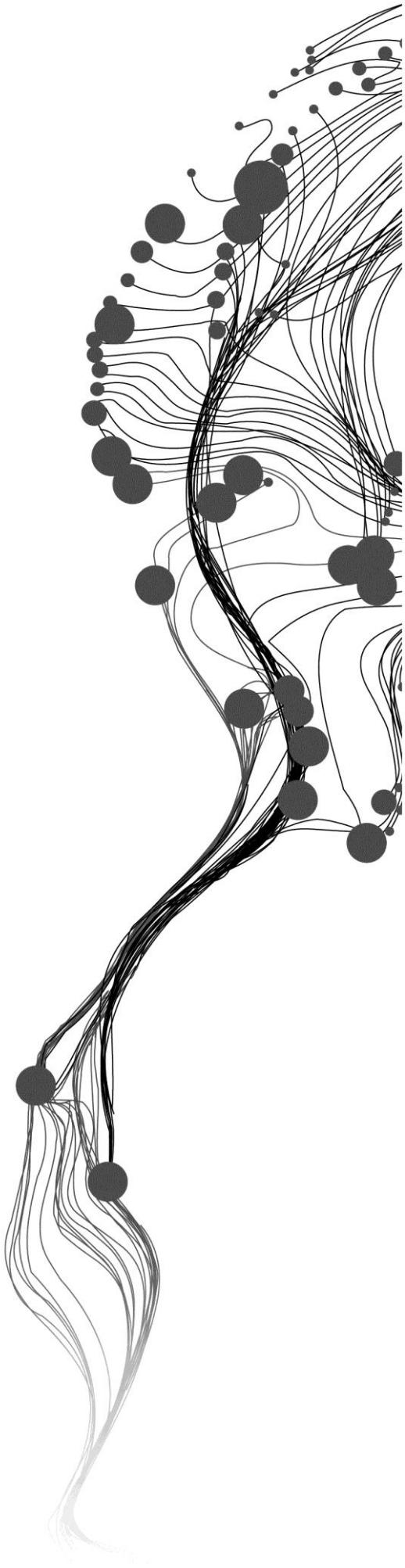
COMPARING THREE CLASSIFIERS FOR DETECTING HYDROCARBON SEEPAGE ALTERATION

WEI WANG
March, 2016

SUPERVISORS:
Mr. W.H. Bakker MSc
Prof. Dr. F.D. van der Meer



COMPARING THREE CLASSIFIERS FOR DETECTING HYDROCARBON SEEPAGE ALTERATION

WEI WANG
Enschede, The Netherlands, March, 2016

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Applied Earth Sciences

SUPERVISORS:
Mr. W.H. Bakker MSc
Prof. Dr. F.D. van der Meer

THESIS ASSESSMENT BOARD:
Dr. M. van der Meijde (Chair)
Dr. Friedrich Kuehn (External Examiner, Federal Institute for Geosciences and Natural Resources (BGR), Germany (retired))

ABSTRACT

Hydrocarbon seepages are effective indicators of oil and gas presence in the underground. They may alter the rocks and cause mineral alterations at the surface. Through detecting the changes of minerals in the surface seeps can be identified by remote sensing technology. In this research Advance Spaceborne Thermal Emission and Reflection Radiometer (ASTER) and WorldView-2 (WV-2) data were used to detect gas-induced alteration in the marly limestone formation in the Dezful Embayment, southwest Iran.

In this research, first a knowledge-based approach (band ratio and relative absorption band depth) was applied to detect alterations. Box plots were made for selecting mineral indexes that could be used for detecting alterations. The combination of laboratory and image-driven spectral analysis illustrates that the alterations are dominated by gypsum, clays, sulfur and ferrous minerals. Furthermore, false color composition images, composed by selected mineral indexes, could be used to identify the alteration zones. As gypsum was observed as an indicator for alterations, the SWIR bands in ASTER were most important. Thus, although WorldView-2 data improved the spatial resolution of ASTER data, it did not improve the classification result.

In the second part of this work, a data driven approach was introduced to classify altered and unaltered areas. Three classifiers, the Supported Vector Machine (SVM), Random Forest (RF) and Gradient Boosted Regression Trees (GBRT), were trained by two training sets of different sizes. The training sets were selected by the spatial–spectral endmember extraction tool (SSEE) with the help of alteration maps produced by knowledge-based approach applied in the first part of the research. However, the altered areas are obviously bright in ASTER data. To eliminate the influence of image intensity, the ASTER data was transformed to Principle component analysis (PCA) image. The new imagery was converted back by PCA image without first component which contained the intensity information. The performance of these classifiers was compared by testing the two different training sets and images. With the significant learning ability using small training sets and a good stability, the SVM method is observed to be the most suitable classifier for detecting hydrocarbon seepage alteration.

Finally, the trained SVM classifier was used to produce a regional gas-induced alteration map. Two previously unknown areas were interpreted as potential hydrocarbon seepage alterations. The geologic models were built to interpret the occurrences of the potential hydrocarbon seeps. In addition, this research was compared with Salati (2014)' work and improved the classification accuracy.

Keywords: hydrocarbon seepage, mineral alteration, ensemble classifier, classifier, remote sensing, Supported Vector Machine, Random Forest, Gradient Boosted Regression Trees

ACKNOWLEDGEMENTS

I am very grateful to have the chance to study in ITC. I have learnt a lot and had a happy life in the past one and a half years. Hereby, I would like to thank each person for your generous help.

First of all, many thanks to my supervisors, Mr. Wim Bakker and Prof. Freek van der Meer. At the beginning, I had little knowledge on classifiers and hydrocarbon seeps. During the thesis period, they guided me to think and encouraged me to express. It is their guidance that supervise me to accomplish the research. The fruitful discussions are like a guiding light showing the right way. Their critical comments improve my professional knowledge and written skill.

I am very grateful to all my teachers. Under their critical lectures, I learned a lot remote sensing geological knowledge which lay a good foundation for completing this research.

I would like to thank my parents. My mother and father are very tolerant and open-minded. They respect and support all my decisions. I cannot go so far without their love.

Lastly, I sincerely thank my classmates and friends. We share the information with each other and learn from each other. I cannot have such a happy time without you. I wish all of you have a bright future.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
WV-2	WorldView-2
VNIR	Visible and near-infrared
SWIR	Short wave infrared
TIR	Thermal infrared
DEM	Digital elevation model
SVM	Supported vector machine
RF	Random forest
GBRT	Gradient boosted regression trees
FCC	False color composite
TCC	True color composite
PCA	Principle component analysis
PC_1	First PCA component
SSEE	Spatial–spectral endmember extraction tool
RBF	Radial basis function
OOB	Out of bag
ntree	Number of trees in 'randomForest' package
mtry	Predictor variables in 'randomForest' package

1. INTRODUCTION

1.1. Background

As petroleum is one of the most important energy sources, various methods are proposed constantly and applied for exploration, such as seismic prospecting, gravity prospecting, magnetic prospecting, electrical prospecting, geochemical exploration and remote sensing(Ceron et al., 2001). Nowadays, seismic prospecting is one of the most common and important methods. However, this method is expensive and time consuming.

Near-surface hydrocarbon seepages are effective indicators of subsurface oil and gas. Liquid and gaseous hydrocarbons escape to the surface through imperfect and leaking seals or cap rocks to form natural springs, which are called oil and gas seepages. Oil and gas seepages have two phases: macro-seeps (visible seeps) and micro-seeps (invisible seeps) respectively(van der Meer et al., 2002). In about 75% basins of oil and gas, seepages were found(Etiope, 2015). Selley(1992) concluded that the most locations of hydrocarbon seepages in UK and Macgregor(1993) showed a relationship between seepages and subsurface petroleum reserves. Through fractures and faults, oil and gas escape to the surface and form hydrocarbon seepages(Macgregor, 1993). Six migration types of seepages are distinguished: unconformity related seepages, salt dome related seepages, anticline related seepages, mud volcano related seepages, normal fault, and thrust fault related seepages respectively. Therefore, seeps are associated with the source and structure of a basin at regional scale, but the near surface migration of seeps is usually more complex and controlled by fracture systems. Although there is quite a complicated relation between hydrocarbon seepages and subsurface petroleum(Abrams, 2005), seeps still have an obvious value for petroleum exploration (Pirkle & Jones, 2006).

In the process of upward migration of oil and gas, hydrocarbons react with the surrounding rock, soil, and vegetation, and cause alterations of rock and soil. Many researchers studied the anomalous patterns in vegetation or changes in vegetation diversity and type to detect hydrocarbon seepages, but this is not a suitable method for sparsely vegetated areas, and seepages can only at the very near surface interact with vegetation. Whereas, at the surface, hydrocarbons oxidize and form a reducing and slight acid environment which is associated with red bed bleaching (the conversion of Fe^{3+} to Fe^{2+}), clay minerals (the conversion of feldspar to clay minerals like kaolinite) and carbonates (the presence of Fe^{2+} rich carbonates like siderite) (Lammoglia & de Souza Filho, 2013). A variety of alterations could happen depending on the original rock composition, the type of gas, and the pressure-temperature conditions. For instance, mineral alterations in the evaporite formation in Zagros, which are affected by interaction of hydrocarbon and evaporite, contain jarosite, alunite, natroalunite and sulfur (Tangestani & Validabadi, 2014). Salati (2014) confirmed that gypsum, jarosite and sulphur are associated with hydrocarbon seeps in Gath-e-tursh. Therefore, hydrocarbon seepages have a high correlation with mineral alterations. That means it is feasible to detect seeps depending on mineral alteration.

Remote sensing is one of approaches of mineral exploration(Bedini, 2011; Sabins, 1999) and has advantages of low cost and saving time. Moreover, the shortages is that it cannot detect the depth and the quality of subsurface reservoir. The Advanced Spaceborne Thermal Emission and Reflection Radiometer

(ASTER) has 14 spectral bands, including 3 bands in the range of VNIR wavelength, 6 bands of SWIR, 5 bands of TIR and a DEM, which is a very useful sensor for geology(Gomez et al., 2005). Most research focuses on lithologic mapping, economic minerals exploration and detecting vegetation changing caused by hydrocarbons by using ASTER. Moreover, ASTER images can be used to map iron oxides, clays, carbonates, quartz and chlorite, which makes it suitable to create mineral maps at regional scale(van der Meer et al., 2012). CSIRO(2013) show their ASTER geoscience maps on their website. To detect more details and improve reliability, a high-resolution sensor is better used to enhance the spatial and spectral resolution of ASTER. Since 1980's, hyperspectral imagery has been a hot topic in mineral mapping(van der Meer et al., 2012).

For mineral distribution, most classification methods are based on subpixel unmixing analysis (Mulder et al.2011), such as Successive Projection Algorithm (SPA), Linear Spectral Unmixing (LSU), Iterative Spectral Mixture Analysis (ISMA), Mixture-tuned matched filtering (MTMF), Matched filtering (MF) and Constrained Energy Minimization (CEM), Support Vector Machines (SVM) with sigmoid (Cui et al., 2015; McCarthy et al., 2015; Van der Meer & Jia, 2012; Vicente & de Souza Filho, 2011; Zhang & Li, 2014; Platt, 1999). However, there are two restrictions in my research: the area of hydrocarbon alterations is relatively small and the amount of field data is limited in my study area. These limitations make it difficult for classifiers to get a reliable result.

Ensemble classifiers were developed to combine multiple classifiers for supervised or unsupervised learning and aims at improving the accuracy and reliability of single classifiers (Mao et al., 2015). A supervised learning algorithm is used to find a good hypothesis. These hypotheses are joined up to produce a better hypothesis by the ensemble algorithm (Rokach, 2009). Delgado et al. (2014) used 121 data sets to evaluated 17 families of 179 classifiers. According to this paper, the best families of classifiers are random forest, and support vector machines are second best. RF is an ensemble classifier that contains multiple decision trees. This method combines Bootstrap aggregating and a random subspace method to build an ensemble of decision trees (Walton, 2008). Lowe & Kulkarni (2015) compared the accuracy of RF, SVM, neural network and maximum likelihood for classification in multispectral imagery, and according to this paper RF performed best while SVM occupied second position. Salati et al.(2014) applied an ensemble classifier for hydrocarbon alteration detection. These authors chose the Boosted Regression Trees (BRT) classification method, which was successful in detecting seeps. There were four reasons for the authors to choose BRT (Salati et al., 2014): (1) the classification result is visualized; (2) the predictors can be various types; (3) irrelevant predictors are avoided; (4) it is not sensitive to outliers.

1.2. Research problem

In this research projection, two main problem should be solved. First, selecting and comparing for hydrocarbon seeps detection. Second, applying optimal classifier compared in this research to detect potential hydrocarbon seep alterations in a regional scale

In sparse vegetation areas, remote sensing can be used to detect hydrocarbon seepages based on the alteration minerals associated with such seepages. Good classifiers can improve the accuracy and reliability of mineral classification. Thus, the key point is how to choose the best classifier. In recent years, although the ensemble technology is applied successfully and performs better than single classifiers in many fields including remote sensing geology (Gao & Xu, 2015; Knudby et al., 2014; Merdith et al., 2015; Zhang et al., 2015), only BRT used in hydrocarbon alteration detection(Salati et al., 2014). However, as far as we know other classifierswith outstanding performance haven't been used, such as RF. Therefore, in this research

good classifiers will be chosen and applied to the test area. Moreover, depending on the performance of these classifiers and choosing the best classifier, we aim to produce a hydrocarbon alteration map at a regional scale, which has never been done before as far as we know.

1.3. Scientific significance and innovative aspects

The scientific significance and innovative aspects are shown following:

For mineral mapping, airborne hyperspectral imagery is optimum. But we don't have hyperspectral imagery in the study area. Hyperspectral imagery are expensive, and acquiring new airborne hyperspectral imagery is time-consuming. Instead, multispectral imagery is cost-saving. In this study, I will try to use multispectral and very high resolution data (ASTER & World View-2) to map hydrocarbon seeps related alteration.

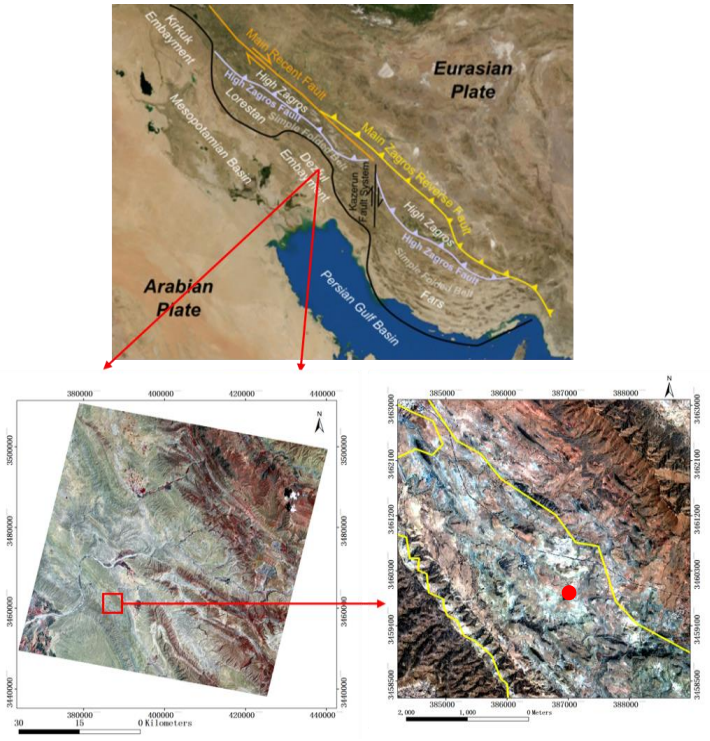
As ASTER has 14 bands and World View-2 only has VNIR bands, producing mineral map by multi-spectral imagery is harder than hyperspectral imagery. It is very important to choose suitable classifiers which have good potential in mineral classification by using multispectral and very high resolution imagery. Based on amount of literatures and trial, I will choose classifiers and these classifiers will be compared in this research. After that, a more reliable way to classify hydrocarbon seeps related alterations based on multi-spectral data will be highlighted.

1.4. Study area and datasets

1.4.1. Study area

The study area is located in the Dezful Embayment, central-southern Zagros fold thrust belt, in south-west Iran. The Zagros Basin, the second largest basin in middle East (Nairn & Alsharhan, 1997), is the largest structure controlled oil and gas field group (ZOU et al., 2015). Almost all the oil fields are located in the Dezful Embayment. Hydrocarbon seepages are very common here, but only a small number of studies have been carried out. There are three reasons: (1) this area is largest structure controlled basin; (2) hydrocarbon seepages are common; (3) it is arid. In Dezful Embayment, hydrocarbon seepages are mainly associated with limestone and evaporite formation. Salati et al. (2014) worked on a local scale area in the north of Dezful Embayment, which is sparsely vegetated, and chose Boosted Regression Trees (BRT) classification method, which was successful in detecting seeps.

In the study area, Gachsaran cap rock underlies Mishan Formation so that hydrocarbon seepages and their alterations might occur in Mishan Formation. Gacharan formation is dominated by evaporite while Mishan Formation is dominated by marl and the marly limestone (Salati, 2014). Moreover, many oil beds in Zagros thrust belt have gas caps. Therefore, it is meaningful for oil and gas exploration to study gas seepages on the surface of the marly limestone in Zagros. In the study area (figure1.1), field work has been done and samples have been collected around active hydrocarbon seepages (Salati, 2014). Field samples and known seepages can be used to evaluate hydrocarbon alteration maps.



(a) (b)

Figure1.1 Study area. (a)is used in chapter5, while (b) is the test area that is used in chapter 3 and 4. The red circle in (b) shows the location of field samples

1.4.2. Datasets

1.4.2.1. Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER)

ASTER has 14 spectral bands, and the wavelengths and the bandwidths of the bands make it a very useful sensor for geology (Gomez et al., 2005). VNIR and SWIR bands allow to map ferrous and ferric minerals, clays, carbonates and sulfur. These minerals are associated with hydrocarbon seepages(Shi et al., 2012).

Table 1.1 ASTER bands information

Band	Label	Wavelength(μm)	Resolution(m)
B1	VNIR	0.520–0.600	15
B2		0.630–0.690	15
B3		0.760–0.860	15
B4		0.760–0.860	15
B4	SWIR	1.600–1.700	30
B5		2.145–2.185	30
B6		2.185–2.225	30
B7		2.235–2.285	30
B8		2.295–2.365	30
B9		2.360–2.430	30
B10	TIR	8.125–8.475	90
B11		8.475–8.825	90
B12		8.925–9.275	90
B13		10.250–10.950	90
B14		10.950–11.650	90

1.4.2.2. WorldView-2

WorldView-2 has a panchromatic imagery (resolution: 0.46 m; wavelength: 0.450-0.800μm), and eight-band multispectral imagery (resolution: 1.84 m). The size of Hydrocarbon seepages and their alteration is small, and World View-2 is a very high spatial resolution satellite. The eight VNIR bands is useful to detect iron and sulfur. Unfortunately, World View-2 doesn't have SWIR bands for mapping some seepages related minerals like clays and carbonates. Consequently, World View-2 can improve the hydrocarbon alteration zones and accuracy of hydrocarbon alteration mapping (Salati, 2014).

Table 1.2 WorldView-2 bands information

Band	Wavelength(μm)
Coastal Blue	0.400 - 0.450
Blue	0.450 - 0.510
Green	0.510 - 0.580
Yellow	0.585 - 0.625
Red	0.630 - 0.690
Red Edge	0.705 - 0.745
Near Infrared (NIR1)	0.770 - 0.895
Near Infrared (NIR2)	0.860 - 1.040

1.5. Research objectives

The main objective is to compare the performance of a number of classifiers and apply them to ASTER and WorldView-2 imagery for detecting hydrocarbon alteration in the Dezful Embayment.

Sub-objectives:

1. Spectral analysis of alteration associated with gas-induced seepages in the marly limestone formation in the Dezful Embayment.
2. Map hydrocarbon seeps related alterations in local scale by Aster and World View imagery.
3. Analyze and compare performance of a number of classifiers for hydrocarbon seepages detection.
4. Extrapolate hydrocarbon alterations to map at regional scale using the classifier which has the best performance.

1.6. Research questions

1. Which minerals associated with hydrocarbon seepages should be chosen for target mineral mapping?
2. How to select the best classifier?
3. Do ensemble classifiers work better than single classifiers in the terms of hydrocarbon alteration detection?
4. How do ensemble classifiers perform with a limited amount of field data?
5. Which classifier has the best performance? Why does this classifier perform so well for hydrocarbon alteration detection?
1. Comparing the accuracy of hydrocarbon seepages detection, could ASTER imagery be used to map hydrocarbon seeps related alterations if WorldView-2 imagery is not available?

1.7. Methodology

This research involves 6 main stages shown in figure 1.2:

1. Pre-processing of the ASTER and WV-2 images. Operations such as cross-talk correction, rotation, layer stacking, geometric correction, FLAASH Model were performed in ENVI 5.2 and

CrossTalk3 software. For ASTER data, we resampled the six 30 m SWIR bands to 15 m VNIR spatial resolution.

2. Analyze spectra of the altered and unaltered samples that were measured using the ASD instrument, in order to indicate which minerals are associated with gas-induced alterations. The laboratory spectra were resampled to the ASTER bands. Compare the spectra of field samples with the spectra picked from pixels of ASTER and World View-2 imagery that have same location as the field samples. Compare with the USGS spectral library; the spectra based on the range of absorption features, depth and pattern of spectral features were visually studied to identify minerals. This step should be very carefully done to decide which minerals would be chosen for classification. These minerals would influence the accuracy of hydrocarbon alteration detection.

3. Based on the altered mineral assemblages, suitable band ratios, relative absorption band depth and false color composite of World View-2 and ASTER were selected and produced to give insight in altered minerals.

4. To classify ASTER, World View-2. There are three problems: (1) this study will use multispectral data in which it is easy to make mistakes in endmember extraction beause of the low spectral resolution; (2) the amount of field data is limited in the study area; (3) the area of the hydrocarbon alterations is relatively small. Thus, the selection of the training sets and the classifiers should be done carefully.

   Meanwhile, the classifier selection is cumbersome. As the research time is limited, how to choose good classifiers from hundreds of classifiers is a big problem. Based on three problems mentioned in stage 4, there are three criteria for classifier selection: (1) these classifiers should be insensitive to outliers; (2) these classifiers should have good performance when the training sets are small; (3) these classifiers should be pixel-based or sub-pixel analysis; (4) overfitting should not be common in these classifiers. In this research, SVM, RF and GBRT were selected, and these classifiers were all available in the software R (version 3.2.0).

   Further, after testing the optimal parameters for each classifier, apply these classifiers and two training sets to ASTER imagery. Then hydrocarbon alteration maps were produced.

5. To validate these classifiers (SVM, RF and GBRT) and measure the quality of the alteration map, a test set is used that consists of ground truth data. The evaluation involves two aspects: (1) the overall accuracy of confusion matrix that is calculated using the test set; (2) comparing the spatial alteration patterns of classification result with a false-color image produced in stage 3.

6. According to the validation result in stage 5, advantages and disadvantages of these classifiers were concluded and the best classifier will be identified. Apply the best classifier to produce hydrocarbon alteration map at regional scale (figure 1.1a).

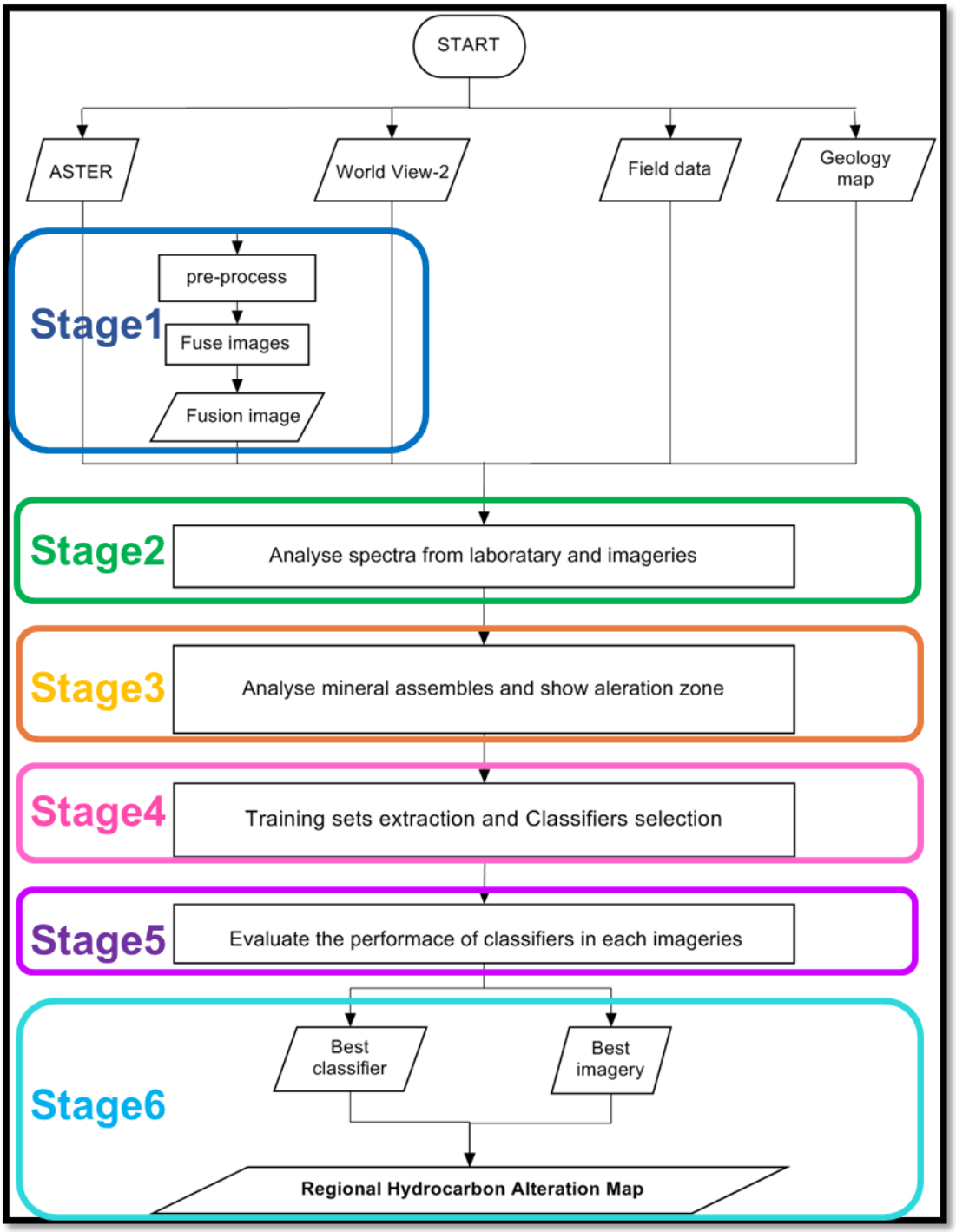


Figure 1.2 Flow chart of methodology

2. LABORATORY SPECTRAL ANALYSIS OF FIELD SAMPLES

2.1. Laboratory spectral analysis of field samples

Field samples and laboratory spectra were collected and acquired with the ASD instrument by Ms. Salati in 2012 (Salati, 2014). In this study, 20 samples which were collected in the marly limestone in the Dezful Embayment are chosen and the locations are shown in figure 2.1. These samples contain both altered and unaltered rocks and show the upper boundary of gas-induced alteration.

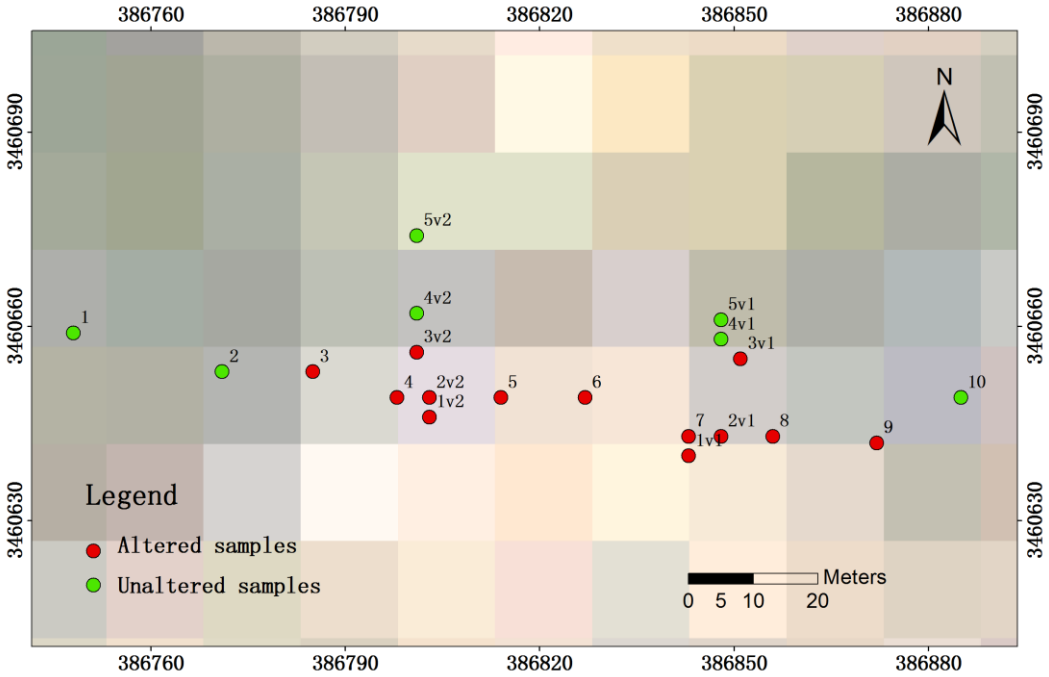


Figure 2.1 Locations of field samples in the Dezful Embayment (see red circle in figure 1.1(b)). Samples (3, 4, 5, 6, 7, 8, 9, 1v1, 1v2, 2v1, 2v2, 3v1 and 3v2) of red color are gas-induced altered while samples (1, 2, 10, 4v1, 4v2, 5v1 and 5v2) of green color are in unaltered areas.

When hydrocarbon escaped, sulfur minerals in the cap rock reduced to hydrogen sulfide. Hydrogen sulfide reacted with calcite of the limestone to produce gypsum and native sulfur (Salati, 2014). Based on the result of geochemical analysis, the altered samples have high concentration of sulfur and gypsum, but low concentrations of calcite. Meanwhile, the unaltered samples have high concentrations of calcite and lack gypsum and sulfur (Salati, 2014).

Laboratory reflectance spectra of field samples were acquired with ASD FieldSpec and spectral library were created by Salati (2014). We will also use the USGS digital spectral library, which was assembled by Clark et al. (2007). As shown in figure 2.2, the USGS digital spectral library shows: (a) the absorption features of gypsum are at 1.45μm, 1.75μm, 1.94μm, 2.21μm and 2.42μm. (b) the absorption features of illite

are at 1.41μm, 1.91μm, 2.21μm, 2.34μm and 2.44μm. (c) the absorption features of smectite are at 1.41μm, 1.91μm, 2.21μm. (d) the typical absorption feature of calcite is 2.34μm.

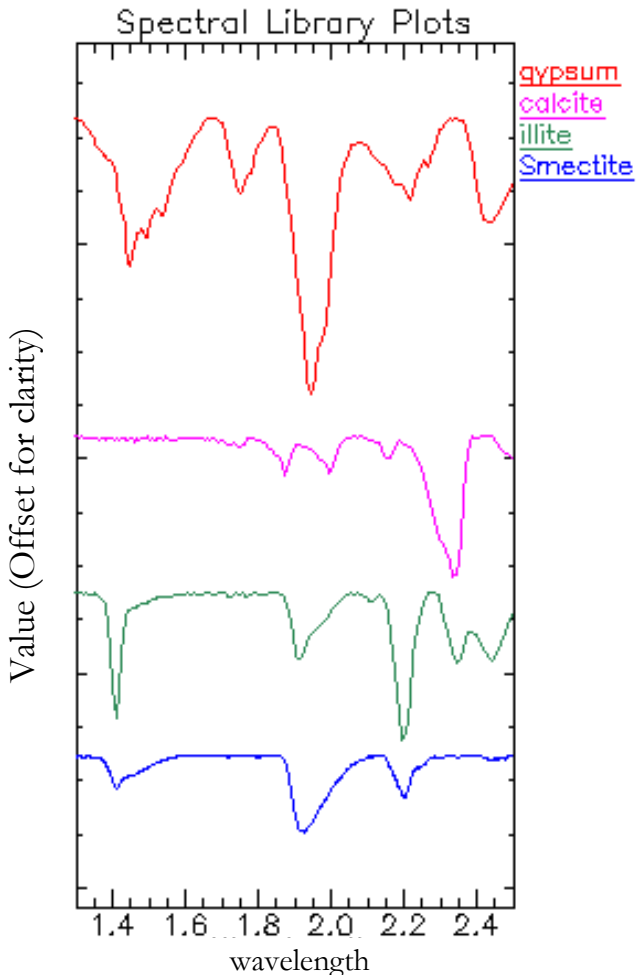


Figure 2.2 Gypsum, calcite, illite and smectite spectra from USGS spectral library (Clark et al., 2007)

Laboratory spectra are compared with the USGS digital spectral library, and the analysis of the absorption features of minerals is used to identify the mineralogy of laboratory spectra. Based on the geochemical result(Salati, 2014) and the USGS digital spectral library, the mineralogy of the field samples is shown as follow.

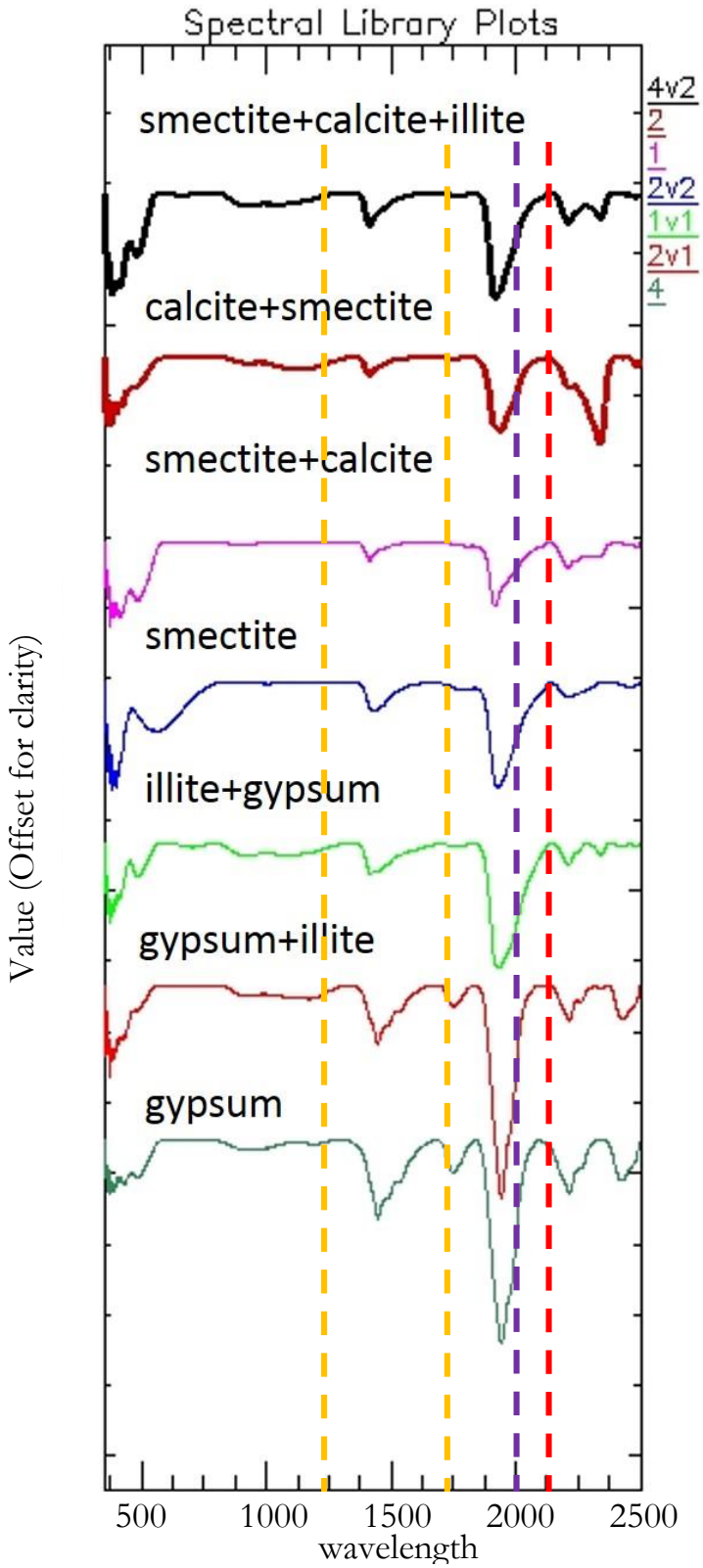


Figure 2.3 Laboratory spectra of selected samples after using continuum removed and the mineralogy of these spectra. Sample 4v2, 2 and 1 show the spectra of unaltered samples, while others show the spectra of altered samples. Two dotted orange lines draw the wavelength of 1.4μm and 1.9μm (absorption feature of water). The dotted purple line draws the wavelength of 2.2μm (absorption feature of AlOH) and dotted

red line highlight the wavelength of 2.35μm (absorption feature of calcite). The locations of these samples are shown in figure 2.1.

Table 2.1 Mineralogy of field samples

	Mineralogy	Sample No.
altered	Gypsum	4,5,6,7,9,1v2,3v2
	gypsum+illite	3,8,2v1,3v1
	illite+gypsum	1v1
	Smectite	2v2
unaltered	smectite+calcite	1
	calcite+smectite	2,4v1,5v1
	smectite+calcite+illite	4v2,5v2,10

For understanding of how the minerals changed when the rocks were altered by gas, we compare the absorption feature and mineralogy of both altered and unaltered samples. As figure 2.3 and table 2.1 showed, unaltered samples contain calcite and clays. Meanwhile gas-induced altered samples contain gypsum and clays. Moreover, two water absorption features of unaltered samples, 1.4μm and 1.9μm, are shallower than altered samples.

2.2. Conclusion

This chapter utilized a laboratory spectra approach to analyze the absorption features and to identify mineral assemblages in altered and unaltered field samples respectively. The results of this chapter demonstrated that gypsum and clays are dominated in gas-induced altered samples, while unaltered samples contain calcite and clays predominately in the marly limestone formation. And the altered samples show a deeper absorption in the water absorption features of 1.4 and 1.9 micron.

3. ASTER AND WORLDVIEW-2 DATA PROCESSING

Laboratory spectra of altered and unaltered samples were analyzed in chapter 2. Mineral assemblages were identified and gypsum was suggested as the indicator mineral in alterations. In this chapter, ASTER and worldview-2 data were processed. In order to identify the minerals in ASTER and WV-2, we resample the laboratory spectra to the ASTER and WV-2 scenes and compare these spectra with the spectra of pixels from ASTER and WV-2 respectively. Moreover, the result of chapter 2 is regarded as prior knowledge for selecting band ratios and relative absorption band depths. The method and result of satellite data processing are shown in the following sections. Based on the result of ASTER and WorldView-2, gas-induced alterations in the study area are identified. Furthermore, the alteration map can be used to guide choosing the training set for classifiers.

3.1. ASTER data processing

3.1.1. ASTER data pre-processing

For this research, the ASTER level 1B data was processed. The ENVI software contains a correction function for ASTER, and radiance calibration is automatically applied when ASTER level 1B data is opened in the ENVI software. Then, crosstalk, rotation, layer stacking, FLAASH MODEL, quick statistic, band math and dark subtraction were done in the ENVI software for atmospheric and geometric correction.

Photons leak from one detector to another, and will cause a radiance offset. Because the solar output of band 4 is significantly higher than other bands and detectors of band 5 and band 9 are nearest to the detectors of band 4, this means that band 5 and band 9 are easier affected by crosstalk. However, the photons of band 4 may leak to all SWIR bands so that the crosstalk of ASTER affects the accuracy of SWIR bands(Alimohammadi, et al., 2015). Referring to the ASTER Mineral Index Processing Manual(Kalinowski, 2004), the ERSDAC Crosstalk software was used to correct crosstalk of all the ASTER SWIR bands used in this study.

The FLAASH Model is an atmospheric correction method for VNIR and SWIR bands in ENVI. And it can be used in both hyperspectral and multispectral data. The FLAASH Model was developed based on MODTRAN4 calculations(Adler-Golden et al., 1999). Tian et al. (2008) indicated this model offered the best calibration for hyperspectral data radiometric calibration in their research. The input file should be a radiometrically calibrated radiance image. The sensor information, image acquired date, elevation and location must be filled in. For this purpose the elevation was obtained from google earth, and other information was obtained from the metadata. Moreover, an atmospheric aerosol model must be selected. As the tropospheric aerosol model is used for open ground with a small-particle component of the rural model, and the study area meets these conditions, the tropospheric aerosol model was chosen in this study(Visual & Solutions, 2009). The output is a reflectance image. Therefore, in this study, the FLAASH Model in ENVI software corrects ASTER imagery based on above steps.

3.1.2. Analyze field data and ASTER data co-located with field data

To analyze the spectra information and compare the spectral difference between altered and unaltered samples in ASTER, laboratory spectra of field samples are resampled to the ASTER spectral resolution. Moreover, we pick spectra from ASTER pixels co-located with field samples. The newly constructed spectral library is created by using these ASTER spectra.

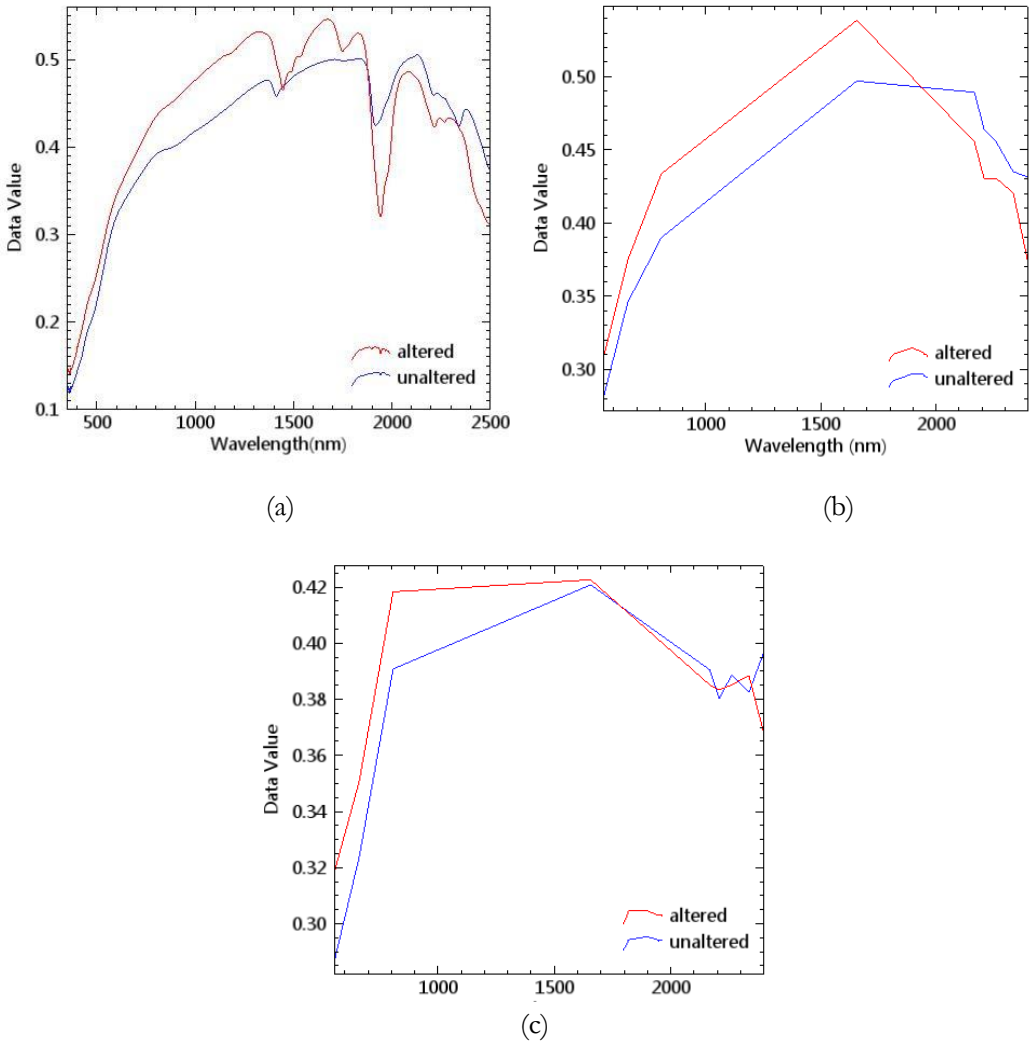


Figure 3.1 (a) Mean laboratory reflectance spectra of both altered and unaltered field samples, (b) mean spectra resampled to ASTER and (c) mean spectra of pixels from ASTER co-located with field samples.

As figure 3.1 shows, the unaltered spectrum has an absorption feature at 2.34μm which is a typical absorption feature of calcite, while the altered spectrum doesn't have this absorption feature. Both altered mean spectrum and unaltered mean spectrum have an obvious AlOH absorption feature at 2.2μm. Moreover, the reflectance of the altered samples is higher than the unaltered samples in VNIR bands. Therefore, we conclude that the result of spectra derived from ASTER imagery agree with the result of laboratory spectra.

3.1.3. Band ratios and False Color Composition images

ASTER has 14 spectral bands with wavelengths from 0.52μm to 11.65μm which means that ASTER doesn't provide a blue band. Although ASTER cannot produce true color images, the 14 spectral bands are useful to calculate mineral indexes. As hydrocarbons escape to surface, hydrocarbons oxidize and form a reducing and slight acid environment at the surface, which is associated with red bed bleaching (the conversion of Fe^{3+} to Fe^{2+}), clay minerals (the conversion of feldspar to clay minerals like kaolinite) and carbonates (the presence of Fe^{2+}-rich carbonates like siderite) (Lammoglia & de Souza Filho, 2013). Moreover, geochemical analysis (Salati, 2014) and laboratory spectral of the field samples showed that the altered samples have a high concentration of sulfur and gypsum and low concentrations of calcite, while the unaltered samples have high concentrations of calcite and lack gypsum and sulfur. Therefore, the band ratios (Territory, 2012) listed in table 1 were selected with which band ratio images were made using ENVI band math. As table 3.1 shows, the ratio b2/b1 is used to indicate ferric iron, b5/b4 is used to indicate ferrous iron, the band depth (b5+b7)/b6 is used to indicate clays, (b7+b9)/b8 is used to strengthen calcite, and the formula b4/(b9+b6) is used to indicate gypsum.

Table 3.1 ASTER band ratios

Code name	Description	Ratio
Ferric iron index	Ferric oxide composition	B2/B1
Ferrous iron index	Ferrous silicate or carbonate	B5/B4
Clays index	Clays	(B5+B7)/B6
Calcite index	Calcite	(B7+B9)/B8
Gypsum index	Gypsum	B4/(B9+B6)
NDVI	Vegetation	(B3-B2)/(B3+B2)
Ferrous iron index*	Ferrous oxide (Salati, 2014)	B3/B1

Box plots of above-mentioned indexes were calculated to compare ferric oxide index, ferrous index, clay index, calcite and gypsum between unaltered and altered field samples. As shown in figure 3.2, the value of the Ferric oxide index (b2/b1), Ferrous iron index (b5/b4) and Calcite ((b7+b9)/b8) in unaltered field samples are higher than altered field samples, while the value of AlOH group content ((b5+b7)/b6) and Gypsum (b4/(b9+b6)) in unaltered field samples are lower than altered field samples. From prior knowledge we knew that the altered area should have a high concentration of ferrous iron, clays and gypsum. But figure 3.2 shows that ferrous iron in altered samples have a lower index value than unaltered samples. This phenomenon is due to the limitation of these band ratios. Band5/band4 is used to identify ferrous iron in silicates and carbonates, and it is hard to use Band5/band4 to detect ferrous iron if it is associated with oxide and sulphate. The altered marly limestone formation contains less carbonates. This explains why the value of band5/band4 in altered samples is lower than in unaltered samples.

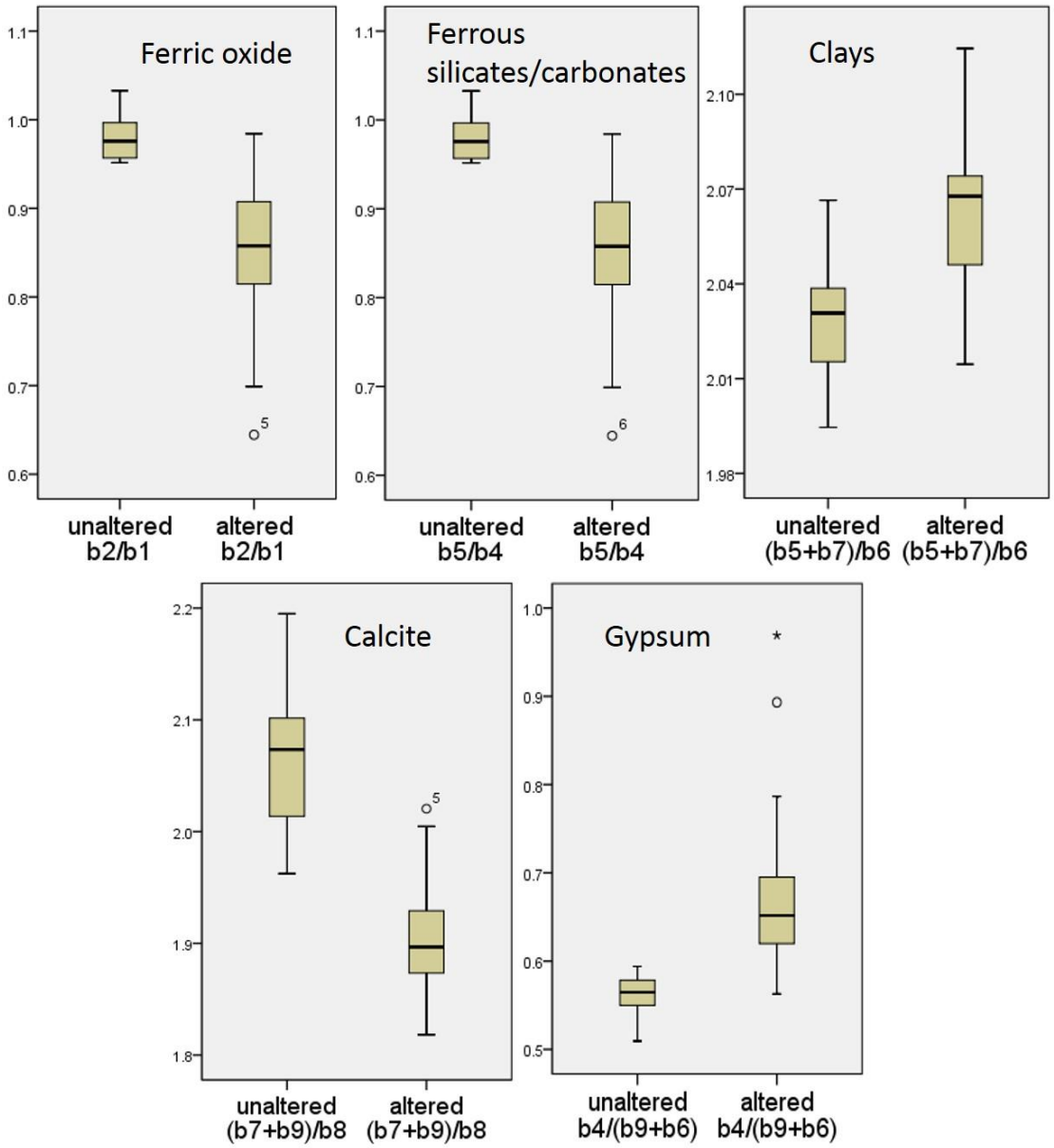


Figure 3.2 Box plots of band ratios show the comparison of the ferric oxide index, ferrous index, clay index, calcite index and gypsum index between unaltered and altered field samples. Plots show, generally, clays and gypsum indexes in the altered samples are higher than the unaltered samples, while ferric oxide, ferrous and calcite indexes in the altered samples are lower than the unaltered samples.

Table 3.2 False color composite images

Code name	R	G	B
FCC1	3N	2	1
FCC2	gypsum index	calcite index	clay index
FCC3	gypsum index	calcite index	ferric index

Therefore, the altered area in our study area should contain high a concentration of gypsum and clays, and a low concentration of calcite and ferric iron. Next, three false color composition images (see table 3.2) were made to show difference between gas-induced alteration and unaltered areas in ASTER data.

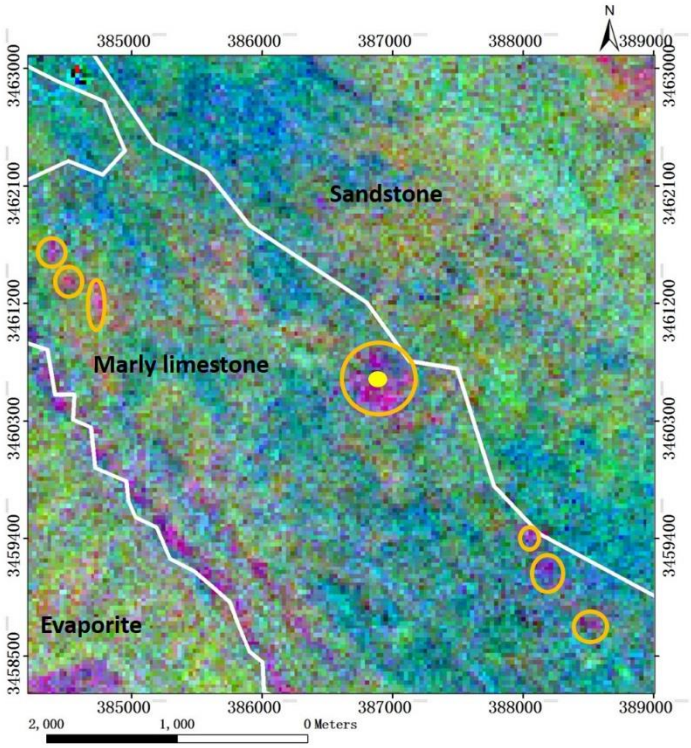


Figure 3.3 FCC2 of gypsum index (red) - calcite index (green) - clay index (blue) obtained from the ASTER image. The yellow dot represents the position of field samples. Orange circles represent the gas-induced alterations.

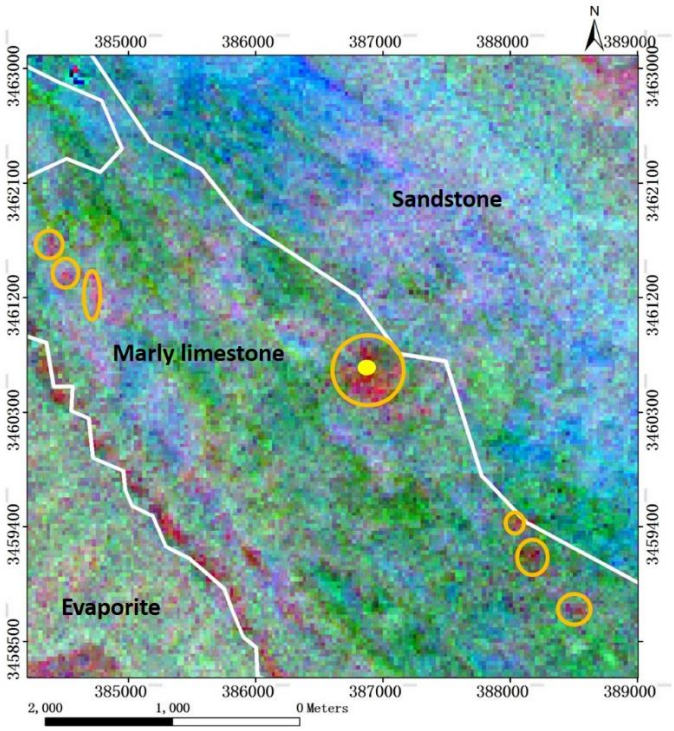


Figure 3.4 FCC3 of gypsum index (red) - calcite index (green) – ferric iron index (blue) obtained from the ASTER image. The yellow dot represents the position of field samples. Orange circles represented the gas-induced alterations.

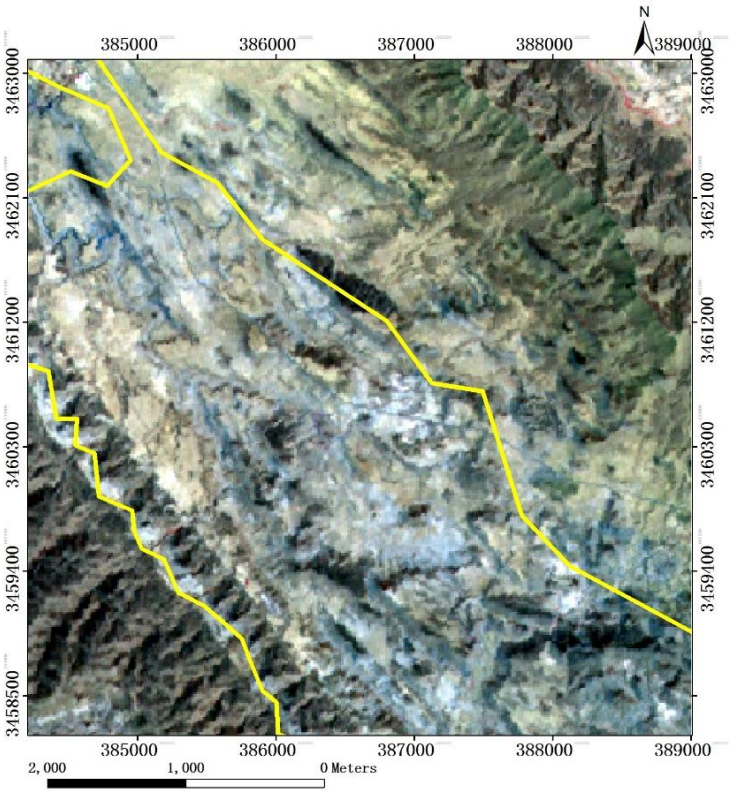


Figure 3.5 ASTER image in band3N (red) – band2 (green) – band1 (blue) false color composite (vegetation shows red color)

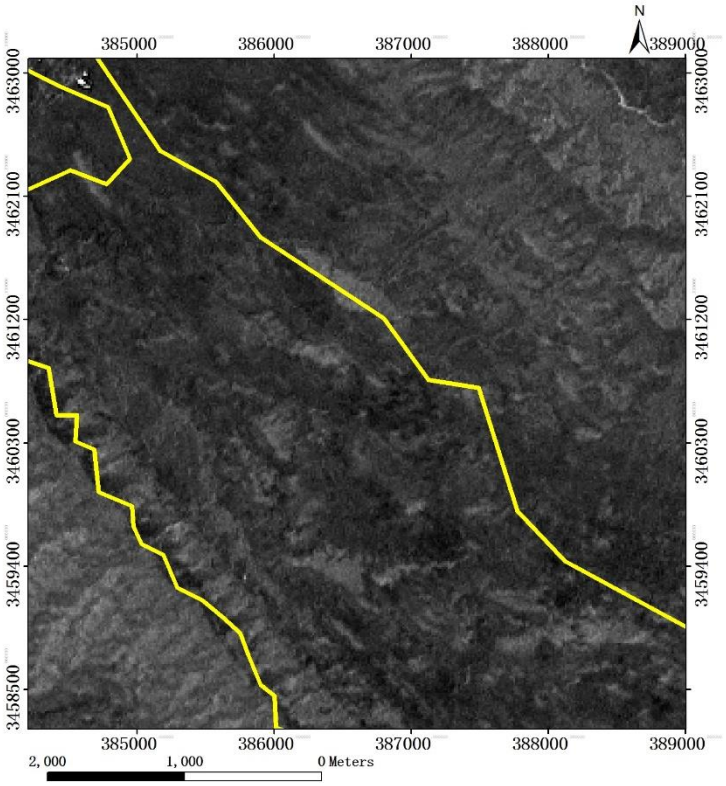


Figure 3.6 NDVI image (vegetation shows white color) obtained from the ASTER image

False color composite (FCC) images were created using the index images. The FCCs can be used to indicate alteration zones. As figure 3.3 shows, a false color composite of gypsum index (red), calcite index (green) and clay index (blue) was created. Since altered areas in our study area should contain a high concentration of gypsum and clays and a low concentration of calcite and ferric iron, the purplish area(s) in the marly limestone formation in figure 3.3 might be gas-induced alteration. In figure 3.4, a false color composite of gypsum index (red), calcite index (green) and ferric index (blue) was created. The reddish area in the marly limestone formation in figure 3.4 might be gas-induced alterations. In figure 3.4, there are reddish areas along the border of the evaporite formation and the marly limestone formation. However, these reddish areas are not gas-induced alterations, because gypsum is a common mineral in evaporite formation (Tangestani & Validabadi, 2014). Thus, gypsum along the western boundary of the marly limestone formation might be transported from the evaporite formation. Moreover, since vegetation has high reflectance in wavelength of 0.8μm, vegetation should show red color in FCC1 (see figure 5). Figure 3.5 and figure 3.6 illustrate the study area lacks vegetation which means that vegetation will not have a major effect on the image analysis. Hydrocarbon alteration zones were indicated by orange circles.

3.2. WorldView-2 data processing

3.2.1. WorldView-2 data preprocessing

For this research, WorldView-2 data which was orthorectified image was acquired on 24 August 2011 under cloud-free conditions, and was geometrically corrected by DigitalGlobe. Gains and offsets were applied and recorded in the metadata which was used to convert digital number to radiance by ENVI software (Robinson et al., 2016). The image was atmospherically corrected and radiometrically calibrated to reflectance using the FLAASH model (Mutanga, Adam, & Cho, 2012; Whiteside & Bartolo, 2015).

3.2.2. Analyze field data and WorldView-2 data co-located with field data

Here we resample the laboratory spectra of field samples to WorldView-2 spectral resolution, pick spectra from WorldView-2 pixels co-located with field samples and create a new spectral library by using image spectra. The mean spectra of altered and unaltered samples are shown in figure 3.7.

As figure 3.7 shows the unaltered spectrum has a slightly deeper absorption feature in 0.48μm than altered spectrum. Moreover, the overall value of the altered spectrum is higher than the unaltered spectrum. Therefore, altered areas are brighter than unaltered areas.

3.2.3. Band ratios and False Color Composition images

When hydrocarbon escaped, sulfur minerals in cap rock reduced to hydrogen sulfide. Hydrogen sulfide reacted with calcite from the limestone to produce gypsum and native sulfur(Salati, et al., 2014). WorldView-2 has 8 spectral bands and the wavelengths range from 0.40μm to 1.04μm. Visible and near-infrared bands can be used to detect sulfur and iron minerals (Horgan, et al., 2014). Sulfur has an absorption feature in about 0.4μm so that it can be enhanced by dividing the right shoulder (about 0.56 μm) by the absorption feature. Thus, band2/band1 was used to indicate sulfur in this research. Band5/band3 (red band/ blue band) was used to indicate ferric iron (Kalinowski, 2004). Band3*band4/band2 (green*yellow/blue) was used to indicate iron. Therefore, the band ratios listed in table 3.3 were selected and using these band ratio images were made using ENVI band math.

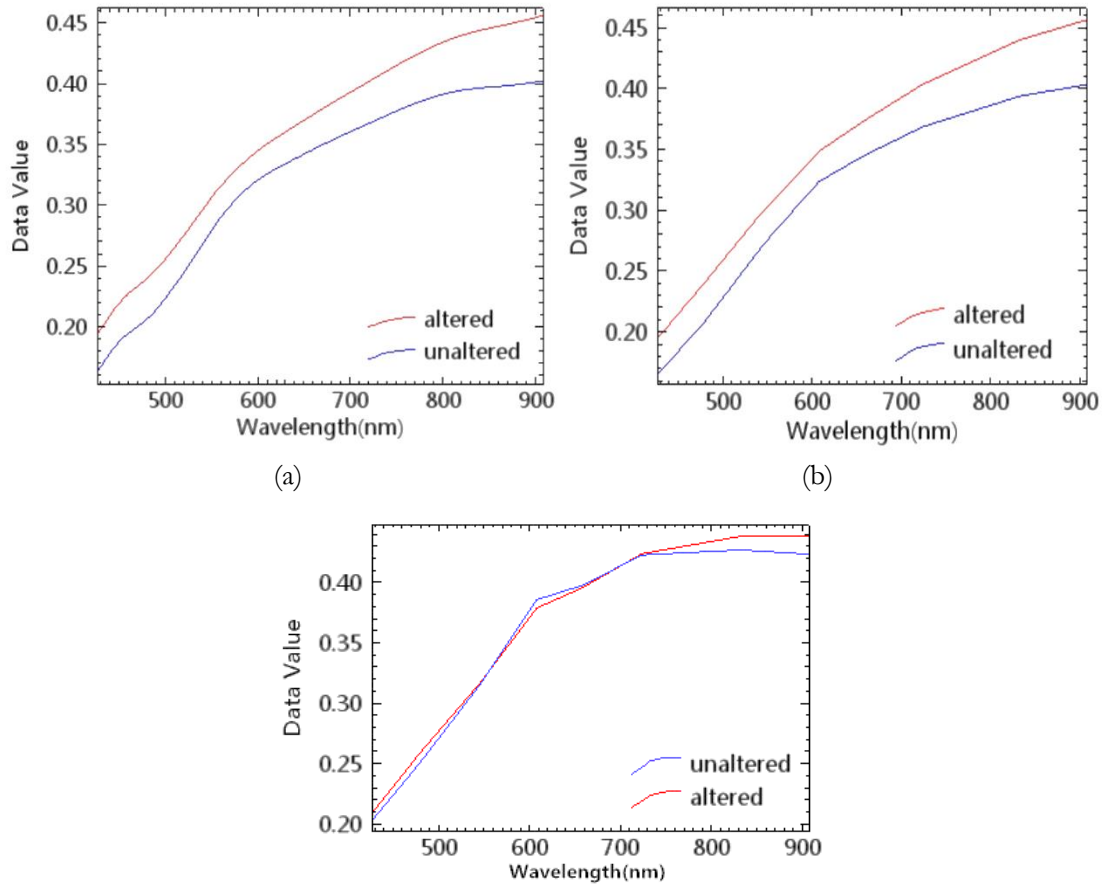


Figure 3.7 (a) Mean laboratory reflectance spectra of both altered and unaltered field samples, (b) mean spectra resampled to WorldView-2 and (c) mean spectra of pixels from WorldView-2 co-located with field samples.

Table 3.3 WorldView-2 band ratios

Code name	Description	Ratio
Sulfur index	Sulfur	B2/B1
Ferric iron index	Ferric oxide composition	B5/B3
Iron index	Iron oxide	(B3*B4)/B2
Ferrous iron index*	Ferrous iron (Salati, 2014)	(B3+B5)/B1
Ferric iron index*	Ferric iron (Salati, 2014)	B4/B7

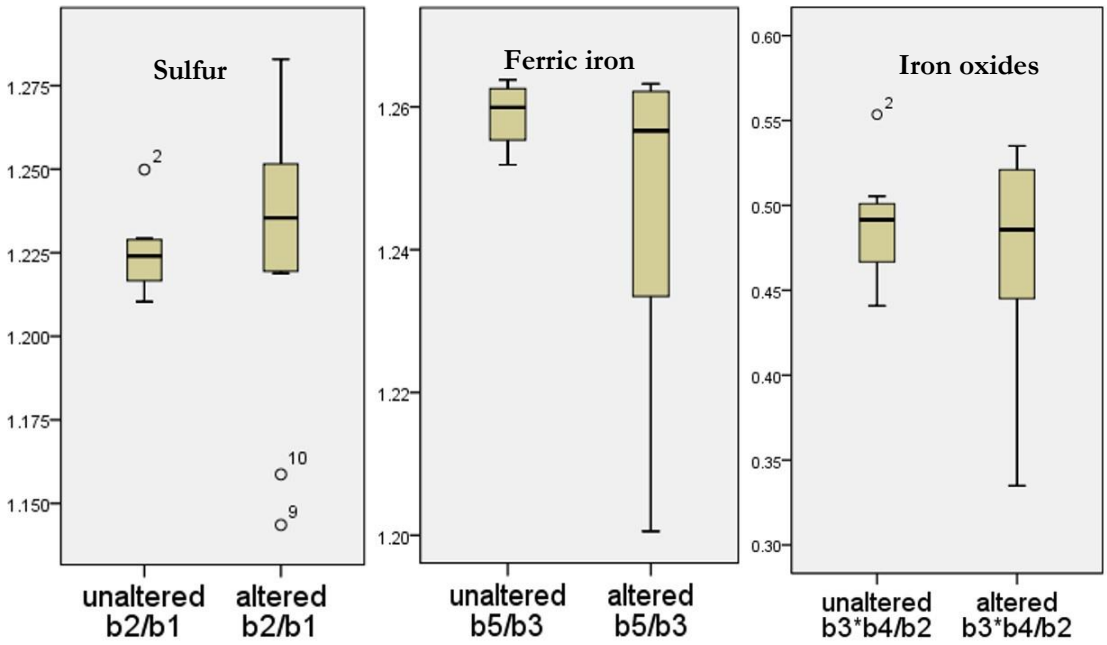


Figure 3.8 Box plots of band ratios showed comparison of sulfur, ferric index, iron oxides between unaltered and altered field samples. Plots show sulfur index are higher in the altered samples than the unaltered samples.

Box plots of the above-mentioned band ratios were made to compare sulfur and iron between unaltered and altered field samples. As shown in figure 3.8, the value of the Ferric iron Index (b5/b3) in unaltered field samples is higher than in altered field samples, while the value of sulfur index (b2/b1) in unaltered field samples are lower than in altered field samples. From prior knowledge we knew that altered areas should have a high concentration of ferrous iron, clays and gypsum. But ferrous iron, clays and gypsum cannot be detected by WorldView-2. This is due is due to the restricted wavelength range of WorldView-2 data. Ferrous minerals mainly have an absorption feature at 1μm, and the spectral characteristics of ferrous minerals have a strong relation with their composition and crystal structure. Moreover, clays and gypsum have obvious spectral characteristics in shortwave infrared bands. Thus, WorldView-2 data cannot be used to detect ferrous minerals, clays and gypsum.

From the above, it is clear that the altered area in our study area should contain a higher concentration of sulfur, and less ferric iron. In addition to this, three false color composition images (see table 3.4) were made to show the difference between gas-induced alterations and unaltered areas in WorldView-2 data.

Table 3.4 True and false color composite images

Code name	R	G	B
TCC1	5	3	2
FCC4	Sulfur index	Iron index	Ferric index

Figure 3.7 illustrates that the reflectance of altered samples is higher than in unaltered samples so that the altered area in true color composite images are brighter than the unaltered area. Gas-induced alterations exist in the whitish areas (see figure 3.9) in the marly limestone formation. In figure 3.10, a false color composite of sulfur index (red), iron index (green) and ferric iron index (blue) was created. Gas-induced alterations exist in the brightly yellowish areas (see figure 3.10) in the marly limestone formation. Alteration zones were circled by red lines in this figure.

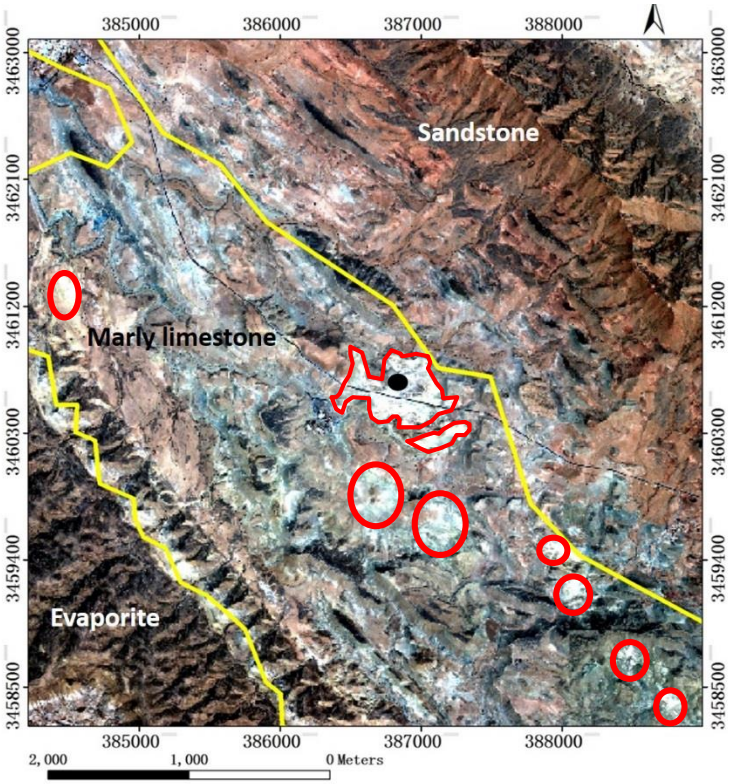


Figure 3.9 True color composite WorldView-2 image. The black dot represents the position of the field samples. Red circles represent gas-induced alterations.

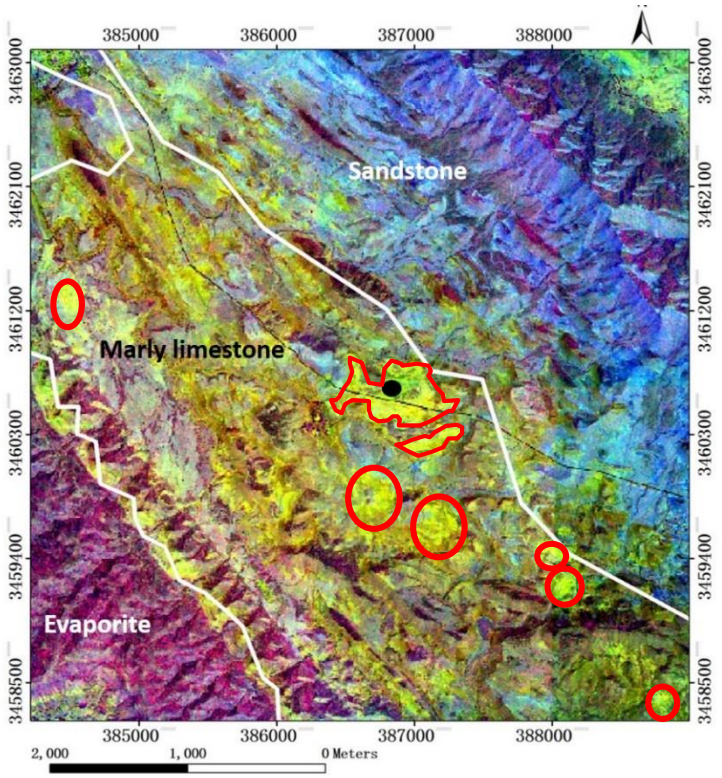


Figure 3.10 FCC4 of sulfur index (red) - iron index (green) – ferric iron index (blue) obtained from WorldView-2 image. The black dot represents the position of field samples. Red circles represent gas-induced alterations.

3.3. Conclusion

This chapter took advantage of the knowledge-based approach (band ratios and relatively absorption band depths) carried out earlier to detect gas-induced alterations in the study area. The results of this chapter indicated that the FLAASH model could successfully be used in ASTER and WV-2 pre-processing. Moreover, in the marly limestone formation, the altered areas are rich in sulfur, ferrous iron, gypsum and clays, while they have a low concentration of carbonates and ferric iron. The gas-induced alteration maps (figure 3.3, 3.4 and 3.10) were produced by using ASTER and WorldView-2 imagery. Since WorldView-2 imagery can only detect sulfur and iron minerals instead of all the alteration indicator-minerals (gypsum and calcite), the alteration map produced by WorldView-2 data shows less details than the map obtained from the ASTER data. Thus we conclude that although WorldView-2 data has high spatial resolution and more VNIR bands, it cannot replace ASTER for mineral mapping and hydrocarbon seeps alteration detecting.

4. IMAGE CLASSIFICATION

Gas-induced alteration maps have been produced by ASTER and WV-2 imagery in chapter 3. Although WV-2 data improves the spatial accuracy of ASTER data, it cannot replace ASTER in mineral mapping field due to the lack of SWIR bands. To test and improve the alteration maps produced in chapter 3, three classifiers will be used to do image classification for detecting hydrocarbon seepages alterations in this chapter.

This chapter mainly describes training set extraction, and the SVM, RF, GBRT, classifiers validation and comparison. Two different sizes of training sets are selected to test the amount of samples needed for training the classifiers. Training data is selected by spectral and spatial analysis using the prior knowledge from chapter 2 and 3. Furthermore, the main parameters of each classifier are tested. With the optimal parameters, classification results are produced. To prevent classifiers focusing too much on imagery intensity value, an ASTER image without the PC_1 component is constructed by deleting first component of PCA. By comparing the learning ability of the three classifiers for different training sets and images, results are validated regarding computational and pattern aspects. Results show that all these three classifiers are successful for this study. Of the three, SVM is the most stable classifier for detecting hydrocarbon seepages. SVM also works well with small training data sets. It is suggested to apply GBRT to test and improve the classification result of SVM.

4.1. Training set extraction

In this study, two training sets are extracted based on ASTER imagery. Endmembers of both training sets are chosen by ROIs from pixels of the ASTER imagery. And the two training sets belong to two classes respectively: the altered class and the unaltered class.

There are three main steps for endmember extraction. Firstly, the spatial–spectral endmember extraction tool (SSEE) is used to choose endmembers with unique spectral information (Rogge et al., 2007). Secondly, we compare the location of endmembers extracted by SSEE with gas-induced alteration maps (figure 3.3, 3.4 and 3.10) shown in chapter 3. Only endmembers located in alteration zones of alteration maps have the chance to be selected to the altered class. Lastly, compare the spectra of endmembers with spectra of ASTER data co-located with altered field data. Endmembers with the similar spectra as figure 4.1 are selected for the altered class, which is regarded as having altered minerals (gypsum, gypsum with clays and clays), while others are selected for the unaltered class.

Based on the spectral and spatial relation of endmembers, similar endmembers are left out manually. Finally, 40 endmembers are selected for the first training set (small training set). And 100 endmembers are selected for the second training set (large training set). These two training sets are used training the classifiers for finding alterations and test sensitivity of classifiers for size of training sets. The location of the endmembers is shown in figure 4.2.

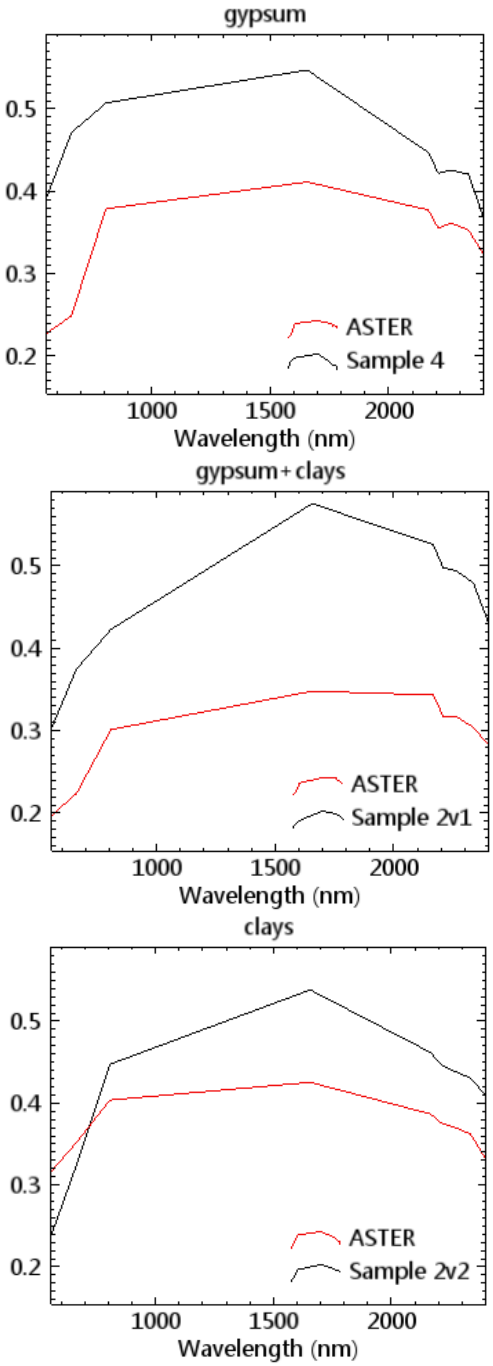
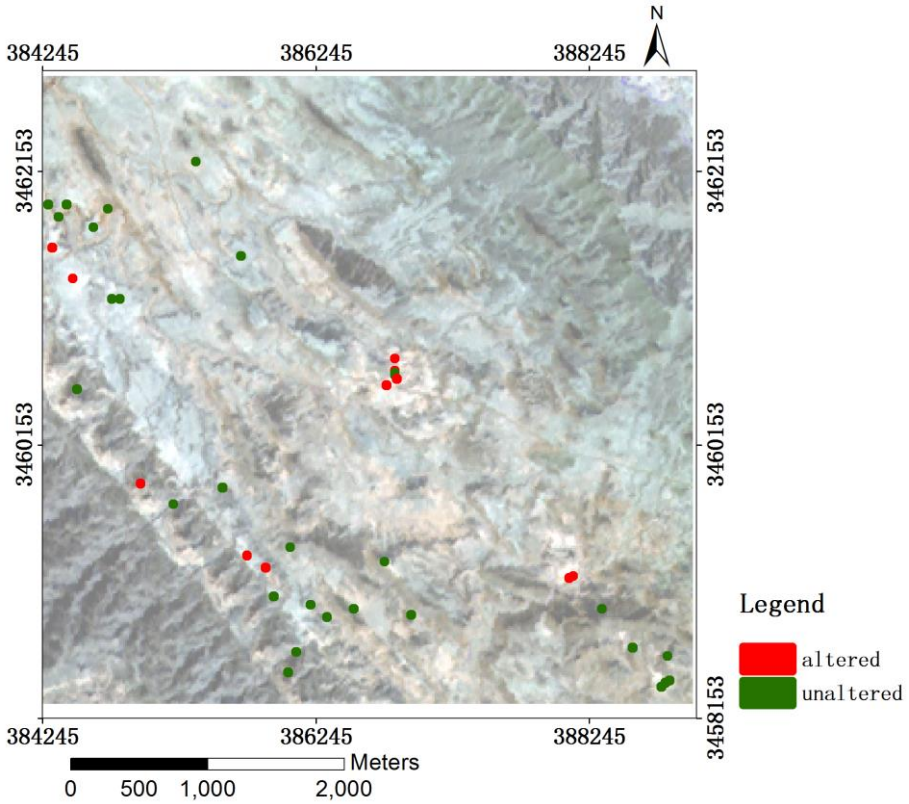
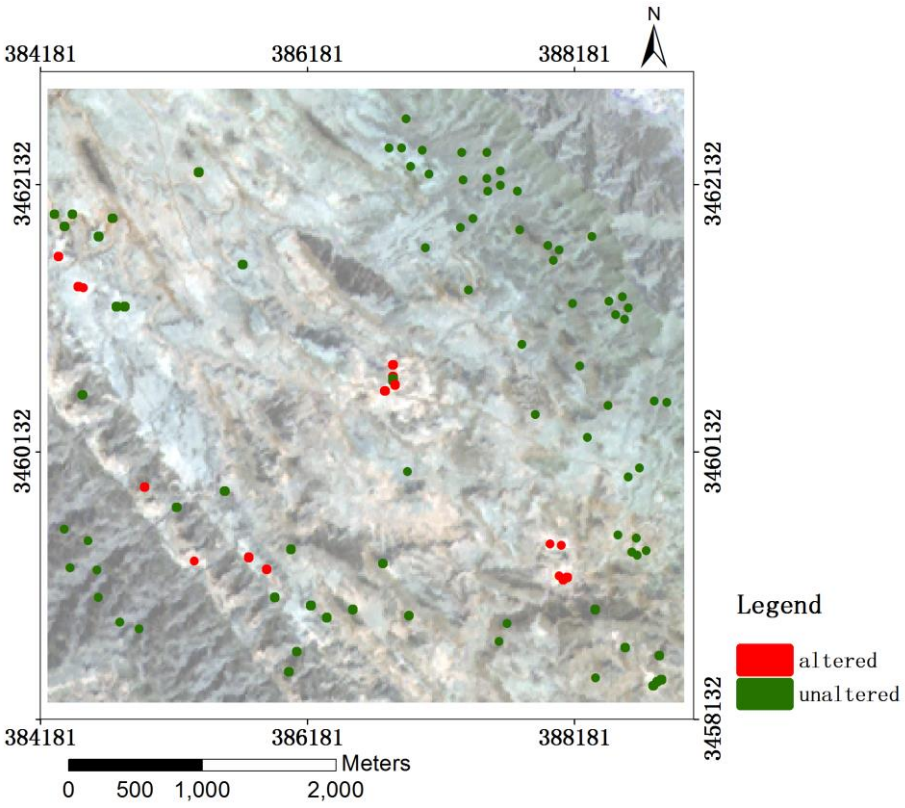


Figure 4.1 Comparison of the spectra of endmembers with spectra of ASTER data co-located with altered field data. Black lines show the spectra of ASTER data co-located with altered field data, while red lines show the spectra of endmembers.



(a)



(b)

Figure 4.2 (a) the location of small training set, (b) the location of large training set

4.2. Supported Vector Machine

4.2.1. Introduction

The Support Vector Machines (SVM), supervised machine learning method for classification, is based on the "Statistical Learning Theory". SVM builds one or more high-dimensional hyperplanes to classify training data, and the hyperplane forms the classification margin. In the other words, the distance from the optimal classification margin to the training data should be as far as possible. Combine SVM and ensemble algorithm, the binary SVM classifier can be extended to multi-classes classifier. Based on the number of classes, this method builds N*(N-1)/2 machines. For each pixel, these machines will vote for a class and the pixel will be labelled by the class which has most votes (Huang et al., 2002). This method is called one against one, which has a good performance. Even the training set can have a small size (Pal & Mather, 2006).

As a popular machine learning method, SVM is successfully applied in many disciplines, such as chemistry, economics and geology (Ali Sebtosheikh & Salehi, 2015). Moreover, this classifier has a good performance in both multispectral and hyperspectral image classification (Ma et al., 2016; Pal & Mather, 2006). There are advantages in the following points (Cortes & Vapnik, 1995; Ma et al., 2016; Pal & Mather, 2006).

(1) Based on sound mathematics theory
(2) Learning result is robust
(3) Over-fitting is not common
(4) Not trapped in local minima
(5) Fewer parameters to consider
(6) Works well with fewer training samples (number of support vectors do not matter much).).

4.2.2. Software

In this study, the support vector machines classifier is built by using the 'kernlab' package (version 0.9-22) in R (version 3.2.0).

4.2.3. Major parameters

The major parameter for the SVM classifier is the kernel function. In the 'kernlab' package, 8 kinds of kernels can be selected: (1) the Gaussian RBF kernel, (2) the Polynomial kernel, (3) the Linear kernel, (4) the Hyperbolic tangent kernel, (5) the Laplacian kernel, (6) the Bessel kernel, (7) the ANOVA RBF kernel, (8) the Spline kernel. To select kernel, the first step is to analyze whether the data is linearly separable or not. Figure 4.3 shows the feature space of training data. All feature spaces were produced and it is observed that any two bands do not have a linear relationship with each other. Thus, linear kernel is not considered. Among kernels which can process linearly inseparable problem, the most widely used kernel is Gaussian RBF kernel(He, Liu, Deng, & Shen, 2016). It can apply to both small and large training sets, and it is also suitable to both high-dimension and low-dimension. Compared with polynomial kernel, RBF need less parameter which can reduce the complexity of models. Therefore, in this study, RBF kernel is chosen.

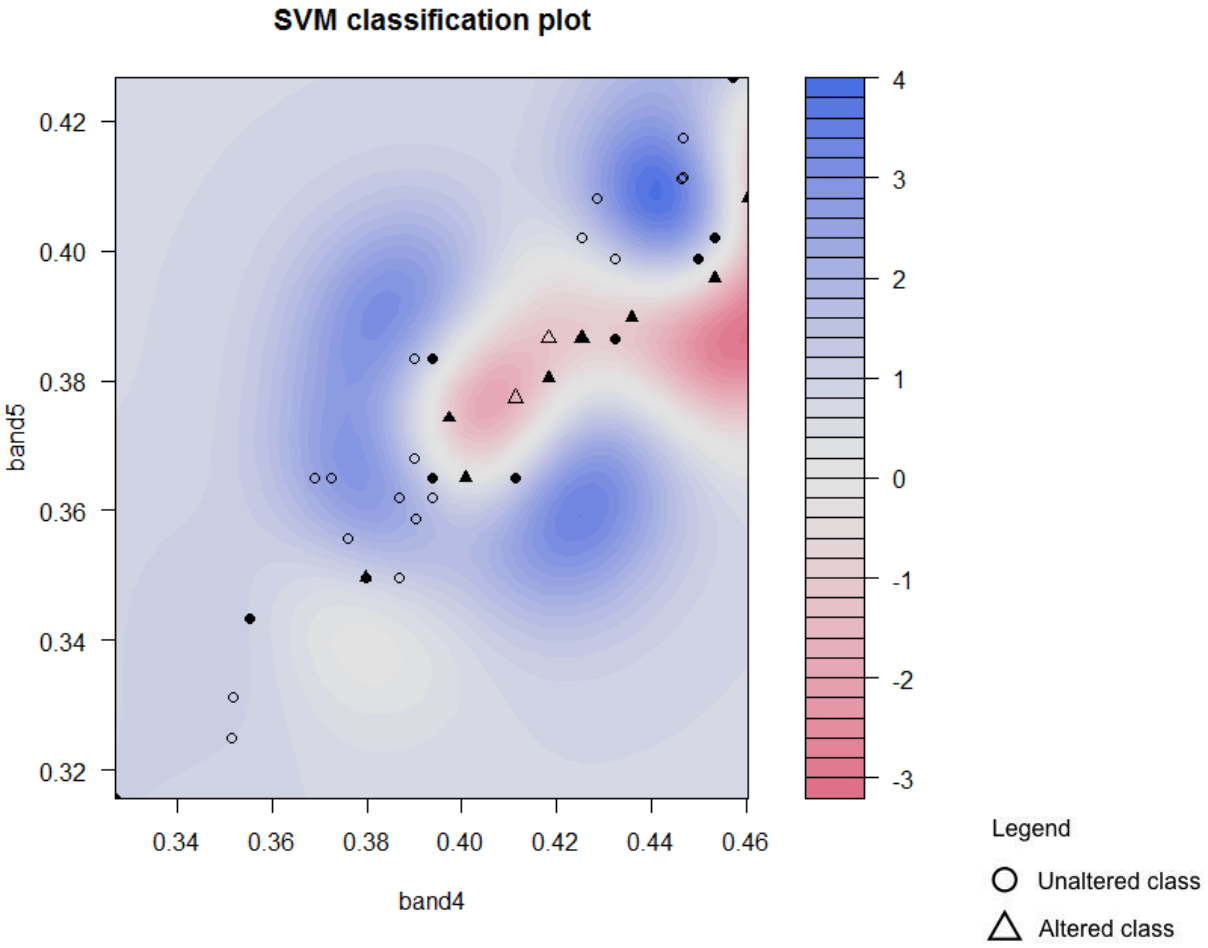


Figure 4.3 band4-band5 feature space of training data shows the relation between band4 and band5 is not linearly separable.

4.2.4. Principle component analysis method

Through comparison between figure 3.4 and figure 3.5, it appears that the alteration areas are significantly brighter in VNIR bands of ASTER. To prevent classifiers only paying attention on VNIR bands instead of SWIR bands, a PCA method is used to normalize the data. Principle component analysis is a spectral transformation method, which is used to compress data, enhance image, reduce noise and fuse image (Shahdoosti & Ghassemian, 2016). After PCA transforming, correlated components in the data are transformed to uncorrelated components. The first principle component has the highest information content and usually contains the intensity information. The spectral information is then present in the other principle components.

Gas-induced altered areas in our images are significantly brighter. However, shortwave infrared bands contain more information for distinguishing altered and unaltered areas. In case the classifiers pay only attention to the visible bands, the following three processing steps are carried out:

(1) ASTER imagery is converted to PCA imagery.

(2) The first principle component is deleted.

(3) PCA imagery without first principle component is transformed back. This inverse PC transform imagery is called ASTER without PC_1 in this thesis.

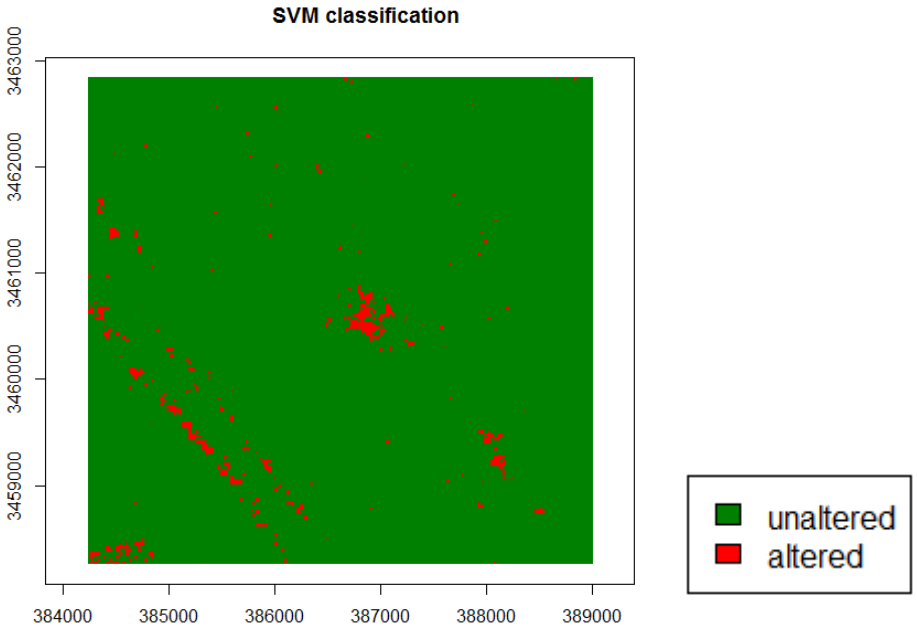


Figure 4.4 SVM classification result by using the original ASTER imagery and the small training set

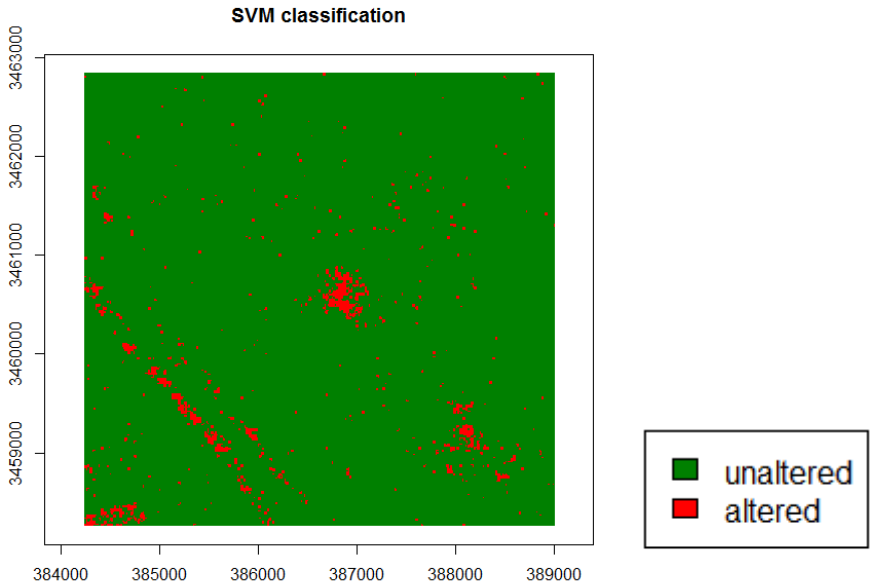


Figure 4.5 SVM classification result by using the ASTER without PC_1 and the small training set

The patterns in figure 4.4 and 4.5 are similar, but figure 4.4 is smoother than figure 4.5. Compared with figure 3.2, the SVM classification result by using ASTER and ASTER without PC_1 it is shown that both can be used to identify the alterations. Thus, it is observed that SVM classifier can be used to discriminate the difference between altered and unaltered areas. .

4.2.5. Training sets test

Figure 4.4 shows the classification result of using small training set, while figure 4.6 shows the classification result of using large training set. The pattern of figure 4.4 and figure 4.6 is similar. Both these

two results have a similar pattern when compared to the gas-induced alteration map (figure 3.3). This illustrates that SVM has a good performance even when the training set is small.

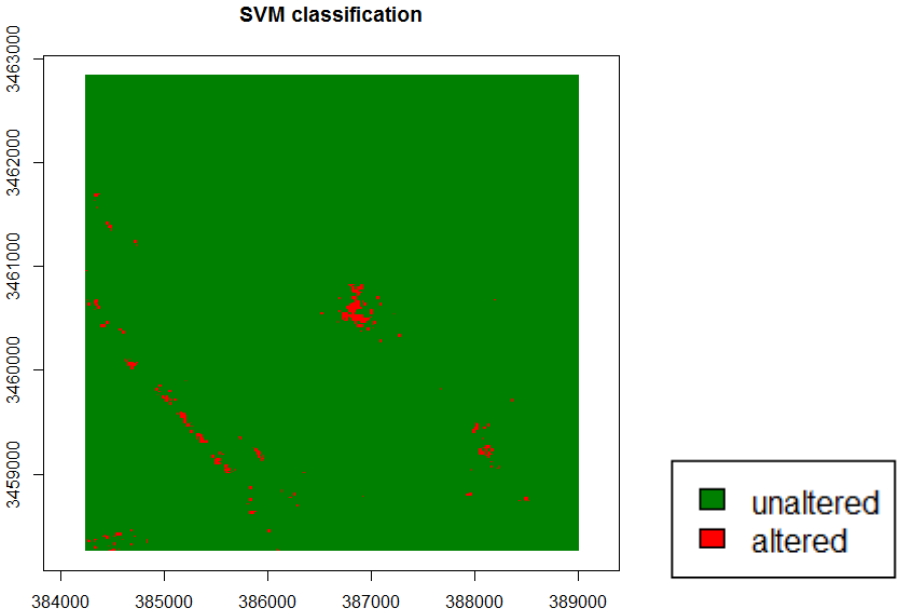


Figure 4.6 SVM classification result by using the original ASTER imagery and the large training set

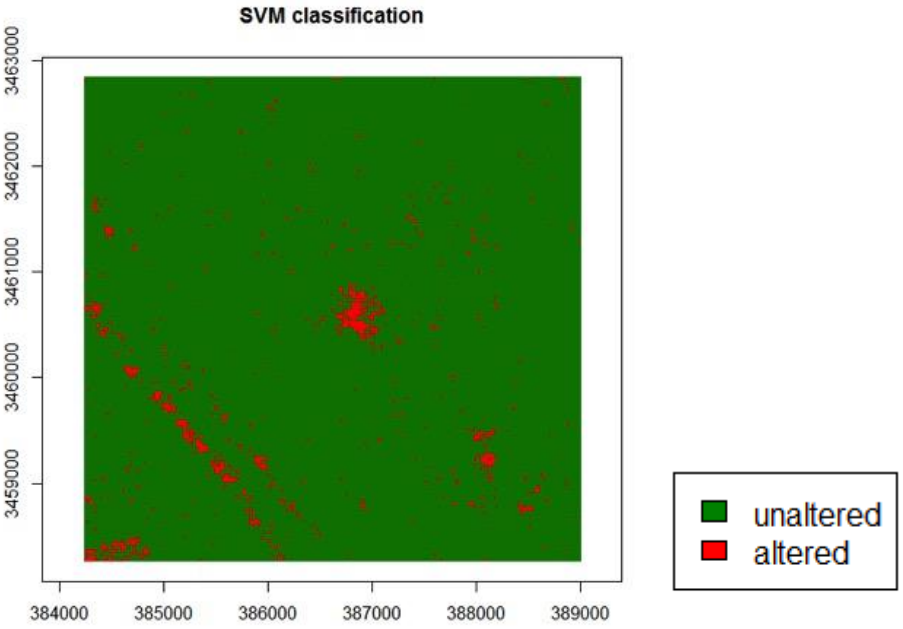


Figure 4.7 SVM classification result by using the ASTER without PC_1 and the large training set

4.3. Random forest

4.3.1. Introduction

Random forest is an ensemble machine learning method for classification and regression, which is built by a large number of decision trees. The feature combinations for each node are split randomly and independently. The most popular class of output is voted by each tree without weight (Breiman, 2001).

The decision tree plays a major role in random forest. It is a decision model whose structure is like a structured tree, which is easy to distribute and understand. This algorithm, decision tree, can only split but not converge. Decision treein general have a big overfitting problem, and it is not sensitive enough for the use small training sets. It also lacks stability so that the result will be different if the training set changes a little (Etemad-Shahidi and Mahjoobi, 2009).

Overfitting means the algorithm pays more attention to irrelevant features and creates an over-complex model to fit the training data. Although it can get a correct classification on training data, it will perform worse on test data. Especially, overfitting often happens when the training set is small.

Random forest randomly produces a large number of decision trees (weak classifier) and uses a bagging method to ensemble them. Therefore, random forest reduces the instability of a single decision tree, which reduces the probability of overfitting and performs better when the training set is small.

Assuming the number of trees in random forest is 's' then 's' datasets should be generated, and the number of endmenbers in each dataset is as same as original data. Datasets are chosen randomly with replacement. Therefore, when compared to the original data, each dataset has duplicate data and lacks some of the data. About 2/3 training of the data is used to train each tree. The remaining 1/3 data is called out of bag data, which is used to calculate the out of bag error rate (misclassification rate). Based on the error of each tree, the overall out of bag error rate, which is used to evaluate the model, is calculated. Lastly, each tree votes for one class. The classification result of each pixel is labelled by the class having the maximum number of votes.

4.3.2. Software

In this study, the random forest classifier is built by using the 'randomForest' package (version 4.8-12) in R (version 3.2.0).

4.3.3. Major parameters

Random forest has two important factors, the number of trees (ntree) and number of variables randomly distribute to each node (mtry).

4.3.3.1. Number of trees selection

Breiman (2001) states that the random forest classifier does not have the overfitting problem, in other words a large number of trees will not cause overfitting. But Mark R. Segal (2004) says that RF still can suffer from overfitting when the training set is a noisy dataset. Therefore, the optimal number of trees should be tested by using the out of bag error rate.

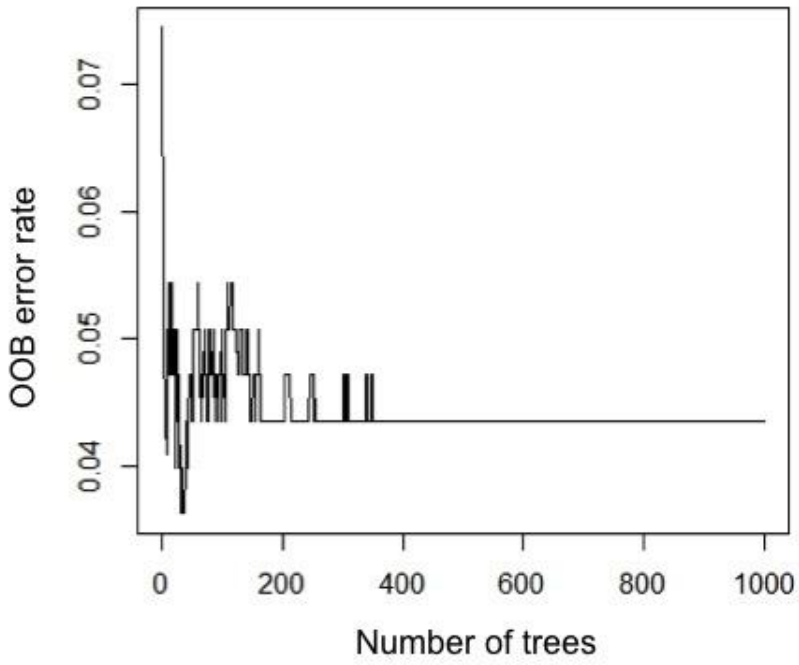


Figure 4.8 Relation between the number of trees and out of bag error rate

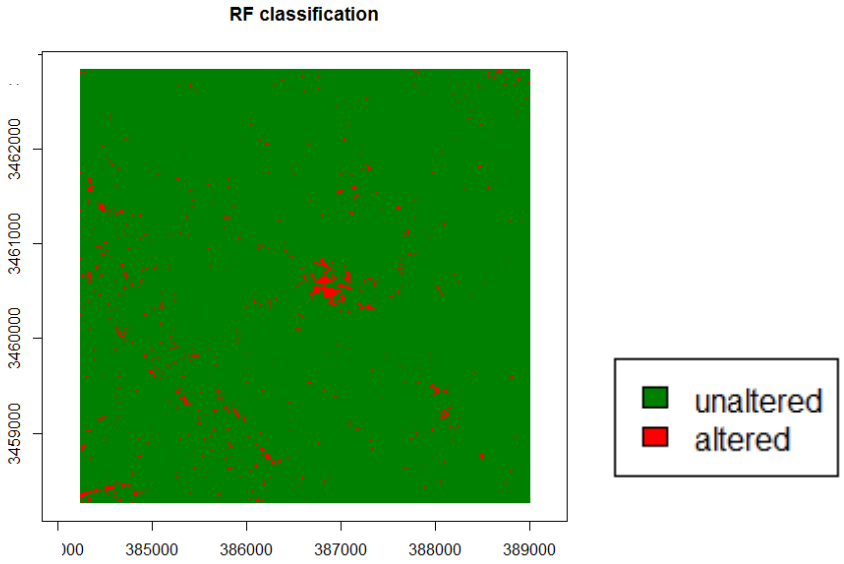
As figure 4.8 shown, when the number of trees is equal to 33, the out of bag error rate is lowest. And after 351 trees, the out of bag error rate does not vary any more. So the number of trees 33, 351 and 1000 are chosen to test for the optimal number of trees.

The main patterns in figure 4.9 are similar, but figure 4.9(c) has less noise than (a) and (b). This illustrates that a large number of trees does not cause overfitting, but that a small number of trees can suffer from overfitting and instability like a single decision tree. Moreover, trees are built randomly so that the out of bag error rate always shows slight changes. The number of trees is chosen by two principles: (1) it cannot be too large, because this will make the training time too long, (2) it cannot too small, because this may cause overfitting and instability, (3) it should be selected behind the point where the out of bag error rate stabilizes. Since large number of trees would not cause overffing problem and figure 4.9(c) is smoother than figure 4.9(b), 1000 trees are chosen for this study.

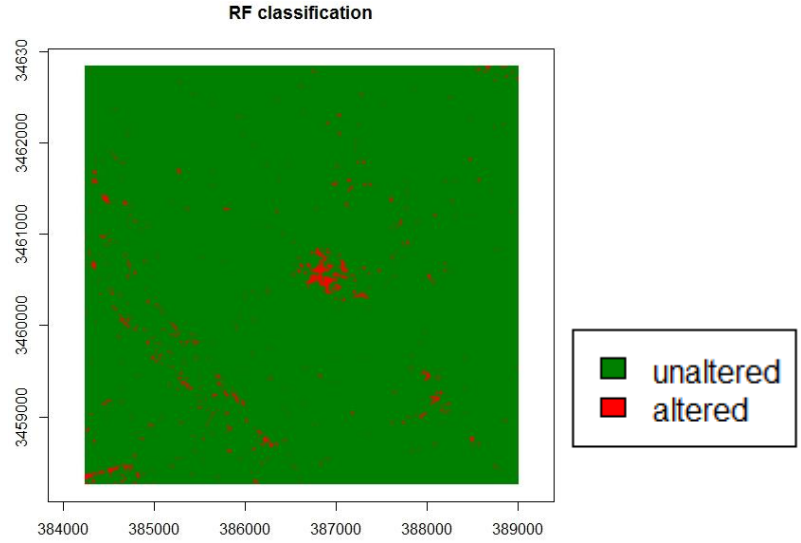
4.3.3.2. Predictor variables selection

Some predictor variables (say, mtry) are selected at random out of all the possible predictor variables and the best split on these mtry is used to split the node. By default, mtry is taken to be the square root of the total number of all predictors for the classification. Because the ASTER image has 9 bands, it has 9 predictors. Therefore, by default, the mtry is set to 3.

(a) 33 trees



(b) 351 trees



(c) 1000 trees

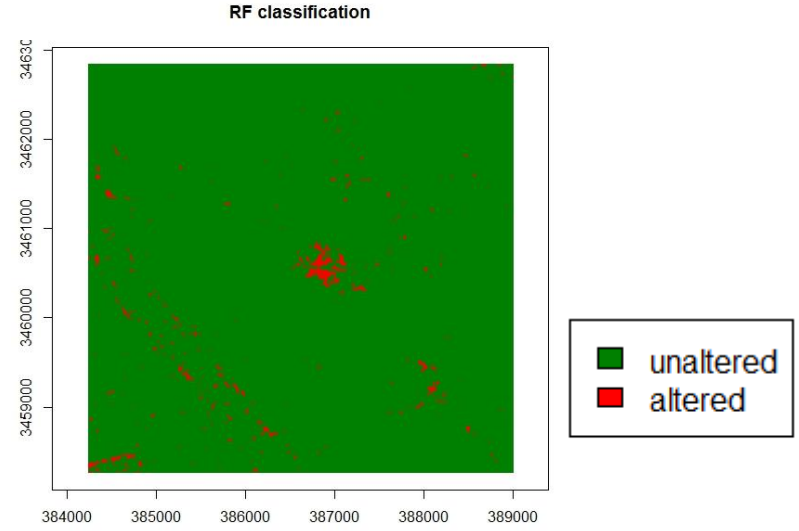


Figure 4.9 RF classification result of different trees by using the ASTER without PC_1 and the small training set

The parameter, mtry, is sensitive and has an obvious effect for RF modelling. It has a close relation with forest error rate. The value of mtry is selected by calculating the relation between mtry and OOB error rate (Breiman, 2001).

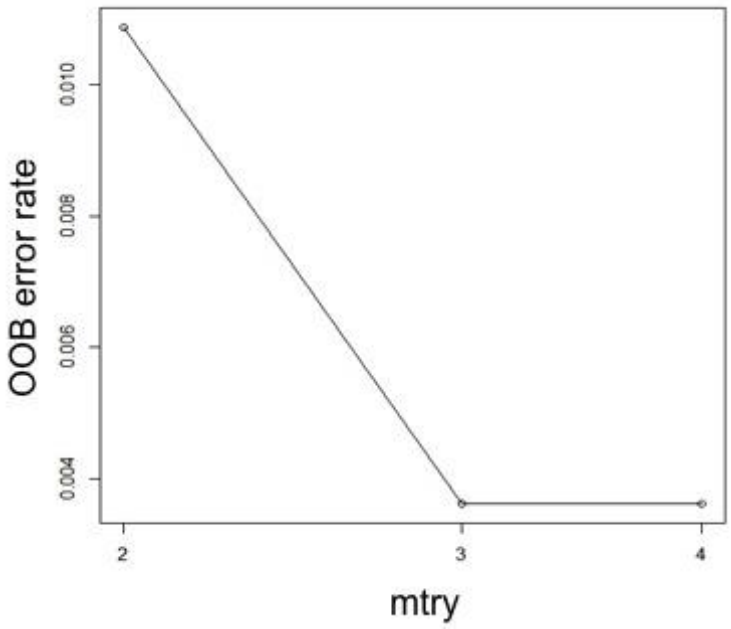


Figure 4.10 Relation between predictor variables (mtry) and out of bag error

As figure 4.10 shows, after mtry is 3, the out of bag error does not reduce any longer. To avoid large trees building a complex model, which could lead to overfitting training data (Liaw & Wiener, 2002), the mtry is set as 3.

4.3.4. Principle component analysis method

Comparing the importance of each variable measured in RF between original ASTER imagery and ASTER without PC_1. Figure 4.11 shows the importance of each variable of original ASTER imagery. The random forest classifier considers the visible bands to be more important than the shortwave infrared bands, which is contrary to the spectral analysis in chapter 2 and 3. However, figure 4.12 shows the importance of each variable of the ASTER without PC_1, which gives an opposite answer when compared to figure 4.11. Moreover, the alteration area (in red) in figure 4.14 is much larger than alteration zones in gas-reduced alteration map (figure 3.3), while figure 4.9(c) has similar patterns when compared to figure 3.3. Therefore, the ASTER without PC_1 prevents classifiers from focusing too much on the intensity of the original imagery, which does improve the performance of classifiers. In general, random forest is insensitive to irrelative variables. But in this study, we can see the shortcoming of random forest regarding the selection of important information. At least when training set is small, random forest focus more on VNIR bands instead of SWIR bands.

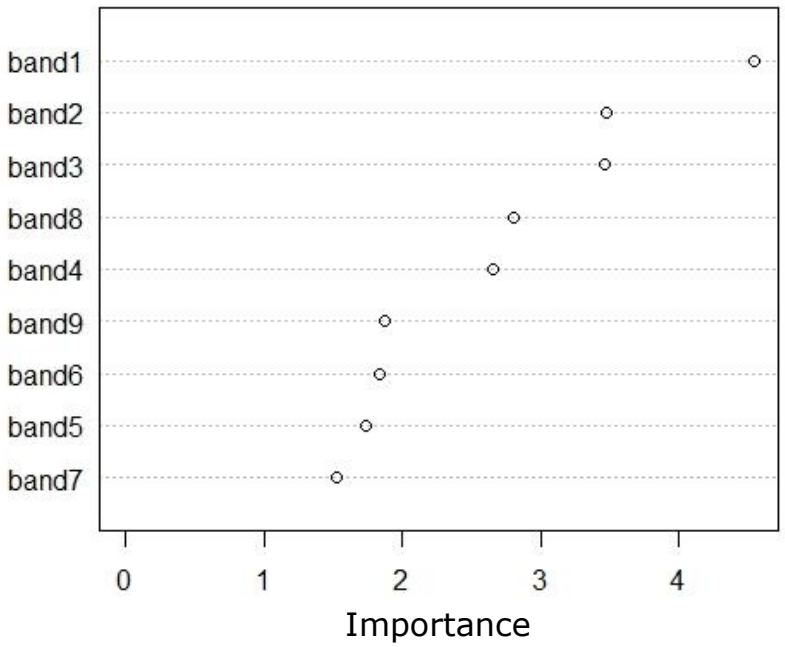


Figure 4.11 The importance of each band of original ASTER imagery, which is calculated by RF model.

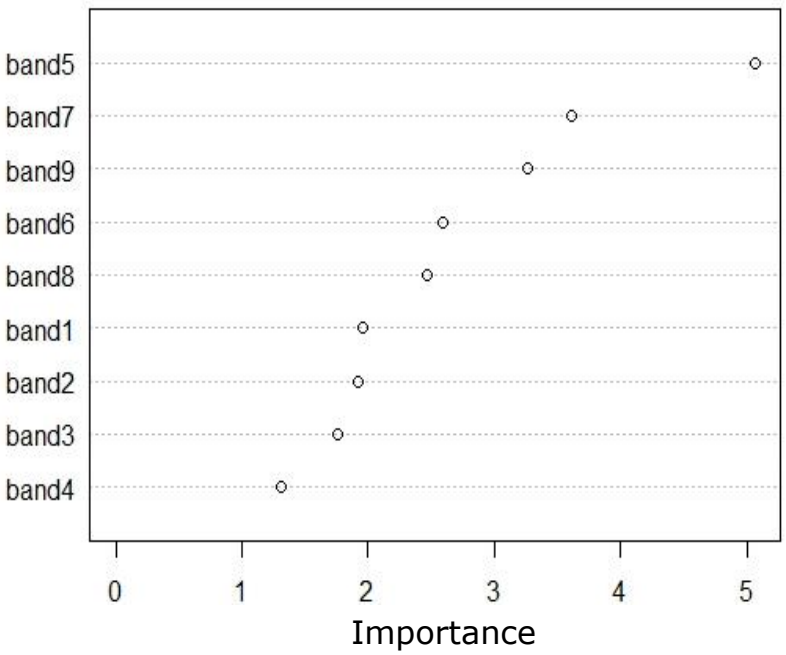


Figure 4.12 The importance of each band of original ASTER imagery, as calculated by the RF classifier.

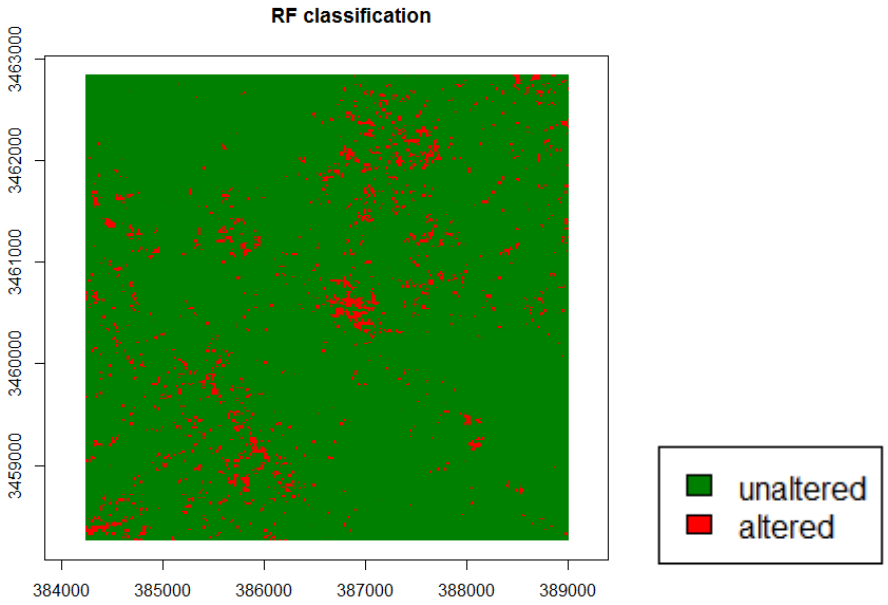


Figure 4.13 The importance of each band of ASTER without PC_1, as calculated by the RF classifier

4.3.5. Training sets test

Figure 4.13 shows the classification result of using the small training set, while figure 4.15 shows the classification result of using the large training set. The patterns of figure 4.13 show more false alarms in the alteration class, but figure 4.15 has similar patterns with gas-induced alteration map (figure 3.3). Based on section 4.2.4, the random forest classifier focuses more on visible bands. However, the large training set can fix this problem and gives a good performance.

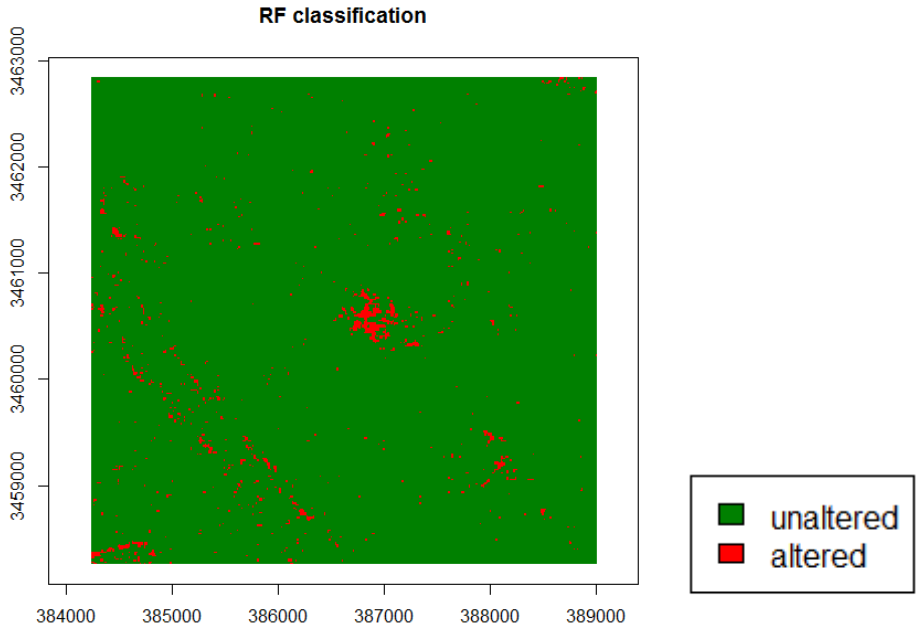


Figure 4.14 RF classification result by using the ASTER without PC_1 and the large training set

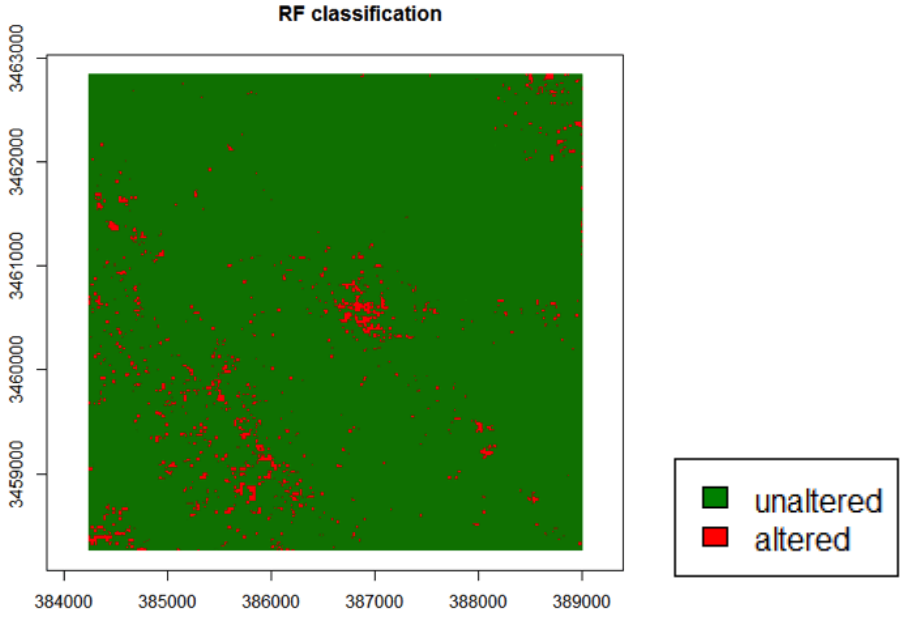


Figure 4.15 RF classification result by using the original ASTER imagery and the large training set

4.4. Gradient boosted regression trees

4.4.1. Introduction

Gradient boosted regression trees is an ensemble classifier which applies a boosting method to a number of regression trees. To improve the performance of a single regression tree, the result of the BRT classification and regression is voted by all the trees. They have successfully been applied in many fields, such as geophysics (Parisien & Moritz, 2009), biology (Friedman & Meulman, 2003) and geosciences (Lawrence et al., 2004)

In general, the gradient boosted algorithm is a process of iterations, and the new training step is to improve the result of the previous model. Every calculation is to reduce the previously obtained residual error. Therefore, the new model is built to eliminate residual in residual reduction gradient orientation.

Gradient boosted regression trees combine most advantages of tree-like machine learning methods and improve the performance of single trees whose biggest problem is instability and a relatively bad performance. There are five significant benefits listed in the following points (Elith et al., 2008).
(1) For predictors, they adapt to different variable types.
(2) They can fit phenomena in which variables contain missing data.
(3) They don't need to eliminate outliers.
(4) They have the ability to build non-linear classifiers.
(5) The interactions, of certain complexity, which are inbetween predictors, can be modeled.

4.4.2. Software

In this study, the GBRT classifier is built by using Generalized Boosted Regression Models, the 'gbm' package (version 2.1.1) in R (version 3.2.0).

4.4.3. Major parameters

There are two major parameters in the 'gbm' package, shrinkage and the number of trees (ntree). The principle of shrinkage is to avoid overfitting. When shrinkage is small, the result is gradually approached. Using this method it is easier to avoid overfitting than when shrinkage is large. In other words, if shrinkage is set, this model does not fully trust every residual tree. It thinks each tree only learns part of truth. Therefore, it needs to build more trees to make up the model for the shortfall. But too small a shrinkage will add more processing time. Based on experience (Elith et al., 2008), shrinkage is set as 0.01.

The number of trees (ntree) is a sensitive factor in boosted regression tree. The out of bag error rate is used to detect the optimal number of trees.

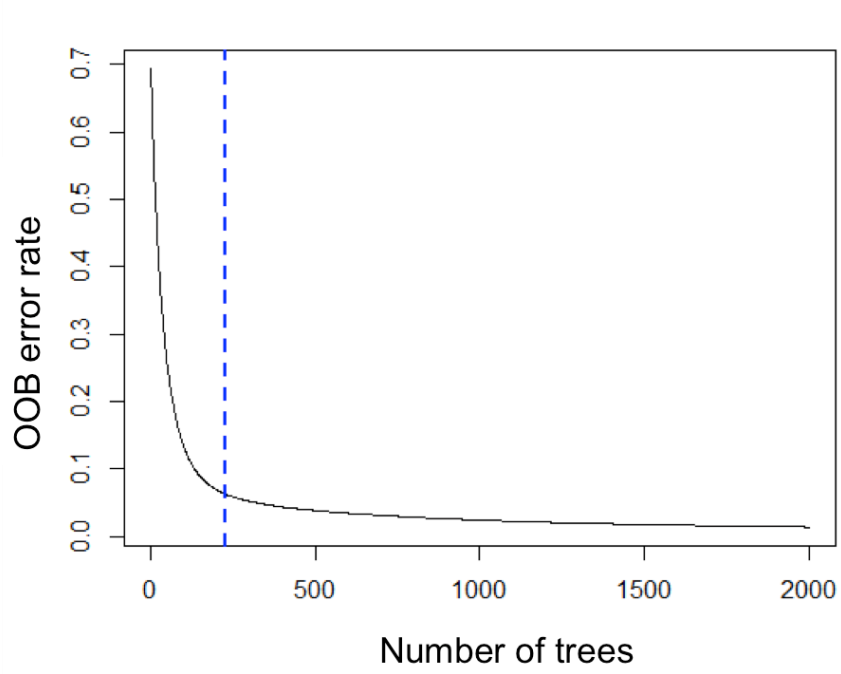
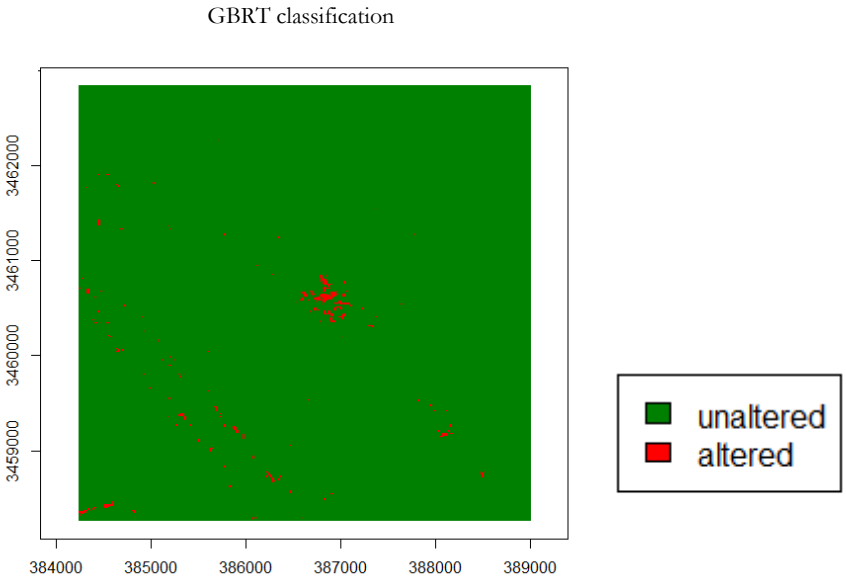


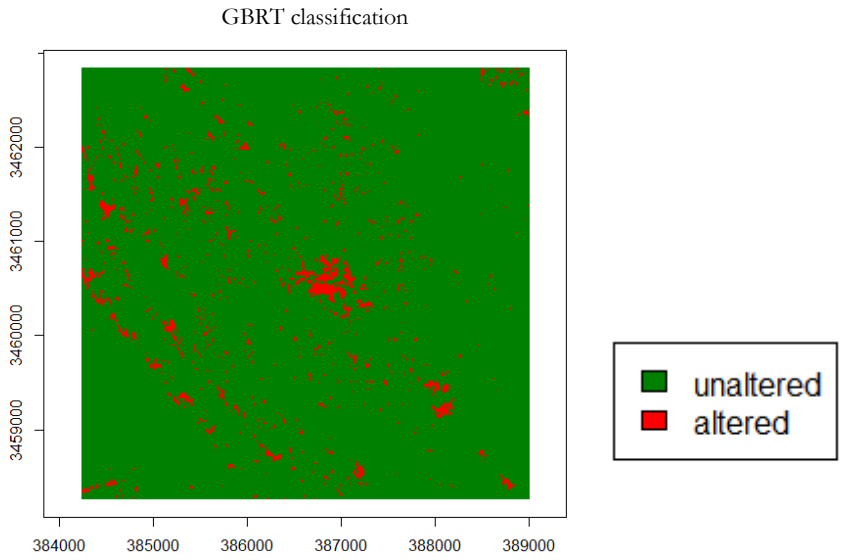
Figure 4.16 Relation between the number of trees and the out of bag error rate. The dashed line, calculated by the 'gbm.perf' function in the 'gbm' package (version 2.1.1), shows the position of the optimal number of trees.

As figure 4.16 shows, the dashed line shows the position of the optimal number of trees (227), where gradient is close to zero. In order to detect the overfitting problem of GBRT, 100, 227 and 1000 are chosen to test the relation between the number of trees and the classification result.

(a) 100 trees



(c) 1000 trees



(c) 1000 trees

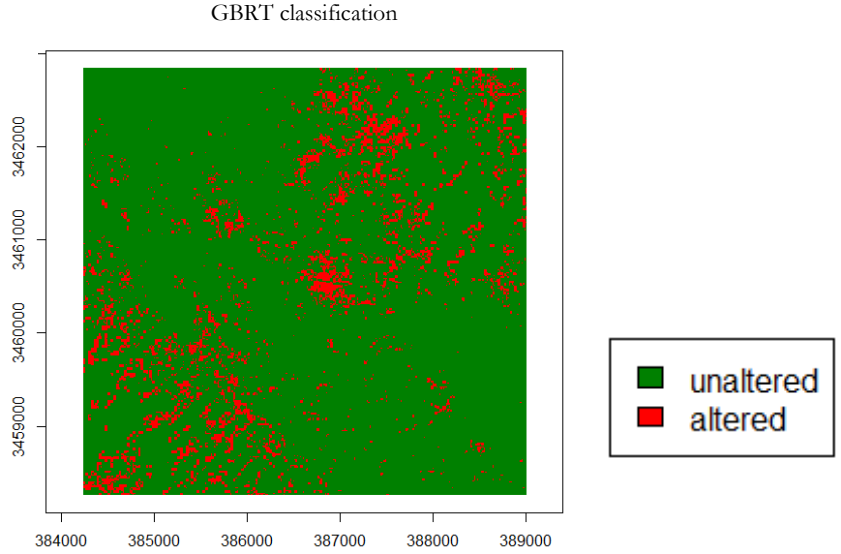


Figure 4.17 GBRT classification result of different trees by using the original ASTER and the small training set

As figure 4.17 shown, with the development of trees, the number of prospected altered areas increases. Figure 4.17(c) has much more noise than (a) and (b). This illustrates that a large number of trees will cause overfitting, but a small number of trees easily cause to underfitting and instability. The number of trees is chosen by two principles: (1) it cannot be too large, because this will cause overfitting in the result, (2) it cannot too small, becasue this will cause underfitting and instability. In the gbm package, the function 'gbm.perf', is used to find the tradeoff between bias and variance (training error and model complexity). In this position, the gradient degree is close to zero. Therefore, the number of 227 trees are chosen for this study.

4.4.4. Principle component analysis method

Comparing the importance of each variable measured in GBRT between original ASTER imagery and ASTER without PC_1. Figure 4.18 shows the importance of each variable of original ASTER imagery, GBRT model considers band8 as the most important variable, while band1, band4 and band3 occupy other important positions. This phenomenon illustrates GBRT model is not only sensitive to visible bands but also band8 in the process of modelling. Thus GBRT is more suitable than RF when variable combinations are complex. Figure 4.19 shows the importance of each variable of ASTER without PC_1, and figure 4.21 shows clearer patterns of altered areas. Therefore, for GBRT, ASTER without PC_1 also prevents classifiers from focusing too much on intensity of original imagery, which improve the performance of GBRT.

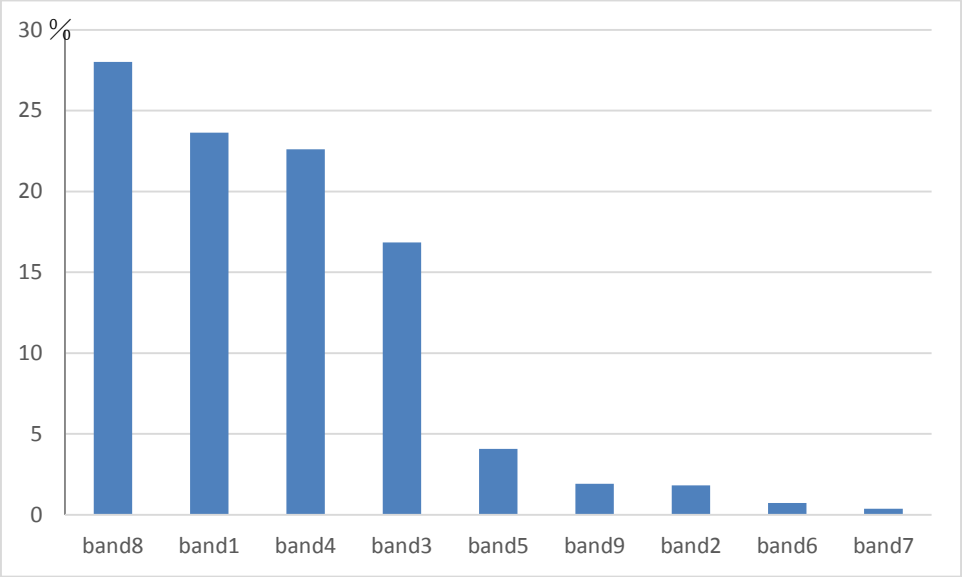


Figure 4.18 The importance of each bands of original ASTER imagery, which is calculated by GBRT model.

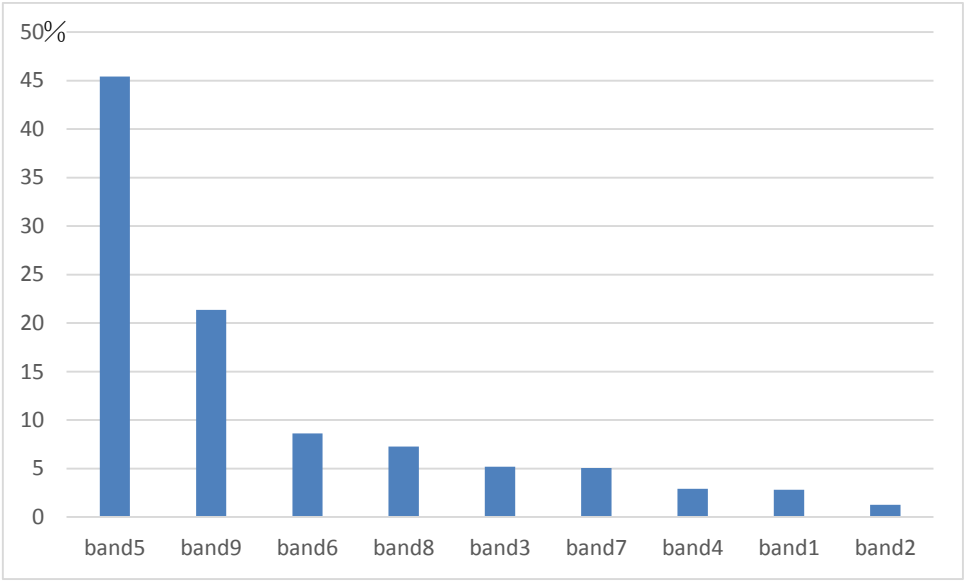


Figure 4.19 The importance of all the bands of the ASTER without PC_1, as calculated by the GBRT classifier

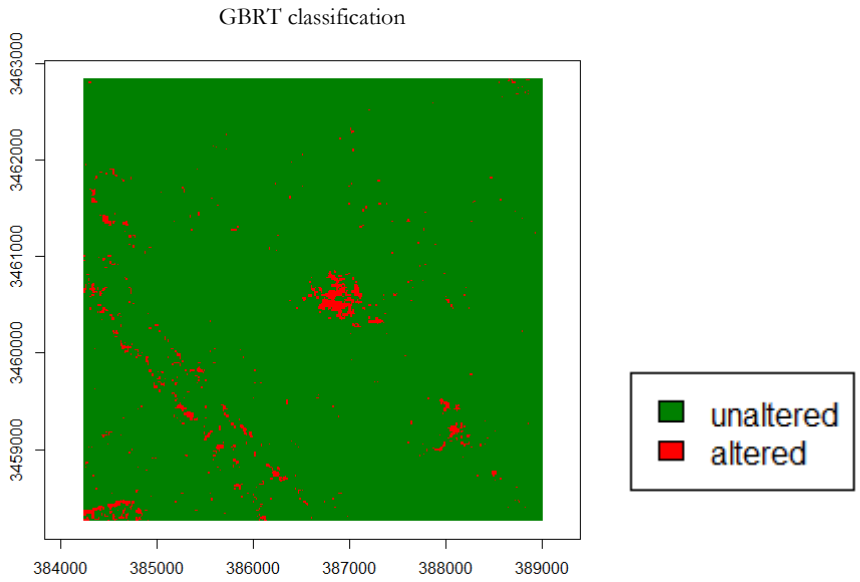


Figure 4.20 BRT classification result of the original ASTER without PC1 and the small training set

4.4.5. Training sets test

Figure 4.20 shows the GBRT classification result of the using small training set, while figure 4.22 shows the GBRT classification result of using the large training set. The pattern of figure is overfitting, but figure 4.22 has less noise. The large training set can show a better performance, but the result is not as obvious as in RF. Moreover, when the training set is small, the performance of GBRT is better than RF.

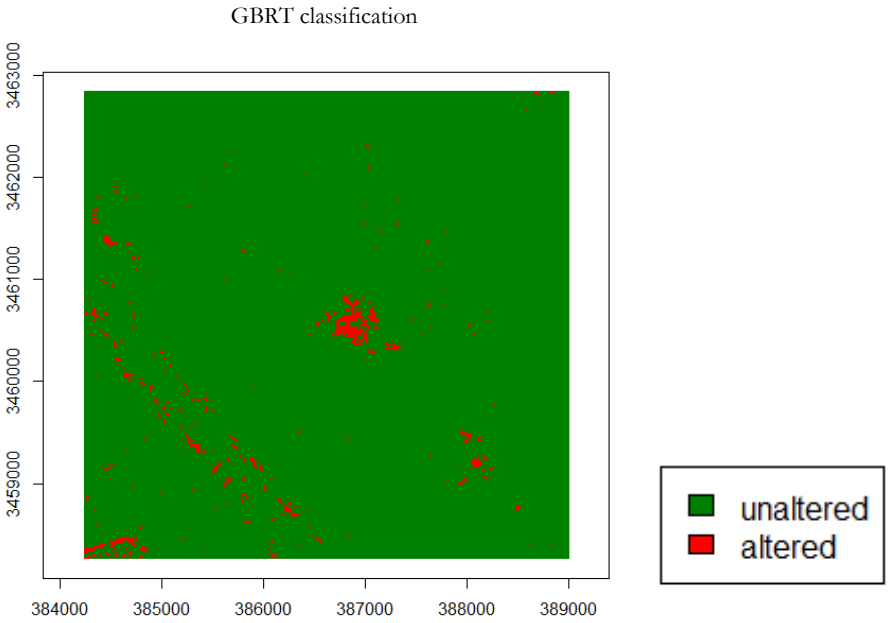


Figure 4.21 BRT classification result by using the ASTER without PC1 and the large training set

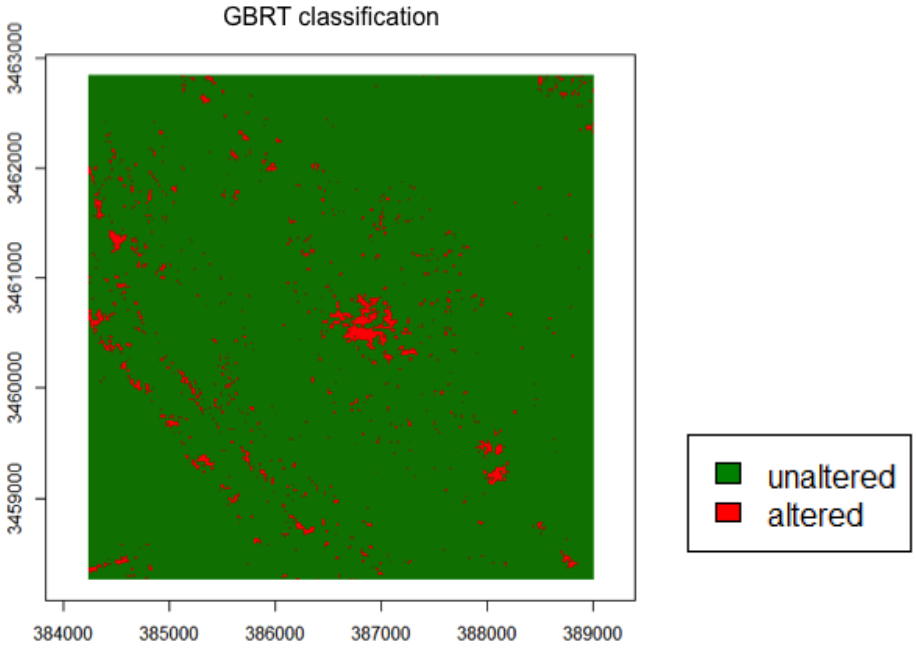


Figure 4.22 GBRT classification result by using the original ASTER and the large training set

4.5. Model validation and comparison

4.5.1. Validation of models

To validate the performance of SVM, RF and GBRT models, there are two aspects being considered: computational aspect and pattern aspect. The selection of the main parameters has been described in the sections above. This section mainly shows the validation of results by using different images and different

training sets. In the following we will validate models regarding computational and pattern aspects respectively.

Table 4.1 Main parameters and result of SVM, RF and GBRT, showing the accuracy of each classifier for different training sets and images.

		Optimal major parameter		**Result**	
SVM		Kernel	C	Image	Overall Accuracy
Small training set	Original ASTER	RBF	20	see figure4.4	80%
	ASTER without PC_1	RBF	20	see figure4.5	80%
Large training set	Original ASTER	RBF	15	see figure4.6	80%
	ASTER without PC_1	RBF	15	see figure4.7	80%
RF		ntree	mtry	Image	Overall Accuracy
Small training set	Original ASTER	1000	3	see figure4.13	75%
	ASTER without PC_1	1000	3	see figure4.9(C)	85%
Large training set	Original ASTER	1000	3	see figure4.15	80%
	ASTER without PC_1	1000	3	see figure4.14	85%
GBRT		ntree	shrinkage	Image	Overall Accuracy
Small training set	Original ASTER	227	0.01	see figure4.17(b)	85%
	ASTER without PC_1	227	0.01	see figure4.20	90%
Large training set	Original ASTER	346	0.01	see figure4.22	85%
	ASTER without PC_1	346	0.01	see figure4.21	90%

Aiming at showing objectively the merits of the classifiers, the confusion matrix is a popular validation method for remote sensed data. It compares the classified data with reference data by calculating the percentage of similarity. The overall accuracy is calculated by using all the reference data. Only using the training set to evaluate models is less convincing, which is why here we choose ground truth as the test set. The Test set is composed of 20 samples, all from the field data, the locations are as shown in figure 2.1. Among them, there are 7 unaltered samples and 13 altered samples. The overall accuracy of each classifier is shown in table 4.1. Purely from the point of view of numbers, we get the following results:

1. For SVM, the overall accuracy of each training set and image is the same. In other words, the

SVM classifier works well with the small training set, and the large training set does not give a better result. Moreover, compared to the original ASTER, the performance of the ASTER without PC_1 also does not increase.

2. For RF, the overall accuracy of the large training set applied in the original ASTER imagery is 5% higher than using the small training set. In terms of the small training set, the overall accuracy of the ASTER without PC_1 is 10% higher than the original ASTER. However, for the ASTER without PC_1 image, there is no improvement when we use large training set instead of small training set.

3. For GBRT, the overall accuracy of the ASTER without PC_1 imagery is 5% higher than the original ASTER imagery. And the large training set gives the same accuracy as the small training set.

However, the ground truth of this research is only 20 samples and they are located in a small area. Therefore, we need to compare the patterns of the classification results to the gas-induced alteration map (figure 3.3). Combining the overall accuracy of the confusion matrix together with the visualization result of altered patterns, the following results are summarized:

1. SVM has a good performance when the training set is small. And the learning ability for predictor variables is high.

2. When the training set is small, the classification result of RF is not so accurate. RF is easily confused by the intensity information of the images, and it requires a relatively high-quality image and training data set.

3. The result of GBRT classifier is sensitive to parameters, so we need to pay special attention to the parameter setting. However, it learns predictor variables accurately, and the accuracy of the resulting classification is high.

4.5.2. Comparison of models

As observed from the result of SVM, RF and GBRT, all these three classifiers have a relative good performance. In order to avoid unilateral evaluation of these classifiers, this section sums up both advantages and disadvantages of each in table 4.2.

For producing hydrocarbon seepages maps, there are two significant difficulties: the training data is difficult to obtain and the seepage size is relatively small. Depending on the different characteristics of SVM, RF and GBRT, in chapter 4 we summarized a classification process to foster strengths and circumvent weaknesses of these classifiers. Firstly, as SVM works well with a small number of training data and have smooth patterns, we think that this classifier is best for this application. However, the smooth patterns n the classification results of SVM means that this method might be not sensitive to relatively small seepages. Thus, if we need a more detailed alteration map, GBRT, which has best overall accuracy, might be applied as the second step. Lastly, one could compare the classification results of SVM and GBRT. If the main patterns are similar, then the results may be more reliable. Although GBRT has best overall accuracy in this study, the number of parameters that need to be tested are higher than SVM and the stability of GBRT is not as good as SVM. RF is not considered because of its moderate and unstable performance for detecting seepage alterations in this study. Therefore, in our view, SVM occupies

the most important position among these three classifiers for producing a hydrocarbon seepage alteration map.

Table 4.2 Comparison of SVM, RF and GBRT

	SVM	**RF**	**GBRT**
Size of training set	Work well with small training set	Work well with large training set	Large training set is better than small training set, but result by using small training set is better than RF
Attention of image intensity	Low	High	Medium
Overfitting or underfitting	Not common	Not common	Overfit when ntree is more; underfit when ntree is less
Complexity of predictor variables	Not as good as GBRT	Can process high-dimensional data, but not as good as GBRT when variables combination is complex.	Work well with complex variables combination
Bootstrap sampling	No	Yes	No
Time	Slowest	Fastest	Medium
Model stability	Stable	Randomly modeling	Randomly modeling
Importance of each variables (bands)	It is impossible to get the effect of each bands	Original ASTER: see figure 4.11 ASTER without PC_1: see figure 4.12	Original ASTER: see figure 4.18 ASTER without PC_1: see figure 4.19

5. APPLYING SVM TO DETECT GAS-INDUCED ALTERATION AT REGIONAL SCALE

Three classifiers, SVM, RF and GBRT, were compared in chapter 4. The SVM method was found to be the most stable and suitable classifier for detecting gas-induced alteration. In this chapter, the SVM classifier is applied to produce regional alteration maps in the marly limestone formation. Potential alterations can be detected and mapped.

5.1. Methodology

To identify additional hydrocarbon seepages and to give this research more practical significance, the study area is enlarged to a regional scale, with a total area of 46*70 km^2. It mainly contains the Gachsaran formation, Mishan formation, Alluvium formation and Bakhtiari formation. The lithology mainly consists of evaporite, marl, limestone, shale, sandstone and alluvium (Salati, 2014). The classifier SVM, is used to distinguish altered and unaltered classes. In chapter 4, we saw that SVM has a good performance when the training set is small. And in a real-world situation the training data is difficult and costly to collect. The training set is a small training set (section 4.1.1) and the image is the original ASTER image whose pre-processing is as same as test area (section 3.3.1). Similarly, the overall accuracy of the confusion matrix and pattern of classification result are used to validate the result. Because the training set is selected to identify alterations in the marly limestone, the geological map is used to mask out the other lithologies.

5.2. Result

Figure 5.1 shows the SVM classification result using the ASTER image. The locations of the Mishan formation are indicated (yellow polygons). TThis formation predominately consists of the marly limestone, while the the Gachsaran formation (blue polygons) is dominated by evaporite, marl and limestone. The most important altered mineral in the marly limestone is gypsum. However, gypsum is also a common mineral in evaporite. Figure 5.1 shows that altered minerals (red color class) are mainly the evaporite (blue polygons), which is consistent with the actual situation. Moreover, the overall accuracy is 85% and the user's accuracy of altered class is up to 91.7% (see table 5.1 and figure 5.3). In the following, each location of potential seepages (indicated by boxes in black) is described respectively.

Table 5.1 SVM classification accuracy result

Classes					
		Altered	Unaltered	Total	Correct %
Observed	Altered	11	1	12	91.7
	Unaltered	2	6	8	75
	Total	13	7	20	
	Omission	15.4	14.3		
Overall accuracy		85%			

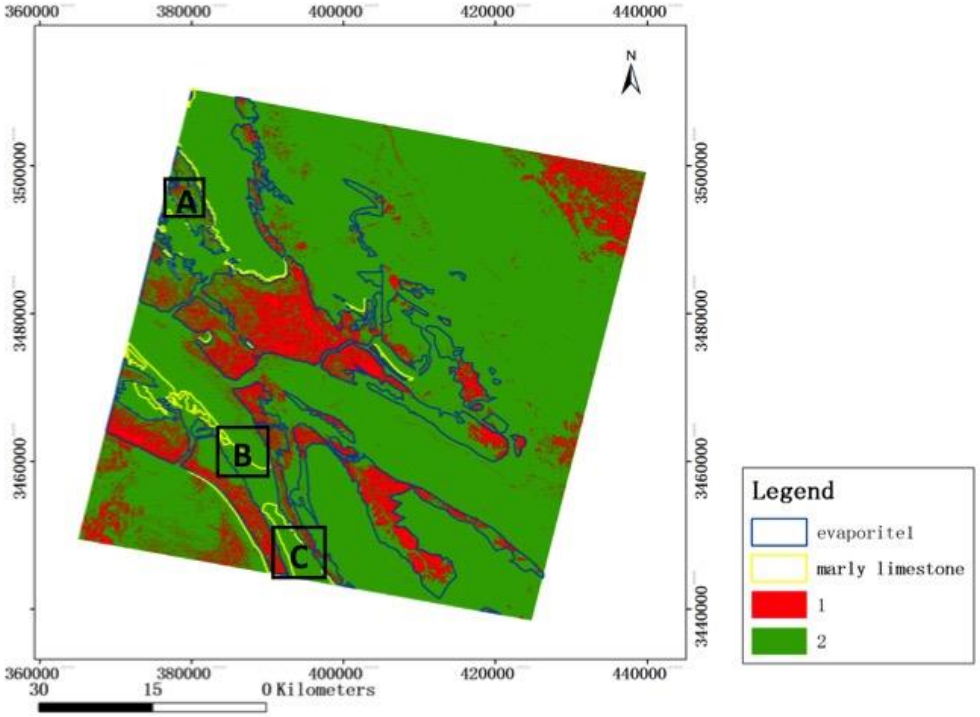


Figure 5.1 SVM classification result at regional scale. The location is shown in figure 1.1. Boxes in black are indicating potential alterations.

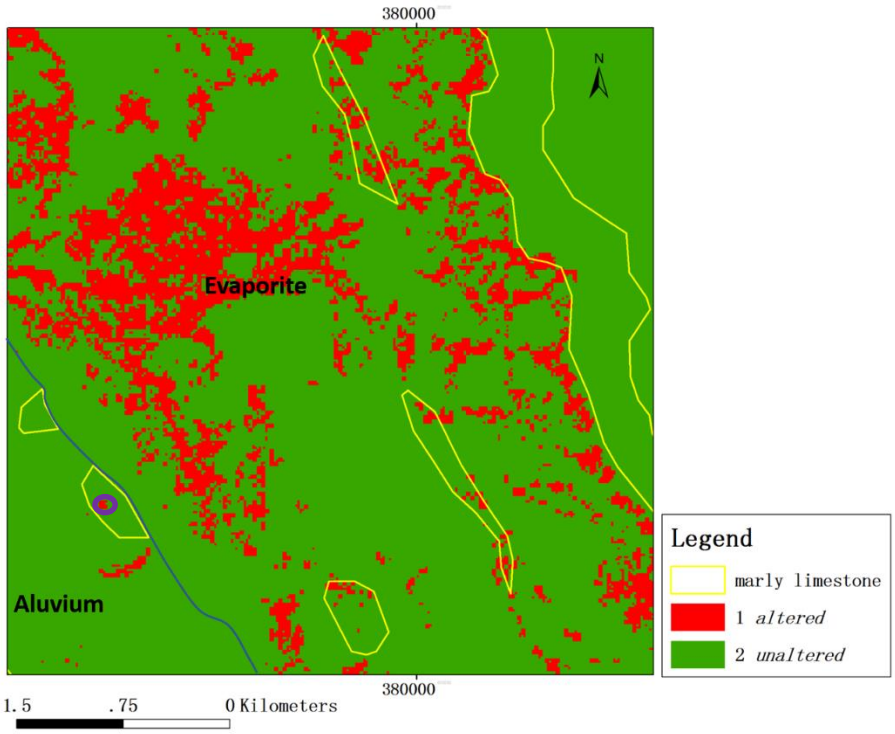


Figure 5.2 Box A, the purple circle shows the potential alteration.

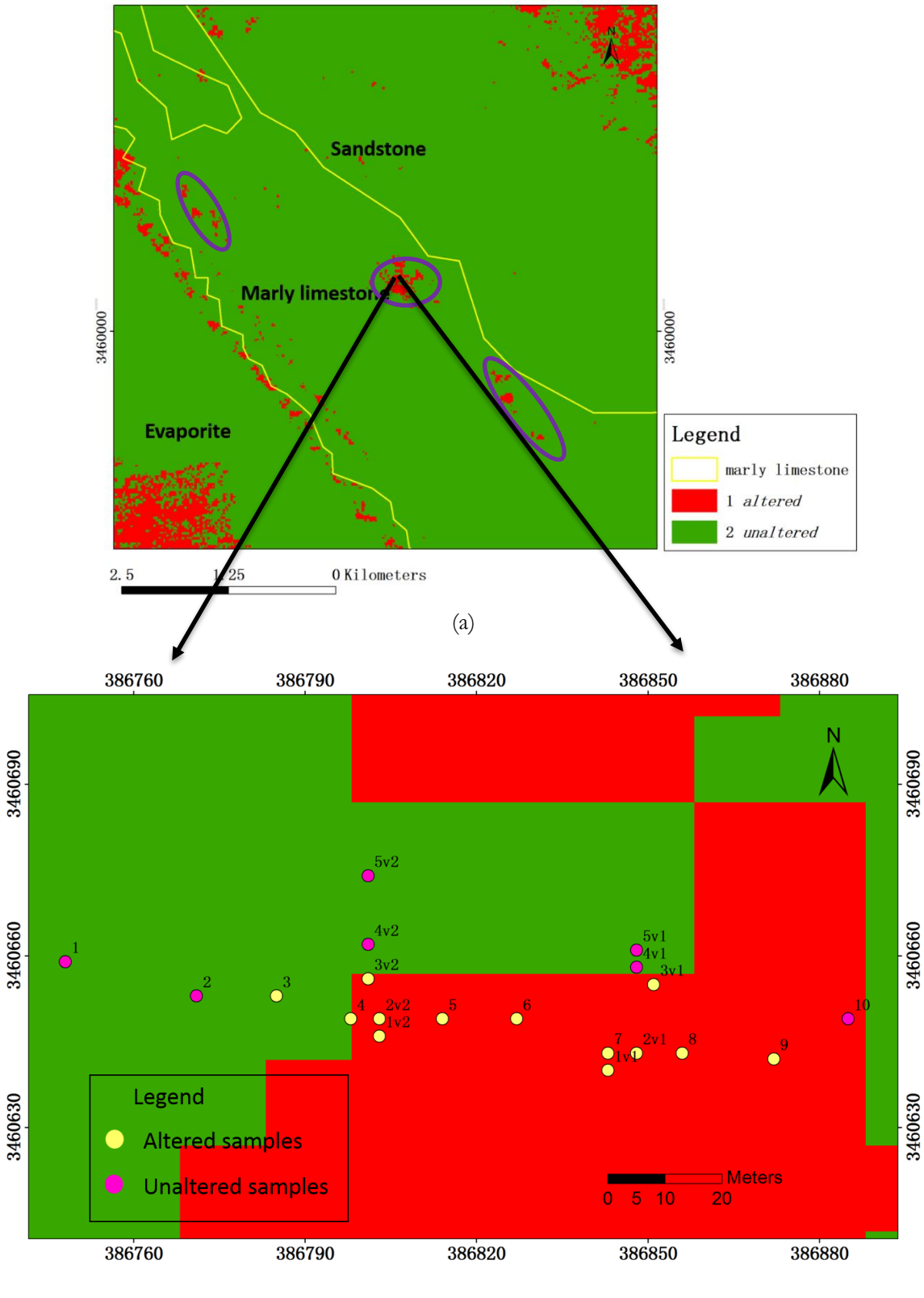


Figure 5.3 (a) Box B, the purple circle indicates the alteration. (b) yellow circles show the altered field samples and pink circles show the unaltered field samples. Samples 3, 4 and 10 are incorrectly classified.

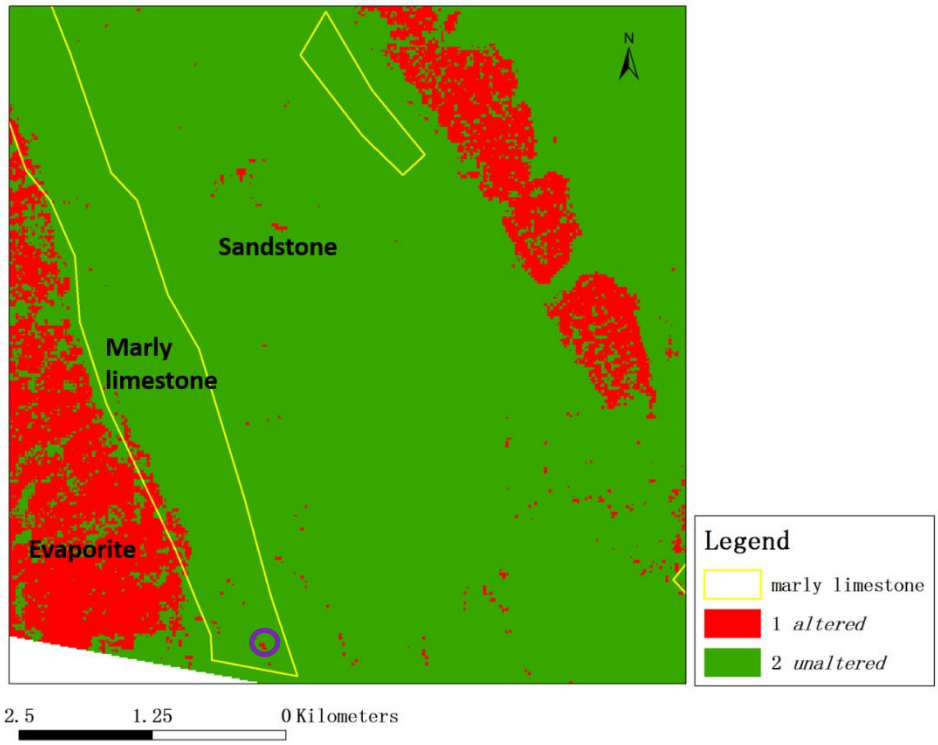


Figure 5.4 Box C, the purple circle shows the potential alteration.

In figure 5.2, there are six marly limestone zones, but only one area is selected as potentially having a gas-induced alteration. Most marly limestone formations in this figure are surrounded by the evaporite formations and gypsum is one of common evaporate minerals. Furthermore, the origin of the gypsum is unknown. It might be from alterations but it might also be transported from the evaporites into the areas having a limestone lithology. We suggest only one area in figure 5.2, which is drawn by purple circle, might be gas-induced alteration, because in this area the evaporite is far away and therefore the gypsum in this area might be from gas-induced alteration. In addition to that, the spectra of this area are indeed interpreted to be gypsum (see figure 5.7). Similarly, red areas in figure 5.3 and figure 5.5 are not considered to be alterations.

Area B (see figure 5.1), indicated in figure 5.3 shows the location of the known alterations that were used for the training of the classifiers (see chapter 3 and 4). It is obvious that the patterns in figure 5.3 are as similar as SVM classification map at local scale (figure 4.4). Since WV-2 imagery has 2 meters spatial resolution, the ASTER image used in chapter 3 and 4 was resampled to 2 meters. Meanwhile, the spatial resolution of ASTER image used in this chapter is 15 meters. Thus, the patterns in 5.3 are not completely same as figure 4.4. Furthermore, the accuracy of SVM is independent of the size of the input data set. The potential alterations are in the purple circle.

In figure 5.4 one area is found, of which the spectra are shown as gypsum (see figure 5.5). This suggests alteration, because it is away from the evaporite. We can see from the boundary between the evaporite and the marly limestone that some gypsum crosses the border and exists in the marly limestone. This suggests that the geological map we use may be inaccurate at some locations.

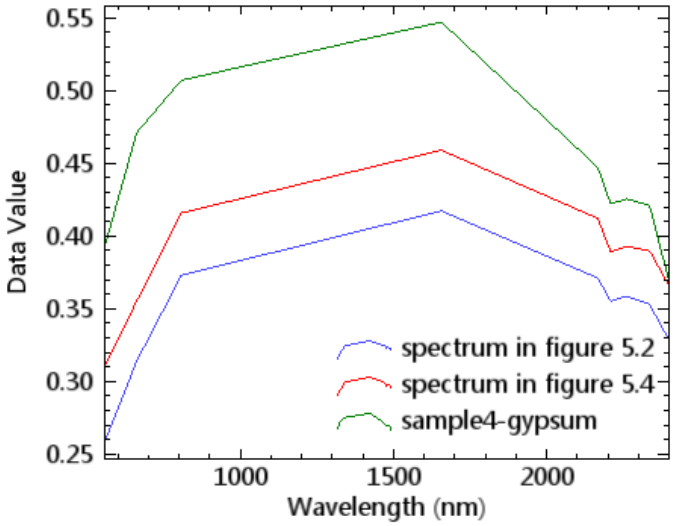


Figure 5.5 Gypsum spectra from alterations highlighted by purple circle in figure 5.2 and 5.4. Compared with gypsum spectrum of field sample 4 which was interpreted as gypsum in chapter 2, the spectrum from pixel of alteration with purple circle in figure 5.2 is interpreted as gypsum.

5.3. Conclusion

The SVM classifier is successfully applied on a regional scale by using small training set (see section 4.1). The overall accuracy is 85%. The gypsum boundary accurately fits the boundary of the evaporite. Therefore, we state that the SVM approach is also a good method for detecting lithology. Moreover, two new areas are suggested as potential alterations, besides the known alterations detected in test area

6. DISSCUSION

In this thesis, SVM, RF and GBRT classifiers have been compared. This is done through using remote sensing data (ASTER, WorldView-2) as input to be classified for detecting hydrocarbon alterations and validation by limited field data. Moreover, a regional mapping of potential alteration areas is performed. In the sections below, several findings on the contributions of remote sensing and machine learning for use in identifying hydrocarbon seepages will be discussed and compared to the results obtained by Salati et al. (2014).

6.1. ASTER and WorldView-2 data processing

In chapter 3 we confirmed that ASTER is useful for mineral mapping, and that WorldView-2 can improve the spatial resolution of ASTER. However, the improvement using WorldView-2 is observed to be minimal. Using WorldView-2 only it cannot discriminate the important minerals such as gypsum, and it also cannot distinguish ferric and ferrous minerals. However, the alterations show the high brightness in image, so we can see clear patterns in WorldView-2. Furthermore, we observed that there exist a shift of about 60 meters between ASTER and WorldView-2, which may not have been observed in Salati et al. (2014). In the following sections, the differences with the previous work by Salati are going to be discussed in two aspects, using ASTER and WorldView-2 respectively.

6.1.1. ASTER data processing

Compared with the result of Salati (2014), there are several differences. In the paper (Salati et al., 2014), the gas-induced alterations are said to have high concentration of ferrous iron and gypsum and low concentration of carbonates and clays. However, in this research it is observed that gas seepages in the marly limestone formation are rich in gypsum and clays, while they seem to contain less ferric iron and carbonates. There might be various reasons for these differences in observations. These will be discussed below.

Firstly, the algorithms used for atmospheric correction are different. In this study, the FLAASH model was chosen instead of the logarithmic residuals correction. The logarithmic residuals method results in pseudo-reflectance (Tian et al., 2008), which makes it difficult to compare image spectra with field measurements or laboratory spectra. In addition, the log residuals correction is a conversion model based on the characteristics of the image itself, while FLAASH model is a calibration model based on the theory of atmospheric radiation. So from this perspective, the FLAASH model is the best approach in hyperspectral data radiometric calibration(Tian et al., 2008).

Moreover, band5/band4 (ferrous index) was chosen to detect ferrous iron in this study, while band3/band1 (ferrous index*) was chosen by Salati (2014). As mentioned in section 3.1.3, band5/band4 is used to identify ferrous iron in silicates and carbonates, and it is impossible to use band5/band4 to detect ferrous iron if it is associated with oxide and sulphate. As shown in figure 6.1, with the ferrous index* (band3/band1) it is impossible to distinguish altered and unaltered samples. Therefore, neither ferrous index (band5/band4) nor ferrous index* (band3/band1) can be used for mapping ferrous iron in this study area.

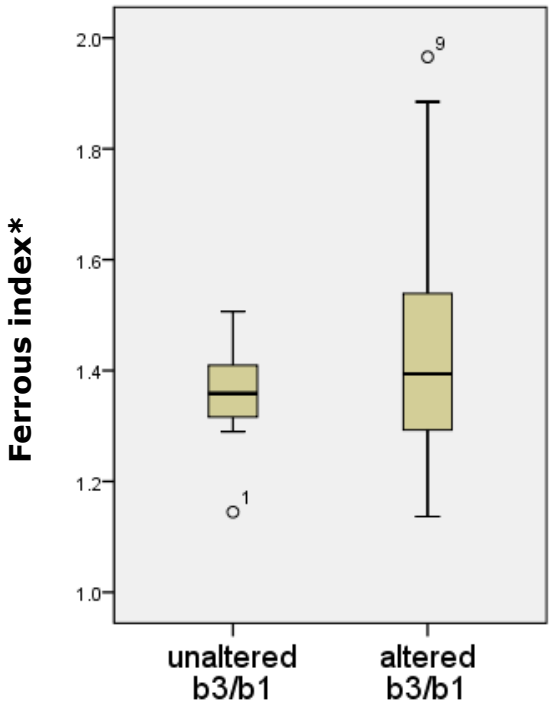


Figure 6.1 Box plots of band ratios showed comparison of ferrous index* (band3/band1) between unaltered and altered field samples. The ferrous index* value in unaltered samples is in the range of altered samples.

Finally, clay minerals have a high abundance in the entire area. According to the result of spectral analysis, clays exist in both altered and unaltered samples. Meanwhile, figure 3.2 illustrates that the concentration of clays in altered areas are higher than unaltered areas. This result does not replicate the finding, that unaltered areas contain more clays than altered areas, reported by Salati (2014).

Thus it is concluded that the FLAASH model is successfully used to pre-process ASTER data, and that the ferrous iron & clay indexes are not good indicators for mapping hydrocarbon seepages in this study area.

6.1.2. WorldView-2 data processing

Salati et al. (2014) observed that the gas-induced alterations have high concentrations of ferric iron and sulphur. However, in this research it is observed that gas seepages in the marly limestone formation are rich in gypsum and clays, while they contain less ferric iron and carbonates. The following will discuss these differences.

Firstly, the algorithms for atmospheric correction are different. As the previous section discussed, the FLAASH model was chosen in this research instead of log residuals correction in Salati's research.

Moreover, ferrous minerals are not detected in this research, while band4/band7 was chosen by Salati et al. (2014). As above mentioned, WorldView-2 data does not have the spectral bands for detecting ferrous minerals. As figure 6.2 shows, using the ferrous iron index*(band4/band7) it is impossible to distinguish altered and unaltered samples. Therefore, the ferrous iron index* (band4/band7) cannot be used for mapping ferrous iron in this study area.

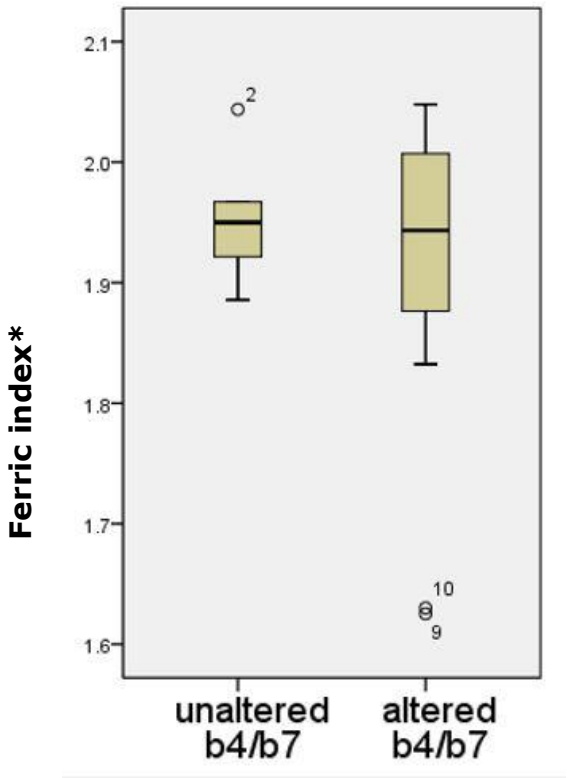


Figure 6.2 Box plots of band ratios showed comparison of ferric index* between unaltered and altered field samples. The ferric index* value in unaltered samples is in the range of altered samples.

Finally, band5/band3 was chosen to detect ferric iron in this study, while band4/band7 was chosen by Salati et al. (2014). As figure 6.3 shows, the value of altered areas is lower than unaltered areas in ferric iron index (band5/band3) image. Meanwhile, ferric iron index*(band4/band7) have a higher value in altered areas than unaltered areas. According to the result of laboratory spectral analysis and ASTER data analysis, unaltered samples contain more ferric iron. Therefore, ferric iron index is more suitable than ferric iron index* to identify ferric minerals in this study area.

Consequently, the FLAASH model is successfully used to pre-process WorldView-2 data, and altered areas are observed to be rich in sulphur, but they lack ferric minerals. In addition, although WorldView-2 data has a high spatial resolution and has more VNIR bands than ASTER data, it cannot replace ASTER for mapping alteration areas because of the missing SWIR bands in WV-2.

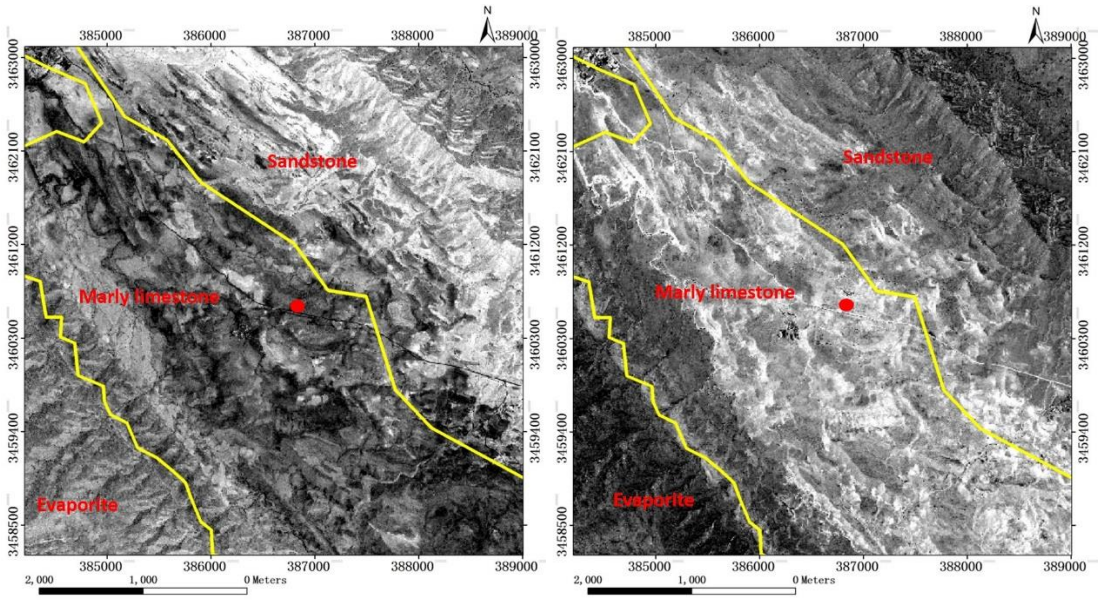


Figure 6.3 (a) Ferric iron index (band5/band3) image, (b) ferric iron index* (band4/band7) image. Ferric iron index is observed to be higher in altered area than unaltered areas, ferric iron index* is lower in altered area than unaltered areas.

6.2. Image classification

Our findings confirm that all these classifiers, SVM, RF and GBRT, perform well and have a similar overall accuracy in this study. Where the alteration area is known, the patterns of all classification maps produced by SVM, RF and GBRT are comparable, although the result of RF has an obvious overfitting problem. Nevertheless, one significant advantage of RF as claimed by Breiman (2001) would be that RF does not overfit. The result of this thesis indicates that the test error does not increase with the rise of model complexity (number of trees). In contrast, RF is significantly influenced by intensity values of the remote sensing image and it needs a large training set to decrease the problem of overfitting.

Furthermore, the overall accuracy of GBRT is the highest among these classifiers. GBRT have an apparent underfitting and overfitting problem when the number of trees is not optimal. Salati et al. (2014) used 1000 trees, which may be suboptimal for this problem. However, we use an optimal number of trees obtained by the 'gbm.perf' function in the 'gbm' package (version 2.1.1) and therefore the result may have been improved compared to the result of Salati.

Compared with the randomness of RF and GBRT, the result of SVM is very stable (see section 4.5). The patterns detected using SVM in ASTER data are extremely similar with the hydrocarbon alteration map produced by the knowledge-based approach (band ratio and relative absorption band depth) and much more smooth than RF and GBRT. In most data sets, GBRT has a better accuracy than SVM. However, we observe that SVM is more suitable for detecting mineralogy with limited field data.

Since our classification maps only have two classes, the SVM method used in this research is not an ensemble classifier. These classifiers are only a selection of the available popular classifiers. This was done in order to minimize the complexity of selecting classifiers and coding time.

Table 6.1 Comparison between our work and Salati' work

	Salati' work	Our work
Pre-processing method	Logarithmic residuals	The FLAASH Model; Geometric correct the bias between ASTER and WV-2
Mineral indexes	Ferrous iron index*: B3/B1 Ferric iron index*: B4/B7	Ferrous iron index: B5/B4 Ferric iron index: B5/B3
Clays	Clay minerals were identified to be less in altered samples than unaltered samples	Clay minerals were identified to be higher in altered samples than unaltered samples.
Ferric iron	Ferric iron minerals were identified to be higher in altered samples than unaltered samples	Ferric iron minerals were identified to be less in altered samples than unaltered samples.
GBRT	Number of trees is 1000	Number of trees was optimal and calculated by 'gbm.perf' function in 'gbm' package. The accuracy is higher than Salati' work.

6.3. Application to other area

The strength of remote sensing and machine learning is that limited field data can already yield detailed information. In other words, as was demonstrated in this thesis, a well training SVM model can predict the alteration map for the entire extent of an ASTER image in the same lithology. Therefore the obtained method might decrease the amount of fieldwork by highlighting the most interesting areas.

Next to the known alterations, our results point out two additional areas that might be altered by gas seeps. Hydrocarbon seepage alterations in this study area are spatially strongly associated with structures in the Gachsaran formation and the Mishan formation. Oil seeps are rare in the Gachsaran and Mishan formation. Nevertheless, there are some known gas seeps. Gas might be separated from oil underground. Afterwards, it could come out along the penetrable fractures, even in the case that seal is effective and there isn't an obvious fault crossing the seal. Gas might migrate along faults in NW-SE direction. Gas-induced alterations have been observed in the Gachsaran formation and the Mishan formation (Salati, 2014).

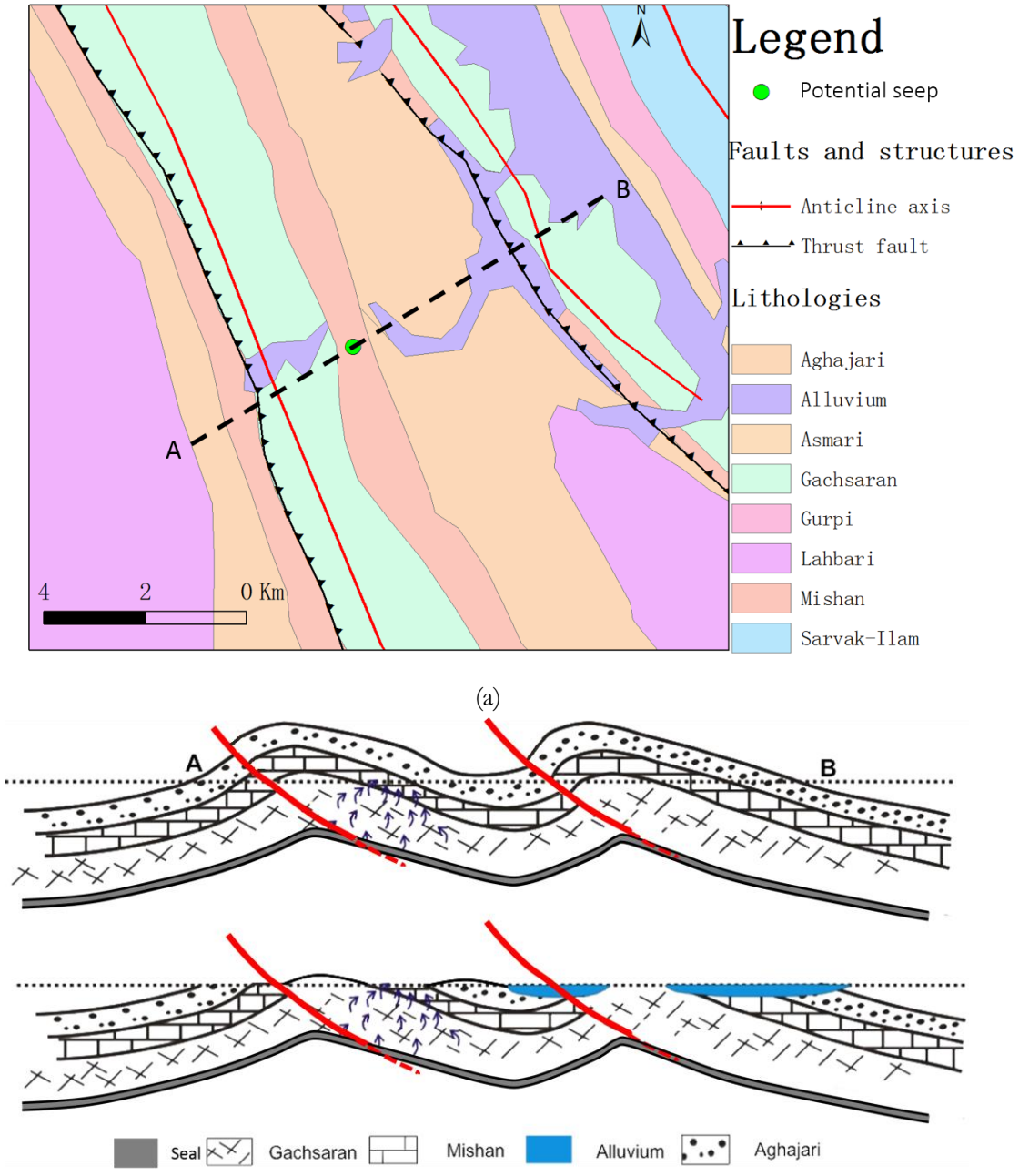
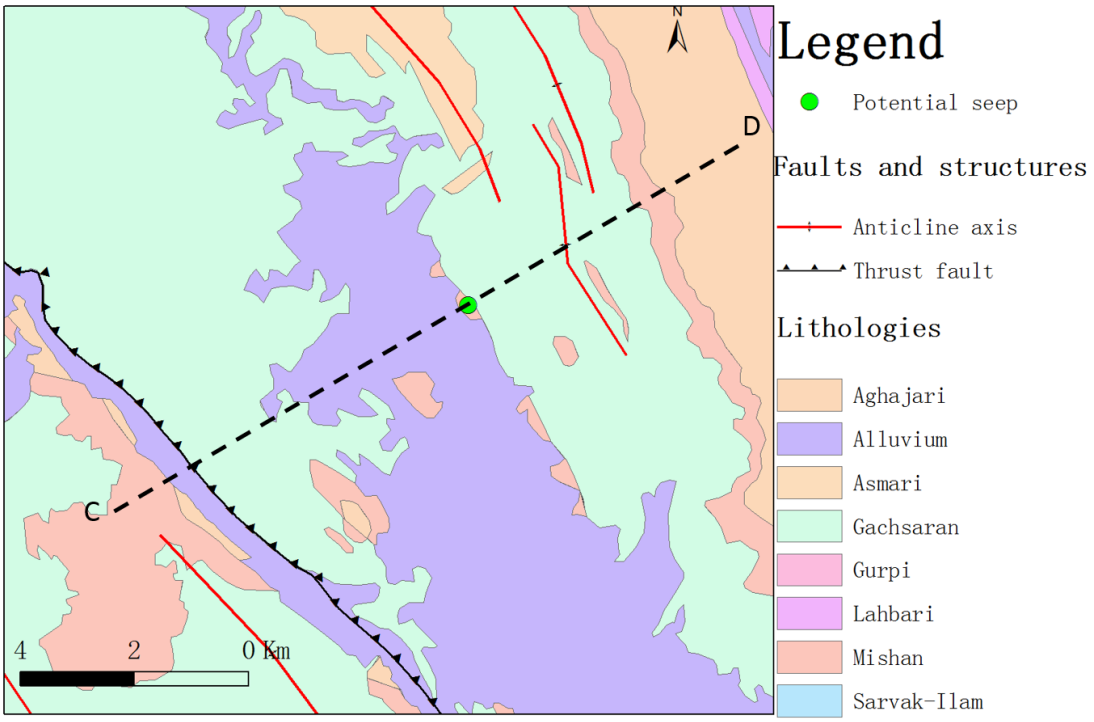
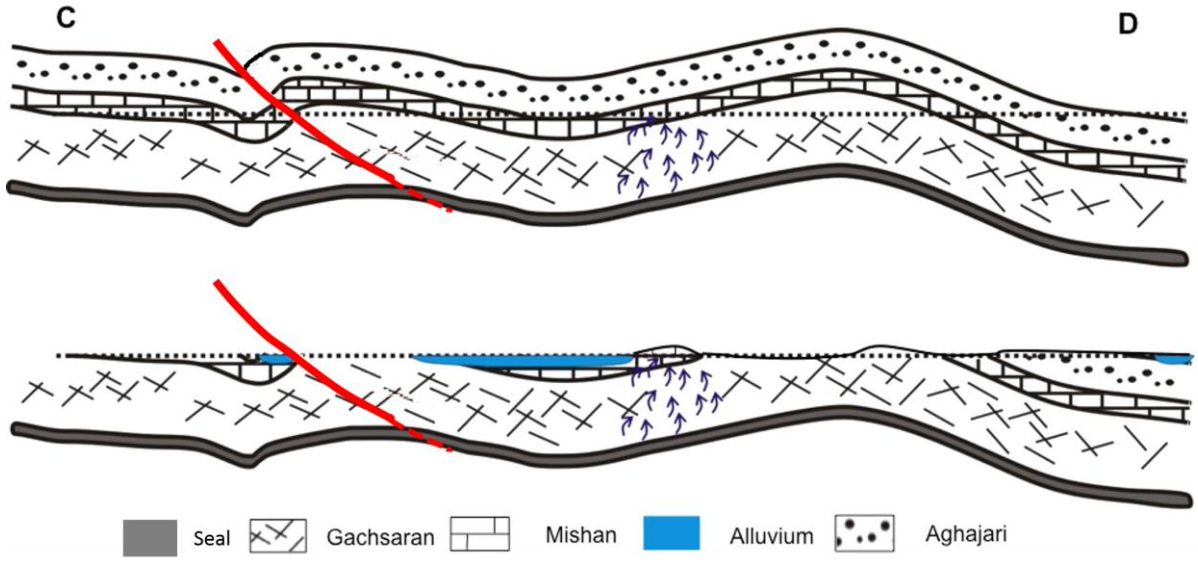


Figure 6.4 (a) Geological map of area C (see figure 5.4), (b) AB cross section through the center of potential seep. The location of the cross-section is shown in (a).

Figure 6.4 shows a geological model explaning how gas might escape from the seal and enter the Gachsaran formation. Due to strong squeezing action a thrust fault was formed. Moreover, a fracture system was formed in Gachsaran formation. Gas migrated upward from these fractures to the permeable limestone formation (Mishan formation). Furthermore, gas seeps alterations were formed in the outcrop of Mishan formation. Thus, Area C is interpreted as potential alteration.



(a)



(b)

Figure 6.5 (a) Geological map of area A (see figure 5.4), (b) CD cross section through the center of potential seep, the location of section line is shown in (a).

Area A has the same oil-gas migration system as area C. But the distance between potential seep and known fault is long. However, the SVM classification result (figure 5.1 and 5.4) and the spectrum information (figure 5.5) imply that area A might be gas-induced alteration. Probably, there might be unknown faults not indicated in the geological map, or the gypsum in area A might be transported from other formations.

6.4. Conclusion

In this chapter, we compared our work with Salati (2014)'s work. We found there was a bias between ASTER and WV-2 imagery. For the atmospheric correction method, the FLAASH Model replaced logarithmic residuals method. Furthermore, altered samples contain clay minerals than unaltered samples in this research, which is opposite to the interpretation of Salati. In addition, the number of trees in GBRT were tested in this research so that the optimal trees were used to build GBRT model. Calculating the optimal trees would improve the classification accuracy.

Moreover, two geologic models were built depending on geological map. That area C was observed to be potential alterations are supported by geologic model. However, the fault is far from area A so that area A need find new evidence to proof it is altered.

7. CONCLUSION

In this research remote sensing data and machine learning methods were used to produce hydrocarbon seepages induced alterations maps. Based on the above chapters conclusions are listed below:

Our laboratory analysis showed that the altered samples could be distinguished from the unaltered samples in two ways. First, gypsum and clays are two kinds of dominating minerals in altered field samples, while unaltered samples were identified as calcite with clays. Second, two water absorption features in altered samples, 1.4μm and 1.9μm, were observed to be deeper than in unaltered samples.

Based on the result of both laboratory and image-driven spectral analysis (see chapter 3), alterations were observed to have high concentration of ferrous iron, gypsum and clays, and have low concentration of carbonates and ferric iron. Among these minerals, gypsum is a typical indicator of alteration. For ASTER data, through comparing the difference of each index between altered and unaltered field samples, the ferric oxide index, clay index, calcite index and gypsum index (see table 3.1) were found to be most suitable indicators to show the difference between altered and unaltered areas. For WorldView-2 data, the sulfur index, iron index and ferric iron index (see table 3.3) were selected with the same method with ASTER data. Through composition of these indexes, a number of colour composites (see table 3.2 and 3.4) were produced to show the alterations. The result confirms that ASTER and WorldView-2 can be used in a meaningful way for mapping alterations. The SWIR bands of ASTER contain the most useful wavelengths for distinguishing gypsum and carbonates. Although WorldView-2 data can be used to improve the spatial resolution and improve the result by comparing result with ASTER data, the classification maps produced by ASTER did not improve when combining with WV-2. Therefore, alteration maps could be produced by using ASTER imagery only, which is more economical.

Furthermore, both the knowledge-based approach (band ratio and relative absorption band depth) and the data driven approach (SVM, RF and GBRT) used in this research proved to be successful in detecting alterations. The alteration maps (figure 4.4 and 4.5) produced by the knowledge-based approach were used to select endmembers for training the data driven models. Through comparing the overall accuracy and patterns of alterations, the SVM method was found to be the most suitable classifier we used for detecting hydrocarbon seepage alteration. SVM is especially suited for small training sets, and the classification result is stable, unlike RF and GBRT. Because our classification maps only have two classes, the SVM method used in this research is not an ensemble classifier. In other words, if the feature space is not complex, in the case of two classes, the performance of SVM is observed to be better than the ensemble classifiers RF and GBRT. Therefore, this research cannot confirm that ensemble classifiers perform better than a single classifier.

Comparing the results of this research with the work of Salati et al. (2014), the classification accuracy has improved considerably. In general four observation can be made. (1) The FLAASH model was successfully used to pre-process ASTER and WorldView-2 data and may be superior to the logarithmic residuals correction used by Salati. (2) There is an about 60 meters bias between ASTER and WorldView-2 data. This observed geometric shift may have led to misclassifications in the work of Salati. (3) The comparison of alteration maps produced by the knowledge-based approach used in this research with the alteration maps obtained by Salati et al. (2014) shows similar seepage locations, but different discrimination of minerals associated with the alterations. That clay minerals were identified to be higher in altered samples than unaltered samples in this research is opposite to the interpretation of Salati. This difference could lead to the selection of incorrect endmembers so that the accuracy of classification result may have been reduced in Salati's work. (4) Since the GBRT classifier has an overfitting problem, this classifier must be tested for the optimal number of trees instead of simply choosing a number of trees. The number of 1000 trees may have led to overfitting in Salati's work. We found that 227 trees gave an optimal classification result.

In this research, regional gas-induced alteration map has been produced by SVM and ASTER data with an overall accuracy of 85 percent. Moreover, next to the known alterations, two new areas were interpreted as potential hydrocarbon seepage alterations.

LIST OF REFERENCES

Abrams, M. A. (2005). Significance of hydrocarbon seepage relative to petroleum generation and entrapment. *Marine and Petroleum Geology*, *22*(4), 457–477. http://doi.org/10.1016/j.marpetgeo.2004.08.003

Adler-Golden, S. M., Matthew, M. W., Bernstein, L. S., Levine, R. Y., Berk, A., Richtsmeier, S. C., Acharya, Prabhat K., Anderson, Gail P., Felde, Jerry W., Gardner, J. A., Hoke, Michael L., Jeong, Laila S., Pukall, Brian, Ratkowski, Anthony J., Burke, H. K. (1999). Atmospheric correction for shortwave spectral imagery based on MODTRAN4. In M. R. Descour & S. S. Shen (Eds.), *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation* (pp. 61–69). International Society for Optics and Photonics. http://doi.org/10.1117/12.366315

Anonymous (2009). *Atmospheric Correction Module : QUAC and FLAASH User's Guide* (p. 44). ITT Visual Information Solutions. Retrieved October 10, 2015, from https://www.exelisvis.com/portals/0/pdfs/envi/Flaash_Module.pdf

Ali Sebtosheikh, M., & Salehi, A. (2015). Lithology prediction by support vector classifiers using inverted seismic attributes data and petrophysical logs as a new approach and investigation of training data set size effect on its performance in a heterogeneous carbonate reservoir. *Journal of Petroleum Science and Engineering*, PETROL5323. http://doi.org/10.1016/j.petrol.2015.08.001

Alimohammadi, M., Alirezaei, S., & Kontak, D. J. (2015). Application of ASTER data for exploration of porphyry copper deposits: A case study of Daraloo–Sarmeshk area, southern part of the Kerman copper belt, Iran. *Ore Geology Reviews*, *70*, 290–304. http://doi.org/10.1016/j.oregeorev.2015.04.010

Bedini, E. (2011). Mineral mapping in the Kap Simpson complex, central East Greenland, using HyMap and ASTER remote sensing data. *Advances in Space Research*, *47*(1), 60–73. http://doi.org/10.1016/j.asr.2010.08.021

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. http://doi.org/10.1023/A:1010933404324

Ceron, J., Lombo, C., Williams, S., & Bain, J. (2001). Effective use of non-seismic methods for petroleum exploration. *First Break*, *19*(9), 523–528. Retrieved from https://www.engineeringvillage.com/blog/document.url?mid=geo_1630ab910af2577bd7e1120613774210&database=geo

Clark, R. N., Swayze, G. A., Wise, R., Livo, K. E., Hoefen, T. M., Kokaly, R. F., Sutley, S J. (2007). USGS digital spectral library splib06a. Retrieved from http://speclab.cr.usgs.gov/spectral.lib06

Cortes, C., & Vapnik, V. (1995). Support-Vector networks. *Machine Learning*, *20*(3), 273–297. http://doi.org/10.1023/A:1022627411411

CSIRO. (2013). CSIRO data access portal - satellite ASTER geoscience map of australia. Retrieved May 25, 2015, from https://data.csiro.au/dap/landingpage?pid=csiro:6182

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *The Journal of Animal Ecology*, *77*(4), 802–13. http://doi.org/10.1111/j.1365-2656.2008.01390.x

Etiope, G. (2015). *Natural Gas Seepage : The Earth's Hydrocarbon Degassing*. Cham: Springer International Publishing.

Fern, M., & Cernadas, E. (2014). Do we need hundreds of classifiers to solve real world classification problems ?, *Journal of Machine Learning Research*, 15(2014), 3133-3181.

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, *22*(9), 1365–81. http://doi.org/10.1002/sim.1501

Gomez, C., Delacourt, C., Allemand, P., Ledru, P., & Wackerle, R. (2005a). Using ASTER remote sensing data set for geological mapping, in Namibia. *Physics and Chemistry of the Earth, Parts A/B/C*, *30*(1-3), 97–108. http://doi.org/10.1016/j.pce.2004.08.042

Horgan, B. H. N., Cloutis, E. a., Mann, P., & Bell, J. F. (2014). Near-infrared spectra of ferrous mineral mixtures and methods for their identification in planetary surface spectra. *Icarus*, *234*, 132–154. http://doi.org/10.1016/j.icarus.2014.02.031

Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, *23*(4), 725–749. http://doi.org/10.1080/01431160110040323

Kalinowski, A., & Oliver, S. (2004). *ASTER Mineral Index Processing Manual Compiled* (p. 37). Retrieved July 25, 2015, from http://www.ga.gov.au/webtemp/image_cache/GA7833.pdf

Lammoglia, T., & de Souza Filho, C. R. (2013). Unraveling hydrocarbon microseepages in onshore basins using spectral–spatial processing of advanced spaceborne thermal emission and reflection radiometer (ASTER) data. *Surveys in Geophysics*, *34*(3), 349–373. http://doi.org/10.1007/s10712-013-9225-3

Lawrence, R. (2004). Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, *90*(3), 331–336. http://doi.org/10.1016/j.rse.2004.01.007

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News: The Newsletter of the R Project*, *2(3)*, 18–22. Retrieved from http://www.bios.unc.edu/~dzeng/BIOS740/randomforest.pdf

Lowe, B., & Kulkarni, A. (2015). Multispectral image analysis using random forest. *International Journal on Soft Computing*, *6*(1), 1–14. http://doi.org/10.5121/ijsc.2015.6101

Ma, F., Qin, H., Shi, K., Zhou, C., Chen, C., Hu, X., & Zheng, L. (2016). Feasibility of combining spectra with texture data of multispectral imaging to predict heme and non-heme iron contents in pork sausages. *Food Chemistry*, *190*, 142–9. http://doi.org/10.1016/j.foodchem.2015.05.084

Macgregor, D. S. (1993). Relationships between seepage, tectonics and subsurface petroleum reserves. *Marine and Petroleum Geology*, *10*(6), 606–619. http://doi.org/10.1016/0264-8172(93)90063-X

Mao, S., Jiao, L., Xiong, L., Gou, S., Chen, B., & Yeung, S.-K. (2015). Weighted classifier ensemble based on quadratic form. *Pattern Recognition*, *48*(5), 1688–1706. http://doi.org/10.1016/j.patcog.2014.10.017

Mulder, V. L., de Bruin, S., Schaepman, M. E., & Mayr, T. R. (2011). The use of remote sensing in soil and terrain mapping — A review. *Geoderma*, *162*(1-2), 1–19. http://doi.org/10.1016/j.geoderma.2010.12.018

Mutanga, O., Adam, E., & Cho, M. A. (2012). High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation*, *18*, 399–406. http://doi.org/10.1016/j.jag.2012.03.012

Nairn, A. E. M., & Alsharhan, A. S. (1997). *Sedimentary Basins and Petroleum Geology of the Middle East*. Burlington: Elsevier Science.

Pal, M., & Mather, P. M. (2006). Some issues in the classification of DAIS hyperspectral data. *International Journal of Remote Sensing*, *27*(14), 2895–2916. http://doi.org/10.1080/01431160500185227

Parisien, M.-A., & Moritz, M. A. (2009). Environmental controls on the distribution of wildfire at multiple spatial scales. *Ecological Monographs*, *79*(1), 127–154. http://doi.org/10.1890/07-1289.1

Pirkle, R. J., & Jones, V. T. (2006). Applications of petroleum exploration and environmental geochemistry to carbon sequestration. *American Geophysical Union*. Retrieved from http://adsabs.harvard.edu/abs/2006AGUFM.H14A..02P

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*. Retrieved February 08, 2016, from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639

Robinson, T. P., Wardell-Johnson, G. W., Pracilio, G., Brown, C., Corner, R., & van Klinken, R. D. (2016). Testing the discrimination and detection limits of WorldView-2 imagery on a challenging invasive plant target. *International Journal of Applied Earth Observation and Geoinformation*, *44*, 23–30. http://doi.org/10.1016/j.jag.2015.07.004

Rogge, D. M., Rivard, B., Zhang, J., Sanchez, A., Harris, J., & Feng, J. (2007). Integration of spatial–spectral information for the improved extraction of endmembers. *Remote Sensing of Environment*, *110*(3), 287–303. http://doi.org/10.1016/j.rse.2007.02.019

Rokach, L. (2009). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1–39. http://doi.org/10.1007/s10462-009-9124-7

Sabins, F. F. (1999). Remote sensing for mineral exploration. *Ore Geology Reviews*, *14*(3-4), 157–183. http://doi.org/10.1016/S0169-1368(99)00007-4

Salati, S. (2014). Characterization and remote detection of onsore hydrocarbon seep-induced alternation. Enschede, The Netherlands: University of Twente Faculty of Geo-Information and Earth Observation (ITC). http://doi.org/10.3990/1.9789036536295

Salati, S., van Ruitenbeek, F., van der Meer, F., & Naimi, B. (2014, April 10). Detection of alteration induced by onshore gas seeps from ASTER and WorldView-2 Data. *Remote Sensing*. Retrieved from http://www.mdpi.com/2072-4292/6/4/3188

Segal, M. R., Dahlquist, K. D., & Conklin, B. R. (2004). Regression approaches for microarray data analysis. *Journal of Computational Biology*, *10*(6), 961–980. http://doi.org/10.1089/106652703322756177

Selley, R. C. (1992). Petroleum seepages and impregnations in Great Britain. *Marine and Petroleum Geology*, *9*(3), 226–244. http://doi.org/10.1016/0264-8172(92)90072-M

Shahdoosti, H. R., & Ghassemian, H. (2016). Combining the spectral PCA and spatial PCA fusion methods by an optimal filter. *Information Fusion*, *27*, 150–160. http://doi.org/10.1016/j.inffus.2015.06.006

Shi, P., Fu, B., Ninomiya, Y., Sun, J., & Li, Y. (2012). Multispectral remote sensing mapping for hydrocarbon seepage-induced lithologic anomalies in the Kuqa foreland basin, south Tian Shan. *Journal of Asian Earth Sciences*, *46*, 70–77. http://doi.org/10.1016/j.jseaes.2011.10.019

Tangestani, M. H., & Validabadi, K. (2014). Mineralogy and geochemistry of alteration induced by hydrocarbon seepage in an evaporite formation; a case study from the Zagros Fold Belt, SW Iran. *Applied Geochemistry*, *41*, 189–195. http://doi.org/10.1016/j.apgeochem.2013.12.015

Thomas Cudahy (2012). *Geoscience Product Notes for Australia* (pp. 1–26). Retrieved May 08, 2015, from http://c3dmm.csiro.au/Australia_ASTER/Australian ASTER Geoscience Product Notes FINALx.pdf

Tian, S., Chen, J., Jiang, M. (2008). Spaceflight Hyperion data radiation calibration preliminary. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVII, 363–368.

Van der Meer, F. D., van der Werff, H. M. A., van Ruitenbeek, F. J. A., Hecker, C. A., Bakker, W. H., Noomen, M. F., van der Meijde, Mark, Carranza, E. John M., Smeth, J. Boudewijn de, Woldai, T. (2012). Multi- and hyperspectral geologic remote sensing: A review. *International Journal of Applied Earth Observation and Geoinformation*, *14*(1), 112–128. http://doi.org/10.1016/j.jag.2011.08.002

Van der Meer, F., van Dijk, P., van der Werff, H., & Yang, H. (2002). Remote sensing and petroleum seepage: a review and case study. *Terra Nova*, *14*(1), 1–17. http://doi.org/10.1046/j.1365-3121.2002.00390.x

Walton, J. T. (2008). Subpixel urban land cover estimation: comparing cubist, random forests and support vector regression. *Photogrammetric Engineering & Remote Sensing*, *74*(10), 1213–1222.

Whiteside, T. G., & Bartolo, R. E. (2015). Use of WorldView-2 time series to establish a wetland monitoring program for potential offsite impacts of mine site rehabilitation. *International Journal of Applied Earth Observation and Geoinformation*, *42*, 24–37. http://doi.org/10.1016/j.jag.2015.05.002

Zou, C., Zhai, G., Zhang, G., Wang, H., Zhang, G., Li, J., Wang, Zhaoming, Wen, Zhixin, MA, Feng,Liang, Yingbo, Yang, Zhi, Li, Xin, Liang, K. (2015). Formation, distribution, potential and prediction of global conventional and unconventional hydrocarbon resources. *Petroleum Exploration and Development*, *42*(1), 14–28. http://doi.org/10.1016/S1876-3804(15)60002-7

APPENDIX 1

The entire code contains variables and functions definition, legend drawing, image reading and displaying, training set and test set reading, svm, rf and gbrt modeling, image prediction and displaying and accuracy assessment. To show the parameters and functions in SVM, RF and GBRT, Pseudo-code of modelling part are shown in the following:

Pseudo-code of SVM, RF and GBRT

Classifier1: Supported Vector Machine

Package: kernlab
Input:
ASTER image
Traning set
Test set
Output:
SVM classification result
Confusion matrix
Overall accuracy
Parameters:

```
class            # read from training set
image            # read from ASTER image
Features <- c(1:9)    # 9 bands of ASTER
Features1 <- c(1:10)  # 9 bands of ASTER and class
C_SVM <- 20       # complexity of model
Formula <- TR$ class
```

Kernel: Gaussian RBF kernel
Feature space plotting:

```
for(j in 1:9) {
   for(k in 2:9) {
      windows()
      Sel_Features2 <- c(j,k,10)
      svm_model <- ksvm(as.factor(class)~., data=TR[,Features1], type="C-svc",
      kernel="rbfdot", C=C_SVM, cross=3, prob.model=TRUE)
      # plot function in kernlab package used to display feature space in this research
      plot(svm_model, data=TR[,Sel_Features1]) }
}
```

SVM modeling:

```
# ksvm function (formula, data, type, kernel, C, cross, prob.model = TRUE or FALSE)
svm_model <- ksvm(class ~., data=TR[,Features1], type="C-svc", kernel=" rbfdot ",
C=C_SVM, cross=3, prob.model=TRUE)
```

SVM applying:

```
# predict function (model, data, type)
SVM <- predict(svm_model, image, type="probabilities")
```

Display and save SVM classification result
Accuracy assessment:
Confusion_matrix <- confusion_matrix(TS, Class_legend)
END

Classifier2: Random Forest

Package: randomForest
Input:
ASTER image
Training set
Test set
Output:
RF classification result
Confusion matrix
Overall accuracy
Parameters:

```
class          # read from training set
image          # read from ASTER image
Features <- c(1:9)     # 9 bands of ASTER
Features1 <- c(1:10)   # 9 bands of ASTER and class
ntree          # number of trees
mtry           # predictor variables
Formula <- as.factor (TR$ class)
```

RF modelling:

```
# randomForest function (formula, data, ntree, mtry, type, norm.votes =TRUE or FALSE,
proximity= TRUE or FALSE)
rf_model <- randomForest(as.factor(class)~., data = TR[,Features1], ntree=1000, mtry=3,
type="classification", norm.votes=TRUE, proximity=TRUE)
```

Major parameter testing:

```
#plot the relation between out of bag error and number of trees
plot(oob[], rf_model$err.rate[,1],type='l')
# plot the relation between out of bag error and mtry
tuneRF(TR, as.factor(TR$class_id), mtryStart= 3, stepFactor=1.5)
# plot the variable importance
varImpPlot(rf_model)
```

RF applying:

```
# predict function (model, data, type)
RF <- predict(rf_model, image, type="response")
```

Display and save SVM classification result
Accuracy assessment:
Confusion_matrix <- confusion_matrix(TS, Class_legend)
END

Classifier3: Gradient Boosted Regression Tree

Package: gbm

Input:

ASTER image
Training set
Test set

Output:

RF classification result
Confusion matrix
Overall accuracy

Parameters:

```
class          # read from training set
image          # read from ASTER image
Features <- c(1:9)    # 9 bands of ASTER
n.tree         # number of trees
y <- TR$ class
```

RF modelling:

```
# gbm.fit function (y, x, n.tree, distribution, interaction.depth, n.minobsinnode, shrinkage,
bag.fraction)
gbrt_model <- gbm.fit(y=y,x = TR[,Sel_Features], n.tree=2000,
                distribution = "multinomial", # classification type
                interaction.depth = 3,   # model with three interaction
                n.minobsinnode = 10,   # ten observations in each tree
                shrinkage = 0.01,      # learning rate
                bag.fraction = 0.5)     # randomly select half training set to build next tree
```

Major parameter testing:

```
# test the optimal number of trees
best.iter <- gbm.perf(brt_model,plot.it= TRUE, method="OOB")
print(best.iter)
# print the variable importance
Prob_train <- array(0,c(nrow(tmp_train),Ncl))
for(k in 1:Ncl)Prob_train[,k] <- tmp_train[,k,1]
TR$classifier=max.col(Prob_train)
summary(gbrt_model, n.trees=best.iter)
```

RF applying:

```
# predict function (model, data, n.trees, type)
GBRT <- predict(gbrt_model, image ,n.trees = best.iter, type="response")
```

Display and save SVM classification result

Accuracy assessment:

```
Confusion_matrix <- confusion_matrix(TS, Class_legend)
```

END