Australian Government
Geoscience Australia

Predicting Seabed Sand Content across the Australian Margin Using Machine Learning and Geostatistical Methods

Record
2012/48
GeoCat # 74030

J. Li, A. Potter, Z. Huang and A. D. Heap

Predicting Seabed Sand Content across the Australian Margin Using Machine Learning and Geostatistical Methods

GEOSCIENCE AUSTRALIA
RECORD 2012/48

by

J. Li, A. Potter, Z. Huang and A. D. Heap

Australian Government
Geoscience Australia

Department of Resources, Energy and Tourism
Minister for Resources and Energy: The Hon. Martin Ferguson, AM MP
Secretary: Mr Drew Clarke

Geoscience Australia
Chief Executive Officer: Dr Chris Pigram

This paper is published with the permission of the CEO, Geoscience Australia

ISSN 1448-2177
ISBN 978-1-922103-47-5 (hardcopy)
ISBN 978-1-922103-48-2 (CD/DVD)
ISBN 978-1-922103-49-9 (web)

GeoCat # 74030

Bibliographic reference: Li, J., Potter, A., Huang, Z. and Heap, A. D., 2012. Predicting seabed sand content across the Australian margin using machine learning and geostatistical methods. Record 2012/48. Geoscience Australia: Canberra.

Contents

List of Figures

List of Tables

Abbreviations

AEEZ: Australian Exclusive Economic Zone
AMJ: Australian Marine Jurisdiction
BDT: boosted decision tree
BDTIDS: the combination of BDT and IDS
BDTOK: the combination of BDT and OK
BRT: boosted regression trees
IDS: inverse distance squared
GRNN: general regression neural network
GRNNIDS: the combination of GRNN and IDS
GRNNOK: the combination of GRNN and OK
KED: kriging with an external drift
LSVM: support vector machine with a linear kernel
LSVMIDS: the combination of LSVM and IDS
LSVMOK: the combination of LSVM and OK
LSVMOKSVMIDS: the average of LSVMOK and LSVMIDS
LSVMSVMOKSVMIDS: the average of LSVM, LSVMOK and LSVMIDS
MAE: mean absolute error
MARS: Marine Samples Database
mtry: the number of variables randomly sampled as candidates at each split
OK: ordinary kriging
RF: random forest
6RF: RF with 6 secondary variables
6RFIDS: the combination of 6RF and IDS
6RFOK: the combination of 6RF and OK
6RFOKRFIDS: the average of 6RFOK and 6RFIDS
6RFRFOKRFIDS: the average of 6RF, 6RFOK and 6RFIDS
i4RF: RF with interactions among the secondary variables excluding second and third order terms
i4RFIDS: the combination of i4RF and IDS
i4RFOK: the combination of i4RF and OK
i4RFOKRFIDS: the average of i4RFOK and i4RFIDS
i4RFRFOKRFIDS: the average of i4RF, i4RFOK and i4RFIDS
iRF: RF with interactions among the secondary variables
iRFIDS: the combination of iRF and IDS
iRFOK: the combination of iRF and OK
iRFOKRFIDS: the average of iRFOK and iRFIDS
iRFRFOKRFIDS: the average of iRF, iRFOK and iRFIDS
RFIDS: the combination of RF and IDS
RFOK: the combination of RF and OK
RFOKRFIDS: the average of RFOK and RFIDS
RFRFOKRFIDS: the average of RF, RFOK and RFIDS;
RMAE: relative MAE
RMSE: root mean squared error
RRMSE: relative RMSE

RPART: regression tree (in rpart)
RPARTIDS: the combination of RPART and IDS
RPARTOK: the combination of RPART and OK
SVM: support vector machine with a radial basis kernel
SVMIDS: the combination of SVM and IDS
SVMOK: the combination of SVM and OK
SVMOKSVMIDS: the average of SVMOK and SVMIDS
SVMSVMOKSVMIDS: the average of SVM, SVMOK and SVMIDS
WGS84: World Geodetic System 1984

Executive Summary

PREDICTING SEABED SAND CONTENT ACROSS THE AUSTRALIAN MARGIN USING MACHINE LEARNING AND GEOSTATISTICAL METHODS

Geoscience Australia often produces spatially continuous marine environmental information products using spatial interpolation methods to support Australian marine zone management. The accuracy of such information is critical for making well-informed decisions for marine environmental management and conservation. Improving the accuracy of these data products by searching for the most accurate method is essential, but it is a difficult task since no method is best for all variables, and the predictive accuracy is affected by many factors. From 2008, we introduced machine learning methods (e.g., random forest: RF) into spatial statistics by combining them with existing spatial interpolation methods because of their high predictive accuracy. We experimentally examined their performance in relation to a number of factors using seabed mud content data from the Australian Exclusive Economic Zone (AEEZ). These studies generated several novel and robust methods, which have opened up a new source of tools for spatial interpolation in environmental sciences, and have formed a solid scientific foundation for this study.

In this study, we aim to identify the most appropriate methods for spatial interpolation of seabed sand content for the AEEZ using samples archived in Geoscience Australia's Marine Samples Database (www.ga.gov.au/oracle/mars). Data was cleaned according to seven criteria. We then used the clean datasets in three regions to experimentally examine:

- the performance of machine learning and geostatistical methods,
- the effects of the input secondary variables on the performance of the methods,
- the accuracy of averaging the predictions of the most accurate methods,
- the effects of 'the number of variables randomly sampled as candidates at each split' (*mtry*) on the best performing methods, and
- the effects of search neighbourhood size on the best performing methods.

The performance of the methods was assessed using 10-fold cross-validation. The predictive accuracy was compared based on information extracted from 4,740 prediction datasets generated in this experiment. We visually examined and analysed the prediction patterns of the most accurate methods based on their prediction maps.

The predictive accuracy changes with method, input secondary variable, model averaging, search window size and the study region but not the choice of mtry. No single method performs best for all scenarios. Of the 18 methods compared, RFIDS and RFOK are the most accurate methods in all three regions. Overall, of the 36 combinations of input secondary variables, methods and regions, RFIDS, 6RFIDS and RFOK were among the most accurate methods in all three regions. Model averaging further improved the prediction accuracy. The most accurate methods reduced the prediction error by up to 7%. If a single method is required for predicting sand content across the AEEZ, RFOKRFIDS (i.e., averaging RFOK and RFIDS) with a search window size of 5 and an mtry of 4 is recommended.

This study will assist in producing more accurate physical data of seabed sand content for the AEEZ which in turn will better inform management and conservation of Australian marine zone by government, industry and the community. This study provides suggestions and guidelines for improving the spatial predictions of various environmental variables.

Chapter 1. Introduction

Geoscience Australia often produces spatially continuous marine environmental information products using spatial interpolation methods based on point samples of environmental variables for environmental management and conservation. These methods fall into three groups: 1) non-geostatistical methods (e.g., inverse distance weighting), 2) geostatistical methods (e.g., ordinary kriging: OK), and 3) combined methods (e.g. regression kriging) (Li and Heap, 2008). The accuracy of such information is critical for making well-informed decisions for marine environmental management and conservation (McArthur et al., 2009; Pitcher et al., 2008; Post, 2008) (Figs. 1.1 and 1.2). Improving the accuracy of these data products by searching for robust methods is essential, but it is a vexed task since no method is best for all variables and the predictive accuracy is affected by many factors (Li and Heap, 2008, 2011).

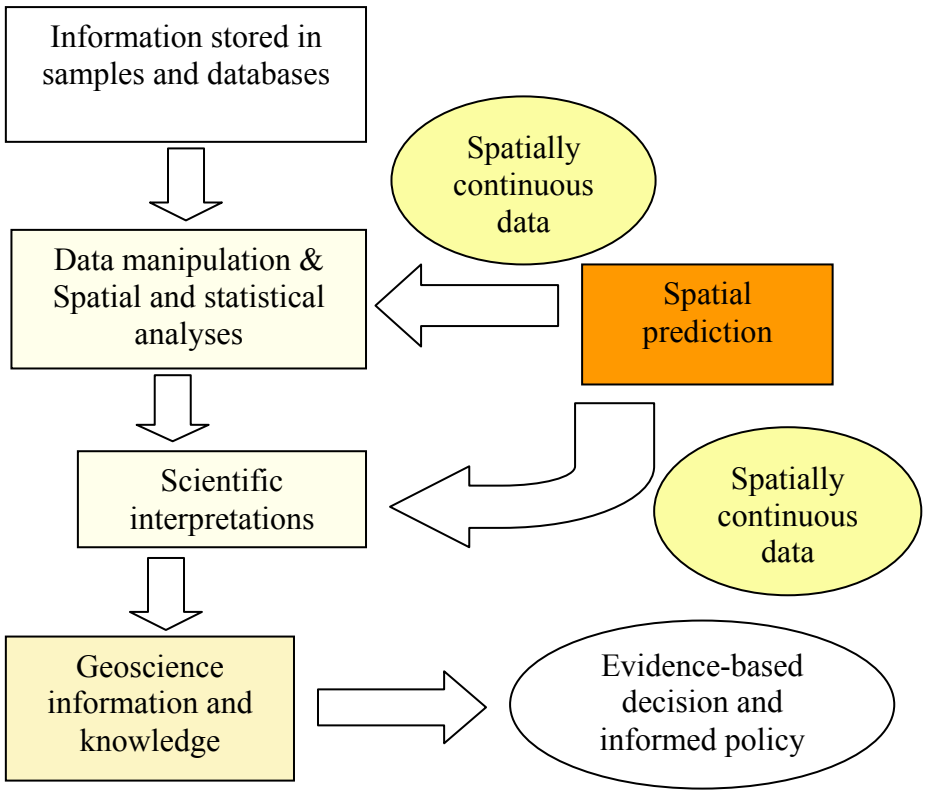


Figure 1.1. *The role of spatially continuous data in generation geoscience information and knowledge.*

Machine learning methods, such as random forest (RF) and support vector machine, have proven their accuracy in the field of data mining (Cutler et al., 2007; Diaz-Uriarte and de Andres, 2006; Drake et al., 2006; Shan et al., 2006), so we introduced them into spatial statistics by combining them with existing spatial interpolation methods (Li et al., 2011a; Li et al., 2010). We experimentally compared the performance of a variety methods/sub-methods using seabed mud content data from the Australian Exclusive Economic Zone (AEEZ) (Li et al., 2011a; Li et al., 2010). In these studies, we

developed a few novel and robust methods that significantly increased the accuracy of spatial predictions. This development can be viewed as an extension of the combined methods from statistical methods to machine learning field, which opened a new research direction and an alternative source of tools for spatial interpolation in environmental sciences (Li, 2011; Li et al., 2011b; Li et al., 2011c). Consequently, more accurate spatial information of the physical properties of Australia's marine jurisdiction was provided by Geoscience Australia to the government, industry, community and its various stakeholders and clients (Li et al., 2011d).

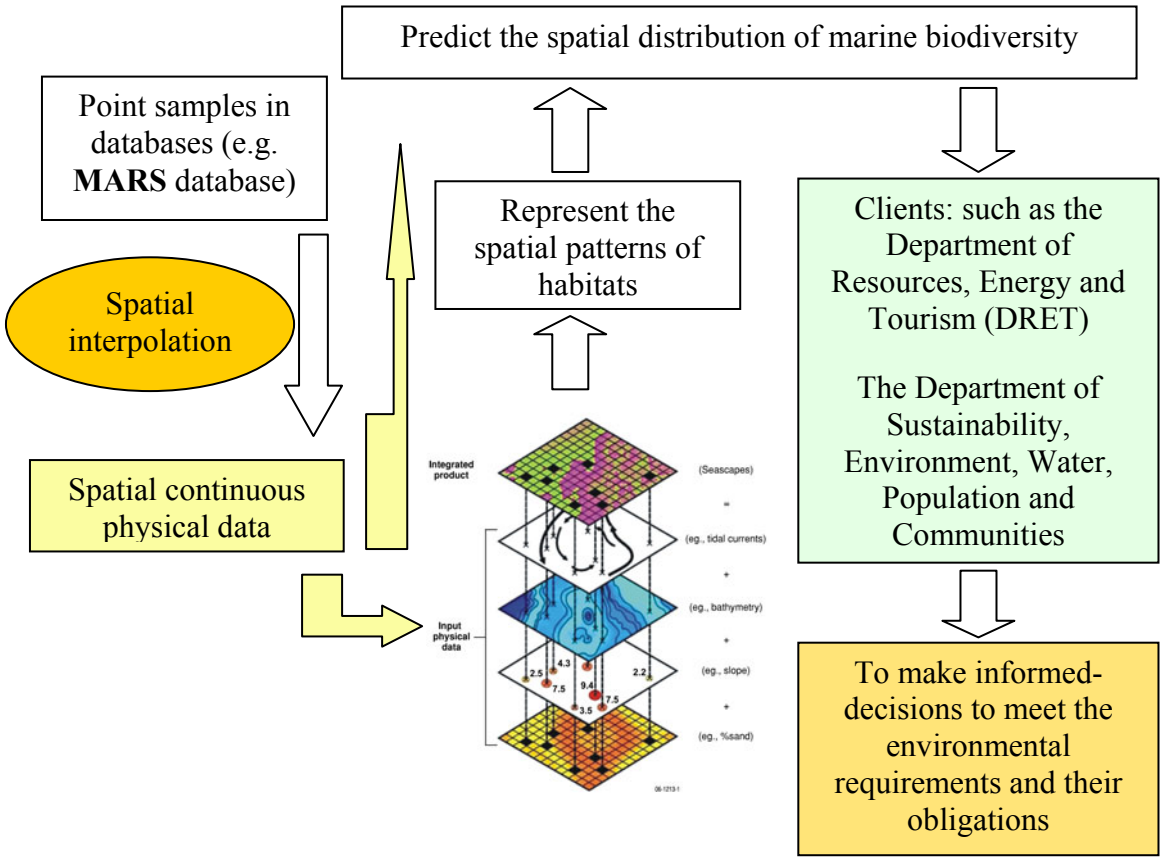


***Figure 1.2**. The role of spatially continuous data in predicting marine biodiversity.*

There are many factors that affect prediction accuracy of spatial interpolation methods (Li and Heap, 2011). Some of these factors have been examined in previous studies based on the seabed mud content from three regions on the Australian marine margin (Li et al., 2011a; Li et al., 2011b; Li et al., 2011c; Li et al., 2010). In these studies, we examined the influence of the following factors: 1) sample stratification by geomorphic provinces, 2) sample density (or sample size), 3) data variation, 4) search window size, 5) secondary information, 6) the combination of two machine learning methods and two commonly used spatial interpolation methods, 7) exclusion of least important secondary variables and 8) model averaging. Findings in these studies and advancement in spatial interpolation methods have formed a solid scientific foundation for studies to generate reliable spatial predictions and these methods can be applied to various environmental properties in both marine and terrestrial disciplines.

In this study, we aimed to identify the most appropriate methods for spatial interpolation of seabed sand content for the AEEZ using samples extracted in August 2010 from Geoscience Australia's Marine Samples Database (MARS) database (www.ga.gov.au/oracle/mars). The performance of 18 statistical methods/sub-methods for spatial interpolation was experimentally compared. This study examines:

- the performance of machine learning and geostatistical methods,
- the effects of input secondary variables on the performance of the methods,
- the accuracy of averaging the predictions of the most accurate methods,
- the effects of 'the number of variables randomly sampled as candidates at each split' (*mtry*), and
- the effect of search neighbourhood size on the best performing methods.

The input variables consisted of from six secondary variables only, to the combinations of these variables and some derived variables including the second and third orders and/or possible two-way interactions of these six predictors. These derived predictors could be regarded as redundant and irrelevant variables because they are correlated with these six predictors and because RF performs implicit variable selection (Okun and Priisalu, 2007) and can model complex interactions among predictors (Cutler et al., 2007; Diaz-Uriarte and de Andres, 2006; Okun and Priisalu, 2007). Inclusion of these derived variables, however, was hoped to compensate for the small number of predictors available for this study and thus increase the predictive accuracy.

We also visually examined and analysed the prediction patterns of the most accurate methods based on their prediction maps as suggested by previous findings (Li et al., 2011a; Li et al., 2011b; Li et al., 2011c). This study is an extension of a previous experiment by Li *et al*. (2010). It incorporates the findings and suggestions from our previous work (Li et al., 2011a; Li et al., 2010).

This record is presented in five chapters. Chapter 2 contains a brief description of study methods including data quality control, secondary variables, experimental design and data analysis. We analyse the experimental results, visually compare a few high performance methods based on the results, and illustrate their applications in Chapter 3. Chapter 4 discusses the findings and their implications. Finally, in Chapter 5, we summarise our findings and provide recommendations for the application of the methods tested.

Chapter 2. Methods

2.1. SAND CONTENT DATA AND DATA QUALITY CONTROL

The sand content data used in this study was extracted from the MARS database in August 2010. The accuracy and precision of attributes assigned to sample points in the MARS database varies, which can result in data noise (Li et al., 2010). Hence data quality control needs to be employed to clean relevant data noise.

2.1.1. Mars database

The MARS database was created in 2003 with the vision of collating all existing seabed sediment sample data for the Australian Marine Jurisdiction (AMJ) into a single publicly accessible database (http://www.ga.gov.au/oracle/mars). The content and structure of the database, its data sources and definitions of sediment data types have been detailed in Li et al. (2010). Preliminary data quality control was taken according to Geoscience Australian Data Standards, Validation and Release Handbook (Li et al., 2010). This resulted in a total of 14,204 surface sediment data points in the MARS database on 26 August 2010.

2.1.2. Data quality control

To remove possible data noise, a similar data quality control approach was adopted as detailed in Li et al. (2010). Seven criteria were used to clean the data:

- within the AEEZ,
- non-dredge,
- non-positive bathymetry,
- >=3 digits after decimal points in lat/long,
- with base depth <=5 cm,
- without duplicates, and
- samples with sand content.

Of these criteria, the criterion 4 selects samples with accurate location information. The criterion 6 removes duplicated samples. The remaining criteria are identical to those in Li et al. (2010). Given that geomorphological features of the Australian margin and adjacent seafloor created by Heap and Harris (2008) are categorically expressed bathymetry (Li et al., 2010), they were not considered in this study. Consequently geomorphology related criterion for data quality control was not considered in this study.

The sample size was gradually reduced from 14,204 to 6,810 after application of the data quality control criteria (Fig. 2.1). The spatial distribution of samples is illustrated in Figure 2.2. This final dataset of 6,810 samples was then used in this study.

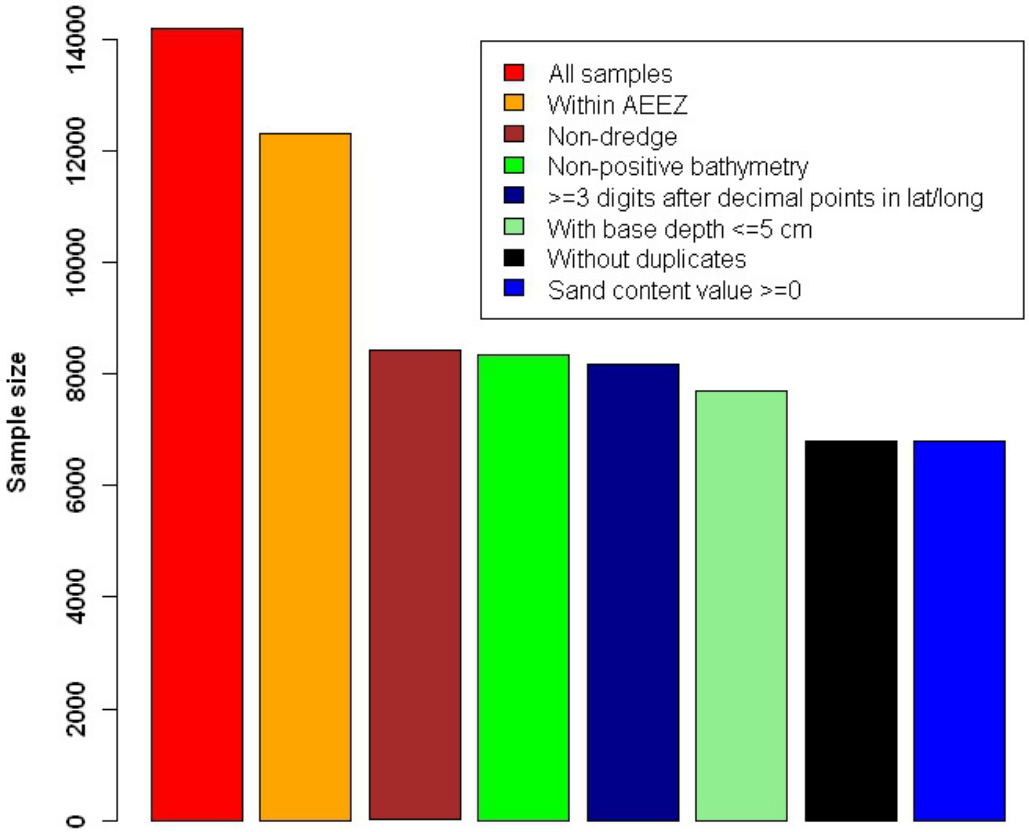


Figure 2.1. *Changes of sand sample size with data quality control criteria.*

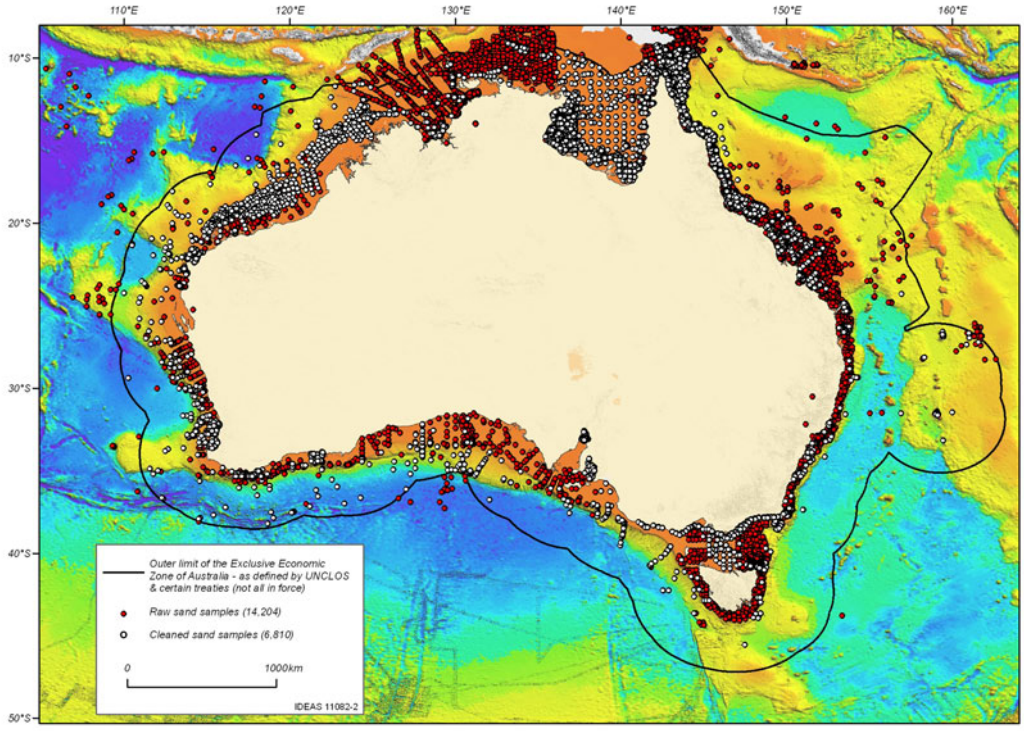


Figure 2.2. *Spatial distribution of sand samples in the AEEZ, with the original 'raw' (red) and 'cleaned' datasets (white).*

2.2. STUDY AREA

Samples in three regions (northwest, northeast and southwest) were selected from the AEEZ for this study (Fig. 2.3). The areas of the later two regions are the same as those used in previous studies. The north region was removed because its complex and contrasting coastal orientations provided misleading secondary information as discussed in Li et al. (2011a). The selection of the northwest region is because of its coastal orientation that contrasts the other two regions. The physical properties of these three regions are summarised in terms of area, orientation, geomorphic composition, and bathymetry (Tables 2.1, 2.2 and 2.3). Sample coverage (sample size and spatial distribution of samples) is also different in the three regions and their spatial distribution is also uneven, with most samples acquired from area near the shore (Fig. 2.4). Sample density is very low, varying from 0.5 to 1.9 samples per 1,000 km^2 (Table 2.1).

Samples in each region were then randomly divided into 10 datasets for cross validation based on the reasons detailed in Li et al. (2010), generating 30 datasets in total for this study.

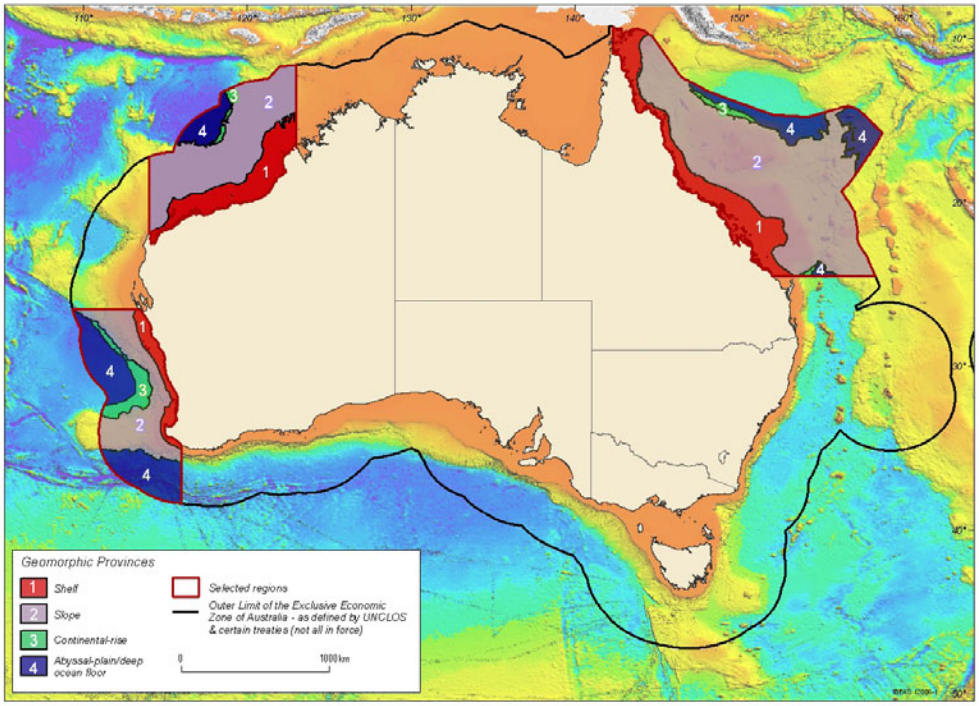


Figure 2.3. *Three regions selected for testing the performance of spatial interpolation methods from the AEEZ, including spatial distribution of geomorphic provinces.*

Table 2.1. Summary of features of each selected region*

REGION	ORIENTATION	BATHYMETRY (M)	AREA (KM^2)	SAMPLE NO	SAMPLE DENSITY (PER 1000 KM2)
Northwest	NE-SW	-5,976	553,422	574	1.0
Northeast	NW-SE	-4,948	1,165,078	2,157	1.9
Southwest	N-S	-6,269	500,915	262	0.5

** The difference in bathymetry and area in the northeast and southwest region from a previous study (Li et al., 2010) is because an updated version of bathymetry data (2009) was used in this study while 2005 version was used in the previous study and thus changes in bathymetry, coastline, the border of AEEZ and consequently areas are expected.*

Table 2.2. Sample size in each region for each geo-province

REGION	ABYSSAL-PLAIN/ DEEP OCEAN FLOOR	RISE	SHELF	SLOPE	TOTAL
NW	2	0	302	270	574
NE	0	0	2146	11	2157
SW	8	6	125	123	262

Table 2.3. Summary statistics of sand dataset in each region

REGION	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	CV (%)
NW	0	70.45	65.33	100	23.89	36.57
NE	1.5	65.5	63.49	100	21.59	34.01
SW	3.64	45.59	52.48	99.94	32.43	61.79

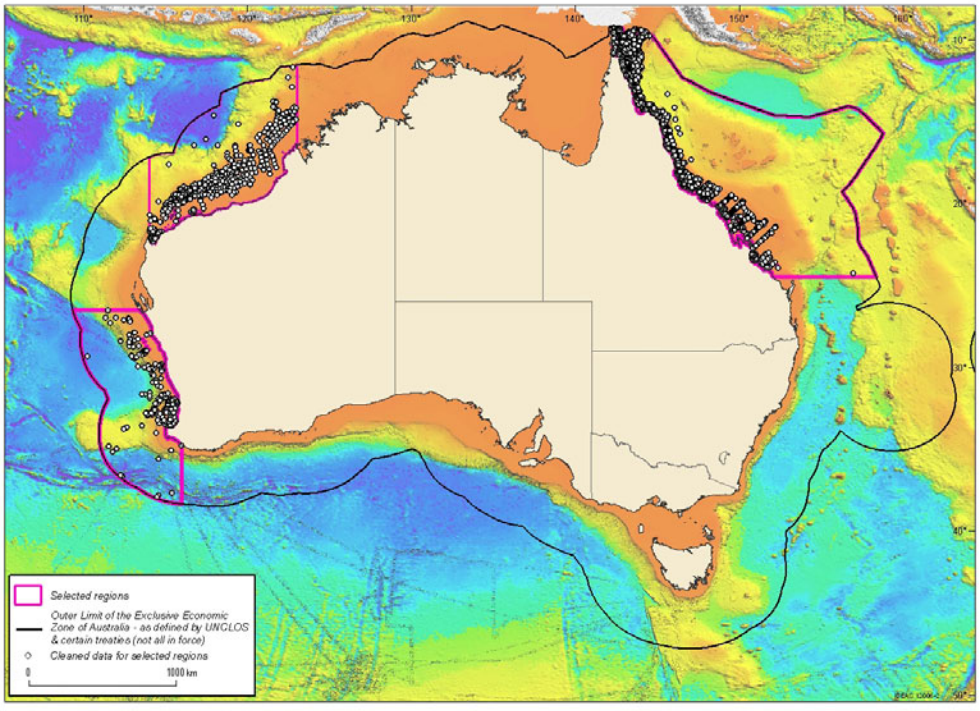
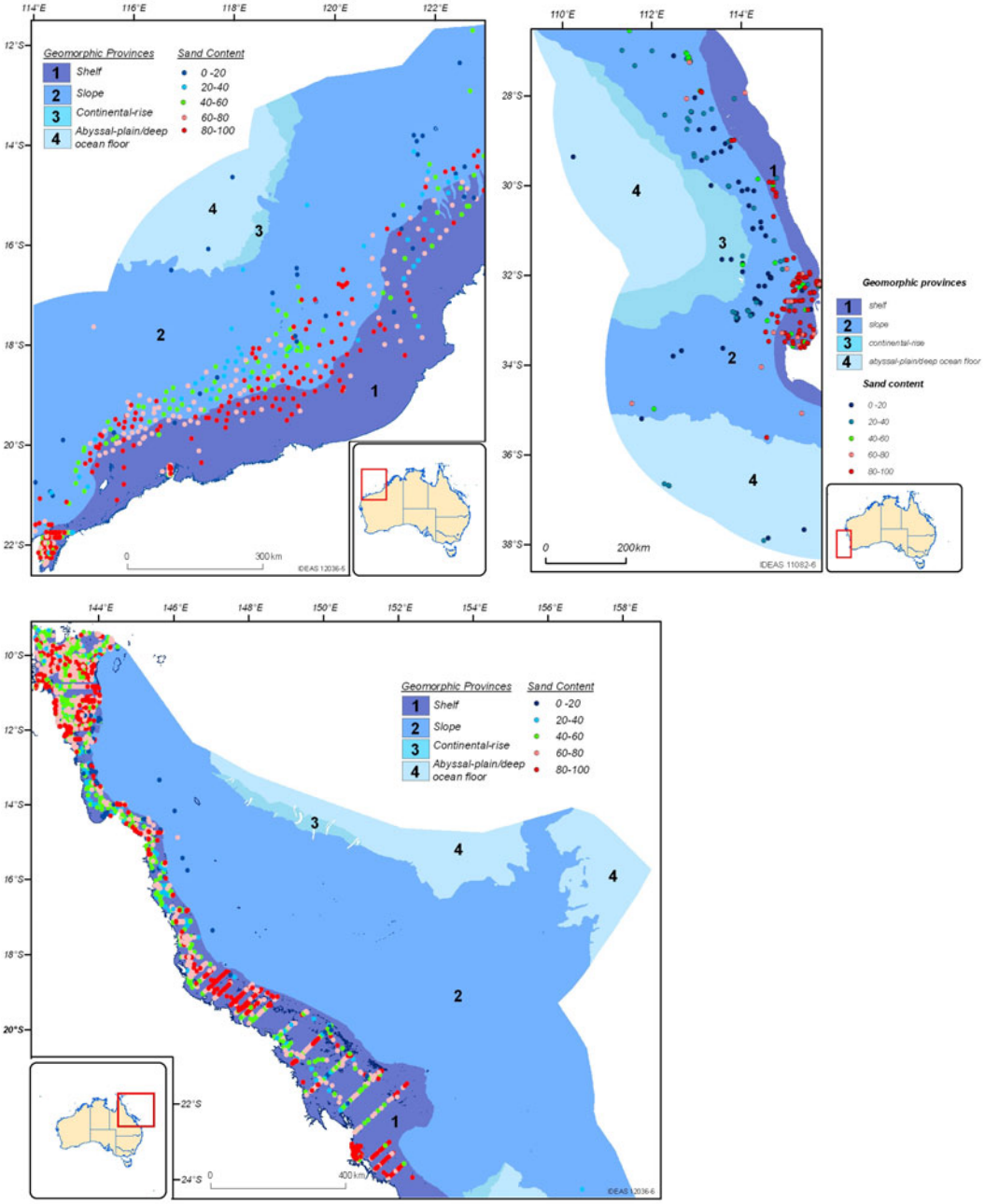


Figure 2.4. *Spatial distribution of samples with sand content for the three selected regions, including their occurrence and sand content in the geomorphic provinces.*

Fig. 2.4. (cont.):


Geomorphic Provinces
1 Shelf
2 Slope
3 Continental-rise
4 Abyssal-plain/deep ocean floor
Sand Content
0 -20
20-40
40-60
60-80
80-100
300 km
200km
400 km
Geomorphic provinces
Sand content

2.3. SECONDARY INFORMATION

The secondary variables used in this study are identical to those used for RF in our previous study (Li et al., 2010), except that 1) they were reproduced based on the upgraded bathymetry (2009 version) and 2) relief was considered as a secondary variable. In total, six secondary variables were used in this study: bathymetry, slope, distance to coast, relief, latitude and longitude. Relief represents the elevation (bathymetry) difference within a specified spatial window. The relief data was generated from the Australian bathymetry and topographic grid (2009 version) using the "focal range" statistical function in ArcGIS desktop. The spatial resolution of the relief data is 0.0025 decimal degrees (dd), which is the same as the input bathymetry and topographic grid. The window size was set at 4 by 4 pixels of a square shape (effectively created a square of 0.01 dd by 0.01 dd). To match a grid of 0.01 dd spatial resolution of other variables, we assigned the value of one of the 16 (0.0025 dd) cells to the sampled cell based on its centre location. This was done for all cells of a 0.01 dd grid in the AEEZ; and the extracted values were used for generating the predictions of sand content in the AEEZ.

The spatial distribution of bathymetry, slope, distance to coast and relief were illustrated in Figures 2.5-2.8. The spatial patterns of slope and relief were similar. The correlations among these variables (Table 2.4) showed that: bathymetry is highly correlated with distance to coast, and slope is correlated with relief, but there are some differences among the correlated variables (Figure 2.9), so they were all used as the secondary information in this study.

Table 2.4. The correlation coefficients among the secondary variables in each region, with sample size varying from 574 in the northwest region, 2,157 in the northeast region and 262 in the southwest region.

REGION	VARIABLES	BATHYMETRY	SLOPE	DISTANCE TO COAST	RELIEF
NW	Bathymetry	1	-0.3298	-0.8658	-0.3348
	Slope	-0.3298	1	0.3669	0.919
	Distance to coast	-0.8658	0.3669	1	0.3718
	Relief	-0.3348	0.919	0.3718	1
NE	Bathymetry	1	-0.1464	-0.6664	-0.1509
	Slope	-0.1464	1	0.1162	0.8975
	Distance to coast	-0.6664	0.1162	1	0.1075
	Relief	-0.1509	0.8975	0.1075	1
SW	Bathymetry	1	-0.0995	-0.7825	-0.1065
	Slope	-0.0995	1	-0.01	0.8465
	Distance to coast	-0.7825	-0.01	1	-0.0094
	Relief	-0.1065	0.8465	-0.0094	1

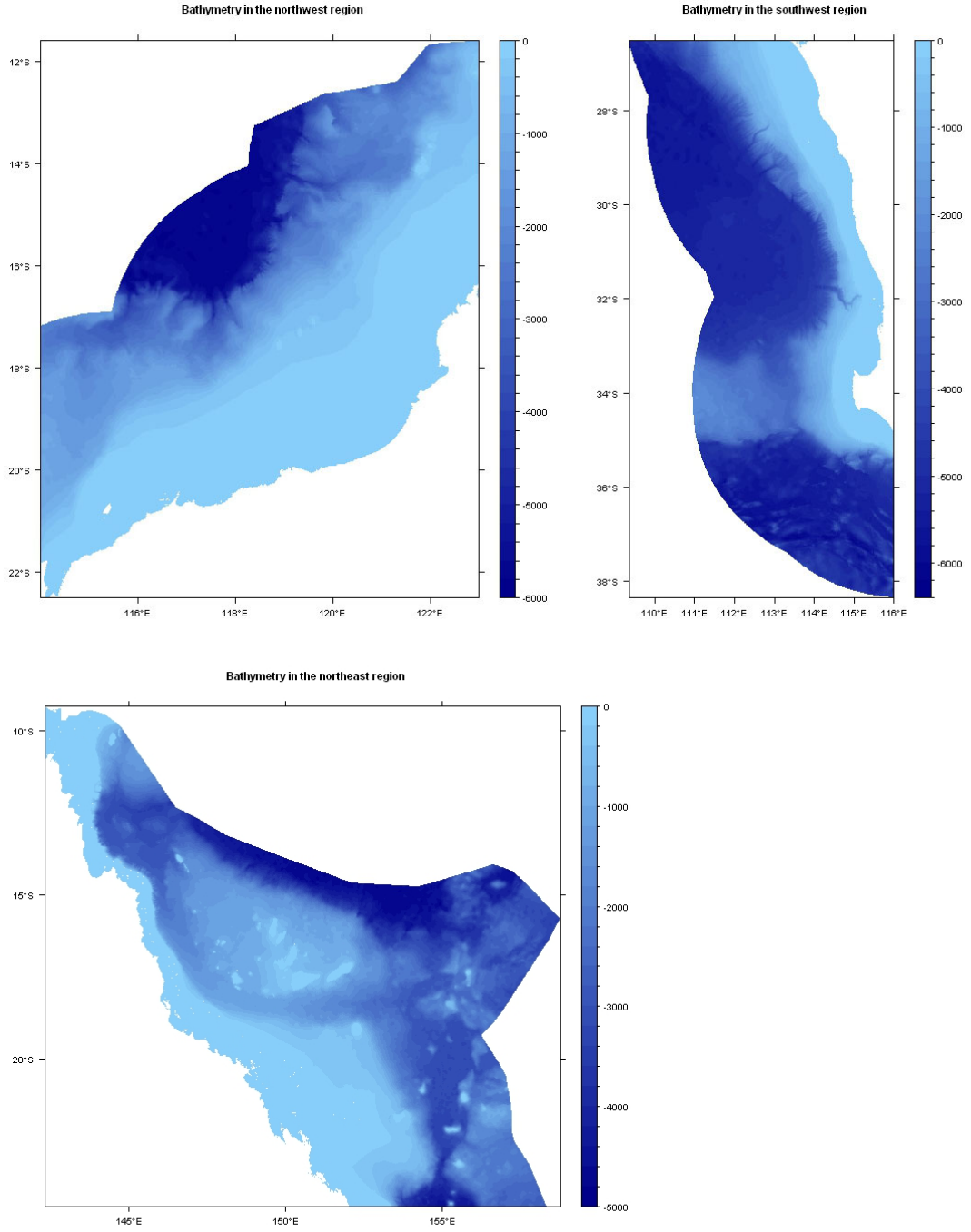


Figure 2.5. *Spatial pattern of bathymetry in the northwest, southwest and northeast regions.*

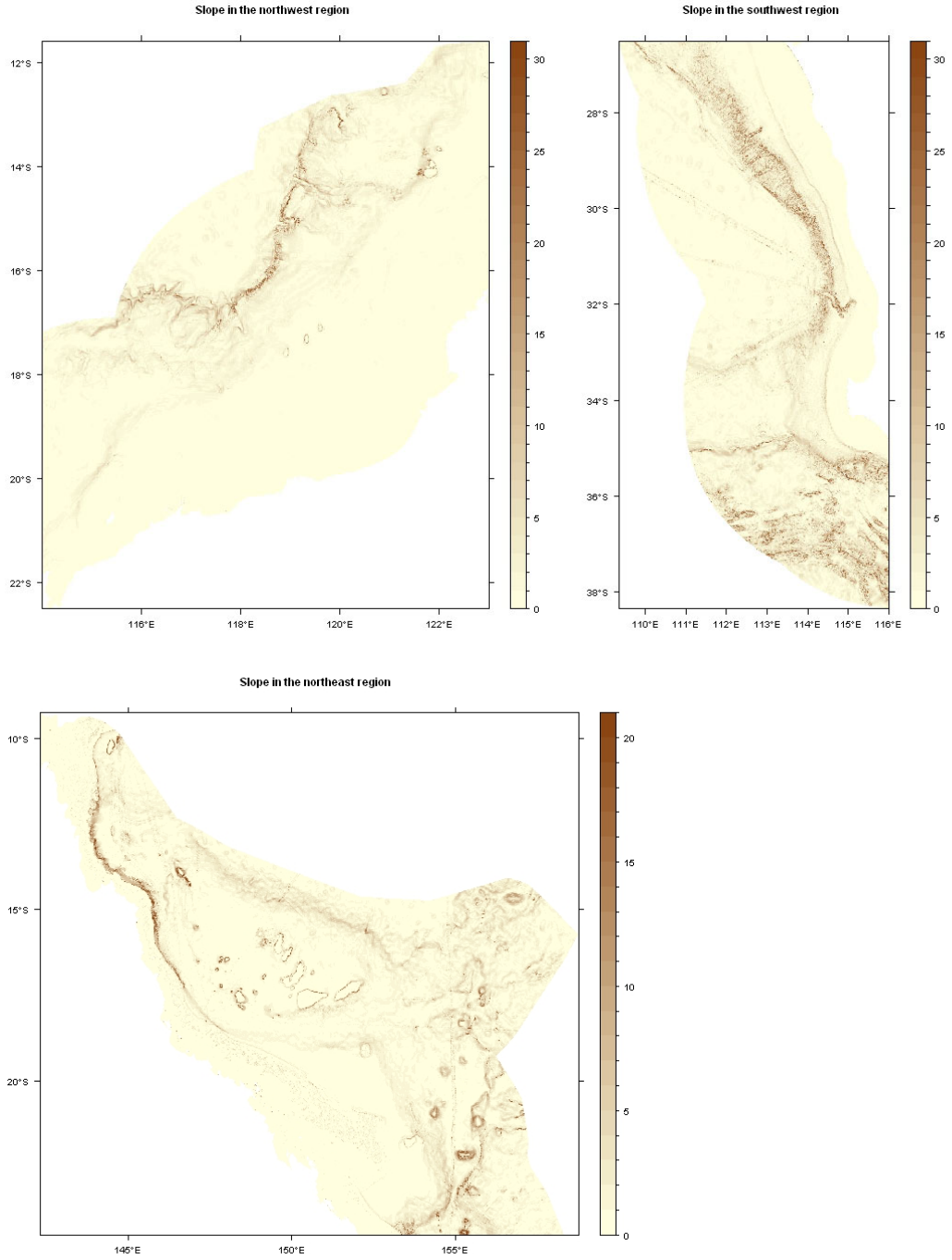


Figure 2.6. *Spatial pattern of slope in the northwest, southwest and northeast regions. To display the patterns of slope in the majority area, values over 30 were converted to 31 in the northwest and southwest regions and values over 20 were converted to 21 in the northeast region.*

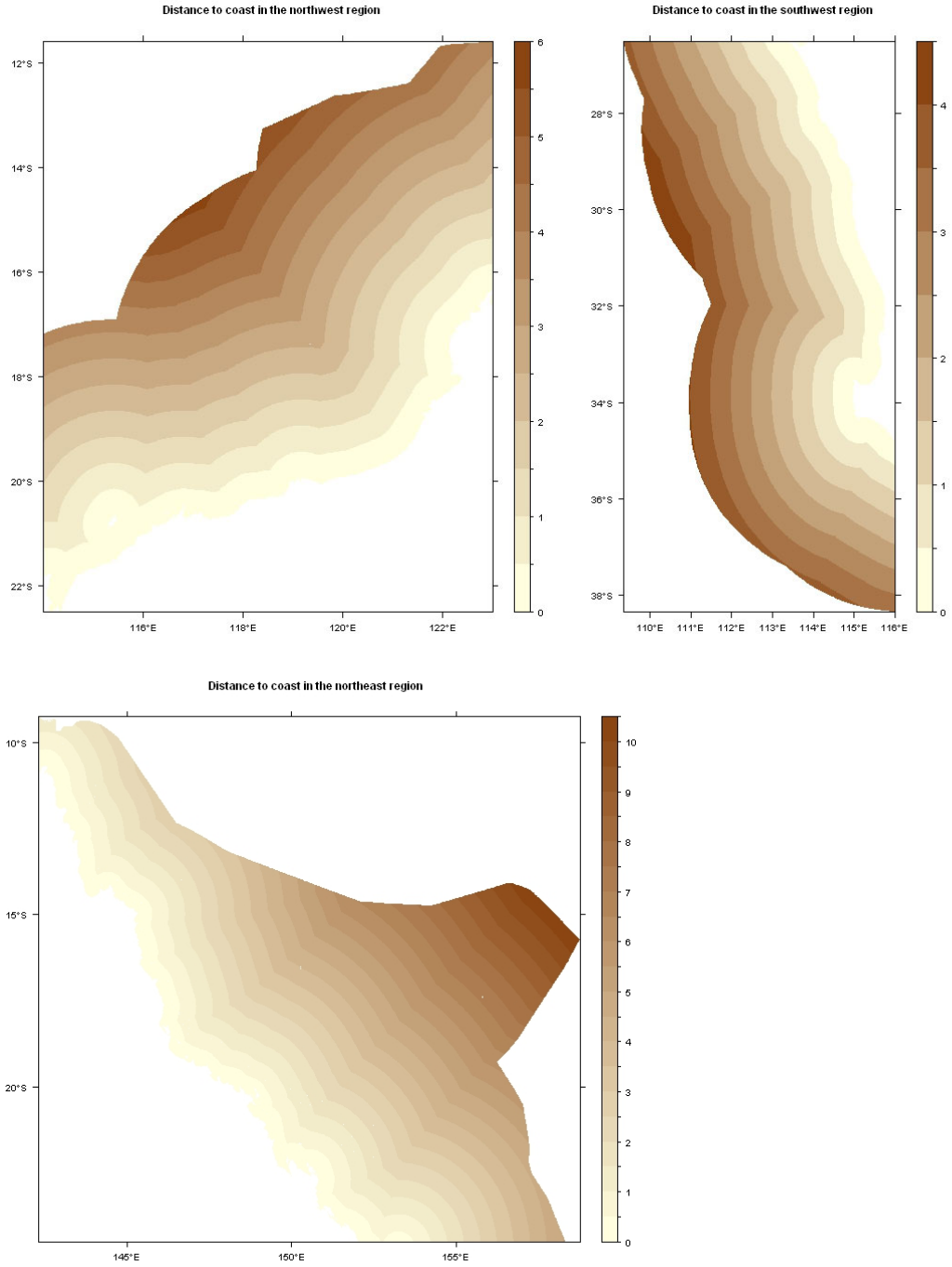


Figure 2.7. *Spatial pattern of distance to coast in the northwest, southwest and northeast regions.*

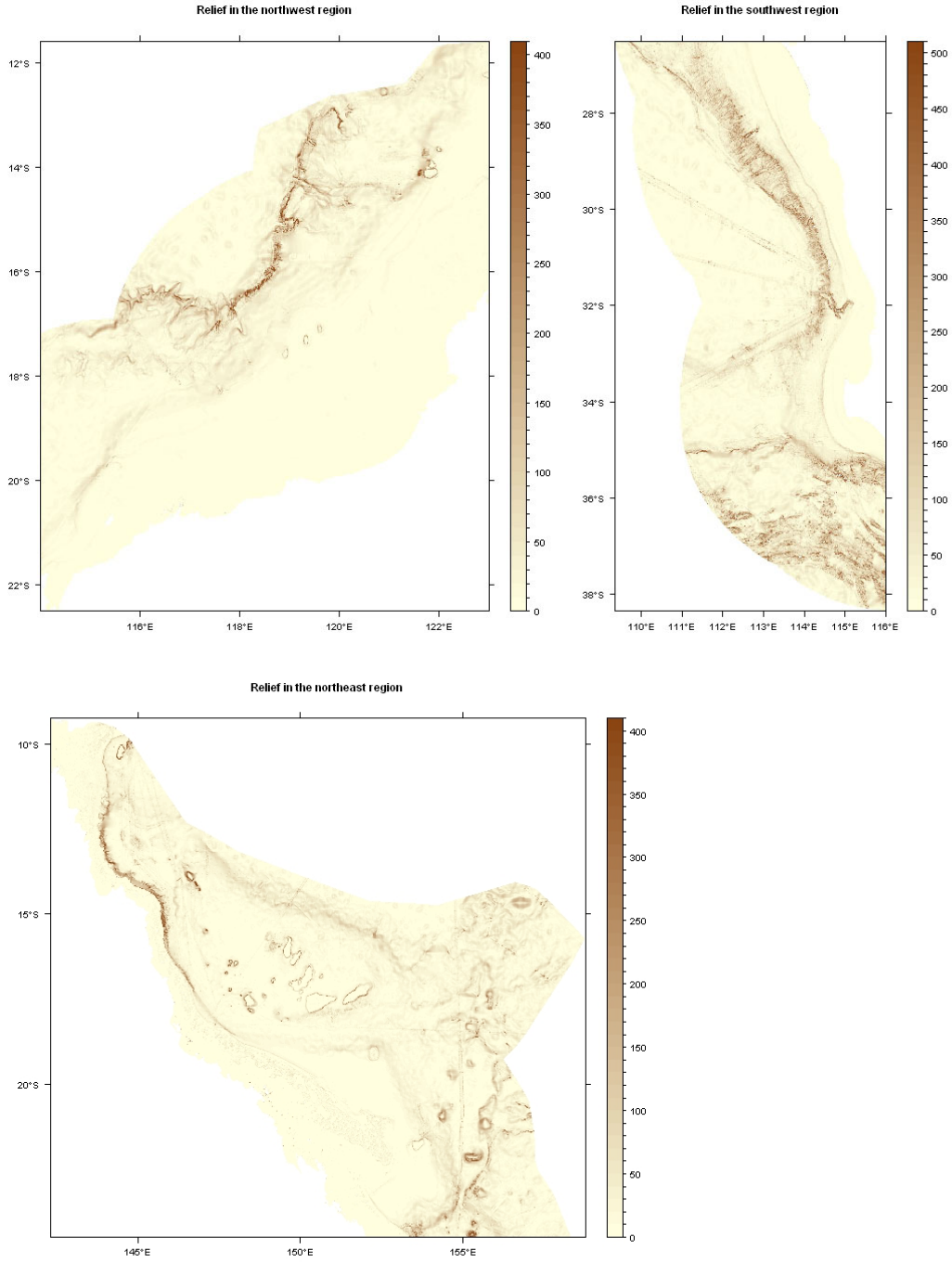


Figure 2.8. *Spatial pattern of relief in the northwest, southwest and northeast regions. To display the patterns of relief in the majority area, values over 400 were converted to 401 in the northwest and northeast regions and values over 500 were converted to 501 in the southwest region.*

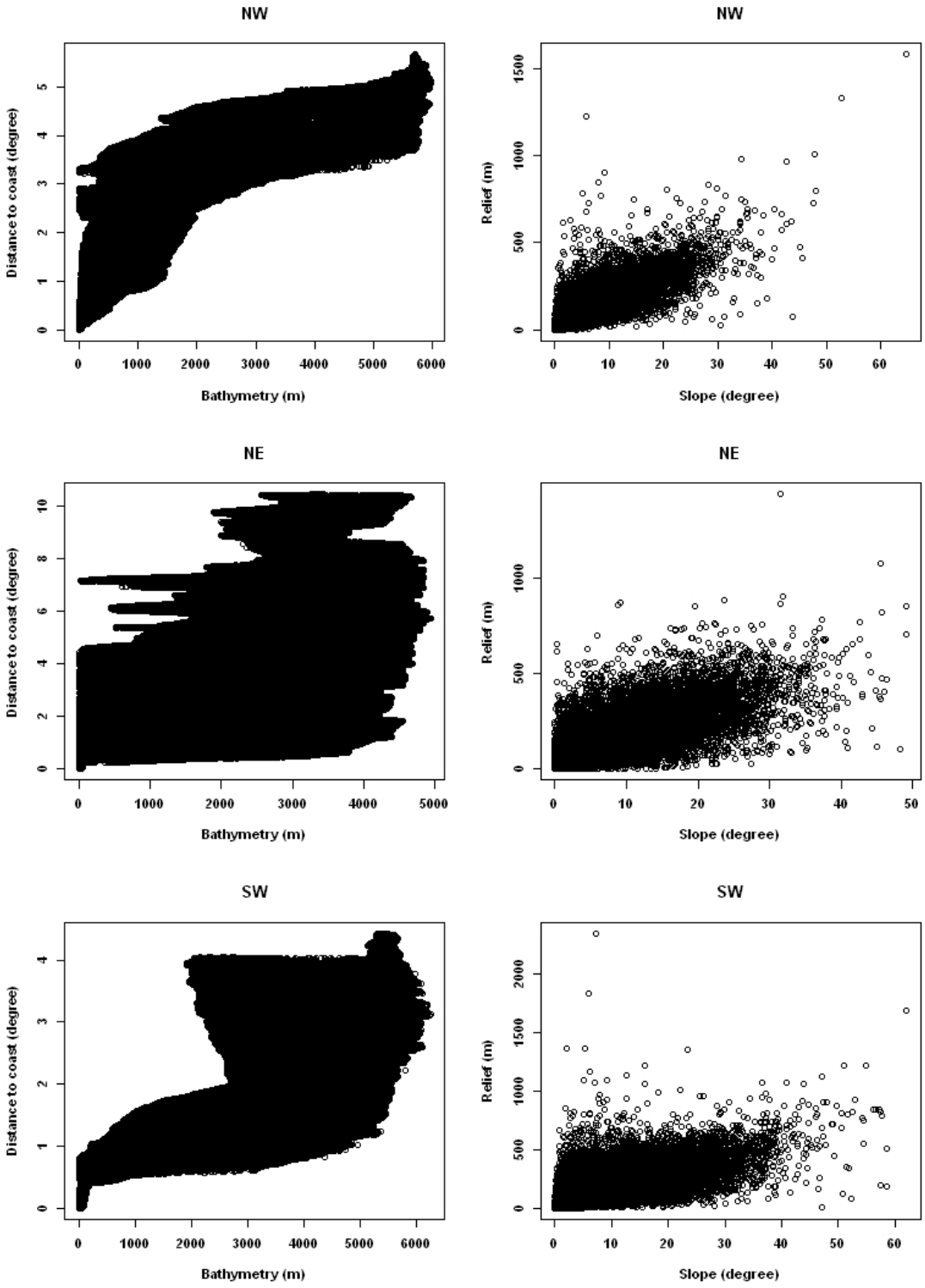


Figure 2.9. *The scatter plot of the highly correlated variables (i.e., bathymetry vs. distance to coast, and slope vs. relief) in the northwest, southwest and northeast regions.*

2.4. MACHINE LEARNING METHODS AND THE COMBINED METHODS

2.4.1. Methods

In this study, we firstly applied the following methods to the datasets in the three regions to test their predictive performance:

- Inverse distance squared (IDW2/IDS);
- Ordinary kriging (OK);
- Kriging with an external drift (KED);
- RF;
- support vector machine with a radial basis kernel (SVM);
- support vector machine with linear kernel (LSVM);
- Boosted decision tree (BDT);
- General Regression Neural Network (GRNN);
- the combination of RF and OK (RFOK);
- the combination of RF and IDS (RFIDS);
- the combination of SVM and OK (SVMOK);
- the combination of SVM and IDS (SVMIDS);
- the combination of LSVM and OK (LSVMOK);
- the combination of LSVM and IDS (LSVMIDS);
- the combination of BDT and OK (BDTOK);
- the combination of BDT and IDS (BDTIDS);
- the combination of GRNN and OK (GRNNOK); and
- the combination of GRNN and IDS (GRNNIDS).

These methods except BDT have been briefly described in previous studies (Li and Heap, 2008; Li et al., 2010). Please refer to these references for details. BDT is also called boosted regression trees (BRT) and briefly introduced; and GRNN is also further described in appendix A. Each of these methods was applied to the 30 datasets, which lead to the generation of 540 prediction datasets.

Then we tested the effects of input secondary variables on the predictive performance of RF and its combined methods. Three additional sets of input secondary variables were considered: 1) interactions among the secondary variables (i-variable), 2) interactions among the secondary variables excluding second and third order terms (i4-variables), and 3) 6 secondary variables (6-variables) (Table B.8).

- Random forest with interactions among the secondary variables (iRF);
- the combination of iRF and OK (iRFOK);
- the combination of iRF and IDS (iRFIDS);
- Random forest with interactions among the secondary variables excluding second and third order terms (i4RF);
- the combination of i4RF and OK (i4RFOK);
- the combination of i4RF and IDS (i4RFIDS);
- Random forest with 6 secondary variables (6RF);
- the combination of 6RF and OK (6RFOK); and
- the combination of 6RF and IDS (6RFIDS).

Each of these nine methods was applied to 30 datasets, which lead to the generation of 270 prediction datasets. In this study, the input variables for RF, RFIDS and RFOK were considered as the control to examine the effects of the choice of input secondary variables.

We further tested the effects of averaging the predictions of the most accurate methods and their combined methods on their predictive performance:

- the average of RFOK and RFIDS (RFOKRFIDS);
- the average of RF, RFOK and RFIDS (RFRFOKRFIDS);
- the average of SVMOK and SVMIDS (SVMOKSVMIDS);
- the average of SVM, SVMOK and SVMIDS (SVMSVMOKSVMIDS);
- the average of LSVMOK and LSVMIDS (LSVMOKSVMIDS);
- the average of LSVM, LSVMOK and LSVMIDS (LSVMSVMOKSVMIDS);
- the average of iRFOK and iRFIDS (iRFOKRFIDS);
- the average of iRF, iRFOK and iRFIDS (iRFRFOKRFIDS);
- the average of i4RFOK and i4RFIDS (i4RFOKRFIDS);
- the average of i4RF, i4RFOK and i4RFIDS (i4RFRFOKRFIDS);
- the average of 6RFOK and 6RFIDS (6RFOKRFIDS); and
- the average of 6RF, 6RFOK and 6RFIDS (6RFRFOKRFIDS).

In this study, we averaged the predictions of the most accurate methods and their combined methods according to discussion in Li et al. (2011a). Each of the 12 methods was applied to 30 datasets, which led to the generation of 360 prediction datasets.

We also examined the effects of the choice of *mtry* on the prediction accuracy of RF in the northwest region:

- RandomForest with *mtry* being 4 (4mRF);
- the combination of RF with *mtry* being 4 (4mRFOK);
- the combination of RF with *mtry* being 4 (4mRFIDS);
- the average of 4mRFOK and 4mRFIDS (4mRFOKRFIDS); and
- the average of 4mRF, 4mRFOK and 4mRFIDS (4mRFRFOKRFIDS).

These applications resulted in 50 prediction datasets.

Finally we tested the effect of search neighbourhood size on the best performing methods, RFOK, RFIDS, RFOKRFIDS, RFRFOKRFIDS, 4mRFOK, 4mRFIDS, 4mRFOKRFIDS, and 4mRFRFOKRFIDS, with search neighbourhood size varying from 4 to 25 and all samples, leading to 23 sizes.

For RFOK, RFIDS, RFOKRFIDS, RFRFOKRFIDS, an additional 2640 prediction datasets were generated. Given that 4mRF, 4mRFOK, 4mRFIDS, 4mRFOKRFIDS, and 4mRFRFOKRFIDS were only applied to the northwest region, 880 additional prediction datasets were generated.

2.4.2. Statistical and mathematical modelling

Statistical and mathematical modelling is an essential and key section of this study. Given its technical nature, we put the main body of this section in the Appendix B for

readers interested. This Appendix covers a number of issues: identification of appropriate data transformation for sand content and relevant secondary variables; analyses of the correlations between sand content and secondary variables; variogram modelling, anisotropy and variogram model selection; and model and parameter specification of all relevant methods. The modelling procedures were outlined in Figure 2.10. All modelling work was conducted in R (R Development Core Team, 2010), except the predictions and residuals of BDT and GRNN that were implemented in DTREG.

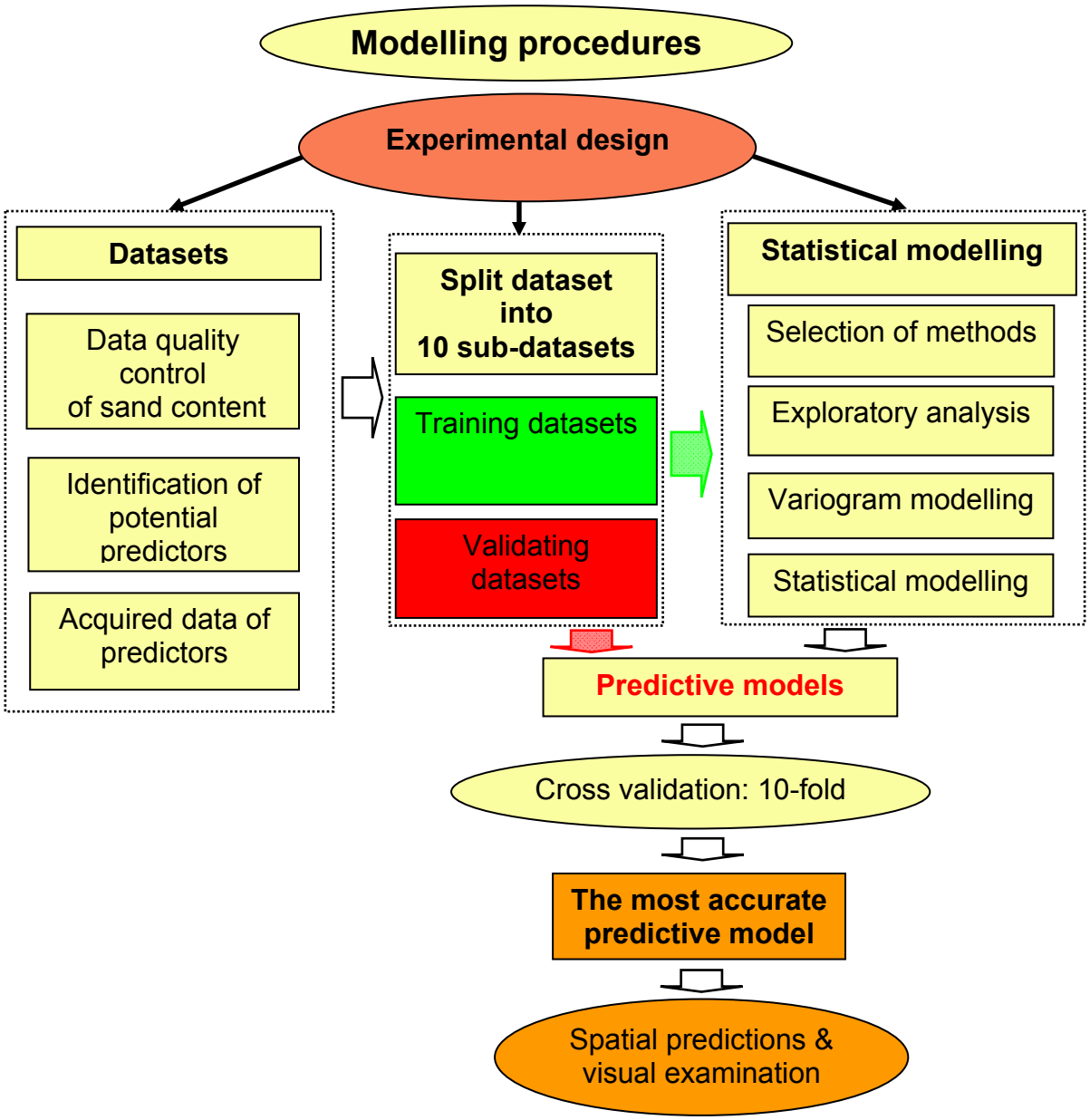


Figure 2.10. *Modelling procedures adopted in this study.*

2.5. ASSESSMENT OF METHOD PERFORMANCE

Ten-fold cross-validation was used to validate the predictions of each method (see Figure 2.10). The validation procedure is detailed in Li *et al*. (2010). In total, 4,740 prediction datasets were generated in this study, which formed a basis for assessing the performance of the methods under various testing conditions. Summary statistics of these datasets were derived for each method, window search size and region as in Appendix C. Relative mean absolute error (RMAE) and relative root mean squared error (RRMSE) (Li and Heap, 2008, 2011) were used to measure the prediction error.

Chapter 3. Results

3.1. BEST PERFORMING METHODS IN EACH REGION

The performance of the methods was dependent on the study region in which they were applied. The performance of the methods showed that RFIDS, KED, and BDTIDS were the most accurate in terms of both RRMSE and RMAE, while RF and RFOK were the most accurate in terms of RRMSE in the northwest region (Fig. 3.1). In the northeast region, RFIDS and RFOK were the most accurate in terms of both RRMSE and RMAE, while RF was the most accurate in terms of RRMSE (Fig. 3.2). RFIDS, RFOK, RF and BDTIDS were the most accurate methods in the southwest region (Fig. 3.3). All the remaining methods performed relatively poorly compared to the control (i.e., IDS) in terms of either of RMAE or RRMSE or both.

The predictive accuracy of the best performing methods varied with regions (Figs. 3.1-3.3). In the northwest region, the predictive errors ranged from 22% to 22.5% for RMAE and 30% to 31% for RRMSE; in the northeast region, they varied from 18% to 18.5% for RMAE and 25% to 25.5% for RRMSE; while in the southwest region, they changed from 26% to 26.5% for RMAE and from 35.5% to 36.5% for RRMSE. The predictive accuracy was the highest in the northeast region, followed by the northwest region, and the least in the southwest region.

The predictive accuracy of the support machine learning methods varied considerably and also changed with regions; and their combination with IDS or OK always improved the prediction accuracy. LSVM was the least accurate method in all three regions, but its combination with IDS performed relatively well and was one of the most accurate methods in terms of RRMSE in the northwest region.

The predictive accuracy of the GRNN and BDT varied considerably with regions. Their combination with IDS or OK generally improved the prediction accuracy. The predictive accuracies of GRNN and its combination with either IDS or OK were always lower than that of IDS.

The results indicate that RFIDS and RFOK are, on average, more accurate than other methods in all three regions.

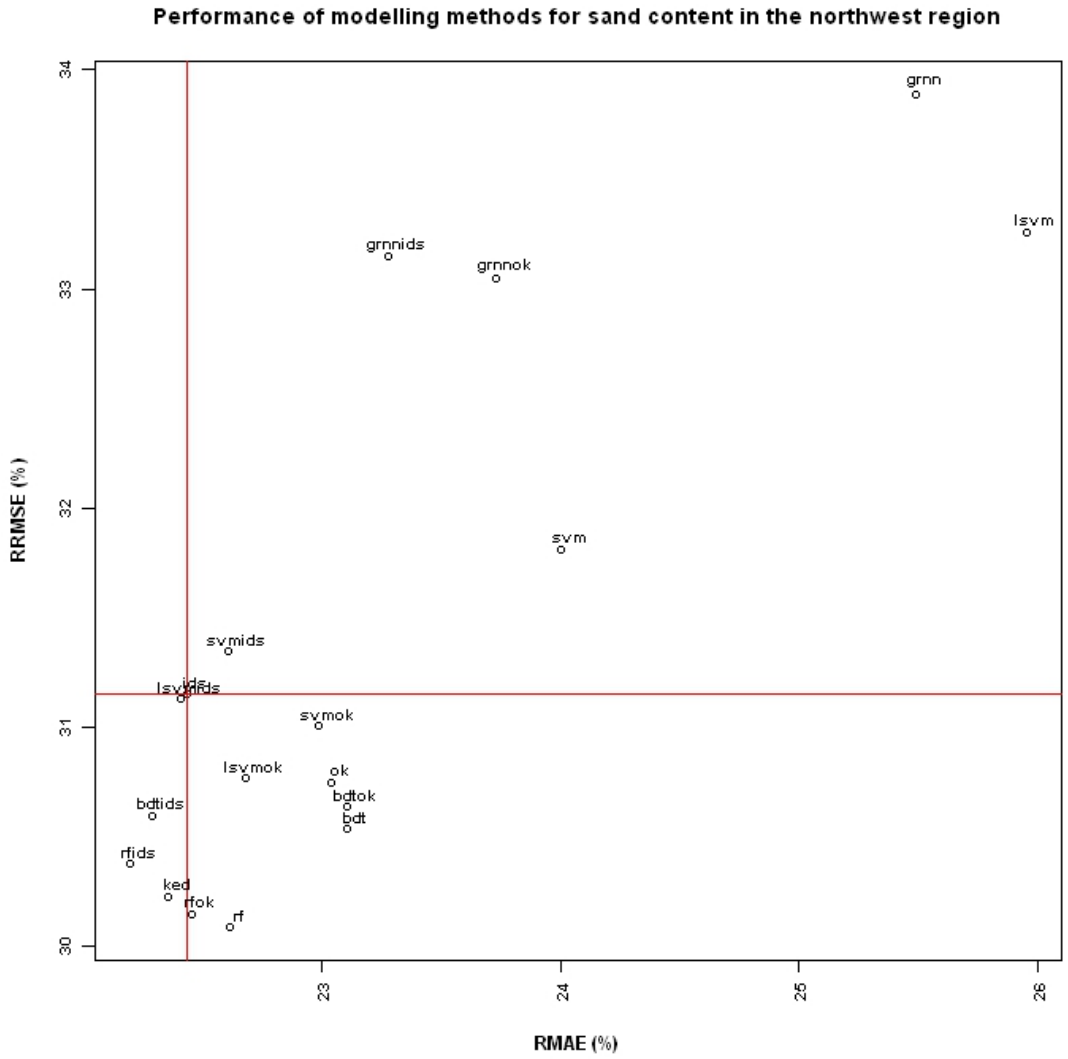


Figure 3.1. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of modelling methods for sand content in the northwest region. The horizontal and vertical lines (red) indicate the accuracy of the control (IDS).*

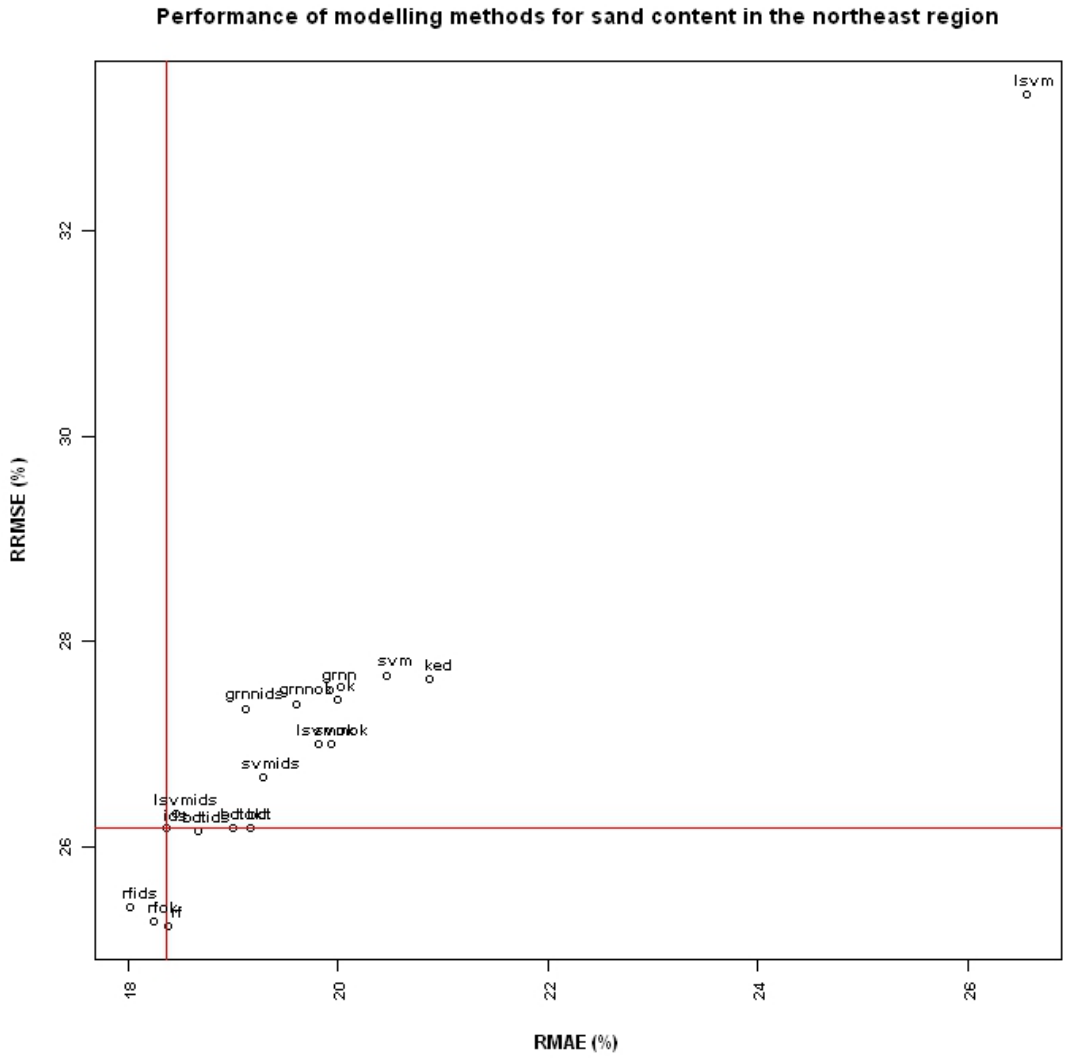


Figure 3.2. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of modelling methods for sand content in the northeast region. The horizontal and vertical lines (red) indicate the accuracy of the control (IDS).*

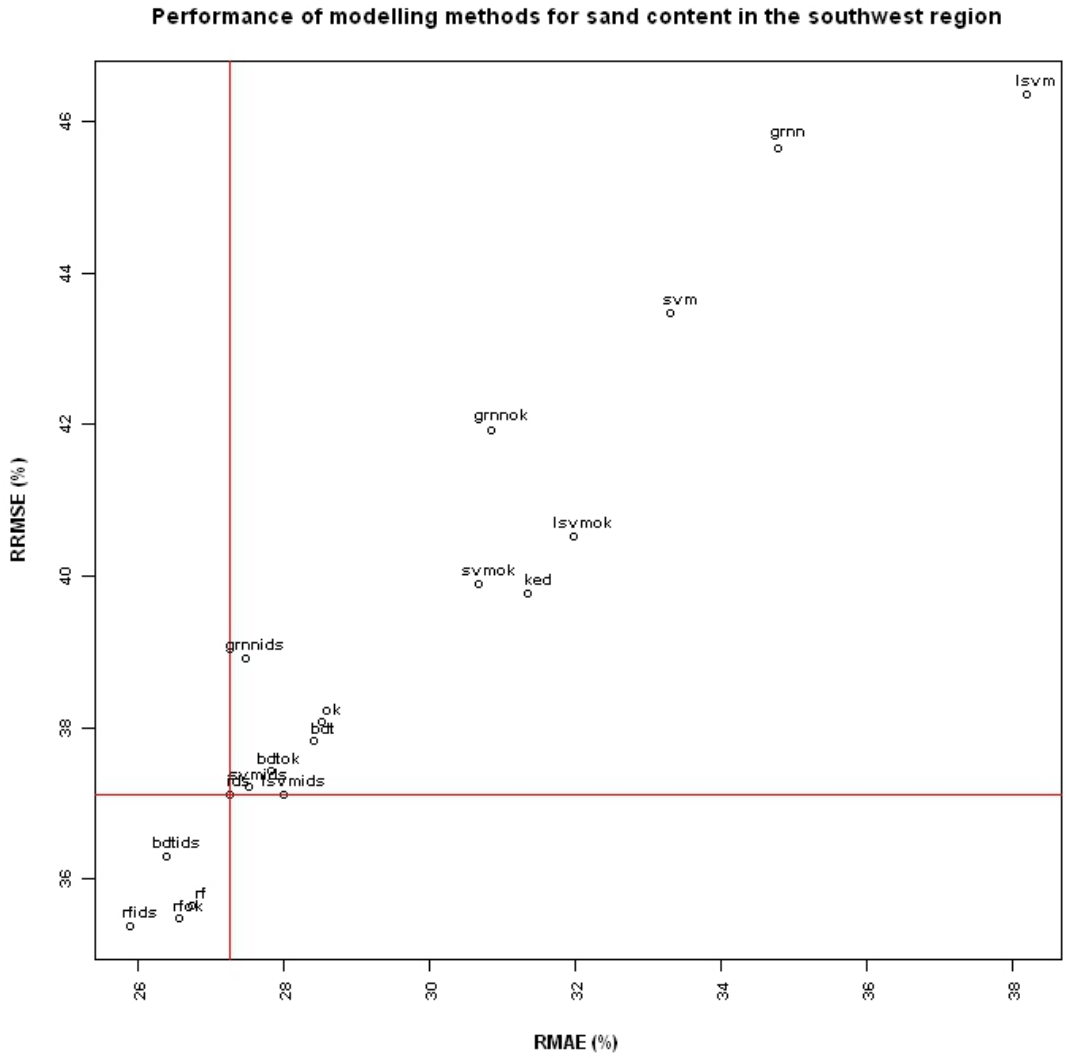


Figure 3.3. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of modelling methods for sand content in the southwest region. The horizontal and vertical lines (red) indicate the accuracy of the control (IDS).*

3.2. EFFECTS OF INPUT SECONDARY VARIABLES

The input secondary variables affected the predictive error, varying with the method used (RF, RFOK and RFIDS) and study region (Figs. 3.4-3.6). In the northwest region, the predictive errors ranged from 22% to 24% in terms of RMAE and from 30% to 31% in terms of RRMSE (Fig. 3.4). The range increased from RFIDS, RFOK to RF under different choice of input secondary variables. Methods with i4-variables produced predictions with the least accuracy, followed by methods with i-variables, and then methods with 6-variables and with control performed the best. Generally speaking, RFIDS performed slightly better than RFOK, and RFOK outperformed RF under different choice of input secondary variables. Overall, RFIDS was the most accurate method in terms of RMAE, 6RFOK was the most accurate method in terms of RRMSE, while RFIDS, 6RFIDS and iRFIDS were the most accurate method in terms of both of RRMAE and RRMSE.

In the northeast region, the predictive errors ranged from 18-19.5% in terms of RMAE and from 25-26% in terms of RRMSE (Fig. 3.5). The ranges of predictive errors also increased from RFIDS, RFOK to RF under different choice of input secondary variables. As was observed in the northwest region, methods with i4-variables performed worst, followed by methods with i-variables, then methods with 6-variables, and the control methods performed the best in the northeast region. RFIDS also performed slightly better than RFOK that again outperformed RF under different choice of input secondary variables. Overall RFIDS was the most accurate method in terms of RMAE, RF was the most accurate method in terms of RRMSE, while RFIDS, RFOK, 6RFIDS and iRFIDS were the most accurate methods in terms of both of RRMAE and RRMSE.

In the southwest region, the predictive errors ranged from 25.5-28% in terms of RMAE and from 35-36.5% in terms of RRMSE (Fig. 3.6). The ranges also increased from RFIDS, RFOK to RF under different choice of input secondary variables. Methods with 6-variables performed worst, followed by the control methods, and then methods with i-variables, and methods with i4-variables performed the best. As in the other two regions, RFIDS performed better than RFOK that again outperformed RF under different choice of input secondary variables. Overall there is an apparent interactive effect between methods and the choice of input secondary variables and all methods except 6RF outperformed the control method. Of which, i4RFIDS, iRFIDS, RFIDS and 6RFIDS were the most accurate.

The choice of input secondary variables affected predictive errors of random forest and its combined methods. Overall, of the 36 combinations of input secondary variables (4 levels), method (3 levels) and region (3 levels), RFIDS, 6RFIDS and RFOK were among the most accurate methods in all three regions.

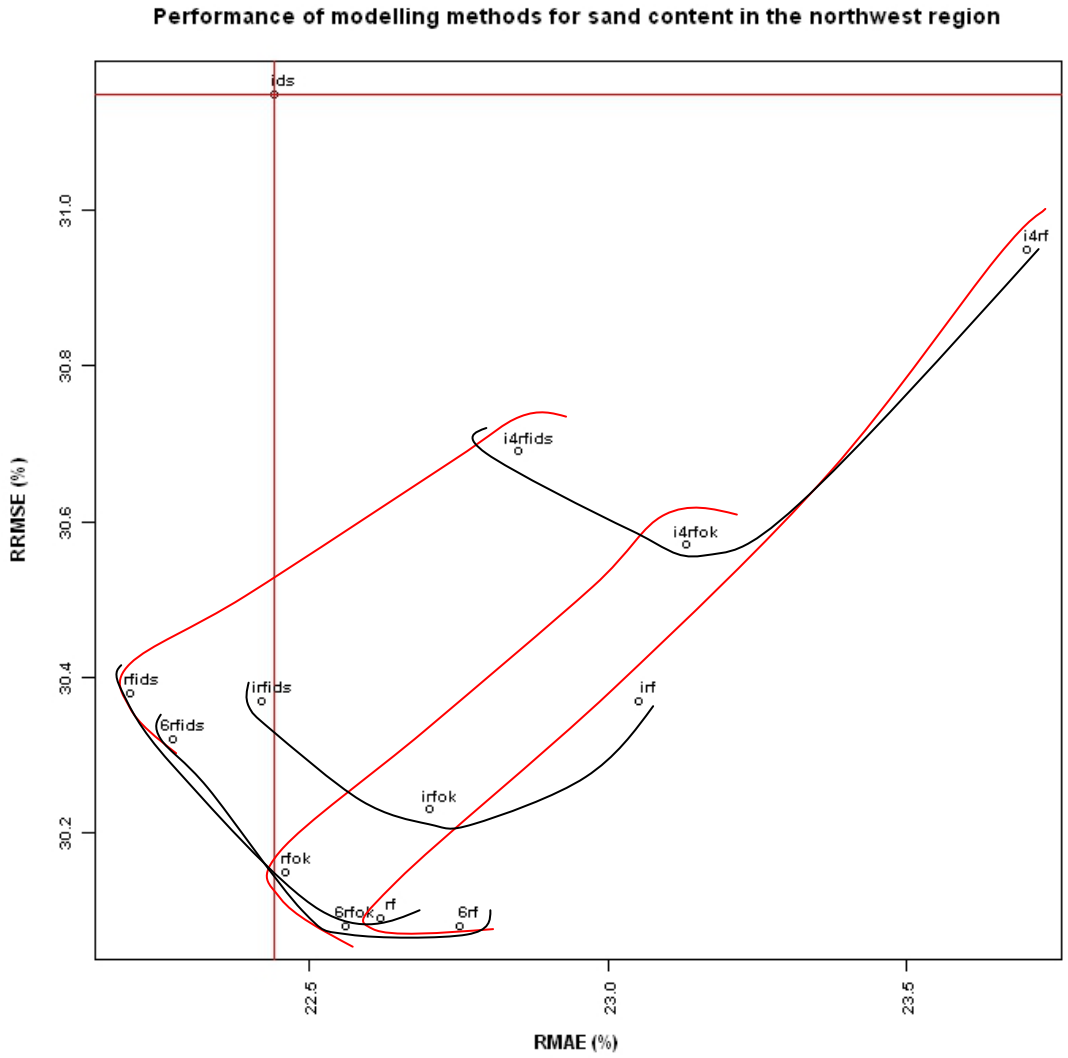


Figure 3.4. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of modelling methods with varying input secondary variables for sand content in the northwest region. The horizontal and vertical lines (red) indicate the accuracy of the control (IDS). The red lines group the same method with different input secondary variables; the black lines group different methods with the same input secondary variables.*

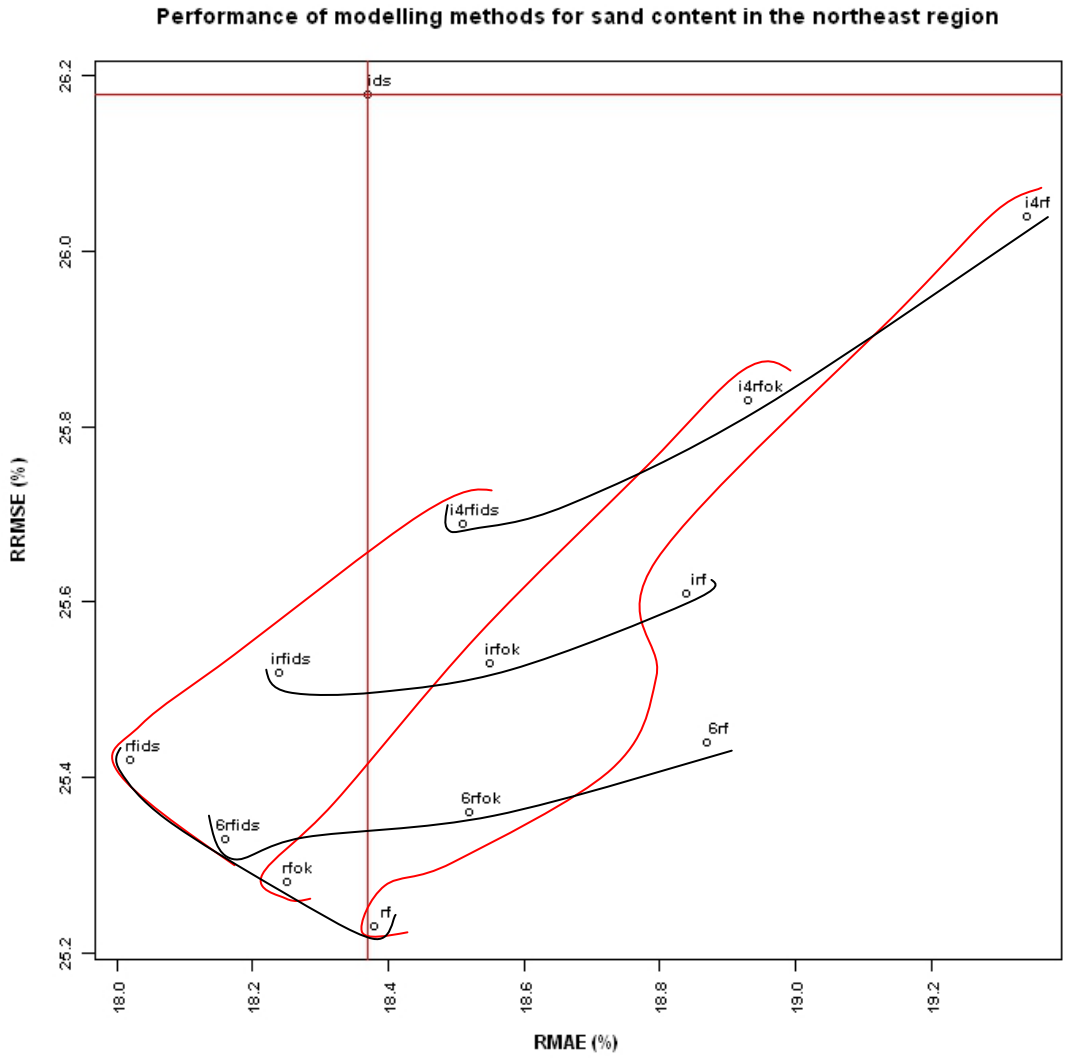


Figure 3.5. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of modelling methods with varying input secondary variables for sand content in the northeast region. The horizontal and vertical lines (red) indicate the accuracy of the control (IDS). The red lines group the same method with different input secondary variables; the black lines group different methods with the same input secondary variables.*

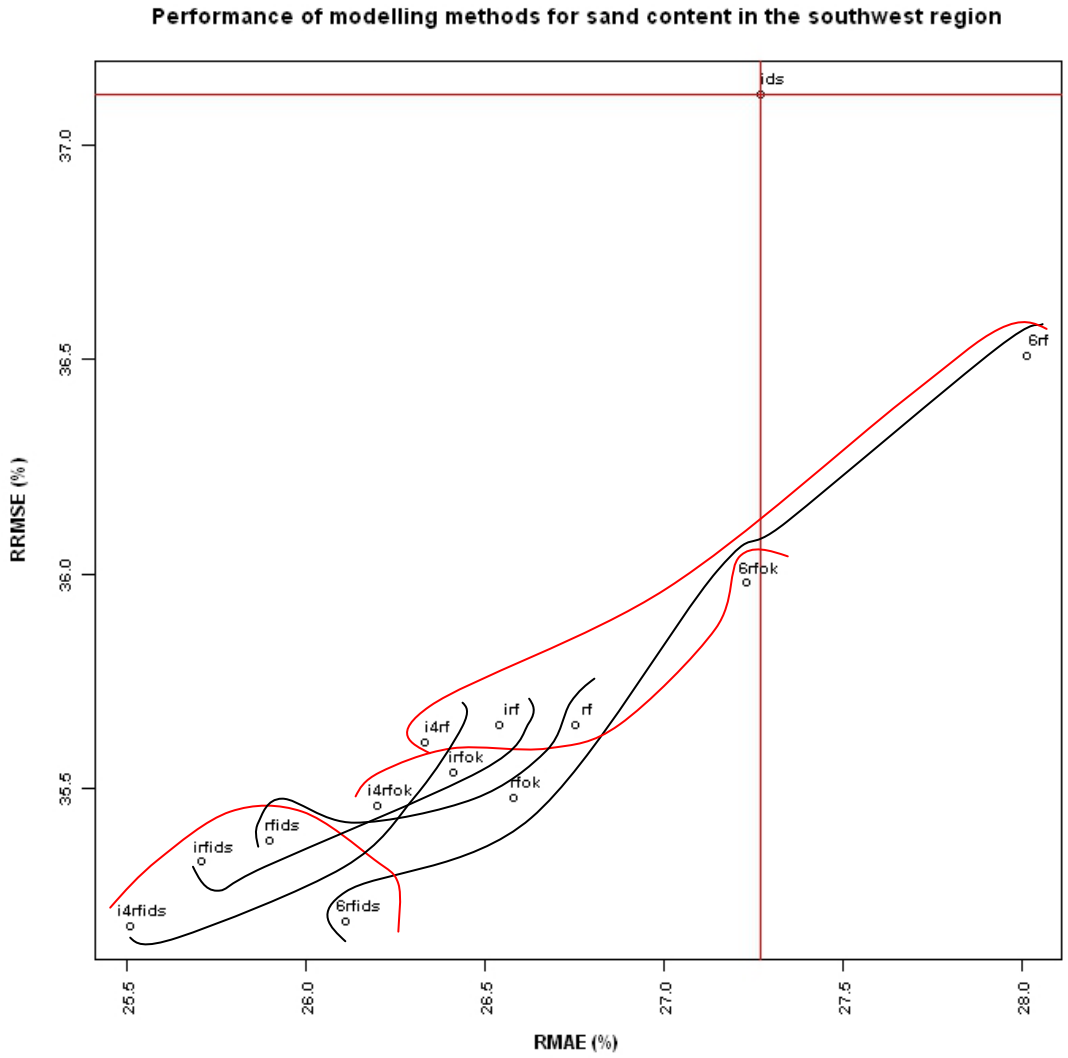


Figure 3.6. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of modelling methods with varying input secondary variables for sand content in the southwest region. The horizontal and vertical lines (red) indicate the accuracy of the control (IDS). The red lines group the same method with different input secondary variables; the black lines group different methods with the same input secondary variables.*

3.3. EFFECTS OF AVERAGING THE PREDICTIONS OF THE MOST ACCURATE METHODS

The effects of averaging the predictions of the two or three most accurate methods (*i.e.*, RF, RFOK and RFIDS) changed with regions in terms of the predictive accuracy (Figs. 3.7-3.9). The choice of two or three most accurate methods was mainly based on the results from section 3.1; and the results of section 3.2 were also considered. SVM and LSVM and their combinations were also considered to examine whether averaging the predictions of these methods can improve their accuracy.

In the northwest region, averaging the predictions of two or three most accurate methods like RFOKRFIDS, RFRFOKRFIDS, 6RFOKRFIDS and 6RFRFOKRFIDS marginally improved the prediction accuracy in terms of RRMSE in comparison with RFIDS, the most accurate identified in previous two sections, while they were slightly less accurate than RFIDS in terms of RMAE. All other averaged methods were less accurate than RFIDS.

In the northeast region, averaging the predictions of two or three most accurate methods like RFOKRFIDS, RFRFOKRFIDS, 6RFOKRFIDS, 6RFRFOKRFIDS and iRFOKRFIDS marginally improved the prediction accuracy in terms of RRMSE in comparison with RFIDS, while they were slightly less accurate than RFIDS in terms of RMAE. All other averaged methods were less accurate than RFIDS.

In the southwest region, averaging the predictions of two or three most accurate methods like i4RFOKRFIDS, i4RFRFOKRFIDS, iRFOKRFIDS were slightly more accurate than RFIDS in terms of both RMAE and RRMSE. RFOKRFIDS, RFRFOKRFIDS, 6RFOKRFIDS, and iRFRFOKRFIDS marginally improved the prediction accuracy in terms of RRMSE in comparison with RFIDS, while they were slightly less accurate than RFIDS in terms of RMAE. All other averaged methods were less accurate than RFIDS.

Overall, model averaging marginally improved the prediction accuracy in the southwest region, while in other regions no apparent effects were observed. RFIDS, RFOKRFIDS and RFRFOKRFIDS performed better than other methods in all three regions in terms of both RMAE and RRMSE. This suggests that the effects of model averaging depend on methods averaged, region and also on error measurement.

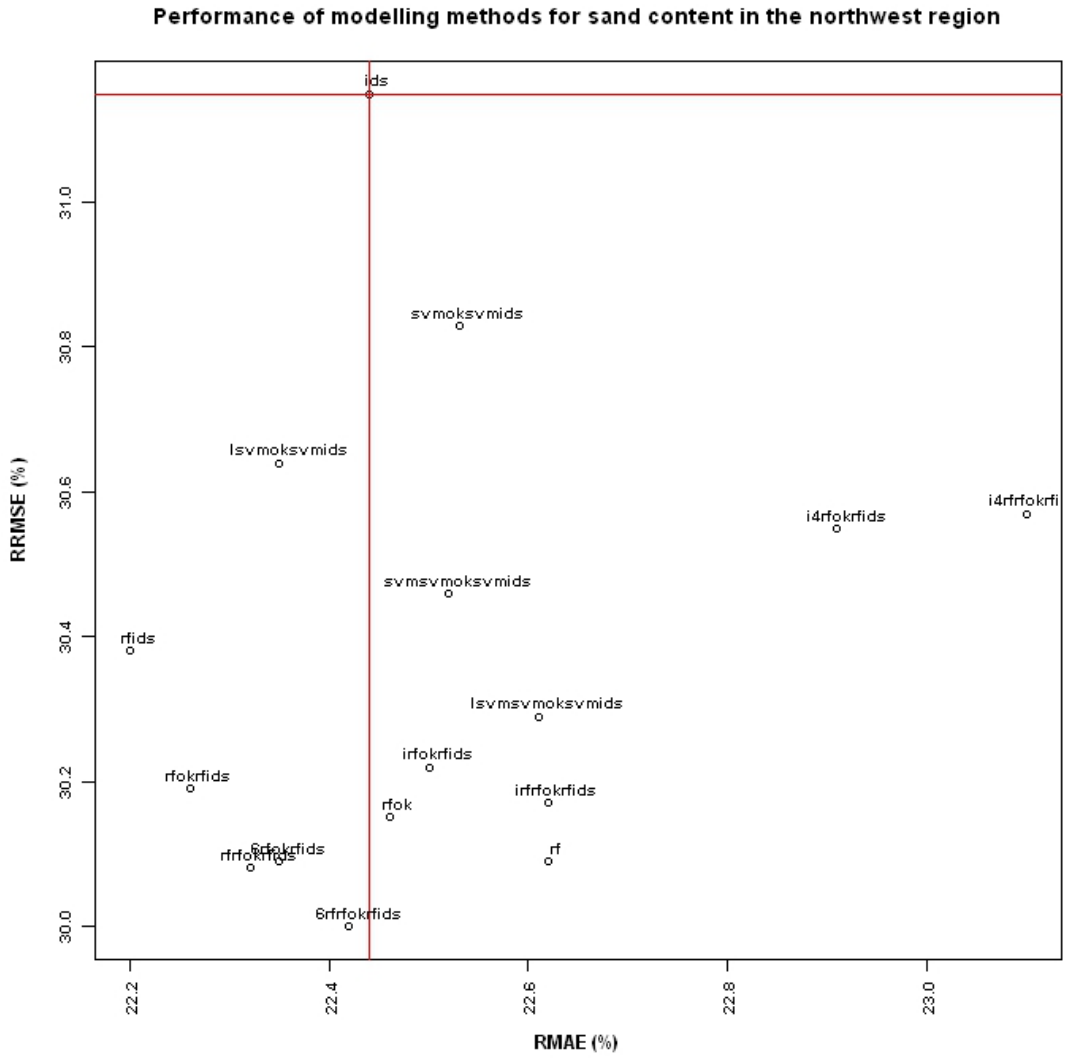


Figure 3.7. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of averaging predictions of two or three modelling methods for sand content in the northwest region. The horizontal and vertical lines (red) indicate the accuracy of the control (IDS).*

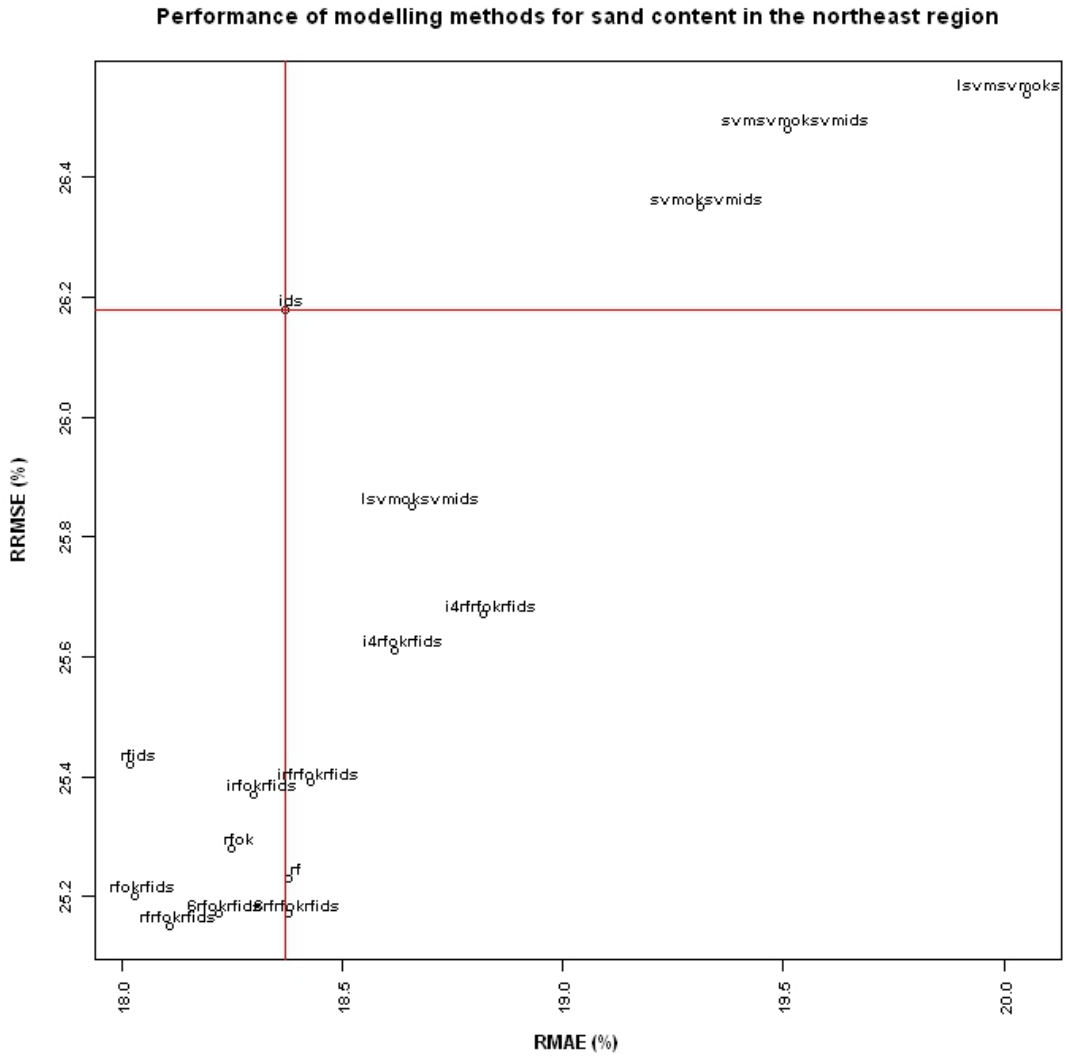


Figure 3.8. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of averaging predictions of two or three modelling methods for sand content in the northeast region. The horizontal and vertical lines (red) indicate the accuracy of the control (IDS).*

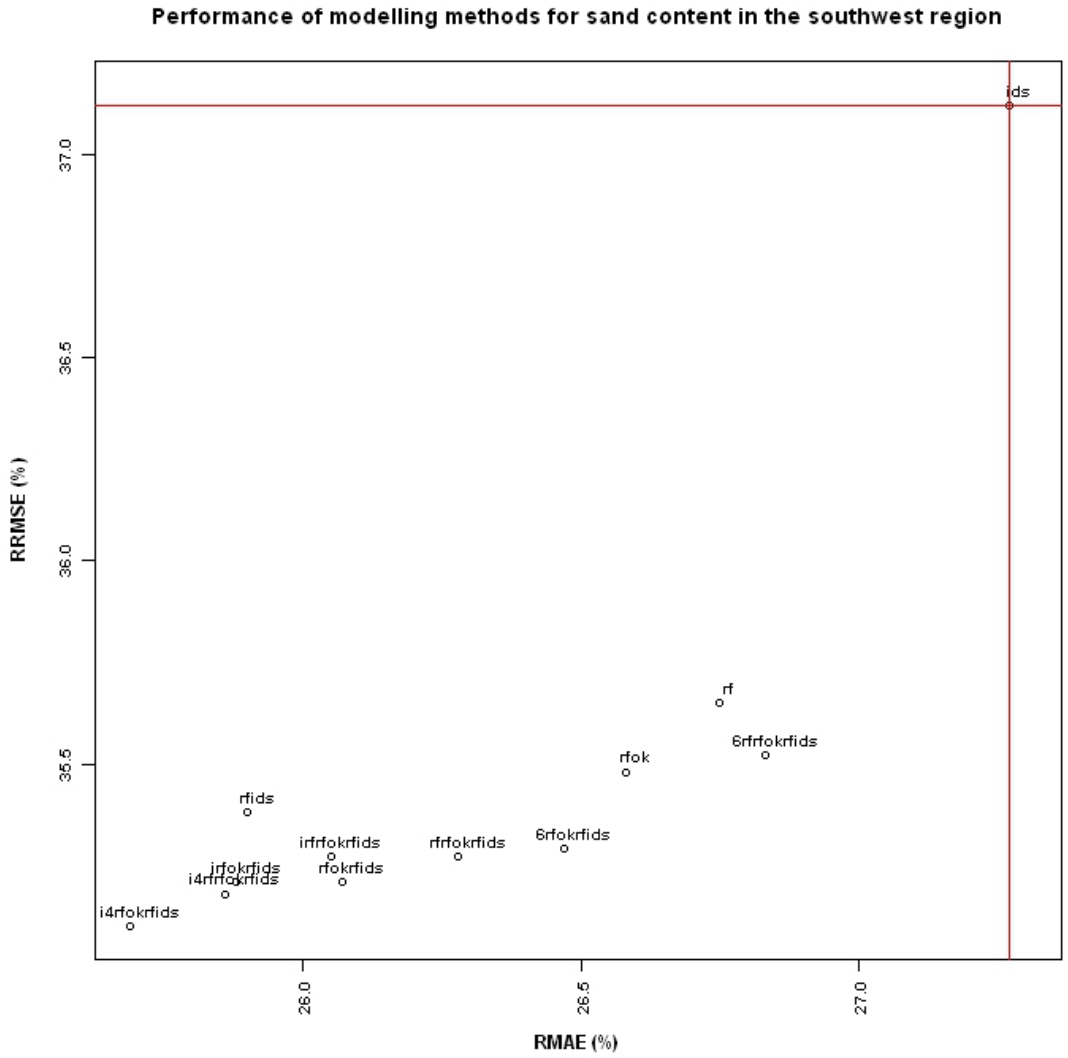


Figure 3.9. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of averaging predictions of two or three modelling methods for sand content in the southwest region. The horizontal and vertical lines (red) indicate the accuracy of the control (IDS). To enhance the presentation, all other averaged methods that were less accurate than IDS were excluded.*

3.4. EFFECTS OF 'THE NUMBER OF VARIABLES RANDOMLY SAMPLED AS CANDIDATES AT EACH SPLIT'

Since the optimal *mtry* was 7 for the northwest region and 4 for the remaining two regions, if the choice of *mtry* does not significantly affect the prediction accuracy, we can choose one single number for *mtry* for all regions in the AEEZ. So it is necessary to test such effect. The effects of *mtry* on the performance of RF, RFOK, RFIDS, RFOKRFIDS and RFRFOKRFIDS were marginal (Table 3.1). The choice of an *mtry* of 4 reduced the prediction accuracy for all methods by 0.03-0.05 for RMAE and 0.07-0.1 for RRMSE in comparison with the optimal number of an *mtry* of 7. Such reductions in the predictive accuracy are negligible.

Table 3.1. The effects of 'the number of variables randomly sampled as candidates at each split' (*mtry*) on the performance of RF, its combinations and averaging their predictions in the northwest region.

METHOD	*MTRY*	RMAE (%)	Δ RMAE	RRMSE (%)	Δ RRMSE
rf	7	22.62	0.05	30.09	0.08
4mrf	4	22.67		30.17	
rfok	7	22.46	0.03	30.15	0.08
4mrfok	4	22.49		30.23	
rfids	7	22.2	0.04	30.38	0.1
4mrfids	4	22.24		30.48	
rfokrfids	7	22.26	0.03	30.19	0.09
4mrfokrfids	4	22.29		30.28	
rfrfokrfids	7	22.32	0.03	30.08	0.07
4mrfrfokrfids	4	22.35		30.15	

3.5. OPTIMAL SEARCH WINDOW SIZE OF THE MOST ACCURATE METHODS

3.5.1. RF with an optimal *mtry* for each region

There was no single optimal search window size for all regions for the most accurate methods (*i.e.*, RFOK, RFIDS, RFOKRFIDS and RFRFOKRFIDS) in terms of RMAE and RRMSE (Figs 3.10-3.21). In the northwest region, the accuracy of the RFIDS varied with the search window size, and the best search window size was 5 (RMAE: 22.06% and RRMSE: 30.38%), followed by 19 (RMAE: 22.18% and RRMSE: 30.32%) and than 25 (RMAE: 22.20% and RRMSE: 30.29%) (Fig. 3.10). For RFOK, the relationship between the accuracy and the search window size was not apparent, but the best sizes were 5 (RMAE: 22.21% and RRMSE: 30.08%) and followed by Inf (RMAE: 22.35% and RRMSE: 30.06%) (Fig. 3.11). For RFOKRFIDS, the relationship between the accuracy and the search window size was similar to RFOK, with the best sizes being 5 (RMAE: 22.06% and RRMSE: 30.14%) and followed by 25 (RMAE: 22.21% and RRMSE: 30.12%) and Inf (RMAE: 22.24% and RRMSE: 30.11%) (Fig. 3.12). For RFRFOKRFIDS, the relationship between the accuracy and the search window size was similar to RFOK, with the best sizes being 5 (RMAE: 22.15% and RRMSE: 29.99%) and followed by 4 (RMAE: 22.20% and RRMSE: 30.03%) (Fig. 3.13). Overall, the best search window size was 5 while the best methods were RFOKRFIDS and RFRFOKRFIDS.

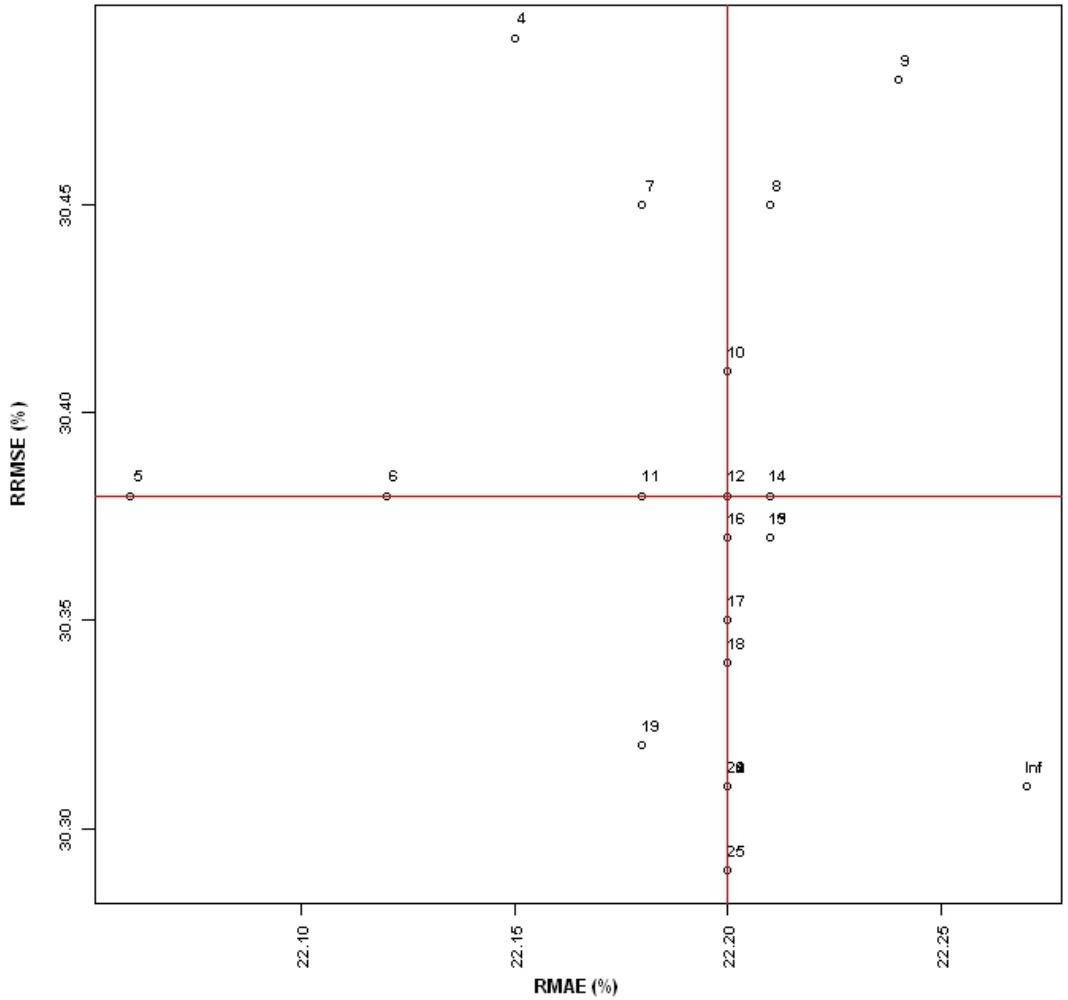


Figure 3.10. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFIDS for sand content in relation to search window size in the northwest region.*

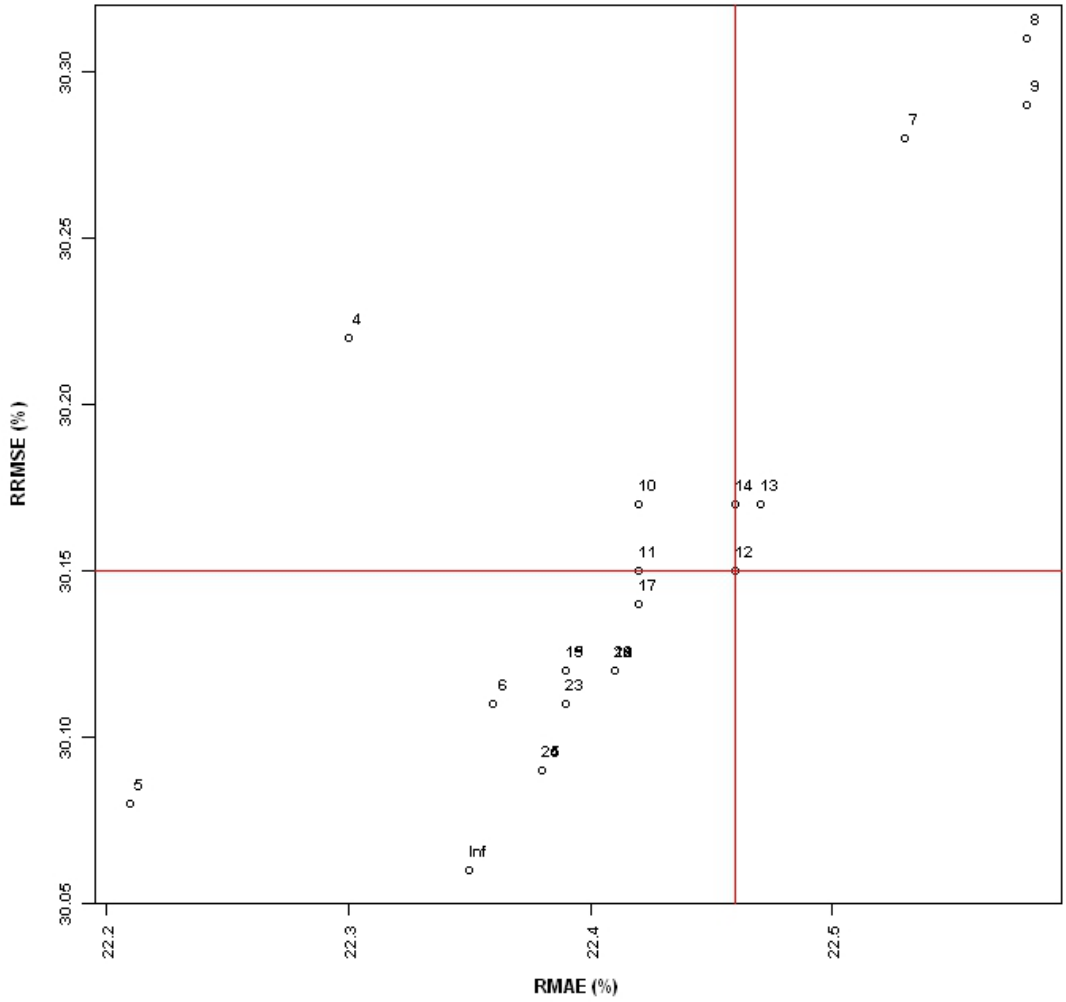


Figure 3.11. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFOK for sand content in relation to search window size in the northwest region.*

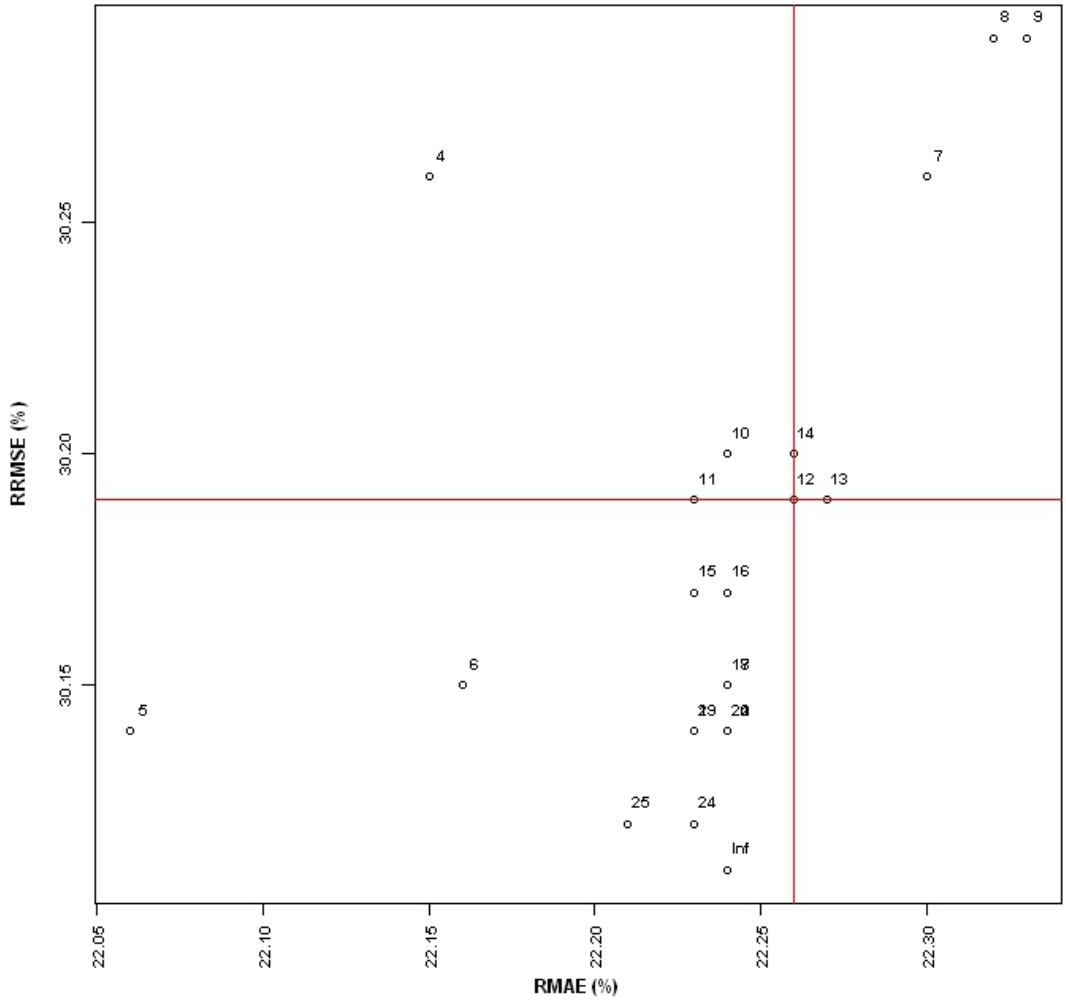


Figure 3.12. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFOKRFIDS for sand content in relation to search window size in the northwest region.*

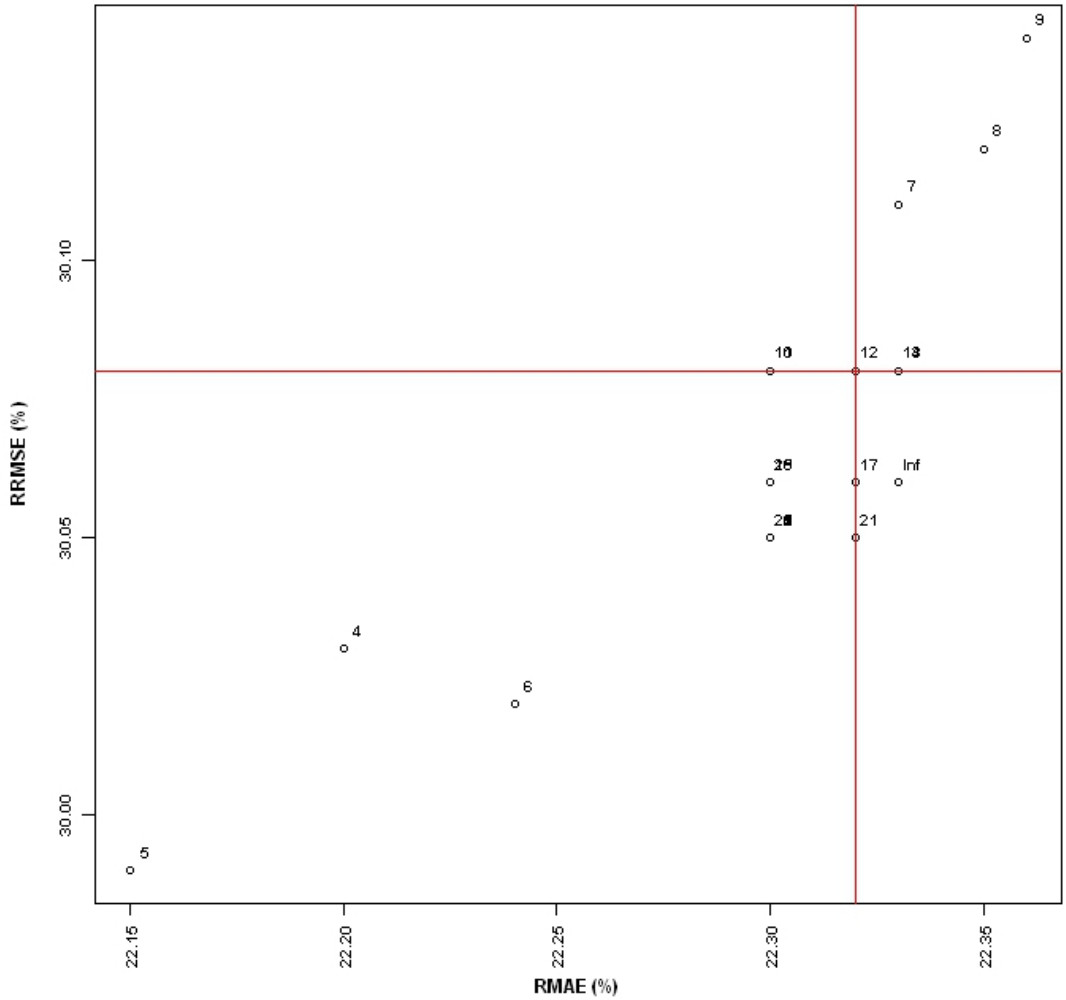


Figure 3.13. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFRFOKRFIDS for sand content in relation to search window size in the northwest region.*

In the northeast region, there was little relationship between the prediction error and the search window size for RFIDS, and the best sizes were 16 and 17 (RMAE: 18.00% and RRMSE: 25.39%) (Fig. 3.14). The prediction error of RFOK increased with the increasing search window size in terms of RMAE while decreased in terms of RRMSE; and the best size was 14 (RMAE: 18.22% and RRMSE: 25.23%) (Fig. 3.15). For RFOKRFIDS, the relationship between the accuracy and the search window size was similar to RFOK, with the best sizes being 14 (RMAE: 18.02% and RRMSE: 25.17%) (Fig. 3.16). The prediction error of RFRFOKRFIDS increased with the increasing search window size in terms of RMAE while displayed no patterns in terms of RRMSE. The best size was 7 (RMAE: 18.08% and RRMSE: 25.14%) (Fig. 3.17). Overall, the effects of the search window size were marginal; and the most accurate prediction was from RFOKRFIDS with a search window size of 14.

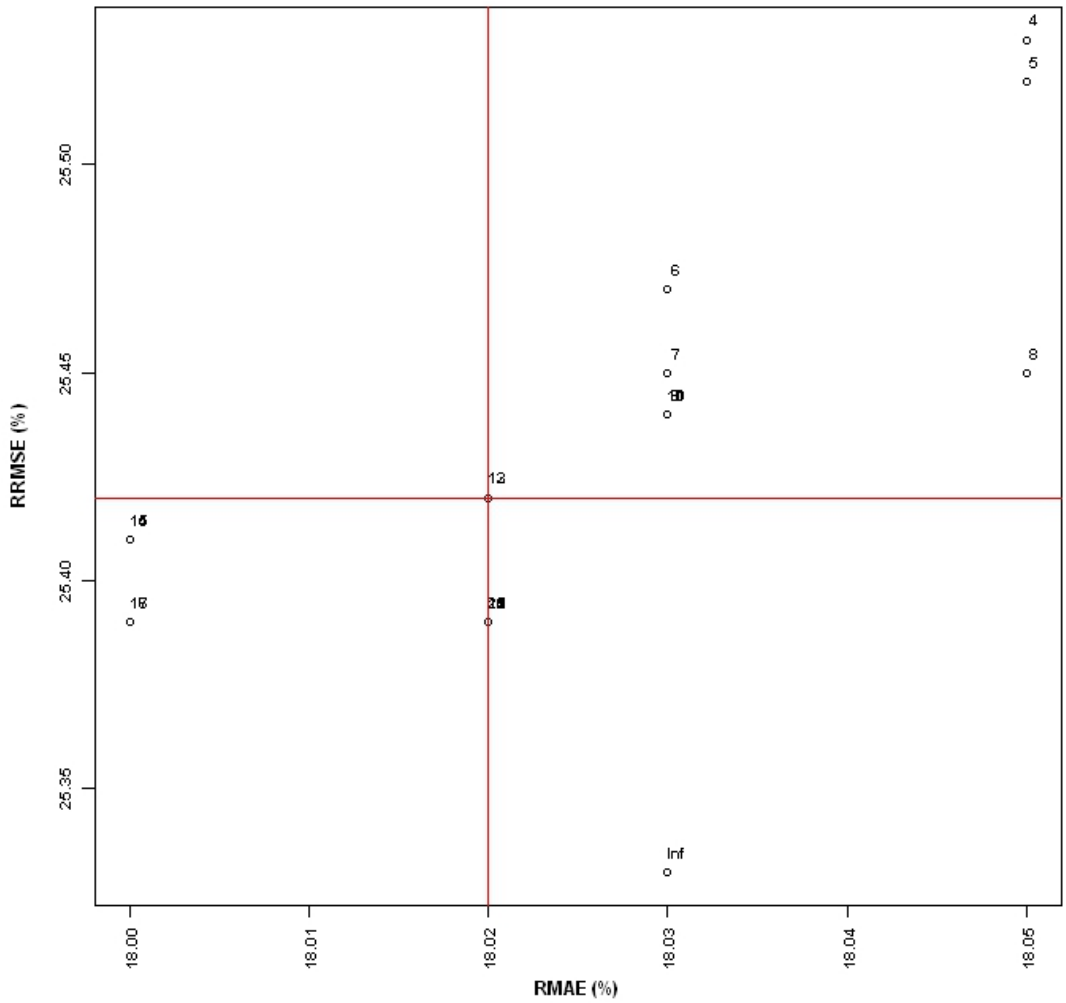


Figure 3.14. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFIDS for sand content in relation to search window size in the northeast region.*

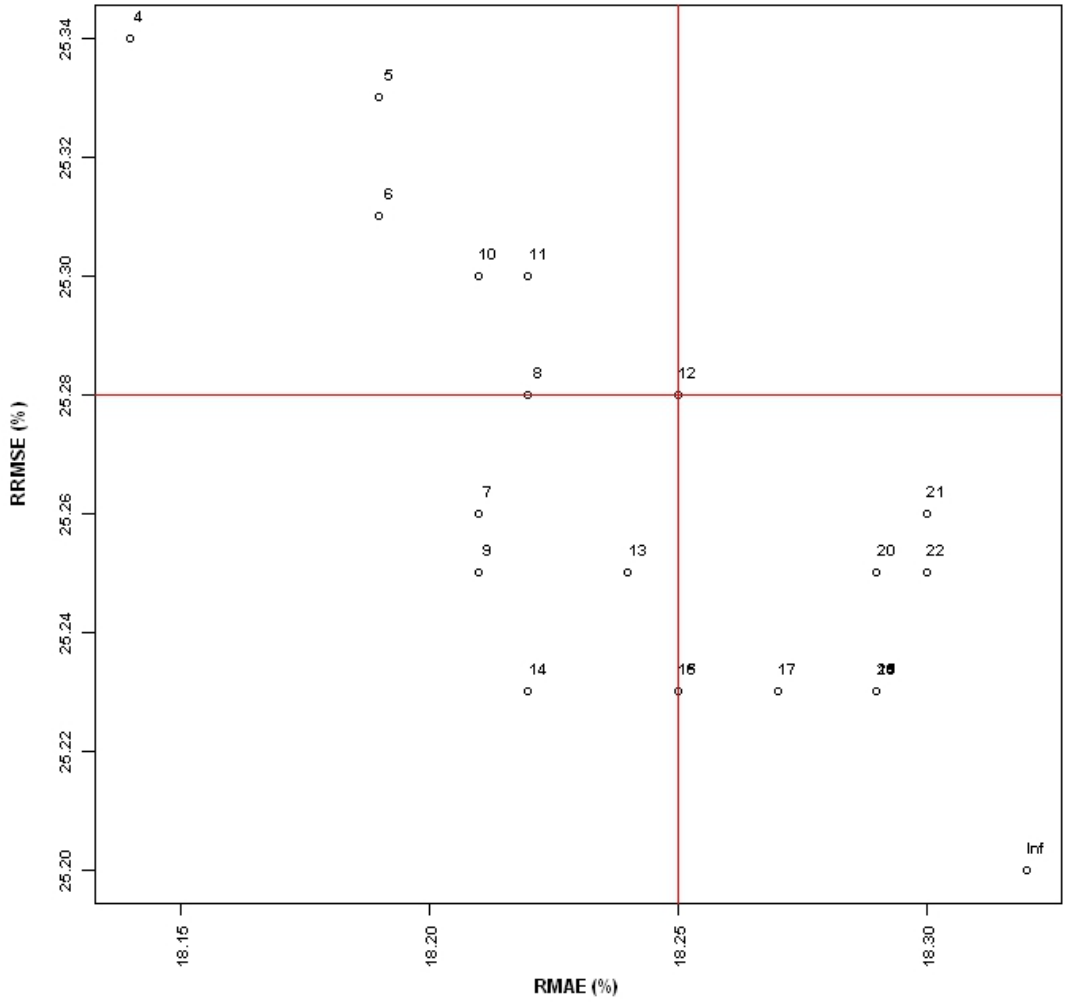


Figure 3.15. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFOK for sand content in relation to search window size in the northeast region.*

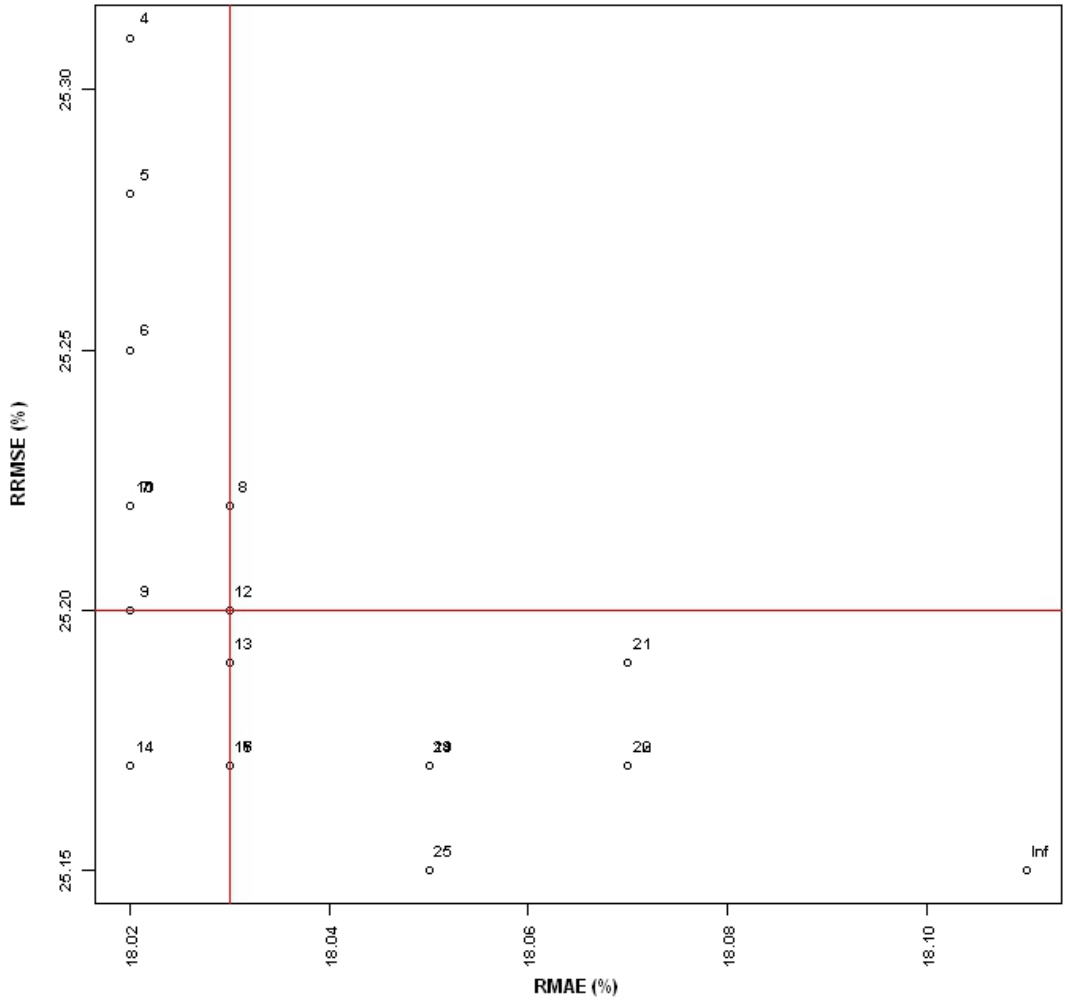


Figure 3.16. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFOKRFIDS for sand content in relation to search window size in the northeast region.*

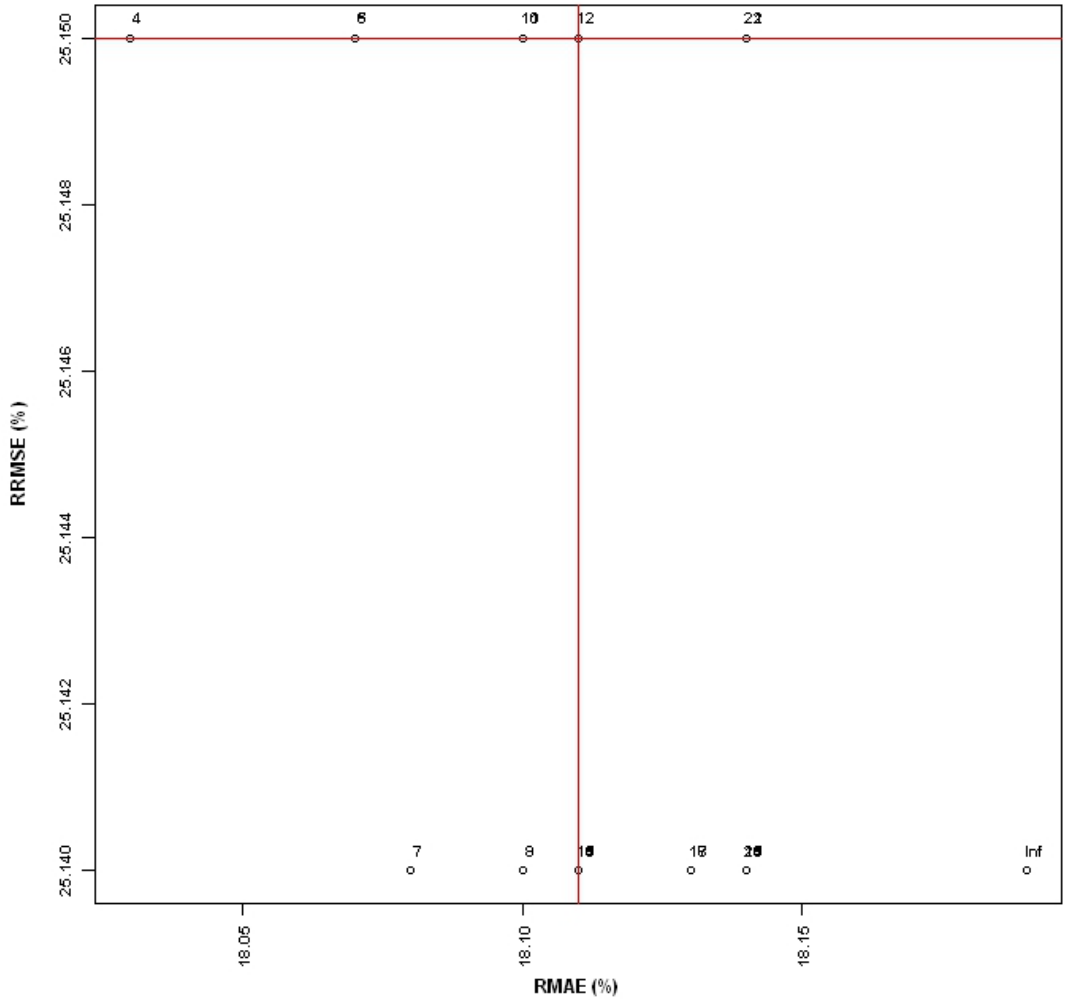


Figure 3.17. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFRFOKRFIDS for sand content in relation to search window size in the northeast region.*

In the southwest region, the prediction error of RFIDS showed a weak linear relationship with the search window size. For RFIDS the best size was 5 (RMAE: 25.69% and RRMSE: 35.37%) (Fig. 3.18). For RFOK, there was little relationship between the prediction accuracy and the search window size; and the best sizes were 4 (RMAE: 25.46% and RRMSE: 34.87%) and 5 (RMAE: 25.47% and RRMSE: 34.83%) (Fig. 3.19). For RFOKRFIDS, the relationship between the accuracy and the search window size was similar to RFOK, with the best sizes being 5 (RMAE: 25.50% and RRMSE: 34.93%) and 4 (RMAE: 25.46% and RRMSE: 34.98%) (Fig. 3.20). RFRFOKRFIDS also displayed a similar pattern to RFOK (Fig. 3.21); with the best size being 4 (RMAE: 25.72% and RRMSE: 34.95%). The difference in predictive accuracy among these methods was marginal. Overall, most accurate prediction in the southwest region was from RFOK with a search window size of 5.

If only one method needs to be selected for all regions with a single search window size, the method is RFOKRFIDS with a size of 5 based on the RMAE and RRMSE.

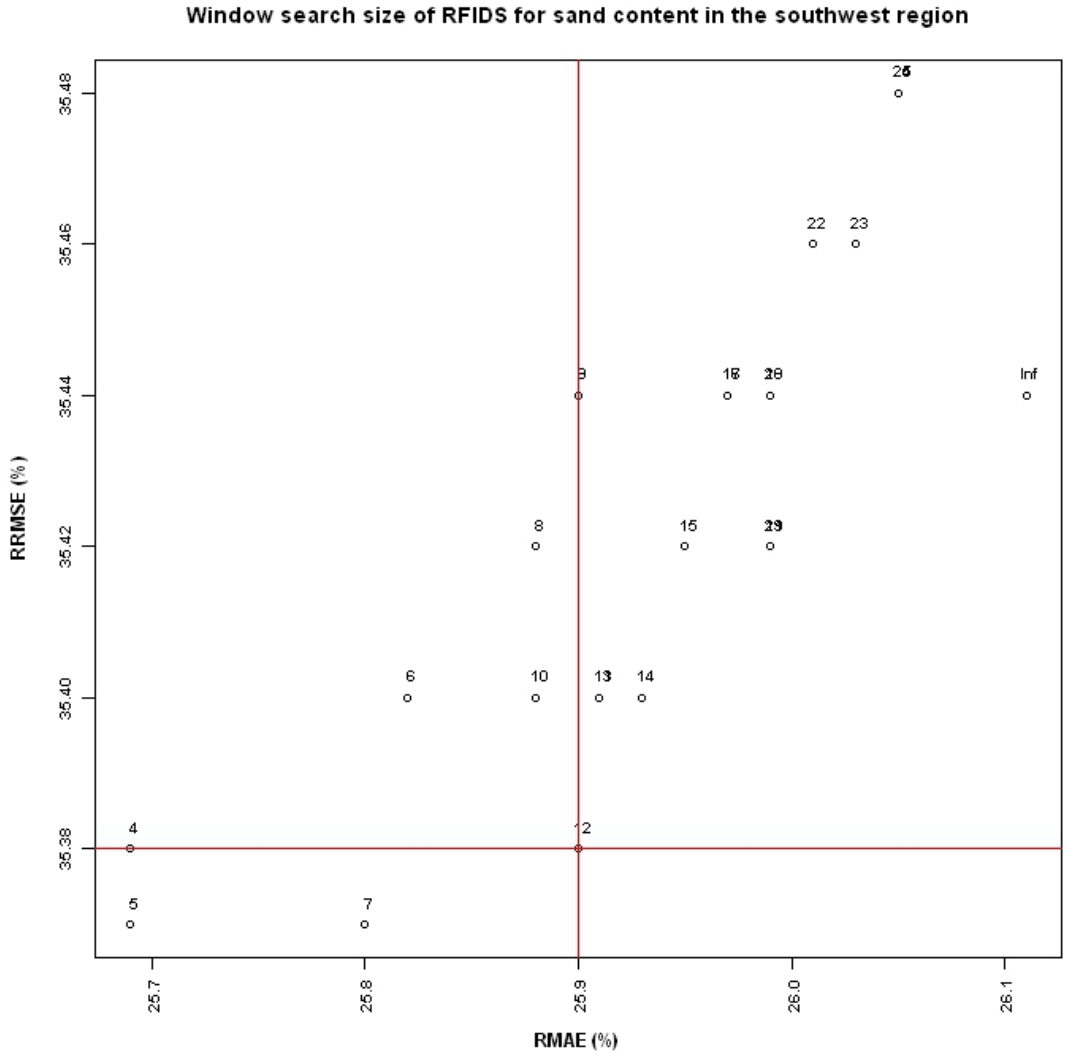


Figure 3.18. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFIDS for sand content in relation to search window size in the southwest region.*

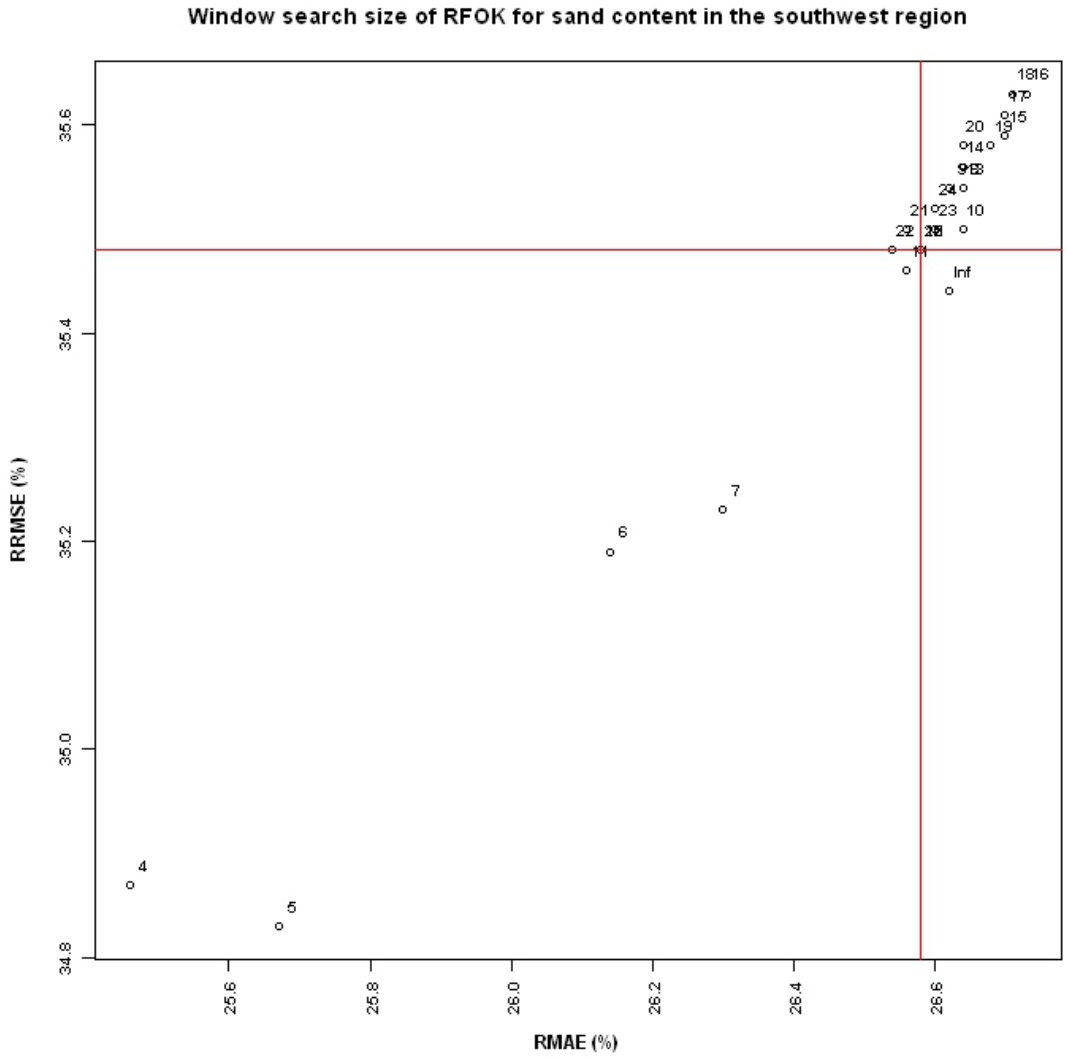


Figure 3.19. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFOK for sand content in relation to search window size in the southwest region.*

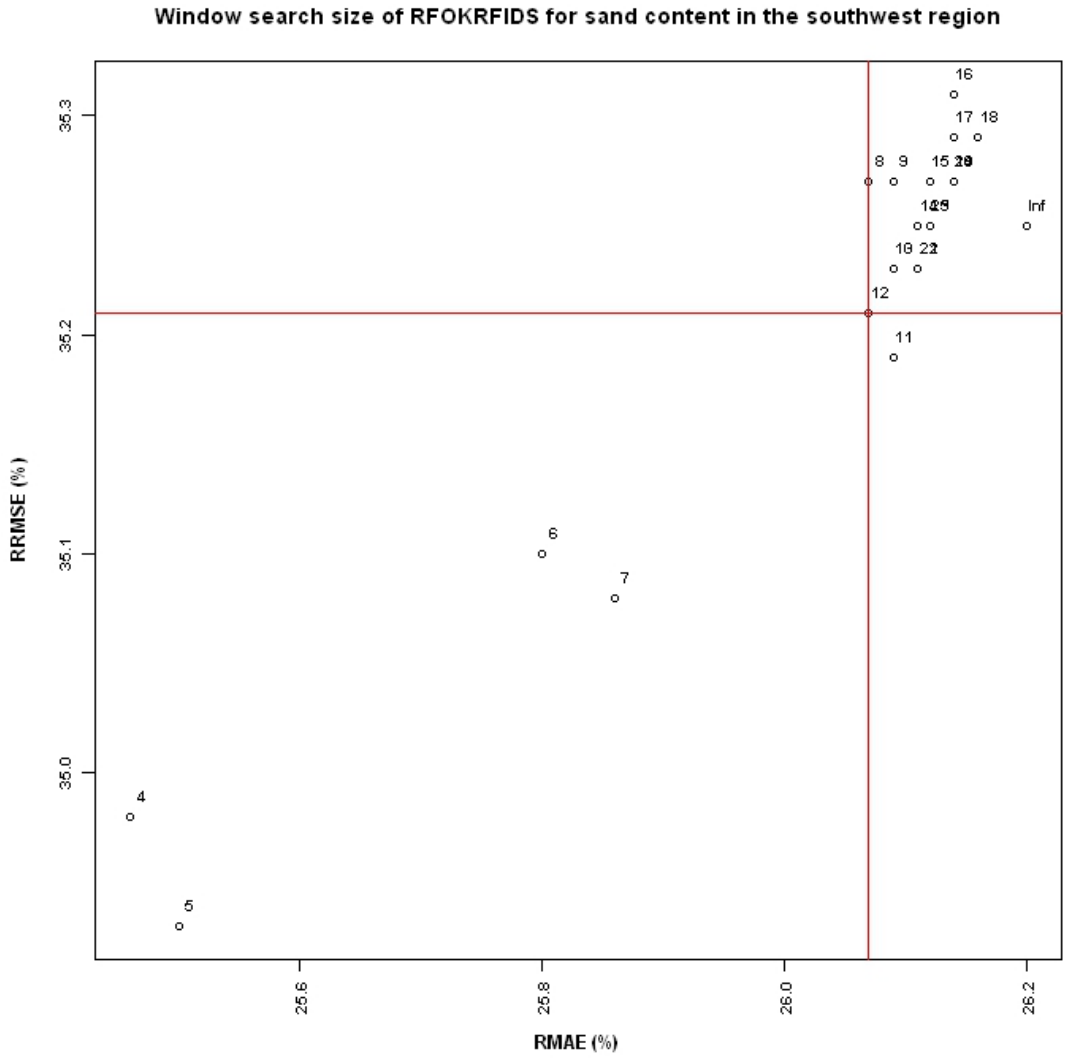


Figure 3.20. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFOKRFIDS for sand content in relation to search window size in the southwest region.*

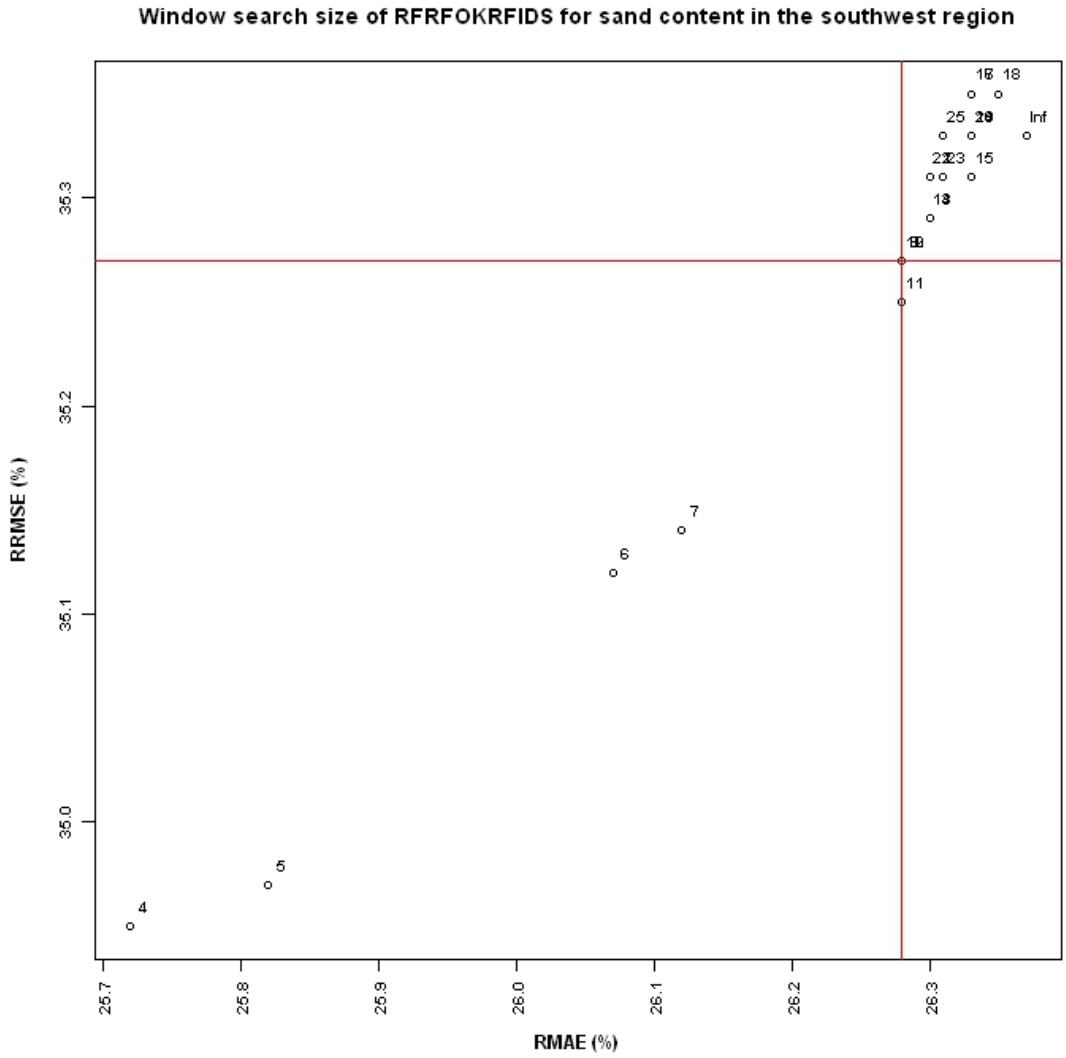


Figure 3.21. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFRFOKRFIDS for sand content in relation to search window size in the southwest region.*

3.5.2. RF with a mtry of 4 in the northwest region

In the northwest region, the accuracy of the RFOKRFIDS varied with the search window size with no apparent pattern, and the best search window size was 5 (RMAE: 22.09% and RRMSE: 30.20%) (Fig. 3.22).

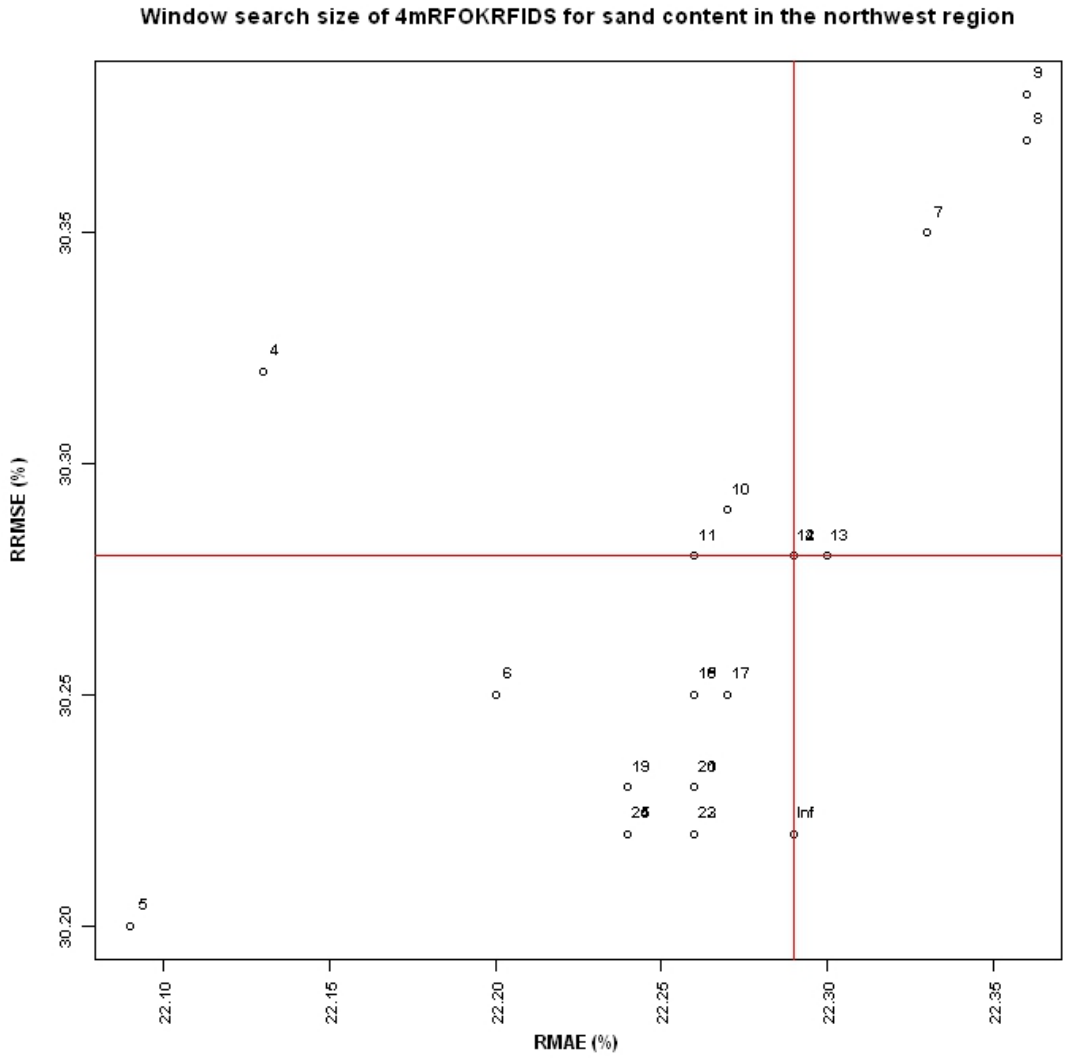


Figure 3.22. *The relative absolute mean error (RMAE (%)) and relative root mean square error (RRMSE (%)) of RFOKRFIDS for sand content in relation to search window size in the northwest region.*

3.6. VISUAL EXAMINATION

3.6.1. Northwest region

The spatial predictions of the most accurate method (*i.e.*, RFOKRFIDS) and control method (i.e., IDS) are shown in Figs. 3.23-3.28. In the northwest region, both methods captured the major spatial patterns and trends of sand content, with higher sand content in areas closer to the coast (Figs. 3.23 and 3.24). However, apparent 'bull's eye' patterns at sample points of high or low values were evident for the predictions of IDS. The predictions of RFOKRFIDS displayed more detailed patterns reflecting both local and regional trends, and with much weaker 'bull's eye' patterns that were only noticeable at a few sample points; and weak horizontal and vertical banding patterns were also perceptible.

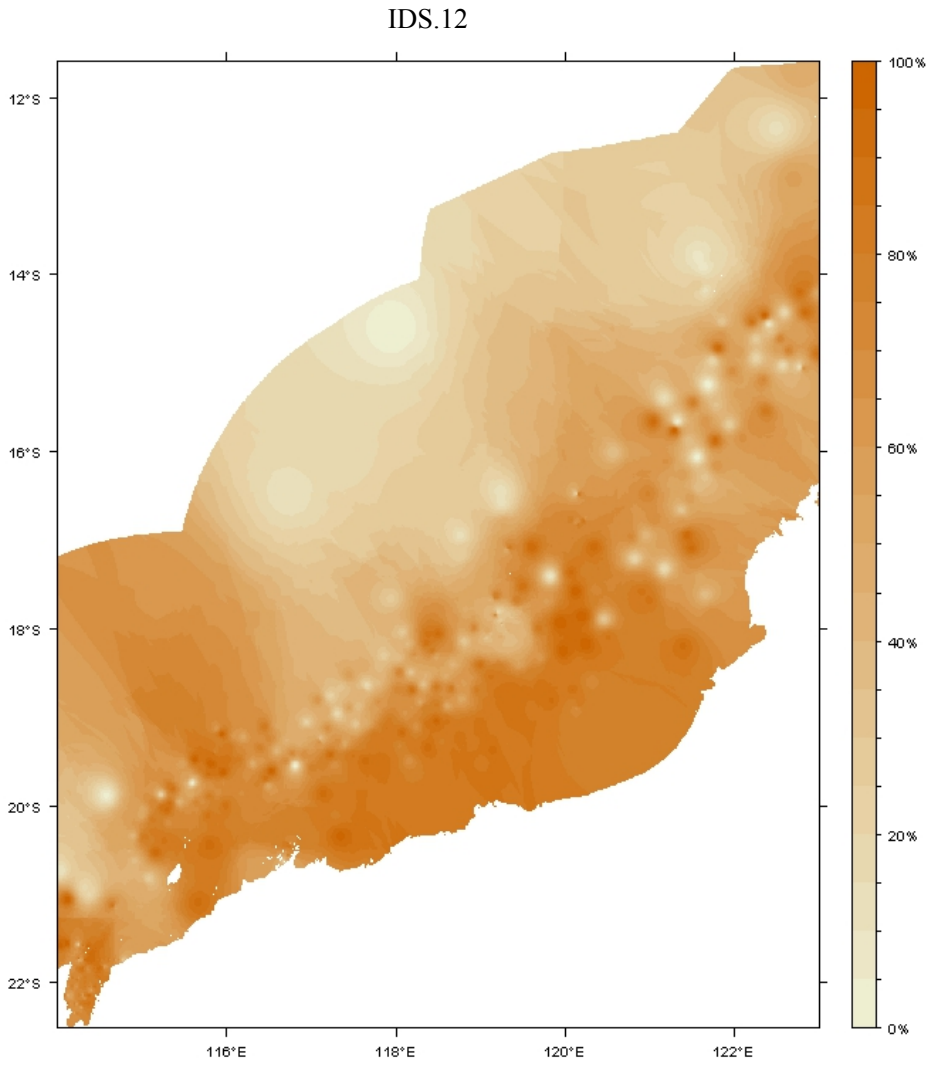


Figure 3.23. *The predictions of the control method (IDS with a search window size of 12) in the northwest region.*

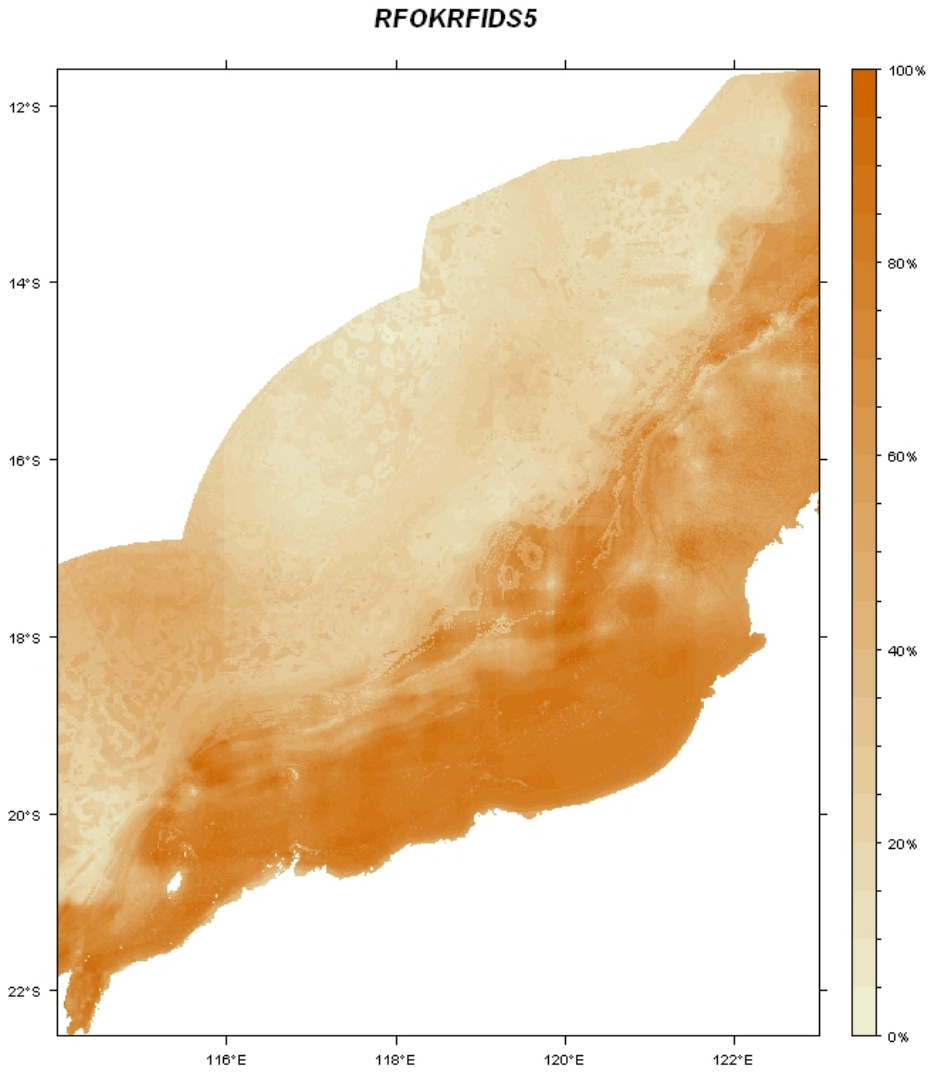


Figure 3.24. *The predictions of the most accurate method (i.e., RFOKRFIDS with a search window size of 5) in the northwest region.*

3.6.2. Northeast region

In the northeast region, both methods captured the major spatial patterns and trends of sand content in areas on the shelf and predicted quite different patterns in the deeper sea areas (Figs. 3.25 and 3.26). The 'bull's eye' patterns at sample points were evident for the predictions of IDS. In the deeper sea areas, its predictions mainly reflected the influences of samples located on the shelf area and displayed artefacts. The predictions of RFOKRFIDS displayed more detailed patterns reflecting both local and regional trends, and with much weaker 'bull's eye' patterns that were only noticeable at a few sample points; and banding patterns including horizontal, vertical and diagonal were perceptible.

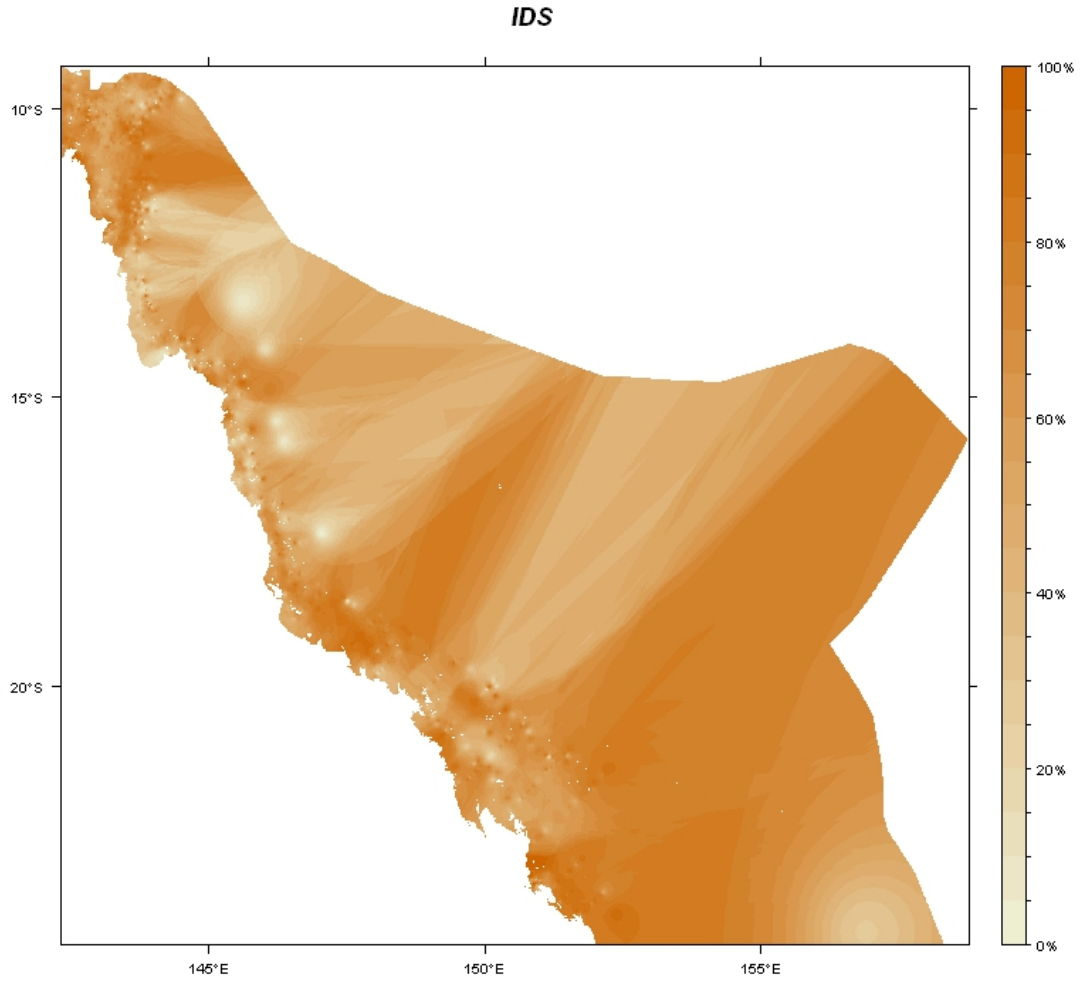


Figure 3.25. *The predictions of the control method (IDS with a search window size of 12) in the northeast region.*

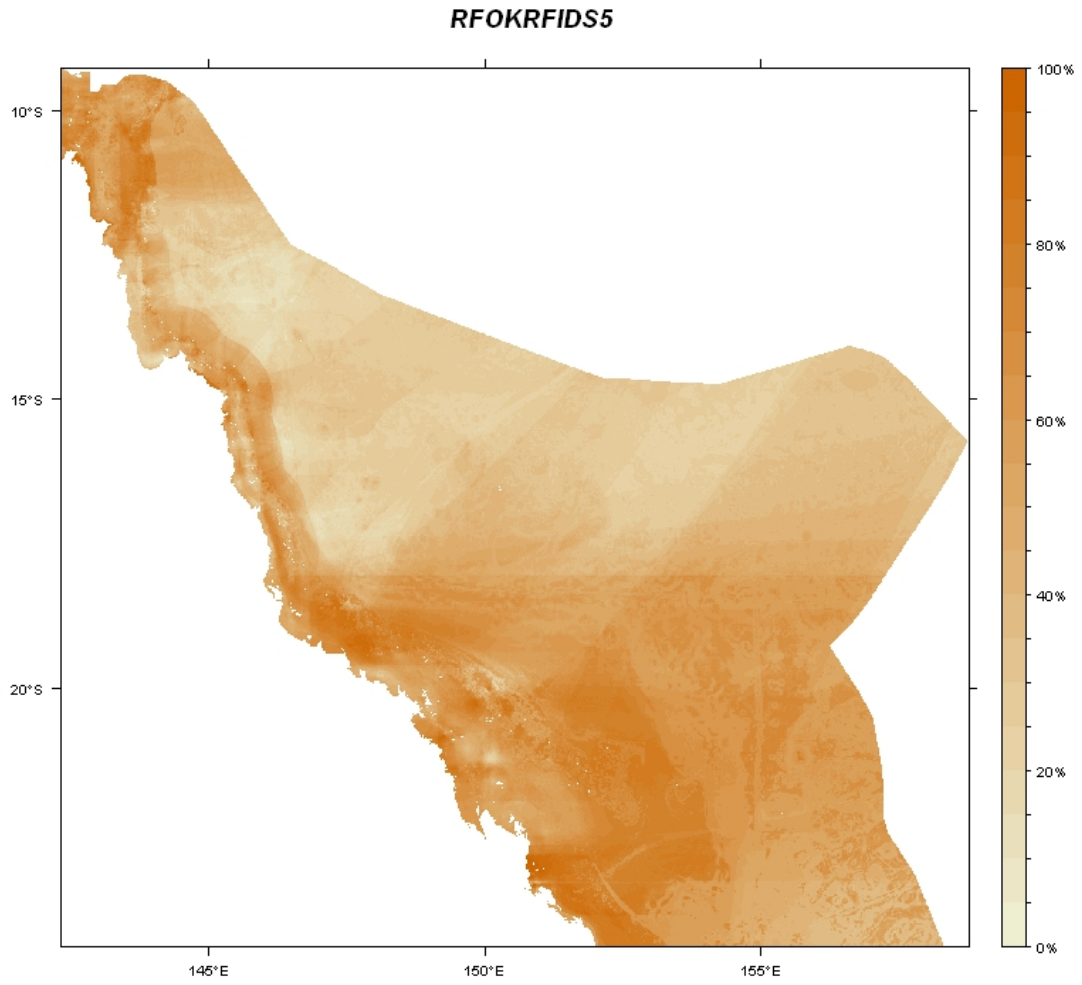


Figure 3.26. *The predictions of the most accurate method (i.e., RFOKRFIDS with a search window size of 5) in the northeast region.*

3.6.3. Southwest region

In the southwest region, both methods depicted major spatial patterns and trends of seabed sand content, with higher sand content on the shelf and lower sand content in the deeper sea areas (Figs. 3.27 and 3.28). Again the predictions of IDS displayed 'bull's eye' patterns, while weak 'bull's eye' patterns were only detectable at a few locations of sample points for RFOKRFIDS; and some linear tracks and weak banding patterns were noticeable in its predictions.

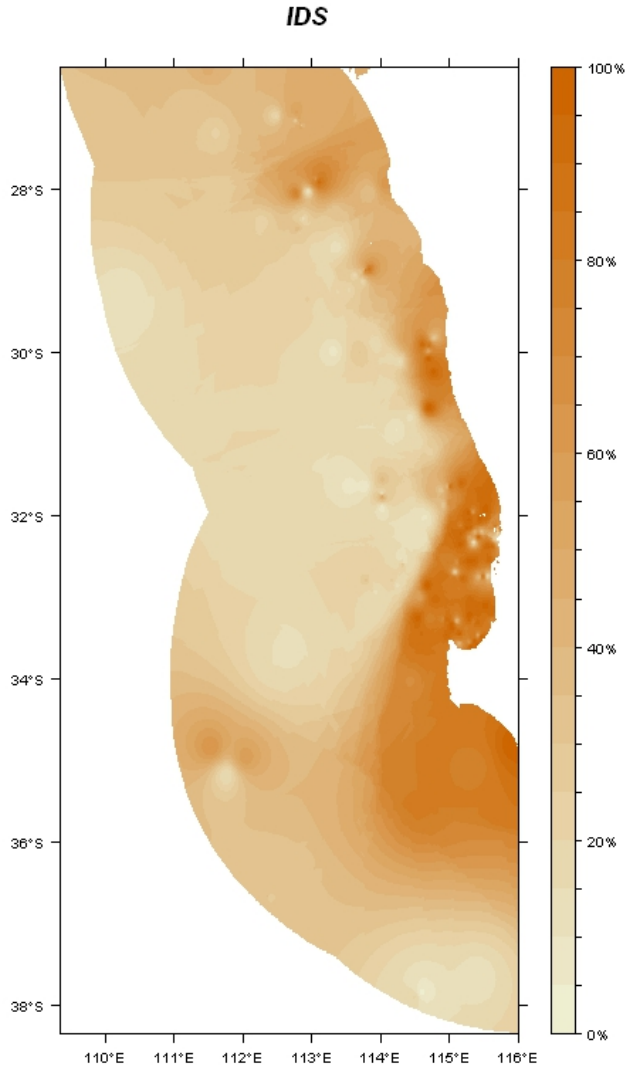


Figure 3.27. *The predictions of the control method (IDS with a search window size of 12) in the southwest region.*

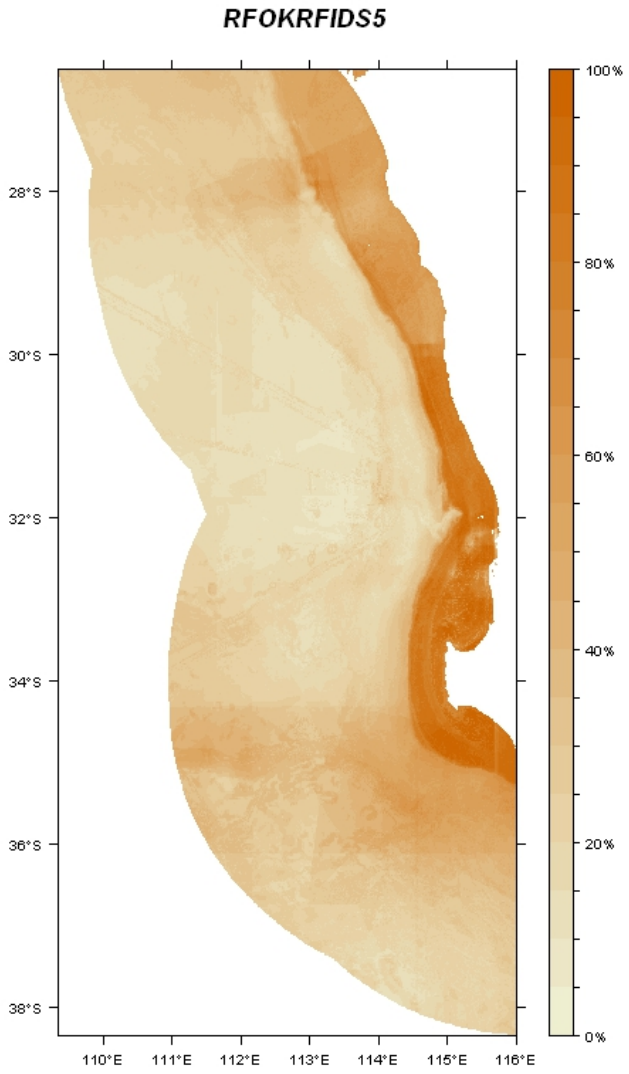


Figure 3.28. *The predictions of the most accurate method (i.e., RFOKRFIDS with a search window size of 5) in the southwest region.*

Chapter 4. Discussion

4.1. OPTIMAL MODELLING METHODS

The performance of the statistical and mathematical methods for the spatial interpolation of sand content was region-dependent and no single method was always superior to the other methods in terms of RMAE and RRMSE. This phenomenon has been observed in a number of previous studies (Elith et al., 2006; Li and Heap, 2011; Li et al., 2011a; Li et al., 2010; Marmion et al., 2009; Olden and Jackson, 2002). It is consistent with the 'no free lunch theorems' in optimisation (Wolpert and Macready, 1997). This may be attributed to the differences in a number of aspects between the study regions as discussed in previous studies (Li et al., 2011a; Li et al., 2010), and also due to that the physical processes that influence the formation and accumulation of sand sediment may differ between regions. This suggests that the optimal interpolation method changes with regions and efforts should be made to identify the optimal method for each individual region if possible.

The accuracies of the best performing methods varied with region. The prediction accuracies of the methods are the highest in the northeast region, followed by the northwest region, and the least in the southwest region. In our previous studies (Li et al., 2011a; Li et al., 2010), such differences were mainly attributed to the higher correlations between the primary variable and the secondary variables, but this is not the case in this study, since the correlations between sand content and the secondary variables (i.e., bathymetry, distance to coast, slope and relief) are lower in the northeast region than in other regions (Table B.2-B.5). The difference in the predictive accuracies is mainly due to the differences in the data variation of sand content among the three regions (Table 2.3), because the data variation in the northeast is the lowest followed by the northwest region and then the southwest region and because the predictive accuracy is often negatively correlated with data variation (Figs. 4.1 and 4.2) (Li and Heap, 2011). Other factors such as sample density and distribution patterns of sample are similar in these two studies, thus their contributions are expected to be similar as well and would not expected to cause such a difference. Therefore, data variation seems more important than the correlations between the primary and secondary variables for sand content and has played a major role in the resultant predictive accuracies.

RFIDS and RFOK were, on average, more accurate than other methods in all three regions. This finding is in consistent with our previous studies for mud content (Li et al., 2011a; Li et al., 2010). This suggests that RF successfully predicted the general trend in each region based on the secondary information, and OK and IDS made further contributions to the observed patterns at a local scale as was discussed in our previous studies. The slightly superior performance of RFIDS to RFOK in the southwest region was also observed in our previous studies (Li et al., 2011a; Li et al., 2010). Since all methods were compared under the same conditions in this study and the previous study, we conclude that the superior performance of RFOK and RFIDS is attributed to the methods themselves rather than to other factors as discussed previously (Li et al., 2011a). RF has been proven to be a robust method in several previous studies (Diaz-Uriarte and de Andres, 2006; Knudby et al., 2010; Marmion et al., 2009; Pino-Mejías et al., 2010), which supports the excellent performance of RFOK and RFIDS in this study.

Such performance can be attributed to a number of features associated with the methods as discussed in Li et al. (2011b) and Li et al. (2011a).

All other machine learning methods (i.e., SVM, LSVM, GRNN and BDT) performed better when combined with IDS or OK. This finding is consistent with previous findings (Li et al., 2011a) and further confirms the effectiveness of the combining machine learning methods with commonly used spatial interpolation methods such as OK or IDS and supports the development of an alternative source of methods for the discipline of spatial statistics (Li et al., 2011a).

In addition, the simplification processes for GRNN for the northwest and northeast regions were not optimal (see B.4.3 for details), which might have resulted in the inferior performance of this method in these two regions. However, this method performed equally poor in the southwest region where all of the unnecessary neurons were removed to optimise the predictive model. Therefore, it is fair to conclude that this method is among the poorest performing methods tested in this study. However, given the difficulty involved in developing a predictive model using artificial neural networks (Özesmi et al., 2006) and limited resources allocated in this study, this finding should not discourage attempts to test and apply this method in future studies.

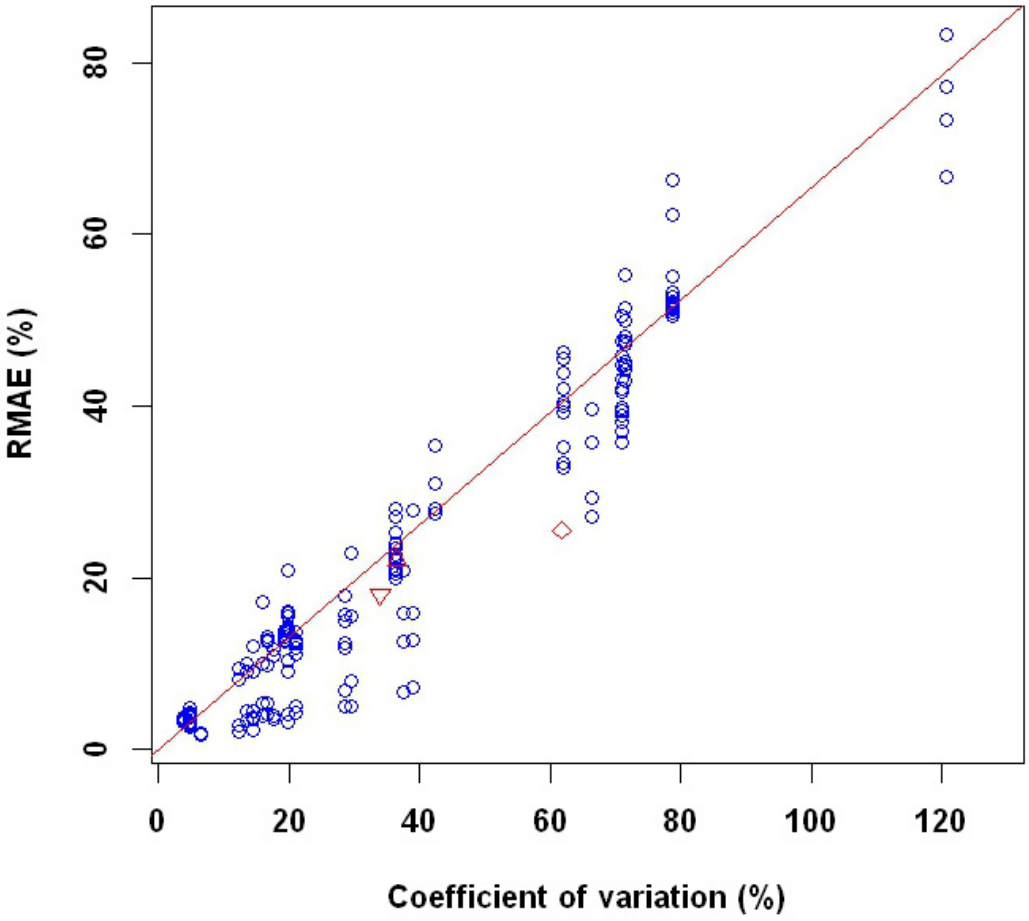


Figure 4.1. *The relative mean absolute error (RMAE(%)) of RFOKRFIDS5 in three regions (northwest: red triangle; northeast: upside down red triangle; and southwest: red diamond) in comparison with the results of previous studies (Li and Heap, 2008, 2011). The fitted line for resistant regression with LTS with a slope of 0.65 was derived using a R library MASS (Venables and Ripley, 2002).*

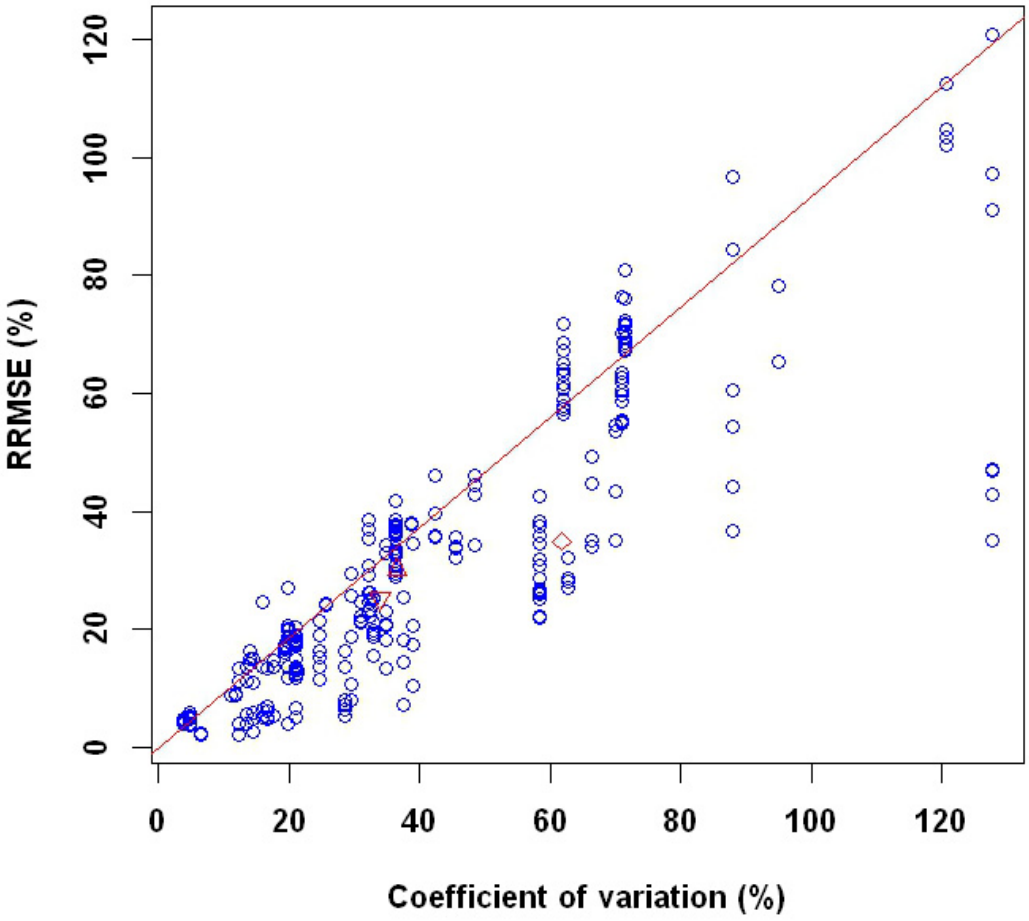


Figure 4.2. *The relative mean absolute error (RRMSE(%)) of RFOKRFIDS5 in three regions (northwest: red triangle; northeast: upside down red triangle; and southwest: red diamond) in comparison with the results of previous studies (Li and Heap, 2008, 2011). Please note that this figure is updated from Figure 6.16 in Li and Heap (Li and Heap, 2008) by correcting the results from two references and by adding results from two new references. The fitted line for resistant regression with LTS with a slope of 0.93 was derived using a R library MASS (Venables and Ripley, 2002).*

4.2. DO INPUT SECONDARY VARIABLES MATTER FOR RANDOM FOREST?

The choice of secondary variables (Table B.8) was found to affect the performance of random forest related methods. In the northwest and northeast regions, methods with the i4-variables performed worst, followed by methods with the i-variables, and then methods with the 6-variables and with the control input variables performed the best. In the southwest region, methods with the 6-variables performed worst, followed by methods with control, then methods with the i-variables, and methods with the i4-variables performed the best. The magnitude of the changes in the predictive errors of the methods tested also varied among regions: with the northwest region having the least changes, and the southwest region the largest. These findings highlight that 1) the effects of the choice of input secondary variables for RF related methods depend on methods (i.e., RF, RFOK and RFIDS) and study regions and there is probably an interactive effect between region and the choice of the input secondary variables; and 2)

the second and third order of relevant secondary variables and their interactions may have contributed to the predictive accuracy which again changes with regions.

Two different phenomena were observed. On one hand, 6RF is more accurate than iRF and i4RF in the northwest and northeast regions, suggesting excluding the irrelevant correlated variables would improve the predictive accuracy, which is consistent with our previous studies (Li et al., 2011a; Li et al., 2011b), where it was found that when the least important variables were excluded, the predictive errors were significantly reduced for the methods tested. This is also supported by findings regarding feature selection for machine learning methods in a previous study (Guyon et al., 2009). This phenomenon was explained in Li et al. (2011b) as that "the inclusion of noisy predictive variables can reduce the probability of the inclusion of good predictive variables at each node split when a portion of predictive variables are randomly selected for each individual tree, thus the chance of the contribution of good predictive variables to the tree is reduced. Consequently, the tree developed produces less accurate predictions." On the other hand, 6RF is slightly less accurate than RF in the northwest and northeast regions, suggesting that inclusion of some correlated variables could improve the predictive accuracy. This implies that those correlated variables may contain some useful information as we hoped initially. These two phenomena seem contradictory to each other. They suggest that we need to find an optimal set of predictors for RF for each region. Therefore, the influence of the inclusion of 'irrelevant variables' is data-/region-specific and 'irrelevant variables' should be considered according to individual situation. "No free lunch theorems' in optimisation (Wolpert and Macready, 1997) are still valid.

The ranges of predictive errors increased from RFIDS, RFOK to RF under different choice of input secondary variables in all three regions. RFIDS performed slightly better than RFOK, and RFOK outperformed RF under different choice of input secondary variables in all three regions. These findings suggest that the effects of the choice of the input secondary variables are method-dependent.

Random forest was claimed to implicitly perform variable selection (Okun and Priisalu, 2007) because it selects the most important variable to split the samples at each node split for each individual trees. It was also claimed that RF can also deliver good predictive performance even when most predictive variables are noise (Diaz-Uriarte and de Andres, 2006). However, this study, in conjunction with our previous studies, suggests that the choice of the input secondary variables for random forest is important and warrants consideration. Model selection is essential for RF in order to find an optimal predictive model.

4.3. CAN MODEL AVERAGING IMPROVE THE ACCURACY OF SPATIAL PREDICTIONS?

In this study, we averaged the predictions of two or three modelling methods to improve the predictions. The effects of model averaging changed with the methods averaged, study region and also error measurement. Overall, model averaging marginally improved prediction accuracy in the southwest region, while in other regions no apparent effects were observed. RFOKRFIDS and RFRFOKRFIDS performed better

than other methods in all three regions in terms of both RMAE and RRMSE. The model averaging had little influence on the prediction accuracy of mud content in the southwest region in our previous study (Li et al., 2011b), while a marginally improved accuracy was observed in the northeast region (Li et al., 2011a). However, improved predictive accuracy as a result of model averaging was observed in previous findings (Goswami and O'Connor, 2007; Hoeting et al., 1999; Marmion et al., 2009; Raftery et al., 2005).

4.4. DOES THE CHOICE OF MTRY FOR RANDOM FOREST AFFECT ITS PREDICTIVE ACCURACY?

The effects of the choice of *mtry* (4 vs. 7) on the performance of RF, RFOK, RFIDS, RFOKRFIDS and RFRFOKRFIDS were marginal in the northwest region. The predictive accuracy of RFOKRFIDS with an *mtry* of 4 is slightly lower than that with an *mtry* of 7, by increasing the predictive error from 22.06% to 22.09 % in terms of RMAE and from 30.14% to 30.20% in terms of RRMSE. These differences are negligible. This finding supports that the performance of RF is not much influenced by parameter choices (Diaz-Uriarte and de Andres, 2006; Liaw and Wiener, 2002; Okun and Priisalu, 2007). Therefore, we can choose one single number for *mtry* for all regions in the AEEZ. Given that the optimal *mtry* for both northwest and northeast regions is 4, if we need to select only one *mtry* for RF across all AEEZ, an *mtry* of 4 is recommended.

4.5. OPTIMAL SEARCH WINDOW SIZE

This study found that the relationship between the prediction error and the search window size changes with method and region; and the optimal search window size also changes with method and region. There is no single optimal search window size for all regions for the most accurate methods (*i.e.*, RFOK, RFIDS, RFOKRFIDS and RFRFOKRFIDS) in terms of RMAE and RRMSE. These findings are consistent with previous studies (Li et al., 2011a; Li et al., 2010).

Overall, the best search window size is 5 while the best method is RFOKRFIDS and RFRFOKRFIDS in the northwest region. In the northeast region, the effects of the search window size were marginal; and the most accurate prediction was from RFOKRFIDS with a search window size of 14. In the southwest region, the difference in predictive accuracy among methods is marginal, and RFOK with a search window size of 5 is the most accurate. The most accurate methods reduced the prediction error by up to 7%. If only one method needs to be selected for all regions with a single search window size, we would recommend RFOKRFIDS with a size of 5 based on the RMAE and RRMSE.

4.6. VISUAL EXAMINATION OF THE PREDICTIONS OF THE METHODS

4.6.1. Northwest

The spatial patterns of predicted seabed sand content from RFOKRFIDS mainly reflect the effect of bathymetry (Fig. 2.5) and its related variables like geomorphic features as the patterns were similar to those of the geo-features (Fig. 4.3) identified in a previous study (Heap and Harris, 2008). This is because of that bathymetry is the most important variable for random forest (Fig. 4.4). The influences of geomorphic features, such as

shelf, reef, bank/shoals, and basins are obvious. Sand content is high on shelf and low on bank/shoals and basins. The predicted patterns of seabed sand content capture the relatively sandy nature of shelf.

Distance to coast also played a role as sand content decreases with the distance from the coastline to the deeper water. An interactive effect exists between bank/shoals and distance to coast since sand content on the bank/shoals is higher on the shelf than on the slope further away from the coastline. The sand content of the basins is relatively stable, having lower values than the shelf, but higher than that the slope, suggesting that there is no apparent interaction between the basins and their distance to the coast.

The contributions of slope and relief are minimal. They are only detectable around reefs where the sand content is lower than that on the reefs.

Some artefacts are also associated with the predictions, such as faint horizontal and vertical lines reflecting the effects of latitude and longitude. This implies that in these areas sand content was more related to longitude and latitude than other secondary variables; and this may also result from the coincidence of the spatial patterns of samples along lines of latitude and longitude, as discussed in the previous studies (Li et al., 2011b).

While relatively less distinct compared to IDS, ‘bull’s eye’ patterns are still detectable in the predictions of RFOKRFIDS. The weak ‘bull’s eye’ patterns at some sample locations are due to the application of IDS, because it is a common feature associated with predictions of IDW as has been discussed in previous studies (Li et al., 2010).

The major patterns of sand content are opposite to those of mud content as predicted in our previous studies (Li et al., 2011a; Li et al., 2011c; Li et al., 2010). This is to be expected, as they are usually negatively correlated.

Given the data variation of sand content, the accuracy of the predictions in this region is less accurate than that of the predictions of mud content in the southwest region (Li et al., 2011a; Li et al., 2010) according to the relationship between the accuracy and data variation. This could be attributed to that the correlations between sand content and the secondary variables are weaker than the correlations between mud content and the secondary variables. Nevertheless, the accuracy of the predictions in this region is comparable with previous studies as reviewed by Li and Heap (2008; 2011) in terms of both RMAE and RRMSE (Figs. 4.1 and 4.2).

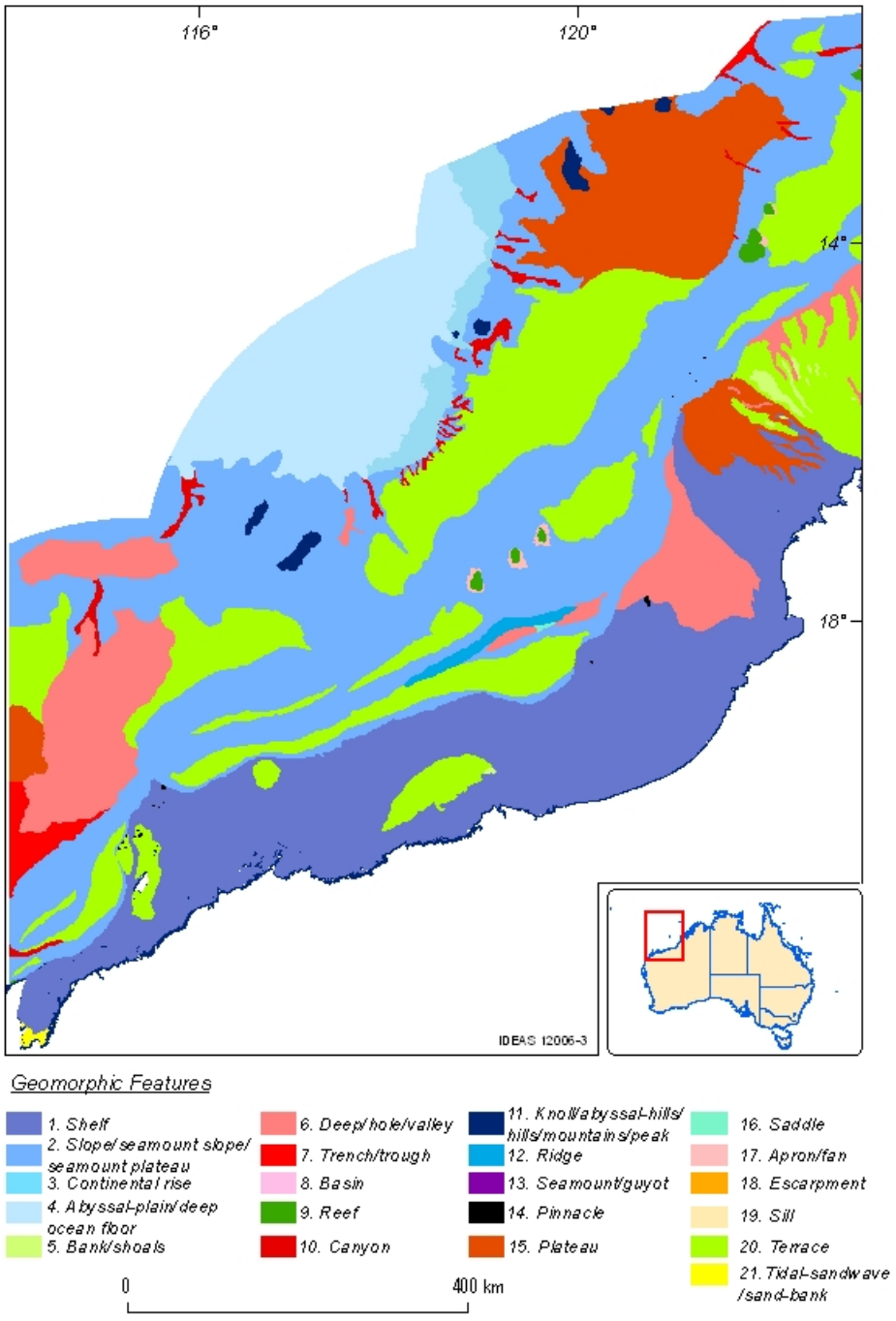


Figure 4.3. *The spatial distribution of geomorphic features in the northwest region (Li et al., 2010).*

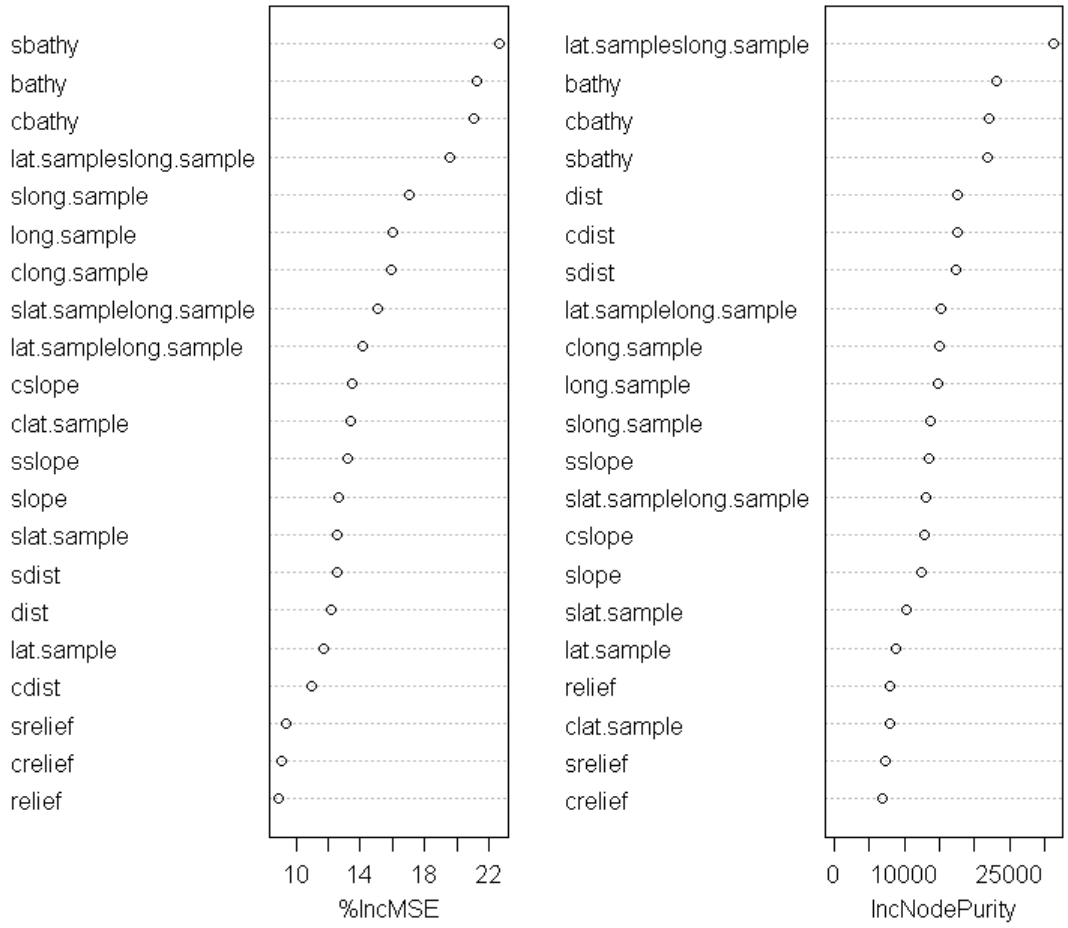


Figure 4.4. *Variable importance produced by random forest in the northwest region (Li et al., 2010).*

4.6.2. Northeast

The spatial patterns of predicted seabed sand content from RFOKRFIDS primarily reflect the effect of bathymetry (Fig. 2.5) and its related variables like geomorphic features as the patterns were similar to those of the geo-features (Fig. 4.5) identified by Heap and Harris (2008). This is because of that bathymetry is the most important variable for random forest (Fig. 4.6). The influences of geomorphic features, such as shelf and reef are obvious. Sand content is high on shelf and low on reefs when they locate on shelf. The predicted patterns of seabed sand content capture the relatively sandy nature of shelf. These phenomena are similar to what was observed in the northwest region.

Distance to coast also played a role as sand content decreases away from the coast into deeper water.

The contributions of slope and relief are minimal and were only detectable around reefs where the sand content is lower than the level associated with reefs in the northwest region.

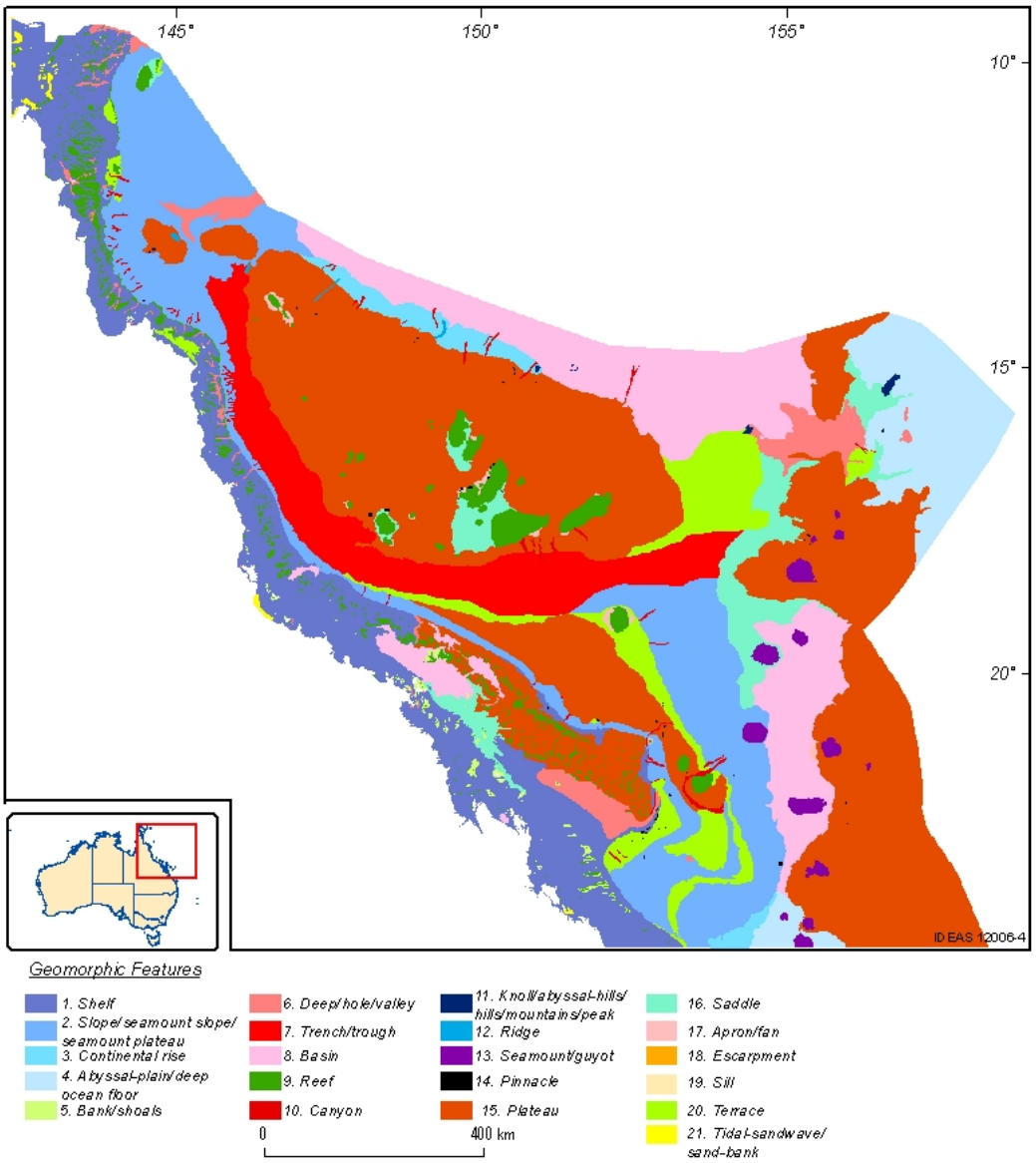


Figure 4.5. *Spatial distribution of geomorphic features in the northeast region.*

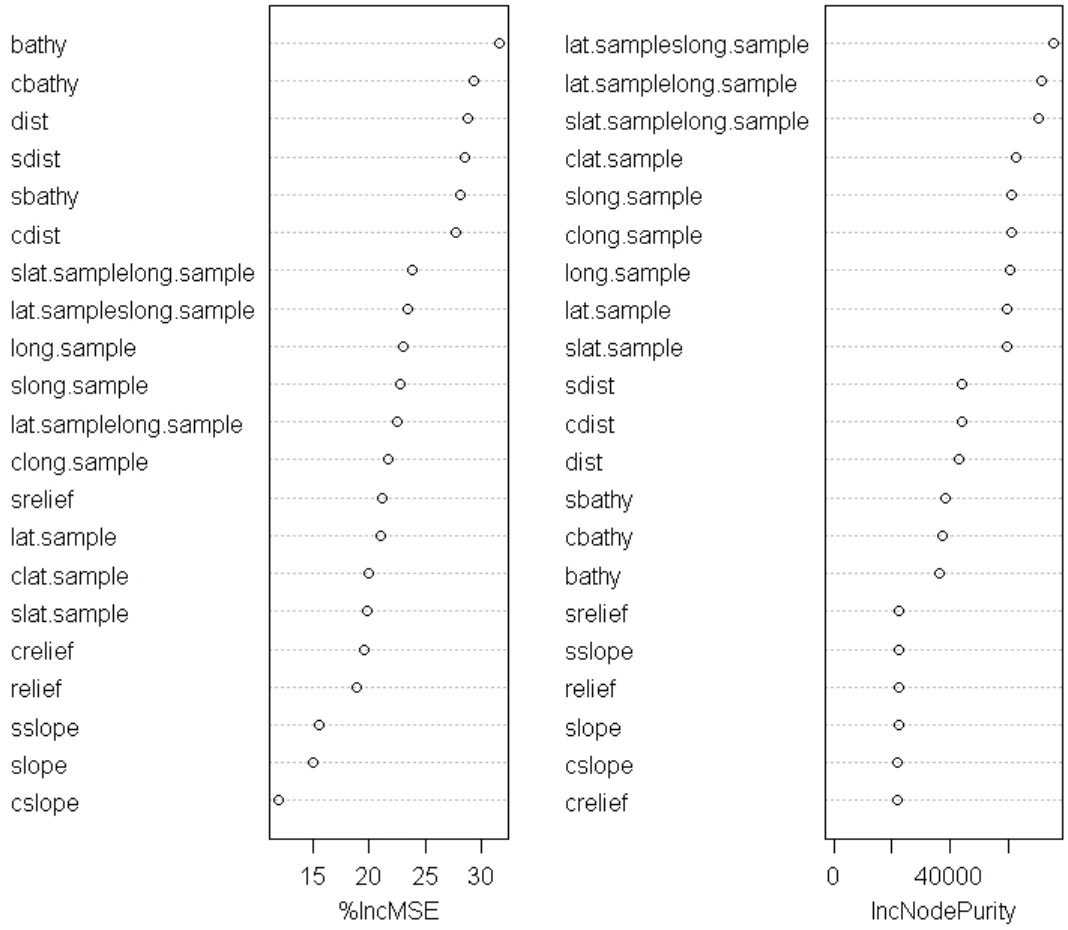


Figure 4.6. *Variable importance produced by random forest in the northeast region (Li et al., 2010).*

Some artefacts are also associated with the predictions. They are horizontal and vertical lines reflecting the effects of latitude and longitude and north-eastern banding patterns demonstrating the interactive impact of latitude and longitude. The explanations for the effects of longitude and latitude on the observed patterns in the northwest region and in our previous studies (Li et al., 2011b) are equally applicable to the observed patterns here.

While relatively less distinct compared to IDS, 'bull's eye' patterns are faint but detectable in the predictions of RFOKRFIDS.

The major patterns of sand content are opposite to those of mud content as predicted in our previous studies (Li et al., 2011a; Li et al., 2011c; Li et al., 2010), which is expected because of their negative relationship as discussed above.

The accuracy of predictions for this region is comparable to that for the northwest region. Given the relatively small data variation, it is relatively less accurate than that of the predictions of mud content in the southwest region (Li et al., 2011a; Li et al., 2010), which may be the resultant of weaker correlations between sand content and the

secondary variables in comparison with those for mud content. Nevertheless, the accuracy of the predictions in this region is comparable with previous studies as reviewed by Li and Heap (2008; 2011) in terms of both RMAE and RRMSE (Figs. 4.1 and 4.2).

4.6.3. Southwest

As with the above two regions, the spatial patterns of predicted seabed sand content from RFOKRFIDS in this region primarily reflect the effect of bathymetry (Fig. 2.5) and its related variables like geomorphic features because the patterns were similar to those of the geo-features (Fig. 4.7) identified in a previous study (Heap and Harris, 2008). This is because bathymetry is the most important variable for random forest (Fig. 4.8). As a result, the influences of geomorphic provinces are obvious.

Sand content is high on the shelf, lower on the slope, and lowest on the deep sea floor. The predicted patterns of seabed sand content capture the sandy nature of shelf, and are similar to what was observed in the other two regions. The impact of Perth Canyon is obvious due to its lower sand content relative to the surrounding area.

The role of distance to coast is not apparent because sand content decreases with distance from the coastline to the deeper water in areas between 29-34^{o} latitude, but this trend is not obvious in other areas.

The contributions of slope and relief are minimal and are only detectable near 35^{o} latitude where the sand content is higher than that of the adjacent area. A few faint NW-SE and one NE-SW trending linear features mainly reflect the influence of slope and relief (Figs. 2.6 and 2.8), which was also observed in the predictions of mud content (Li et al., 2011a; Li et al., 2010).

There are weak horizontal and vertical lines reflecting the effects of latitude and longitude and north-eastern banding patterns demonstrating the interactive impact of latitude and longitude. The explanations for the effects of longitude and latitude on the observed patterns given above and in our previous studies (Li et al., 2011b) are equally applicable to the observed patterns here.

While less distinct compared to IDS, ‘bull’s eye’ patterns are faint but still detectable in the predictions of RFOKRFIDS.

The major patterns of sand content is opposite to those of mud content as predicted in our previous studies (Li et al., 2011a; Li et al., 2011c; Li et al., 2010) as observed in the other two regions.

The accuracy of the predictions in this region is relatively higher than in the northwest and northeast regions if their associated data variation is considered (Figs. 4.1 and 4.2) although its absolute value is higher than those in the other two regions. This is probably because the correlation between sand content and bathymetry in this region is higher than in the other two regions. It is relatively less accurate than that of the predictions of mud content in the southwest region (Li et al., 2011a; Li et al., 2010),

which may again resulted from the weaker correlations between sand content and the secondary variables in comparison with that for mud content as discussed above. However, the accuracy of the predictions in this region is considerably higher than what observed in previous studies as reviewed by Li and Heap (2008; 2011) in terms of both RMAE and RRMSE (Figs. 4.1 and 4.2).

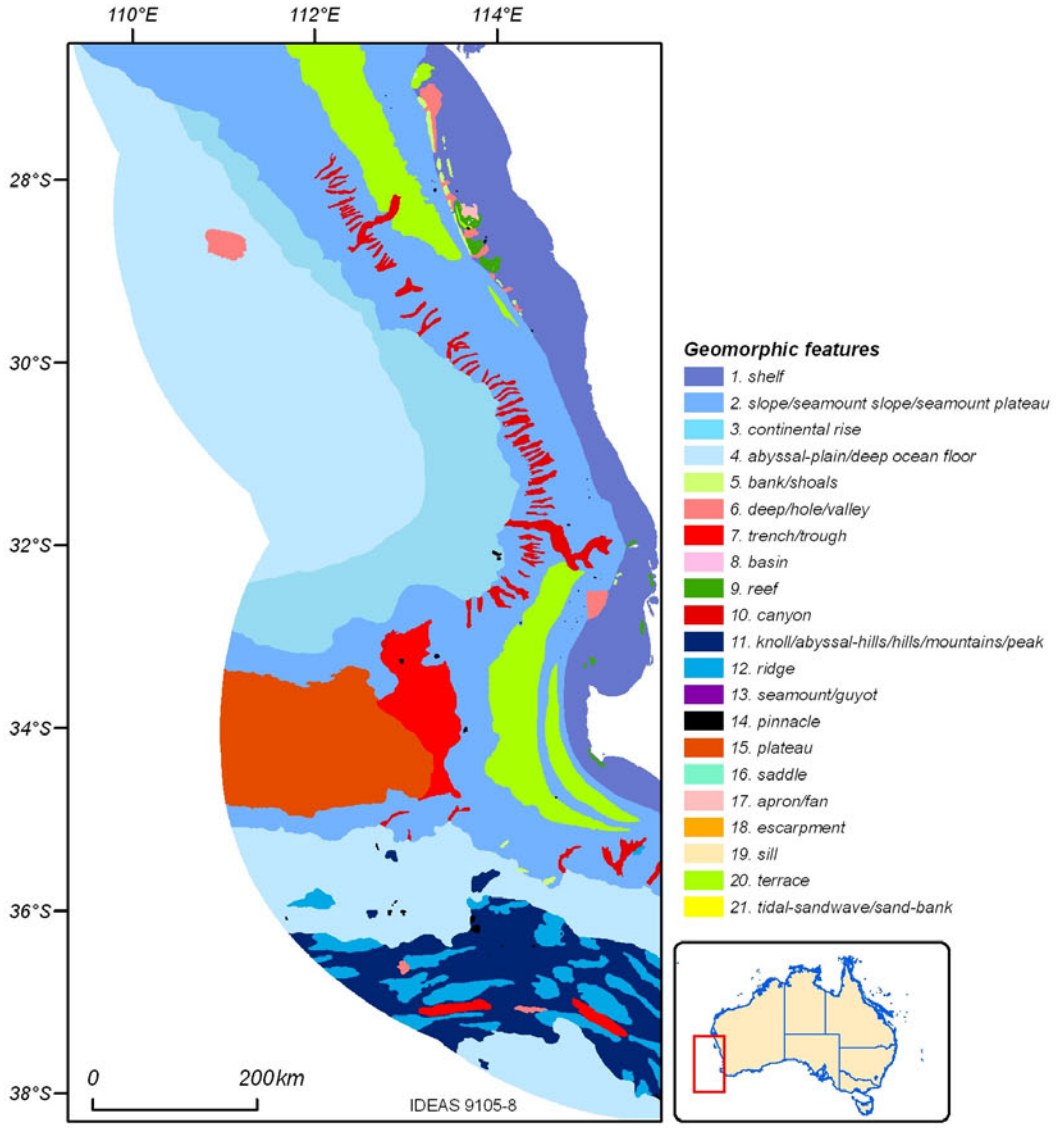


Figure 4.7. *Spatial distribution of geomorphic features in the southwest region.*

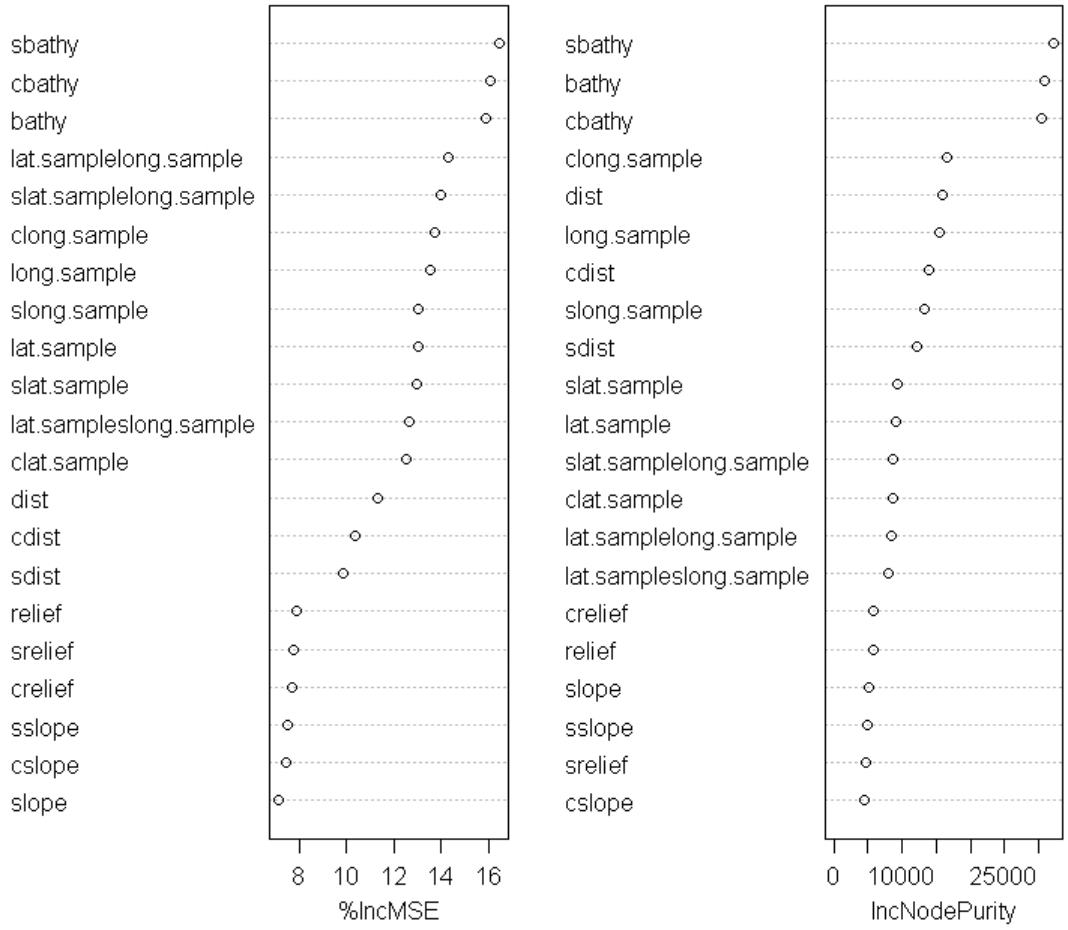


Figure 4.8. *Variable importance produced by random forest in the southwest region (Li et al., 2010).*

4.7. LIMITATIONS

This study is part of a continuing effort to identify the most accurate methods for predicting the spatial distribution of seabed sediments following our previous studies (Li et al., 2011a; Li et al., 2010). Given that all samples were from the MARS database, this study shares similar limitations as discussed in previous studies in regarding the data sources, data projection, data sampling and time span of samples collected, although sample size has been increased since 2008. The problems associated with the automation of the experiment are also existing as discussed by Li et al. (2010). Despite these limitations, the predictions achieved in this study are more accurate than those of the control method and are comparable or even more accurate than those reported in previous studies (Li and Heap, 2008, 2011).

Chapter 5. Conclusions

The best performing method changes with regions. Of the 18 compared methods, RFIDS and RFOK are the most accurate in all three regions. All other machine learning methods (i.e., SVM, LSVM, GRNN and BDT) performed better when combined with IDS or OK. This finding confirms the effectiveness of combining machine learning methods with commonly used spatial interpolation methods such as OK or IDS, and supports an alternative source of methods for the discipline of spatial statistics.

The choice of the input secondary variables affects the performance of RF, RFOK and RFIDS and such effects vary with the method and study region. Overall, of the 36 combinations of input secondary variables (4 levels), method (3 levels) and region (3 levels), RFIDS, 6RFIDS and RFOK were among the most accurate methods in all three regions. Model selection is essential for RF in order to find an optimal predictive model.

The effects of model averaging depend on methods averaged, regions and error measurements. Overall, model averaging marginally improved the prediction accuracy in the southwest region, while in other regions no apparent effects were observed. RFIDS, RFOKRFIDS and RFRFOKRFIDS perform better than other methods in all three regions in terms of both RMAE and RRMSE. Of these three methods, RFOKRFIDS is the most accurate method.

The effects of the choice of *mtry* on the performance of RF, RFOK, RFIDS, RFOKRFIDS and RFRFOKRFIDS are small and negligible in this study. Therefore, one choice of *mtry* (4) can be used for all regions.

The optimal search window size is also region-dependent. If the best method with its optimal search window size for each region needs to be found, then similar experiments for each region will need to be run. If only one method needs to be selected for all regions with a single search window size, the method is RFOKRFIDS with a size of 5 based on the RMAE and RRMSE.

Given that the performance of the methods tested changes with region, input secondary variables, *mtry* and search window size, we recommend that the most accurate method could only be identified by considering all relevant impact factors for each study region in order to achieve the most accurate predictions. These factors may include those factors recommended in our previous studies (Li et al., 2011a; 2010).

The small number of secondary variables used in this study is a limitation for further improving the predictive accuracy. Other secondary variables should be identified and used, and thus the artefacts associated with influence of latitude and longitude could be removed from the predictions and the predictive accuracy could be further improved. Given that it is difficult to acquire secondary variables covering the whole AEEZ, we may shift our focus onto the shelf where the most industrial activities occur and where some secondary variables like oceanographic information are more readily available.

RFOKRFIDS, with a search window size of 5 and an *mtry* of 4, is recommended for predicting sand content across the AEEZ if a single method is required. The 'best' methods identified and relevant findings in this study are conditioned on the resources allocated and conditions applied in the experiment. This should be taken into account when making final decision.

This study has provided suggestions and guidelines for improving the spatial interpolations of marine environmental data in general, which can be equally applicable to terrestrial environmental variables. A more accurate physical dataset of seabed sand content for the AEEZ is expected from applying the method recommended, which would result in more accurate mapping and characterisation of seabed across the AEEZ. Consequently, it will assist in the development of management and conservation strategies of marine zones by government, industry and community.

Acknowledgements

John Wilford and Johnathan Kool from Geoscience Australia provided valuable comments on this record. Tanya Whiteway carefully proofread and also commented the record. Shoaib Burq (Geoscience Australia) prepared datasets from MARS. Chris Lawson, Brad Cook and Amanda Spalding (all from Geoscience Australia) provided bathymetry, slope and distance to coast data and producing several maps. Scott Nichol provided comments on the experimental design. This record is published with permission of the Chief Executive Officer, Geoscience Australia.

References

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.

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

Drake J.M., Randin C., Guisan A., 2006. Modelling ecological niches with support vector machines. Journal of Applied Ecology 43: 424-432.

Elith J., Graham C.H., Anderson R.P., Dulik M., Ferrier S., Guisan A., Hijmans R.J., Huettmann F., Leathwick J.R., Lehmann A., Li J., Lohmann L.G., Loiselle B.A., Manion G., Moritz C., Nakamura M., Nakazawa Y., Overton J.M., Peterson A.T., Phillips S.J., Richardson K., Scachetti-Pereira R., Schapire R.E., Soberon J., Williams S., Wisz M.S., Zimmermann N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151.

Friedman J. H., 1999. Stochastic Gradient Boosting. Technical report, Dept. of Statistics, Stanford University.

Goovaerts P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press: New York.

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

Guyon I., Lemaire V., Boullé M., Dror G., Vogel D., 2009. Analysis of the KDD Cup 2009: Fast scoring on a large Orange customer database. In JMLR: Workshop and Conference Proceedings, Lawrence N. (ed.); 1-22.

Heap A.D., Harris P.T., 2008. Geomorphology of the Australian margin and adjacent seafloor. Australian Journal of Earth Sciences 55: 555-585.

Hoeting J.A., Madigan D., Raftery A.E., Volinsky C.T., 1999. Bayesian model averaging: a tutorial. Statistical Science 14: 382-417.

Knudby A., Brenning A., LeDrew E., 2010. New approaches to modelling fish-habitat relationships. Ecological Modelling 221: 503-511.

Legendre P., Legendre L., 1998. Numerical Ecology. ELSEVIER: Amsterdam.

Li J., 2011. Novel spatial interpolation methods for environmental properties: using point samples of mud content as an example. The Survey Statistician: The Newsletter of the International Association of Survey Statisticians No. 63: 15-16.

Li J., Heap A., 2008. A Review of Spatial Interpolation Methods for Environmental Scientists. Geoscience Australia: Canberra; 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., 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: Canberra; 69.

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

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

Li J., Heap A.D., Potter A., Huang Z., Daniell J., 2011d. Seabed mud content across the Australian continental EEZ 2011. Geoscience Australia.

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: Canberra; 146.

Liaw A., Wiener M., 2002. Classification and regression by radomForest. R News 2: 18-22.

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

McArthur M.A., Brooke B., Przeslawski R., Ryan D.A., Lucieer V.L., Nichol S., McCallum A.W., Mellin C., Cresswell I.D., Radke L.C., 2009. A review of surrogates for marine benthic biodiversity. Geoscience Australia, Record 2009/42. Geoscience Australia, Canberra. 61pp.

Okun O., Priisalu H., 2007. Random forest for gene expression based cancer classification: overlooked issues. In Pattern Recognition and Image Analysis: Third Iberian Conference, IbPRIA 2007 Martí J., Benedí J.M., Mendonça A.M., Serrat J. (eds.); Lecture Notes in Computer Science: Girona, Spain; 4478: 4483-4490.

Olden J.D., Jackson D.A., 2002. A comparison of statistical approaches for modelling fish species distributions. Freshwater Biology 47: 1976-1995.

Özesmi S.L., Tan C.O., Özesmi U., 2006. Methodological issues in building, training, and testing artificial neural networks in ecological applications. Ecological Modelling 195: 83-93.

Pino-Mejías R., Cubiles-de-la-Vega M.D., Anaya-Romero M., Pascual-Acosta A., Jordán-López A., Bellinfante-Crocci N., 2010. Predicting the potential habitat of oaks with data mining models and the R system. Environmental Modelling & Software 25: 826-836.

Pitcher C.R., Doherty P.J., Anderson T.J., 2008. Seabed environments, habitats and biological assemblages. In The Great Barrier Reef: biology, environment and management, Hutchings P., Kingsford M., Hoegh-Guldberg O. (eds.); CSIRO Publishing: Collingwood; 377.

Post A.L., 2008. The application of physical surrogates to predict the distribution of marine benthic organisms. Ocean & Coastal Management 51: 161-179.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna.

Raftery A.E., Gneiting T., Balabdaoui F., Polakowski M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review 133: 1155-1174.

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.

Venables W.N., Ripley B.D., 2002. Modern Applied Statistics with S-Plus. Springer-Verlag: New York.

Wolpert D., Macready W., 1997. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation 1: 67-82.

Appendix A. Description of BDT

A.1. BOOSTED DECISION TREE (BDT)

Boosted Decision Tree (BDT) (Friedman, 1999) consists of a series of single decision trees, the first of which is fitted to the data. The residuals (error values) from the first tree are then fed into the second tree which attempts to reduce the error. This process is repeated through a series of successive trees. The final predicted value is formed by adding the weighted contribution of each tree.

A.2. GENERAL REGRESSION NEURAL NETWORK (GRNN)

General Regression Neural Network (GRNN) (Specht, 1990) has an input layer (one neuron for each predictor), a hidden layer (one neuron for each training sample), a pattern layer (two neurons; one is the denominator summation unit, the other is the numerator summation unit) and output layer (one neuron). A hidden neuron computes the distance between the point being evaluated and a training sample, then applies a radial basis function (e.g., Gaussian function) using the sigma value to the distance to compute the weight (influence) for each point. The output layer divides the value accumulated in the numerator summation unit by the value in the denominator summation unit and uses the result as the predicted target value. Please refer to Li et al. (2010) for further information.

Appendix B. Statistical and mathematical Modelling

In this appendix we discuss a number of issues in relation to variogram modelling and statistical and mathematical modelling. As indicated by Li and Heap (2008), geostatistics has a few assumptions and limitations. Geostatistical methods assume stationarity of data, which is often not true, although this assumption can be relaxed with specific forms of kriging. The definition of the required variogram model can be time consuming and somewhat subjective; and definition of neighbourhoods is also required and is difficult to do objectively.

B.1. DATA TRANSFORMATION

Geostatistical methods assume data stationarity of the primary variable. Distributions of sand content are skewed and non-normal in the three regions (Figs. B.1), so several transformations were tested and the appropriate transformation was identified for each variable in each region (Table B.1). The distributions of the transformed data for most variables are approximately normal, but not for all. Although other transformation methods may be able to produce more normalised data (Legendre and L., 1998), but appropriate back-transformation methods are not readily available (Legendre and Legendre, 1998; Li et al., 2010), which thwarts the use of these methods when predictions need to be made on the original scale.

The secondary information was also transformed to: 1) explore the correlation between sand and secondary information, 2) find the linear relationship between sand and transformed secondary information and 3) find the linear relationship between transformed sand content (normalised) and transformed secondary information (Table. B.1). The best transformations of the secondary variables are identical for the last two purpose, so they were combined and presented in Table B.1. The information in Table B.1 was based on the analyses in section B.2 and the relationships between sand content and the secondary variables are also illustrated in section B.2.

Table B.1. Normalised sand and transformed secondary variables to normalise the data of sand content and to improve their correlation with sand content.

REGION	SAND CONTENT	BATHYMETRY	DISTANCE TO COAST	SLOPE	RELIEF
NW	No	sqrt	No	sqrt	sqrt
NE	Squared(sand/100)	sqrt	Squared	sqrt	sqrt
SW	arcsin	sqrt	No	sqrt	sqrt

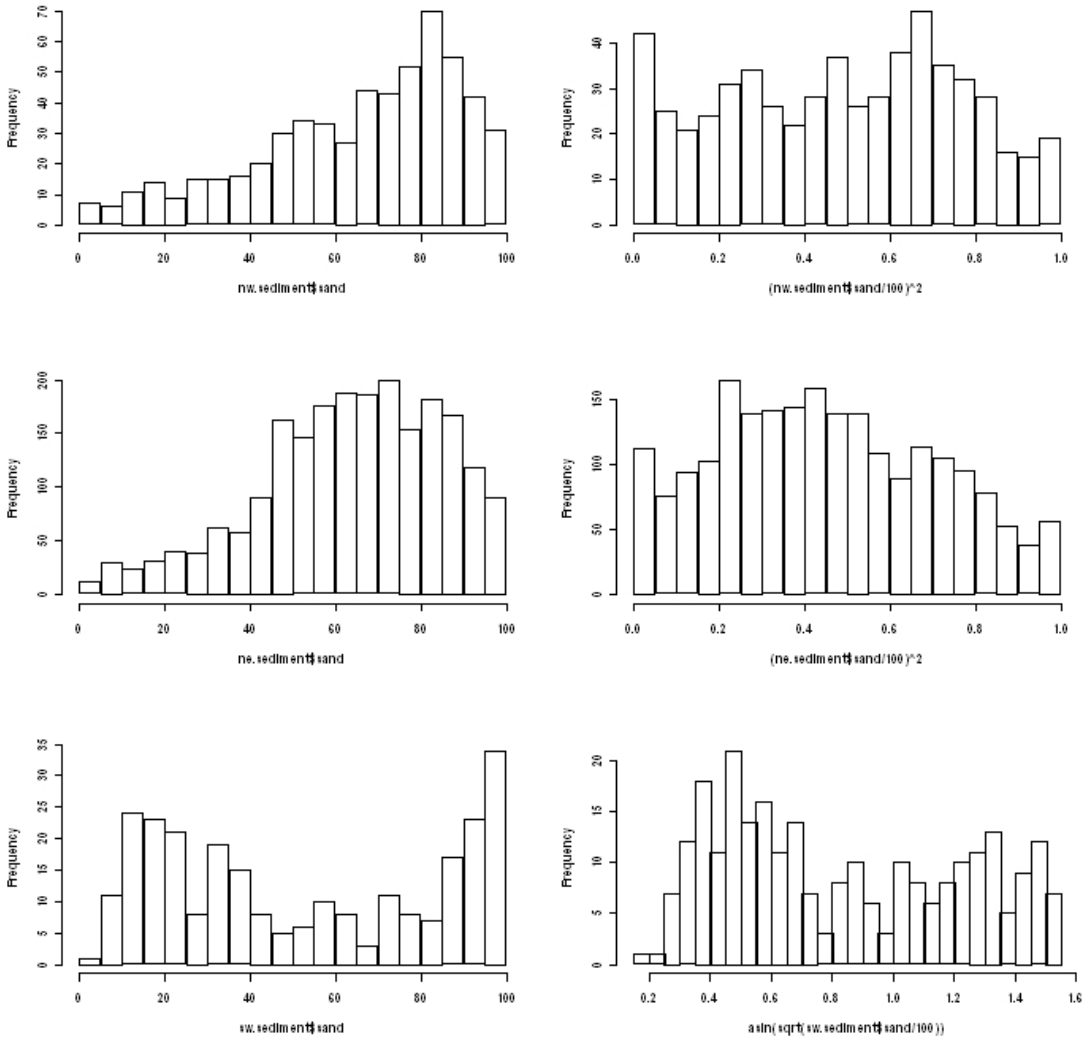


Figure B.1. *Data distribution of sand content in the three study regions before and after transformation.*

B.2. CORRELATION BETWEEN SAND CONTENT AND SECONDARY VARIABLES

B.2.1. Correlation between untransformed data

Correlation between the primary and secondary variables is critical for spatial interpolation methods 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, the correlation of sand content and the secondary variables changes with variables and regions in terms of Pearson's product-moment correlation (r) (Table B.2) and Spearman's rank correlation rho (ρ) (Table B.3). Bathymetry has a high correlation with sand content, particularly in the southwest region (Tables B.2 and B.3, and Fig. B.2), except the northeast region in terms of ρ. Distance-to-coast is highly correlated with sand content in the northwest and southwest

regions and not significantly in the northeast region (Tables B.2 and B.3, and Fig. B.3) in terms of r and ρ. Slope displayed a similar pattern as bathymetry in terms of *r* and *ρ* (Tables B.2 and B.3, and Fig. B.4). Relief is significantly correlated with sand content in all three regions in terms of *r* and *ρ* (Tables B.2 and B.3, and Fig. B.5). In most cases, the relationship between sand content and the secondary variables is non-linear (Figure B.2-B.5).

Table B.2. Pearson's product-moment correlation of sand content with bathymetry, distance-to-coast, slope and relief in the three study regions. The test for correlation between paired samples was conducted in R (R Development Core Team, 2010).

VARIABLE	REGION	T VALUE	DEGREE OF FREEDOM	CORRELATION COEFFICIENT	P-VALUE
Bathymetry	NW	8.8400	572	0.3467	0.0000
	NE	5.8531	2155	**0.1251**	0.0000
	SW	10.7260	262	**0.5539**	0.0000
Distance-to-coast	NW	-10.3758	572	**-0.3980**	0.0000
	NE	-0.0208	2155	-0.0004	0.9834
	SW	-7.6884	262	-0.4304	0.0000
Slope	NW	-5.2696	572	-0.2152	0.0000
	NE	-1.5826	2155	-0.0341	0.1137
	SW	-6.4106	262	-0.3694	0.0000
Relief	NW	-5.8793	572	-0.2387	0.0000
	NE	-2.6251	2155	-0.0565	0.0087
	SW	-7.4476	262	-0.4193	0.0000

Table B.3. Spearman's rank correlation ρ of sand content with bathymetry, distance-to-coast, slope and relief in the three study regions. The test for correlation between paired samples was conducted in R (R Development Core Team, 2010).

VARIABLE	REGION	*P*	P-VALUE
Bathymetry	NW	0.3189	0.0000
	NE	0.0288	0.1805
	SW	0.4424	0.0000
Distance-to-coast	NW	-0.3022	0.0000
	NE	0.0157	0.4656
	SW	-0.4344	0.0000
Slope	NW	-0.2477	0.0000
	NE	-0.0557	0.0097
	SW	-0.4174	0.0000
Relief	NW	-0.2376	0.0000
	NE	-0.0584	0.0066
	SW	-0.3921	0.0000

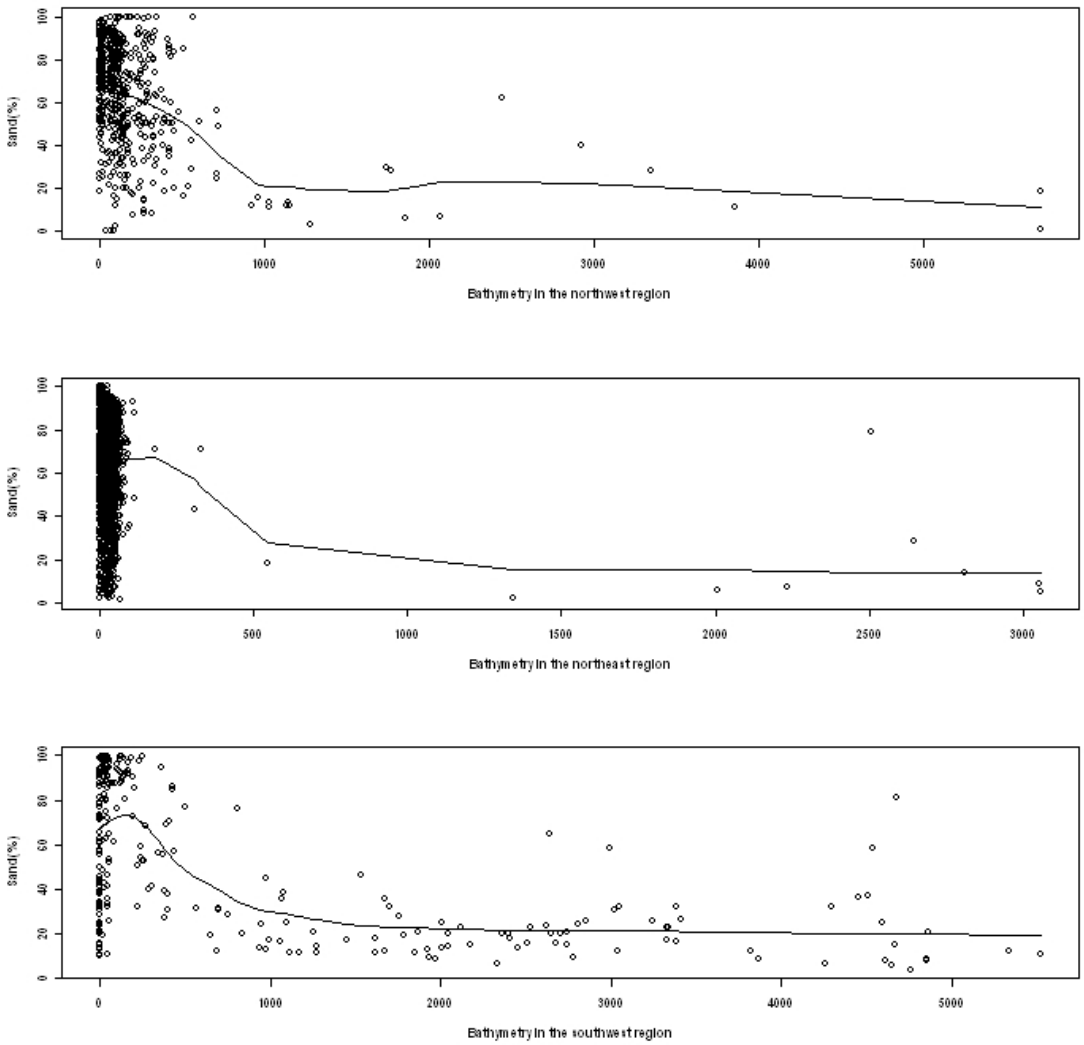


Figure B.2. *Relation between sand data and bathymetry in the three study regions and the curve was fitted using lowess in R (R Development Core Team, 2010).*

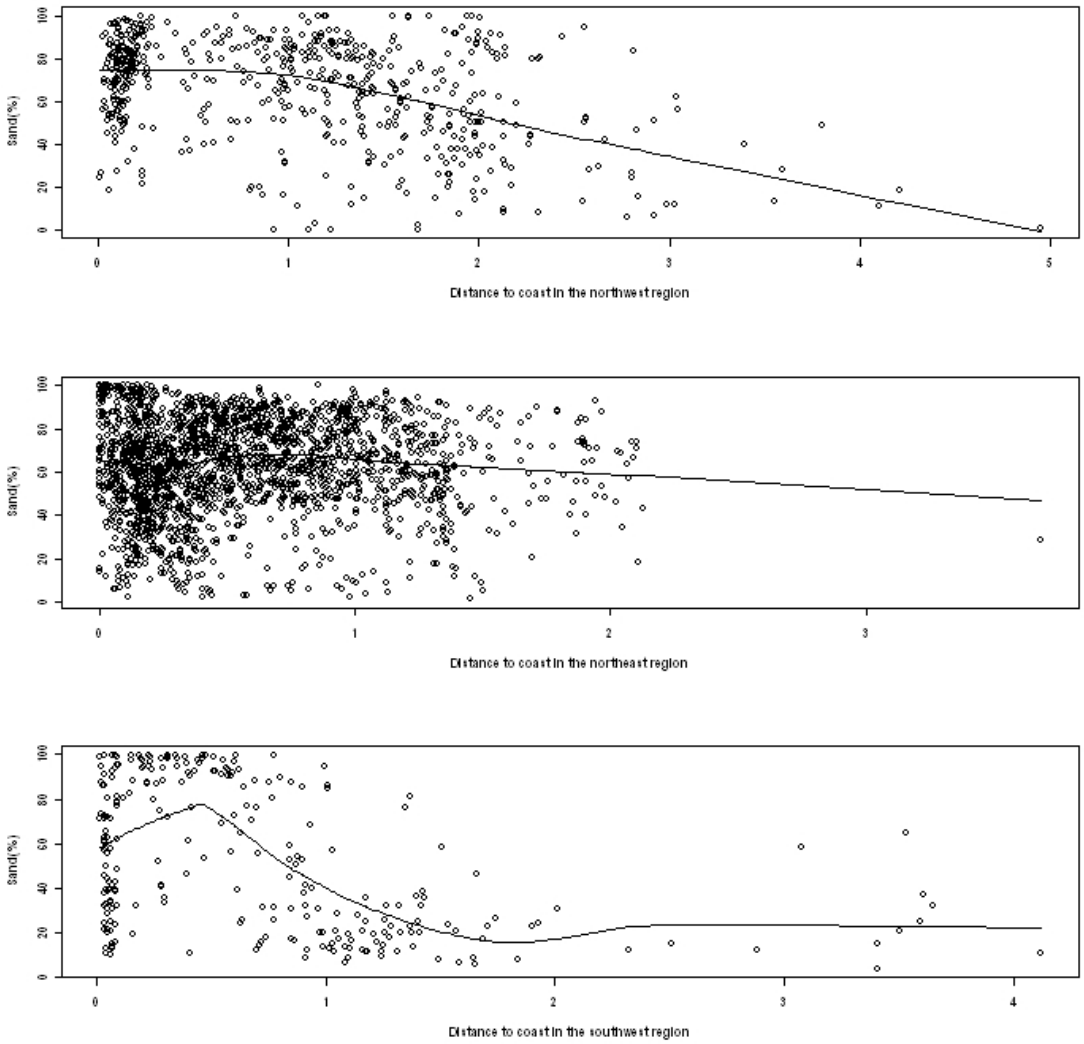


Figure B.3. *Relation between sand data and distance to coast in the three study regions and the curve was fitted using lowess in R (R Development Core Team, 2010).*

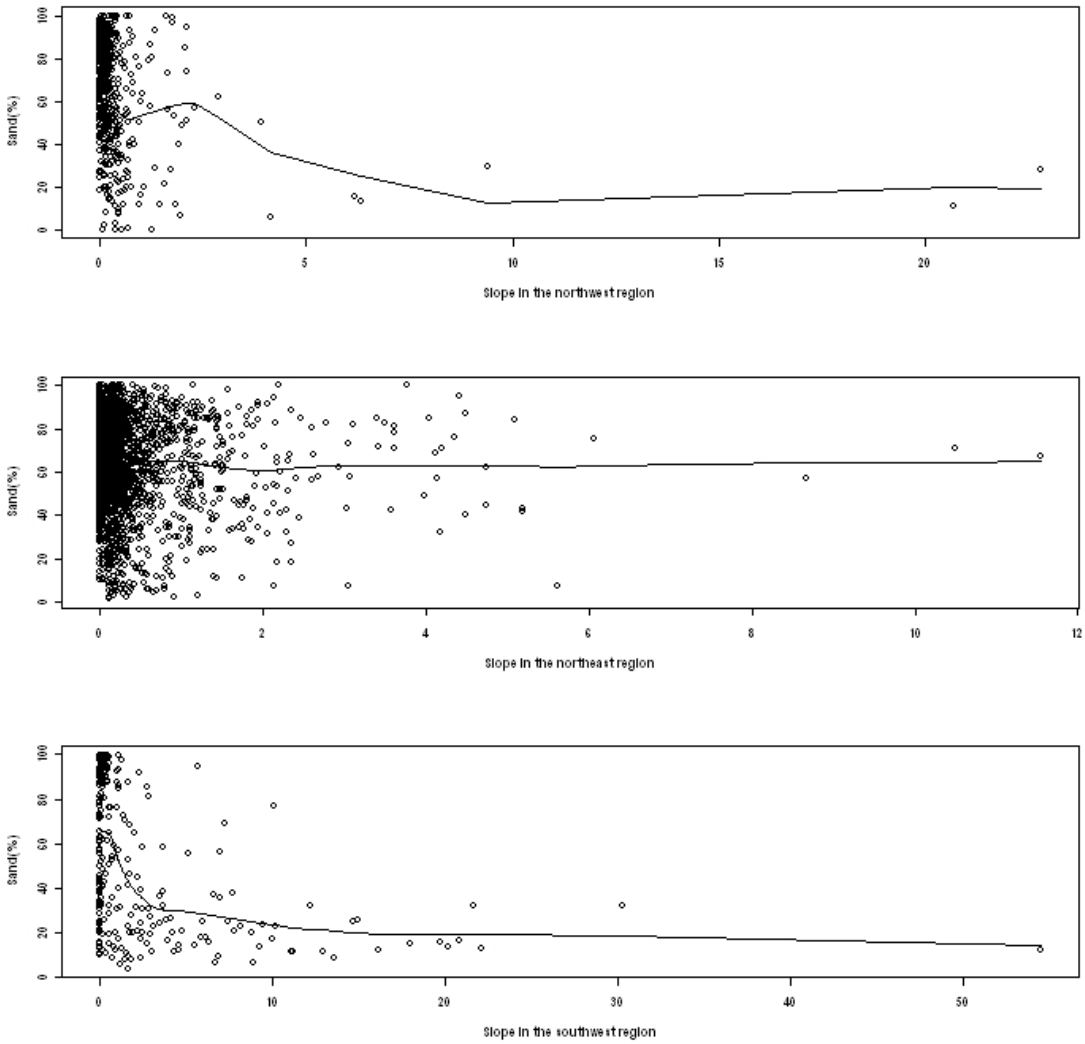


Figure B.4. *Relation between sand data and slope in the three study regions and the curve was fitted using lowess in R (R Development Core Team, 2010).*

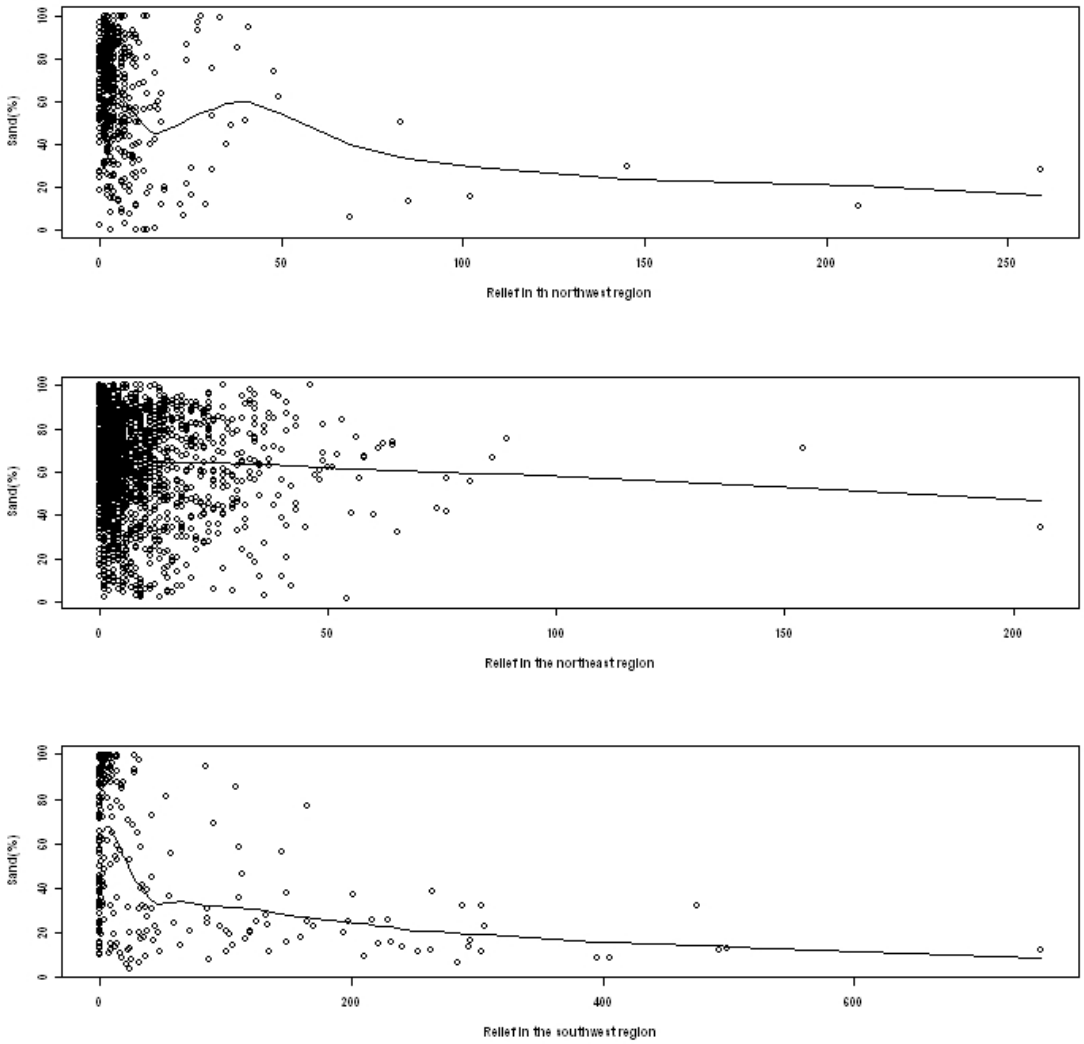


Figure B.5. *Relation between sand data and relief in the three study regions and the curve was fitted using lowess in R (R Development Core Team, 2010).*

B.2.2. Correlation between sand content with transformed secondary variables

As shown in Figure B.2-B.5, the relationship between sand content and the secondary variables is non-linear. To linearise the correlation and maximise the correlation coefficient, the best form of transformation was selected for each secondary variable (Table B.4 and Figure B.6-B.9). In comparison with Table B.2, the coefficients were increased for all variables, with an exception of that there is no best transformation for distance to coast in the northwest and southwest regions.

Table B.4. Pearson's product-moment correlation of sand content with transformed secondary variables: sqrt(abs(bathymetry)) in the three study regions, squared (distance-to-coast) in the northeast region, and sqrt(slope) and sqrt(relief) in the three study regions. The test for correlation between paired samples was conducted in R (R Development Core Team, 2010).

VARIABLE	REGION	T VALUE	DEGREE OF FREEDOM	CORRELATION COEFFICIENT	P-VALUE
Bathymetry	NW	-10.9210	572	-0.4154	0.0000
	NE	-6.5076	2155	-0.1388	0.0000
	SW	-11.1226	262	-0.5678	0.0000
Distance-to-coast	NW	-10.3758	572	-0.3980	0.0000
	NE	-1.073	2155	-0.0231	0.2834
	SW	-7.6884	262	-0.4304	0.0000
Slope	NW	-7.5993	572	-0.3028	0.0000
	NE	-2.9319	2155	-0.0630	0.0034
	SW	-8.2537	262	-0.4556	0.0000
Relief	NW	-7.4699	572	-0.2981	0.0000
	NE	-3.5433	2155	-0.0761	0.0004
	SW	-8.4438	262	-0.4639	0.0000

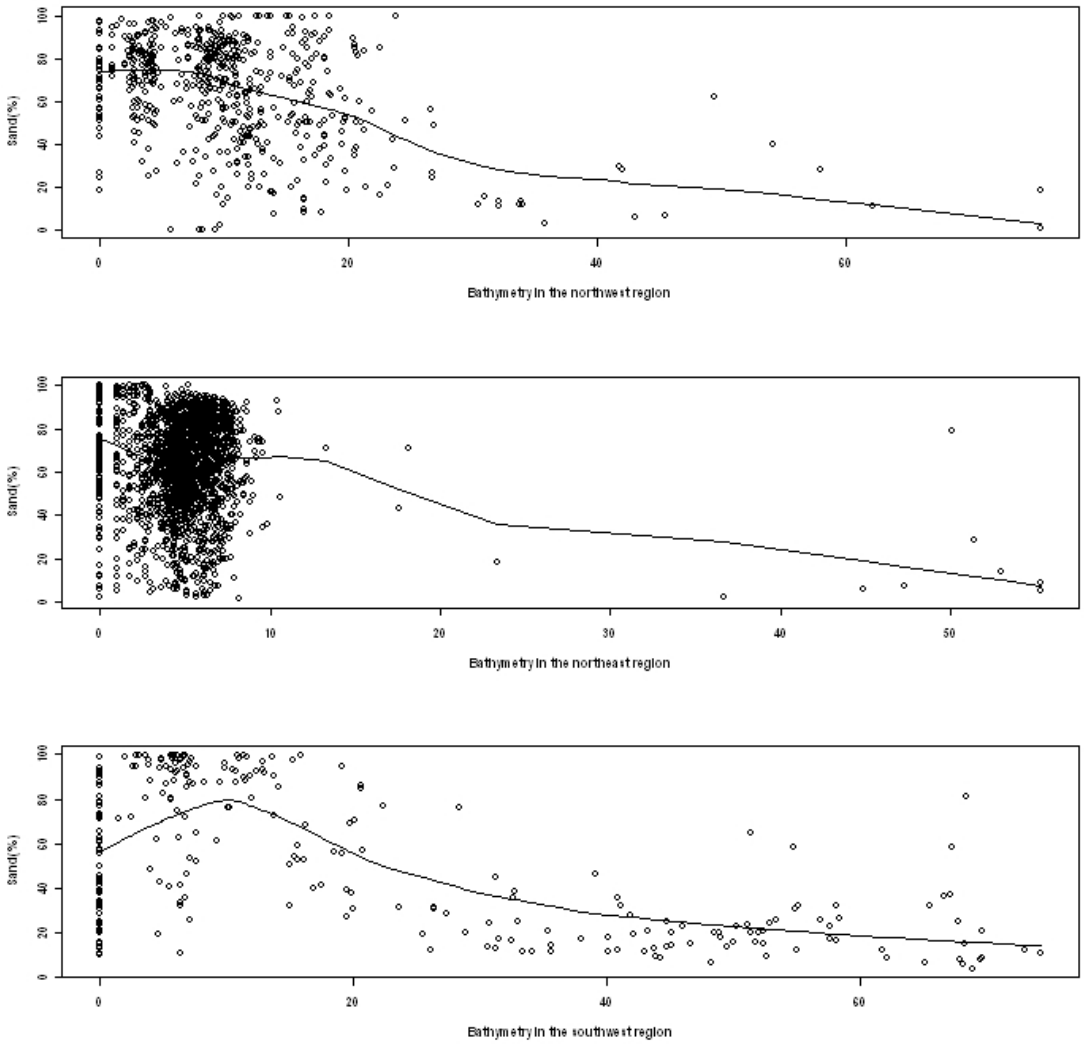


Figure B.6. *Relation between sand data and transformed bathymetry in the three study regions and the curve was fitted using lowess in R.*

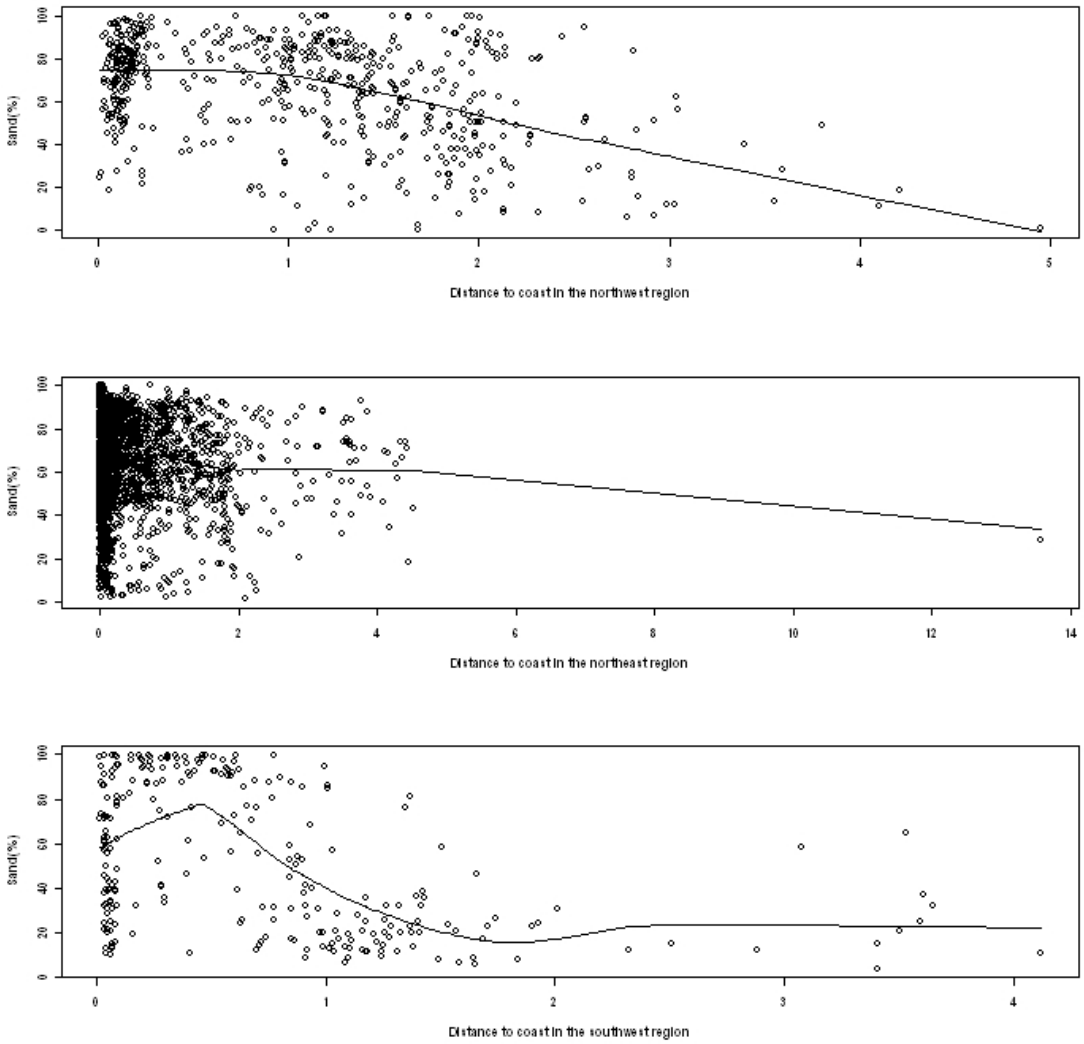


Figure B.7. *Relation between sand data and transformed distance to coast in the three study regions and the curve was fitted using lowess in R (R Development Core Team, 2010).*

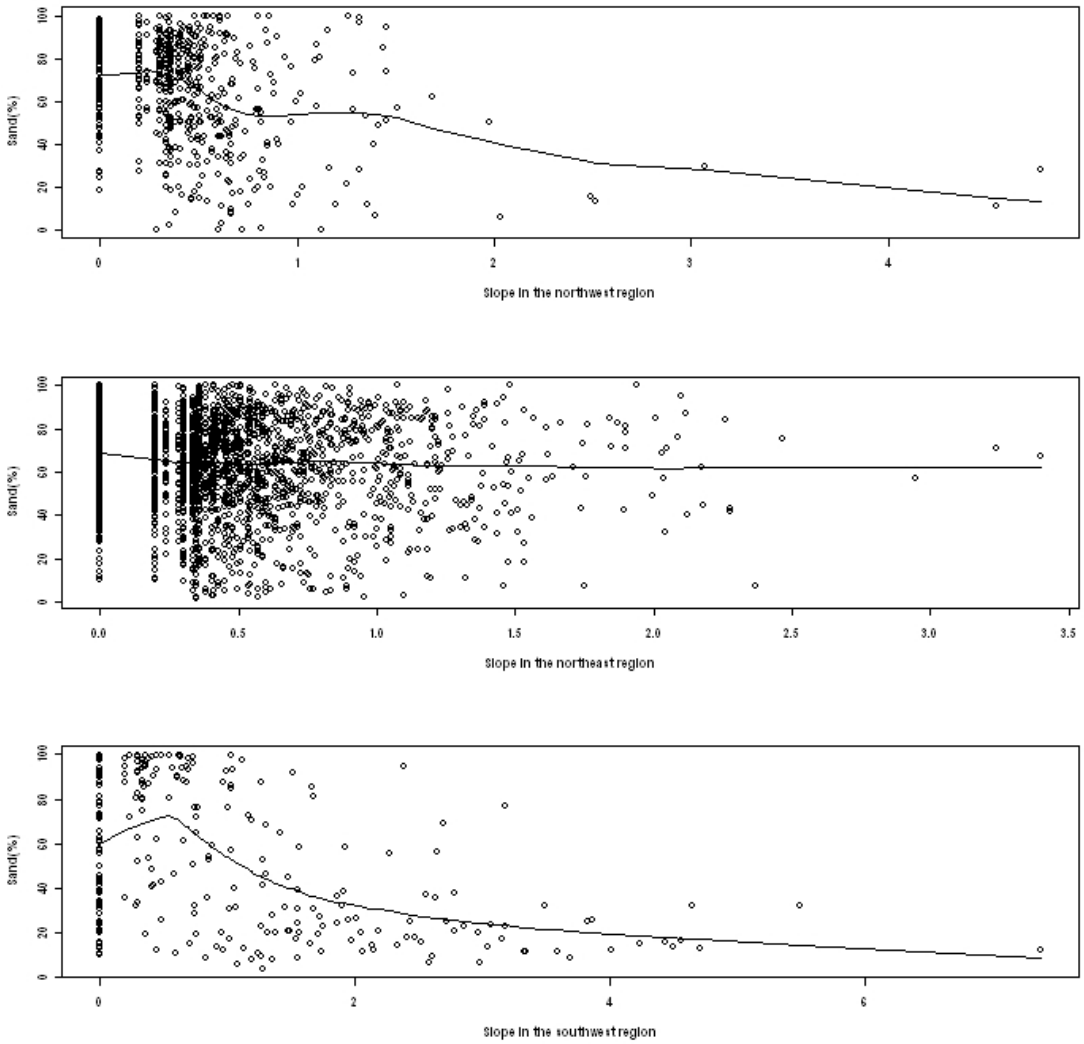


Figure B.8. *Relation between sand data and transformed slope in the three study regions and the curve was fitted using lowess in R (R Development Core Team, 2010).*

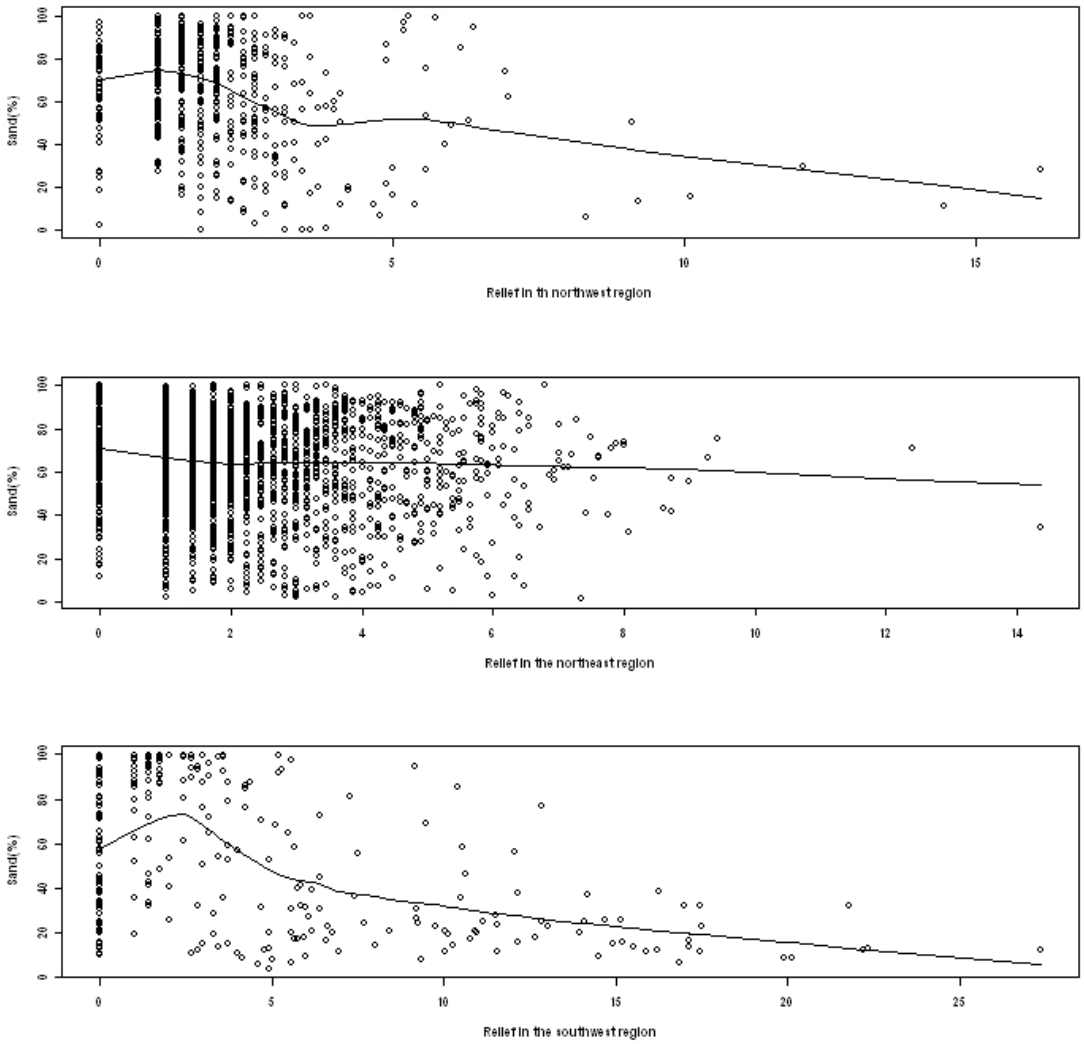


Figure B.9. *Relation between sand data and transformed relief in the three study regions and the curve was fitted using lowess in R (R Development Core Team, 2010).*

B.2.3. Correlation between normalised sand content with transformed secondary variables

The correlation of the normalised sand data and transformed secondary variables is analysed and summarised in Table B.5 and illustrated in Figure B.10-B.13. In comparison with Table B.2, the coefficients were increased for most variables, while compared with Table B.4, there is only a marginal increase in the coefficient of sand and distance to coast in the northeast region and the coefficient slightly decreased for other variables in northeast and southwest regions. Despite of this, the normalised sand data was used to meet the assumption of relevant methods.

Table B.5. Pearson's product-moment correlation of normalised sand content with transformed secondary variables: sqrt(abs(bathymetry)) in the three study regions, squared(distance-to-coast) in the northeast region, and sqrt(slope) and sqrt(relief) in the three study regions. The test for correlation between paired samples was conducted in R (R Development Core Team, 2010).

VARIABLE	REGION	SAND (x)	T VALUE	DEGREE OF FREEDOM	CORRELATION COEFFICIENT	P-VALUE
Bathymetry	NW	none	-10.921	572	-0.4154	0.0000
	NE	squared	-5.8516	2155	-0.1251	0.0000
	SW	asin(sqrt(x /100))	-10.8804	262	-0.5593	0.0000
Distance-to-coast	NW	none	-10.3758	572	-0.3980	0.0000
	NE	squared	-1.6616	2155	-0.0358	0.0967
	SW	asin(sqrt (x /100))	-7.6302	262	-0.4277	0.0000
Slope	NW	none	-7.5993	572	-0.3028	0.0000
	NE	squared	-2.6025	2155	-0.0560	0.0093
	SW	asin(sqrt (x /100))	-8.0847	262	-0.4482	0.0000
Relief	NW	none	-7.4699	572	-0.2981	0.0000
	NE	squared	-3.0233	2155	-0.0650	0.0025
	SW	asin(sqrt (x /100))	-8.2344	262	-0.4548	0.0000

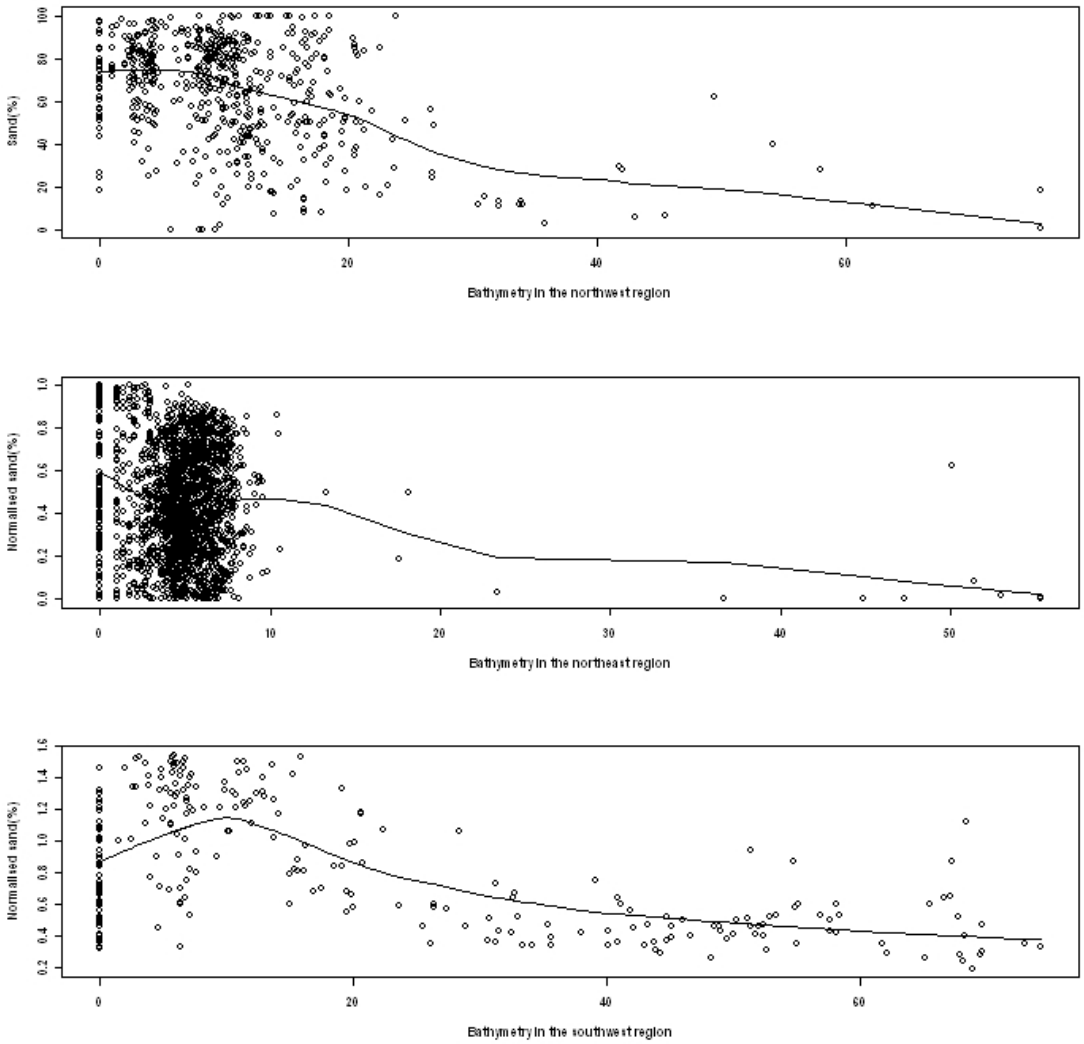


Figure B.10. *Relation between normalised sand data and transformed bathymetry in the three study regions and the curve was fitted using lowess in R (R Development Core Team, 2010).*

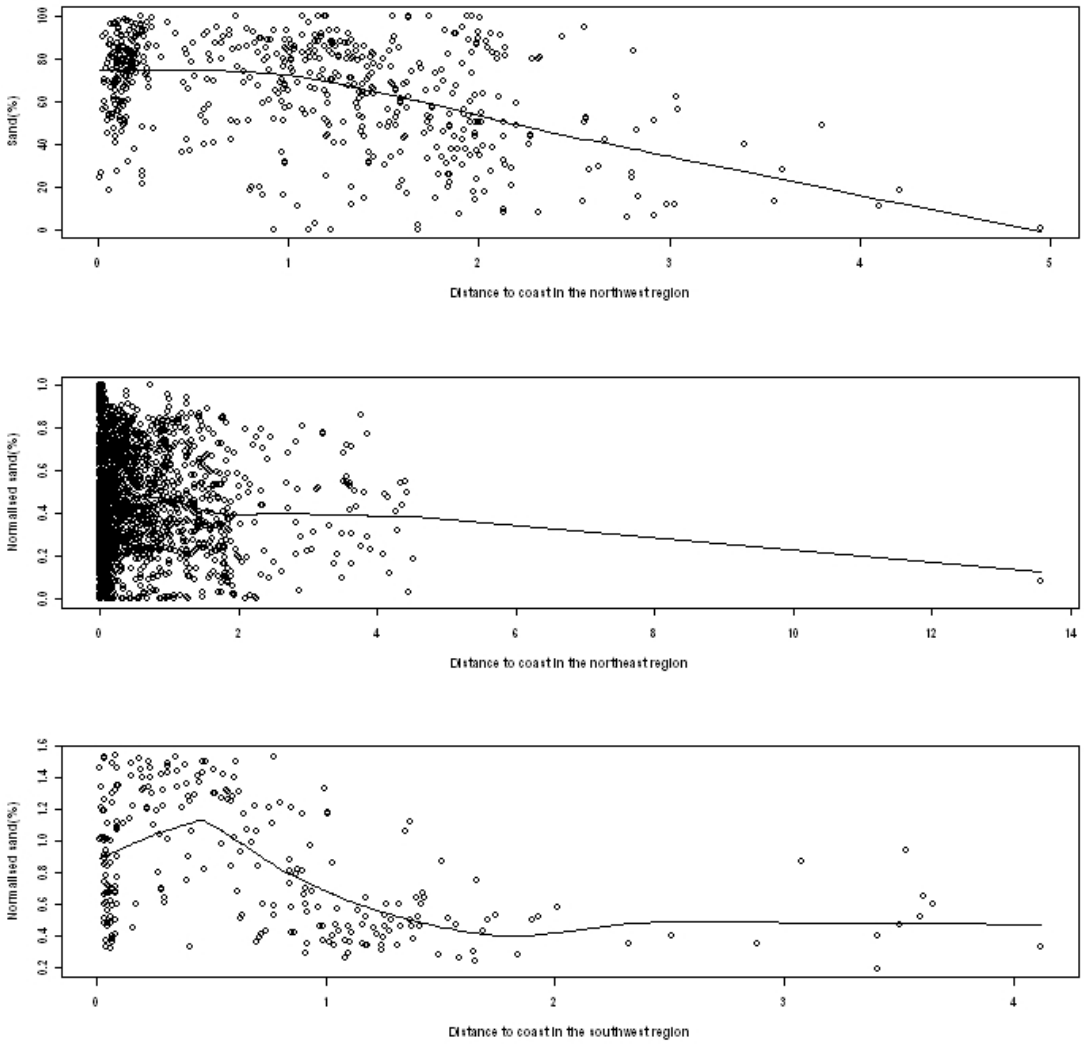


***Figure B.11**. Relation between normalised sand data and transformed distance to coast in the three study regions and the curve was fitted using lowess in R (R Development Core Team, 2010).*

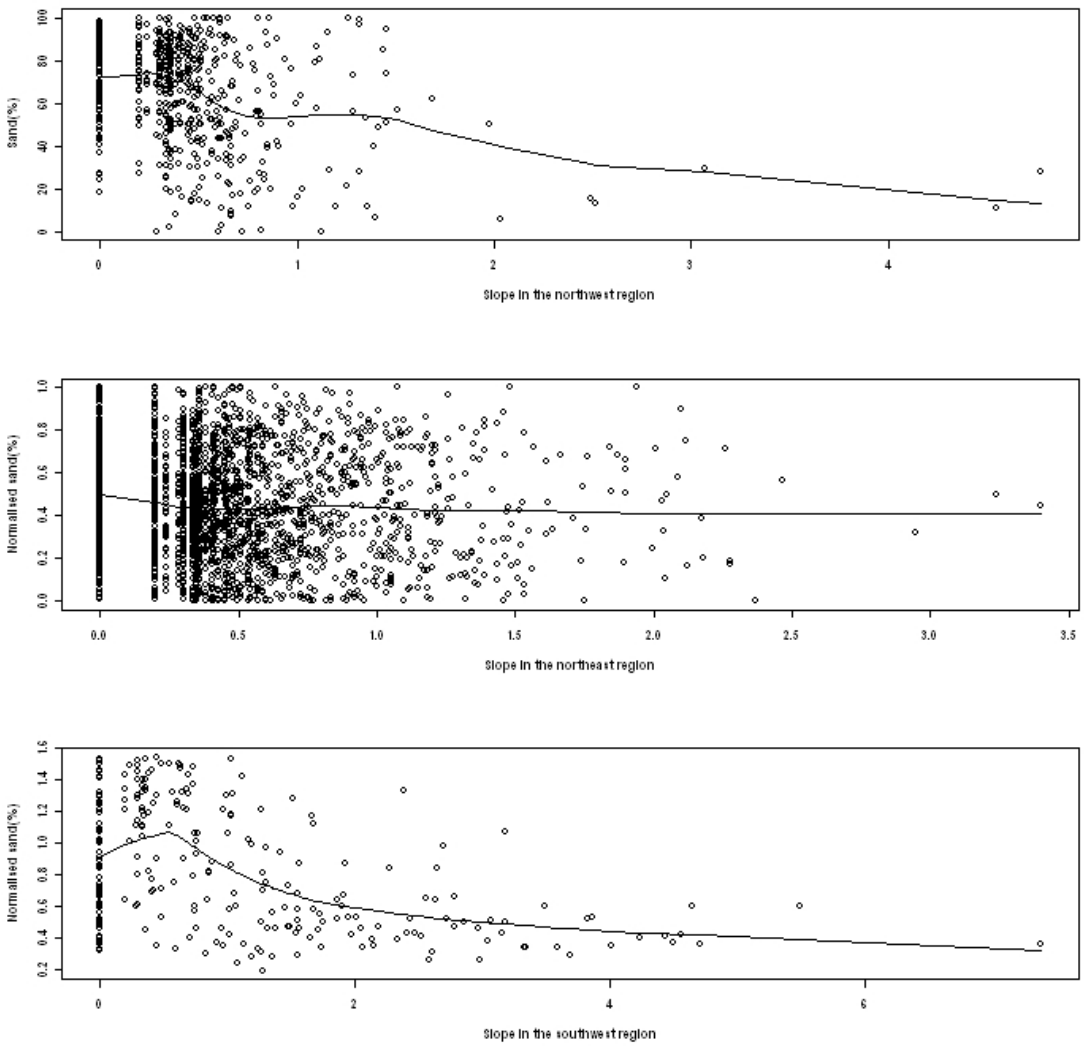


Figure B.12. *Relation between normalised sand data and transformed slope in the three study regions and the curve was fitted using lowess in R (R Development Core Team, 2010).*

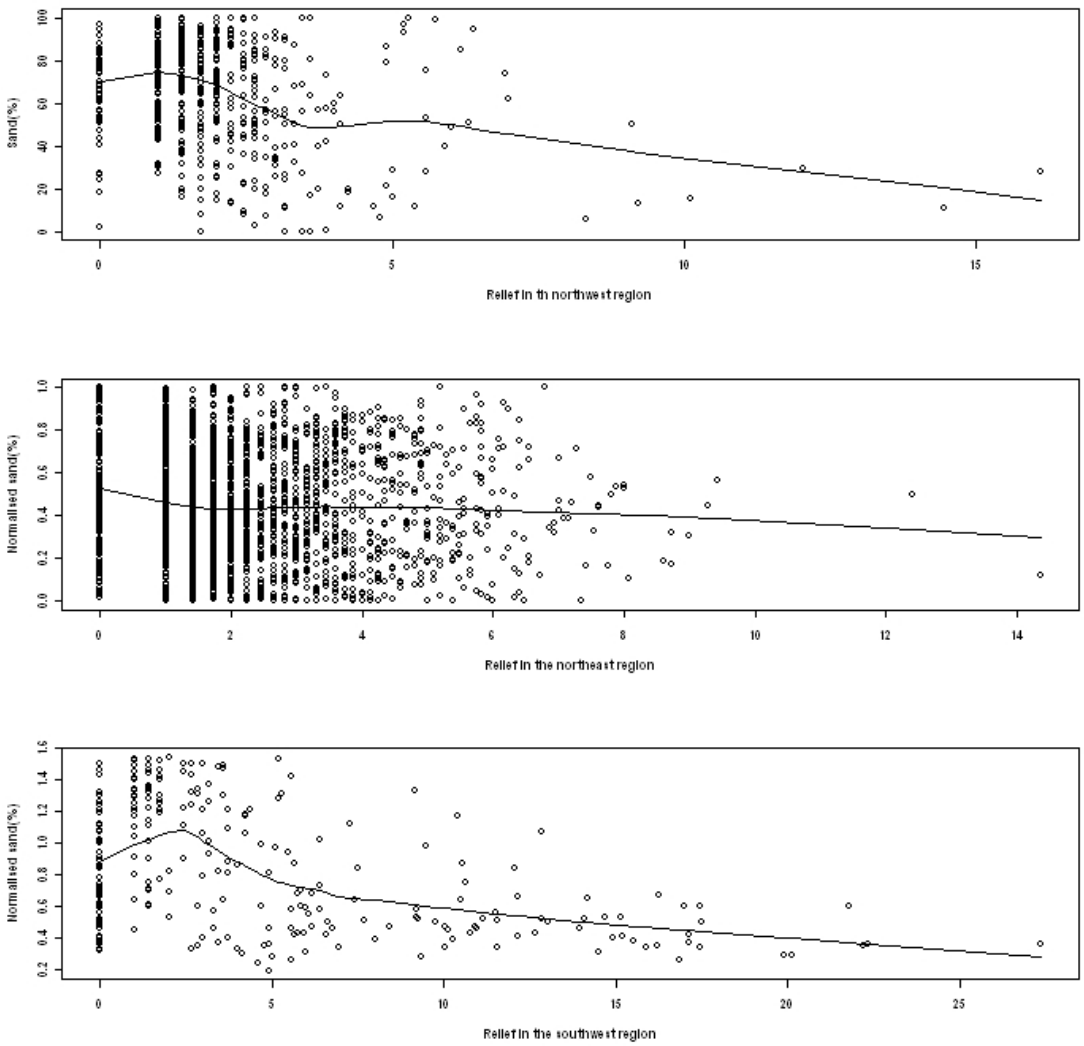


Figure B.13. *Relation between normalised sand data and transformed relief in the three study regions and the curve was fitted using lowess in R (R Development Core Team, 2010).*

B.3. VARIOGRAM MODELLING

Variogram modelling is an essential component for geostatistical methods. The result of variogram modelling depends on the selection of data projection, variogram models and the nature of the data. For the selection of data projection and variogram models please refer to Li et al. (2011b; 2011c; 2010). In both variogram and statistical/mathematical modelling, either bathymetry or all secondary variables were used depending on the statistical method used. The spatial structure of the data affects the performance of geostatistical interpolators.

B.3.1. Variogram anisotropy

To test for anisotropy, semivariogram maps were generated for each of the three regions using gstat in R (R Development Core Team, 2010) (Fig. B.14). There is a strong trend at 135^{o} in the northwest region, with an anisotropy ratio 1/3.5=0.3. For the normalised sand data in the northeast region the maps showed weak directional changes at 45^{o}, suggesting a weak anisotropy. There is no obvious trend for the normalised sand data in

the southwest region. Therefore, sand content was modelled as anisotropic in the northwest region and isotropic in the other two regions (Table B.6). The anisotropy was also examined for KED, RF, SVM, BDT and GRNN and their sub-methods (Table B.6). A directional change was detected for KED in the northwest region. There is no directional change for all other methods in all regions.

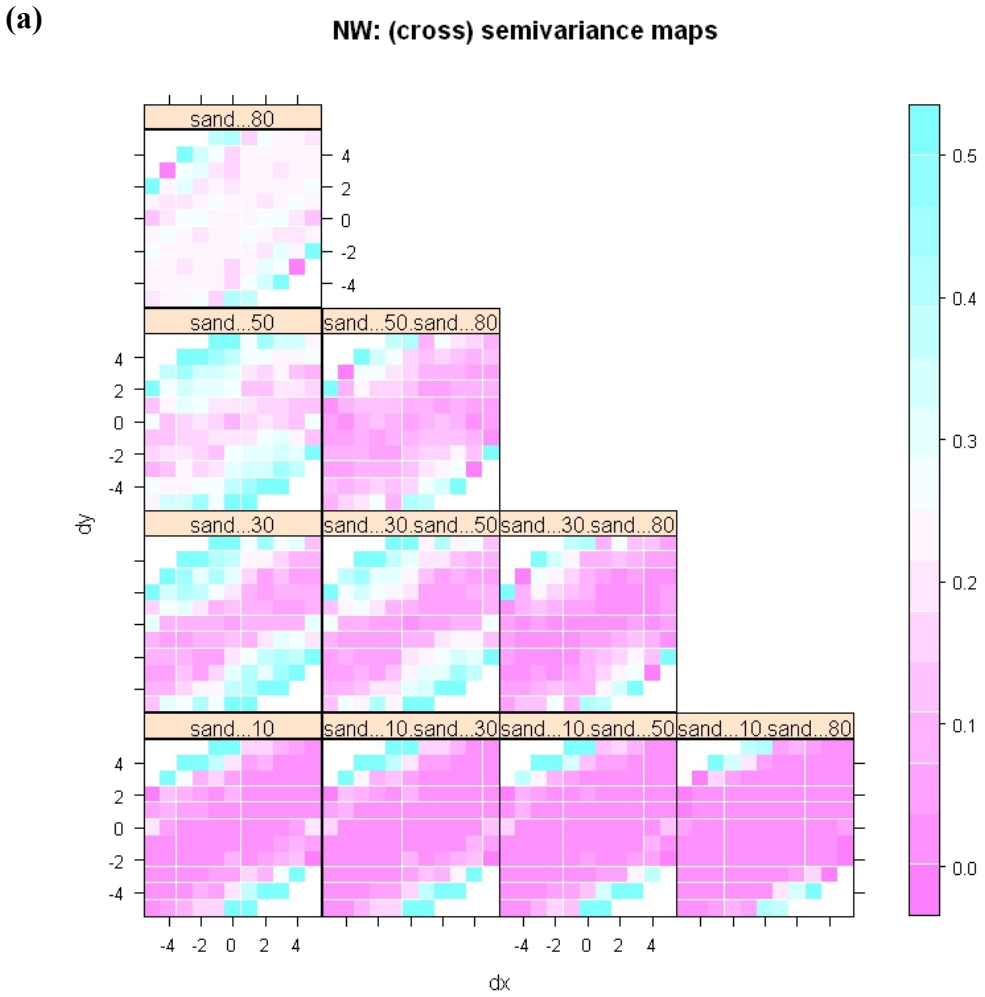


Figure B.14. *Variogram maps: a) northwest, b) northeast and c) southwest.*

Fig. B.14. (cont.):

(b)

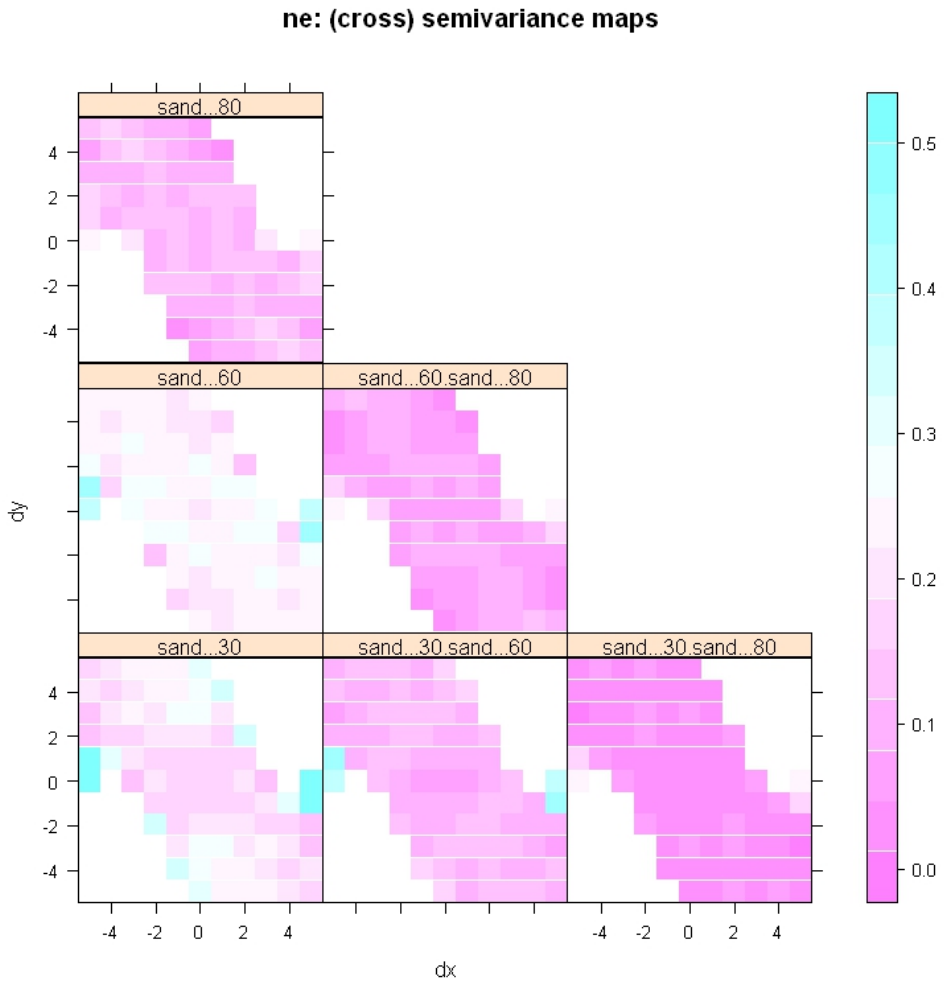
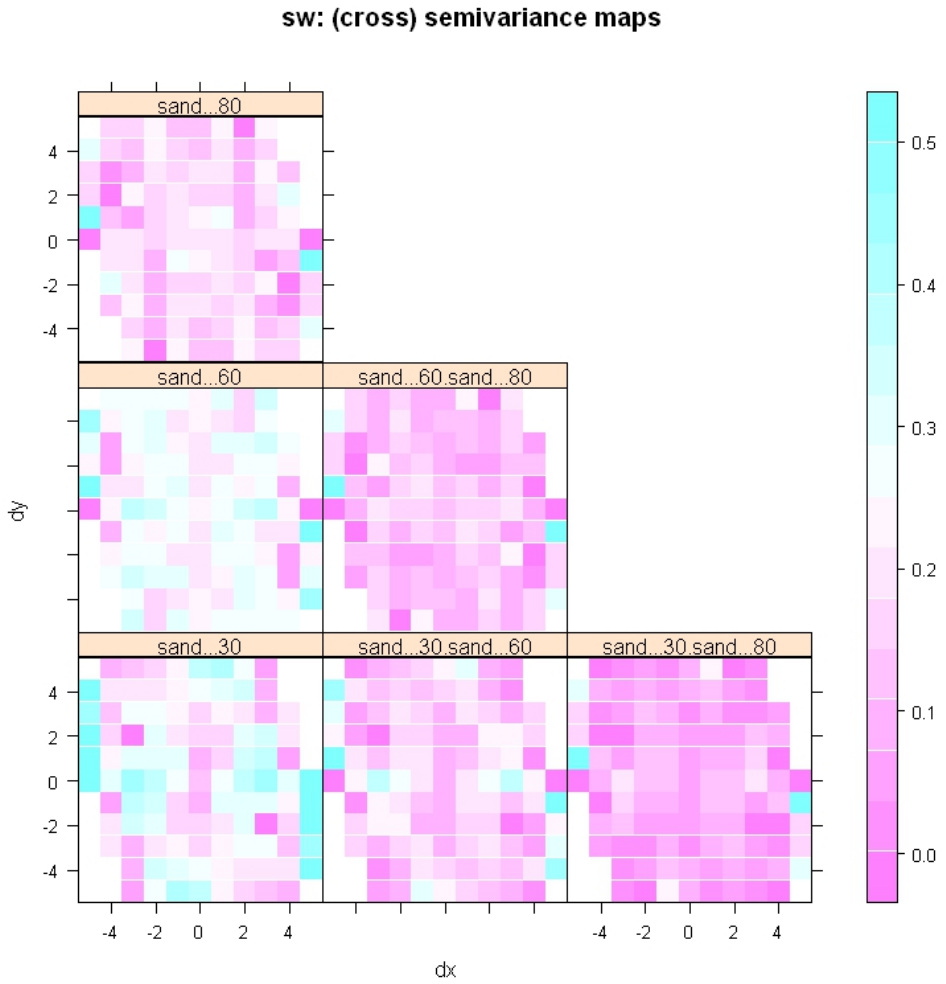


Fig. B.14. (cont.):

(c)



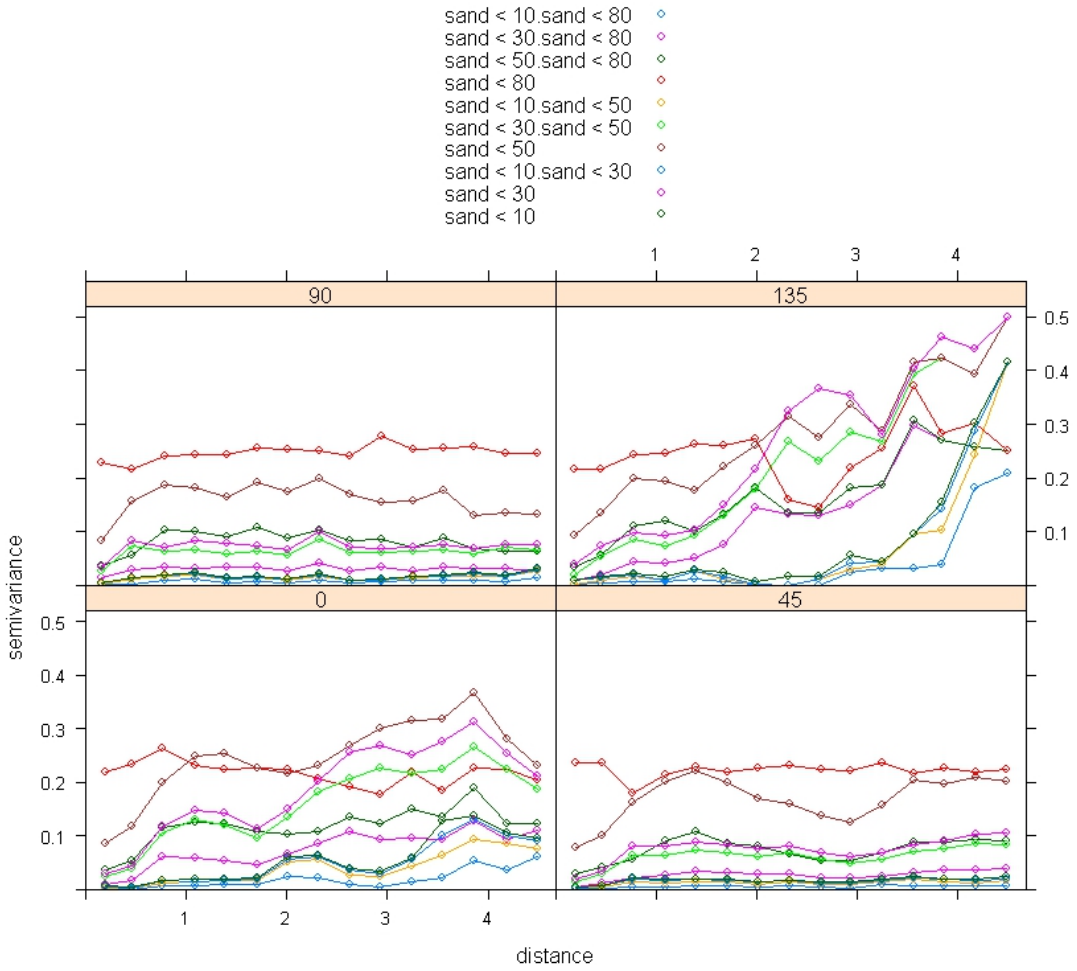


Figure B.15. *Semivariance of sand data at different directions in the northwest region.*

Table B.6. Anisotropic analyses of sand content (sand in the NW, (sand/100)2 in the NE and arcsin(sand) in the SW) under various conditions and residuals of various methods as derived in R (R Development Core Team, 2010).

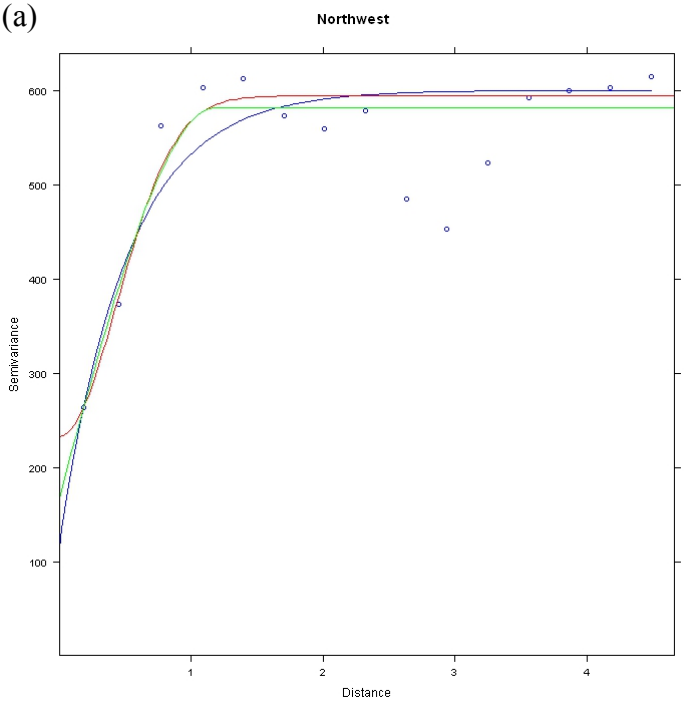
METHOD	ANISOTROPY		
	NW	NE	SW
OK	135, 0.3	no	no
KED	135, 0.6	no	no
RF, iRF, i4RF, 4mRF, SVM, LSVM, BDT, GRNN	no	no	no
6RF,	135, 0.6	no	no

B.3.2. Variogram model selection

There are a number of variogram models that can be employed, and different variogram models may lead to different interpolations (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 Exponential, Gaussian and Spherical using gstat in R (R Development Core Team, 2010). Of these models, Spherical model performed better than the others in terms of range, nugget and sill for sand data in the northwest region and the normalised sand data in the other two regions (Figure B.16). These models were also applied to KED and the

residuals of all other methods. The Spherical model was selected for all methods in all regions.

(a)



(b)

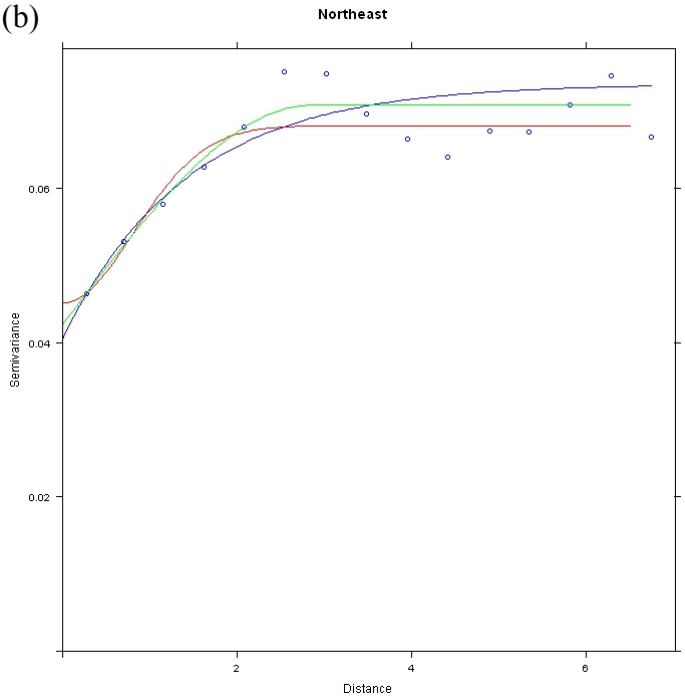


Figure B.16. *Variogram models of normalised sand content with no trend (Exponential: blue, Gaussian: red, and Spherical: green) in (a) northwest and (b) northeast regions.*

Fig. B.16. (cont.):
(c)

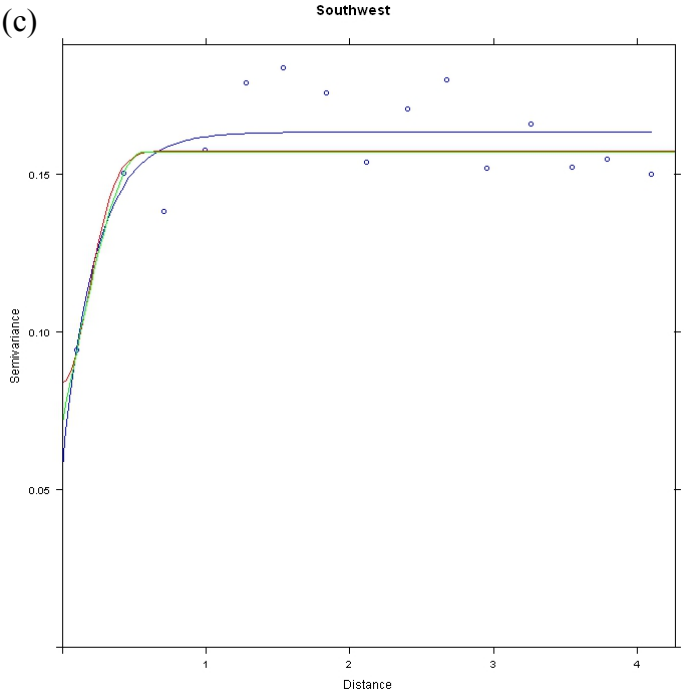


Figure B.17. *Variogram models of normalised sand content with no trend (Exponential: blue, Gaussian: red, and Spherical: green) in the SW region.*

B.4. STATISTICAL AND MATHEMATICAL MODELLING

B.4.1. Model specification

For each method, data transformation, variogram model, size of searching neighbourhood and anisotropy change with study regions (Table B.7). For KED and OK, relevant transformation was applied to sand content data to normalise the data.
A Spherical (Sph) variogram model was used for all geostatistical methods and their combinations with other methods.

Searching window size, that is the number of nearest samples used for making prediction, was set at two levels for most methods. The window size was 12 for all methods except KED. For KED, it does not produce any results when the window search size is less than 40, so Inf (all) was used instead. In order to search for the optimal size, 4 to 25 and all samples were used for RF, its combinations with OK and IDS, and their averages.

A singular model in variogram fit was observed for: 1) KED for some sub-datasets in the southwest region, 2) the residuals of RF, iRF, 6RF, SVM and LSVM for some sub-datasets in the southwest region, and 3) the residuals of BDT and GRNN for some datasets in all regions.

Anisotropy was detected and specified for KED and OK in the northwest region and for 6RF in the northwest region.

Table B.7. Data transformation, variogram model, searching window and anisotropy for each spatial interpolation method in each region.

METHOD	REGION	DATA TRANSFORMATION	VARIOGRAM MODEL	WINDOW SEARCH SIZE	ANISOTROPY
IDW2	all	no	na	12	na
KED	nw	no	Sph	Inf	135, 0.6
	ne	squared(sand/100)	Sph	Inf	no
	sw	arcsine	Sph	Inf	no
OK	nw	no	Sph	12	135, 0.3
	ne	squared(sand/100)	Sph	12	no
	sw	arcsine	Sph	12	no
RF	all	no	Sph	na	na
RFOK, RFIDS, RFOKRFIDS, RFRFOKRFIDS	nw	no	Sph	4:25, Inf	no
	ne	no	Sph	4:25, Inf	no
	sw	no	Sph	4:25, Inf	no
iRF	all	no	Sph	na	na
iRFOK, iRFIDS, iRFOKRFIDS, iRFRFOKRFIDS	nw	no	Sph	12	no
	ne	no	Sph	12	no
	sw	no	Sph	12	no
i4RF	all	no	Sph	na	na
i4RFOK, i4RFIDS,	nw	no	Sph	12	no
i4RFOKRFIDS,	ne	no	Sph	12	no
i4RFRFOKRFIDS	sw	no	Sph	12	no
6RF	all	no	Sph	na	na
6RFOK, 6RFIDS,	nw	no	Sph	na	135, 0.6
6RFOKRFIDS,	ne	no	Sph	na	no
6RFRFOKRFIDS	sw	no	Sph	na	no
4mRF	all	no	Sph	na	na
4mRFOK, 4mRFIDS,	nw	no	Sph	4:25, Inf	no
4mRFOKRFIDS,	ne	no	Sph	4:25, Inf	no
4mRFRFOKRFIDS	sw	no	Sph	4:25, Inf	no
SVM	all	no	Sph	na	na
SVMOK, SVMIDS, SVMOKSVMIDS, SVMSVMOKSVMIDS	all	no	Sph	12	no
LSVM	all	no	Sph	na	na
LSVMOK, LSVMIDS, LSVMOKSVMIDS, LSVMSVMOKSVMIDS	all	no	Sph	12	no
BDT	all	no	na	na	na
BDTIDS, BDTOK	all	no	Sph	12	no
GRNN	all	no	na	na	na
GRNNOK, GRNNIDS	all	no	Sph	12	no

B.4.2. Secondary variables and parameters specification

The secondary variables used for each method are summarised in Table B.8. Given that bathymetry (bathy) was the most strongly correlated variable with sand content in all three regions, it was used as a secondary variable in all methods that consider secondary information. Although distance to coast is weakly correlated with sand content in the northeast region, it was used in relevant models in all three regions. In RF, SVM and LSVM, we used all six secondary variables, their second and third power, and the latitude and longitude terms in the Legendre and Legendre (1998). The 4mRF is RF with 'the number of variables randomly sampled as candidates at each split' being 4 (i.e., *mtry*=4). In iRF, besides those terms in RF we further included all possible interactions among these six variables. In i4RF we excluded all second and third ordered terms and their interactions that were used in iRF. In 6RF, BDT and GRNN, only six variables were used. All these modelling work was implemented using *gstat*, *randomForest* and *e1071* in R (R Development Core Team, 2010) and the searching neighbourhood size is 12 if applicable, with an exception of BDT and GRNN as further discussed in section B.4.3. Predictions were corrected by resetting the faulty estimate to the nearest bound of the data range (*i.e.*, 0 or 100%) (Goovaerts, 1997).

The parameter specified for each method are summarised in Table B.9. A distance power of 2 was used in IDW (i.e., IDS, a standard method used in GA), which was used as the control in this study. The *mtry* changes with RF methods and with regions. For SVM and LSVM, the cost and gamma also vary with regions, which were selected based on 10-fold cross validation.

Table B.8. Variables used for each spatial interpolation methods*.

METHOD	TREND/SECONDARY VARIABLES
KED	bathy
RF, RFOK, RFIDS, RFOKRFIDS, RFRFOKRFIDS	bathy, dist.coast, slope, relief, lat, long,
4mRF, 4mRFOK, 4mRFIDS, 4mRFOKRFIDS, 4mRFRFOKRFIDS	bathy^2, bathy^3, dist.coast^2, dist.coast^3, slope^2, slope^3, relief^2, relief^3, lat^2, long^2,
SVM, SVMMOK, SVMIDS, SVMOKSVMIDS, SVMSVMOKSVMIDS	lat*long, lat*long^2, long*lat^2, lat^3, long^3
LSVM, LSVMMOK, LSVMIDS, LSVMOKSVMIDS, LSVMSVMOKSVMIDS	
iRF, iRFOK, iRFIDS, iRFOKRFIDS, iRFRFOKRFIDS	bathy, dist.coast, slope, relief, lat, long, bathy^2, bathy^3, dist.coast^2, dist.coast^3, slope^2, slope^3, relief^2, relief^3, lat^2, long^2, lat*long, lat*long^2, long*lat^2, lat^3, long^3, bathy*dist.coast, bathy*slope, bathy*relief, bathy*lat, bathy*long, dist.coast*slope, dist.coast*relief, dist.coast*lat, dist.coast*long, slope*relief, slope*lat, slope*long, relief*lat, relief*long
i4RF, i4RFOK, i4RFIDS, i4RFOKRFIDS, i4RFRFOKRFIDS	bathy, dist.coast, slope, relief, lat, long, bathy*dist.coast, bathy*slope, bathy*relief, bathy*lat, bathy*long, dist.coast*slope, dist.coast*relief, dist.coast*lat, dist.coast*long, slope*relief, slope*lat, slope*long, relief*lat, relief*long
6RF, 6RFOK, 6RFIDS, 6RFOKRFIDS, 6RFRFOKRFIDS	bathy, dist.coast, slope, relief, lat, long
BDT, BDTOK, BDTIDS,	
GRNN, GRNNOK, GRNNIDS	

*: All modelling work was completed in R except that BDT and GRNN were implemented in DTREG.

Table B.9. Parameters used for each spatial interpolation methods.

METHOD/PARAMETER	NW	NE	SW
IDS/ Distance power	2	2	2
RF/*mtry*	7	4	4
iRF/*mtry*	6	6	6
i4RF	4	7	7
6RF	2	2	2
4mRF	4	4	4
SVM/cost	1	1	4
SVM/gamma	0.25	4	0.125
LSVM/cost	0.015625	0.25	2

B.4.3. Model and parameter specification of BDT and GRNN

The parameter selection process was complex for BDT (Table B.10). It was different for the three regions. Four model parameters were varied; while other parameters were kept as default. The "maximum number of trees" parameter was set as 2000 for all experiments. The other three parameters, "Depth of individual trees", "Minimum size node to split" and "Number of cross-validation folds" were obtained from a tuning process. For the NE and NW regions, the tuning process was as follows:

1. Tune the "Number of cross-validation folds" parameter. The "Depth of individual trees" parameter was set as 4. The "Minimum size node to split" parameter was set as 5. The "Number of cross-validation folds" parameter was varies between 3 and 10. The number that gave the "highest" validation performance was selected for the "Number of cross-validation folds" parameter.
2. Tune the "Depth of individual trees" and "Minimum size node to split" parameters. The "Number of cross-validation folds" parameter selected from the step 1 was used. The "Depth of individual trees" parameter was varied between 4 and 6; the "Minimum size node to split" parameter was varied between 5 and 10. The combination that gave the "highest" validation performance was selected for the "Depth of individual trees" and "Minimum size node to split" parameters.
3. Fine-tune the "Depth of individual trees" and "Minimum size node to split" parameters. The "Number of cross-validation folds" parameter selected from the step 1 was used. The "Depth of individual trees" parameter was set as 10; the "Minimum size node to split" parameter was set as 5. The validation result was compared to that of the step 2. If the result is better, the "Depth of individual trees" and "Minimum size node to split" parameters were set as 10 and 5 respectively. Otherwise, the parameters resulting from the step 2 were selected.
4. The selected model was then used to predict the prediction datasets.

For the SW region, due to its much smaller sample size we tested different values for the "Minimum size node to split" parameter. First, the "Minimum size node to split" parameter was varied between 3 and 8 in the step 2. Second, the "Minimum size node to split" parameter was set as 3 in the step 3.

Six explanatory variables were used for all experiments: latitude, longitude, bathymetry, slope, distance to coast and relief. BDT was implemented in DTREG.

Table B.10. Modelling parameters used for BDT in each region.

REGION	CV	MAXIMUM NUMBER OF TREES	DEPTH OF INDIVIDUAL TREES	MINIMUM SIZE NODE TO SPLIT	NUMBER OF CROSS-VALIDATION FOLDS
NE	1	2000	10	5	8
	2	2000	10	5	8
	3	2000	10	5	8
	4	2000	10	5	7
	5	2000	10	5	10
	6	2000	10	5	9
	7	2000	10	5	10
	8	2000	10	5	6
	9	2000	10	5	10
	10	2000	10	5	9
NW	1	2000	6	9	9
	2	2000	10	5	8
	3	2000	6	5	10
	4	2000	6	8	10
	5	2000	6	5	8
	6	2000	5	5	6
	7	2000	5	8	4
	8	2000	6	8	7
	9	2000	6	9	5
	10	2000	10	5	8
SW	1	2000	4	4	10
	2	2000	5	8	8
	3	2000	5	7	8
	4	2000	6	7	7
	5	2000	5	7	10
	6	2000	6	8	10
	7	2000	5	3	4
	8	2000	4	4	9
	9	2000	4	3	9
	10	2000	4	8	8

GRNN was also implemented in DTREG. For all experiments, we used “leave-one-out” validation for model development. The “Minimum Sigma” parameter was set as 0.0001; the “Maximum Sigma” parameter was set as 10; the “Search steps” parameter was set as 20. We also chose to remove unnecessary neurons to simplify and optimise model, which is a very resource demanding process. Due to resource limitations, for the NE region, the number of neurons to remove was set as 22 (out of 1942 neurons in the hidden layer); for NW region, the number of neurons to be removed was set as 37 (out of 517 neurons in the hidden layer). The simplification processes for the two regions were therefore not optimal, which would impact on the models’ prediction performance. For SW region, instead of pre-setting the number of neurons to remove, all of the unnecessary neurons were removed to minimise validation error. The models developed from the development datasets were used to predict the prediction datasets. Again, six explanatory variables were used for all experiments: latitude, longitude, bathymetry, slope, distance to coast and relief.

Appendix C. Basic statistical summaries of the predictions of and statistics measuring the performance of each modelling method.

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
1	6rf	ne	na	12.97	63.69	63.51	98.56	13.75	11.98	16.15	18.87	25.44
2	6rfids	ne	12	1.5	63.76	63.52	99.99	15.94	11.53	16.08	18.16	25.33
3	6rfok	ne	12	7.49	63.95	63.59	100	14.59	11.76	16.1	18.52	25.36
4	6rfokrfids	ne	12	7	63.72	63.56	100	15.16	11.57	15.98	18.22	25.17
5	6rfrfokrfids	ne	12	8.99	63.73	63.54	99.52	14.65	11.67	15.98	18.38	25.17
6	Bdt	ne	na	6.72	64.08	63.53	101.94	14.86	12.17	16.63	19.17	26.19
7	bdtids	ne	12	1.5	63.69	63.43	100	16.63	11.85	16.61	18.66	26.16
8	bdtok	ne	12	5.64	63.92	63.51	100	15.46	12.06	16.63	19	26.19
9	Grnn	ne	na	0	65.15	63.65	97.84	16.25	12.65	17.47	19.92	27.52
10	grnnids	ne	12	0	65.14	63.53	99.98	18.45	12.14	17.36	19.12	27.34
11	grnnok	ne	12	0	65.31	63.62	100	17.07	12.45	17.39	19.61	27.39
12	i4rf	ne	na	12.35	63.41	63.46	99.58	13.76	12.28	16.53	19.34	26.04
13	i4rfids	ne	12	1.5	63.68	63.5	99.99	15.72	11.75	16.31	18.51	25.69
14	i4rfok	ne	12	8.84	63.65	63.57	100	14.57	12.02	16.4	18.93	25.83
15	i4rfokrfids	ne	12	6.45	63.6	63.53	100	15.06	11.82	16.26	18.62	25.61
16	i4rfrfokrfids	ne	12	8.42	63.66	63.51	99.84	14.58	11.95	16.3	18.82	25.67
17	Ids	ne	12	1.5	64.39	63.66	99.99	17.13	11.66	16.62	18.37	26.18
18	Irf	ne	na	12.39	63.61	63.44	99.47	14.27	11.96	16.26	18.84	25.61
19	Irfids	ne	12	1.5	63.6	63.47	99.99	16.13	11.58	16.2	18.24	25.52
20	Irfok	ne	12	9.04	63.7	63.52	99.87	14.93	11.78	16.21	18.55	25.53
21	irfokrfids	ne	12	6.95	63.57	63.49	99.93	15.45	11.62	16.11	18.3	25.37
22	irfrfokrfids	ne	12	8.76	63.69	63.47	99.67	15.02	11.7	16.12	18.43	25.39
23	Ked	ne	Inf	0	66.25	66.15	91.83	10.95	13.25	17.55	20.87	27.64
24	Lsvm	ne	na	3.5	64.64	65.51	100	6.45	16.86	21.16	26.56	33.33

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
25	lsvmids	ne	12	1.5	64.37	63.59	100	17.34	11.72	16.72	18.46	26.33
26	lsvmok	ne	12	7.11	64.37	63.89	100	14.17	12.58	17.14	19.81	27
27	lsvmoksvmids	ne	12	8.07	64.39	63.74	100	15.27	11.85	16.41	18.66	25.85
28	lsvmsvmoksvmids	ne	12	6.55	64.64	64.33	100	11.19	12.73	16.85	20.05	26.54
29	ok	ne	12	20.85	66.19	66.18	98.29	12.84	12.69	17.42	19.99	27.44
30	rf	ne	na	10.51	63.58	63.42	99.46	14.79	11.67	16.02	18.38	25.23
31	rfids	ne	4	1.5	64.06	63.41	100	16.92	11.46	16.21	18.05	25.53
32	rfids	ne	5	1.5	64.06	63.42	99.99	16.84	11.46	16.2	18.05	25.52
33	rfids	ne	6	1.5	63.83	63.42	99.99	16.76	11.45	16.17	18.03	25.47
34	rfids	ne	7	1.5	63.87	63.43	99.99	16.71	11.45	16.16	18.03	25.45
35	rfids	ne	8	1.5	63.89	63.44	99.99	16.67	11.46	16.16	18.05	25.45
36	rfids	ne	9	1.5	63.84	63.45	99.99	16.64	11.45	16.15	18.03	25.44
37	rfids	ne	10	1.5	63.75	63.46	99.99	16.61	11.45	16.15	18.03	25.44
38	rfids	ne	11	1.5	63.69	63.45	99.99	16.6	11.45	16.15	18.03	25.44
39	rfids	ne	12	1.5	63.79	63.46	99.99	16.58	11.44	16.14	18.02	25.42
40	rfids	ne	13	1.5	63.84	63.46	99.99	16.56	11.44	16.14	18.02	25.42
41	rfids	ne	14	1.5	63.88	63.46	99.99	16.55	11.43	16.13	18	25.41
42	rfids	ne	15	1.5	63.9	63.47	99.99	16.54	11.43	16.13	18	25.41
43	rfids	ne	16	1.5	63.91	63.47	99.99	16.53	11.43	16.12	18	25.39
44	rfids	ne	17	1.5	63.89	63.47	99.99	16.52	11.43	16.12	18	25.39
45	rfids	ne	18	1.5	63.82	63.47	99.99	16.51	11.44	16.12	18.02	25.39
46	rfids	ne	19	1.5	63.86	63.47	99.99	16.5	11.44	16.12	18.02	25.39
47	rfids	ne	20	1.5	63.94	63.46	99.99	16.5	11.44	16.12	18.02	25.39
48	rfids	ne	21	1.5	63.92	63.46	99.99	16.49	11.44	16.12	18.02	25.39
49	rfids	ne	22	1.5	63.92	63.46	99.99	16.48	11.44	16.12	18.02	25.39
50	rfids	ne	23	1.5	63.95	63.47	99.99	16.48	11.44	16.12	18.02	25.39
51	rfids	ne	24	1.5	63.95	63.46	99.99	16.47	11.44	16.12	18.02	25.39
52	rfids	ne	25	1.5	63.92	63.46	99.99	16.46	11.44	16.12	18.02	25.39
53	rfids	ne	Inf	1.5	64.02	63.49	99.99	16.22	11.45	16.08	18.03	25.33

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
54	rfok	ne	4	6.41	63.8	63.42	100	16.08	11.52	16.09	18.14	25.34
55	rfok	ne	5	4.96	63.77	63.4	100	15.91	11.55	16.08	18.19	25.33
56	rfok	ne	6	5.31	63.77	63.4	100	15.79	11.55	16.07	18.19	25.31
57	rfok	ne	7	4.76	63.66	63.44	100	15.66	11.56	16.04	18.21	25.26
58	rfok	ne	8	8.52	63.9	63.44	100	15.57	11.57	16.05	18.22	25.28
59	rfok	ne	9	7.89	63.75	63.49	100	15.49	11.56	16.03	18.21	25.25
60	rfok	ne	10	8.27	63.74	63.52	100	15.43	11.56	16.06	18.21	25.3
61	rfok	ne	11	8.99	63.69	63.51	99.78	15.38	11.57	16.06	18.22	25.3
62	rfok	ne	12	9.18	63.71	63.5	99.85	15.33	11.59	16.05	18.25	25.28
63	rfok	ne	13	9.58	63.71	63.51	99.92	15.3	11.58	16.03	18.24	25.25
64	rfok	ne	14	9.66	63.73	63.51	100	15.26	11.57	16.02	18.22	25.23
65	rfok	ne	15	9.94	63.74	63.51	100	15.23	11.59	16.02	18.25	25.23
66	rfok	ne	16	9.47	63.66	63.51	100	15.21	11.59	16.02	18.25	25.23
67	rfok	ne	17	9.69	63.74	63.51	100	15.17	11.6	16.02	18.27	25.23
68	rfok	ne	18	9.81	63.92	63.52	100	15.16	11.61	16.02	18.29	25.23
69	rfok	ne	19	10.08	63.7	63.52	100	15.16	11.61	16.02	18.29	25.23
70	rfok	ne	20	10.29	63.82	63.52	100	15.16	11.61	16.03	18.29	25.25
71	rfok	ne	21	10	63.74	63.52	100	15.14	11.62	16.04	18.3	25.26
72	rfok	ne	22	9.73	63.75	63.52	100	15.13	11.62	16.03	18.3	25.25
73	rfok	ne	23	9.99	63.79	63.52	100	15.12	11.61	16.02	18.29	25.23
74	rfok	ne	24	10.11	63.77	63.52	100	15.11	11.61	16.02	18.29	25.23
75	rfok	ne	25	10.38	63.65	63.52	100	15.11	11.61	16.02	18.29	25.23
76	rfok	ne	Inf	9.78	63.62	63.45	99.8	14.98	11.63	16	18.32	25.2
77	rfokrfids	ne	4	5.3	63.93	63.41	100	16.43	11.44	16.07	18.02	25.31
78	rfokrfids	ne	5	4.58	63.87	63.41	99.97	16.29	11.44	16.05	18.02	25.28
79	rfokrfids	ne	6	4.75	63.81	63.41	99.99	16.2	11.44	16.03	18.02	25.25
80	rfokrfids	ne	7	4.48	63.71	63.44	99.99	16.1	11.44	16.01	18.02	25.22
81	rfokrfids	ne	8	6.36	63.82	63.44	99.99	16.04	11.45	16.01	18.03	25.22
82	rfokrfids	ne	9	6.04	63.68	63.47	99.99	15.98	11.44	16	18.02	25.2

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
83	rfokrfids	ne	10	6.24	63.67	63.49	99.98	15.94	11.44	16.01	18.02	25.22
84	rfokrfids	ne	11	6.6	63.72	63.48	99.8	15.91	11.44	16.01	18.02	25.22
85	rfokrfids	ne	12	6.69	63.82	63.48	99.92	15.87	11.45	16	18.03	25.2
86	rfokrfids	ne	13	6.89	63.74	63.48	99.95	15.85	11.45	15.99	18.03	25.19
87	rfokrfids	ne	14	6.93	63.71	63.49	99.99	15.82	11.44	15.98	18.02	25.17
88	rfokrfids	ne	15	7.07	63.75	63.49	99.99	15.8	11.45	15.98	18.03	25.17
89	rfokrfids	ne	16	6.84	63.8	63.49	99.99	15.79	11.45	15.98	18.03	25.17
90	rfokrfids	ne	17	6.94	63.72	63.49	99.99	15.76	11.45	15.98	18.03	25.17
91	rfokrfids	ne	18	7.01	63.76	63.49	99.99	15.75	11.46	15.98	18.05	25.17
92	rfokrfids	ne	19	7.14	63.71	63.49	99.99	15.75	11.46	15.98	18.05	25.17
93	rfokrfids	ne	20	7.24	63.74	63.49	99.98	15.75	11.47	15.98	18.07	25.17
94	rfokrfids	ne	21	7.1	63.75	63.49	99.96	15.73	11.47	15.99	18.07	25.19
95	rfokrfids	ne	22	6.97	63.73	63.49	99.97	15.72	11.47	15.98	18.07	25.17
96	rfokrfids	ne	23	7.09	63.77	63.49	99.97	15.72	11.47	15.98	18.07	25.17
97	rfokrfids	ne	24	7.16	63.76	63.49	99.97	15.71	11.46	15.98	18.05	25.17
98	rfokrfids	ne	25	7.29	63.76	63.49	99.99	15.71	11.46	15.97	18.05	25.15
99	rfokrfids	ne	Inf	7	63.76	63.47	99.87	15.54	11.5	15.97	18.11	25.15
100	rfrfokrfids	ne	4	7.04	63.74	63.42	99.8	15.82	11.45	15.97	18.03	25.15
101	rfrfokrfids	ne	5	6.55	63.61	63.41	99.8	15.74	11.47	15.97	18.07	25.15
102	rfrfokrfids	ne	6	6.67	63.7	63.42	99.8	15.68	11.47	15.97	18.07	25.15
103	rfrfokrfids	ne	7	6.49	63.68	63.43	99.8	15.62	11.48	15.96	18.08	25.14
104	rfrfokrfids	ne	8	7.74	63.72	63.44	99.8	15.58	11.49	15.96	18.1	25.14
105	rfrfokrfids	ne	9	7.53	63.64	63.45	99.76	15.55	11.49	15.96	18.1	25.14
106	rfrfokrfids	ne	10	7.66	63.65	63.47	99.74	15.52	11.49	15.97	18.1	25.15
107	rfrfokrfids	ne	11	7.9	63.65	63.46	99.68	15.5	11.49	15.97	18.1	25.15
108	rfrfokrfids	ne	12	7.96	63.74	63.46	99.71	15.48	11.5	15.97	18.11	25.15
109	rfrfokrfids	ne	13	8.09	63.73	63.46	99.74	15.47	11.5	15.96	18.11	25.14
110	rfrfokrfids	ne	14	8.12	63.8	63.47	99.76	15.45	11.5	15.96	18.11	25.14
111	rfrfokrfids	ne	15	8.22	63.86	63.47	99.77	15.44	11.5	15.96	18.11	25.14

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
112	rfrfokrfids	ne	16	8.06	63.82	63.47	99.8	15.43	11.51	15.96	18.13	25.14
113	rfrfokrfids	ne	17	8.13	63.74	63.47	99.78	15.41	11.51	15.96	18.13	25.14
114	rfrfokrfids	ne	18	8.17	63.88	63.47	99.76	15.41	11.51	15.96	18.13	25.14
115	rfrfokrfids	ne	19	8.26	63.91	63.47	99.76	15.41	11.52	15.96	18.14	25.14
116	rfrfokrfids	ne	20	8.33	63.82	63.47	99.75	15.41	11.52	15.96	18.14	25.14
117	rfrfokrfids	ne	21	8.24	63.82	63.47	99.77	15.4	11.52	15.97	18.14	25.15
118	rfrfokrfids	ne	22	8.15	63.86	63.47	99.8	15.39	11.52	15.97	18.14	25.15
119	rfrfokrfids	ne	23	8.23	63.82	63.47	99.8	15.39	11.52	15.96	18.14	25.14
120	rfrfokrfids	ne	24	8.27	63.79	63.47	99.8	15.38	11.52	15.96	18.14	25.14
121	rfrfokrfids	ne	25	8.36	63.78	63.47	99.8	15.38	11.52	15.96	18.14	25.14
122	rfrfokrfids	ne	Inf	8.17	63.76	63.46	99.73	15.27	11.55	15.96	18.19	25.14
123	svm	ne	na	6.66	64.23	64.83	100	13.59	12.99	17.57	20.46	27.67
124	svmids	ne	12	1.5	63.81	63.65	99.99	16.59	12.24	16.94	19.28	26.68
125	svmok	ne	12	8.27	63.79	63.91	100	14.57	12.66	17.15	19.94	27.01
126	svmoksvmids	ne	12	9.36	63.69	63.78	99.99	15.27	12.26	16.73	19.31	26.35
127	svmsvmoksvmids	ne	12	10.95	63.93	64.13	100	14.5	12.39	16.81	19.51	26.48
128	4mrf	nw	na	13.01	67.78	65.03	91.66	15.01	14.81	19.71	22.67	30.17
129	4mrfids	nw	4	7.87	68.85	65.06	94.04	17.51	14.45	19.96	22.12	30.55
130	4mrfids	nw	5	7.87	68.56	65	94.11	17.38	14.42	19.9	22.07	30.46
131	4mrfids	nw	6	7.87	68.27	65	94.35	17.24	14.47	19.91	22.15	30.48
132	4mrfids	nw	7	7.87	68.22	64.97	94.18	17.18	14.52	19.95	22.23	30.54
133	4mrfids	nw	8	7.87	68.33	65.01	94.28	17.11	14.53	19.95	22.24	30.54
134	4mrfids	nw	9	7.87	68.29	65.02	94.37	17.06	14.55	19.97	22.27	30.57
135	4mrfids	nw	10	7.87	68.29	65.01	94.32	16.98	14.52	19.93	22.23	30.51
136	4mrfids	nw	11	7.87	68.29	65.01	94.27	16.97	14.51	19.91	22.21	30.48
137	4mrfids	nw	12	7.87	68.15	64.99	94.29	16.93	14.53	19.91	22.24	30.48
138	4mrfids	nw	13	7.87	67.98	64.97	94.28	16.92	14.53	19.9	22.24	30.46
139	4mrfids	nw	14	7.87	67.96	64.98	94.1	16.89	14.53	19.91	22.24	30.48
140	4mrfids	nw	15	7.87	67.92	64.98	94.05	16.88	14.53	19.91	22.24	30.48

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
141	4mrfids	nw	16	7.87	67.94	64.98	93.96	16.86	14.52	19.9	22.23	30.46
142	4mrfids	nw	17	7.87	67.99	64.98	93.93	16.85	14.52	19.89	22.23	30.45
143	4mrfids	nw	18	7.87	67.97	64.97	93.96	16.84	14.52	19.88	22.23	30.43
144	4mrfids	nw	19	7.87	67.99	64.97	93.99	16.82	14.51	19.87	22.21	30.41
145	4mrfids	nw	20	7.87	67.95	64.97	93.98	16.8	14.51	19.87	22.21	30.41
146	4mrfids	nw	21	7.87	67.93	64.97	93.94	16.79	14.52	19.86	22.23	30.4
147	4mrfids	nw	22	7.87	67.78	64.96	93.92	16.79	14.52	19.86	22.23	30.4
148	4mrfids	nw	23	7.87	67.84	64.97	93.91	16.78	14.52	19.87	22.23	30.41
149	4mrfids	nw	24	7.87	67.77	64.97	93.93	16.78	14.52	19.86	22.23	30.4
150	4mrfids	nw	25	7.87	67.76	64.97	93.96	16.76	14.51	19.86	22.21	30.4
151	4mrfids	nw	Inf	7.87	68.11	64.96	93.77	16.49	14.57	19.86	22.3	30.4
152	4mrfok	nw	4	10.29	68.63	65.23	93.21	16.59	14.56	19.78	22.29	30.28
153	4mrfok	nw	5	10.81	68.21	65.14	93.62	16.39	14.53	19.7	22.24	30.15
154	4mrfok	nw	6	9.86	68.17	65.12	94.7	16.23	14.63	19.73	22.39	30.2
155	4mrfok	nw	7	10.53	67.99	65.1	93.66	16.19	14.75	19.84	22.58	30.37
156	4mrfok	nw	8	10.48	68.4	65.16	94.24	16.1	14.77	19.86	22.61	30.4
157	4mrfok	nw	9	11.19	68.25	65.18	94.73	16.05	14.78	19.85	22.62	30.38
158	4mrfok	nw	10	11.66	68.33	65.19	94.25	15.93	14.67	19.76	22.46	30.25
159	4mrfok	nw	11	11.21	68.28	65.17	93.88	15.96	14.68	19.76	22.47	30.25
160	4mrfok	nw	12	11.2	68.18	65.16	94.04	15.92	14.69	19.75	22.49	30.23
161	4mrfok	nw	13	10.98	68.29	65.16	94.04	15.92	14.69	19.77	22.49	30.26
162	4mrfok	nw	14	10.58	68.26	65.18	92.88	15.92	14.69	19.76	22.49	30.25
163	4mrfok	nw	15	10.52	68.2	65.16	92.7	15.92	14.65	19.73	22.42	30.2
164	4mrfok	nw	16	10.83	68.15	65.16	92.16	15.91	14.65	19.74	22.42	30.22
165	4mrfok	nw	17	10.24	68.26	65.17	92.11	15.92	14.67	19.74	22.46	30.22
166	4mrfok	nw	18	9.8	68.21	65.16	92.36	15.93	14.66	19.74	22.44	30.22
167	4mrfok	nw	19	9.82	68.2	65.17	92.66	15.92	14.65	19.74	22.42	30.22
168	4mrfok	nw	20	9.28	68.21	65.17	92.68	15.92	14.66	19.74	22.44	30.22
169	4mrfok	nw	21	9.8	68.24	65.18	92.54	15.91	14.66	19.74	22.44	30.22

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
170	4mrfok	nw	22	9.9	68.17	65.16	92.52	15.94	14.66	19.73	22.44	30.2
171	4mrfok	nw	23	10.87	68.15	65.17	92.53	15.91	14.65	19.73	22.42	30.2
172	4mrfok	nw	24	10.88	68.19	65.17	92.67	15.92	14.64	19.72	22.41	30.19
173	4mrfok	nw	25	10.83	68.16	65.18	92.87	15.9	14.64	19.72	22.41	30.19
174	4mrfok	nw	Inf	11.34	68.18	65.22	93.2	15.88	14.63	19.7	22.39	30.15
175	4mrfokrfids	nw	4	9.85	69.06	65.15	93.62	16.99	14.46	19.81	22.13	30.32
176	4mrfokrfids	nw	5	10.17	68.41	65.07	93.86	16.82	14.43	19.73	22.09	30.2
177	4mrfokrfids	nw	6	9.63	68.02	65.06	94.52	16.67	14.5	19.76	22.2	30.25
178	4mrfokrfids	nw	7	10.62	68.22	65.04	93.92	16.62	14.59	19.83	22.33	30.35
179	4mrfokrfids	nw	8	10.56	68.26	65.08	94.26	16.54	14.61	19.84	22.36	30.37
180	4mrfokrfids	nw	9	10.76	68.26	65.1	94.55	16.49	14.61	19.85	22.36	30.38
181	4mrfokrfids	nw	10	11.42	68.25	65.1	94.28	16.4	14.55	19.79	22.27	30.29
182	4mrfokrfids	nw	11	10.82	68.39	65.09	94.08	16.4	14.54	19.78	22.26	30.28
183	4mrfokrfids	nw	12	11.18	68.31	65.07	94.16	16.36	14.56	19.78	22.29	30.28
184	4mrfokrfids	nw	13	10.61	68.16	65.07	94.16	16.36	14.57	19.78	22.3	30.28
185	4mrfokrfids	nw	14	10.45	68.29	65.08	93.49	16.35	14.56	19.78	22.29	30.28
186	4mrfokrfids	nw	15	10.39	68.07	65.07	93.37	16.34	14.54	19.76	22.26	30.25
187	4mrfokrfids	nw	16	10.95	68.14	65.07	93.06	16.32	14.54	19.76	22.26	30.25
188	4mrfokrfids	nw	17	10.61	68.17	65.08	93.02	16.32	14.55	19.76	22.27	30.25
189	4mrfokrfids	nw	18	10.34	68.13	65.06	93.16	16.33	14.54	19.76	22.26	30.25
190	4mrfokrfids	nw	19	10.35	68.14	65.07	93.32	16.31	14.53	19.75	22.24	30.23
191	4mrfokrfids	nw	20	10.03	68.06	65.07	93.33	16.3	14.54	19.75	22.26	30.23
192	4mrfokrfids	nw	21	10.32	68.15	65.07	93.24	16.29	14.54	19.75	22.26	30.23
193	4mrfokrfids	nw	22	10.38	68.07	65.06	93.22	16.3	14.54	19.74	22.26	30.22
194	4mrfokrfids	nw	23	10.92	68.11	65.07	93.22	16.29	14.54	19.74	22.26	30.22
195	4mrfokrfids	nw	24	10.92	68.14	65.07	93.3	16.29	14.53	19.74	22.24	30.22
196	4mrfokrfids	nw	25	10.91	68.1	65.07	93.42	16.27	14.53	19.74	22.24	30.22
197	4mrfokrfids	nw	Inf	11.21	68.21	65.09	93.48	16.13	14.56	19.74	22.29	30.22
198	4mrfrfokrfids	nw	4	11.08	68.69	65.11	92.97	16.22	14.5	19.66	22.2	30.09

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
199	4mrfrfokrfids	nw	5	11.29	68.16	65.06	93.13	16.12	14.49	19.63	22.18	30.05
200	4mrfrfokrfids	nw	6	10.93	67.99	65.05	93.57	16.04	14.55	19.66	22.27	30.09
201	4mrfrfokrfids	nw	7	11.41	67.93	65.03	93.16	16.01	14.61	19.72	22.36	30.19
202	4mrfrfokrfids	nw	8	11.39	67.99	65.06	93.39	15.96	14.63	19.73	22.39	30.2
203	4mrfrfokrfids	nw	9	11.68	67.94	65.07	93.58	15.93	14.64	19.74	22.41	30.22
204	4mrfrfokrfids	nw	10	11.95	67.99	65.08	93.41	15.87	14.6	19.7	22.35	30.15
205	4mrfrfokrfids	nw	11	11.72	68.01	65.07	93.27	15.88	14.59	19.7	22.33	30.15
206	4mrfrfokrfids	nw	12	11.79	67.96	65.06	93.33	15.85	14.6	19.7	22.35	30.15
207	4mrfrfokrfids	nw	13	11.58	67.95	65.06	93.33	15.86	14.61	19.7	22.36	30.15
208	4mrfrfokrfids	nw	14	11.48	67.99	65.06	92.88	15.85	14.61	19.71	22.36	30.17
209	4mrfrfokrfids	nw	15	11.44	68.04	65.06	92.8	15.84	14.6	19.7	22.35	30.15
210	4mrfrfokrfids	nw	16	11.64	68.06	65.06	92.59	15.84	14.6	19.7	22.35	30.15
211	4mrfrfokrfids	nw	17	11.41	68.09	65.06	92.57	15.84	14.6	19.7	22.35	30.15
212	4mrfrfokrfids	nw	18	11.23	68.01	65.05	92.66	15.84	14.6	19.69	22.35	30.14
213	4mrfrfokrfids	nw	19	11.24	68.03	65.06	92.77	15.83	14.59	19.69	22.33	30.14
214	4mrfrfokrfids	nw	20	11.03	68.04	65.05	92.77	15.82	14.6	19.69	22.35	30.14
215	4mrfrfokrfids	nw	21	11.22	68.04	65.06	92.71	15.82	14.6	19.69	22.35	30.14
216	4mrfrfokrfids	nw	22	11.26	67.99	65.05	92.7	15.83	14.6	19.69	22.35	30.14
217	4mrfrfokrfids	nw	23	11.62	68	65.06	92.7	15.81	14.6	19.69	22.35	30.14
218	4mrfrfokrfids	nw	24	11.63	68	65.05	92.75	15.82	14.59	19.68	22.33	30.12
219	4mrfrfokrfids	nw	25	11.61	67.99	65.06	92.83	15.8	14.59	19.68	22.33	30.12
220	4mrfrfokrfids	nw	Inf	11.9	68.12	65.07	92.87	15.72	14.62	19.69	22.38	30.14
221	6rf	nw	na	14.54	67.55	65.08	90.04	14.63	14.86	19.65	22.75	30.08
222	6rfids	nw	12	7.87	68.25	65.02	93.71	16.72	14.55	19.81	22.27	30.32
223	6rfok	nw	12	12	68.38	65.18	93.19	15.61	14.74	19.65	22.56	30.08
224	6rfokrfids	nw	12	11.65	68.37	65.1	93.45	16.09	14.6	19.66	22.35	30.09
225	6rfrfokrfids	nw	12	12.61	68.27	65.09	92.31	15.54	14.65	19.6	22.42	30
226	bdt	nw	na	11.99	69.31	65.65	90.86	15.01	15.1	19.95	23.11	30.54
227	bdtids	nw	12	7.87	68.59	65.24	95.02	17.32	14.56	19.99	22.29	30.6

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
228	bdtok	nw	12	12.22	68.17	65.21	93.05	15.88	15.1	20.02	23.11	30.64
229	grnn	nw	na	0	71.34	65.02	97.53	15.85	16.65	22.14	25.49	33.89
230	grnnids	nw	12	0	68.54	64.64	96.21	19.47	15.21	21.66	23.28	33.15
231	grnnok	nw	12	0	68.79	64.8	96.17	18.04	15.5	21.59	23.73	33.05
232	i4rf	nw	na	14.68	67.47	65	89.48	14.46	15.48	20.22	23.7	30.95
233	i4rfids	nw	12	7.87	68.65	64.95	93.14	16.33	14.93	20.05	22.85	30.69
234	i4rfok	nw	12	11.01	67.84	65.14	92.39	15.47	15.11	19.97	23.13	30.57
235	i4rfokrfids	nw	12	11.19	68.28	65.04	92.76	15.84	14.97	19.96	22.91	30.55
236	i4rfrfokrfids	nw	12	12.35	68.05	65.03	91.67	15.3	15.09	19.97	23.1	30.57
237	ids	nw	12	7.89	69	65.52	93.66	16.74	14.66	20.35	22.44	31.15
238	irf	nw	na	14.03	67.81	65.01	90.65	14.74	15.06	19.84	23.05	30.37
239	irfids	nw	12	7.87	68.08	64.93	93.5	16.55	14.65	19.84	22.42	30.37
240	irfok	nw	12	11.65	68.12	65.1	93.1	15.68	14.83	19.75	22.7	30.23
241	irfokrfids	nw	12	11.74	68.3	65.01	93.3	16.06	14.7	19.74	22.5	30.22
242	irfrfokrfids	nw	12	12.5	68.03	65.01	92.42	15.55	14.78	19.71	22.62	30.17
243	ked	nw	Inf	0	69.59	65.54	93.56	15.52	14.61	19.75	22.36	30.23
244	lsvm	nw	na	0	70.05	68.03	80.5	11.74	16.95	21.73	25.95	33.26
245	lsvmids	nw	12	0	68.66	65.13	93.61	17.73	14.64	20.34	22.41	31.13
246	lsvmok	nw	12	0	69.38	65.33	91.69	16.02	14.82	20.1	22.68	30.77
247	lsvmoksvmids	nw	12	0	69.43	65.23	91.81	16.65	14.6	20.02	22.35	30.64
248	lsvmsvmoksvmids	nw	12	0	69.69	66.16	86.82	14.13	14.77	19.79	22.61	30.29
249	ok	nw	12	21.09	68.36	65.57	93.51	14.79	15.05	20.09	23.04	30.75
250	rf	nw	na	12.52	67.74	65	91.56	15.1	14.78	19.66	22.62	30.09
251	rfids	nw	4	7.87	69.15	65.02	93.87	17.53	14.47	19.92	22.15	30.49
252	rfids	nw	5	7.87	68.72	64.97	93.73	17.4	14.41	19.85	22.06	30.38
253	rfids	nw	6	7.87	68.25	64.97	94	17.26	14.45	19.85	22.12	30.38
254	rfids	nw	7	7.87	68.26	64.94	93.85	17.21	14.49	19.89	22.18	30.45
255	rfids	nw	8	7.87	68.26	64.97	93.93	17.14	14.51	19.89	22.21	30.45
256	rfids	nw	9	7.87	68.24	64.98	94	17.09	14.53	19.91	22.24	30.48

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
257	rfids	nw	10	7.87	68.16	64.98	93.94	17.01	14.5	19.87	22.2	30.41
258	rfids	nw	11	7.87	68.2	64.98	93.9	16.99	14.49	19.85	22.18	30.38
259	rfids	nw	12	7.87	67.95	64.95	93.91	16.96	14.5	19.85	22.2	30.38
260	rfids	nw	13	7.87	68.01	64.94	93.92	16.94	14.51	19.84	22.21	30.37
261	rfids	nw	14	7.87	68.05	64.95	93.73	16.92	14.51	19.85	22.21	30.38
262	rfids	nw	15	7.87	68	64.95	93.68	16.9	14.51	19.84	22.21	30.37
263	rfids	nw	16	7.87	68.08	64.95	93.6	16.88	14.5	19.84	22.2	30.37
264	rfids	nw	17	7.87	68.09	64.95	93.58	16.87	14.5	19.83	22.2	30.35
265	rfids	nw	18	7.87	68.09	64.94	93.6	16.86	14.5	19.82	22.2	30.34
266	rfids	nw	19	7.87	68.08	64.94	93.64	16.84	14.49	19.81	22.18	30.32
267	rfids	nw	20	7.87	68.01	64.94	93.63	16.83	14.5	19.8	22.2	30.31
268	rfids	nw	21	7.87	67.98	64.94	93.59	16.82	14.5	19.8	22.2	30.31
269	rfids	nw	22	7.87	67.86	64.94	93.58	16.82	14.5	19.8	22.2	30.31
270	rfids	nw	23	7.87	67.89	64.94	93.57	16.81	14.5	19.8	22.2	30.31
271	rfids	nw	24	7.87	67.97	64.94	93.59	16.8	14.5	19.8	22.2	30.31
272	rfids	nw	25	7.87	67.94	64.94	93.63	16.79	14.5	19.79	22.2	30.29
273	rfids	nw	Inf	7.87	68.04	64.94	93.46	16.52	14.55	19.8	22.27	30.31
274	rfok	nw	4	9.87	69.14	65.2	92.69	16.63	14.57	19.74	22.3	30.22
275	rfok	nw	5	10.84	68.27	65.11	93.19	16.42	14.51	19.65	22.21	30.08
276	rfok	nw	6	9.28	68.15	65.09	94.22	16.26	14.61	19.67	22.36	30.11
277	rfok	nw	7	9.93	67.81	65.06	93.31	16.23	14.72	19.78	22.53	30.28
278	rfok	nw	8	9.87	68.17	65.12	93.81	16.13	14.75	19.8	22.58	30.31
279	rfok	nw	9	10.58	68.04	65.13	94.19	16.08	14.75	19.79	22.58	30.29
280	rfok	nw	10	11.01	68.18	65.15	93.7	15.97	14.65	19.71	22.42	30.17
281	rfok	nw	11	10.63	68.02	65.13	93.33	16	14.65	19.7	22.42	30.15
282	rfok	nw	12	10.63	67.89	65.11	93.52	15.95	14.67	19.7	22.46	30.15
283	rfok	nw	13	10.47	68.11	65.12	93.59	15.96	14.68	19.71	22.47	30.17
284	rfok	nw	14	10.12	68.12	65.14	92.44	15.95	14.67	19.71	22.46	30.17
285	rfok	nw	15	10.09	68.09	65.12	92.24	15.95	14.63	19.68	22.39	30.12

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
286	rfok	nw	16	10.35	68.13	65.12	91.73	15.94	14.64	19.68	22.41	30.12
287	rfok	nw	17	9.83	68.12	65.13	91.69	15.95	14.65	19.69	22.42	30.14
288	rfok	nw	18	9.53	68.11	65.12	91.93	15.96	14.64	19.68	22.41	30.12
289	rfok	nw	19	9.56	68.16	65.14	92.28	15.95	14.63	19.68	22.39	30.12
290	rfok	nw	20	9.13	68.15	65.13	92.3	15.95	14.64	19.68	22.41	30.12
291	rfok	nw	21	9.61	68.09	65.14	92.17	15.94	14.64	19.68	22.41	30.12
292	rfok	nw	22	9.69	68.1	65.13	92.2	15.96	14.64	19.68	22.41	30.12
293	rfok	nw	23	10.62	68.15	65.13	92.21	15.94	14.63	19.67	22.39	30.11
294	rfok	nw	24	10.66	68.14	65.14	92.35	15.95	14.62	19.66	22.38	30.09
295	rfok	nw	25	10.63	68.12	65.14	92.56	15.93	14.62	19.66	22.38	30.09
296	rfok	nw	Inf	11.1	68.06	65.2	92.8	15.91	14.6	19.64	22.35	30.06
297	rfokrfids	nw	4	10.03	69.32	65.11	93.19	17.02	14.47	19.77	22.15	30.26
298	rfokrfids	nw	5	10.63	68.37	65.04	93.38	16.85	14.41	19.69	22.06	30.14
299	rfokrfids	nw	6	9.69	68.06	65.03	94.11	16.7	14.48	19.7	22.16	30.15
300	rfokrfids	nw	7	10.07	68.09	65	93.58	16.65	14.57	19.77	22.3	30.26
301	rfokrfids	nw	8	10.03	68.26	65.04	93.84	16.57	14.58	19.79	22.32	30.29
302	rfokrfids	nw	9	10.49	68.11	65.06	94.1	16.52	14.59	19.79	22.33	30.29
303	rfokrfids	nw	10	10.81	68.11	65.06	93.82	16.43	14.53	19.73	22.24	30.2
304	rfokrfids	nw	11	10.59	68	65.05	93.61	16.43	14.52	19.72	22.23	30.19
305	rfokrfids	nw	12	10.61	68.02	65.03	93.72	16.4	14.54	19.72	22.26	30.19
306	rfokrfids	nw	13	10.5	68.11	65.03	93.75	16.39	14.55	19.72	22.27	30.19
307	rfokrfids	nw	14	10.29	68.07	65.05	93.08	16.38	14.54	19.73	22.26	30.2
308	rfokrfids	nw	15	10.26	67.94	65.04	92.96	16.37	14.52	19.71	22.23	30.17
309	rfokrfids	nw	16	10.43	68.13	65.04	92.66	16.36	14.53	19.71	22.24	30.17
310	rfokrfids	nw	17	10.12	68.06	65.04	92.63	16.36	14.53	19.7	22.24	30.15
311	rfokrfids	nw	18	9.92	67.96	65.03	92.76	16.36	14.53	19.7	22.24	30.15
312	rfokrfids	nw	19	9.94	68.11	65.04	92.96	16.34	14.52	19.69	22.23	30.14
313	rfokrfids	nw	20	9.69	68.01	65.04	92.97	16.33	14.53	19.69	22.24	30.14
314	rfokrfids	nw	21	9.96	68.1	65.04	92.88	16.32	14.53	19.69	22.24	30.14

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
315	rfokrfids	nw	22	10	68	65.03	92.89	16.34	14.53	19.69	22.24	30.14
316	rfokrfids	nw	23	10.52	68.04	65.04	92.89	16.32	14.52	19.69	22.23	30.14
317	rfokrfids	nw	24	10.55	68.09	65.04	92.97	16.32	14.52	19.68	22.23	30.12
318	rfokrfids	nw	25	10.53	68.1	65.04	93.1	16.3	14.51	19.68	22.21	30.12
319	rfokrfids	nw	Inf	10.96	68.09	65.07	93.13	16.16	14.53	19.67	22.24	30.11
320	rfrfokrfids	nw	4	10.86	68.33	65.07	92.65	16.27	14.5	19.62	22.2	30.03
321	rfrfokrfids	nw	5	11.26	68.36	65.03	92.77	16.17	14.47	19.59	22.15	29.99
322	rfrfokrfids	nw	6	10.63	67.72	65.02	93.26	16.09	14.53	19.61	22.24	30.02
323	rfrfokrfids	nw	7	10.88	67.79	65	92.91	16.06	14.59	19.67	22.33	30.11
324	rfrfokrfids	nw	8	10.86	67.95	65.03	93.08	16.01	14.6	19.68	22.35	30.12
325	rfrfokrfids	nw	9	11.17	67.9	65.04	93.25	15.99	14.61	19.69	22.36	30.14
326	rfrfokrfids	nw	10	11.38	67.95	65.04	93.07	15.93	14.57	19.65	22.3	30.08
327	rfrfokrfids	nw	11	11.24	67.94	65.04	92.93	15.93	14.57	19.65	22.3	30.08
328	rfrfokrfids	nw	12	11.24	67.9	65.02	93	15.91	14.58	19.65	22.32	30.08
329	rfrfokrfids	nw	13	11.17	67.87	65.02	93.02	15.91	14.59	19.65	22.33	30.08
330	rfrfokrfids	nw	14	11.03	67.96	65.03	92.57	15.9	14.59	19.65	22.33	30.08
331	rfrfokrfids	nw	15	11.01	67.96	65.02	92.49	15.9	14.57	19.64	22.3	30.06
332	rfrfokrfids	nw	16	11.12	68.07	65.02	92.29	15.89	14.57	19.64	22.3	30.06
333	rfrfokrfids	nw	17	10.92	67.91	65.03	92.27	15.89	14.58	19.64	22.32	30.06
334	rfrfokrfids	nw	18	10.79	67.98	65.02	92.36	15.89	14.57	19.64	22.3	30.06
335	rfrfokrfids	nw	19	10.8	67.95	65.03	92.49	15.88	14.57	19.64	22.3	30.06
336	rfrfokrfids	nw	20	10.63	67.95	65.02	92.5	15.88	14.57	19.64	22.3	30.06
337	rfrfokrfids	nw	21	10.81	67.9	65.03	92.44	15.87	14.58	19.63	22.32	30.05
338	rfrfokrfids	nw	22	10.84	67.92	65.02	92.45	15.88	14.57	19.63	22.3	30.05
339	rfrfokrfids	nw	23	11.18	67.89	65.02	92.45	15.87	14.57	19.63	22.3	30.05
340	rfrfokrfids	nw	24	11.21	67.92	65.02	92.5	15.87	14.57	19.63	22.3	30.05
341	rfrfokrfids	nw	25	11.19	67.92	65.03	92.58	15.85	14.57	19.63	22.3	30.05
342	rfrfokrfids	nw	Inf	11.48	68.02	65.05	92.61	15.77	14.59	19.64	22.33	30.06
343	svm	nw	na	13.59	72.23	67.15	88.21	13.52	15.68	20.78	24	31.81

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
344	svmids	nw	12	7.89	68.93	65.36	93.82	17.09	14.77	20.48	22.61	31.35
345	svmok	nw	12	18.35	68.68	65.64	92.59	15.09	15.02	20.26	22.99	31.01
346	svmoksvmids	nw	12	16.62	69.6	65.5	91.92	15.83	14.72	20.14	22.53	30.83
347	svmsvmoksvmids	nw	12	15.61	70	66.05	88.16	14.5	14.71	19.9	22.52	30.46
348	6rf	sw	na	12.14	51.55	52.01	93.27	24.92	14.7	19.16	28.01	36.51
349	6rfids	sw	12	6.52	51.57	52.38	97.5	27.47	13.7	18.47	26.11	35.19
350	6rfok	sw	12	8.88	52.38	52.02	94.2	25.94	14.29	18.88	27.23	35.98
351	6rfokrfids	sw	12	9.13	51.95	52.2	94.75	26.61	13.89	18.52	26.47	35.29
352	6rfrfokrfids	sw	12	10.23	52.13	52.13	94.09	25.99	14.08	18.64	26.83	35.52
353	bdt	sw	na	10.64	50.64	52.08	97.87	26.68	14.91	19.85	28.41	37.82
354	bdtids	sw	12	6.94	52.1	52.52	100	28.9	13.85	19.05	26.39	36.3
355	bdtok	sw	12	9.09	52.12	52.17	98.76	27.32	14.6	19.64	27.82	37.42
356	grnn	sw	na	0	49.95	50.67	99.55	28.59	18.25	23.95	34.78	45.64
357	grnnids	sw	12	0	53.98	52.4	100	30	14.42	20.42	27.48	38.91
358	grnnok	sw	12	1.09	53.19	50.98	100	28.46	16.19	22	30.85	41.92
359	i4rf	sw	na	12.43	53.46	51.86	93.45	26.19	13.82	18.69	26.33	35.61
360	i4rfids	sw	12	7.72	52.66	52.26	98.05	28.02	13.39	18.46	25.51	35.18
361	i4rfok	sw	12	10.41	53.04	52.05	93.95	26.8	13.75	18.61	26.2	35.46
362	i4rfokrfids	sw	12	9.69	52.67	52.16	95.64	27.33	13.48	18.42	25.69	35.1
363	i4rfrfokrfids	sw	12	10.78	51.99	52.06	94.44	26.92	13.57	18.46	25.86	35.18
364	ids	sw	12	4.54	52.87	53.59	95.95	27.65	14.31	19.48	27.27	37.12
365	irf	sw	na	12.17	52.1	51.97	93.87	26.27	13.93	18.71	26.54	35.65
366	irfids	sw	12	7.1	53.12	52.38	98.23	28.05	13.49	18.54	25.71	35.33
367	irfok	sw	12	9.25	53.22	52.08	94.48	26.79	13.86	18.65	26.41	35.54
368	irfokrfids	sw	12	8.89	52.31	52.23	96.09	27.35	13.58	18.48	25.88	35.21
369	irfrfokrfids	sw	12	9.98	53.25	52.14	95.05	26.96	13.67	18.51	26.05	35.27
370	ked	sw	Inf	3.44	55.56	54.43	94.06	25.46	16.45	20.87	31.35	39.77
371	lsvm	sw	na	0	67.17	55.08	89.87	25.17	20.04	24.33	38.19	46.36
372	lsvmids	sw	12	0	53.94	52.71	96.95	29.2	14.7	19.48	28.01	37.12

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
373	lsvmok	sw	12	0	54.31	51.23	94.32	26.73	16.79	21.27	31.99	40.53
374	lsvmoksvmids	sw	12	0	55.01	51.97	92.68	27.48	15.43	19.67	29.4	37.48
375	lsvmsvmoksvmids	sw	12	0	57.67	53.01	88.32	25.85	16.61	20.17	31.65	38.43
376	ok	sw	12	10.73	51.21	54.12	96.92	26.86	14.97	19.98	28.53	38.07
377	rf	sw	na	11.6	53.62	52.19	93.71	26.21	14.04	18.71	26.75	35.65
378	rfids	sw	4	6.53	53.83	52.75	97.85	28.59	13.48	18.57	25.69	35.38
379	rfids	sw	5	6.55	53.96	52.68	97.71	28.43	13.48	18.56	25.69	35.37
380	rfids	sw	6	6.56	53.29	52.63	97.62	28.29	13.55	18.58	25.82	35.4
381	rfids	sw	7	6.55	53.65	52.58	97.61	28.21	13.54	18.56	25.8	35.37
382	rfids	sw	8	6.56	53.65	52.55	97.56	28.17	13.58	18.59	25.88	35.42
383	rfids	sw	9	6.57	53.63	52.57	97.56	28.14	13.59	18.6	25.9	35.44
384	rfids	sw	10	6.58	53.64	52.56	97.5	28.09	13.58	18.58	25.88	35.4
385	rfids	sw	11	6.58	53.56	52.55	97.46	28.07	13.6	18.58	25.91	35.4
386	rfids	sw	12	6.58	53.67	52.57	97.4	28.05	13.59	18.57	25.9	35.38
387	rfids	sw	13	6.59	53.68	52.56	97.39	28.03	13.6	18.58	25.91	35.4
388	rfids	sw	14	6.59	53.76	52.54	97.38	28.01	13.61	18.58	25.93	35.4
389	rfids	sw	15	6.6	53.8	52.54	97.36	27.99	13.62	18.59	25.95	35.42
390	rfids	sw	16	6.6	53.78	52.52	97.37	27.97	13.63	18.6	25.97	35.44
391	rfids	sw	17	6.6	53.74	52.53	97.36	27.96	13.63	18.6	25.97	35.44
392	rfids	sw	18	6.61	53.76	52.52	97.36	27.94	13.64	18.6	25.99	35.44
393	rfids	sw	19	6.61	53.8	52.53	97.31	27.93	13.64	18.59	25.99	35.42
394	rfids	sw	20	6.61	53.86	52.53	97.31	27.92	13.64	18.6	25.99	35.44
395	rfids	sw	21	6.61	53.8	52.53	97.26	27.9	13.64	18.59	25.99	35.42
396	rfids	sw	22	6.62	53.73	52.53	97.26	27.9	13.65	18.61	26.01	35.46
397	rfids	sw	23	6.62	53.73	52.53	97.27	27.89	13.66	18.61	26.03	35.46
398	rfids	sw	24	6.63	53.8	52.53	97.26	27.88	13.67	18.62	26.05	35.48
399	rfids	sw	25	6.63	53.79	52.53	97.26	27.87	13.67	18.62	26.05	35.48
400	rfids	sw	Inf	6.79	53.06	52.51	97.1	27.68	13.7	18.6	26.11	35.44
401	rfok	sw	4	8.07	53.51	52.64	97.44	27.82	13.36	18.3	25.46	34.87

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
402	rfok	sw	5	11.16	53.26	52.47	95.93	27.4	13.47	18.28	25.67	34.83
403	rfok	sw	6	10.06	53.37	52.4	96.49	27.09	13.72	18.47	26.14	35.19
404	rfok	sw	7	9.76	53.66	52.18	94.37	26.91	13.8	18.49	26.3	35.23
405	rfok	sw	8	9.81	53.42	52.11	94.97	26.86	13.98	18.65	26.64	35.54
406	rfok	sw	9	10.1	53.62	52.2	95.94	26.84	13.97	18.65	26.62	35.54
407	rfok	sw	10	8.87	53.44	52.19	94.73	26.72	13.98	18.63	26.64	35.5
408	rfok	sw	11	9.18	53.35	52.19	95.09	26.75	13.94	18.61	26.56	35.46
409	rfok	sw	12	8.39	52.97	52.27	93.94	26.69	13.95	18.62	26.58	35.48
410	rfok	sw	13	8.35	53.23	52.23	93.3	26.64	13.98	18.65	26.64	35.54
411	rfok	sw	14	8.95	53.53	52.22	93.62	26.6	13.98	18.66	26.64	35.56
412	rfok	sw	15	10.49	53.78	52.23	93.96	26.54	14.01	18.68	26.7	35.59
413	rfok	sw	16	10.83	53.82	52.18	93.67	26.51	14.03	18.7	26.73	35.63
414	rfok	sw	17	10.47	53.71	52.18	93.93	26.5	14.01	18.69	26.7	35.61
415	rfok	sw	18	10.77	53.61	52.13	94.04	26.45	14.02	18.7	26.71	35.63
416	rfok	sw	19	10.62	53.75	52.19	94.07	26.44	14	18.67	26.68	35.58
417	rfok	sw	20	10.67	53.73	52.23	94.18	26.43	13.98	18.67	26.64	35.58
418	rfok	sw	21	10.57	53.66	52.21	94.36	26.44	13.94	18.63	26.56	35.5
419	rfok	sw	22	10.56	53.51	52.21	94.21	26.46	13.93	18.62	26.54	35.48
420	rfok	sw	23	10.52	53.53	52.22	94.21	26.44	13.96	18.63	26.6	35.5
421	rfok	sw	24	10.5	53.68	52.2	94.43	26.46	13.96	18.64	26.6	35.52
422	rfok	sw	25	10.55	53.45	52.19	94.63	26.45	13.95	18.62	26.58	35.48
423	rfok	sw	Inf	10.65	53.78	52.26	94.35	26.39	13.97	18.6	26.62	35.44
424	rfokrfids	sw	4	7.3	53.71	52.7	97.3	28.15	13.36	18.36	25.46	34.98
425	rfokrfids	sw	5	8.86	53.82	52.57	96.08	27.86	13.38	18.33	25.5	34.93
426	rfokrfids	sw	6	9.09	53.44	52.52	96.46	27.62	13.54	18.42	25.8	35.1
427	rfokrfids	sw	7	8.16	53.36	52.38	95.06	27.5	13.57	18.41	25.86	35.08
428	rfokrfids	sw	8	8.76	53.5	52.33	95.45	27.45	13.68	18.51	26.07	35.27
429	rfokrfids	sw	9	9.28	53.43	52.38	95.24	27.43	13.69	18.51	26.09	35.27
430	rfokrfids	sw	10	9.1	53.12	52.37	95.13	27.33	13.69	18.49	26.09	35.23

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
431	rfokrfids	sw	11	9.27	53.05	52.37	95.35	27.34	13.69	18.47	26.09	35.19
432	rfokrfids	sw	12	8.84	53.24	52.42	94.63	27.3	13.68	18.48	26.07	35.21
433	rfokrfids	sw	13	8.82	53.17	52.4	94.46	27.26	13.69	18.49	26.09	35.23
434	rfokrfids	sw	14	9.15	53.3	52.38	94.58	27.23	13.7	18.5	26.11	35.25
435	rfokrfids	sw	15	9.99	53.69	52.38	94.59	27.19	13.71	18.51	26.12	35.27
436	rfokrfids	sw	16	9.95	53.49	52.35	94.8	27.17	13.72	18.53	26.14	35.31
437	rfokrfids	sw	17	9.92	53.46	52.35	94.84	27.16	13.72	18.52	26.14	35.29
438	rfokrfids	sw	18	10.08	53.65	52.32	94.99	27.12	13.73	18.52	26.16	35.29
439	rfokrfids	sw	19	10.1	53.61	52.36	94.78	27.11	13.72	18.51	26.14	35.27
440	rfokrfids	sw	20	10.13	53.64	52.38	94.72	27.1	13.72	18.51	26.14	35.27
441	rfokrfids	sw	21	10.01	53.47	52.37	94.93	27.1	13.7	18.49	26.11	35.23
442	rfokrfids	sw	22	10.08	53.44	52.37	94.85	27.11	13.7	18.49	26.11	35.23
443	rfokrfids	sw	23	10.05	53.54	52.38	94.84	27.09	13.71	18.5	26.12	35.25
444	rfokrfids	sw	24	10.04	53.6	52.36	94.97	27.1	13.72	18.51	26.14	35.27
445	rfokrfids	sw	25	10.08	53.52	52.36	95.09	27.09	13.71	18.5	26.12	35.25
446	rfokrfids	sw	Inf	10.36	53.25	52.38	94.84	26.98	13.75	18.5	26.2	35.25
447	rfrfokrfids	sw	4	9.5	53.7	52.53	95.99	27.43	13.5	18.34	25.72	34.95
448	rfrfokrfids	sw	5	10.54	53.77	52.45	95.09	27.25	13.55	18.35	25.82	34.97
449	rfrfokrfids	sw	6	10.37	53.64	52.41	95.34	27.1	13.68	18.43	26.07	35.12
450	rfrfokrfids	sw	7	10.07	53.67	52.32	94.46	27.02	13.71	18.44	26.12	35.14
451	rfrfokrfids	sw	8	10.29	53.96	52.29	94.75	27	13.79	18.51	26.28	35.27
452	rfrfokrfids	sw	9	10.4	53.8	52.32	94.62	26.98	13.79	18.51	26.28	35.27
453	rfrfokrfids	sw	10	9.94	53.67	52.31	94.45	26.92	13.79	18.51	26.28	35.27
454	rfrfokrfids	sw	11	10.05	53.41	52.31	94.6	26.93	13.79	18.5	26.28	35.25
455	rfrfokrfids	sw	12	9.76	53.77	52.34	94.12	26.91	13.79	18.51	26.28	35.27
456	rfrfokrfids	sw	13	9.74	53.86	52.33	93.93	26.88	13.8	18.52	26.3	35.29
457	rfrfokrfids	sw	14	9.97	53.78	52.32	94.03	26.86	13.8	18.52	26.3	35.29
458	rfrfokrfids	sw	15	10.53	53.87	52.32	94.1	26.84	13.82	18.53	26.33	35.31
459	rfrfokrfids	sw	16	10.66	53.73	52.3	94.02	26.82	13.82	18.55	26.33	35.35

NO	METHOD	REGION	WINDOW SEARCH SIZE	MINIMUM	MEDIAN	MEAN	MAXIMUM	STANDARD DEVIATION	MAE	RMSE	RMAE (%)	RRMSE (%)
460	rfrfokrfids	sw	17	10.53	53.71	52.3	94.03	26.82	13.82	18.55	26.33	35.35
461	rfrfokrfids	sw	18	10.65	53.76	52.28	94.3	26.79	13.83	18.55	26.35	35.35
462	rfrfokrfids	sw	19	10.6	53.78	52.3	94.3	26.79	13.82	18.54	26.33	35.33
463	rfrfokrfids	sw	20	10.62	53.79	52.32	94.27	26.78	13.82	18.54	26.33	35.33
464	rfrfokrfids	sw	21	10.59	53.72	52.31	94.41	26.78	13.8	18.53	26.3	35.31
465	rfrfokrfids	sw	22	10.58	53.5	52.31	94.35	26.79	13.8	18.53	26.3	35.31
466	rfrfokrfids	sw	23	10.57	53.55	52.31	94.35	26.78	13.81	18.53	26.31	35.31
467	rfrfokrfids	sw	24	10.56	53.67	52.3	94.43	26.78	13.82	18.54	26.33	35.33
468	rfrfokrfids	sw	25	10.59	53.68	52.3	94.51	26.78	13.81	18.54	26.31	35.33
469	rfrfokrfids	sw	Inf	10.78	53.62	52.32	94.28	26.7	13.84	18.54	26.37	35.33
470	svm	sw	na	0	55.37	55.26	97.04	27.05	17.48	22.81	33.31	43.46
471	svmids	sw	12	2.92	53.35	53.19	95.34	28.07	14.45	19.53	27.53	37.21
472	svmok	sw	12	2.55	52.63	51.82	93.74	25.74	16.1	20.94	30.68	39.9
473	svmoksvmids	sw	12	3.34	52.41	52.5	92.51	26.51	15.1	19.68	28.77	37.5
474	svmsvmoksvmids	sw	12	2.23	52.42	53.42	91.38	26.2	15.56	20.09	29.65	38.28