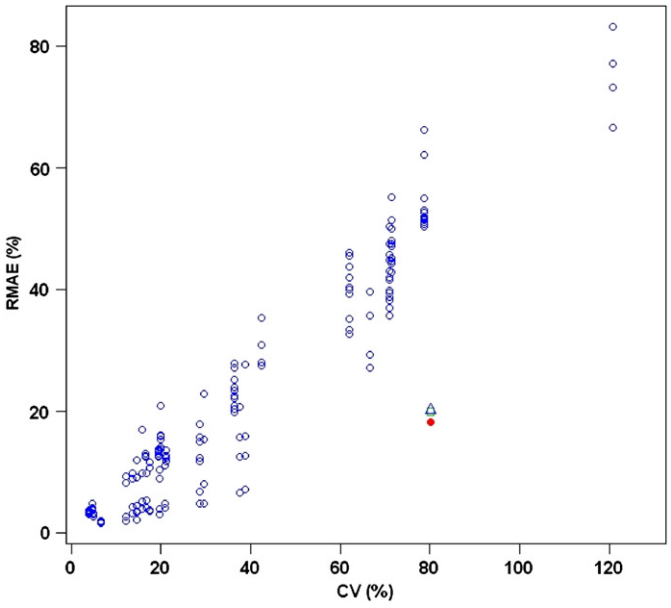

Fig. 8. The relative mean absolute error (RMAE (%)) of the best three methods (the best: red solid circle; the second-best: green square; the third best: blue triangle) in relation to CV (%) compared results (blue open circles) of previous studies (Li and Heap, 2011). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

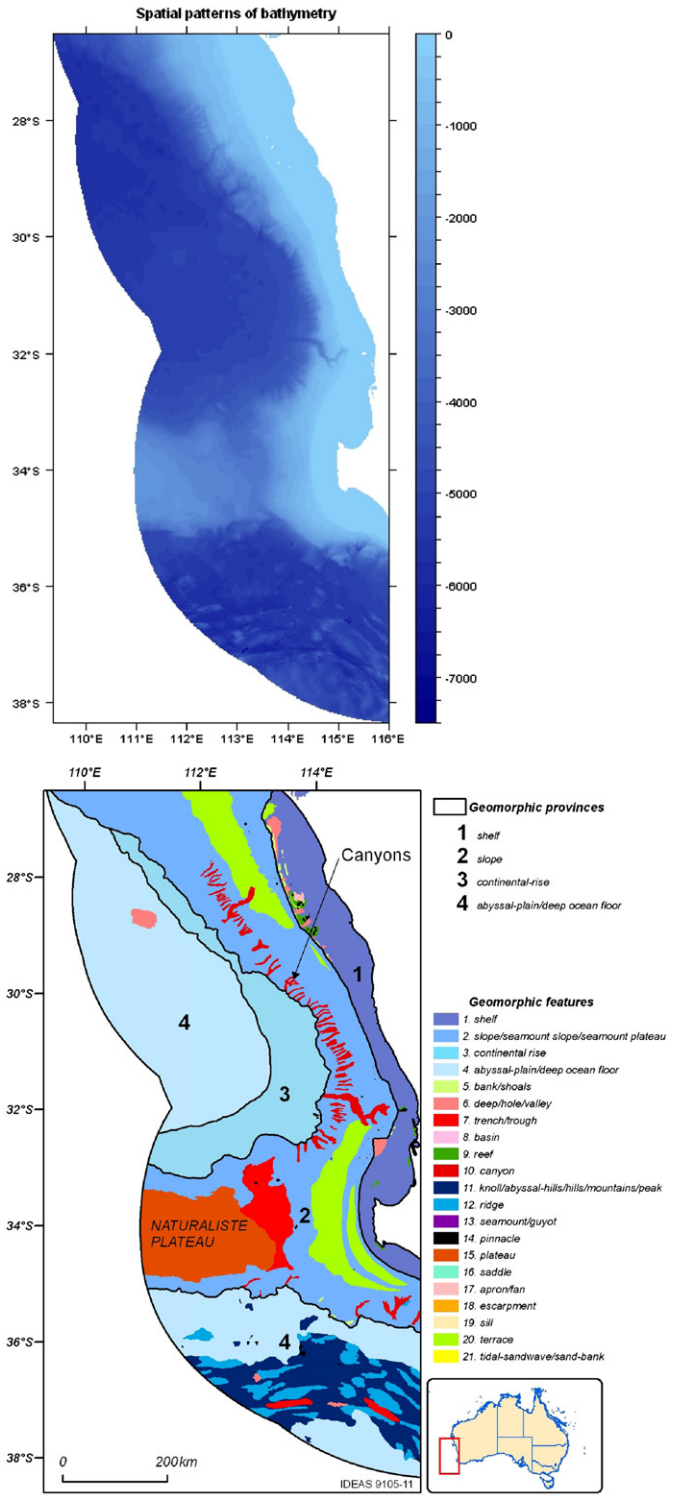

Fig. 9. The spatial pattern of bathymetry and the spatial distribution of geomorphic features in southwest Australian margin (Heap and Harris 2008).

performance of the methods using such secondary information or using variogram modelling.

4.5.2. Data sampling

To test the effect of sample density, we randomly selected a portion of samples from the full dataset. The resultant sub-datasets of higher sample density do not necessarily contain all samples in sub-datasets of lower sample density. This could result in that predictions using a dataset with lower sample density being more accurate than those using a dataset with higher sample density as observed in this study. This phenomenon was mainly caused by the difference in the mean and CV of samples of the two datasets. Differences in spatial distribution of samples in two datasets may also be crucial (Isaaks and Srivastava, 1989; Laslett, 1994; Zimmerman et al., 1999). Therefore, in future studies for testing the effect of sample density, to avoid such random error, samples from the lower sample density dataset should be included in the higher sample density dataset.

4.5.3. Searching neighbourhood

The local searching window size was specified based either on previous studies for methods computed in *R* or on the default value for methods computed in ArcGIS. This size may not lead to an optimised prediction.

4.5.4. Validation

The 10-fold cross validation was applied to the predictions of each method compared using the spatially clustered samples because samples are denser on shelf than on other sections. The spatially clustered patterns may affect the validation results, which may depend on a number of influencing factors such as the distance between samples in space, distance between samples along the environmental gradient, correlation between primary and secondary variables, whether the method can model trend, local effect or both, and possible interactions between various factors. For example, with an assumption that the primary and secondary variables are correlated, if two samples are far apart in space, the local effect becomes minimal, but the role of a trend model depends on their locations in the environmental space; if two samples are apart with short spatial distance, local effect becomes important and the trend model may be important depending on their locations in the environmental space. Therefore, the role of the trend model depends on the changes in the environmental space, while the role of the local effect may change with both the spatial distance and the trend model if it is combined with the trend model such as RK. No quantitative study on how spatially clustered samples affect the results of cross validation has been reported and it is worth further exploitation.

5. Summary and recommendations

5.1. Important findings

- *The most robust methods:* Considering both the prediction accuracy and visual examination, we found that RKrf is the most robust method for predicting mud content across the southwest Australian margin.
- *Sample density:* Increases in sample size (or sample density) improves the prediction accuracy of the methods for spatial prediction. No threshold in sample size was found in relation to prediction accuracy.
- *Data variation:* No consistent patterns of method performance were observed in relation to data variation. However, the accuracy of the best three methods increased by up to 30% in terms of RMAE in comparison with performance of the best methods in previous publications, highlighting the robustness of the methods selected in this study.
- *Secondary information:* The effects of inclusion of secondary information are method dependent.
- *Visual examination:* Visual examination is equally important to the accuracy measurement like RMAE and is an essential step in assessing the predictions of a method.

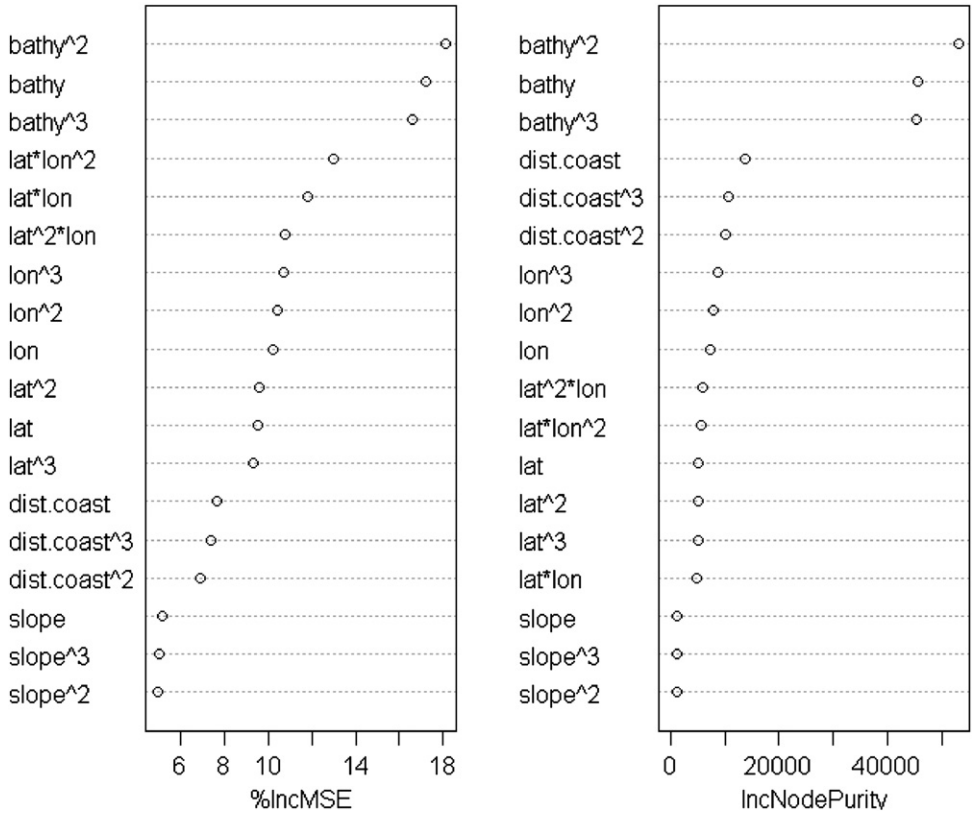

Fig. 10. Variable importance measured by random forest in the randomForest package in *R* (R Development Core Team, 2007).

5.2. Recommendations for future study

- *Machine learning methods:* The combination of machine learning methods, like random forest, with other methods for spatial prediction is novel. Our findings demonstrate that there is potential in applying machine learning methods to predict spatially continuous variables.
- *Searching neighbourhood:* We have only considered one level of searching window size. To find an optimal size we would recommend that more levels need to be tested.
- *Sample size:* No threshold in sample size was found in relation the accuracy of the method, suggesting that the number of samples collected so far is still below the threshold if it exists. Therefore more samples need to be collected.
- *Secondary information:* Secondary information considered in this study was only limited to bathymetry, distance-to-coast, slope, latitude and longitude. Other correlated secondary variables should be identified and employed as they provide essential information for machine learning methods.
- *Visual examination:* For spatial predictions, visual examination is an essential step to assess the predictions of a method.
- *Directions and requirements for future studies:* Although this simulation experiment has identified a few more robust methods for spatial predictions than the method used previously in Geoscience Australia, it is just a beginning in searching methods to optimise the spatial prediction. The combined methods of machine learning methods and other techniques provide a new direction for future studies. Further studies are warranted for testing their performance in different scenarios. To achieve the optimal spatial predictions of physical variables for Australian margin, additional experiments need to be conducted.

Acknowledgements

We thank Hideyasu Shimadzu for the valuable comments. We also thank Christina Baker, Mark Webster, Shoaib Burq and Tanya Whiteway for preparing the datasets used in this study. Chris Lawson is appreciated for providing bathymetry data and producing several maps. Scott Nichol, David Ryan, and Frederic Saint-Cast are thanked for their suggestions on the experimental design. We also thank Edzer Pebesma, Michael Sumner, Paul Hiemstra and Roger Bivand for their help in using various functions in *gstat*, *maptools* and *sp* packages in *R*. This paper is published with permission of the Chief Executive Officer, Geoscience Australia.

References

Arthur, A.D., Li, J., Henry, S., Cunningham, S.A., 2010. How the spatial distribution of native vegetation influences the ecosystem service of pollination in south-eastern Australian agricultural landscapes. Basic and Applied Ecology 11, 406–414.

Asli, M., Marcotte, D., 1995. Comparison of approaches to spatial estimation in a bivariate context. Mathematical Geology 27, 641–658.

Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V., 2008. Applied Spatial Data Analysis with R. Springer, New York.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont.

Brus, D.J., de Gruijter, J.J., Marsman, B.A., Visschers, B.A., Bregt, A.K., Breeuwsma, A., 1996. The performance of spatial interpolation methods and choropleth maps to estimate properties at points: a soil survey case study. Environmetrics 7, 1–16.

Collins, F.C., Bolstad, P.V., 1996. A comparison of spatial interpolation techniques in temperature estimation. In: Proceedings of the Third International Conference/Workshop on Integrating GIS and Environmental Modeling, Santa Fe, NM; National Center for Geographic Information and Analysis, Santa Barbara, CA.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

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, 2783–2792.
Dambolena, I.G., Eriksen, S.E., Kopcso, D.P., 2009. Logarithmic transformations in regression: do you transform back correctly? Primus 19, 280–290.
Diaz-Uriarte, R., de Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3.
Drake, J.M., Randin, C., Guisan, A., 2006. Modelling ecological niches with support vector machines. Journal of Applied Ecology 43, 424–432.
Englund, E., Weber, D., Leviant, N., 1992. The effects of sampling design parameters on block selection. Mathematical Geology 24, 329–343.
Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York.
Goovaerts, P., 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology 228, 113–129.
Hartkamp, A.D., De Beurs, K., Stein, A., White, J.W., 1999. Interpolation Techniques for Climate Variables.. CIMMYT, Mexico, D.F.
Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag.
Heap, A.D., Harris, P.T., 2008. Geomorphology of the Australian margin and adjacent seafloor. Australian Journal of Earth Sciences 55, 555–585.
Hengl, T., 2007. A Practical Guide to Geostatistical Mapping of Environmental Variables. Office for Official Publication of the European Communities, Luxembourg 143.
Isaaks, E.H., Srivastava, R.M., 1989. Applied Geostatistics. Oxford University Press, New York.
Jarvis, C.H., Stuart, N., 2001. A comparison among strategies for interpolating maximum and minimum daily air temperature. Part II: the interaction between number of guiding variables and the type of interpolation method. Journal of Applied Meteorology 40, 1075–1084.
Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann, pp. 1137–1143.
Laslett, G.M., 1994. Kriging and splines: an empirical comparison of their predictive performance in some applications. Journal of the American Statistical Association 89, 391–400.
Laslett, G.M., McBratney, A.B., Pahl, P.J., Hutchinson, M.F., 1987. Comparison of several spatial prediction methods for soil pH. Journal of Soil Science 38, 325–341.
Legendre, P., L.L., 1998. Numerical Ecology. Elsevier, Amsterdam.
Li, J., Heap, A., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists. Geoscience Australia Record 2008/23, Canberra, p. 137.
Li, J., Heap, A., 2011. A review of comparative studies of spatial interpolation methods: performance and impact factors. Ecological Informatics 6, 228–241.
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 Record 2010/11, Canberra, p. 146.
Li, Y., Shi, Z., Wu, C., Li, H., Li, F., 2007. Improved prediction and reduction of sampling density for soil salinity by different geostatistical methods. Agricultural Science in China 6, 832–841.
Martínez-Cob, A., 1996. Multivariate geostatistical analysis of evapotranspiration and precipitation in mountainous terrain. Journal of Hydrology 174, 19–35.
Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67, 215–226.
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, 4478. Lecture Notes in Computer Science, Girona, Spain, pp. 4483–4490.
Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computer & Geosciences 30, 683–691.
Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-PLUS. Springer, New York.
Pitcher, C.R., Doherty, P.J., Anderson, T.J., 2008. Seabed environments, habitats and biological assemblages. In: Hutchings, P., Kingsford, M., Hoegh-Guldberg, O. (Eds.), The Great Barrier Reef: biology, environment and management. CSIRO Publishing, Collingwood, pp. 377.
R Development Core Team, 2007. R: A Language and Environment for Statistical Computing.. R Foundation for Statistical Computing, Vienna.
Schloeder, C.A., Zimmerman, N.E., Jacobs, M.J., 2001. Comparison of methods for interpolating soil properties using limited data. Soil Science Society of American Journal 65, 470–479.
Schuurmans, J.M., Bierkens, M.F.P., Pebesma, E.J., 2007. Automatic prediction of high-resolution daily rainfall fields for multiple extents: the potential of operational radar. Journal of Hydrometeorology 8, 1204–1224.
Shan, Y., Paull, D., McKay, R.I., 2006. Machine learning of poorly predictable ecological data. Ecological Modelling 195, 129–138.
Specht, D.F., 1990. Probabilistic neural networks. Neural Networks 3, 109–118.
Stahl, K., Moore, R.D., Floyer, J.A., Asplin, M.G., McKendry, I.G., 2006. Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density. Agricultural and Forest Meteorology 139, 224–236.
Statnikov, A., Wang, L., Aliferis, C.F., 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9, 319.
Strobl, C., Boulesteix, A., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8, 25.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S-Plus. Springer-Verlag, New York.
Verfaillie, E., van Lancker, V., van Meirvenne, M., 2006. Multivariate geostatistics for the predictive modelling of the superficial sand distribution in shelf seas. Continental Shelf Research 26, 2454–2468.
Wang, H., Liu, G., Gong, P., 2005. Use of cokriging to improve estimates of soil salt solute spatial distribution in the Yellow River delta. Acta Geographica Sinica 60, 511–518.
Whiteway, T., Heap, A., Lucieer, V., Hinde, A., Ruddick, R., Harris, P.T., 2007. Seascapes of the Australian margin and adjacent sea floor: methodology and results. Geoscience Australia, 133.
Yamamoto, J.K., 2007. On unbiased backtransform of lognormal kriging estimates. Computer & Geosciences 11, 219–234.
Zimmerman, D., Pavlik, C., Ruggles, A., Armstrong, M.P., 1999. An experimental comparison of ordinary and universal kriging and inverse distance weighting. Mathematical Geology 31, 375–390.