Remote Sensing Application for Assessing Salinity Intrusion in the Mekong Delta, Vietnam

Nguyen Thi Bich Phuong

Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Environmental Management
Technology (International Program)
Prince of Songkla University
2018

Remote Sensing Application for Assessing Salinity Intrusion in the Mekong Delta, Vietnam

Nguyen Thi Bich Phuong

Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Environmental Management
Technology (International Program)
Prince of Songkla University
2018

Thesis Title	Remote Sensing Application for Assessing Salinity Intrusion in the Mekong Delta, Vietnam
Author	Miss. Nguyen Thi Bich Phuong
Major Program	Environmental Management and Technology (International Program)

Major Advisor	**Examining Committee**
.....	..... Chairperson
(Asst. Prof. Dr. Werapong Koedsin)	(Assoc. Prof. Dr. Pun Thongchumnum)
Co-advisor	..... Committee
	(Assoc. Prof. Dr. Raymond J. RITCHIE)
.....	..... Committee
(Dr. Sangdao Wongsai)	(Asst. Prof. Dr. Werapong Koedsin)
	..... Committee
	(Dr. Sangdao Wongsai)
	..... Committee
	(Dr. Pariwate Varnakovida)

The Graduate School, Prince of Songkla University, has approved this thesis as partial fulfillment of the requirements for the Master of Science Degree in Environmental Management and Technology.

.....

(Assoc. Prof. Dr. Damrongsak Faroongsarng)

Dean of Graduate School

This is to certify that the work here submitted is the result of the candidate's own investigations. Due acknowledgement has been made of any assistance recieved.

.....Signature
(Asst. Prof. Dr. Werapong Koedsin)
Major Advisor

.....Signature
(Miss. Nguyen Thi Bich Phuong)
Candidate

I hereby certify that this work has not been accepted in substance for any other degree, and is not being currently submitted in candidature for any degree.

.....Signature

(Miss. Nguyen Thi Bich Phuong)

Candiddate

Thesis Title	Remote Sensing Application for Assessing Salinity Intrusion in the Mekong Delta, Vietnam
Author	Miss Nguyen Thi Bich Phuong
Major Program	Environmental Management Technology (International Program)
Academic year	2017

ABSTRACT

Salinity intrusion is a complex issue in coastal areas. Currently, remote sensing techniques have been widely used to monitor water quality changes, ranging from inland river networks to deep oceans. The Vietnamese Mekong Delta (VMD) is an important rice-growing area and intrusion of saline water into irrigated freshwater-based agriculture areas is one of the most crucial constraints for agriculture development. This study aimed at building a numerical model to realize the salinity intrusion through the relationship between reflectance from the Landsat-8 Operational Land Imager (OLI) images and salinity levels measured in-situ. 103 observed samples were divided into 50% training and 50% test. The Multiple Linear Regression (MLR), Decision Trees (DTs) and Random Forest (RF) approaches were applied in the study. The result showed that the RF approach was the best model to estimate salinity along the coastal river network in the study area. However, the large samples size needed was a significant challenge to circumscribe the predicting ability of the RF models.

The reflectance was found good to have a correlation with salinity when locations (latitude – longitude) of salinity measured station were added as a parameter of the Step-wise model. The R-square values were 77.48% in training and 74.16% in test while RMSE was smaller than 3. The reflectance – location model was employed for mapping salinity intrusion on 24th Jan 2015 and 09th Feb 2015 recognized changes of salinity concentration in the whole study area. However,

locality issue was a limitation for mapping salinity by using latitude and longitude as parameters.

On the other hand, the real data was used for a re-sampling routine where data was performed re-sampling two times and four times by using bootstrap method. Four statistical models including the DTs, the MLR the RF and the Neural Network (ANN and ANT) were applied. Larger sample sizes that are regularly updated are needed to more fully develop the model. The best model was performed by the RF in re-sampling data four times which was employed for mapping salinity in early dry season 2015. The salinity map on 24th Jan and 09th Feb distinguished the tendency of salinity level as well as salinity dynamics and recognized changes of salinity concentration from upstream to downstream. This study proved the possibility of using the Landsat-8 images for mapping salinity as a useful tool to support the early warning system in the future in the VMD.

Keywords: *Salinity intrusion model, samples size, remote sensing and the Vietnamese Mekong Delta*

ACKNOWLEDMENT

This work was supported by the Higher Education Research Promotion and the Thailand's Education Hub for Southern Region of ASEAN Countries Project Office of the Higher Education Commission. I am grateful to acknowledge my advisor Asst. Prof. Dr. Werapong Koedsin for guiding me step by step to make this research a success and making me the best abilities. It was great learning experience that has enabled me to be equipped with academic skills.

I'm thankful to Assoc. Prof. Dr. Raymond J Ritchie who gave a lot of useful advice for this research; especially, Prof. Don McNeil who has a lot of contribution to develop statistic model in the R program and many important recommendations.

I would like to thank to all my lecturers who gave many useful course works and gains my motivation. Finally, I greatly appreciate all of my friends in Faculty of Technology and Environmental Management who are always by my side during the time I am studying in Prince of Songkla University, Phuket campus.

Nguyen Thi Bich Phuong

Contents

Contents (continued)

List of Tables

List of Figures

Figures **Page**

List of Figures (continued)

List of Abbreviations and Symbols

ANN	Artificial Neural Network
AVIRIS	Airborne Visible Infrared Imaging Spectrometer
BART	Bayesian Additive Regression Tree
CART	Categorical and Regression Tree model
CGIAR	Climate Change, Agriculture and Food Security
DTs	Decision Trees
EO-1 ALI	Earth Observing-1
FLASH	Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes
GAM	Generalized Additive Model
GLM	Generalized Linear Model
Landsat- TM	The Landsat Thematic Mapper
Landsat-8 OLI	Landsat-8 Operational Land Imager
MLR	Multiple Linear Regression
MARS	Multivariate Adaptive Regression Spline
MODIS	Moderate-resolution Imaging Spectroradiometer
OLS	Ordinary Least Square
RF	Random Forest
Sea WiFS	Sea-viewing Wide Field-of-view Sensor
SPOT	Satellite Pour l'Observation de la Terre
TIRS	Thermal Infrared Sensor
VMD	Vietnamese Mekong Delta

Chapter 1

Introduction

1.1 Salinity intrusion in Vietnamese Mekong Delta

Saline intrusion has been a difficult issue for fresh water supply in coastal areas in SE-Asia (Trung, *et al*., 2016;Himi, *et al*., 2017). Extraordinary saline, as well as short-term salinity intrusion, occurs earlier and longer as a slow-onset hazard and is difficult to predict even though it causes huge damages (CGIAR Research Centers in Southeast Asia, 2016; Binh, 2015). In the Vietnamese Mekong Delta (VMD), the salinity intrusion trend has increased in terms of concentration and duration, leading to damage to sustainable development (Binh,2015;Trung, *et al*., 2016). The Agriculture in the delta includes paddy rice, fruit and vegetables and are all vulnerable to salinity damage due to the lack of freshwater input and abnormal weather (Binh,2015). However, in providing salinity information the lack of the agility is a critical problem because farmers cannot keep in touch with information on salinity (CGIAR Research Centers in Southeast Asia, 2016). On the other hand, increasing uncertainties and complexities exacerbated by climate change and abnormal weather, guarantee that accurate information for salinity forecasting will become more difficult to obtain (CGIAR Research Centers in Southeast Asia, 2016). Improved early warning systems could disseminate forecasts and early warning messages to local communities and farmers. The usual salinity alert level for agriculture is about 5 ppt (Thai and Dung, 2013).

Vietnam is the world's second largest rice exporter of which is ninety percent comesfrom the Vietnamese Mekong Delta (VMD) (CGIAR Research Centers in Southeast Asia, 2016). Agriculture plays a main role in the economy of the VMD. The

average of paddy field areawas 4,295.2 thous.ha of which 1,555.7 thous.ha is in the Spring crop and the total aquaculture area was 772 thous.ha (General Statistic Office of Vietnam, 2016). However, the delta was damaged due to lack of freshwater from upstream in the dry season. In addition, rising sea water level is warned to be an issue which will enhance water scarcity in the area. Around 2.1 million hectares of the Mekong Delta coastal areas (or about 50%) are affected by salinity during the dry season (from December to May) (Tuan, *et al*., 2007). Salinity intrusion is considerd a natural phenomenon in the Vietnamese Mekong Delta which happens every year but the severity of intrusion varies greatly from year to year. However, in a short period during the dry season 2015-2016, salinity intrusion caused disastrous damage to economic and livelihood of local people in the whole VMD particularly in coastal areas. Saline intrusion has a large impact on sustainable development in the area.

The extent of salinity intrusion is a complex process dependent on many factors such as fresh water discharge from the upstream, capacity and morphology of channel, configuration of the drainage network, tidal conditions and the presence of control structures such as sluice gates (Hashimoto, 2001; Nguyen, *et al*., 2008). The magnitude of the floods, summer-autumn paddy production status, timing of the rainy season (Deltares and Delta Alliance, 2011; Nguyen, *et al*., 2008) and combined impact of sea level rise (Khang, *et al*., 2008) all contribute to the severity of salinity intrusion Ability to predict salinity intrusion was an issue of interest in many previous studies. Hydraulic models were assessed as a tool to simulate salinity intrusion in the VMD (Nguyen, *et al*., 2008; Khang, *et al*., 2008; Trung and Tri, 2014). However, these are complex models requiring a heavy set of data and built on the basis of detailed topographical information and heavily depend on the latest knowledge about the infrastructure (Nguyen, *et al*., 2008). Meanwhile, although satellite techniques open new potential applications to monitor freshwater input, coastal water and monitor open ocean salinity they are typically illsuited for high resolution applications to coastlines and estuaries (Urquhart, *et al*., 2012; Roy, *et al*., 2014) and a very effective technique to capture the dynamics of coastal systems is needed (Lira and Rui, 2014). The success of satellite images for predicting salinity intrusion was demonstrated in some previous studies using the Landsat Thematic

Mapper (Landsat TM) (Baban, 1997; Wang and Xu, 2008), the moderate-resolution imaging spectroradiometer (MODIS) (Urquhart, *et al*., 2012), the Earth Observing-1 (EO-1 ALI) (Fang, *et al*., 2010), the Sea-viewing Wide Field-of-view Sensor (Sea WiFS) (D'Sa, *et al*., 2002) and the newly operational Landsat-8 Operational Land Imager (OLI) *(Zhao, et al*., 2017).

Relationships between salinity and water reflectance remote sensing may be used to relate results measuring suspended solids and colored dissolved organic matters from the river discharge (Wang and Xu, 2008) to salinity. Thus, Total Suspended Solids has a good correlation with salinity in brackish water and can be used as an indirect way to detect salinity concentration (Fang, *et al*., 2010). Modelling salinity has also been performed using only single reflectance - remote sensing (Rrs) (Urquhart, *et al*., 2012; Zhao, *et al*., 2017). On the other hand, latitude and longitude (location) of collected data was found to have a strong correlation with salinity. The location was the most significant predictor variable in surface salinity estimation models (Urquhart, *et al*., 2012). However, the locality issue was anticipated in each particular study area and thus complex processes operating upon a water body need to be taken into account in order to develop statistical salinity models that are not heavily dependent uponlocation (Urquhart, *et al*., 2012).

Machine learning has been successfully applied for different studies in the remote sensing field (Waske, *et al*., 2009; Dev, *et al*., 2016). Eight statistic models including a Categorical and Regression Tree model (CART), a Generalized Linear Model (GLM), a Generalized Additive Model (GAM) a Random Forest Model, a Mean model, an Artificial Neural Network (ANN), a Multivariate Adaptive Regression Spline (MARS), and a Bayesian Additive Regression Tree (BART) were applied to find out the best model to predict salinity concentration in Chesapeake Bay, in the United States (Urquhart, *et al*., 2012). In another study, an ordinary least square regression was performed to determine relationships between salinity and reflectance in Lake Pontchartrain, the US Gulf of Mexico (Wang and Xu, 2008). A multivariable linear algorithm was employed for predicting sea surface salinity from remote sensing reflectance (Rrs) in the hypersaline Arabian Gulf (Zhao, *et al*., 2017). Therefore,

choosing a suitable model for detecting salinity depends on the characteristics of each study area.

Lack of in-situ data for supervised Modelling is one of the great challenges leading to unrealistic for application of salinity modelling (Urquhart, *et al*., 2012; Zhou and Zhang, 2016; Wang and Xu, 2012). Small sample sizes is related to overfitting which causes of a serious limitation to developing a useful model (Shcheglovitova and Anderson, 2013; Liu and Gillies, 2016). In such cases, the model does good fitting on the training data points but would not do well in predicting for new task in testing samples (Ratner, 2011). Boosting is considered to be one of the most powerful learning ensemble algorithms to be proposed recently which is a way of increasing the complexity of the primary model (Mohamed, *et al*., 2017). Using the bootstrap approach, it was possible to obtain a better understanding of the relationshipof different variables in the model (Tran and Tran, 2016).

An early warning system is necessary to provide timely and meaningful warning information. It must be appropriate for the preparedness measures available and act appropriately in sufficient time to decrease the possibility of harm or losses (Alessa, *et al*., 2015). The early warning system includes an identification of the risk to manage, planning to maintain an adequate monitoring system; an efficient communication system and preparedness for damage mitigation (Katalin, *et al*., 2016).

1.2 Scope and objective

This study sought to identify risk of salinity intrusion and develop a model using reflectance data from Landsat - 8 OLI as a new approach to predict salinity in the VMD. There are three main objectives:To develop a salinity intrusion model from wavelength - reflectance data (Landsat – 8 OLI) by using a statistic model;

(i) To determine the most useful reflectance wavelength – of the Landsat – 8 OLI image and choosing a suitable statistic model for predict salinity intrusion with limited salinity observations; and to examine trends of saline intrusion by a combination of reflectance wavelength and location factor.

(ii) Apply bootstrap methods for resampling data to examine influences of sample size on the accuracy of models and

(iii) To assess trends of saline intrusion in the short term that will be an important part of support for the early warning system in this study area.

Chapter 2

Literature Review

2.1 Application of remote sensing to manage surface water resource and salinity intrusion

The remote sensing techniques have been widely used to delineate the surface water bodies (Kumar and Reshmidevi, 2013) and allowed measurements on a global scale. In addition, high-resolution satellite data opens great potential to extend satellite remote sensing to create timely and reliable assessments of land and water resources at a local scale (Sawaya, *et al*., 2003). Selecting satellite images depends on objectives, financial conditions and characteristics of each case study which including Hyperspectral analysis: The Hyperion EO-1, Airborne Visible Infrared Imaging Spectrometer (AVIRIS) and Multispectral: Satellite Pour l'Observation de la Terre (SPOT), Sentinel, Landsat, Radar, MODIS. Low cost, and flexible to manage for the large area of the VMD were advantages for using free satellite images. However, poor resolution at local scales, due to policies of environmental managers and the general public (Keith, *et al*., 2014) and the time recycle of the satellite are also factors to take into account when deciding on asuitable satellite images. Factors to consider are thespatial resolution of the satellites, medium range (e.g., Landsat: 30m); coarser spatial resolution satellites (e.g., MODIS: 250–1000m) or high resolution (e.g., SENTINEL-2 resolution of 10–60 m); frequency of overpasses to address temporal variability (e.g., Landsat every 16 days compared to SENTINEL-2 every 5 days and MODIS every 1–2 days) (Hansen, *et al*., 2017). Table 2.1 is a summary of the application of remote sensing in managing water resources.

Table 2.1 Application of remote sensing to manage water resources

Satellite images	**Specific application in manage water resource**
MODIS	- Sea surface salinity (Urquhart, *et al*., 2012) - Surface water turbidity (Constantin, *et al*., 2017) heavy metal pollution in river water
Landsat/Landsat -TM	- Sea surface salinity (Zhao, *et al*., 2017) - total suspended sediment concentration (Dorji and Fearns, 2017) - Water Quality Monitoring in Estuarine Waters (Lavery, *et al*., 1993)
SPOT	- Coastal Water Quality Mapping (Su, *et al*., 2008) - Identification of the Salinity in Gasikule Salt Lake (Wang, 2012)
Hyperion EO-1	- Monitoring water constituents and salinity variations of saltwater (Fang, et al., 2008)
Sentinel	- Observations of Human Impacts in Coastal Waters (Vanhellemont, et al., 2014)
Sea WiFS	- Salinity and Ocean Color (D'Sa, et al., 2002)

Remote sensing technique has been extended to application on monitoring surface water quality and quantity (Matsushita, *et al*., 2014; Jerry, *et al*., 2003), detecting water bodies (flooding) (Ma, *et al*., 2013) and assessing influences of water discharge when there is mixing of fresh water and salt water in coastal areas (Wang and Xu, 2008). Using remote sensing to detect salinity intrusion is also a interesting topic for many previous studies which did not have access to satellite data.

2.2 Landsat-8 processing and its application to managing water resources

The Landsat 8 has great potential for applications in many fields including agriculture; land cover, forest and agricultural land condition, disturbance and change; fresh and coastal water; and snow and ice cover (Roy, *et al*., 2014). Furthermore, advantages of the high spatial resolution L8/OLI data are clear and suitable to employ in coastal and estuarine waters (Vanhellemont and Ruddick, 2015). The Landsat-8 OLI and TIRS spectral bands are stored as geolocated 16-bit digital numbers in the same L1T file. The 100m TIRS bands are resampled by cubic convolution to 30 m and coregistered with the 30 m OLI spectral bands (Roy, *et al*., 2014). The table 2.2 provides information about the wavelength and resolution for 11 bands of Landsat-8. For every single Band have strengths and limitations for each field application following the studies of Acharya, *et al*. (2015).

- Band 1: Coastal and Aerosol studies
- Band 2: Bathymetric mapping, distinguishing soil from vegetation and deciduous from coniferous vegetation (dry season/wet season vegation)
- Band 3: Emphasizes peak vegetation
- Band 4: Discriminates vegetation slopes
- Band 5: Emphasizes biomass content and shorelines
- Band 6: Discriminates moisture content of soil and vegetation; penetrates thin clouds
- Band 7: Improved measurement of moisture content of soil and vegetation and thin cloud penetration
- Band 8: Sharper image definition
- Band 9: Improved detection of cirrus cloud contamination
- Band 10: Thermal mapping and estimated soil moisture
- Band 11: Improved thermal mapping and estimated soil moisture
- BQA: Quality assessments for every pixel in the scene

Table 2.2 Landsat 8 Operational Land Imager and Thermal Infrared Sensor Source Wavelengths (Usgs, 2015)

Bands	Wavelength (micrometers)	Resolution (meters)
Band 1 - Ultra Blue (coastal/aerosol)	0.43 - 0.45	30
Band 2 - Blue	0.45 - 0.51	30
Band 3 - Green	0.53 - 0.59	30
Band 4 - Red	0.64 - 0.67	30
Band 5 - Near Infrared (NIR)	0.85 - 0.88	30
Band 6 - Shortwave Infrared (SWIR) 1	1.57 - 1.65	30
Band 7 - Shortwave Infrared (SWIR) 2	2.11 - 2.29	30
Band 8 - Panchromatic	0.50 - 0.68	15
Band 9 - Cirrus	1.36 - 1.38	30
Band 10 - Thermal Infrared (TIRS) 1	10.60 - 11.19	100 × (30)
Band 11 - Thermal Infrared (TIRS) 2	11.50 - 12.51	100 × (30)

Compisites of bands are a good method to show general characterestics of one objective. Different combination of bands can lead to dissimilar visual impressions of what is there relative to the physical properties of the wavelength, Geospatial Innovation Facility (2008) give a common combination of Landsat EMT.

- Bands 3,2,1 R-G-B: This color composite is close to true color which is also useful for studying aquatic habitats.

- Bands 4,3,2, NIR-R-G: It shows similar qualities to the image with bands 3,2,1 however, since this includes the near infrared channel (Band 4) land water boundaries are clearer and different types of vegetation are more apparent.
- Bands 4,5,3 NIR-SWIR1-R: Different vegetation types can be more clearly defined and the land/water interface is very clear.
- Bands 7,4,2 SWIR-NIR-G: This has similar properties to the 4,5,3 Band combination with the biggest difference being that vegetation appears as green.
- Bands 5,4,1 SWIR1-NIR-B:This band combination has similar characteristics to the Band 7,4,2 combination, however it is better suited in visualizing agricultural vegetation.

The reflectance wavelength of water is more obviousin the Visible and NIR bands; the refectance value of clear water is quite low (Figure 2.1). However, reflectance from water containing other components such as sediment will be higher, in wavelength reflectance ofseawater ranges from 0.01 to 0.14 or 1 to 14% (Xiong, *et al*., 2012).

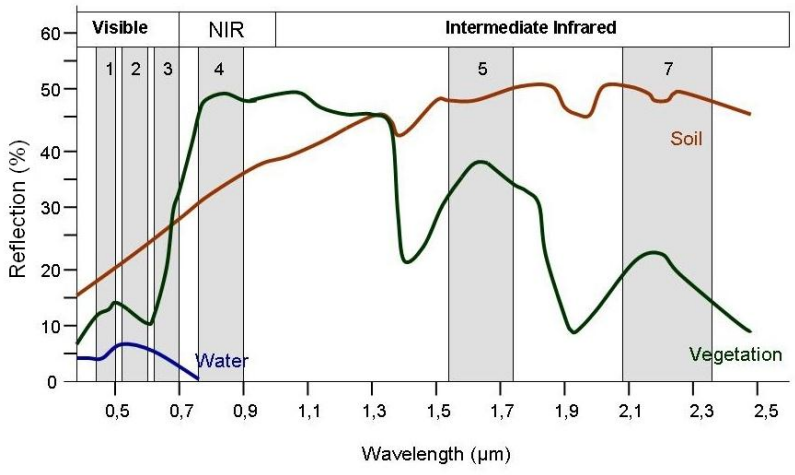


Figure 2.1 Spectral signatures of soil, vegetation and water, compared to the spectral bands of LANDSAT 7 Siegmund, Menz (2005).
(Source: http://www.seos-project.eu/modules/remotesensing/remotesensing-c01-p05.html, 15 May 2016).

2.3 Processing Landsat -8 OLI images

2.3.1 Atmosphere correction

The atmospheric effects might be the cause of inaccuracies interfering with reliable spectral information, it reduces reflectance from the target objects resulting in under or overestimation of remote sensing reflectance (Bernardo, *et al*., 2017). The atmosphere correction routines find the darkest pixel in a spectral band with a non-zero top-of-atmosphere reflectance which should have a high probability of showing a zero reflectance, however, due to atmospheric scattering, some energy is attributed to that pixel (Bernardo, *et al*., 2017). The Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes (FLAASH) (ENVI, 2009) using Metadata file as a input can be used to correct for atmosphere effects. This Metadata file contains much important information for correction of the atmosphere effects (Usgs, 2015) including unique Landsat scene identifier, WRS path and row information, Scene Center Time of the date the image was acquired, Corner longitude and latitude in degrees and map projection values in meters, reflective, thermal, and panchromatic band lines and samples. File names contain image attributes including cloud cover, sun azimuth and elevation, and number of GCPs used. The process for atmosphere correlation includes two steps: convert DN to radiance and radiance to reflectance.

Firstly, converting DN to Radiance based on the formula (1)

$$L\lambda = ML \times Qcal + AL \quad (1)$$

where:

$L\lambda$: Spectral radiance (W/(m2 * sr * μm))

ML: Radiance multiplicative scaling factor for the band (RADIANCE_MULT_BAND_n from the metadata)

AL: Radiance additive scaling factor for the band (RADIANCE_ADD_BAND_n from the metadata).

$Qcal$: L1 pixel value in DN

Secondly, the OLI Top of Atmosphere Reflectance, using the radiance to convert into reflectance.

$$\rho\lambda' = M\rho * Qcal + A\rho \quad (2)$$

where:

$\rho\lambda'$: TOA Planetary Spectral Reflectance, without correction for solar angle (Unit less)

$M\rho$: Reflectance multiplicative scaling factor for the band (REFLECTANCEW_MULT_BAND_n from the metadata).

$A\rho$: Reflectance additive scaling factor for the band (REFLECTANCE_ADD_BAND_N from the metadata).

$Qcal$: L1 pixel value in DN

The output of the FLASH model was the reflectance and the value of reflectance at a given wavelength must be 0-1. On the other hand, the FLASH should determine some parameters to consist with particular study area which includes the column water vapor amount for each pixel in the image;

- 1050-1210 nm (for the 1135 nm water feature)
- 870-1020 nm (for the 940 nm water feature)
- 770-870 nm (for the 820 nm water feature)

The Initial Visibility field, an estimate of the scene visibility value is assumed for the atmospheric correction if the aerosol is not being retrieved. The Table 2.3 showed the scene visibility depending on the weather condition.

Table 2.3 Lists approximate values based on weather conditions (Source: ENVI 2009).

Weather Condition	Scene Visibility
Clear	40 to 100 km
Moderate Haze	20 to 30 km
Thick Haze	15 km or less

Type of surface imagery also should be selected to optimize the effective of correction atmosphere impact. One of the most common surfaces in this study is a rural scene which presents aerosols in areas not strongly affected by urban or industrial sources. The particle sizes are a blend of two distributions, one of large particles and one of small particles.

2.3.2 Remove cloudiness by using F-mask function

Clouds and cloud shadows are significant factors that add to the data noise in the Landsat images (Zhu and Woodcock, 2012). The F-mask algorithm effectively finds clouds and cloud shadows by using the F-mask- function with the Matlab 1.6.0 version (Zhu and Woodcock, 2012). Detecting cloud and cloudy shadows is more of a challenge in water scenes due to water being generally dark. The existence of clouds overwater can increase Band 5 reflectance greatly. The F-mask uses the normalized Band 5 reflectance to calculate the brightness probability for cloud detection overwater (Zhu and Woodcock, 2012).

Meanwhile, the Cirrus routine allows for better detection of cirrus cloud contamination in each scene for Landsat-8 (Acharya, *et al*., 2015). The cloud shadows are easily confused with a body of water or wetland area because of the reduced solar irradiance or low reflectance in the cloud-shaded area (Li, *et al*., 2013).

The input for the F-mask included Top of Atmosphere (TOA) reflectance for Bands 1, 2, 3, 4, 5, 7 and Band 6 Brightness Temperature (BT) from Landsat Bands 4–7 (Zhu and Woodcock, 2012). For Landsat 8, almost all the F-mask algorithm components are the same as for Landsat Bands 4–7, except a new cirrus cloud probability is calculated using the new cirrus band which improves detection of thin cirrus clouds (Zhu, *et al*., 2015).

2.4 Machine learning

Machine learning can be used to identify and exploit patterns in data combining computer science with mathematics and statistics (Stephen, 2007). It includes supervised learning where the training data consist of pairs of input data and target (desired) outputs, while on the other is unsupervised learning, where there is no target output provided (Stephen, 2007).

2.4.1 Multiple Linear regression

Linear and multiple regression are the most commonly applied statistical models (Owen, 2001). The multiple linear regression is simple and often provide an adequate and interpretable description of how the inputs affect the output (Hastie, *et al*., 2009).

2.4.2 Decision Trees

The Decision Trees identification starts from the entire set of available training samples (root node), recursive binary partition is performed for each node until no further split is possible or a certain terminating criteria is satisfied. The output is a tree diagram with the branches determined by the splitting rules and a series of terminal nodes that contain the mean response (Basu and Basu, 2011). Tree size is a tuning parameter governing the model's complexity, and the optimal tree size should be adaptively chosen from the data (Hastie, *et al*., 2009). A smaller tree with a fewer number of splits eases interpretation at the cost of a small bias; too large a tree may lead to overfitting (Basu and Basu, 2011).

2.4.3 Random Forest

The regression RF is used by growing trees, contingently on a random vector that the tree predictor takes on numerical values as opposed to class labels (Breiman, 2001). This randomness comes from randomly choosing a group of mtry predictors to split on at each node and bootstrapping a sample from the training set.

The non-selected cases are called out-of-bag (Epifanio, 2017). When the number of variables is large, but the fraction of relevant variables is small, random forests are likely to perform poorly (Hastie, *et al*., 2009).

2.4.4 Neural network

The Marquardt algorithm for nonlinear least squares presented and incorporated into the backpropagation algorithm can be used for training feed forward neural networks; this algorithm is efficient due to the network contains no more than a few hundred weights (Training Feedforward Networks with the Marquardt Algorithm). The log-sigmoid transfer function was selected which is perfect for learning to output Boolean values with its output range (0 to 1) and so is excellent for data with a range of 0 to 1

The Network has to define parameters before training that include the number of Inputs and Layers; bias connections; input, Layer Weight connections; and output connections. Selecting the training algorithm is very important to limit early stop which may reduce the effectiveness of the modeling. The Bayesian Regularization is the better method to avoid early stop issues. The Levenberg–Marquardt is suitable for training small- and medium-sized networks and patterns. The Training of Neural Networks (‘neuralnet’- ANN),training of neural networks uses multi hidden layers.The Feed-Forward Neural Networks and Multinomial Log-Linear Models (‘nnet’ – ANT) software for feed-forward neural networks have a single hidden layer, and is suitable for multinomial log-linear models.

2.5 Resampling data using Bootstrap method and overfitting problem

Small sample sizes makes for challenges to any statistical analyses and decreases the predictive potential of models (Stockwell and Peterson, 2002 and McPherson, *et al*., 2004). This is also a cause of high dimensionality of the data, increasing bias and resulting in large standard errors of estimated parameters (de

Winter, *et al*., 2009). Thus, model performance should be improved by a bigger dataset (Ließ, *et al*., 2012).

The bootstrap is a resampling method which allows to make a statistical inferences from data without considering strong distributional assumptions about the original data or the statistic being calculated (Haukoos and Lewis, 2005 and Dixon, 2006). Although nonparametric efficiency measures are often criticized for lacking a statistical basis, in fact, nonparametric efficiency measures do have a statistical basis (Simar and Wilson, 1998). The bootstrap allows reusing of samples in training sets and other observations are not included in the sample and are employed to test as validation (Kelly, *et al*., 2011).

Bootstrap samples are created from the original data set by random sampling with replacement (Wehrens, *et al*., 2000). Each bootstrap sample should have the same sample size as the original data set to make sure the calculated estimation for the confidence intervalto keep away from bias issues (Haukoos and Lewis, 2005). The bootstrap datasets play a role as the training samples, while the original training sample works as the test sample, and both of them have observations in common (Hastie, *et al*., 2009). Small perturbations to the data-generating process produce a lot of changing in the sampling distribution, in such cases bootstrapping is not able to work well and can fail spectacularly (Shalizi, 2011). Nonparametric bootstrap re-select many observations, which may be the cause of inconsistent estimators (Dixon, 2006). In addition, increasing the number of bootstrap replicates is less effective for insufficient datasets (Dixon, 2006). Figure 2.2 describesre-sampling real data by using a nonparametric bootstrap function.

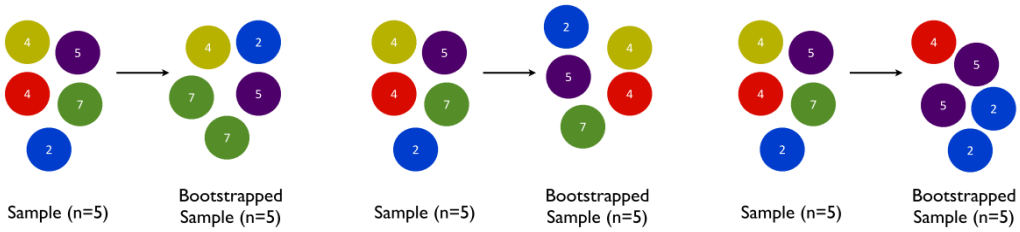


Figure 2.2 Example of bootstrap method: taking 3 bootstrapped samples of n=5 from the original sample (n=5). Notice that the middle bootstrapped sample reproduces the original sample exactly (Source: Ong, 2014).

Chapter 3

Methodology

3.1 Study area

The VMD, located in the downstream of the Mekong River, is a flat and low-lying area of highly complex rivers and channels (Noh, *et al*., 2013). The Tien River is one of the biggest distributary river systems in the VMD. Before entering the East Sea (South China Sea), the Tien River splits into four branches including, the Cua Tieu, Cua Dai, Ham Luong and Co Chien – Cung Hau (Figure 3.1). At a distance of 30 km from the South China Sea, the Co Chien River again splits into two estuary branches: Co Chien – Cung Hau (Nguyen and Savenije, 2006).

Annually, declining discharge from the upstream from the onset of the dry season, strong wind speed from the sea and high tide levels cause incursions of saltwater along the rivers up to about 50km inland. Furthermore, the proportion of discharge entering the delta river has significant roles to determine salinity distribution (both in terms of time and magnitude) in the different branches. In 2005, the discharge contribution at Cua Tieu was about 10% of total discharge in the Tien River while the discharge in Cua Dai and Ham Luong were about 20.8% and 10.2%, respectively. In the Co Chien River, the Co Chien estuary received 10.5% and Cung Hau received 3.3% and the remainder of the discharge was through other parts of the delta (Nguyen, 2008). Thus, about 55% of the total Mekong discharge occurs through the study area.

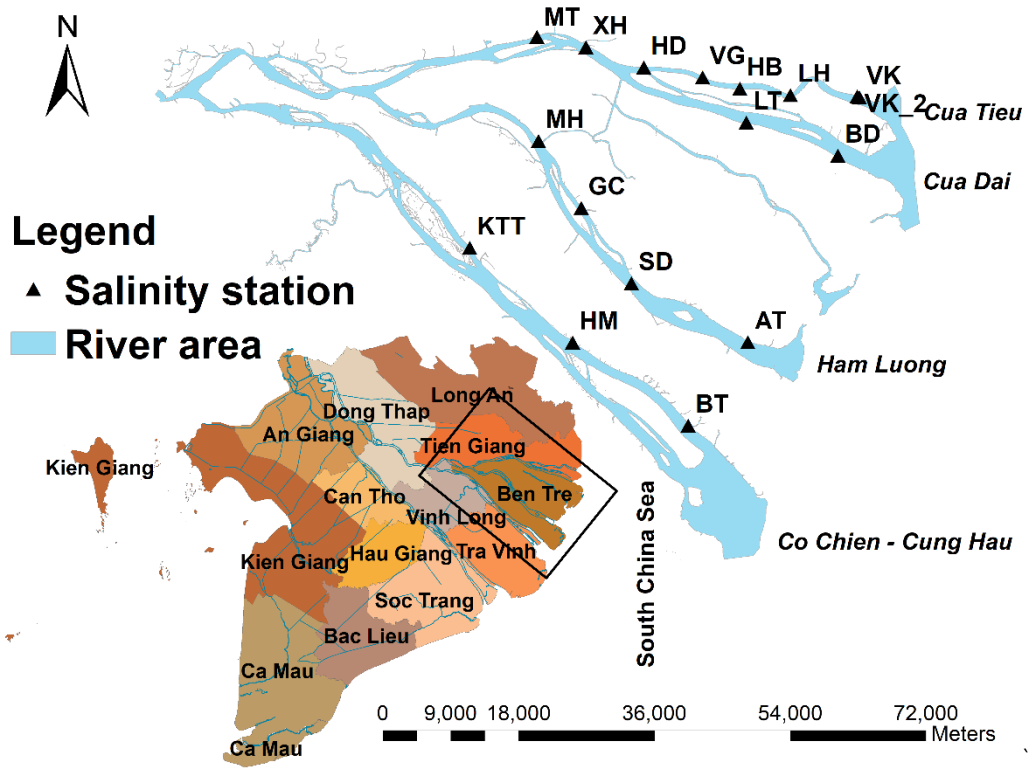


Figure 3.1 Sampling locations of fifteen salinity stations in the Tien River –and also downstream at Cua Dai estuary, Cua Tieu estuary, Ham Luong estuary and Co Chien – Cung Hau estuary.

3.2 Data collection

3.2.1 Landsat-8 collecting and processing

The Landsat-8 OLI has an entire Earth coverage every 16 days and spatial resolution (30 x 30 meters). 25 satellite images were used consisting of mostly cloud-free scenes in the dry season (January – June) from 2013 to 2016. The scenes of images were between path/row: 124/53 and 125/53. These images were used to extract reflectance data for a single pixel that also has data on salinity measured in the field at the time. “Ground truth” from fifteen salinity stations was applied to the Region of Interest (ROI) function in ENVI to identify 103 samples from these images to

prepare these samples for developing models as well as images for mapping salinity intrusion for the whole study area.

Firstly, the radiometric correction was performed to normalize satellite images for factors such as sensor degradation, Earth-Sun distance variation, incidence angle, view angle, and time of data gathering were applied to the image data. This process involved converting Digital Number (DN) into radiance using calibration parameters that accompanied with the images metadata. The image was then atmospherically corrected and transformed to the reflectance using the MOD-TRAN-based FLAASH (Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes) algorithm under ENVI 5.3 program. Secondly, a Function of mask (Fmask) algorithm was applied to detect clouds and cloud shadows to correct the cloudy maps. Finally, the mask of the study area prepared for every single satellite image from the first step was created by overlapping, the river network map and cloudy maps. The completed reflectance-image prepared as inputs was clear of cloud, cloud shadows and focused on the water area.

Depending on each object (water, soil), the spectral reflectance has specified wavelengths but the range should be scaled 0-1(Peddle, *et al*., 2001). Normally, the spectral reflectance of water is quite low. The wavelength reflectance of seawater was small between 400 and 850 nm, ranging from 0.01-0.14 (Xiong, *et al*., 2012). In this study, mean reflectance of water was around 0.03- 0.13 (Figure 3.2).

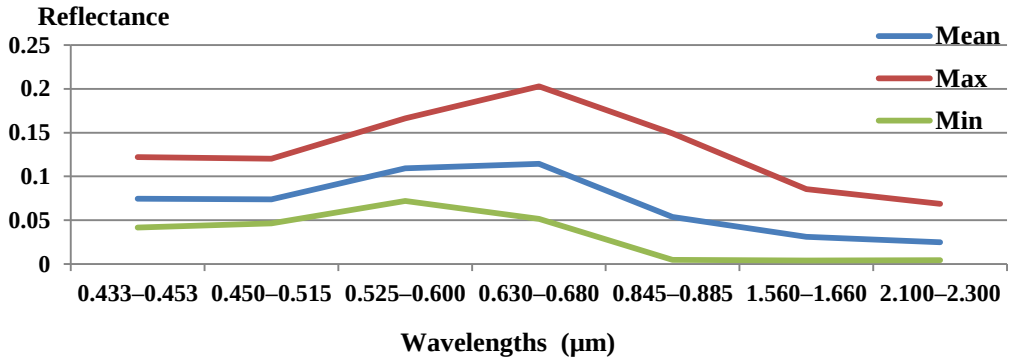


Figure 3.2 Reflectance *vs.* wavelength of water in the river system.

- **Detecting water area from Landsat -8**

Mapping of the area was provided by the government office to extract river area: but this map sometimes did not match with the real map from the satellite image. That may cause big error issue when the model is applied to the study area. On another hand, the study area was changed following the time-frame of this study due to human activities. Thus, water area should combine both available map and imagery.

3.2.2 Salinity measurement

Saltwater data was collected at the same time with satellite imagery at 12 hydrology stations (Figure 3.1): Vam Kenh1 (VK1), Long Hai (LH), Hoa Binh (HB), Hoa Dinh (HD), My Thoa (MT), Binh Dai (BD), Loc Thuan (LT), An Thuan (AT), Son Doc (SD), Giong Chom (GC), My Hoa (MH), Ben Trai (BT), Huong My (HM) and Khanh Thanh Tan (KTT) and 3 sluice gates: Vam Kenh 2 (VK2), Vam Giong (VG) and Xuan Hoa (XH) along the rivers: Tieu River, Dai River, Ham Luong River and Co Chien River. Saltwater data was provided by the Tien Giang and Ben Tre Hydrology and Metrology Centres and Department of Agriculture and Rural Development Tien Giang in the dry season (January to June) from 2013 to 2016. The hydrology stations measured salinity by using the Electrical Conductivity (EC) YSI (Figure 3.3); along the transects measured with saltwater samples were collected by mixing water at 0.5 meter depth and surface water for every two hours.

Figure 3.3 YSI Model Y30 SCT to measure salinity.

3.3 Model development

To assess effective band compositions of the seven bands used in the present study, multi-linear regression was employed. The Multi-stepwise Regression (MLR), the Decision Trees (DTs) Random Forest (RF) and Neural Network model (ANN and ANT) were applied to develop the relationship between salinity and reflectance from Landsat-8 OLI by using R software (version 3.3.2). The best model was used to predict salinity intrusion for the whole study area by using R software version 3.3.2 and ArcGIS (version 9.3) on 24th January 2015 and 9th February 2015. Moreover, this study used the step-wise model to develop reflectance–locations for predicting based on the real data.The Figure 3.4 summarizes the data processing development and evaluate models.

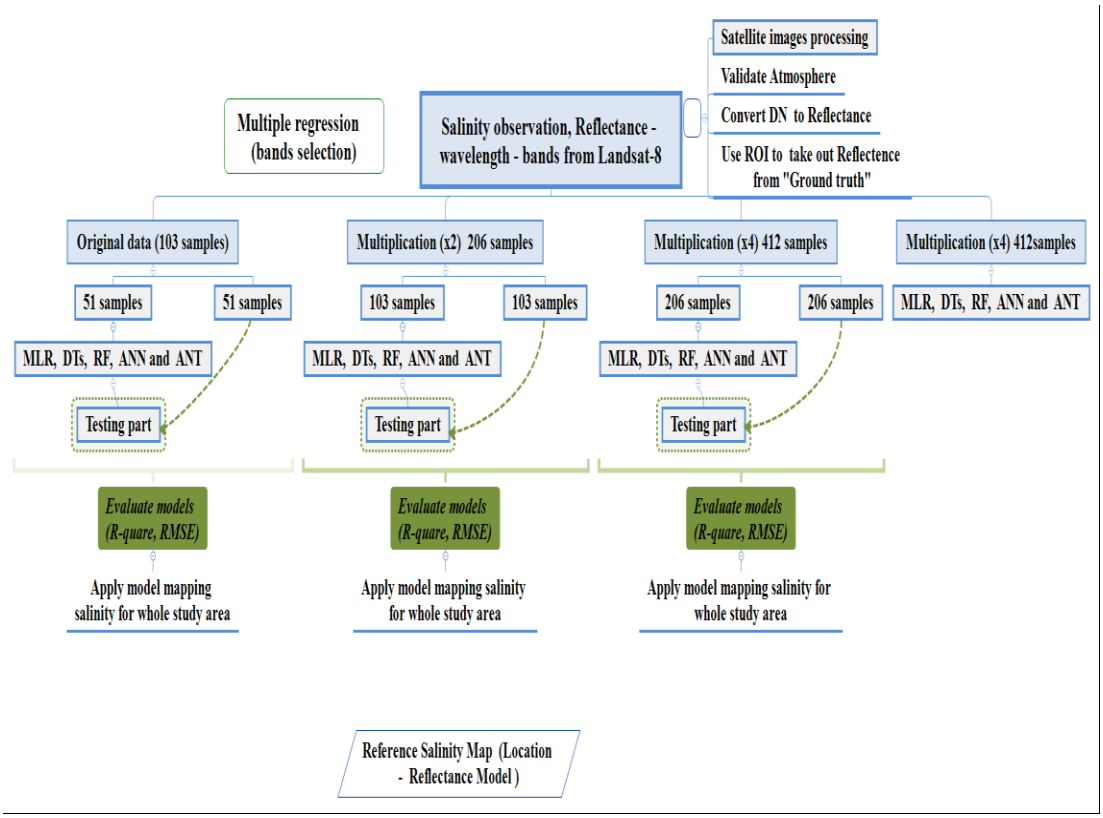


Figure 3.4 Diagram summarizing the whole process for developing the salinity model.

3.3.1 Preparing reflectance wavelength - bands data for modeling salinity intrusion

Selecting suitable bands for model development, reflectance – wavelength data from 7 bands were considered to investigate the correlation of both single bands and multiple bands with salinity observation. The MLR was employed to analyze these relationships. The p-value and R-square (R^2) were factors to assess which were the effective bands to use as an input to apply the best statistical model.

3.3.2 Modeling salinity intrusion with reflectance wavelength in the real samples

Three models: the MLR, the DTs and the RF were used to analyze the real samples. The input for these model was prepared as 103 samples as a real data set and then divided (randomly) into 50:50 for training and test purposes.

3.3.3 Resampling by bootstrap data

Bootstrapping requires repeated random sampling from the data set and estimating the indirect effect in each resampled data set (Preacher and Hayes, 2008). The dependent and independent variables are kept together in pairs (Mackinnon, 2009). From the 103 samples bootstrapping was executed for resampling two times and four times by the bootstrap function in R and divided into 50:50 for training and test purposes.

3.3.4 Modeling salinity intrusion with reflectance wavelength using bootstrap data

The bootstrap data was divided 50% for training and 50% test to prepare as the input for Modeling salinity intrusion by using MLR, DTs, ANN, ANT and RF. These models were employed for choosing the best salinity model based on reflectance-wavelength from the Lansat-8 OLI data and to identify the influence of sample size on the effective for developing salinity model.

- ***Multiple Linear Regression***

The MLR is a highly flexible tool for examining the relationship of independent variables and single dependent variable. The function “lm” was used to perform by MLR in R.

- ***Decision Trees***

The DTs wasbased on nodes and splits; the full sample is known as the root node. At each instance of split, a variable and its level is selected, so that purity at each child node is the highest possible at that level (Lenguajes, 2011). The DTs was integrated in R software by recursive partitioning (Rpart).

- ***Random Forest***

The RF regression algorithm is an ensemble-learning algorithm that combines a large set of regression trees. It was an integrated package ‘randomForest’ in the R program; two parameters need to be optimized in the RF: the number of regression trees (ntree); The RF trees are not prone to over fit, this study chose a high number of 1000 trees and the number of input variables per node (*mtry*); default value is 1/3 of the total number of variables as the study of Wang, *et al*., 2016.

- ***Neural network***

The ANN is often a good fit for identifying complex patterns even though the resulting network of a black box solutionare very difficult to interpret (Vabuena, 2016). The neural network has three basic components: the inputs, the hidden layer, and the outputs. The ANN uses multi-layer connections with the weight and activation function to predict a nonlinear problem (Niu, 2017).

In this study, the library (neuralnet): multinomial log-linear for training of neural networks and the Library (nnet): feed-forward neural networks with a single hidden layer were employed.

3.3.5 Modeling salinity intrusion with reflectance wavelength and location data

The Stepwise Multiple Linear regression was employed for predicting salinity by using reflectance data at different wavelength bands of Landsat-8 OLI and the location (latitude and longitude) of *in situ* salinity measurements.

The Root-mean-square error (RMSE), R-square (R^2) and P-value were used to evaluate the goodness of fit and signification of these models.

Chapter 4

Results and Discussion

4.1 Results

4.1.1 The relationship between reflectance and salinity

The composite spectral of 7 bands including Band 1, Band 2, Band 3, Band 4, Band 5, Band 6 and Band 7 provided high effective prediction significance for the salinity model (Figure 4.1). Besides, the composition of Band 2, 3, 4 and 7 also indicated high significance with p-value (approximately, 3×10-09) which was greater than the combination of 7 bands. Use of single bands was unlikely to be productive and only Band 3 presented high significance to salinity. Therefore, four bands (Band 2, 3, 4 and 7) were used as inputs for developing the salinity intrusion model.

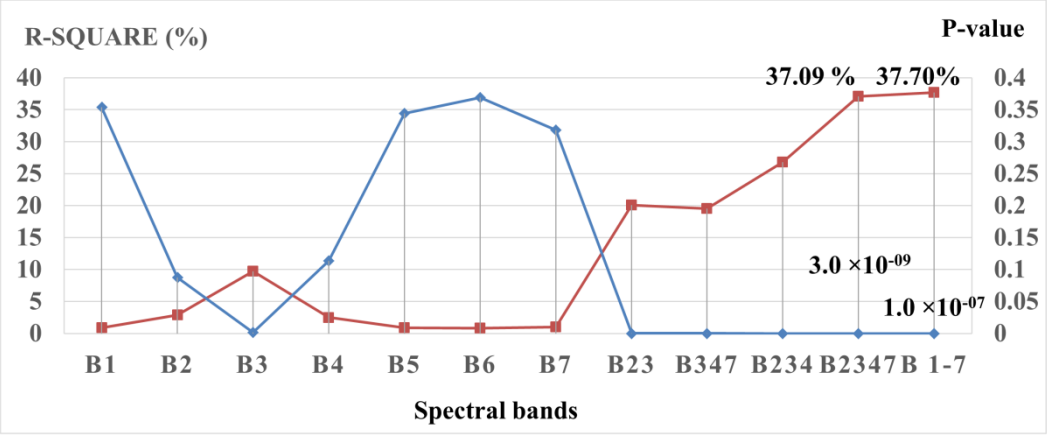
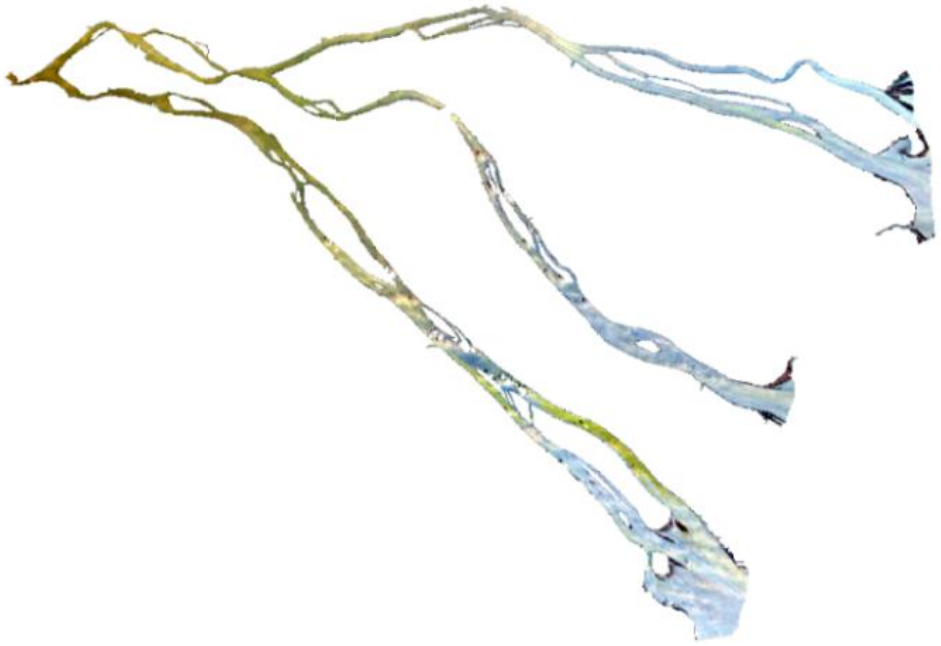


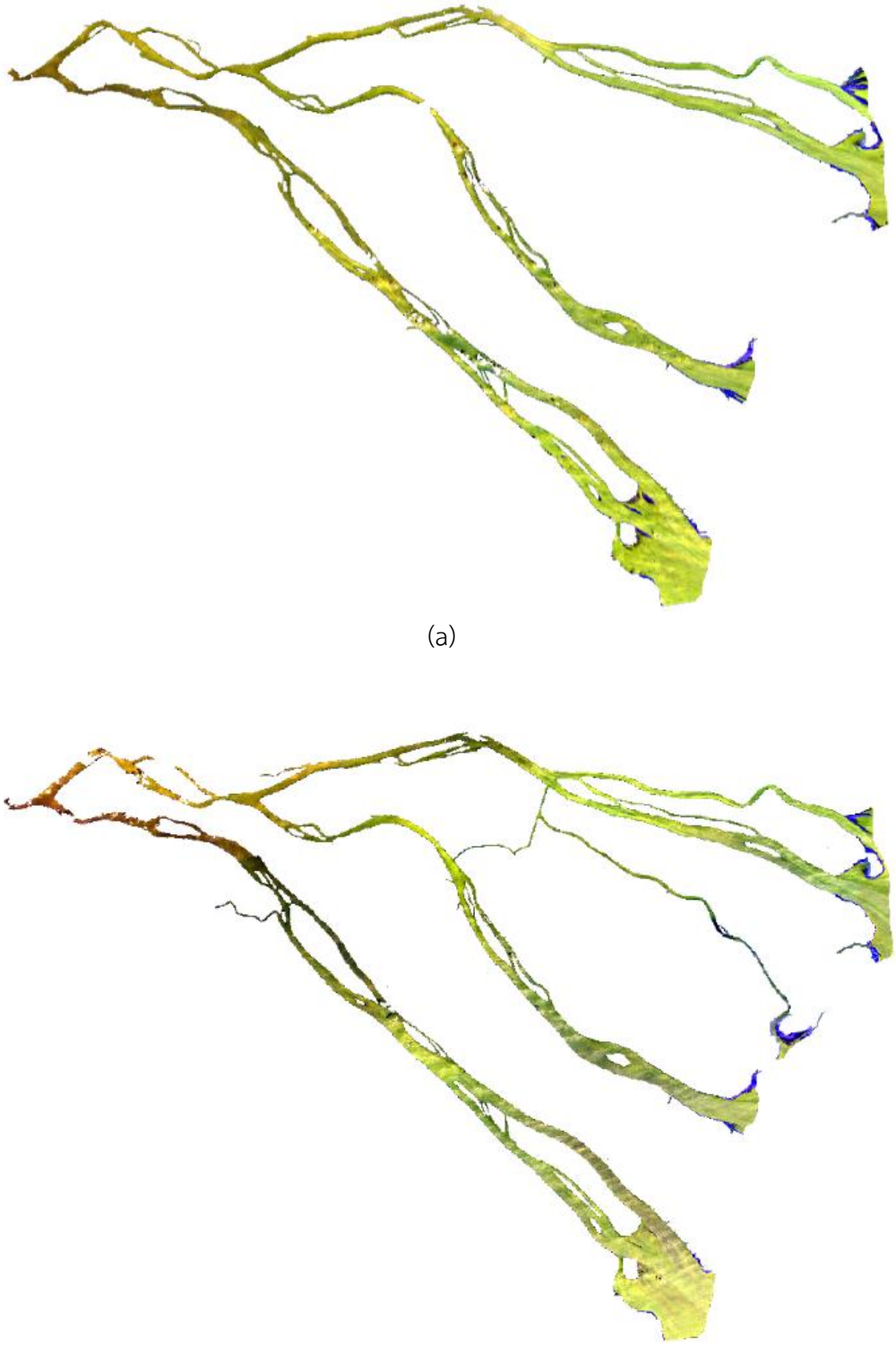
Figure 4.1 Multiple linear analyzed the relationship between salinity and composite bands. Red is P-value, blue is R-square.



(a)

(b)

Figure 4.2 Composite of bands 2, 3, and 4 on 24th Jan 2015 (a) and 09th Feb 2015 (b).



(a)

(b)

Figure 4.3 Composite of bands 2, 3 and 7 on 09th Feb 2015 (a) and 24th Jan 2015 (b).

The Figure 4.2 showed that a composite of band 2, 3 and 4 on of surface water were showed there were separate salinity regimes between upstream and downstream of rivers. It presented strong mixing color on 09th Feb 2015 (Figure 4.2). Besides, the composite of band 2, 3 and 7 showed a clear pattern for the whole area (Figure 4.3).

4.1.2 Modeling salinity intrusion base on the real samples size

The DTs model indicated low correlation in both training and test; the R-square values were 25.4% and 13.09% while the RMSE values were 2.72 and 2.83. Predicted salinity values were around 5-15 ppt in the training and were similar to the test. While the MLR/ LM model presented fairly good R-square in training when the R-square and RMSE were 50.23% and 4.15, respectively but the test showed a poor correlation: the R-square was 22.61% and RMSE was 4.12. The range of predicted salinity value was 0-20 ppt in training, it was higher in the test (Figure 4.4). However, RMSE of the DTs model was better than those of the MLR model for all processing; and the P-value was 9.0×10^{-05} in the DTs and 2.0×10^{-07} in the MLR.

Even though, the RF performed very well in the training but the test part of this model showed a poor correlation: low R-square and high RMSE. R-square was 90.41% and RMSE was 1.1 in training. However, the test figured out the limitation of the model; R-square was 22.55% and RMSE was 2.13. The range of predicted salinity was between 4 - 18 ppt in the training and 6 - 18 ppt in the test (Figure 4.5). Overall, the RF model was the best one to develop salinity intrusion model with the P-value of RF had strongly significance which was smaller than $p < 2 \times 10^{-16}$.

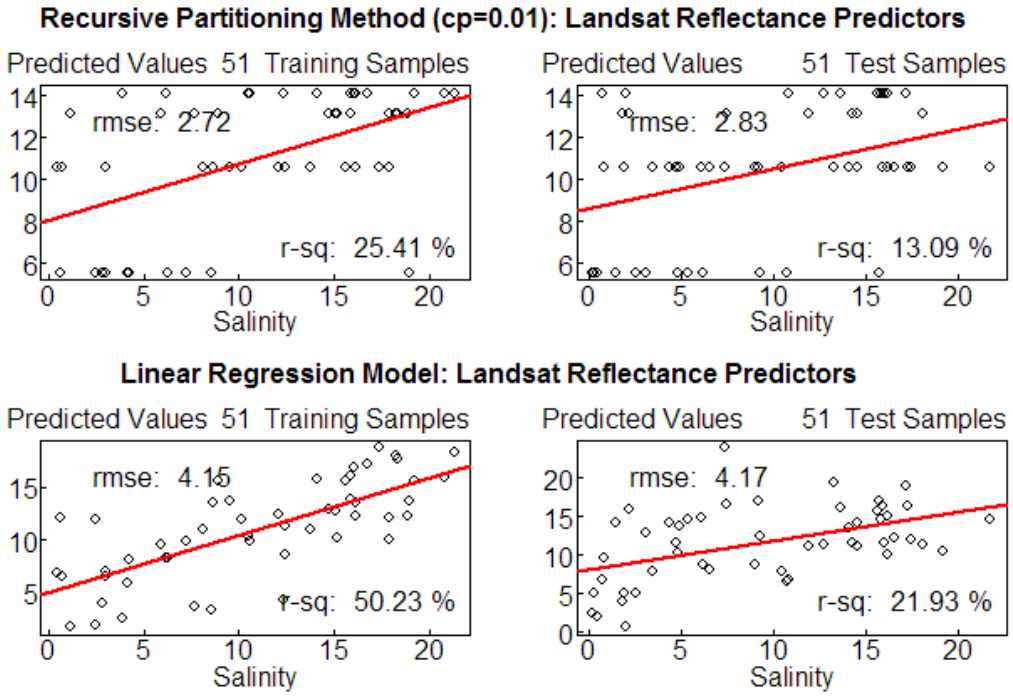


Figure 4.4 Regression of the DTs and the MLR based on the real data correlation in training and test.

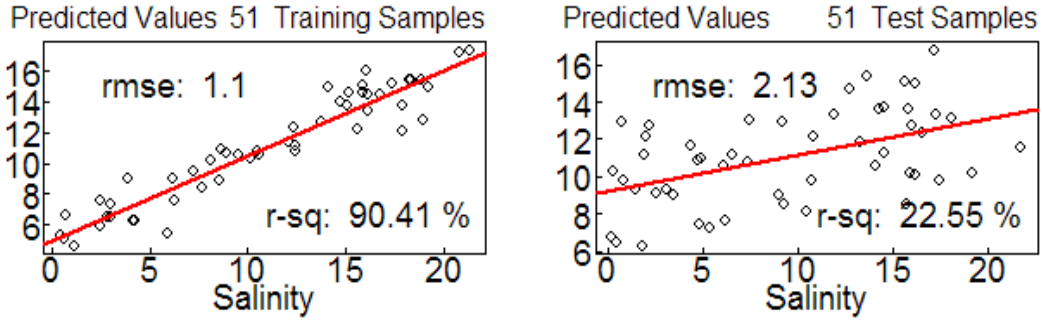


Figure 4.5 Regression of the RF based on the real data correlation in training and test.

4.1.3 Modeling salinity intrusion with bootstrap data

The real samples size was n = 103 re-sampling was taken from the 103 real observations to obtain 412 bootstrap-samples. The DTs and MLR regression was employed in the bootstrap data. The DTs showed better correction for both training and test. The R-square rose sharply for both training part which was smaller than 40% at 51 samples (real samples) but the model with the bootstrap was over 50%, the peak reached over 80% at 412 samples (Figure 4.6) while the RMSE was reduced slightly around 2-3 units. However, although the MLR model improved R-square, the R-square was nevertheless really low and the RMSE was still high for all these case samples sizes. Bootstrapping resulted in some improvement but the correlation was still low like in the case of using the real samples in the training, despite the fact that the test became better it was still not enough for the MLR model to significantly improve even with increased sample size (Figure 4.7). The MLR model has a simple structure and less dimensions than the others. Thus, in case where the relationship variables were complex, the MLR may lack of ability to be able to make a confidence prediction. While the DTs illustrated quite good correlation when R-square rose to almost peak maximization but the RMSE was too high to make useful predictions.

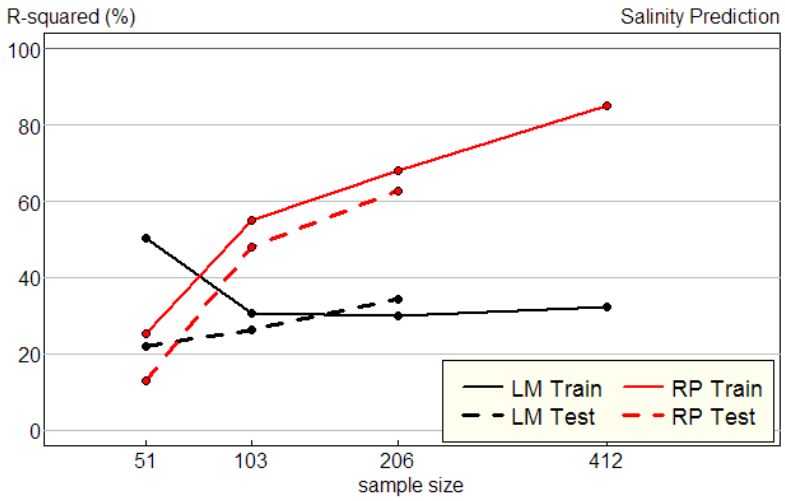


Figure 4.6 R-square from the Multi Linear regression (MLR/LM) and Decision Trees when the sample size was increased by bootstrap sampling.

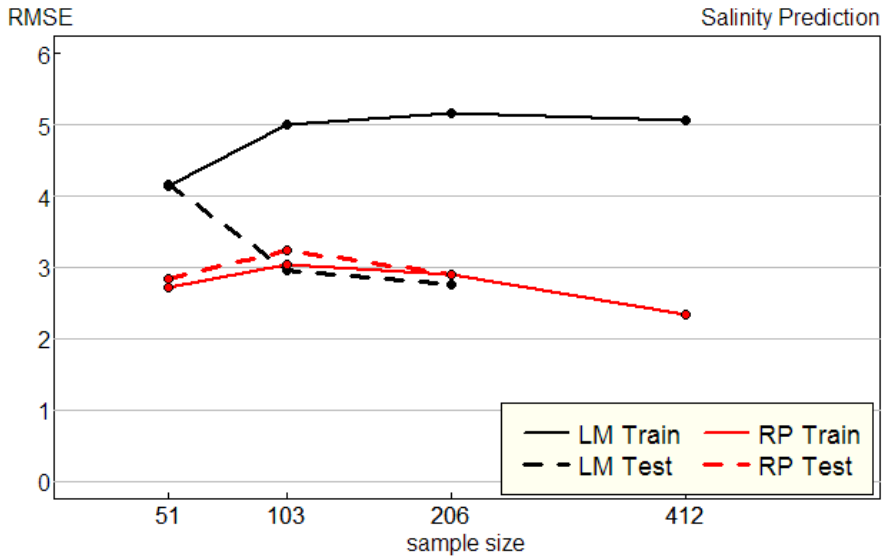


Figure 4.7 RMSE from the Multi Linear Regression and Decision Trees when changing the sample size.

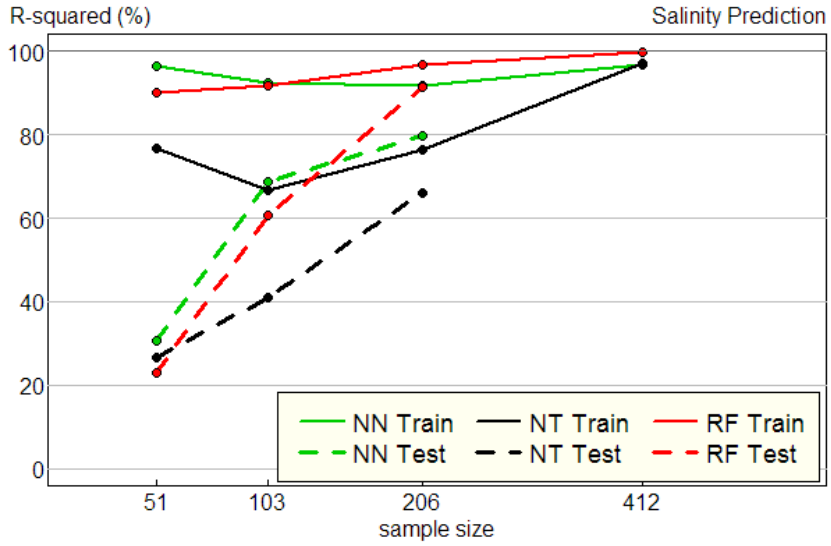


Figure 4.8 R-Square from the Neural Network (ANN and ANT) and Random Forest (RF) for different sample sizes.

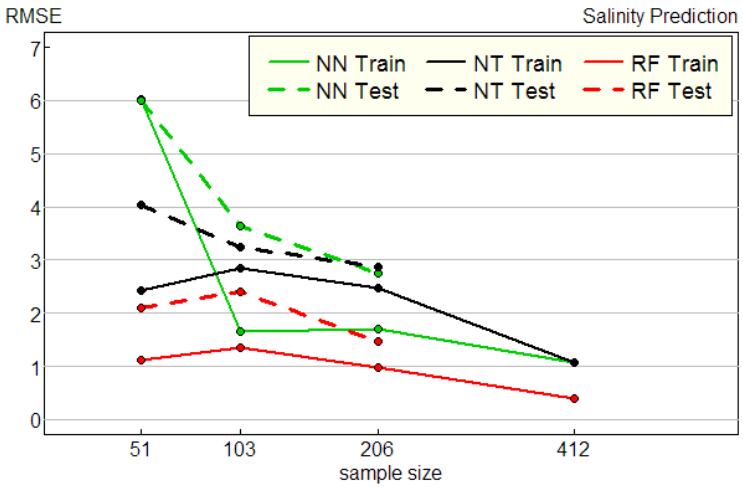


Figure 4.9 RMSE from the Neural Network (ANN and ANT) and the Random Forest for different sample sizes.

The ANN model showed that the RMSE was extraordinary high even though R-square was really low. On the other hand, the samples size is unlikely to have playeda clear role in changing the ANN model in the whole training part. The R-square was decreased in the training part, but the RMSE value clearly improved when comparedto the real data (51 samples) and the bootstrap data (103, 206 and 412 samples). Nevertheless, the ANN model had a significantly increased predicting ability by using bootstrap data in the test. The R-square and RMSE were more optimized with more observations; R-square was enhanced from 20% - 60% (Figure 4.8). In the case of the training step using 51 samples (the real data), the effective of R-square was strange due to the RMSE being very high (Figure 4.9). This may be caused by the many uncertain factors inside the black box of the ANN model which were integrated in the R program. One of the big limitation of ANN in this study when using R to run simulations is that it's hard to set up suitablefunctions for the hidden layers.

Increasing sample size also improved the performance of the NT model. R-square was maximized for the 412 bootstrap samples as well as the RMSE. However, ANT was less effective than ANN for all sample size cases based on R-square and RMSE.

The RF model was the best explanation for the influence of the sample size on the accuracy of the modeling when the sample size was increased from 51 samples (real data) to 103 samples, 206 samples and 412 samples; R-square got closer to 100% (Figure 4.8) and RMSE moved toward to 0 (Figure 4.9). This is what would be expected from a valid model and shows that the other models were less satisfactory.

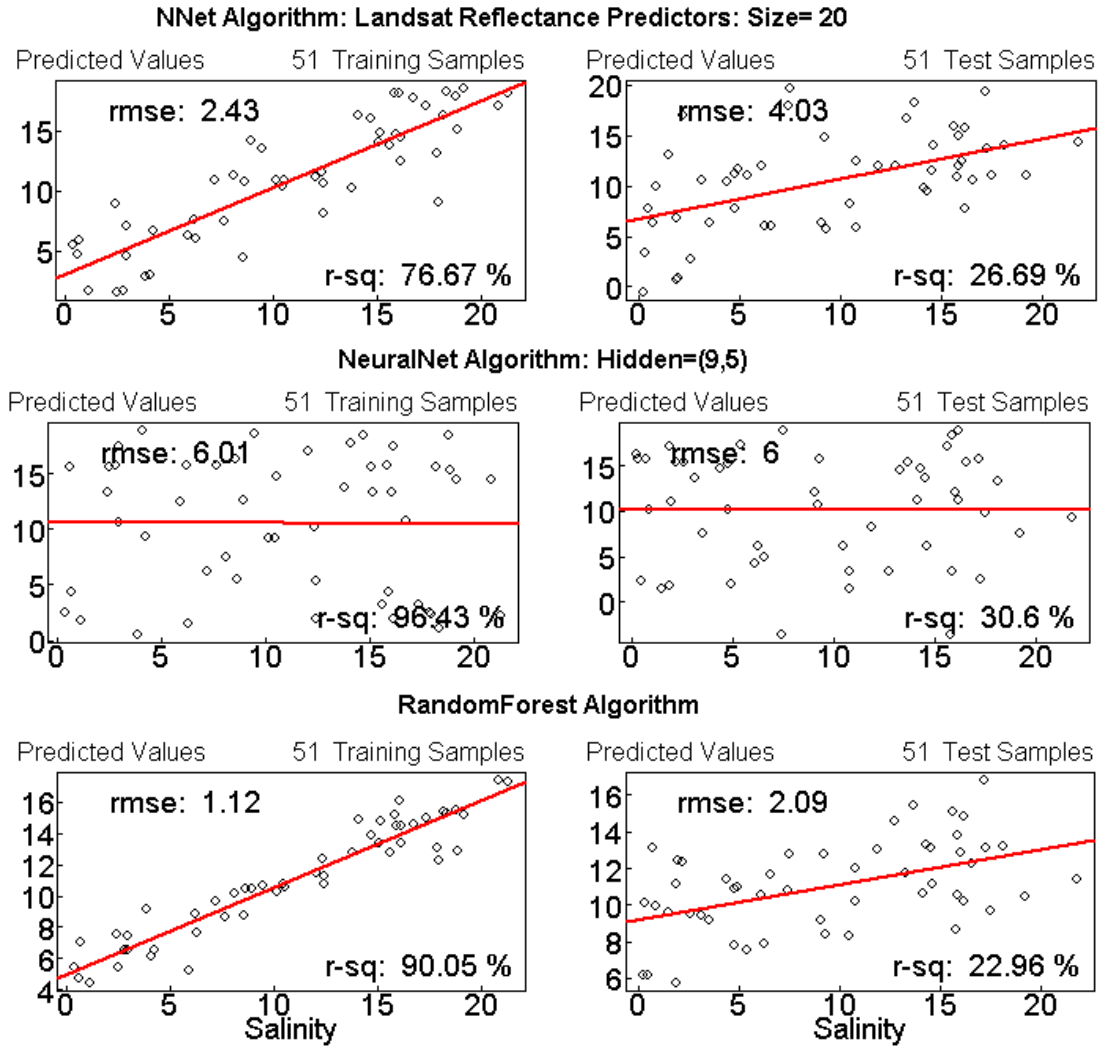


Figure 4.10 Regression model showing the relationship between observation and prediction of real data from ANN, ANT and RF.

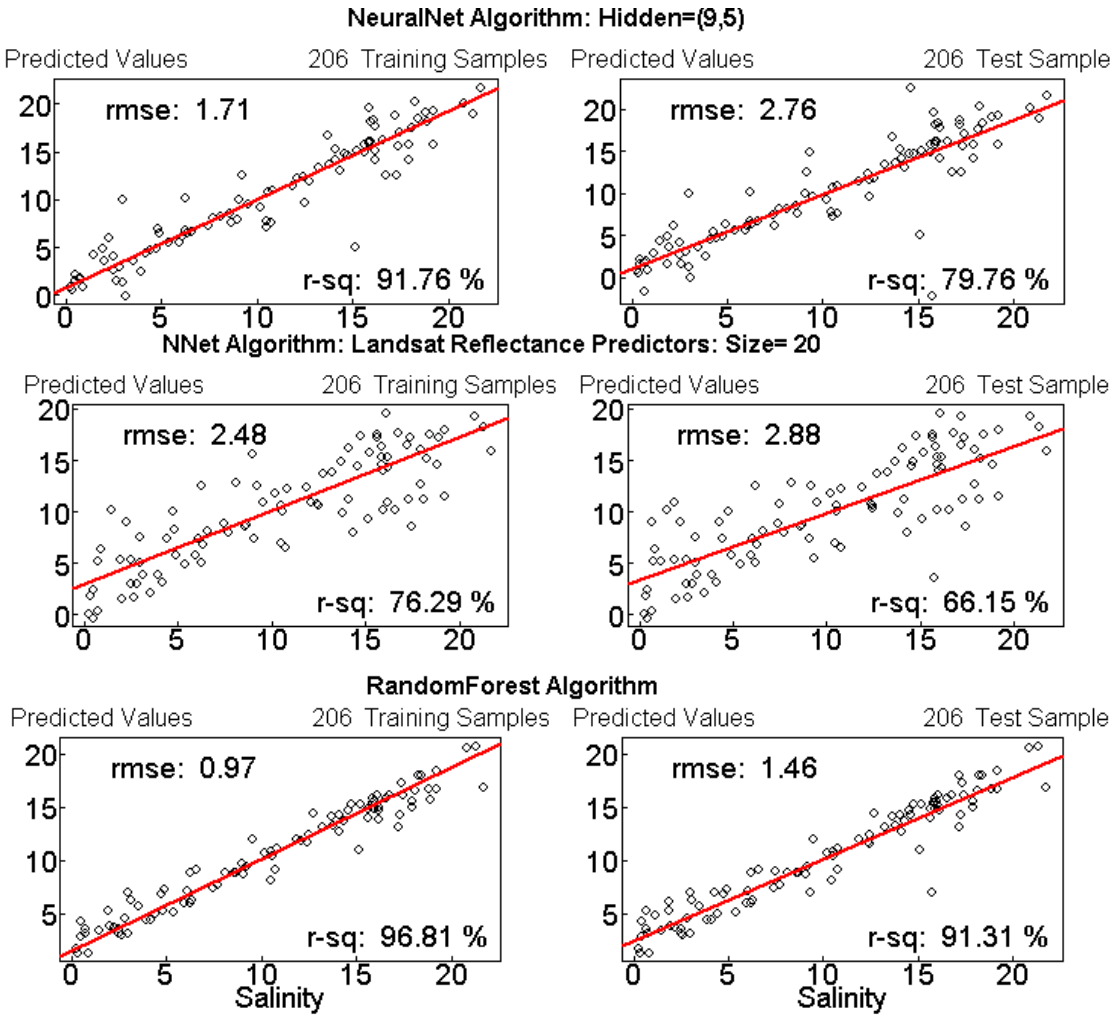


Figure 4.11 Regression model show the relationship between observation and prediction of bootstrapped data from ANN, ANT and RF.

Regressions of the ANN and the ANT were displayed in the real sample sizes in Figure 4.10 which clearly show that over-fitting was a major problem for the ANN and ANT models. All test parts were had poor correlations but the fit of the training parts was misleadingly good. Furthermore, the ANN model showed very strange regression when the R-square was really good in training (96.43%) but the RMSE was extremely high and the base line must have been wrong to show such a high of R-square. The predicted salinity was quite higher than others in the RF (3-16). In practice, the ANN model would therefore be likely to give too many false salinity alerts if the model was not good training enough to find out the best fit model. This problem arose from ANN giving a different result each time it was run. When, the RF

consistently gave similar results each time it was run ANN was pretty much different for each time running model indicating poor predictive value. The conclusion drawn from Figure 4.10 was that the data set was too small to avoid overfitting.

The overfitting problem that was faced on the real sample sizes (Figure 4.10) was solved using bootstrapped data sets (Figure 4.11). ANN, NNT and RF then demonstrated good correlations between salinity and reflectance when set up with bootstrap data. ANN and ANT indicated both R-square and RMSE were strongly significant. However, the RF in the bootstrap data was the best performing. The regression of the RF based on bootstrap data illustrated a good correlation for both training and test which R-square were 96.81%; 91.31% and RMSE were 0.97; 1.46 (Figure 4.11). The salinity concentration was around 0 - 20 ppt on the bootstrap model that was higher than the real data (2 - 16 ppt). Indeed, combining random features with boosting, for the larger data sets, it seems that significantly lower error rates are possible (Breiman, 2001). Thus, the RF model was indicated to be the best model for mapping salinity intrusion.

4.1.4 Comparison ANN by using Matlab analysis in the real samples size

The ANN performing in the previous part demonstrated a good relationship in the real data. However, it was strange that when training R-square showed high correlation but the RMSE was extremely bad (Figure 4.10 - Neural Net model). On the other hand, in the running of ANN the many times the fitting is performed to make sure the best number of hidden layers should be set up and the rate to divide observation in to the model should have achieved better results than actually found. The number of hidden layers was set at 9 and the percent of training-validation-test at 75% - 5% - 20% was found to be the best performance in training (R-square: 77.57% and RMSE: 2.85). The model was also faced with the over fitting which had a good fit in the training but the test did not work well (R-square: 30.63%; RMSE: 5.69 in test part) (Figure 4.12). On the other hand, in using ANN one should be careful to select the number of hidden layers as well as the balance divide between training-validation and test.

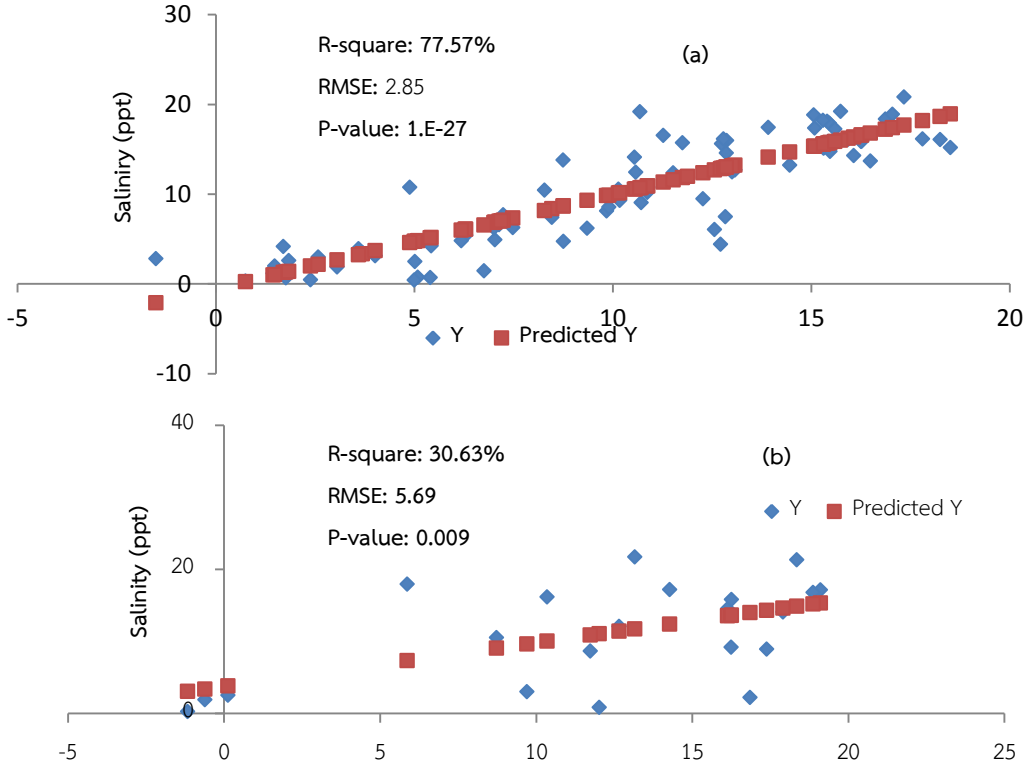


Figure 4.12 Regression of ANN in training (a) and test (b) using Bayesian Regularization.

4.1.5 Mapping salinity intrusion with real data and bootstrap data using Random Forest model

Mapping salinity was applied by using the RF model with the bootstrap data. Overall, mapping salinity clearly separated a trend of salinity in the whole study area. As the result, the predicted salinity level was 4.13-16.46 ppt (Mean: 12.42 and SD: 2.41). However, salinity concentration of the downstream was still too low at Cua Tieu estuary on 24th Jan 2015 (Figure 4.13). Meanwhile, on 09th Feb 2016, the salinity concentration was 3.06 - 16.33ppt (mean: 11.41; SD: 2.65). And at Co Chien - Cung Hau estuaries, salinity concentration was too low even though the location nearby the sea (Figure 4.14). Furthermore, the upstream area had a pretty high salinity, high noise and mix of salinity for both days. These anomalies may be because of lack of observations in the upstream.

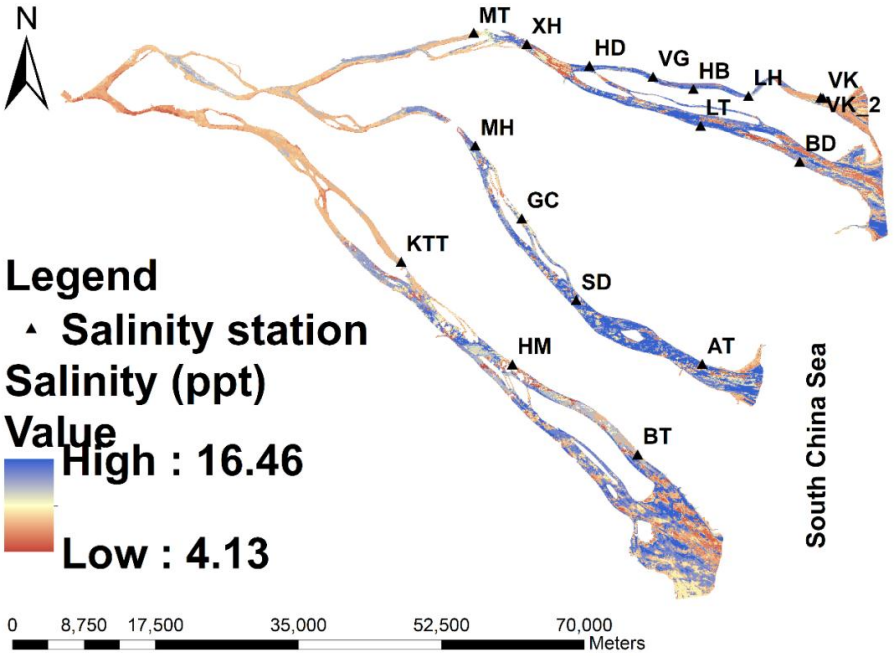


Figure 4.13 Salinity map from model based on bootstrap model Salinity -24 Jan 2015 (mean: 12.06; SD: 2.41).

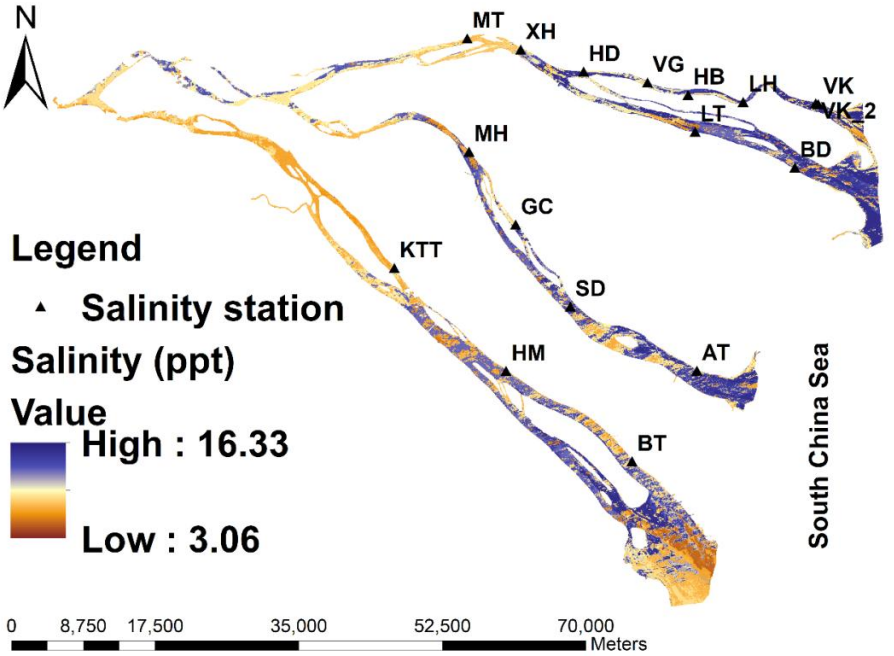


Figure 4.14 Salinity map from model based on bootstrap model Salinity - 9 Feb 2015 (mean: 11.41; SD: 2.65).

The comparison of salinity level on these days showed that mean of salinity concentration changed by nearly 1.1 ppt. Furthermore, saltwater area was more extended in Co Chien - Cung Hau and Cua Tieu. High saltwater moved inland until reaching the KTT and MT stations on 24th Jan while salinity was very low at KTT and XH (Figure 4.15).

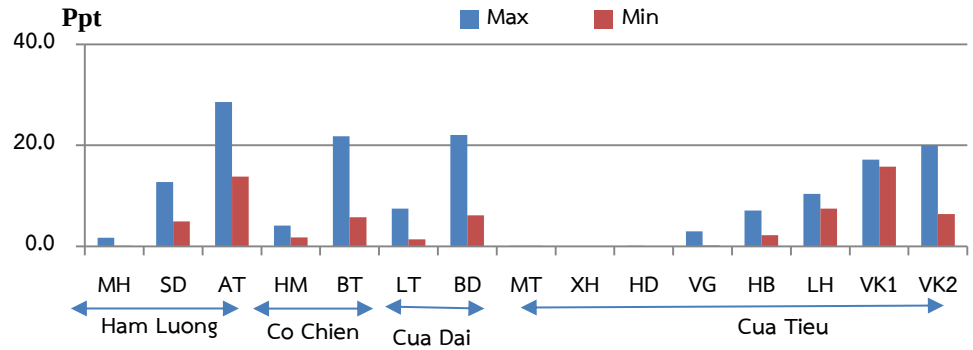


Figure 4.15 Observation of saltwater in Cua Tieu, Cua Dai, Ham Luong and Co Chien-Cung Hau River on 9th Feb 2015.

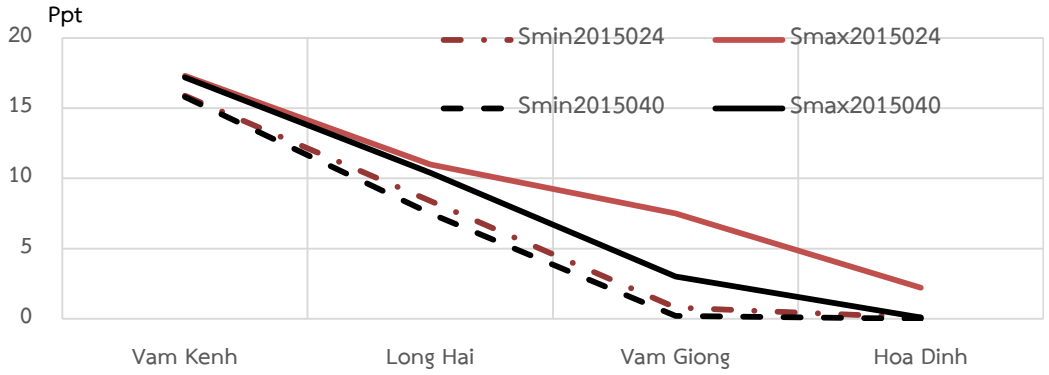


Figure 4.16 Observation of max and min saltwater in Cua Tieu River in 24th Jan 2015 and 09 th Feb 2015.

Figure 4.16 was based on salinity observations at fifteen saltwater stations on 09th Jan. Generally, salinity concentration reduced from downstream to upstream for every branch and salinity concentration on 24th Jan 2015. Salinity overall was higher than salinity on 09th Feb 2015. In Cua Tieu branch, salinity from the observations the maximum salinity reached at VK2 station was around 20 ppt

and minimum at XH around 0 - 0.1 ppt. In Cua Dai branch, maximum salinity reached at BD station was around 22 ppt and minimum at MT around 0.1 ppt. In Ham Luong branch, salinity from the observations reached a maximum salinity at AT station of around 28.6 ppt and minimum at MH of around 0.1 ppt with a maximum of 1.7 ppt. In Co Chien branch, salinity from the observations reached a maximum at BT station of around 21 ppt and minimum at HM station of around 1.8 ppt.

4.1.6 Reflectance - Geographic salinity modeling for mapping salinity

The location and reflectance also showed good relationship with salinity. Figure 4.17 presents a regression between predicted and observed salinity, R-square on training and test ordered 77.48%; 74.16% while RMSE were 2.57; 2.96. In addition, the extraordinary predicted values were smaller than 0 in both training and test which was exceptional in the salinity observations. Thus, locality issues are needed to verify the results in other study areas. That means a formula such as Equation (3) can only be used in this particular study area. If a similar model was used for a different area, the weight of the coefficients would have to be changed. Furthermore, the location played a strong factor in the formula which was unexpected in the performing model. Even though location is helpful in illustrating general trends it was also a cause of unrecovered essential information from reflectance wavelength data.

$$\text{Salinity} = (3.741\times10^{-04})\,x - (8.235\times10^{-05})\,y + (1.575\times10^{2})B_2 - (2.92\times10^{2})B_3 + 40.51B_4 + (1.048\times10^{2})B_7 - 1.366\times10^{2} \quad (3)$$

Where x is latitude; y is longitude; B2, B3, B4 and B7 are Landsat-8OLI spectral bands.

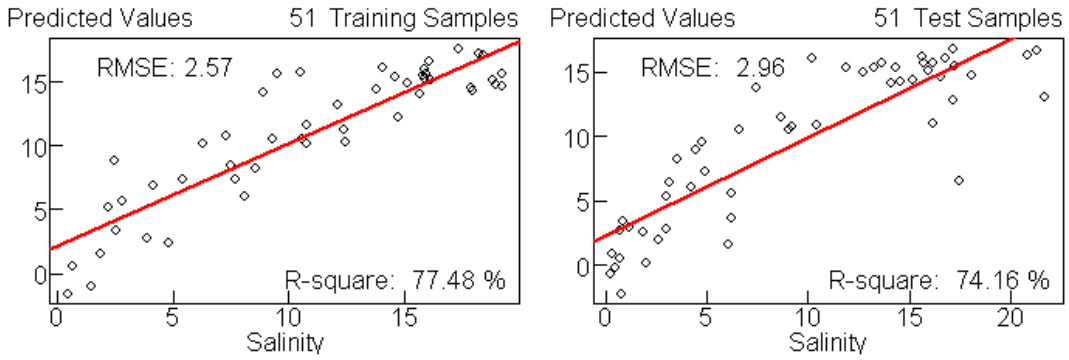


Figure 4.17 The regression between predicted salinity and observed salinity in training and test using reflectance - location.

Salinity level clearly decreased from estuaries to the upstream of the river. The Cua Tieu and Cua Dai rivers presented high salinity intrusion while the Co Chien - Cung Hau River has the lowest salinity concentration and least salt intrusion. The result demonstrated the changing of saltwater concentration on the two sampling days. The mean of salinity level was 7.12 (ppt) and standard deviation was 9.88 (ppt) on 24th Jan (Figure 4.18). It was 6.44 (ppt) in mean and 9.48 (ppt) in standard deviation on 09th Feb (Figure 4.19).

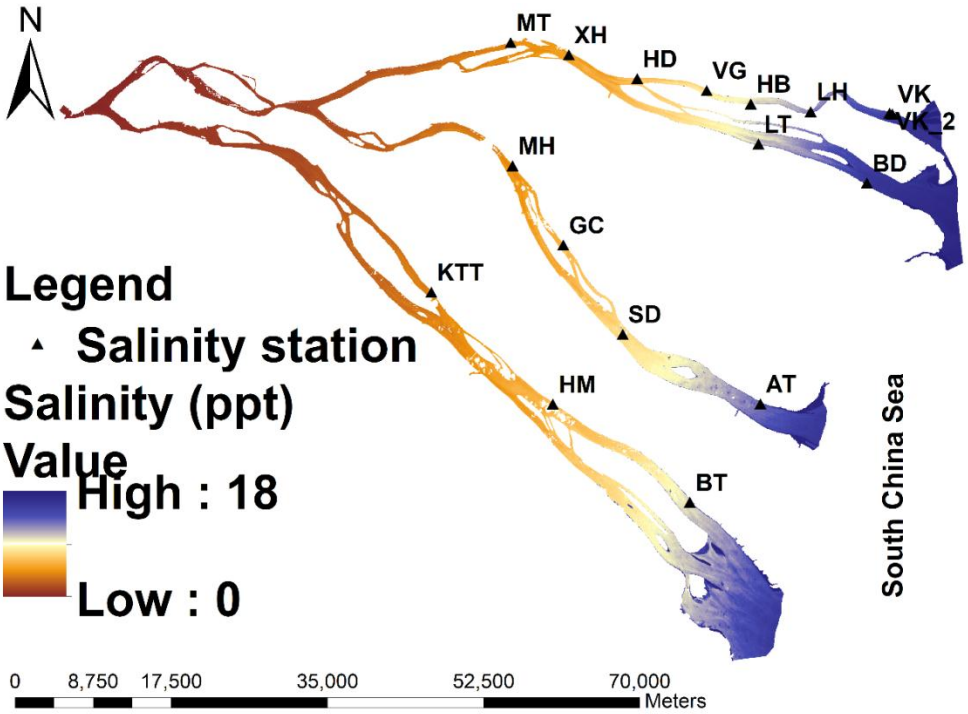


Figure 4.18 Salinity intrusion from step-wise model combination latitude and longitude 24th Jan 2015 (mean: 7.12; SD: 9.88).

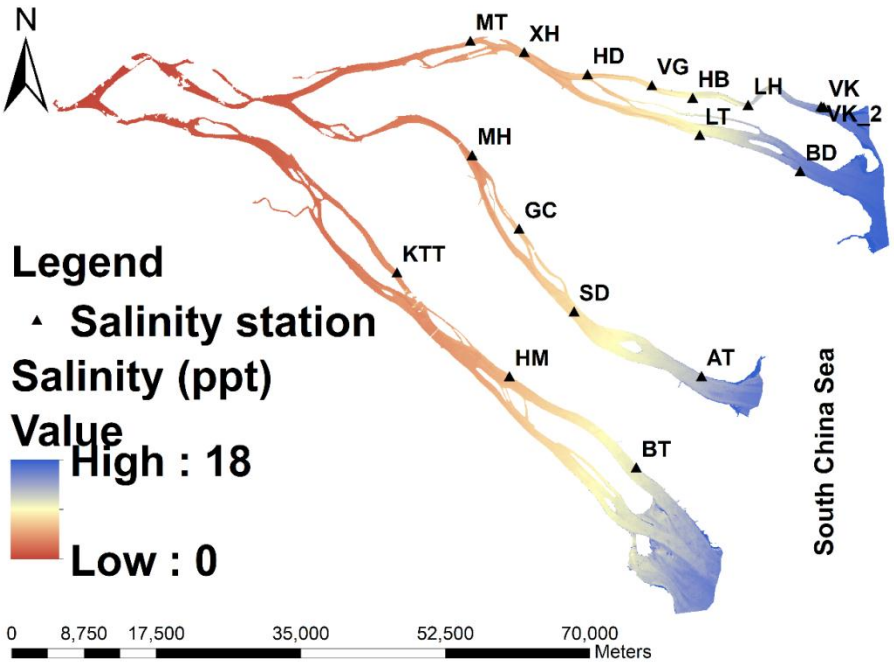


Figure 4.19 Salinity intrusion from step-wise model in combination with latitude and longitude 09th Feb 2015 (mean: 6.44; SD: 9.48).

4.2 Discussion

4.2.1 Bands selection and choosing statistical models for modeling salinity intrusion

Identification bands for setting the model was very important for the accuracy of the predictive salinity map: that means that not only does R-square and RMSE of the Modelling need to be considered but also the composite wavelength data set as inputs for predicting salinity in the whole study area. On the other hand, using too many input variables can lead for fittings which quickly becomes computationally infeasible if attempts are made to use a higher number of potential predictor variables (Messner, *et al*., 2016). Thus, proper selection of the identification bands set up for use in the model is very important for predictive accuracy of the salinity map.

Salinity of the sea surface can be represented by a function of remotely sensed ocean color bands (Urquhart, *et al*., 2012). This study found that the reflectance of wavelengths in band 2, band 3, band 4 and band 7 were important to detecting salinity intrusion. In other studies, specifically, TM band 1-5 in Landsat-5 TM was strongly correlated with salinity (Wang and Xu, 2008). Meanwhile, TM band-3 was found to be the most correlative with salinity (Wang and Xu, 2008 and Baban, 1997). The reflectance – wavelength of MODIS at 488 nm, Rrs at 667 nm and at 443 nm (Urquhart, *et al*., 2012) and the landsat-8 OLI‘s bands 1-4 were associated with salinity (Zhao, *et al*., 2017).

Selecting a suitable model seriouly affects the accuracy of the prediction, and is almost equally important as variables and more important than samples size (Fassnacht, *et al*., 2014). The multi-linear regression method clearly showed the structure of model but in cases where the relationship between variables is complex, linear regression may lack the ability to achieve a confidence level for useful for predictive purposes. The DTs showed very low potential for predicting salinity for both training and test parts. Nevertheless, the RF model is a popular method in many fields since they can be successfully applied to complex

data, with a small sample size, complex interactions and correlations and mixed type predictors (Epifanio, 2017). The RF model also exhibits the highest predictive capability compared with the LMR and DTs models (Chen, *et al*., 2017). As a result, the RF model was the best model to predict salinity intrusion. Unfortunately, where the sample size is not enough to present a sufficiently detailed profile of the study area, the results should be ignored in practice but this lower threshold level needs to be estimated. The overfitting problem was a big issue, which showed even in training tasks the R-square was rarely good enough to make useful predictions (Figure 5).

4.2.2 Modeling salinity intrusion using resampling data with bootstrap method

Resampling real data by using bootstrap method helpfully improved prediction ability of model when there is a lack of sufficient observations. As the result showed increasing samples size significantly impoved the correlation of models (DTs, ANN, ANT and RF) and also overcame the overfitting problem which was faced when using the real sample size.

The RF had its best performce at resampling four times-206 samples training and 206 for test. The RF was explained effective prediction of the data model cannot uncover in a larger sample size (Fassnacht, *et al*., 2014). However, one of the challenges in using bootstrap is the sample size of the real data. If the sample sizeis too small itcan lead to over repeats of the selected data. In this case, the model had a good correlation but its usefulness was limited in application for mapping salinity in the whole study area. The performance of the boot-strap depends on the sample size and it is hard to recommend a minimum sample size because of different situations require a different minimum sample size (Dixon, 2006).

The ANN model runing on Matlab and R program showed good correlation between salinity and reflectance. And the samples size also played a vital role in the predictive accuracy of the Neural Network model (ANT, ANN). However, one of the issues one must bear in mind is that the structure of the modelwas changed in each single setup and re-run. Thus, to find out the best

performance, the number of hidden layer,times running for every single setup as well asmethod should be carefuly considered. The ANN converge to local not overall minima and faces the overfitting problem. The ANN model is unlikely to be useful for prediction (Zou, *et al*., 2011; Niu, *et al*., 2016).

4.2.3 Mapping salinity intrusion single reflectance using bootstrap data and Reflectance- location salinity Modeling

Combination of reflectance - location improved upon the inherent weakness of single reflectance models. Field observations cannot cover enough in the whole study area. Nevertheless, this method may face with locally issues that are mentioned by Urquhart, *et al*. (2012). Thus, using the location can neglect important information on satellite images if the location plays too strong role for predicting salinity.

Furthermore, the location played a strong factor in the formula which was unexpected. Even though location is helpful in illustrating general trend but it was also a cause of unrecovered essential information from reflectance wavelength data. Using the RF model - with bootstrap data presented justifiably for mapping salinity intrusion. The sensitive of reflectance were examined to recognize salinity changes. Salinity mapping distinctly divided low concentration area in upstream and high concentration downstream. However, some areas including at the Cua Tieu on 24th Jan (Figure 4.13) and the Co Chien – Cung Hau on 09th Feb (Figure 4.14) dislocation and magnitude of salinity were still very noisy data. Especially, areas dominated intertidal mudflats which strongly impact by water level from tide phenomenon (Wolanskin, *et al*., 1996). The threshold of salinity presented quite the same tendency, but the concentration was lower than observation. Although mapping salinity intrusion still has limitations, overall using reflectance - wavelength was useful to detect salinity intrusion.

Mapping saltwater intrusion by using single – reflectance adapt proficiently an accuracy when observation enough to cover the characteristic study area. Furthermore, based on the reflectance – wavelength salinity model which was

possible for using satellite images predict salinity concentration even though missing of measured data in a large study area.

Chapter 5

Conclusion

Determining a statistical model has an essential role to optimize prediction of salinity intrusion. The RF model was demonstrated to maximize capacity for exploring the relationship between reflectance and salinity. Band 2, band 3, band 4 and band 7 were vital for developing a successful reflectance - wavelength salinity model. However, small sample size was a cause of the overfitting issue and limited the application of these models for predictive purposes. Improving sample size is the most important work for further study to explore the relationship between salinity level and single reflectance - wavelength from satellite images. Ongoing use of the model incorporating new data as it is collected will progressively improve the model. Using bootstrap technique for re-sampling data was helpful to improve effective Modeling. This study figured out that the Random Forest and bootstrap techniques were powerful methods to deal with the lack of sufficient field observations.

The combination of reflectance - locations for predicting salinity intrusion illustrated a good correlation, which could be used to predict trends of salinity intrusion. However, locality issues should be kept in mind which can negate the reflectance information in the satellite images and limit application when there is a lack of observational data. Even though mapping salinity of reflectance - location model clearly showed the salinity dynamic in study area, it also faced different limitations due to lack of observations in upstream areas and locality issues such as the suitability of monitoring station sites due to the strong weight of the latitude and longitude coefficient in the model.

The Landsat-8 OLI offers a convenient approach to enhance prediction capacity of saline intrusion. The succeess of this study not only illustrated good correlation between reflectance bands and surface saltwater in the river area but also ponited out the potential of satellite - remote sensing application to predict salinity intrusion which was meaningfull for setting up an early warning system. That is vital for solving the problems of unexpected saltwater intrusion events that can have such negative effects on agriculture activities in the VMD. Nevertheless, improving the quality and frequency of satellite images by using other satellite images (multispectral Drone camera or MODIS satellite) will be essential to quickly update salinity information to make a viable early warning system.

References

Acharya, T.D.(2015). "Exploring Landsat 8." *International Journal of IT, Engineering and Applied Sciences Research*, 4, 4–10.

Alessa, L., Kliskey, A., Gamble, J., Fidel, M., Beaujean, G., Gosz, J.(2015). "The role of Indigenous science and local knowledge in integrated observing systems: moving toward adaptive capacity indices and early warning systems." *Sustain Sci*, 11(2016),91–102.

Baban, S.M.J.(1997). "Environmental monitoring of estuaries; estimating and mapping various environmental indicators in Breydon Water Estuary, U.K., using Landsat TM Imagery." *Estuar Coastal Shelf S*, 44, 589–598.

Basu, A., Basu, Si.(2011). *Decision trees: A User's Guide to Business Analytics,* Taylor & Francis Group, CRC Press, New York.

Bernardo, N., Watanabe, F., Rodrigues, T., Alcântara, E.(2017)."Atmospheric correction issues for retrieving total suspended matter concentrations in inland waters using OLI/Landsat-8 image." *Adv Space Res*, 59, 2335–2348.

Binh, N.T.(2015). "Vulnerability and Adaptation to Climate Change." PhD thesis inthe Faculty of Agriculture, Bonn University and Institute for Environment and Human Security, United Nations University (UNU-EHS), Germany.(Online) Available on http://hss.ulb.uni-bonn.de/2015/3924/3924.pdf (15 May 2016)

Biodiversity Informatics & Geospatial Innovation Facilities .(2008). RS/GIS Quick Start Guides Collaborative training materials. (Online) Available on http://gif.berkeley.edu/documents/Landsat%20Band%20Information.pdf (17 June 2017).

Breiman, L.(2001). "Random Forest." *Mach Learn*, 45, 5–32.

CGIAR Research Centers in Southeast Asia.(2016). "The drought and salinity intrusion in the Mekong River Delta of Vietnam Report." (Online) Available on https://cgspace.cgiar.org/rest/bitstreams/78534/retrieve (25 June 2016).

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., Duan, Z., Ma, J.(2017). "A comparative study of logistic model tree, random forest, and classification and

regression tree models for spatial prediction of landslide susceptibility." *Catena*, 151, 147–160.

Constantin, S., Constantinescu, S., Doxaran, D.(2017). "Long-term analysis of turbidity patterns in Danube Delta coastal area based on MODIS satellite data." *J Marine Syst*,170, 10–21.

D'Sa, E.J., Zaitzeff, J.B., Yentsch, C.S., Miller, J.L., Ives, R.(2002). *Rapid remote assessments of salinity and ocean color in Florida Bay: The Everglades, Florida Bay, and Coral Reefs of the Florida Keys: An Ecosystem Sourcebook*.CRC Press. Taylor & Francis Group, New York.

de Winter, J.C.F., Dodou, D., Wieringa, P.A.(2009). "Exploratory Factor Analysis With Small Sample Sizes." *Multivar Behav Res*, 44, 147–181.

Deltares, Delta Alliance.(2011). "Vietnam-Netherlands Mekong Delta Masterplan project Mekong delta water resources assesment studies., Vietnam-Neetherlands Mekong Delta Masterplan project." Online. Available on https://www.wur.nl/upload_mm/2/c/3/b5f2e669-cb48-4ed7-afb6-682f5216fe7d_mekong.pdf (30 May 2016).

Dev, S., Wen, B., Lee, Y.H., Winkler, S.(2016). Machine Learning Techniques and Applications For Ground-based Image Analysis." *IEEE Geosci Remote S*, 79 - 93.

Dixon, P.M.(2006). *Bootstrap Resampling: Statistical and Numerical Computing*. John Wiley & Sons, New York.

Dorji, P., Fearns, P.(2017). "Impact of the spatial resolution of satellite remote sensing sensors in the quantification of total suspended sediment concentration: A case study in turbid waters of Northern Western Australia." *PLoS One*, 12 (4),1–24.

ENVI. (2009). "ENVI Atmospheric Correction Module: QUAC and FLAASH user's guide. Modul". Version 44.

Epifanio, I.(2017). "Intervention in prediction measure: a new approach to assessing variable importance for random forests."*BMC Bioinformatics*, 18 (2017), 1–16.

Fang, L., Chen, S., Li, H., Gu,C.D.(2008). "Monitoring water constituents and salinity variations of saltwater using EO-1 Hyperion satellite imagery in the Pearl River Estuary , China."*Geoscience and Remote Sensing Symposium*,978, 438–441.

Fang, L., Chen, S., Wang, H., Qian, J., Zhang, L.(2010). "Detecting marine intrusion into rivers using EO-1 ALI satellite imagery: Modaomen Waterway, Pearl River Estuary, China." *Int J Remote Sens*, 31, 4125–4146.

Fassnacht, F.E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., Koch, B.(2014). "Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass." *Remote Sens Environ*, 154, 102–114.

Hansen, C., Burian, S., Dennison, P., Williams, G.(2017). "Spatiotemporal Variability of Lake Water Quality in the Context of Remote Sensing Models." *Remote Sens Basel*, 9, (409),1-15.

Hashimoto, T.R.(2001). "Environmental issues and recent Infrastructure development in the Mekong Delta: review , analysis and recommendations with particular reference to largescale water control projects and the development of coastal areas." (Online) Available on http://sydney.edu.au/mekong/documents/wp4.pdf (30 May 2017).

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistic, California.

Haukoos, J.S., Lewis, R.J.(2005). "Advanced statistics: Bootstrapping confidence intervals for statistics with “difficult” distributions." *Acad Emerg Med*, 12, 360–365.

Himi, M., Benabdelouahab, S., Salhi, A., Rivero, L., Elgetta, M., El, A., Stitou, J., Casas, A. (2017). "Geophysical characterization of saltwater intrusion in a coastal aquifer: The case of Martil-Alila plain (North Morocco)." *J Afr Earth Sci*, 126, 136–147.

Jerry, C. R., Paul V. Z, and James H. E. (2003). "Remote Sensing Techniques to Assess Water Quality. " *Photogramm Eng Rem S*. 69(6). 695-704.

Katalin, G., Tamás, M., Éva, F., 2016. *Engineering Tools for Environmental Risk Management –3*.CRC PressTaylor & Francis Group,London.

Keith, D.J., Schaeffer, B. a., Lunetta, R.S., Gould, R.W., Rocha, K., Cobb, D.J.(2014). "Remote sensing of selected water-quality indicators with the hyperspectral imager for the coastal ocean (HICO) sensor." *Int J Remote Sens* , 35, 2927–2962.

Khang, N.D., Kotera, A., Sakamoto, T., Yokozawa, M.(2008). "Sensitivity of Salinity Intrusion to Sea Level Rise and River Flow Change in Vietnamese Mekong Delta-Impacts on Availability of Irrigation Water for Rice Cropping." *J Agric Meteorol*, 64, 167–176.

Kumar, D.N., Reshmidevi, T. V.(2013). "Remote Sensing Applications in Water Resources." *J Indian I Sci*, 93, 163–187.

Li, S., Sun, D., Yu, Y.(2013). "Automatic cloud-shadow removal from flood/standing water maps using MSG/SEVIRI imagery". *Int J Remote Sens*, 34, 5487–5502.

Lira,C., and Rui, T.(2014). *Advances in Applied Remote Sensing to Coastal Environments Using Free Satellite Imagery Cristina*. Coastal Research Library. Springer International Publishing Switzerland, New York. p 431–452.

Ließ, M., Glaser, B., Huwe, B.(2012). "Uncertainty in the spatial prediction of soil texture. Comparison of regression tree and Random Forest models." *Geoderma*, 170, 70–79.

Liu, R., Gillies, D.F.(2016). "Overfitting in linear feature extraction for classification of high-dimensional image data." *Pattern Recogn*, 53, 73–86.

Ma,H., Guo,S., Zhou,Y.(2013). “Detection of Water Area Change Based on Remote Sensing Images.” Proceedings of Geo-Informatics in Resource Management and Sustainable Ecosystem, Part I. Wuhan, China: 8-10 November, 2013.

Mackinnon, J.G.(2009). *Handbook of Computational Econometrics: Chapter 6 Bootstrap Hypothesis Testing, Handbook of Computational Econometrics*. John Wiley & Sons,Chichester.

Messner, J.W., Mayr, G.J., Zeileis, A.(2017). "Nonhomogeneous Boosting for Predictor Selection in Ensemble Postprocessing. Am." *American Meteorological Society*, 145 (2017), 137-147.

Mohamed, H., Salah, M., Nadaoka, K. (2017). "Assessment of proposed approaches for bathymetry calculations using multispectral satellite images in shallow coastal / lake areas: a comparison of five models." *Arab J Geosci*, 42 (2017),1–17.

Nguyen, A.D., Savenije, H.H.G.(2006). "Salt intrusion in multi-channel estuaries: a case study in the Mekong Delta, Vietnam." *Hydrol Earth Syst Sc*, 3, 499–527.

Nguyen, A.D., Savenije, H.H.G., Pham, D.N., Tang, D.T.(2008). "Using salt intrusion

measurements to determine the freshwater discharge distribution over the branches of a multi-channel estuary: The Mekong Delta case." *Estuar Coastal Shelf S*, 77, 433–445.

Niu, X., Yanga, C., Wanga, H., Wanga, Y.(2016). "Investigation of ANN and SVM based on limited samples for performance and emissions prediction of a CRDI-assisted marine diesel engine." *Appl Therm Eng*, 111, 1353–1364.

Noh, S., Choi, M., Kim, E., Dan, N.P., Thanh, B.X., Ha, N.T. Van, Sthiannopkao, S., Han, S. (2013). "Influence of salinity intrusion on the speciation and partitioning of mercury in the Mekong River Delta." *Geochim Cosmochim Ac*, 106, 379–390.

Ong, D.C.(2014). "A primer to bootstrapping and an overview of doBootstrap." (Online) available https://web.stanford.edu/class/psych252/tutorials/do Bootstrap Primer.pdf. (30 May 2016).

Owen, A.B. 2001. *Chap. 4: Empirical Likelihood in Regression and modeling*, Chapman and Hall, New York.

Peddle, D.R., White, H.P., Soffer, R.J., Miller, J.R., Ledrew, E.F.(2001). "Reflectance processing of remote sensing spectroradiometer data." *Comput Geosci*, 27, 203–213.

Preacher, K.J., Hayes, A.F.(2008). "Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models." *Behav Res Methods*, 40, 879–891.

Ratner,B.(2011). *Overfitting: Old Problem, New Solution Statistical and Machine-Learning Data Mining Techniques for Better Predictive Modeling and Analysis of Big Data, Second Edition.* Taylor & Francis Group. CRC Press, Boca Raton.

Roy, D.P., Wulder, M.A., Loveland, T.R., C.E., W., Allen, R.G., Anderson, M.C., Helder, D., Irons, J.R., Johnson, D.M., Kennedy, R., Scambos, T.A., Schaaf, C.B., Schott, J.R., Sheng, Y., Vermote, E.F., Belward, A.S., Bindschadler, R., Cohen, W.B., Gao, F., Hipple, J.D., Hostert, P., Huntington, J., Justice, C.O., Kilic, A., Kovalskyy, V., Lee, Z.P., Lymburner, L., Masek, J.G., McCorkel, J., Shuai, Y., Trezza, R., Vogelmann, J., Wynne, R.H., Zhu, Z.(2014). "Landsat-8: Science and product vision for terrestrial global change research." *Remote Sens Environ*, 145, 154–172.

Roy, P.K., Roy, S.S., Giri, A., Banerjee, G., Majumder, A., Mazumdar A.(2014). "Study of impact on surface water and groundwater around flow fields due to changes in river stage using groundwater modeling system." *Clean Technol Envir*, 17, 145–154.

Sawaya, K.E., Olmanson, L.G., Heinert, N.J., Brezonik, P.L., Bauer, M.E.(2003). "Extending satellite remote sensing to local scales: land and water resource monitoring using high-resolution imagery." *Remote Sens Environ*, 88, 144–156.

Shalizi, C.R.(2011). "The Bootstrap." (Online) Available on http://www.stat.cmu.edu/~cshalizi/402/lectures/08-bootstrap/lecture-08.pdf (15 May 2016).

Shcheglovitova, M., Anderson, R.P.(2013). "Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes." *Ecol Model*, 269, 9–17.

Simar, L., Wilson, P.W.(1998). "Models Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models." *Manage Sci*, 44, 49–61.

Stephen, M.(2007). *Part V- Intelligent System - Machine Learning in Computing Handbook- Computer Science and Software Engineering*. Chapman and Hall, Bacu Raton.

Su, Y.-F., Liou, J.-J., Hou, J.-C., Hung, W.-C., Hsu, S.-M., Lien, Y.-T., Su, M.-D., Cheng, K.-S., Wang, Y.-F.(2008). "A Multivariate Model for Coastal Water Quality Mapping Using Satellite Remote Sensing Images." *Sensors*, 8, 6321–6339.

Tran, P., Tran, L.(2016). "Validating negative binomial lyme disease regression model with bootstrap resampling." *Environ Model Softw*, 82, 121–127.

Thai, L.X.,Dung, T.N. (2013). "Selecting rice varieties tolerant to salinity in the Mekong Delta of Vietnam." Can Tho University Journal, 28 (2013), 79-85.

Trung, N.H., Tri, V.P.D. (2014). *Possible Impacts of Seawater Intrusion and Strategies for Water Management in Coastal Areas in the Vietnamese Mekong Delta in the Context of Climate Change: Coastal Disasters and Climate Change in Vietnam*. Elsevier Inc, London.

Trung, N.H., C.H.Hoanh, T.P. Tuong, X.H. Hien,L.Q. Tri, V.Q. Minh, D.K. Nhan, P.T. Vu, and V.P.D.Tri. 2016. "Climate Change Affecting Land Use in the Mekong Delta:

Adaptation of Rice-based Cropping Systems (CLUES) Theme 5: Integrated adaptation assessment of Bac Lieu Province and development of adaptation master-plan." https://www.researchgate.net/publication/301612048_Climate_change_affecting_land_use_in_the_Mekong_Delta_Adaptation_of_rice-based_cropping_systems_CLUES_ISBN_978-1-925436-36-5. (03 March 2017).

Urquhart, E.A., Zaitchik, B.F., Hoffman, M.J., Guikema, S.D., Geiger, E.F.(2012). "Remotely sensed estimates of surface salinity in the Chesapeake Bay: A statistical approach." *Remote Sens Environ*,123, 522–531.

USGS, 2015). "Landsat 8 (L8) Data Users Handbook." Earth Resources Observation and Science (EROS) Center.(Online) Available on https://landsat.usgs.gov/sites/default/files/documents/Landsat8DataUsersHandbook.pdf (30 May 2016).

Vanhellemont, Q., Ruddick, K. (2015). "Advantages of high quality SWIR bands for ocean colour processing: Examples from Landsat-8." *Remote Sens Environ*, 161, 89–106.

Vanhellemont, Q., Ruddick, K. .(2014)." Landsat-8 As a Precursor To Sentinel-2 : Observations of Human Impacts in Coastal Waters." *Proceedings of the Sentinel-2 for Science Workshop (2008)*, Frascati, ESA Italy: 20-23 May, 2014.

Wang, F., Xu, Y.J.(2008)."Development and application of a remote sensing-based salinity prediction model for a large estuarine lake in the US Gulf of Mexico coast." *J Hydrol*, 360, 184–194.

Wang, F., Xu, Yj.(2012). *Remote Sensing to Predict Estuarine Water Salinity: Environmental Remote Sensing and Systems Analysis.* Taylor & Francis Group. CRC Press,Boca Raton.

Wang, J.(2012). “High Resolution Remote Sensing Information Identification for the Salinity in GASIKULE Salt Lake.” *Proceedings on 2nd International Conference - Remote Sensing, Environment and Transportation Engineering (RSETE)*, Nanjing, China: 1-3 June, 2012.

Wang, L., Zhou, X., Zhu, X., Dong, Z., Guo, W.(2016). "Estimation of biomass in wheat using random forest regression algorithm and remote sensing data." *The Crop*

Journal, 4, 212–219.

Waske, B., Fauvel, M., Benediktsson, J.A., Chanussot, J.(2009). *Machine learning techniques in remote sensing data analysis: Kernel Methods for Remote Sensing Data Analysis.* John Wiley & Sons, Singapore.

Wehrens, R., Putter, H., Buydens, L.M.C.(2000). "The bootstrap in : A tutorial." *Chemometr Intell Lab*, 54 (2000), 35–52.

Wolanski,E., Huan, N.N., Dao, L.T. , Nhan, N.H., and Thuy, N.N. (1996). "Fine-sediment Dynamics in the Mekong River Estuary, Vietnam." *Estuar Coast Shelfs S*, 43,565-582.

Xiong, Y.J., Qiu, G.Y., Chen, X.H., Tan, S.L., Feng, H.X.(2012). "Hyperspectral characteristics of seawater intrusion in Pearl River delta , China based on laboratory experiments." *Proceeding of the IEEE International Conference-Geoscience and Remote Sensing Symposium (IGARSS),* Munich, Germany: 22-27 July 2012.

Zhao, J., Temimi, M., Ghedira, H.(2017). "Remotely sensed sea surface salinity in the hyper-saline Arabian Gulf: Application to landsat 8 OLI data." *Estuar Coast Shelf S*, 187, 168–177.

Zhou, F., Zhang, A.(2016). "Optimal subset selection of time-series MODIS images and sample data transfer with random forests for supervised classification modelling. *Sensors*, 16, 1–18.

Zhu, Z., Wang, S., Woodcock, C.E.(2015). "Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images." *Remote Sens Environ*, 159, 269–277.

Zhu, Z., Woodcock, C.E.(2012). "Object-based cloud and cloud shadow detection in Landsat imagery." *Remote Sens Environ*, 118, 83–94.

Zou, K.H., Aiyi Liu, B., Andriy, I. R., Ohno-Machado, Lucila., H.E.(2011). *Machine Learning and Predictive Modeling 9.1: Statistical Evaluation of Diagnostic Performance Topics in ROC Analysis*, Chapman and Hall, CRC., Taylor & Francis Group, London.

APPENDICES

Appendix I: Reflectance wavelength of Landsat-8 OLI and salinity observations for setting up models in the study

B1	B2	B3	B4	B5	B6	B7	Salinity
0.0894	0.0837	0.1267	0.1351	0.091	0.0657	0.0498	18.122
0.0816	0.0794	0.1179	0.1253	0.0678	0.0409	0.0313	17.935
0.0955	0.0983	0.1419	0.1606	0.0901	0.0689	0.058	17.906
0.0705	0.0685	0.109	0.1218	0.0691	0.0596	0.0498	18.378
0.0847	0.0891	0.1273	0.1372	0.075	0.06	0.0504	15.902
0.1027	0.104	0.1348	0.129	0.0759	0.0623	0.0501	14.311
0.0878	0.0842	0.1212	0.1553	0.1492	0.0425	0.0345	15.727
0.0562	0.0578	0.0981	0.1193	0.0318	0.015	0.013	16.536
0.08	0.0816	0.1283	0.1379	0.0804	0.071	0.0593	18.829
0.0923	0.0912	0.1207	0.1291	0.075	0.0577	0.0476	18.222
0.0502	0.0529	0.0918	0.1024	0.0169	0.0145	0.0145	16.152
0.0912	0.0905	0.1299	0.1553	0.0719	0.0114	0.0072	12.386
0.0813	0.0751	0.1037	0.1222	0.0606	0.0102	0.0068	16.124
0.1009	0.0922	0.1154	0.1394	0.1359	0.019	0.0113	19.205
0.0549	0.0567	0.0852	0.0669	0.0128	0.0111	0.0096	10.517
0.0599	0.0595	0.0919	0.0961	0.0172	0.0092	0.0072	15.964
0.0664	0.0648	0.0942	0.1132	0.0422	0.006	0.0048	16.161
0.0977	0.0996	0.1433	0.1608	0.0844	0.0662	0.0565	14.537
0.087	0.0862	0.1312	0.1507	0.0868	0.0664	0.0521	15.118
0.055	0.0566	0.0981	0.1144	0.0342	0.0176	0.0145	13.799
0.0866	0.088	0.1177	0.1273	0.0745	0.057	0.0471	17.354
0.0502	0.0521	0.0927	0.1054	0.034	0.0227	0.021	9.499
0.0986	0.0894	0.1122	0.1346	0.1331	0.0184	0.011	15.86
0.0639	0.0627	0.0954	0.1028	0.0248	0.0094	0.0074	12.713
0.0679	0.0641	0.0931	0.1119	0.0469	0.0073	0.0058	14.082
0.0611	0.0633	0.0903	0.0943	0.0362	0.031	0.0259	13.239
0.0515	0.0506	0.089	0.1034	0.0335	0.018	0.0162	21.732
0.0624	0.0626	0.091	0.0733	0.035	0.031	0.0258	14.119
0.0679	0.0658	0.0984	0.092	0.0327	0.0306	0.0252	12.089
0.1005	0.0965	0.126	0.1209	0.0851	0.0647	0.0549	8.931
0.089	0.0801	0.1108	0.1312	0.0722	0.0255	0.019	17.198
0.0805	0.0772	0.1028	0.1025	0.0403	0.0289	0.0244	15.634
0.0575	0.061	0.1035	0.1124	0.0222	0.0127	0.0115	0.745
0.0808	0.0796	0.1154	0.1097	0.0414	0.0289	0.024	0.44
0.0534	0.0554	0.0888	0.0828	0.0404	0.0317	0.0264	6.075
0.0626	0.068	0.1184	0.1355	0.0345	0.0217	0.0198	2.816
0.0616	0.0552	0.0821	0.0909	0.0906	0.0478	0.0314	7.412

0.1011	0.0976	0.1311	0.1436	0.1013	0.0464	0.0376	2.482
0.0541	0.0559	0.098	0.129	0.0554	0.0165	0.0148	4.417
0.0871	0.0821	0.1118	0.1037	0.0305	0.0218	0.017	3.498
0.057	0.0568	0.0953	0.0897	0.0548	0.0441	0.0354	4.92
0.0491	0.0508	0.0885	0.0875	0.0295	0.0174	0.0147	8.125
0.0962	0.0931	0.1229	0.1103	0.0576	0.0441	0.0342	6.221
0.0555	0.059	0.1042	0.1057	0.0195	0.0086	0.0072	3.905
0.0481	0.0519	0.0894	0.0847	0.028	0.0258	0.0216	17.451
0.0969	0.094	0.1245	0.104	0.0638	0.0497	0.0414	6.288
0.0785	0.0782	0.1199	0.1019	0.0524	0.0428	0.0347	8.572
0.0499	0.0513	0.0871	0.0707	0.0214	0.0169	0.015	12.462
0.073	0.0715	0.1039	0.0755	0.0413	0.0392	0.0319	6.602
0.053	0.0496	0.0809	0.0673	0.0053	0.0042	0.0044	10.552
0.0446	0.0513	0.0842	0.0623	0.0191	0.0202	0.0176	16.164
0.0859	0.0828	0.1091	0.0841	0.0416	0.0348	0.0272	9.055
0.0583	0.0596	0.0987	0.1104	0.0342	0.0204	0.0171	4.746
0.0781	0.0811	0.1267	0.1467	0.0632	0.0302	0.0239	4.153
0.1158	0.1201	0.1632	0.1536	0.0894	0.0856	0.0687	7.697
0.0756	0.0681	0.0999	0.1287	0.1137	0.0193	0.0125	9.205
0.0478	0.0507	0.0879	0.0818	0.0216	0.0226	0.0188	12.482
0.0689	0.061	0.1	0.0937	0.0528	0.0257	0.0197	10.454
0.0809	0.084	0.1281	0.1392	0.0649	0.0525	0.0447	7.263
0.1739	0.1629	0.1891	0.1961	0.1611	0.0891	0.0725	5.954
0.0769	0.0764	0.1196	0.1562	0.1084	0.0298	0.0243	9.308
0.0539	0.0544	0.0972	0.1246	0.0509	0.0177	0.0155	8.651
0.0612	0.0637	0.1107	0.1183	0.0654	0.0573	0.0484	14.75
0.0885	0.0906	0.1294	0.1256	0.0523	0.0473	0.0385	10.758
0.0798	0.0869	0.1402	0.1729	0.0644	0.0269	0.0241	0.242
0.0848	0.0852	0.1334	0.1719	0.0871	0.0509	0.0409	0.679
0.0875	0.0872	0.1267	0.1437	0.0831	0.0665	0.0536	2.222
0.1168	0.1128	0.1528	0.1691	0.0649	0.0333	0.0267	0.459
0.0908	0.0933	0.129	0.1244	0.039	0.0333	0.0278	2.617
0.1101	0.1127	0.1662	0.2027	0.1059	0.058	0.0493	1.877
0.0784	0.0776	0.1261	0.1565	0.079	0.055	0.0447	3.12
0.1138	0.1159	0.1635	0.1761	0.0801	0.0619	0.0502	1.176
0.1129	0.1141	0.1586	0.1651	0.0755	0.0597	0.0472	2.512
0.0824	0.0825	0.1254	0.1436	0.1033	0.0733	0.059	7.5
0.0835	0.0843	0.1211	0.121	0.0454	0.0376	0.0307	4.248
0.0892	0.0863	0.1234	0.1463	0.0722	0.0467	0.0371	5.403
0.0909	0.0879	0.1246	0.1268	0.0458	0.0343	0.0257	2.991
0.0488	0.0488	0.0885	0.1009	0.032	0.0209	0.0178	14.594
0.1057	0.1091	0.1421	0.1286	0.0891	0.0788	0.0651	11.889

0.0834	0.0838	0.1284	0.1307	0.0732	0.0665	0.0547	15.179
0.0929	0.0906	0.1291	0.1604	0.0838	0.0126	0.0087	10.787
0.0772	0.0709	0.0944	0.1134	0.0521	0.0131	0.0097	17.157
0.1013	0.0906	0.1123	0.1326	0.0841	0.0174	0.0113	13.683
0.0556	0.0577	0.084	0.0643	0.0192	0.014	0.0124	10.192
0.0617	0.06	0.0917	0.1069	0.0326	0.0096	0.0079	15.822
0.0638	0.0607	0.0897	0.1114	0.047	0.0074	0.0059	16.775
0.1222	0.1126	0.1353	0.1521	0.0893	0.0379	0.0272	18.917
0.0757	0.0723	0.1081	0.1387	0.0841	0.0245	0.0216	15.875
0.0718	0.0674	0.0938	0.1015	0.0279	0.0136	0.0115	15.626
0.0672	0.0641	0.0986	0.1279	0.0872	0.0172	0.0127	17.257
0.0747	0.0707	0.0961	0.115	0.0543	0.0162	0.0122	21.306
0.0539	0.063	0.104	0.1013	0.0418	0.039	0.0325	19.192
0.0835	0.0759	0.0969	0.1035	0.0358	0.0103	0.0076	20.844
0.079	0.0723	0.0932	0.1	0.0347	0.01	0.0074	16.086
0.0806	0.0823	0.1201	0.1351	0.0618	0.0476	0.0405	1.474
0.0743	0.077	0.1144	0.1222	0.0376	0.0169	0.0133	0.7
0.0949	0.0938	0.1252	0.1199	0.0602	0.0465	0.0374	4.798
0.0645	0.0668	0.108	0.0988	0.0048	0.0061	0.006	2
0.0416	0.0465	0.082	0.0621	0.005	0.0137	0.0119	6.212
0.0419	0.0525	0.0988	0.0968	0.0429	0.018	0.0162	1.898
0.059	0.0608	0.1004	0.096	0.0357	0.0324	0.0258	0.84
0.0955	0.0958	0.1354	0.1436	0.0559	0.0349	0.0289	0.33
0.0597	0.0623	0.0996	0.0859	0.0255	0.0232	0.0197	3

VITAE

Name NGUYEN Thi Bich Phuong

Student ID 5830222003

Educational Attainment

Degree	Name of Institution	Year Graduation
Engineer of Environment	Can Tho University, Vietnam	2015

List of Publication and Proceedings

1. Phuong, T.B Nguyen, Werapong Koedsin, Donald McNeil and Tri P. D Van. (2017). Remote Sensing Techniques to Predict Salinity Intrusion: Application for a Data-Poor Area of the Coastal Mekong Delta, Vietnam. International Journal of Remote Sensing. (*In press, accepted for publication)*
2. Phuong, T.B. Nguyen, Werapong Koedsin, Donald McNeil and Raymond J Ritchie.(2017). Using bootstrap method for resampling data and statistical modeling to detect salinity intrusion with Satellite images in the Coastal Mekong Delta, Vietnam. 2017 (*in preparation)*
3. Phuong, T.B. Nguyen, Thanh T. Vo and Tri P. D Van. (2015). Impacts of land use change on the hydrological characteristics of the Duong Dong river basin, Phu Quoc island. Journal of Can Tho University, 40 (2015), 81-91.