Statistical analysis of radar and hyperspectral remote sensing data

Deok Han

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Electrical Engineering
in the Department of Electrical and Computer Engineering

Mississippi State, Mississippi

May 2016
Statistical analysis of radar and hyperspectral remote sensing data

By

Deok Han

Approved:

Qian (Jenny) Du
(Major Professor)

Nicolas H. Younan
(Co-Major Professor)

Tung-Lung Wu
(Minor Professor)

James V. Aanstoos
(Committee Member)

Farshid Vahedifard
(Committee Member)

James E. Fowler
(Graduate Coordinator)

Jason M. Keith
Dean
Bagley College of Engineering
In this dissertation, three studies were done for radar and hyperspectral remote sensing applications using statistical techniques. The first study investigated a relationship between synthetic aperture radar backscatter and in situ soil properties for levee monitoring. A series of statistical analyses were performed to investigate potential correlations between three independent polarization channels of radar backscatter and various soil properties. The results showed a weak but considerable correlation between the cross-polarized (HV) radar backscatter coefficients and several soil properties.

The second study performed effective statistical feature extraction for levee slide classification. Images about a levee are often very large, and it is difficult to monitor levee conditions quickly because of high computational cost and large memory requirement. Therefore, a time-efficient method to monitor levee conditions is necessary. The traditional support vector machine (SVM) did not work well on original radar images with three bands, requiring extraction of discriminative features. Gray level co-occurrence matrix is a powerful method to extract textural information from grey-scale images, but it may not be practical for a big data in terms of calculation time. In this study, very efficient feature extraction methods with spatial filtering were used, including
a weighted average filter and a majority filter in conjunction with a nonlinear band normalization process. Feature extraction with these filters, along with normalized bands, yielded comparable results to gray level co-occurrence matrix with a much lower computational cost.

The third study focused on the case when only a small number of ground truth labels were available for hyperspectral image classification. To overcome the difficulty of not having enough training samples, a semisupervised method was proposed. The main idea was to expand ground truth using a relationship between labeled and unlabeled data. A fast self-training algorithm was developed in this study. Reliable unlabeled samples were chosen based on SVM output with majority voting or weighted majority voting, and added to labeled data to build a better SVM classifier. The results showed that majority voting and weighted majority voting could effectively select reliable unlabeled data, and weighted majority voting yielded better performance than majority voting.
DEDICATION

To my mother, anchor of my families.
ACKNOWLEDGEMENTS

I would like to thank Dr. Qian Du for her advice and guidance from her rich experience of prolific research and expertise on machine learning and remote sensing as a renowned scholar. I would like to thank Dr. Nicolas H. Younan for his guidance through my graduate studies, and for being a mentor for many graduate students. I would like to give special thanks to Dr. James V. Aanstoos. Because I was a member of his research team, I was able to obtain hands-on experience with remotes sensing, in addition to having large-scale research opportunities. I thank Dr. Farshid Vahedifard for his suggestion on my research from his energetic attitude and enthusiasm on research. I also thank Dr. Tung-Lung Wu for kindly becoming my committee member for my minor degree in statistics.

Finally, I would like to thank my wife for her support.
# TABLE OF CONTENTS

DEDICATION .............................................................................................................................................. ii

ACKNOWLEDGEMENTS .......................................................................................................................... iii

LIST OF TABLES ......................................................................................................................................... vi

LIST OF FIGURES ...................................................................................................................................... vii

CHAPTER

I. INTRODUCTION ........................................................................................................................................ 1

II. CORRELATION ANALYSIS BETWEEN RADAR BACKSCATTER AND IN SITU SOIL PROPERTY MEASUREMENTS .......................................................... 6  
   2.1 Background ......................................................................................................................................... 6  
   2.2 Materials and Methods .......................................................................................................................... 8  
      2.2.1 Study Area .................................................................................................................................... 9  
      2.2.2 Field Investigation .......................................................................................................................... 10  
      2.2.3 In Situ Soil Property Measurements ............................................................................................... 11  
   2.3 PolSAR Imagery ................................................................................................................................... 18  
   2.4 Statistical Analysis ............................................................................................................................... 20  
   2.5 Results and Discussion ......................................................................................................................... 21

III. CLASSIFICATION OF LEVEE SLIDES FROM AIRBORNE SYNTHETIC APERTURE RADAR IMAGES WITH EFFICIENT SPATIAL FEATURE EXTRACTION ........................................................................ 25  
   3.1 Background ......................................................................................................................................... 25  
   3.2 Data ..................................................................................................................................................... 27  
   3.3 Method ............................................................................................................................................... 29  
      3.3.1 GLCM and Band Normalization ....................................................................................................... 30  
      3.3.2 Spatial Filtering ............................................................................................................................... 32  
   3.4 Experiment .......................................................................................................................................... 33  
   3.5 Discussion .......................................................................................................................................... 40

IV. SEMISUPERVISED CLASSIFICATION OF HYPERSPECTRAL REMOTE SENSING IMAGES .................................................................................. 42
4.1 Background.................................................................................................................42
4.2 Semisupervised classification....................................................................................43
  4.2.1 Self-training method..........................................................................................44
  4.2.2 Co-training method...........................................................................................45
  4.2.3 Generative model...............................................................................................45
  4.2.4 Graph-based method.........................................................................................46
4.3 Proposed method.......................................................................................................46
4.4 Hyperspectral image data..........................................................................................51
  4.4.1 Salinas A-scene..................................................................................................52
  4.4.2 Indian Pines.......................................................................................................52
  4.4.3 Pavia Centre and Pavia University......................................................................52
4.5 Results and discussion...............................................................................................53

V. CONCLUSION AND FUTURE WORK........................................................................64

REFERENCE.....................................................................................................................67

APPENDIX

A. SCATTER PLOT BETWEEN SAR DATA AND IN SITU SOIL PROPERTIES......................77
  A.1 Scatter plot between SAR and in situ soil properties on the river side and the land side.............................................................................................................78
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Statistics of measured in situ soil properties for the river side of study area</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Statistics of measured in situ soil properties for the land side of study area</td>
<td>14</td>
</tr>
<tr>
<td>2.3</td>
<td>Correlation between radar backscatter and soil properties for the surface layer</td>
<td>20</td>
</tr>
<tr>
<td>2.4</td>
<td>Correlation between radar backscatter and soil properties for the subsurface layer</td>
<td>21</td>
</tr>
<tr>
<td>3.1</td>
<td>The major features of GLCM used in this research</td>
<td>32</td>
</tr>
<tr>
<td>3.2</td>
<td>Parameter settings for both GLCM and spatial filtering based feature extraction</td>
<td>33</td>
</tr>
<tr>
<td>3.3</td>
<td>Classification Performance when Using 30% for training (and 70% for testing)</td>
<td>38</td>
</tr>
<tr>
<td>3.4</td>
<td>Execution time in seconds</td>
<td>39</td>
</tr>
<tr>
<td>4.1</td>
<td>Windows size and threshold for confidence level</td>
<td>53</td>
</tr>
<tr>
<td>4.2</td>
<td>Selected unlabeled data</td>
<td>62</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

2.1 Study area. ......................................................................................................10

2.2 A cone penetrometer in conjunction with all-terrain vehicle-mounted........11

2.3 Cone penetrometer sampling locations. ........................................................12

2.4 Histograms of clay and sand fraction for surface and subsurface layers...............................................................................................................15

2.5 Histograms of $q_s$ and $q_c$ for surface and subsurface layers..................16

2.6 Histograms of $K_{sat}$ and $f_c$ for surface and subsurface layers.................17

2.7 Histograms of pwp and n for surface and subsurface layers......................17

2.8 Soil moisture of surface and subsurface layers. ...........................................18

2.9 Soil moisture of selected months of 2011. ...................................................18

2.10 Histogram of SAR magnitude (dB). .............................................................19

3.1 Study area. ....................................................................................................28

3.2 Slide location. .................................................................................................29

3.3 SVM classification results of two subimage scenes.................................35

3.4 Classification map of two subimage scenes. ..................................................36

4.1 Effect of using unlabeled data. .....................................................................44

4.2 Majority voting and confidence level (non-majority center pixel). ..............47

4.3 Majority voting and confidence level (majority center pixel).......................48

4.4 Weighted majority voting and confidence level (majority center pixel)........49

4.5 Procedure for semisupervised classification. ..............................................50

4.6 Algorithm of self-training............................................................................51
4.7 Classification images of Indian Pines ............................................................54
4.8 Classification result of Indian Pines ............................................................55
4.9 Classification images of Salinas-A scene ......................................................56
4.10 Classification result of Salinas-A scene ........................................................57
4.11 Classification images of Pavia University ....................................................58
4.12 Classification result of Pavia University ......................................................59
4.13 Classification images of Pavia Centre ..........................................................60
4.14 Classification result of Pavia Centre .............................................................61
4.15 Comparison of MV and WMV for the same number of added confident unlabeled data ................................................................................63
A.1 Scatter plot of q_s and q_c with HH, HV and VV channels on the land side .................................................................................................................79
A.2 Scatter plot of clay and sand fraction with HH, HV and VV channels on the land side ..............................................................................................80
A.3 Scatter plot of K_{sat}, and f_c with HH, HV and VV channels on the land side .................................................................................................................81
A.4 Scatter plot of pwp and saturation with HH, HV and VV channels on the land side ...........................................................................................................82
A.5 Scatter plot of n with HH, HV and VV channels on the land side ..................83
A.6 Scatter plot of q_s and q_c with HH, HV and VV channels on the river side .........................................................................................................................84
A.7 Scatter plot of clay and sand fraction with HH, HV and VV channels on the river side ........................................................................................................85
A.8 Scatter plot of K_{sat} and f_c with HH, HV and VV channels on the river side .........................................................................................................................86
A.9 Scatter plot of pwp and saturation with HH, HV and VV channels on the river side .........................................................................................................87
A.10 Scatter plot of n with HH, HV and VV channels on the river side ...............88
CHAPTER I
INTRODUCTION

Radar and hyperspectral remote sensing data have been successfully applied to many applications including military, agriculture, mineralogy, and environmental monitoring. In situ measurement has been a basic method to obtain accurate information on a study area. Usually, remote sensing methods can cover large areas, and their data acquisitions are cost affordable and fast. On the contrary, in situ measurement is time consuming, and it is sometimes cost prohibitive.

In this dissertation, three related studies about statistical learning have been conducted [1-3]. A relationship between synthetic aperture radar (SAR) and in situ soil measurement properties has been explored, and classification with fast feature extraction for levee monitoring was performed on SAR data [4]. Finally, an effective semisupervised classification algorithm was developed for hyperspectral images [5].

The first study investigated a statistical relationship between SAR and in situ soil physical properties. Three independent polarization channels (i.e., HH, HV and VV) and eight in situ soil physical properties were examined for correlation. Three polarization data were acquired by an airborne SAR instrument (the NASA JPL’s Uninhabited Aerial Vehicle Synthetic Aperture Radar (UAVSAR) [6-8]) and eight in situ soil physical properties were obtained by a cone penetrometer machine and core sampling. The eight soil properties were penetration resistance parameters (sleeve friction and cone tip...
resistance), clay fraction, sand fraction, saturated hydraulic conductivity, field capacity, permanent wilting point, and porosity. Past studies relating remote sensing data to in situ soil properties have primarily focused on soil moisture, roughness and texture [9-17].

Extensive review for remote sensing methods (including active and passive microwave and optical imaging) to monitor soil moisture and surface roughness can be found in [18]. A study investigating a relationship between TerraSAR-X and clay fraction showed that they had a linear relationship [19]. Other studies have explored a relationship between various soil properties, based on in situ measurements alone [20-23]. No studies have investigated any direct or indirect relationship between radar backscatter and more extensive sets of soil physical properties, such as the eight properties mentioned earlier. The main contribution of this study was to investigate statistical relationships between radar backscatter and extensive eight soil physical properties. Knowledge of possible relationships between these soil properties and radar backscatter could allow estimates of their range to be mapped in radar data, and when combined with limited in situ measurements may reduce uncertainty in such maps economically.

The second study performed effective classification with fast feature extraction along with added normalized bands for levee monitoring [24-28]. About 200,000 km of earth levees stretch through the United States. Formidable failure of levees in New Orleans during Hurricane Katrina along the Mississippi River highlighted the importance of levee monitoring [29-31]. Research on screening levees has been conducted at Mississippi State University [26, 29, 32, 33]. Airborne and spaceborne SAR images were used to monitor the abnormality of study areas along the Mississippi River. In this study, an airborne SAR having three magnitude bands with HH, VV, and HV polarization were
used for levee monitoring, and supervised classification for slide and non-slide levee areas was performed using SVM classifier [1, 2, 34-36]. The original three bands failed to yield a satisfying classification result for slide and non-slide classification. The original low-dimensional data might not include enough discriminant features. Thus, unlike a dimensionality reduction process for high-dimensional data [37-39], a dimensionality expansion process was included by adding additional artificial bands [40, 41], which were simply nonlinear when using normalized bands in this study. In addition, a feature extraction method, i.e., gray-level co-occurrence matrix (GLCM), for spatial information has been considered before classification. GLCM is one of the popular methods to describe spatial features, and it has performed well in many applications [29, 42-46]. However, it requires significant computational power since several position and direction operators have to be applied, and statistical features must be calculated from the GLCM. Thus, it may not be suitable for fast analysis of a large-scale image. In this study, very simple and effective feature extraction methods, i.e., a weighted average filter and a majority filter, were used. The weighted average filter and the majority filter [47] offered slightly lower classification accuracy with much less computational cost. Such spatial lowpass filtering techniques with a sliding window are suitable for parallel computing, since the output of a single pixel is unrelated to other areas of the image [48]. Thus, such spatial feature extraction methods are preferred for fast processing of large-scale remote sensing images. In addition, including additional normalized bands may improve the performance of GLCM and spatial filtering. The contribution of this study was to reduce the computational burden while maintaining comparable performance of GLCM by using spatial lowpass filters and normalized bands of three SAR bands.
The last study conducted semisupervised classification for hyperspectral images. Optical hyperspectral bands were studied for this topic due to the fact that more land cover and land use classes were available in image data for testing classification performance more comprehensively. Unlike SAR imagery with several bands, hyperspectral images can have more than 200 bands, and their rich spectral information has led to many successful applications including astronomy, agriculture, mineralogy, environmental monitoring, and surveillance. The preprocessing step of band selection or feature reduction is required to avoid the problem of curse of dimensionality, known as the Hughes phenomenon [49-52], and feature extraction may be required for better image interpretation [55]. Classical classification involves either supervised or unsupervised methods. Semisupervised learning is relatively new, and it utilizes unlabeled samples when labeled samples are limited [5, 56-58]. Semisupervised learning is about building a better learning machine by using labeled and unlabeled data together during a training process. Most data are unlabeled. However, labeling those data is not easy. It requires expertise on specific data, and it can be time consuming. Sometimes, it is dangerous or impossible to label data. Therefore, it is intuitive to utilize abundant unlabeled data along with labeled data in a learning process. There are several different approaches for semisupervised learning: self-training, co-training, generative model, and graph-based model. There are also special SVMs for semisupervised models. In this study, the self-training method has been explored for hyperspectral images. The self-training method selects reliable or confident unlabeled data from a classifier output, and those confident samples are added to a labeled data set to build a better classifier. This process can be repeated if necessary. For this self-training method, two separate approaches were
developed and compared. One was majority voting (MV), and the other was weighted majority voting (WMV). MV exploited a local spatial relationship of hyperspectral images, and WMV utilized a spectral relationship along with spatial consideration. The self-training method with MV and WMV yielded significant classification improvement while WMV performed better than MV.
CHAPTER II
CORRELATION ANALYSIS BETWEEN RADAR BACKSCATTER AND IN SITU
SOIL PROPERTY MEASUREMENTS

2.1 Background

Several monitoring and modeling applications including agriculture, mining, and civil engineering as well as military tactics require extensive knowledge of soil properties over a large area. The ability to employ remote sensing techniques to extract such soil data can enable large areas to be surveyed economically. While direct measurements of many soil properties may not be possible solely by remote sensing techniques, knowledge of possible relationships between these soil properties and remote observables could allow estimates of their range to be mapped, and when combined with limited in situ measurements may reduce uncertainty in such maps. Therefore, exploring the relationship between what can be mapped by remote sensing methods and in situ soil properties will be of value.

Past studies relating remote sensing data to in situ soil properties have primarily focused on soil moisture, roughness and texture. Ulaby et al. [9, 10] investigated the relationship of microwave backscatter and surface roughness, soil moisture, and soil texture for bare soil as well as vegetation covered soil and its dependence on incidence angle and microwave frequency. Bindlish and Barros [11] incorporated a vegetation parameter into a radar backscatter model for soil moisture estimation. Santanello et al.
estimated surface soil moisture and hydraulic conductivity for troop and vehicle mobility using microwave. Thoma et al. [14] presented a protocol to decide on the appropriate scale with which to retrieve soil moisture from high resolution radar data. Anderson and Croft [18] extensively reviewed remote sensing methods including active and passive microwave and optical imaging to monitor soil moisture and surface roughness. Zribi et al. [19] showed that TerraSAR-X and clay fraction had a linear relationship. Flores et al. [15, 16] explored the possible soil moisture estimation from L-band radar for military mobility applications. Lowe et al. [61] used a radar system to identify clandestine graves in various soil property scenarios. Han et al. [17] used passive microwave simulations to estimate soil moisture, sand and clay fraction, organic density, saturated hydraulic conductivity, and surface energy flux with assimilation models.

Other studies have explored the relationship between various soil properties, based on in situ measurements alone. Pan et al. [20] reported that soil moisture was closely related to soil texture, bulk density, and air dried water content. Ayers and Bowen [21] estimated soil density using cone penetration resistance and soil moisture profile. Vaz et al. [23] tried to find the best regression model among known models with in situ cone penetration resistance, soil moisture, soil bulk density, and soil texture. Vaz et al. [22] found that normalized water content and soil bulk density were effective to reduce regression model variation from soil texture and organic matter content.

Both Liu et al. [60] and Santanello et al. [59] used SAR to estimate soil properties, but they both also used ancillary data to estimate soil moisture which then enabled the separation of soil texture effects on backscatter from moisture (and thus dielectric) effects. A motivating question for our investigation is: What can be learned
about soil properties from a single SAR image without ancillary data related directly to soil moisture? Put another way: Is there significant correlation between the radar backscatter and soil texture and other properties? Furthermore, we test for such correlation with soil properties measured at depths which include those to which the L-band radar used is unlikely to penetrate. Any such correlation would obviously rely on the related correlation between the deeper soil and that nearer the surface. Such correlation is expected to be stronger in constructed soil environments where the soils are fairly homogenous.

Studies of investigating any direct or indirect relationship between radar backscatter data and more extensive sets of soil physical properties such as penetration resistance and hydraulic conductivity have not been found. The main objective of this study was to statistically explore possible relationships between radar backscatter strength and several soil physical properties. The fact that radar backscatter is mostly influenced by roughness and the dielectric constant – which in soil is primarily driven by moisture—led us to speculate that the soil properties investigated here, which clearly affect these characteristics, would show some correlation.

2.2 Materials and Methods

To accomplish the main objective of this study, three polarization channels (i.e., HH, HV and VV) of an airborne SAR instrument (the NASA JPL’s Uninhabited Aerial Vehicle Synthetic Aperture Radar (UAVSAR) [8]), and eight in situ soil physical properties at two depth intervals acquired for the study area were considered in the statistical analyses. The soil properties which were examined include penetration resistance parameters (sleeve friction \(q_s\) and cone tip resistance \(q_c\), clay fraction, sand
fraction, saturated hydraulic conductivity ($K_{\text{sat}}$), field capacity ($f_c$), permanent wilting point ($pwp$), and porosity ($n$).

The study area used in this investigation is a portion of a constructed earthen levee system, and as such is relatively homogeneous—by design—in terms of soil properties, topography, and surface vegetation. This homogeneity allows us to consider the extent of correlation between long-wave radar backscatter and slowly changing soil properties without explicitly accounting for the amount of soil moisture present at the time of the radar acquisition. This assumption is further supported by the observation that the range of spatial and temporal variability of soil moisture over this area is relatively narrow.

In situ soil properties and airborne radar data were obtained as part of a study applying SAR to the problem of monitoring earthen levees [32]. Over the course of that project, it was observed that analysis of the radar image not only detected surface anomalies such as slump slides resulting from slope instability, but also highlighted some areas that had not slid at the time of the radar image but later did show unstable slope characteristics [33]. This led to our speculation that deeper sub-surface soil characteristics, at least in this specific and fairly homogeneous environment, might be detectable in the radar profiles.

2.2.1 Study Area

Fig. 2.1 shows the study area which was used in this work. As shown, the study area is an approximately 3 km long portion of the levee system on the east side of the Mississippi river, north of Vicksburg, Mississippi, USA. Recently, Aanstoos et al. [32] obtained polarimetric synthetic aperture radar (PolSAR) imagery and also measured
several in situ soil properties on earthen levees in this study area as part of the development of remote levee monitoring methods. Sehat et al. [33] analyzed these in situ soil properties to detect differences in soil properties in the vicinity of slump slides versus non-slide areas as determined by an automatic SAR image classification.

Figure 2.1 Study area.
*(a) Map of lower Mississippi River and vicinity. (b) River path color composite from polarimetric UAVSAR data. (c) Study area highlighted in red.

2.2.2 Field Investigation

Several in situ soil properties were collected from Sept. 15, 2010 to Sept. 29, 2010. A detailed description of how this data was collected can be found in [33] and only summary is presented here for completeness. C3 Consulting, LLC made in situ measurements from 1 cm to 120 cm of soil over the entire study area. A hydraulic push system drove a miniaturized cone penetrometer (Fig. 2.2) [33] into the ground to measure various soil properties. The miniaturized cone penetrometer integrates moisture, resistivity and compaction sensors that simultaneously collect data in one centimeter.
increments of depth for moisture content, resistivity to electricity flow and soil compactness, respectively.

![Cone penetrometer in conjunction with all-terrain vehicle-mounted.](image)

Figure 2.2  A cone penetrometer in conjunction with all-terrain vehicle-mounted.

2.2.3  In Situ Soil Property Measurements

Cone penetrometer locations are shown using red points in Fig. 2.3. A total number of 106 testing locations were uniformly selected over the study area. Further, to supplement and verify the data obtained by the cone penetrometer, soil core sampling was performed to the depth of 120 cm at nearly 15 cm away from each cone penetration location. Its purpose was to supplement texture properties of soil and validate the data obtained by the cone penetrometer. The data collected from the cone, sensors and soil corings were then integrated to build soil profile versus depth at each sampling location. The soil profile was broken into surface and subsurface layers. After removing the top 7 cm of the soil to avoid the effect of anthropogenic activities, the surface layer is specified
from the surface to the first horizon break line of the soil profile which varied from 30 cm to 50 cm from the surface. Therefore, the surface layer is described from a top 7 cm to the depth of 30 to 50 cm. The subsurface layer is specified from the bottom of the surface layer down to the depth of around 120 cm. In this study, $q_s$, $q_c$, $K_{sat}$, soil texture (sand and clay fractions), $f_c$, pwp, and n were used for analysis purposes.

$q_s$ measures the average skin friction as the probe is advancing through the soil, and $q_c$ is the measure of the soil resistance on the cone tip as the probe is going down into the soil. Their values vary based on soil texture, soil moisture content, and soil bulk density. Sand and clay fractions represent the ratio of sand and clay contents in the soil. $f_c$ is the amount of soil moisture or water remaining in the soil after water has drained for some time, and $n$ is a measure of empty space in soil or how much water soil can retain water. $f_c$ and $n$ are closely related to soil texture, soil bulk density, and organic matter.

Figure 2.3 Cone penetrometer sampling locations.
content. pwp is the minimal level of soil moisture that prevents a plant from wilting. $K_{\text{sat}}$ is the amount of flow per unit area when the soil is saturated with water and is directly proportional to seepage velocity and is a key factor for seepage analysis. It is also closely correlated to soil texture and soil bulk density.

The soil data was analyzed separately for samples taken from the river side of the levee and from the land side since these areas differ significantly in slope (and thus also radar local incidence angle) and exposure to surface water. Tables 2.1 and 2.2 show the statistics of the in situ soil property data for the surface and subsurface layers. There are 52 and 54 data samples for the river side and land sides, respectively.

Table 2.1   Statistics of measured in situ soil properties for the river side of study area

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Surface Layer</th>
<th>Subsurface Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Clay (%)</td>
<td>10.4</td>
<td>48.1</td>
</tr>
<tr>
<td>Sand (%)</td>
<td>11.2</td>
<td>81.4</td>
</tr>
<tr>
<td>$q_s$ (kPa)</td>
<td>43.8</td>
<td>163.2</td>
</tr>
<tr>
<td>$q_c$ (kPa)</td>
<td>1902</td>
<td>5610</td>
</tr>
<tr>
<td>$K_{\text{sat}}$ (cm/s)</td>
<td>$7.8 \times 10^{-5}$</td>
<td>$9.2 \times 10^{-5}$</td>
</tr>
<tr>
<td>$f_c$ (%v)</td>
<td>17.0</td>
<td>43.0</td>
</tr>
<tr>
<td>pwp (%v)</td>
<td>8.5</td>
<td>27.5</td>
</tr>
<tr>
<td>n (%)</td>
<td>15.1</td>
<td>63.3</td>
</tr>
</tbody>
</table>

*Number of measurements: N = 52.
Table 2.2  Statistics of measured in situ soil properties for the land side of study area

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Surface Layer</th>
<th>Subsurface Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Clay (%)</td>
<td>18.4</td>
<td>59.2</td>
</tr>
<tr>
<td>Sand (%)</td>
<td>8.9</td>
<td>43.2</td>
</tr>
<tr>
<td>$q_s$ (kPa)</td>
<td>35.5</td>
<td>197.4</td>
</tr>
<tr>
<td>$q_c$ (kPa)</td>
<td>1262</td>
<td>5590</td>
</tr>
<tr>
<td>$K_{sat}$ (cm/s)</td>
<td>$7.9 \times 10^{-5}$</td>
<td>$2.8 \times 10^{-4}$</td>
</tr>
<tr>
<td>$f_c$ (%v)</td>
<td>25.4</td>
<td>47.1</td>
</tr>
<tr>
<td>pwp (%v)</td>
<td>12.0</td>
<td>32.1</td>
</tr>
<tr>
<td>n (%)</td>
<td>39.3</td>
<td>67.1</td>
</tr>
</tbody>
</table>

*Number of measurements: N = 54.

Fig. 2.4 through 2.7 show histograms of the in situ soil physical properties along with means and standard deviations. The number of histogram bins is set at 11. Fig. 2.4 shows soil texture in terms of sand and clay fraction in the surface and subsurface layers for the river side (Fig. 2.4a) and land side (Fig. 2.4b), respectively. The land side has higher clay fraction (maximum of 59.2% and 54.4% for the surface and subsurface layers, respectively) than the river side (maximum of 48.1% and 47.0% for surface and subsurface layer respectively). For sand fraction, it has a similar data range excluding an extreme sample or outlier (the bin or bar with star on the top) on the surface layer of the river side. However, the subsurface layer of the river side has a wider sand fraction range. Their distributions roughly fit a normal distribution.
Fig. 2.4  Histograms of clay and sand fraction for surface and subsurface layers.
*(a) river side and (b) land side

Fig. 2.5 shows $q_s$ and $q_c$ on the surface and subsurface layer for the river side (Fig. 2.5a) and land side (Fig. 2.5b). $q_s$ on both the river and land sides shows a roughly normal distribution, and it has a wider range on the land side. On the river side it goes up to about 160 kPa and 170 kPa for the surface and subsurface layers, respectively. On the land side, the averaged $q_s$ of 200 kPa and 285 kPa were obtained for the surface and subsurface layers, respectively. $q_s$ of about 285 kPa seen on the subsurface layer of the land side can be considered an extreme data point or outlier. Even without it, however, the land side has higher $q_s$ values as seen in Fig. 2.4. $q_c$ also shows roughly a normal distribution with skewness because of outliers. The data ranges of $q_c$ are similar if the outliers are excluded, but it has greater variation (standard variation) on the subsurface layer than on the land side.
Fig. 2.5  Histograms of $q_s$ and $q_c$ for surface and subsurface layers.

*(a) river side and (b) land side.

Fig. 2.6 presents histograms for $K_{sat}$ and $f_c$. $K_{sat}$ in the surface layer has a similar data range, without outliers, for the river and land sides; while the river side has a greater data range on the subsurface layer excluding outliers. For $f_c$, the land side has a greater data range for both layers. $K_{sat}$ and $f_c$ can be approximately modeled as a normal distribution.

Data for $pwp$ and $n$ are displayed in Fig. 2.7. $pwp$ has a similar mean for the corresponding layer from both sides with a greater data range on the land side. $n$ also has a similar data distribution excluding outliers on the river side data. Both $pwp$ and $n$ can be approximately modeled as normal distribution.
Figure 2.6  Histograms of $K_{\text{sat}}$ and $f_c$ for surface and subsurface layers. *(a) river side and (b) land side.

Figure 2.7  Histograms of pwp and n for surface and subsurface layers. *(a) river side and (b) land side.
2.3 PolSAR Imagery

The UAVSAR imagery for the study area was taken on January 25, 2010. UAVSAR acquires repeat-track L-Band SAR data for differential interferometric measurements [62-64]. The bandwidth of the radar is 80 MHz, and it has 1.8m slant range resolution and full polarimetry. Also, it has exceptionally low noise [8]. Thus, it has capability to differentiate targets with weak radar backscattering cross section. L-
band SAR can penetrate several meters in an extremely dry soil, but typically the extent of its wavelength (15–30 cm) in more typical dry soil. In this study, the HH, VV, and HV polarization channels were used.

The strength of radar backscatter depends on many factors including local incidence angle [46]. Since the slope of the land and river sides of a levee are in opposite directions, two separate flight segments were scanned and respectively analyzed for the river and the land side of levee. Fig. 2.8 shows the histograms of power (dB) of the HH, HV and VV channels along with mean and standard deviation for the river side (Fig. 2.8a) and land side (Fig. 2.8b). They overall demonstrate a pattern of normal distribution, and the VV channel has a similar mean for both river and land side. In the case of HH, the land side has slightly stronger backscatter while the HV backscatter is stronger on the river side. Backscatter on the river side appears to be right skewed while left skewness is seen on the land side.

Figure 2.10  Histogram of SAR magnitude (dB).
*(a) river side and (b) land side.
2.4 Statistical Analysis

Correlation of different radar backscatter channels with in situ soil data measurements were examined using SAS package. Each radar backscatter channel was converted to power (dB), and in situ soil data measurements were transformed by log transformation to make data more normally distributed. Pearson correlation coefficients and p-value were calculated between each channel and in situ soil data measurements. In Table 2.3 and Table 2.4, each soil data has two values. The first value is Pearson correlation coefficient, and the second one in parentheses is p-value. In this study, correlation coefficients with 0.4 or more and -0.4 or less are considered showing some statistical relationships with 0.05 of p-value. Correlations can be visually examined from the scatter plots from Appendix A.

Table 2.3 Correlation between radar backscatter and soil properties for the surface layer.

<table>
<thead>
<tr>
<th>Side</th>
<th>Band</th>
<th>(q_s)</th>
<th>(q_c)</th>
<th>Clay</th>
<th>Surface Layer</th>
<th>Sand</th>
<th>(K_{sat})</th>
<th>(f_c)</th>
<th>PWP</th>
<th>(n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>River (N=52)</td>
<td>HH</td>
<td>-0.06</td>
<td>-0.34</td>
<td>0.37</td>
<td>-0.34</td>
<td>-0.37</td>
<td>0.37</td>
<td>0.38</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.696)</td>
<td>(0.013)</td>
<td>(0.006)</td>
<td>(0.012)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.006)</td>
<td>(0.036)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HV</td>
<td>0.02</td>
<td></td>
<td></td>
<td>-0.53</td>
<td>0.55</td>
<td>-0.54</td>
<td>-0.53</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.899)</td>
<td></td>
<td></td>
<td>(&lt;.0001)</td>
<td>(&lt;.0001)</td>
<td>(&lt;.0001)</td>
<td>(&lt;.0001)</td>
<td>(&lt;.0001)</td>
<td>(&lt;.0001)</td>
</tr>
<tr>
<td></td>
<td>VV</td>
<td>-0.09</td>
<td>-0.32</td>
<td>0.26</td>
<td>-0.33</td>
<td>-0.20</td>
<td>0.28</td>
<td>0.25</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.534)</td>
<td>(0.020)</td>
<td>(0.065)</td>
<td>(0.018)</td>
<td>(0.155)</td>
<td>(0.047)</td>
<td>(0.070)</td>
<td>(0.174)</td>
<td></td>
</tr>
<tr>
<td>Land (N=54)</td>
<td>HH</td>
<td>0.16</td>
<td>0.15</td>
<td>-0.20</td>
<td>0.15</td>
<td>0.06</td>
<td>-0.18</td>
<td>-0.19</td>
<td>-0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.245)</td>
<td>(0.264)</td>
<td>(0.155)</td>
<td>(0.275)</td>
<td>(0.690)</td>
<td>(0.195)</td>
<td>(0.168)</td>
<td>(0.736)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HV</td>
<td>0.03</td>
<td>0.20</td>
<td>-0.16</td>
<td>0.16</td>
<td>0.04</td>
<td>-0.15</td>
<td>-0.15</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.846)</td>
<td>(0.142)</td>
<td>(0.261)</td>
<td>(0.251)</td>
<td>(0.795)</td>
<td>(0.285)</td>
<td>(0.269)</td>
<td>(0.828)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VV</td>
<td>-0.13</td>
<td>-0.26</td>
<td>0.16</td>
<td>-0.25</td>
<td>-0.27</td>
<td>0.19</td>
<td>0.16</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.362)</td>
<td>(0.053)</td>
<td>(0.234)</td>
<td>(0.072)</td>
<td>(0.046)</td>
<td>(0.176)</td>
<td>(0.253)</td>
<td>(0.143)</td>
<td></td>
</tr>
</tbody>
</table>

*The first value is Pearson correlation coefficient, and the second one in parentheses is p-value.
Table 2.4  Correlation between radar backscatter and soil properties for the subsurface layer.

<table>
<thead>
<tr>
<th>Side</th>
<th>Band</th>
<th>q_t</th>
<th>q_c</th>
<th>Clay Subsurf</th>
<th>Sand</th>
<th>K_sat</th>
<th>f_c</th>
<th>PWP</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>River (N=52)</td>
<td>HH</td>
<td>0.11</td>
<td>-0.34</td>
<td>0.37</td>
<td>-0.36</td>
<td>-0.31</td>
<td>0.38</td>
<td>0.38</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.421)</td>
<td>(0.013)</td>
<td>(0.006)</td>
<td>(0.009)</td>
<td>(0.027)</td>
<td>(0.006)</td>
<td>(0.005)</td>
<td>(0.045)</td>
</tr>
<tr>
<td></td>
<td>HV</td>
<td>0.23</td>
<td>-0.46</td>
<td>0.54</td>
<td>-0.50</td>
<td>-0.48</td>
<td>0.54</td>
<td>0.55</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.103)</td>
<td>(0.0007)</td>
<td>(&lt;.0001)</td>
<td>(0.0002)</td>
<td>(0.0003)</td>
<td>(&lt;.0001)</td>
<td>(&lt;.0001)</td>
<td>(0.0018)</td>
</tr>
<tr>
<td></td>
<td>VV</td>
<td>0.30</td>
<td>-0.20</td>
<td>0.26</td>
<td>-0.25</td>
<td>0.10</td>
<td>0.19</td>
<td>0.27</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.029)</td>
<td>(0.159)</td>
<td>(0.058)</td>
<td>(0.072)</td>
<td>(0.059)</td>
<td>(0.055)</td>
<td>(0.052)</td>
<td>(0.258)</td>
</tr>
<tr>
<td>Land (N=54)</td>
<td>HH</td>
<td>0.06</td>
<td>-0.018</td>
<td>-0.09</td>
<td>0.10</td>
<td>0.06</td>
<td>-0.08</td>
<td>-0.10</td>
<td>-0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.659)</td>
<td>(0.8973)</td>
<td>(0.531)</td>
<td>(0.47)</td>
<td>(0.657)</td>
<td>(0.574)</td>
<td>(0.443)</td>
<td>(0.620)</td>
</tr>
<tr>
<td></td>
<td>HV</td>
<td>-0.04</td>
<td>0.01</td>
<td>-0.11</td>
<td>0.12</td>
<td>0.13</td>
<td>-0.08</td>
<td>-0.13</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.750)</td>
<td>(0.938)</td>
<td>(0.436)</td>
<td>(0.392)</td>
<td>(0.346)</td>
<td>(0.541)</td>
<td>(0.36)</td>
<td>(0.922)</td>
</tr>
<tr>
<td></td>
<td>VV</td>
<td>-0.09</td>
<td>-0.28</td>
<td>0.15</td>
<td>-0.15</td>
<td>0.04</td>
<td>0.18</td>
<td>0.16</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.527)</td>
<td>(0.040)</td>
<td>(0.273)</td>
<td>(0.289)</td>
<td>(0.770)</td>
<td>(0.200)</td>
<td>(0.233)</td>
<td>(0.191)</td>
</tr>
</tbody>
</table>

*The first value is Pearson correlation coefficient, and the second one in parentheses is p-value.

2.5  Results and Discussion

From Figures 2.4~2.8, the soil properties and SAR magnitudes approximately show normal distribution. Table 2.3 (for the surface layer) and Table 2.4 (for the subsurface layer) provide summaries of correlation between SAR image (radar backscatter) and soil properties with log transformation. The results shown in Tables 2.3 and 2.4 indicate that both the surface and subsurface layers exhibit similar correlations; only the river side shows some correlation between HV and some soil properties while weaker correlation can be found with HH. This might come from the fact that the river side could have stronger backscatter signal on HV channel as seen in Fig. 2.8 because of incident angle and slope differences between the river side and the land side.

Clay fraction shows a positive correlation with HV having correlation coefficient, 0.55 on the surface layer and 0.54 on the subsurface layer while sand fraction is negatively correlated with HV with correlation coefficient -0.54 on the surface and -0.50 on the subsurface layer. This can be attributed to the fact that clay can contain more
moisture while sand quickly drains water, so more soil moisture in clay reflects more radar signal on the surface [9, 19]. $q_c$ and $K_{sat}$ show a negative correlation with HV while $f_c$ and pwp have a positive correlation with HV in both layers. In the surface layer, correlation coefficients of $q_c$ and $K_{sat}$ are -0.53 while 0.55 for field capacity and pwp. In the sub surface layer, correlation coefficients are -0.46 and -0.48 for $q_c$ and $K_{sat}$, respectively, while 0.55 for $f_c$ and pwp. $n$ shows a weak positive correlation with HV as 0.40 and 0.42 for the surface and subsurface layers, respectively.

Tables 2.3 and 2.4 show that HV has a negative correlation with $q_c$ and a positive correlation with clay fraction. This observation is consistent with the fact that $q_c$ decreases as clay fraction increases due to the fact that a higher percentage of clay (versus sand) will decrease the penetration resistance of soil. Baziar et al. [65] reported a similar positive correlation between HV and silt fraction.

$f_c$ and pwp are basically different metrics of the moisture amount in soil, and they have similar patterns with soil moisture. As shown in Tables 2.3 and 2.4, $f_c$ and pwp have a positive correlation with HV as radar backscatter has a positive correlation with soil moisture on the surface. $n$ is greater in loosely aggregated soil than tightly packed soil, and, in general, it is greater in clayey soil than in sandy soil. Therefore, the correlation of $n$ with radar backscatter follows the pattern of clay fraction. As shown, both $n$ and clay fraction exhibit a positive correlation with HV.

$K_{sat}$ is mostly related to soil particles size, particle and components of soil mass, and density and viscosity of the fluid. Usually, $K_{sat}$ is greater in a sandy soil than a clay soil. Therefore, it is closely related to soil texture, and has the similar negative correlation with radar backscatter as sand fraction.
The results shown in Tables 2.3 and 2.4 confirm the strong correlation between in situ soil physical properties and soil moisture and radar backscatter, which were previously addressed in the introduction section. In other words, stronger radar backscatter is usually expected from higher moisture content in the soil. Clay fraction, $f_c$, pwp, and n have positive correlations with HV channel. This indirectly implies that they are positively correlated with soil moisture. On the contrary, sand fraction, $q_c$, and $K_{sat}$ have negative correlation with HV channel. This also indirectly shows a tendency that they are inversely related to soil moisture. $K_{sat}$, $f_c$, and n have close relationships with soil texture. The surface and subsurface layers exhibited a similar correlation pattern. This may imply that the surface and subsurface layers have similar physical soil properties. In other words, the surface and subsurface layers are highly correlated for each physical soil property in the study site.

The soil moisture measurements were taken from Sept. 2010 in Fig. 2.8, and Feb., Oct., and Nov. of 2011, as shown in Fig. 2.9. There had been no rainfall for at least 15 days when soil moisture was measured during Sept. 2010. There had been rain of 0.8in 6 days before when soil moisture was obtained on Feb. 2011. There had been no rain for 10 days for soil moisture measurement from Oct. 2011. Finally, there had been 0.5in rainfall 5 days ago when soil moisture measurement was done on Nov. 2011. They show a relatively narrow range of values over time and space in the study area. For this reason, we believe that the spatial variation of radar backscatter can reveal patterns of variation of the related soil properties without making adjustments for exact moisture content at the time of the radar flight.
As noted, most of our in situ soil property measurements excluded the top 7 cm of the surface. This was due to the data collection contractor’s belief that they would not be useful due to contamination by anthropogenic activity. Although we would have preferred to have measurements from this layer regardless, especially since most of the radar backscatter would be expected from that layer, the absence of this data was beyond our control. The presence of correlations between the radar data and deeper layers of soil is likely explained by the vertically homogeneous nature of this constructed earthen environment. Such a connection may not be present in more general soil environments, and our results should be tested more broadly before any such conclusions can be made.

From the results, strong conclusions about correlation between SAR magnitude and soil properties may not be reasonable. However, more controlled data acquisition from lab experiments will help make stronger statements about the results. The contribution of this study is mainly to explore relationships between SAR and in situ soil physical properties. If such relationships are verified with more data, it will be possible for SAR images to be used more confidently and economically with fast data acquisition for a large spatial coverage. Moreover, as the use of unmanned aerial vehicles (UAVs) is increasing with better performance and efficiency, real time monitoring will soon be economically feasible by a drone with radar.
CHAPTER III
CLASSIFICATION OF LEVEE SLIDES FROM AIRBORNE SYNTHETIC
APERTURE RADAR IMAGES WITH EFFICIENT SPATIAL FEATURE
EXTRACTION

3.1 Background

There are about 200,000 km of earth levees in the United States, and even more
with various designs throughout the world. According to a survey on Governing.com,
only 10% of 744 levees from National Levee Database (NLD) were rated as “acceptable”
while the rest were marginally acceptable or unacceptable [65]. Formidable failure of
levees in New Orleans during Hurricane Katrina along the Mississippi River resulted in
great loss economically with human casualties in that area. This vivid catastrophe
highlighted the importance of levee monitoring. Extensive research has been done to
monitor levee status with various different approaches. The main approaches used are
field-based in situ soil property measurements and remote sensing measurements (e.g.,
SAR images and optical images).

Recently, research on screening levees has been conducted at Mississippi State
University [26, 29, 32, 33]. Airborne and spaceborne SAR images are used to monitor the
abnormality of study areas along Mississippi River. Another similar application is
landslide monitoring, where SAR images may be combined with optical images and
digital terrain models [40, 67]. In this research, an airborne SAR having three magnitude
bands with polarization HH, VV, and HV were used for levee monitoring, and supervised classification of slide and non-slide levee areas was performed using the standard support vector machine (SVM) classifier [1, 2, 34, 36]. Although SVM is a powerful classifier, it failed to accurately classify the slide area without any spatial feature extraction [68, 69]. This may be because the original low-dimensional data does not include enough discriminant features. Thus, unlike a dimensionality reduction process for high-dimensional data [38, 39, 70], a dimensionality expansion process is included by adding additional artificial bands [40, 41], which are simply nonlinearly normalized bands in this research. In addition, a feature extraction method for spatial information has been considered before classification. Gray-level co-occurrence matrix (GLCM) is one of the popular methods to describe spatial features, and it performs well on many applications [43-45]. It has been applied to levee slide classification [29, 46]. However, it requires significant computational power, since a relatively large local area has to be considered in order to find accurate texture features. Furthermore, several position operators have to be applied, and statistical features must be calculated from the GLCM before classification. Thus, it may not be suitable to fast analysis of a big image.

In this study, very simple and effective feature extraction methods for SAR images were conducted [42, 71-73]. Specifically, the weighted average filter and the majority filter [47] may offer slightly lower classification accuracy but with much less computational cost. Such spatial lowpass filtering techniques with a sliding window are suitable for parallel computing, since the output of a single pixel is unrelated to other areas of the image [48]. Thus, such spatial feature extraction methods are preferred for fast processing of large-scale remote sensing images. When the GLCM is used, adding
the features from such spatial filters may further improve the performance. For complicated slide features, the use of additional normalized bands improves the performance of GLCM and spatial filtering. The contributions of this study are three-fold:

1. Spatial lowpass filters were successfully used for spatial feature extraction in levee slide classification from SAR images. The center pixel was replaced by the filter output of a small neighborhood, which included spatial information; meanwhile, the spatial lowpass filters could reduce noise for classification performance enhancement. It may also be easily implemented in parallel for fast data processing and analysis.

2. For GLCM, adding the lowpass filtering outputs further improved classification accuracy.

3. Additional normalized bands could generate additional spatial and texture features to improve the classification accuracy.

The remainder of this chapter is organized as follows: Section 3.2 introduces the data used in the experiments, Section 3.3 presents the methods used in this study, and Section 3.4 shows experimental results. The conclusions are drawn in Section 3.5.

### 3.2 Data

As shown in Fig. 3.1, the study area in this work is an approximately 3 km long portion of the levee system on the east side of the Mississippi River, north of Vicksburg, Mississippi, USA. About a 40m wide mask (buffer) on the river side from the levee road was applied to segment the area for classification, as illustrated in Fig. 3.2.
The imagery for this study was taken by the NASA JPL’s Uninhabited Aerial Vehicle Synthetic Aperture Radar (UAVSAR) on June 16, 2009. UAVSAR acquires repeat-track SAR data to get differential interferometric measurements [8]. Reconfigurable polarimetric L-Band SAR sensors are mounted. The bandwidth of the radar is 80 MHz, and it has 1.8 m slant resolution with full polarimetry, in addition to having exceptionally low noise. Thus it has the capability to differentiate targets with weak radar backscattering cross section. L-band SAR can penetrate a few meters in very dry soil, but its penetration is typically a few centimeters. Therefore, the imagery from the UAVSAR is an excellent source for monitoring levee change.

Figure 3.1  Study area.

*(a) Map of lower Mississippi River. (b) Polarimetric UAVSAR image in false color (combination of HH, HV and VV channels). (c) Study area on levee highlighted in red.
Fig. 3.2 Slide location.

*(a) slide location on an optical image. (b) Two AOIs. (c) Ground truth with slide areas in light blue.

Fig. 3.2 shows levee slide locations in two areas of interest (AOI) used in the experiments: AOI1 is a $66 \times 48 \times 3$ image and AOI2 $80 \times 83 \times 3$. AOI1 has more complicated texture features than AOI2. All the pixels are labeled. Different percentages of labeled samples are used as training samples and the rest for testing in the experiments.

3.3 Method

The strength of radar backscatter is heavily affected by surface roughness of the terrain and a slide is identified as rough patch [73]. Features responsive to surface roughness include the magnitudes of the HH, VV, and HV polarimetric backscattering coefficients. Sometimes classification can be done successfully on the original images without spatial feature extraction. But in the case of using SAR images for levee slide classification, it is almost impossible to achieve accurate classification without proper spatial feature extraction.
There are many approaches to extract spatial features. GLCM is commonly used to extract texture features. GLCM is a versatile method with many choices about operators and features. On the contrary, it has many parameters to tune for a specific application, such as window size, level of grey scale, direction and distance from a center pixel, etc. A common approach for GLCM is to choose as many features as possible, followed by a feature dimensionality reduction method. In this study, since processing efficiency is a concern, very simple feature extraction techniques, i.e., the weighted average filter and majority filter, are adopted to capture spatial property as an alternative to GLCM. It is well-known that GLCM is prone to be affected by noise. For noisy SAR images, spatial lowpass filters can reduce noise. The spatial lowpass filters require much lower computational costs, and they can be easily implemented in parallel.

3.3.1 GLCM and Band Normalization

A GLCM relates a pixel to other pixels with specific distance and direction defined by a position operator. The position operator is application-specific. Sometimes, several different operators have to be used together. After a GLCM is generated, several quantitative features, such as homogeneity, uniformity, contrast, entropy, have to be computed before the related texture information can be actually used for classification. Therefore, it may be complicated for the GLCM technique to be applied for large scale images in remote sensing, and it is difficult to be implemented in parallel. There are dozens of quantitative textural features that can be derived from GLCM, but only four major features, as shown in Table 3.1, are used in this study since most of the others are either insensitive to levee slides or are highly correlated to these four.
The magnitude of the SAR image varies within a certain range based on levee condition. This could be problematic with setting a proper level of grey scale for GLCM generation. Therefore, a fractional measure for relative backscatter strength among HH, VV, and HV bands is derived as:

\[ HH_f = \frac{|HH|}{\sqrt{|HH|^2 + |VV|^2 + |HV|^2}} \]  

(3.1)

\[ VV_f = \frac{|VV|}{\sqrt{|HH|^2 + |VV|^2 + |HV|^2}} \]  

(3.2)

\[ HV_f = \frac{|HV|}{\sqrt{|HH|^2 + |VV|^2 + |HV|^2}} \]  

(3.3)

Then, the values in the three bands after normalization are fractional within [0 1]. To be applicable to GLCM generation, all the values are converted to integers (after multiplying by 100). Note that such a normalization process does not actually change the polarization feature in a 3×1 pixel vector, but it does change the neighboring spatial feature of the pixel in each band because the normalization term in equation (3.1~3.3) is varied per pixel. In the experiments, it will be demonstrated that using normalization bands can improve the performance of GLCM-based feature extraction.
Table 3.1 The major features of GLCM used in this research.

<table>
<thead>
<tr>
<th>Features</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homogeneity</td>
<td>[ \sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1 + (i - j)^2} ]</td>
</tr>
<tr>
<td>Uniformity</td>
<td>[ \sum_{i,j=0}^{N-1} P_{i,j}^2 ]</td>
</tr>
<tr>
<td>Contrast</td>
<td>[ \sum_{i,j=0}^{N-1} P_{i,j}(i - j)^2 ]</td>
</tr>
<tr>
<td>Entropy</td>
<td>[ \sum_{i,j=0}^{N-1} P_{i,j} (-\ln P_{i,j}) ]</td>
</tr>
</tbody>
</table>

*P represents a GLCM, i and j are the coordinates of P, and N is the number of image pixels.

3.3.2 Spatial Filtering

Spatial filtering is widely used in digital image processing. A small square window is often employed to slide over an entire image; the filter output at each location is assigned to the center pixel of the window. For multi-dimensional image processing, the filter output includes both spatial and spectral information. A lowpass spatial filter, such as a local averaging filter, can reduce noise but smooth out image details such as edges.

A weighted average filter, which actually is the Gaussian lowpass filter, is often employed, whose weight in a local window is defined by

\[ w = e^{-\frac{d}{\sigma}} \]  \hspace{1cm} (3.4)

where \( d \) is the spatial distance between the center pixel and a neighboring pixel in the window and \( \sigma \) is a user-defined parameter. In this research, a simple average filter with equal weights is adopted to save computational cost in weight multiplication.
Another option is to apply a majority filter. The image needs to be quantized to integer levels similar to GLCM. Proper window size and quantization level should be decided for a specific application. In the case of each 3×3 window, for example, the majority filter assigns the predominant value to the center pixel. If no predominant value is found or when all the 9 input pixels have different values, then the median value in the window is used for the center pixel. Compared to the weighted average filter, the majority filter can better maintain image details.

### Table 3.2 Parameter settings for both GLCM and spatial filtering based feature extraction.

<table>
<thead>
<tr>
<th></th>
<th>Weighted average filter</th>
<th>Majority filter</th>
<th>GLCM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>σ Window size</td>
<td>Level Window size</td>
<td>Level Window size</td>
</tr>
<tr>
<td>AOI1</td>
<td>5 5</td>
<td>9 7</td>
<td>9 7</td>
</tr>
<tr>
<td>AOI2</td>
<td>5 3</td>
<td>9 7</td>
<td>9 7</td>
</tr>
</tbody>
</table>

### 3.4 Experiment

A few combinations of feature sets are investigated for comparison purposes which are described below (with the number of features).

1. **OR:** original 3 bands (3)
2. **ON:** original 3 bands + normalized bands (3 + 3 = 6)
3. **OA:** original 3 bands + weighted average filter (3 + 3 = 6)
4. **OM:** original 3 bands + majority filter (3 + 3 = 6)
5. **OAM:** original 3 bands + weighted average filter + majority filter (3 + 3 + 3 = 9)
6. ONM: original 3 bands + normalized bands + majority filter (3 + 3 + 6 = 12)

7. ONAM: original 3 bands + normalized bands + weighted average filter + majority filter (3 + 3 + 6 + 6 = 18)

8. OG: original 3 bands + GLCM (3 + 3 × 4 = 15)

9. ONG: original 3 bands + normalized bands + GLCM (3 + 3 + 6 × 4 = 30)

10. OGM: original 3 bands + GLCM + majority filter (3 + 3 × 4 + 3 = 18)

11. ONGM: original 3 bands + normalized bands + GLCM + majority filter (3 + 3 + 6 × 4 + 3 = 33)

These combinations (including the original bands, nonlinearly generated bands, and extracted spatial features) present discriminant quantities in different domains (polarizations, their correlations, and spatial information) for better classification performance [53]. For each combination, the SVM with a radial basis function kernel and cross-validation-tuned parameters was applied; 10–50% randomly selected pixels were used as training samples and the remaining for testing, and a total of 20 runs were made and average performance was reported. Table 3.2 summarizes the parameter settings for spatial feature extraction.
Figure 3.3  SVM classification results of two subimage scenes.
Figure 3.4 Classification map of two subimage scenes.

*slide area in green, and non-slide in yellow.
Figure 3.4 (continued)

*slide area in green, and non-slide in yellow.

Fig. 3.3 shows the classification results for AOI1 and AOI2 with different percentages of labeled samples. Table 3.3 further summarizes the performance when using 30% of labeled samples. The classification performance was evaluated in terms of overall accuracy. As shown in Table 3.3, SVM could not effectively classify slide pixels directly from the original three bands (denoted as OR). Applying an average filter or a majority filter to the original bands (i.e., OA and OM) respectively slightly improves the performance while applying both of these two lowpass filters (i.e., OAM) improves the
performance further. After adding the normalized bands (i.e., ONM, ONAM), classification accuracy is increased (compared to the counterparts OM and OAM, respectively). Once GLCM is deployed to generate texture features, the results are enhanced for AOI1 which has more complicated texture features than AOI2; it seems that applying GLCM to the original bands (i.e., OG) works well for simple texture in AOI2. However, when the majority filter and weighted average filter are applied to the original bands and normalized bands (i.e., ONAM), the results are comparable to the results from GLCM (i.e., OG). For AOI1, ONAM has overall accuracy of 0.9312 while OG yields 0.9305; the overall accuracy of ONAM is 0.9820, and 0.9832 for OG in the case of AOI2. Obviously, the difference is marginal. Fig. 3.4 shows classification maps of OAM, ONAM, and ONGM of the two study areas. Compared to the ground truth maps, those from ONAM look quite similar to those from ONGM; although the classification accuracy values of ONAM in Table 3.3 are slightly lower than those of ONGM.

Table 3.3  Classification Performance when Using 30% for training (and 70% for testing).

<table>
<thead>
<tr>
<th></th>
<th>AOI 1</th>
<th>AOI 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR</td>
<td>0.8413</td>
<td>0.9746</td>
</tr>
<tr>
<td>ON</td>
<td>0.8413</td>
<td>0.9746</td>
</tr>
<tr>
<td>OA</td>
<td>0.8423</td>
<td>0.9749</td>
</tr>
<tr>
<td>OM</td>
<td>0.8681</td>
<td>0.9754</td>
</tr>
<tr>
<td>OAM</td>
<td>0.9215</td>
<td>0.9749</td>
</tr>
<tr>
<td>ONM</td>
<td>0.9175</td>
<td>0.9806</td>
</tr>
<tr>
<td>ONAM</td>
<td>0.9312</td>
<td>0.9820</td>
</tr>
<tr>
<td>OG</td>
<td>0.9305</td>
<td>0.9832</td>
</tr>
<tr>
<td>ONG</td>
<td>0.9491</td>
<td>0.9798</td>
</tr>
<tr>
<td>OGM</td>
<td>0.9235</td>
<td>0.9814</td>
</tr>
<tr>
<td>ONGM</td>
<td>0.9501</td>
<td>0.9849</td>
</tr>
</tbody>
</table>
When normalized bands are added, classification accuracy can be improved in general. For instance, ONM is better than OM, and ONAM is better than OAM. The improvement from using normalized bands together with GLCM is also obvious (ONG is better than or similar to OG, and ONGM is better than or similar to OGM). Since the normalization process changes the neighboring spatial feature of a pixel in each band, a spatial filter or a GLCM on these bands provides additional spatial or texture features thereby yielding performance enhancement.

Table 3.4 shows execution time (and feature dimensionality) for the representative combinations in Table 3.3. The experiment was done on a computer with Intel Xeon CPU (3.20GHz) and 6G memory. While the results from ONAM and OG are comparable, the execution time has very big gap for these two small images. ONAM method only takes 3.81 seconds to classify the AOI1 image while 18.86 seconds for OG. If GLCM is applied to the one with a larger feature dimensionality as in ONGM, computing time increases exponentially to 58.16 seconds. To classify the AOI2 image, it costs ONAM 7.08 seconds, OG 37.75 seconds, and ONGM 121.39 seconds. The same tendency could be found for those using GLCM features: if GLCM has to be applied to a
large scale image, it will take a much longer time; however, the weighted average filter and the majority filter can extract spatial features with much less computing time. Therefore, the ONAM method can be a very promising and efficient approach to handle fairly large images, such as spaceborne images, with less computational complexity.

3.5 Discussion

Efficient spatial feature extraction approaches are investigated for levee slide classification using SAR images. GLCM feature extraction performs well as expected but has limitations of high computation cost and storage needs, and the entire process is difficult to implement in parallel. For small datasets, this may not be crucial, but as data size increases, efficient feature extraction is needed. Spaceborne and airborne image applications usually involve with large images. In this work, an average filter and a majority filter were shown to have much lower computational cost and can be easily implemented in parallel to further reduce computing time. Even though their classification performance does not exceed GLCM (in some feature combinations), they can be used for fast screening before more complicated methods are applied on a small selected area. The spatial features and feature combinations investigated in this research may be useful for early abnormality screening with a technical of anomaly detection [75-77].

In high-dimensional data analysis (e.g., hyperspectral image classification), generating spatial features will dramatically increase the feature dimensionality and aggravate the problem of “curse-of-dimensionality”. For a polarized SAR image, this is not a problem; actually, lacking of sufficient discriminant features in a low-dimensional SAR image is the difficulty. The nonlinear band generation process intends to dig out the
information embedded in the original data that can be used to maximize class separability which in this research, is about band normalization although many nonlinear combinations can be tested [41]. Such a normalization step is to generate a new three-dimensional unit vector, so the polarization vector shape (instead of magnitude) can be emphasized; meanwhile, each pixel is divided by a different value, creating three different polarization bands with new spatial information for discrimination.
CHAPTER IV
SEMISUPERVISED CLASSIFICATION OF HYPERSPECTRAL REMOTE SENSING IMAGES

4.1 Background

Unlike SAR imagery with only several bands, hyperspectral images can have more than 200 bands, and their rich spectral information has led to many successful applications including astronomy, agriculture, mineralogy, environmental monitoring and surveillance [78-81]. The preprocessing step of band selection or reduction is required to avoid the problem of curse of dimensionality known as the Hughes phenomenon [49-52], and spectral unmixing may be helpful to more accurately separate materials at the sub-pixel level [53, 54, 82]. In addition, feature extraction may be required for better image interpretation [55, 83].

Classification can be done with either a supervised (with labeled samples) method or an unsupervised (without labeled samples) method. A semisupervised method is relatively new, which utilizes unlabeled samples when labeled samples are limited [5, 53, 56, 58]. Usually, unlabeled data are easily available, and those data can be processed in many ways to extract useful information. However, generating labeled data is not an easy task. It requires expertise on specific data, and it can be time consuming. Therefore, it is intuitive to utilize abundant unlabeled data along with limited labeled data in a learning
process. The goal of semisupervised learning is to devise a better learning machine by using labeled and unlabeled data at the same time.

4.2 Semisupervised classification

In semisupervised classification, there are labeled data, unlabeled data and a classifier. For labeled data set denoted as \((X_l, Y_l) = \{(x_{1:l}, y_{1:l})\}\), each data sample, \(x_j\) has a label, \(y_j\), while unlabeled data set \(X_u = \{x_{l+1:n}\}\) does not have labels. The total \(n\) data samples include \(l\) labeled and \(n-l\) unlabeled samples. In reality, the number of labeled data is very small \((l \ll n)\). A classifier, \(f\), can be built using labeled data (training data) as follows:

\[
f: X_l \rightarrow Y_l
\]  

(4.1)

In traditional classification, a classifier \(f\) is built by only labeled data set, \((X_l, Y_l)\).

In the case of semisupervised classification, a better classifier may be designed by using some of unlabeled data set, \(X_u\), along with labeled data. There exist different semisupervised approaches in the literature. Extensive literature reviews can be found in [5, 56, 58]. Fig. 4.1 illustrates how a better classifier can be built using unlabeled data. The three big blue and red points are considered labeled data while smaller points are unlabeled data (labels of these unlabeled data are assumed to be unknown even though a color label was used to differentiate classes). Classifier (a) performs poorly when only labeled data are used. On the contrary, classifier (b) can yield a better result if abundant unlabeled data can be incorporated with labeled data during classifier training. In this simple case, for example, a distance measure may be used between labeled and unlabeled data.
4.1 Effect of using unlabeled data.

*Classifier (a) is from labeled data (big points), and classifier (b) is from utilizing unlabeled data (small points) along with labeled data (big points) by a semisupervised method.

There is no standard categorization for semisupervised learning, but there may be generally four groups of semisupervised learning methods: self-training, generative model, graph-based model, and co-training model.

4.2.1 Self-training method:

This may be the oldest approach of semisupervised learning [84-87]. A classifier, $f$, is trained using labeled data (training data), $(X_l, Y_l)$. Then it predicts label, $y_j$ of unlabeled data, $x_j \in X_u$. Among the predicted unlabeled data, $(X_u, f(X_u)) = \{(x_i+1:n, f(x_{i+1:n}))\}$, the most reliable or confident predicted data are selected and added to existing labeled data. Reliable or confident predicated data means the probability of correct label prediction is high. This procedure can be repeated if necessary.
4.2.2 Co-training method:

Co-training method splits $m$-dimensional feature space into subfeature sets, and the same classifier may be used for the subfeature sets [88-92]. For example, in the case of two subfeature sets, original data can be divided into $X_n = \{x_{1:n}^{(1:p)}, x_{1:n}^{(p+1:m)}\}$. In this example, consecutive subfeatures are selected, but subsets can be made out of randomly chosen subfeature sets. Then the two subfeature sets are used to train each classifier, $f^{(1:p)}$ and $f^{(p+1:m)}$. Finally, the most agreeable predicted unlabeled data are chosen with a rule (e.g., majority voting) among two predicted unlabeled data

$\left(X_u^{(1:p)}, f\left(X_u^{(1:p)}\right)\right)$ and $\left(X_u^{(1+p:m)}, f\left(X_u^{(1+p:m)}\right)\right)$.

4.2.3 Generative model:

This method assumes data can be modeled mathematically such as a Gaussian distribution model [93-95]. For example, parameters of distribution (i.e., mean and standard deviation) can be estimated from a labeled data set. Expectation-maximization (EM) is one of the most popular generative algorithms [96, 97]. Hidden Markov model (HMM) is also a popular model [98, 99]. Generative model has a clear probabilistic framework, and often results in excellent performance if the probabilistic assumption is correct. However, the danger of this model is that if data does not follow the assumed distribution, a worse classifier will be built. For example, distribution of high dimensional data is difficult to estimate, and the generative model may not be a suitable method for those data.
4.2.4  **Graph-based method:**

This method is based on graph construction [100-102]. Nodes and weights can be made from labeled and unlabeled data. Then a classic graph-based classifier can be used to build a classifier. Graph-cut [103-105] and manifold regularization are representative methods [38, 106]. Graph-based method has clear mathematical framework, and performance will be excellent if data is well modeled by an appropriate graph.

Besides these, there are specialized SVMs of semisupervised learning [107, 108]. They may work better if the original version of SVM can perform well for a given data. However, unlike the classic SVM, they can be stuck in local optima, and it is more difficult to optimize the solution.

4.3  **Proposed method**

In this study, a new self-training method for hyperspectral images was proposed. The selected classifier was SVM because of its robust performance for high dimensional data, but the method can be extended to any classifier. The key of self-training method is to generate an additional confident training set from unlabeled data. Two strategies were proposed to use. The first one was majority voting (MV), and it used a majority filter to consider a spatial relationship in an image, assuming neighboring pixels share the same label. The second one was weighted majority voting (WMV). Equation 4.2 shows how to calculate a weight of a pixel, \( x_i \) with regard to a center pixel, \( x_c \).

\[
w_i = e^{-\frac{\|x_i - x_c\|^2}{\sigma^2}} \tag{4.2}
\]

where \( \sigma^2 \) is a width parameter.
MV utilizes a spatial relationship while WMV adds spectral consideration to the spatial relationship of MV. With an $N \times N$ window, the confidence level (CL) for MV and WMV was calculated as

\[
CL_{mv} = \frac{\text{the number of the most frequent label}}{N \times N} \quad (4.3)
\]

\[
CL_{wmv} = \frac{\sum_{i} w_i}{N \times N} \quad (4.4)
\]

respectively, where $w_i$ is the weight of pixel $x_i$, which has the label with the most frequent occurrence in the window. Here, $CL \in [0,1]$. Figures 4.3 and 4.4 illustrate how to get majority value (the most frequent value), and the corresponding CL. CL is calculated for a center pixel only when the center pixel is from a majority label; otherwise, it is set to zero.

Figure 4.2  Majority voting and confidence level (non-majority center pixel).

*(a) $3 \times 3$ majority filter. (b) Center pixel changes to the most frequent value, 3. (c) Confidence level of majority value at the center pixel is set to 0 because it is not a majority label.
Figure 4.3  Majority voting and confidence level (majority center pixel).

*(a) 3 × 3 majority filter. (b) Center pixel changes to the most frequent value, 3. (c) Confidence level of majority value at the center pixel is $\frac{7}{9}$ by Equation 4.3.
Figure 4.4 Weighted majority voting and confidence level (majority center pixel)

*(a) Original pixels. (b) Weights by Equation 4.2. (c) Labels from SVM output. (d) Majority voting.

The self-training procedure is presented in Fig. 4.5. If CL is greater than a threshold $\tau$, the most frequent label is chosen as reliable unlabeled data and added to the set of labeled data. Finally, SVM is trained again with the new labeled samples along with existing labeled data, and then a new SVM classifier is applied to the data. For a fixed number of labeled data from each class (i.e., 5), the effect of adding selected unlabeled data will be explored based on the number of added unlabeled data to existing labeled data. Fig. 4.6 summarizes the algorithm.

For MV and WMV, two strategies were used for performance assessment. A high amount of confident unlabeled data were selected from the first strategy if their CLs
exceeded threshold. In other words, enough unlabeled data could be chosen. For the second case, a fixed number of unlabeled data were selected to compare the performance of MV and WMV. The selection could be controlled by a window size of a majority filter and a threshold of confidence level, $\tau$.

![Diagram](image)

**Figure 4.5** Procedure for semisupervised classification.

Fig. 4.5 illustrates how semisupervised classification is performed. First, a small number of data are labeled from each class. In this study, 5, 10, 15, and 20 samples from each class were assumed labeled.
4.4 Hyperspectral image data

Four hyperspectral images were considered for this study: Indian Pines, Salinas A-scene, Pavia Centre, and Pavia University. The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) was used to obtain Indian Pines and Salinas A-scene. Pavia Centre and Pavia University were acquired from the ROSIS sensor. The image data are available at

4.4.1 Salinas A-scene

Salinas A-scene is a small portion of Salinas’s image. Salinas was collected by the AVIRIS sensor over Salinas Valley, California, and it has high spatial resolution (3.7m). Salinas’s ground truth consists of 16 classes, and Salinas A-scene contains 6 classes with 86×83 pixels and 224 bands.

4.4.2 Indian Pines

Indian Pines data was acquired over the Indian Pines region in Northwestern Indiana by the AVIRIS sensor. It contains 200 spectral bands with the wave-length range from 0.4 to 2.5µm, and the spatial resolution is 20m. It has 16 classes with 145×145 pixels. The Indian Pines scene contains two-thirds agriculture and one-third forest or other natural perennial vegetation. There are two major dual lane highways and a rail line as well as some low density housing, other built structures, and smaller roads. Since the scene was taken in June with crops present. Corn and soybeans were in early stages of growth with less than 5% coverage. The ground truth available includes 16 classes, which are not all mutually exclusive.

4.4.3 Pavia Centre and Pavia University

Pavia Centre and Pavia University were acquired by the ROSIS sensor over Pavia, northern Italy. The spatial resolution is 1.3m with 1096×1096 pixels for Pavia Centre and 610×610 pixels for Pavia University. The number of spectral bands is 102 for Pavia Centre and 103 for Pavia University with 6 classes.
4.5 Results and discussion

Two cases were examined for MV and WMV. The first case explored how MV and WMV enhanced the performance over supervised SVM with a different number of labeled data. In this case, all reliable and confident unlabeled data were added to a labeled data set. The second case compared the performance between MV and WMV when a fixed number of confident unlabeled data was added to a labeled data set.

For the first experiment, 5, 10, 15 and 20 labeled samples from each class were assumed available, and overall accuracy (OA) was calculated by averaging over 10 runs. In this study, labeled data were randomly selected from the ground truth. For each labeled data, all possible confident unlabeled data were chosen from an original SVM output to expand a labeled data set by the self-training algorithm in Fig. 4.6.

Table 4.1 Windows size and threshold for confidence level.

<table>
<thead>
<tr>
<th></th>
<th>Indian Pines</th>
<th>Salinas A-scene</th>
<th>Pavia University</th>
<th>Pavia Centre</th>
</tr>
</thead>
<tbody>
<tr>
<td>MV Window size</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>WMV Window size</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>MV Threshold $\tau$</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.85</td>
</tr>
<tr>
<td>WMV Threshold $\tau$</td>
<td>0.88</td>
<td>0.88</td>
<td>0.85</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 4.1 shows the parameter set for the experiment, and all parameters were heuristically decided. Obviously, for a homogeneous image scene such as Salinas-A and Indian Pines, a larger window size could be used. Figures 4.7 through 4.14 show the classification results for each image. Overall, the semisupervised approach with MV and WMV yielded increased performance for all images, and WMV also showed marginal improvement over MV. In the case of Indian Pines from Figures 4.7 and 4.8, WMV and
MV produced significant improvement (about 3–6%). For Salinas-A scene in Figures 4.9 and 4.10, it consists of simple classes, and SVM itself yielded excellent performance with a small number of labeled data (training data). It had about 0.8–3% performance increases while it had the greatest increase for the smallest labeled data (5 samples from each class). About 0–5% performance gain could be observed from Pavia University from Figures 4.11 and 4.12, and WMV displayed marginal improvement over MV. For the last image of Pavia Centre in Figures 4.13 and 4.14, 6–9% improvement was observed for both MV and WMV.

Figure 4.7 Classification images of Indian Pines.

*(a) Ground truth, (b) SVM (47.1%), (C) MV (49.5%) and (d) WMV (49.8%) for 5 labeled samples from each class.
Figure 4.8 Classification result of Indian Pines.
Figure 4.9  Classification images of Salinas-A scene.

*(a) Ground truth, (b) SVM (95.1%), (C) MV (98.0%) and (d) WMV (98.4%) for 5 labeled samples from each class.*
Figure 4.10  Classification result of Salinas-A scene.
Figure 4.11  Classification images of Pavia University.

*(a) Ground truth, (b) SVM (55.7%), (C) MV (55.7%) and (d) WMV (56.8%) for 5 labeled samples from each class.
Figure 4.12  Classification result of Pavia University.
Figure 4.13  Classification images of Pavia Centre.

*(a) Ground truth, (b) SVM (78.6%), (C) MV (86.1%) and (d) WMV (86.5%) for 5 labeled samples from each class.*
Table 4.2 shows the number of confident unlabeled data were added to a labeled data set. The number of selected confident data was decided based on a threshold, $\tau$. If a lower threshold was set, a lot of unreliable unlabeled data were selected, and a poor classifier was built. However, a higher threshold would discard some of the reliable unlabeled data. Therefore, the best threshold should be the one achieving the balance.

Fig. 4.15 shows the comparison between MV and WMV when the same number of confident unlabeled data was added to the number of labeled data set (10 samples from each class). The result of WMV from Indian Pines only yielded significant improvement (about 1–4 %) over MV, while there was marginal improvement (less than 1%) in other images. WMV was expected to have better results since it considered both a spatial and a
spectral relationship among local neighbors. An additional spectral relationship to spatial consideration slightly impacted the results. Sometimes, a spectral relationship might not be a significant addition to local spatial relationship. Interestingly, performance increase between MV and WMV from Indian Pines is relatively large compared to the others. This may come from the fact that additional spectral consideration helped to effectively select reliable unlabeled samples in the case of Indian Pines.

In summary, self-training with MV and WMV based on unlabeled sample selection yielded significant improvement in terms of OA. WMV outperformed MV in the selection of more reliable unlabeled samples.

Table 4.2 Selected unlabeled data.

<table>
<thead>
<tr>
<th>Total samples</th>
<th>Indian Pines</th>
<th>Salinas A-scene</th>
<th>Pavia University</th>
<th>Pavia Centre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>MV</td>
<td>WMV</td>
<td>MV</td>
<td>WMV</td>
</tr>
<tr>
<td>5</td>
<td>627</td>
<td>714</td>
<td>3111</td>
<td>3500</td>
</tr>
<tr>
<td>10</td>
<td>1023</td>
<td>820</td>
<td>3251</td>
<td>3551</td>
</tr>
<tr>
<td>15</td>
<td>1088</td>
<td>1081</td>
<td>3380</td>
<td>3611</td>
</tr>
<tr>
<td>20</td>
<td>1411</td>
<td>1258</td>
<td>3431</td>
<td>3660</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MV</th>
<th>WMV</th>
<th>MV</th>
<th>WMV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11480</td>
<td>11484</td>
<td>12279</td>
<td>12265</td>
</tr>
<tr>
<td></td>
<td>3611</td>
<td>3660</td>
<td>12220</td>
<td>12227</td>
</tr>
<tr>
<td></td>
<td>12591</td>
<td>12566</td>
<td>62036</td>
<td>61198</td>
</tr>
</tbody>
</table>
Figure 4.15  Comparison of MV and WMV for the same number of added confident unlabeled data.
CHAPTER V
CONCLUSION AND FUTURE WORK

In this dissertation, statistical analysis and machine learning methods were applied to radar and hyperspectral images. In the first study, in situ soil property data and SAR data were investigated to discover a statistical relationship. Usually, the two have been separate research areas. Until now, most of the studies examined radar response to soil moisture on the soil and mapped the soil moisture for a large area. However, more research interest has been drawn to how to relate radar response to soil physical properties. Radar can economically cover a large area, while in situ soil property data can cover a small area with more accurate information. If a certain relationship between two methods can be well established, the impact will be great, because it may help efficiently map the soil property data for a large area. The following soil parameters were used in the analysis: soil texture (sand and clay fraction), penetration resistance (sleeve friction and cone tip resistance), saturated hydraulic conductivity, field capacity, permanent wilting point, and porosity. For the radar data used in this research, HH, VV and HV polarization from SAR were considered.

The results of statistical analysis showed weak but considerable correlation between the cross-polarized (HV) radar backscatter coefficients and most of these properties. This observation demonstrates the possibility of using radar to estimate various soil properties over a large area economically. Among the soil properties
examined, sand fraction, cone tip resistance, and saturated hydraulic conductivity exhibited negative correlations with HV channel, while clay fraction, field capacity, permanent wilting point, and porosity showed positive correlations with HV channel. However, more studies are needed to further investigate and quantify the possible correlation.

The second study focused on efficient feature extraction and classification of SAR images. Efficient spatial feature extraction approaches were investigated for a levee slide classification using SAR images. GLCM feature extraction performed well as expected but had limitations of high computation cost and storage needs, and the entire process is difficult to implement in parallel. For small datasets, this may not be crucial, but as data size increases, efficient feature extraction is needed. In this work, an average filter and a majority filter were shown to have much lower computational cost and could be easily implemented in parallel to further reduce computing time. Even though their classification performance did not exceed GLCM (in some feature combinations), they can be used for fast screening before more complicated methods are applied on a small selected area. The spatial features and feature combinations investigated in this study may be useful for early abnormality screening with a method of anomaly detection.

The nonlinear band generation process could dig out the information embedded in the original data, and it could be used to maximize class separability, creating three different polarization bands with new spatial information for discrimination. For future work, parallel implementation of the proposed methods and anomaly detection for screening of potential levee slide will be conducted.
The last study explored semisupervised learning. There are abundant unlabeled data, but it is difficult to label enough data to understand whole data. The semisupervised learning approach is valuable for this real life problem. The key of semisupervised learning is to extract as much information as possible from a few labeled data to distinguish classes by relating them to unlabeled data in one way. In this work, a new self-training method was developed in conjunction with majority voting (MV) or weighted majority voting (WMV) based unlabeled sample selection. Basically, no assumption was made on the hyperspectral images except that a limited number of labeled data from each class was assumed to be available. This situation can easily be found in real life. The experimental results showed the possibility of significant improvement for all images used in the study as more reliable and confident unlabeled data were added to labeled data. When the same number of confident unlabeled data was added, the improvement of WMV over MV was marginal with exception of Indian Pines. Future work includes algorithm assessment using different classifiers and comparison with other types of semisupervised methods.


APPENDIX A

SCATTER PLOT BETWEEN SAR DATA AND IN SITU SOIL PROPERTIES
A.1 Scatter plot between SAR and in situ soil properties on the river side and the land side.

The following figures show scatter plots between SAR backscatter and in situ soil properties. Correlation can be visually inspected from these scatter plots. A correlation cannot be seen from the land side. However, possible correlations can be observed from the river side. In Fig. A.6, cone tip resistance, $q_c$, clearly displays a negative relationship with HV channel while sleeve friction, $q_s$, fails to show a relationship. In Fig. A.7, clay fraction has a positive relationship while sand fraction has a negative relationship with the HV channel. Interestingly, HH and VV channels also show similar patterns. In Figures A.8 through A.10, saturated hydraulic conductivity, $K_{sat}$, shows an apparent negative relationship with HH, HV and VV channels, while field capacity, $f_c$, and permanent wilting point, $pwp$, display a positive relationship with HH, HV and VV channels. HH and VV channels display similar patterns as HV channel again while their correlation coefficients are very small compared to HV channel. Extreme points or outliers may make the difference.
Figure A.1  Scatter plot of $q_s$ and $q_e$ with HH, HV and VV channels on the land side.

*Histograms on the diagonal show distribution of each data.
Figure A.2  Scatter plot of clay and sand fraction with HH, HV and VV channels on the land side.
**Figure A.3**  Scatter plot of $K_{\text{sat}}$, and $f_c$ with HH, HV and VV channels on the land side.
Figure A.4 Scatter plot of pwp and saturation with HH, HV and VV channels on the land side.
Figure A.5  Scatter plot of n with HH, HV and VV channels on the land side.
Figure A.6  Scatter plot of $q_s$ and $q_c$ with HH, HV and VV channels on the river side.
Figure A.7 Scatter plot of clay and sand fraction with HH, HV and VV channels on the river side.
Figure A.8  Scatter plot of $K_{\text{sat}}$ and $f_c$ with HH, HV and VV channels on the river side.
Figure A.9 Scatter plot of pwp and saturation with HH, HV and VV channels on the river side.
Figure A.10  Scatter plot of $n$ with HH, HV and VV channels on the river side.